มหาวิทยาลัยมหิดล

รายงานวิจัยฉบับสมบูรณ์

โครงการ แบบจำลองทางคอมพิวเตอร์เพื่อใช้ในการออกแบบและพัฒนาสายเปบไทด์ที่มีประสิทธิภาพเพื่อใช้ในการทำลายเนื้องอก

วัชระ ชุ่มบัวตอง
มหาวิทยาลัยมหิดล

พฤษภาคม 2563

สัญญาเลขที่ MRG6180226

รายงานวิจัยฉบับสมบูรณ์

โครงการ แบบจำลองทางคอมพิวเตอร์เพื่อใช้ในการออกแบบและพัฒนาสายเปบไทด์ที่มีประสิทธิภาพเพื่อใช้ในการทำลายเนื้องอก

สนับสนุนโดยสำนักงานกองทุนสนับสนุนการวิจัยและมหาวิทยาลัยมหิดล

บทคัดย่อ

Project Code : MRG6180226
(รหัสโครงการ)

Project Title : แบบจำลองทางคอมพิวเตอร์เพื่อใช้ในการออกแบบและพัฒนาสายเปปไทด์ที่มีประสิทธิภาพเพื่อใช้ในการทำลายเนื้องอก

Investigator : ผศ. ดร. วัชระ ชุ่มบัวตอง เทคนิคการแพทย์ มหาวิทยาลัยมหิดล
Mentors : รศ. ดร. ชนินทร์ นันทเสนามาตร์ เทคนิคการแพทย์ มหาวิทยาลัยมหิดล
รศ. ดร. สุรีรัตน์ พรธาดาวิทย์ เทคนิคการแพทย์ มหาวิทยาลัยมหิดล

E-mail Address : watshara.sho@mahidol.ac.th

Project Period : 2 ปี (2 พฤษภาคม พ.ศ. 2561 ถึง 1 พฤษภาคม พ.ศ. 2563)
(ระยะเวลาโครงการ)

บทคัดย่อ

สายเปปไทด์ที่มีประสิทธิภาพเพื่อใช้ในการทำลายเนื้องอกได้ถูกนำมาใช้ในการรักษาโรคมะเร็ง เนื่องจากสายเปปไทด์ตัวนี้มีค่าความเป็นพิษต่ำและมีความจำเพาะต่อเซลล์มะเร็งมากกว่ายาชนิดอื่น แต่อย่างไรก็ตามวิธีการทางทดลองเพื่อใช้ในการระบุสายเปปไทด์ที่มีประสิทธิภาพเพื่อใช้ในการทำลายเนื้องอกนั้นใช้เวลาและเสียค่าใช้จ่ายค่อนข้างสูง ดังนั้นจึงมีความจำเป็นที่ต้องพัฒนาตัวแบบการทำนายทางคอมพิวเตอร์เพื่อใช้ในการวิเคราะห์ข้อมูลดังกล่าวอย่างรวดเร็วและมีประสิทธิภาพ โดยงานวิจัยชิ้นนี้ผู้วิจัยจึงได้มีการพัฒนาตัวแบบการทำนายทางคอมพิวเตอร์ที่สามารถวิเคราะห์ผ่านทางสายลำดับเปปไทด์เพื่อใช้ในการทำนายและวิเคราะห์สายเปปไทด์ที่มีประสิทธิภาพเพื่อใช้ในการทำลายเนื้องอกที่มีชื่อว่า ACPred โดยพัฒนาจากซัพพอร์ตเวกเตอร์แมชชิน และ แรนด้อมฟอร์เรส จากผลการทำนายโยใช้ jackknife test พบว่า ACPredให้ค่าความถูกต้อง 95.61% ยิ่งไปกว่านั้นเพื่อความสะดวกแก่นักวิจัยที่ได้ทำแลปผู้วิจัยจึงได้สร้างเวปไซส์ http://codes.bio/acpred/

คำหลัก : สายเปปไทด์ที่มีประสิทธิภาพเพื่อใช้ในการทำลายเนื้องอก, เปปไทด์ที่ใช้ในการรักษา, การจำแนก, การเรียนรู้จากแมทชีน, แรนด้อมฟอร์เรสท์

Abstract

Project Code : MRG6180226

Project Title : *In silico* approaches for designing and developing highly effective anticancer peptides

Investigator : Assist. Prof. Dr. Watshara Shoombuatong Faculty of Medical Technology, Mahidol University

Mentors : Assoc. Prof. Dr. Chanin Nantasenamat Faculty of Medical Technology, Mahidol University
Assoc. Prof. Dr. Sureerut Porntadavit Faculty of Medical Technology, Mahidol University

E-mail Address : watshara.sho@mahidol.ac.th

Project Period : 2 years (2 May 2018 – 1 May 2020)

Abstract:

Anticancer peptides (ACPs) have emerged as a new class of therapeutic agent for cancer treatment due to their lower toxicity as well as greater efficacy, selectivity and specificity when compared to conventional small molecule drugs. However, the experimental identification of ACPs still remains a time-consuming and expensive endeavor. Therefore, it is desirable to develop and improve upon existing computational models for predicting and characterizing ACPs. In this study, we present a bioinformatics tool called the ACPred, which is an interpretable tool for the prediction and characterization of the anticancer activities of peptides. ACPred was developed by utilizing powerful machine learning models (support vector machine and random forest) and various classes of peptide features. It was observed by a jackknife cross-validation test that ACPred can achieve an overall accuracy of 95.61% in identifying ACPs. Finally, for the convenience of experimental scientists, the ACPred web server was established and made freely available online at http://codes.bio/acpred/.

Keywords : Anticancer peptide, therapeutic peptide, classification, machine learning, random forest, support vector machine

Executive summary

Previously, many studies have reported that anticancer peptides (ACPs) display several advantages over small molecule drugs. For instance, they exhibit a shorter half-life that consequently leads to decreased probability of resistance, lowered toxicity (not harming normal cells as much), higher specificity, good solubility as well as good tumor penetration capabilities thereby indicating their great potential in cancer therapy. Accurate prediction of ACPs is of great importance for the exploration of their mechanism of action and for the development of therapeutic ACPs. As the experimental identification and development of novel ACPs is a very time-consuming and labor-intensive process, therefore bioinformatics tools are needed for effective analysis of the available big data on existing peptides so as to allow the identification of novel ACPs while also shedding light about their mechanism of action. There are four major steps that are involved in the development of bioinformatics tools as follows: (i) data preparation, (ii) model construction (feature extraction and feature combination), (iii) feature importance analysis and (iv) web server construction.

Introduction

Cancer remains one of the leading causes of death worldwide, accounting for an estimated 18.1 million new cases and resulting in 9.6 million deaths in 2018 (1). Of these, the most common cancer types causing mortality worldwide include lung, colorectal, stomach, liver and breast cancer, respectively (1, 2). The global burden of cancer has risen dramatically in the last few decades thereby making the prevention of cancer the most significant and challenging public health concern of the 21st century (3). In spite of the many advanced clinical methods available for the treatment of cancer (e.g., chemotherapy, radiation therapy, hormonal therapy, etc.) their recurrence rate still remains high. In addition, anti-cancer drugs are known to exert their therapeutic effects by killing off normal cells and tissues, which exacerbates the immunodeficiency of patients. Thus, the discovery and development of novel anti-cancer drugs are crucial for decreasing premature death as well as for increasing the survival of many populations. In this regard, peptide-based therapeutics have gained much attention as a new drug class since they are considered to be relatively safe, highly selective, possess good tolerability, lower production cost, easy to modify and synthesize while also exhibiting promising pharmacological profiles (4-7). In the last decade, more than 7000 naturally occurring peptides have been discovered and they are found to exhibit a wide range of bioactivities including tumor homing, antihypertensive, antiparasitic, antiviral, antiangiogenic, antibiofilm, antimicrobial and anticancer (8-10). Currently, 60 peptide-based drugs have been FDA-approved (11) with preclinical and clinical studies in the pipeline for another 150 peptides (6).

Anticancer peptides (ACPs) are a subset of antimicrobial peptides (AMPs) that are have been shown to exhibit anticancer activities. These peptides are typically made up of short (5–30 amino acids) cationic residues that mostly adopts the α-helical structure (e.g., BMAP-27, BMAP-28, Cecropin A and B, LL-37, Magainins etc.) or folds into a β-sheet (e.g., Defensins, Lactoferricin and Tachylepsin I) while extended linear forms (e.g., Tritrpticin and Indolicidin) have also been reported (12, 13), as shown in Figure 1.

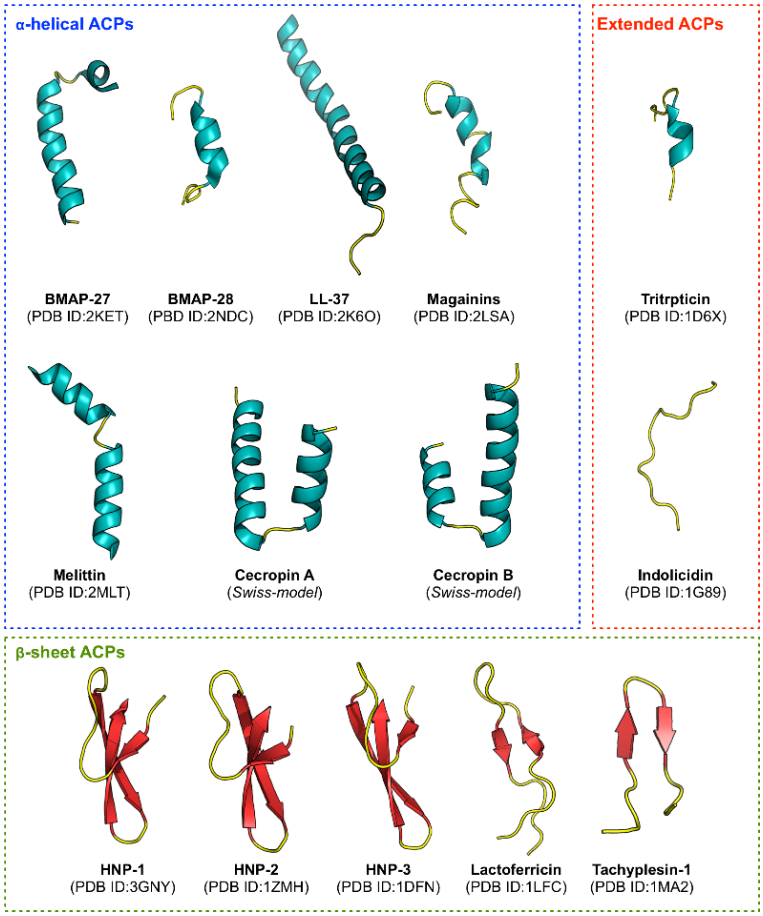

Figure 1. Overview of the structural diversity of three classes of anticancer peptides. Each structure is labeled by its common name followed by the Protein Data Bank (PDB ID) in parenthesis on the subsequent line. In cases where the PDB ID was not available, the SWISS-MODEL server (available at: https://swissmodel.expasy.org/) was used to construct the structure.

In comparison to normal cells, the cancer cell membrane exhibits different properties (i.e., larger surface area due to the higher number of microvili, high negative charge of the cell membrane, higher membrane fluidity, etc.) (14-17). Thus, the inherent cationic property of ACPs is responsible for the electrostatic interactions between ACPs and the anionic cell membrane of cancer cells, and hence allow the selective killing of cancer cells (18). Previously, many studies have reported that ACPs display several advantages over small molecule drugs. For instance, they exhibit a shorter half-life that consequently leads to decreased probability of resistance, lowered toxicity (not harming normal cells as much), higher specificity, good solubility as well as good tumor penetration capabilities thereby indicating their great potential in cancer therapy (19-22).

Accurate prediction of ACPs is of great importance for the exploration of their mechanism of action and for the development of therapeutic ACPs. As the experimental

identification and development of novel ACPs is a very time-consuming and labor-intensive process, therefore bioinformatics tools are needed for effective analysis of the available big data on existing peptides so as to allow the identification of novel ACPs while also shedding light about their mechanism of action. Recently, many sequence-based computational methods, for example, AntiCP (23), Hajisharifi et al.'s method (24), ACPP (25), iACP (26), Li and Wang's method (27), iACP-GAEnsC (28), MLACP (29), SAP (30) and TargetACP (31), have been developed using a wide range of machine learning methods and peptide features as summarized in Table 1. Although, iACP-GAEnsC and TargetACP yielded encouraging performances with reasonably high prediction accuracies, the overall utility of these two methods is limited in terms of interpretability and practical utility thereby warranting further improvements (32).

Table 1. Summary of existing methods for predicting anticancer peptides.

Method (Year)	Classifier [a]	Sequence Features [b]	Interpretable	Web Server
AntiCP (2013)	SVM	AAC, DPC, BP	No	✓
Hajisharifi et al. (2014)	SVM	PseACC, LAK	No	
ACPP (2015)	SVM	PRM	No	✓ [c]
iACP (2016)	SVM	g-gap DPC	**Yes**	✓
Li and Wang (2016)	SVM	AAC, RACC, acACS	No	
iACP-GAEnsC (2017)	Ensemble method	Pse-g-gap DPC, Am-PseAAC, RACC	No	
MLACP (2017)	RF	AAC, ATC, DPC, PCP	**Yes**	✓
SAP (2018)	SVM	g-gap DPC	No	
TargetACP (2018)	SVM	CPSR, SAAC, PsePSSM	No	
ACPred (this study)	SVM	AAC, DPC, PCP, PseAAC, Am-PseAAC	**Yes**	✓

[a] RF: random forest, SVM: support vector machine. [b] AAC: amino acid composition, ATC: atomic composition, acACS: auto covariance of the average chemical shift, Am-PseAAC: amphiphilic pseudo amino acid composition, BP: binary profile, CPDR: composite protein sequence representation, DPC: dipeptide composition, g-gap DPC: G-Gap dipeptide composition, LAK: local alignment kernel, PCP: Physicochemical properties, PseACC: Pseudo amino acid composition, Pse-g-gap DPC: Pseudo G-Gap dipeptide composition, PRM: protein relatedness measure, RACC: reduce amino acid composition, SAAC: split amino acid composition. [c] The web server is not accessible.

One of the major limitations of iACP-GAEnsC and TargetACP was that there was no web server provided for these two methods. Therefore, their utility is limited to researchers with informatics background who can develop in-house prediction models. In addition, the underlying mechanisms of the investigated bioactivity for these two methods affords limited interpretability for experimental scientists. Owing to the heterogeneous computational architecture and design of the study as well as differences in the benchmark dataset used, it is not easy to identify and assess which methods and features provided the greatest contribution to good prediction. Motivated by aforementioned issues presented in previous models (32-36), in order to establish an interpretable ACP predictor we propose a systematic effort for the prediction and analysis of anticancer activities of peptides called the ACPred. The overall framework of ACPred is illustrated in Figure 2.

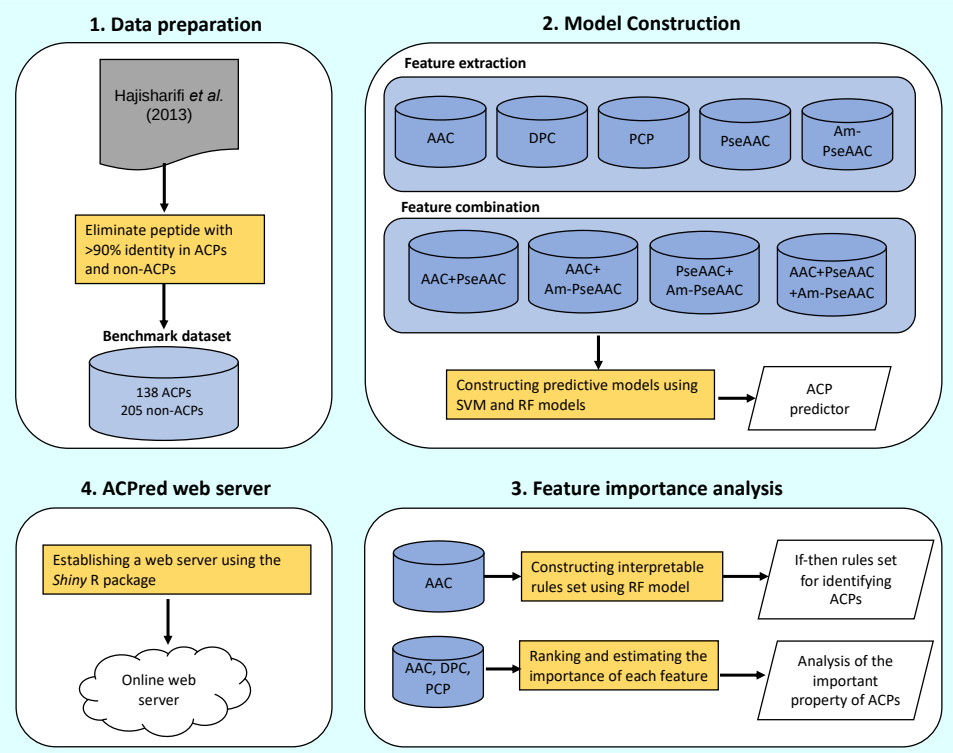

Figure 2. Schematic framework of ACPred.

Materials and Methods

As elaborated by a series of publications from our group (33, 34, 37-41), to develop a useful and interpretable sequence-based predictor, we should consider the following five procedures: (i) establish or select a reliable benchmark dataset for training

and validating the predictor and also calculating the dataset modelability (MODI) to evaluate the feasibility of obtaining a predictor with significant predictive power; (ii) represent protein or peptides sequences that can truly reflect their intrinsic properties to be predicted; (iii) introduce or develop an easily interpretable predictor that can provide a better understanding of the biological system; (iv) evaluate the predictive power of the predictor by performing rigorous cross-validation methods; (v) develop a user-friendly web-server that can easily reveal the desired results without needing to go through the mathematical and statistical details. Figure 1 presents the workflow of the systems used to predict and analyze THPs, which involves dataset compilation, feature extraction, data splitting, model construction and evaluation and model post-analysis.

Benchmark datasets

A high-quality benchmark dataset can ensure the reliability of the proposed method. Therefore, the investigated dataset was retrieved from the work of Hajisharifi et al. *(24)*, which contains 192 ACPs and 215 non-ACP. As to guarantee the quality of the benchmark dataset, the following steps were considered. Firstly, in order to avoid a dataset containing redundant peptides, the CD-HIT program was used to eliminate ACPs and non-ACPs with >90% identity. Secondly, ACPs and non-ACP containing special characters, such as X and U, were removed. After such screening procedures, a set of 138 ACPs and 205 non-ACPs was attained for prediction and analysis of ACPs.

Feature representation

In order to build an effective prediction model, it is necessary to represent each peptide as numerical vectors that can encompass the perspective of its biophysical and biochemical properties. For peptide or protein sequences representation, amino acid composition (AAC), dipeptide composition (DPC), physicochemical property (PCP), pseudo amino acid composition (PseAAC) and amphiphilic pseudo amino acid composition (Am-PseAAC) were employed.

Modelability indices (MODI)

After feature representation, the Modelability indices (MODI) (42, 43) were checked using an in-house developed R code. Prior to development of an empirical model, the calculations of MODI values help the modelers to evaluate the feasibility of obtaining models with significant predictive power as well as to determine the types of

features that can produce a predictor with a guarantee on reliability. A threshold value of 0.65 separates the dataset into modelable or non-modelable.

Classifier selection

In this study, we employed two successful machine learning algorithms namely random forest (RF) and support vector machine (SVM). Previously, these approaches have been successfully used in the prediction of various functions and properties of peptides and proteins (33, 36, 41, 44-49) as well as other biological or chemical entities (35, 40, 50-53).

Feature importance

The RF method is useful for evaluating feature importance with the use of the OOB set. In order to provide a better understanding of the tumor homing activity of peptides, herein, the RF models were utilized for the identification of informative features for each type of peptide sequential feature. This is due to its efficient and effective built-in feature importance estimator. Generally, in the system of RF model construction, two measures based on the mean decrease of the Gini index (MDGI) and prediction accuracy are available for ranking feature importance. It was reported by Calle and Urrea (54) that the Gini index provided more robust results as compared to accuracy, therefore we utilized the value of MDGI to rank the importance of each type of peptide sequential features.

Rule Extraction

This work presented an interpretable rule extraction of ACPs (IR-ACP) based on the RF method cooperating with AAC features for determining how the features work in combination within the learning process. A set of rules from an individual tree is derived from the root to the leaves. In this study, only 100 decision trees were used to extract if-then interpretable rules for explaining the prediction results by means of the *RRF*, *inTrees* and *xtable* packages in the R software (55-57). More details of the rule extraction process can be found in previous related works (56, 58).

Performance evaluation

Three testing methods consisting of (i) sub-sampling test (2-, 5- or 10-fold cross-validation), (ii) jackknife test or also known as leave-one-out cross-validation (LOO-CV), and (iii) independent (or external) testing dataset are often used to evaluate the

prediction performance. The sub-sampling test is one of the popular cross-validation methods for assess the predictive capability of a model. As elucidated in (59, 60) and demonstrated by Equation (50) of (61), among the three cross-validation methods, the jackknife test is considered as one of the most rigorous and objective methods for cross-validation in statistics and can provide a unique result for a given benchmark dataset. Thus, in this study, the 5-fold cross-validation (5-fold CV) and jackknife test were used to evaluate the prediction performance of our models. The former set (5-fold CV) makes use of data from the training set where the data set is separated into five subsets. Practically, one subset from a total of five subsets is left out as the testing set while the remaining are used for training the model. This process is repeated iteratively until all data samples have had the chance to be left out as the testing set. During the jackknifing process, each sample in the dataset (N samples) is left out as an independent sample (one sample) and the remaining sequences (N-1 samples) are used for training the model. This process is repeated iteratively until all samples have had the chance to be left out as the independent sample. For this kind of binary classification problem, the following set of metrics, i.e. accuracy (Ac), sensitivity (Sn), specificity (Sp) and Matthews coefficient correlation (MCC), are often utilized to determine the predictive power of the model. Furthermore, the receiver operating characteristic (ROC) curves were plotted to show the predictive capability of our models by the *pROC* package in the R software (62). The ROC curve represents the model behaviour of the true positive rate (sensitivity) against the false positive rate (1-specificity) in a visual way. The area under the ROC curve was calculated to quantitatively and objectively measure the performance of the proposed models. A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1.

Results and Discussion

Prediction performance

Tables 2 and 3 shows the performance comparison amongst different feature types with RF and SVM on the benchmark dataset via 5-fold CV and jackknife test, respectively. In addition, Figure 3 shows the receiver operating characteristic (ROC) curve of the five considered feature sets obtained from RF (top and bottom left) and SVM (top and bottom right) as evaluated by 5-fold CV (top left and right) and jackknife test (bottom left and right). As seen in Tables 2 and 3, SVM model with the Am-

PseAAC feature afforded the highest accuracy of 95.03% as evaluated by the jackknife test. Meanwhile, RF model with PseACC feature and SVM model with AAC feature performed well with the second and third the highest accuracy of 93.28% and 92.98%, respectively. These results showed that ACC, PseACC and Am-PseAAC were effective for ACP prediction. Since using a combination of various features might yield better prediction performance, four combinations of the three effective features were also considered. The highest accuracy and *MCC* of 95.61% and 0.91, respectively, was achieved by using SVM model cooperating with the combination feature of AAC and Am-PseAAC, while using the combination of the three effective features, i.e., ACC, PseACC and Am-PseAAC, performed slightly worse than the combination feature of AAC and Am-PseAAC with an accuracy and *MCC* of 94.74% and 0.89, respectively. In addition, our prediction results were well consistent with previous studies (27-29) and related studies (33, 34, 36, 63).

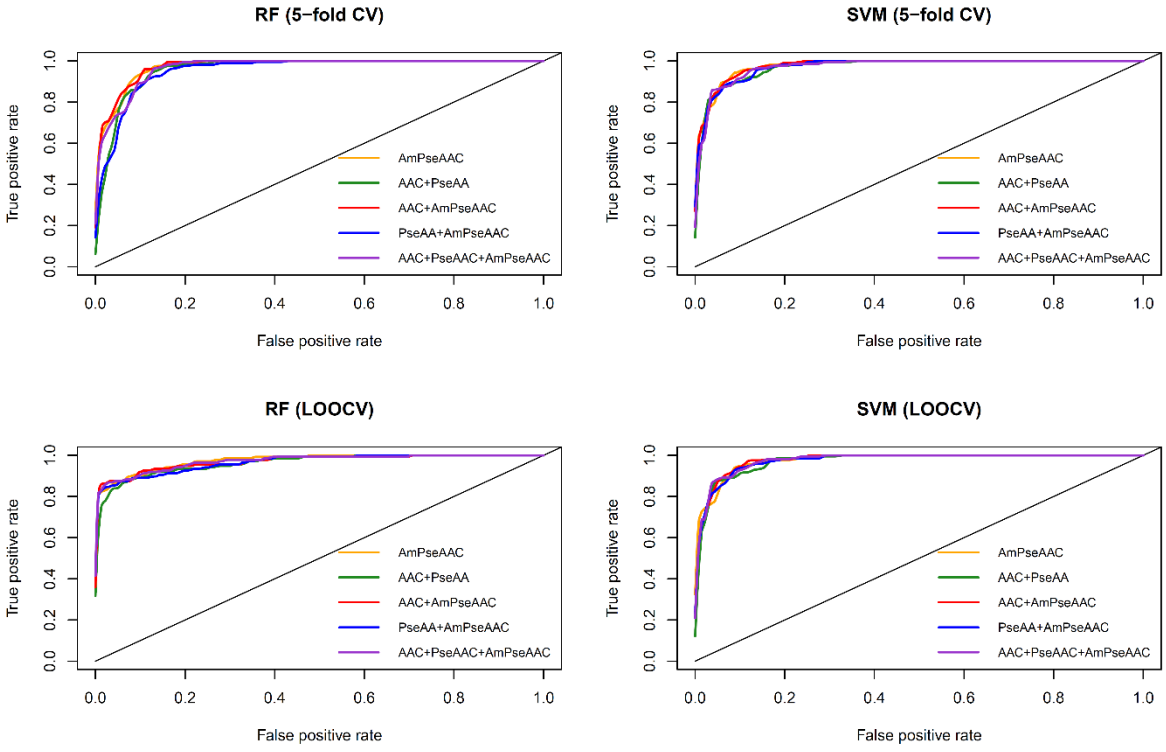

Figure 3. ROC curve of RF (top and bottom left) and SVM (top and bottom right) models as assessed by 5-fold cross-validation (top left and right) and jackknife test or leave-one-out cross-validation (bottom left and right).

Comparison of ACPred with the existing methods

As for performance comparisons on the benchmark dataset, it was found that ACPred was comparable with that of iACP-GAEnsC and TargetACP as indicated by

four statistical parameters. Moreover, ACPred yielded a greater prediction performance than the method proposed by Hajisharifi et al. and iACP. To the best of our knowledge, iACP-GAEnsC is the only method that utilizes the synthetic minority oversampling technique (SMOTE) (64) technique for constructing a balanced dataset from the original benchmark dataset *(24)*. From the perspectives of machine learning, it is not fair to directly compare prediction results between TargetACP and ACPred because TargetACP was trained and tested on the balanced dataset consisting of 205 ACPs and 205 non-ACPs. In order to perform a fair comparison, the herein proposed ACPred method was applied on the balanced dataset (called ACPred-modified), which was generated by means of the SMOTE technique. Herein, the oversampling method was used to add synthetic samples for the minority class (i.e., ACPs) presented in the benchmark dataset by setting the parameters of the number of nearest neighbours (*k*), the number of extra cases (oversampling) to add to the minority class (perc.over) and the number of cases to reduce (undersampling) from the majority class (perc.under) were set to 9, 50 and 300, respectively. As seen in Table 4, ACPred-modified was found to outperform Hajisharifi et al.'s method, iACP and iACP-GAEnsC with an improvement of 2%–4% and 3%–7% on *Ac* and *MCC*, respectively. However, it was observed that ACPred-modified performed slightly worse than TargetACP by approximately 2%–3% where ACPred-modified and TargetACP provided 97.56% *Ac*/0.95 *MCC* and 98.78% *Ac*/0.97 *MCC*, respectively. Nevertheless, TargetACP was constructed with a sophisticated design thus, it is not easy to identify and assess which features offer the most contribution to the prediction improvement. On the other hand, ACPred was designed in a systematic manner for prediction and characterization of anticancer activities of peptides. In addition, a user-friendly web server ACPred was developed to facilitate high-throughput prediction of ACP. Therefore, the proposed ACPred model could become a practical tool for predicting and interpreting the anticancer activity of peptides, or at least as a complementary tool to existing methods in the field.

Analysis of feature importance

The identification of feature importance from AAC and DPC descriptors can provide a better understanding on the biochemical and biophysical properties of anticancer activities of peptides. Previously, AAC and DPC features have been analyzed as to further gain insights on how to characterize therapeutic peptides *(29, 34, 36, 46)* and protein functions (33, 49, 65). In this study, the value of MDGI was adopted to rank

and estimate the importance of each AAC and DPC feature. Such information is derived from the analysis of the benchmark dataset that consists of 138 ACPs and 205 non-ACPs. Table 5 lists the percentage values of the 20 amino acids for both ACPs and non-ACPs along with amino acid composition difference between the two classes as well as their MDGI values. In addition, a heatmap showing the feature importance for DPC features is shown in Figure 4. From Table 5, it can be observed that the ten informative amino acids with the highest MDGI values are Lys, Cys, Gln, Arg, Ile, Leu, His, Pro, Glu and Gly (29.54, 19.71, 15.89, 11.42, 10.13, 7.84, 7.80, 7.37, 6.88 and 6.56, respectively). Meanwhile, Figure 4 shows that the 20 top-ranked dipeptides according to their MDGI value are KK, KI, CY, IK, AK, KV, IP, LK, TC, SC, CG, LF, FK, GL, FL, LR, PP, LG, IG and AQ.

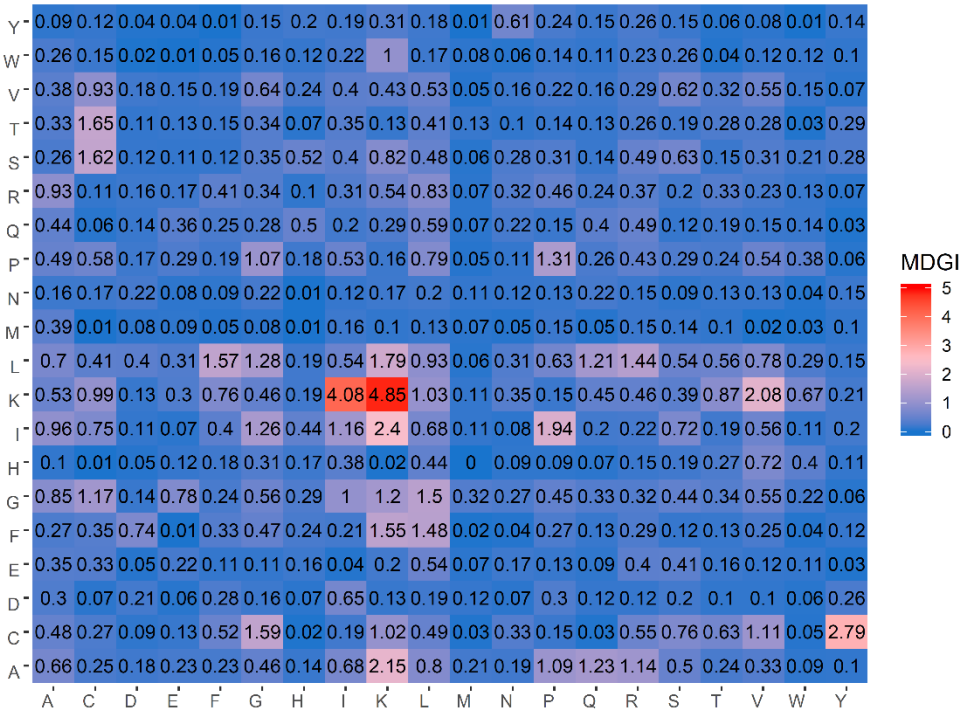

Figure 4. Heat map of the mean decrease of Gini index of dipeptide compositions.

ACPred web server

In order to make the prediction model presented herein a practical tool that can be widely used by the scientific community, we have constructed a web server called the ACPred using the best model as described in previous sections. The web interface has been established using the Shiny package under the R programming environment. The web server is freely accessible at http://codes.bio/acpred/. Screenshots of the ACPred

web server is shown in Figure 5 in which panel A shows the web server prior to submission of input data and panel B shown the web server after the prediction has been made.

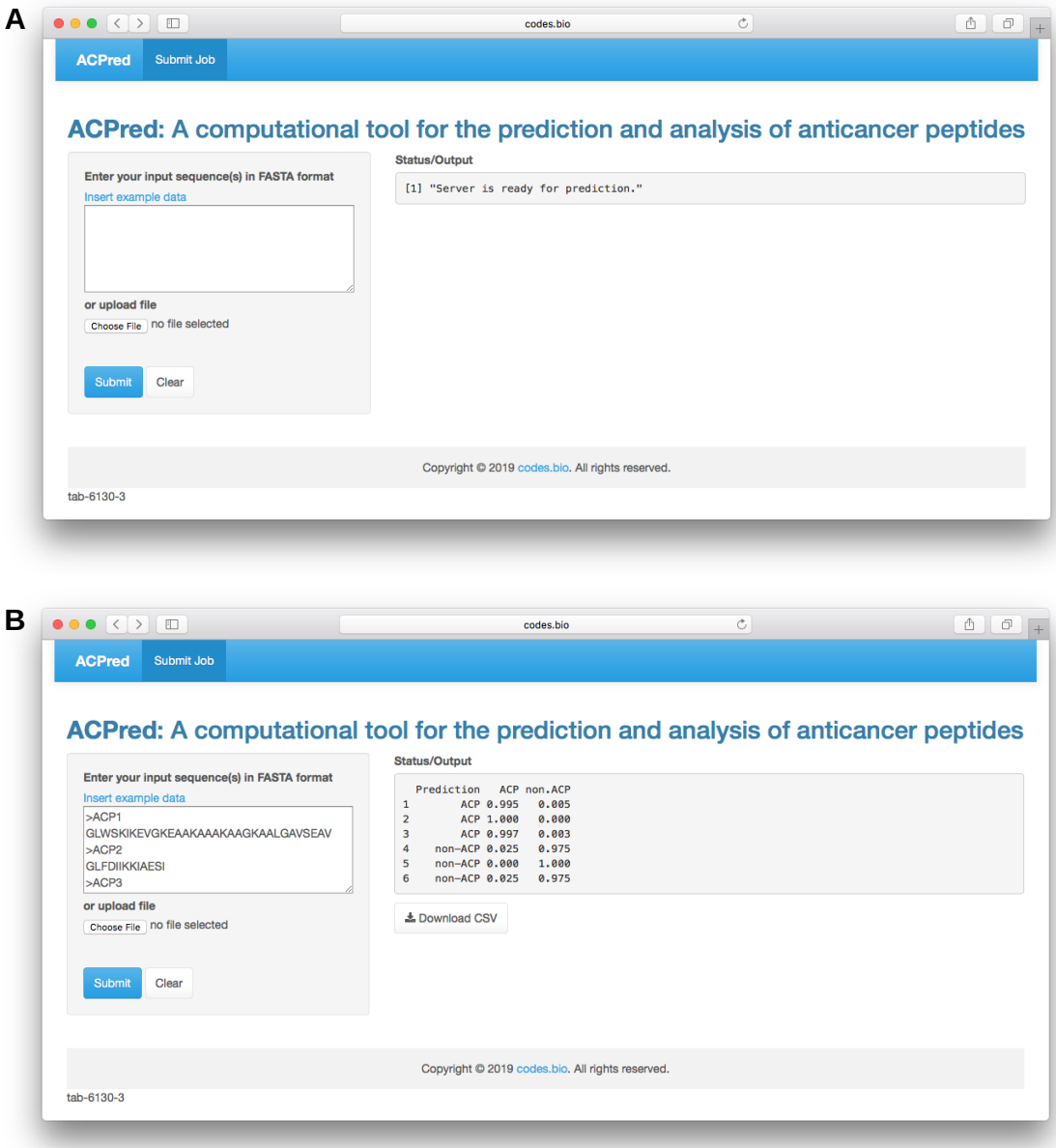

Figure 5. Screenshots of the ACPred web server before (**A**) and after (**B**) submission of sequence data for prediction.

Conclusion

There is increasing evidence indicating that ACPs could be a potential, valuable therapeutic agent for cancer treatment. With the avalanche of peptide sequences generated from many sources, there is significant demand for computational methods

that can support the timely discrimination of ACPs from other peptides, as well as offer insights from analysis using only their primary sequences. In this study, we have developed a the ACPred model in a systematic manner by taking advantage of two powerful machine learning methods consisting of support vector machine (SVM) and random forest (RF) for the prediction and characterization of anticancer activities of peptides. The best model for further development of ACPred was achieved by using the SVM method coupled with the combination of amino acid composition and pseudo amino acid composition features. By comparing with the state-of-the-art method, the TargetACP, it was found that the performance of ACPred was comparable to that of TargetACP (95.61% vs. 96.22% accuracies) as assessed by the rigorous jackknife test. Moreover, the feature importance analysis and interpretable rules extraction were carried out by using the RF model to identify which feature was the most important in predicting ACPs while also providing a better understanding of the anticancer activities of investigated peptides. Results from model interpretation revealed that the hydrophobic residues on the α-helical structure, Cys residues on the β-sheet structure and the formation of amphipathic α-helix played a crucial role in the anticancer activity of the investigated peptides. Finally, a web server named ACPred was established and made freely available online at https://codes.bio/acpred/.

References

1. WHO. Cancer 2018 [Available from: https://www.who.int/news-room/fact-sheets/detail/cancer.

2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394-424.

3. Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Pineros M, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer. 2018.

4. Craik DJ, Fairlie DP, Liras S, Price D. The future of peptide-based drugs. Chem Biol Drug Des. 2013;81(1):136-47.

5. Vlieghe P, Lisowski V, Martinez J, Khrestchatisky M. Synthetic therapeutic peptides: science and market. Drug Discov Today. 2010;15(1-2):40-56.

6. Lau JL, Dunn MK. Therapeutic peptides: Historical perspectives, current development trends, and future directions. Bioorg Med Chem. 2018;26(10):2700-7.

7. Fosgerau K, Hoffmann T. Peptide therapeutics: current status and future directions. Drug Discov Today. 2015;20(1):122-8.

8. O'Brien-Simpson NM, Hoffmann R, Chia CSB, Wade JD. Editorial: Antimicrobial and Anticancer Peptides. Front Chem. 2018;6:13.

9. Jin G, Weinberg A. Human antimicrobial peptides and cancer. Semin Cell Dev Biol. 2018.

10. Karpinski TM, Adamczak A. Anticancer Activity of Bacterial Proteins and Peptides. Pharmaceutics. 2018;10(2).

11. Usmani SS, Bedi G, Samuel JS, Singh S, Kalra S, Kumar P, et al. THPdb: Database of FDA-approved peptide and protein therapeutics. PLoS One. 2017;12(7):e0181748.

12. Deslouches B, Di YP. Antimicrobial peptides with selective antitumor mechanisms: prospect for anticancer applications. Oncotarget. 2017;8(28):46635-51.

13. Gaspar D, Veiga AS, Castanho MA. From antimicrobial to anticancer peptides. A review. Front Microbiol. 2013;4:294.

14. Ran S, Downes A, Thorpe PE. Increased exposure of anionic phospholipids on the surface of tumor blood vessels. Cancer Res. 2002;62(21):6132-40.

15. Yoon WH, Park HD, Lim K, Hwang BD. Effect of O-glycosylated mucin on invasion and metastasis of HM7 human colon cancer cells. Biochem Biophys Res Commun. 1996;222(3):694-9.

16. Sok M, Sentjurc M, Schara M. Membrane fluidity characteristics of human lung cancer. Cancer Lett. 1999;139(2):215-20.

17. Chaudhary J, Munshi M. Scanning electron microscopic analysis of breast aspirates. Cytopathology. 1995;6(3):162-7.

18. Dobrzynska I, Szachowicz-Petelska B, Sulkowski S, Figaszewski Z. Changes in electric charge and phospholipids composition in human colorectal cancer cells. Mol Cell Biochem. 2005;276(1-2):113-9.

19. Domalaon R, Findlay B, Ogunsina M, Arthur G, Schweizer F. Ultrashort cationic lipopeptides and lipopeptoids: Evaluation and mechanistic insights against epithelial cancer cells. Peptides. 2016;84:58-67.

20. Gaspar D, Freire JM, Pacheco TR, Barata JT, Castanho MA. Apoptotic human neutrophil peptide-1 anti-tumor activity revealed by cellular biomechanics. Biochim Biophys Acta. 2015;1853(2):308-16.

21. Figueiredo CR, Matsuo AL, Massaoka MH, Polonelli L, Travassos LR. Anti-tumor activities of peptides corresponding to conserved complementary determining regions from different immunoglobulins. Peptides. 2014;59:14-9.

22. Riedl S, Zweytick D, Lohner K. Membrane-active host defense peptides--challenges and perspectives for the development of novel anticancer drugs. Chem Phys Lipids. 2011;164(8):766-81.

23. Tyagi A, Kapoor P, Kumar R, Chaudhary K, Gautam A, Raghava G. In silico models for designing and discovering novel anticancer peptides. Scientific reports. 2013;3:2984.

24. Hajisharifi Z, Piryaiee M, Beigi MM, Behbahani M, Mohabatkar H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. Journal of Theoretical Biology. 2014;341:34-40.

25. Vijayakumar S, Lakshmi P. ACPP: a web server for prediction and design of anti-cancer peptides. International Journal of Peptide Research and Therapeutics. 2015;21(1):99-106.

26. Chen W, Ding H, Feng P, Lin H, Chou K-C. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget. 2016;7(13):16895.

27. Li F-M, Wang X-Q. Identifying anticancer peptides by using improved hybrid compositions. Scientific reports. 2016;6:33910.

28. Akbar S, Hayat M, Iqbal M, Jan MA. iACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space. Artificial intelligence in medicine. 2017;79:62-70.

29. Manavalan B, Basith S, Shin TH, Choi S, Kim MO, Lee G. MLACP: machine-learning-based prediction of anticancer peptides. Oncotarget. 2017;8(44):77121.

30. Xu L, Liang G, Wang L, Liao C. A Novel Hybrid Sequence-Based Model for Identifying Anticancer Peptides. Genes. 2018;9(3):158.

31. Kabir M, Arif M, Ahmad S, Ali Z, Swati ZNK, Yu D-J. Intelligent computational method for discrimination of anticancer peptides by incorporating sequential and evolutionary profiles information. Chemometrics and Intelligent Laboratory Systems. 2018;182:158-65.

32. Shoombuatong W, Schaduangrat N, Nantasenamat C. Unraveling the bioactivity of anticancer peptides as deduced from machine learning. EXCLI journal. 2018;17:734.

33. Pratiwi R, Malik AA, Schaduangrat N, Prachayasittikul V, Wikberg JE, Nantasenamat C, et al. CryoProtect: a web server for classifying antifreeze proteins from nonantifreeze proteins. Journal of Chemistry. 2017;2017.

34. Win TS, Malik AA, Prachayasittikul V, S Wikberg JE, Nantasenamat C, Shoombuatong W. HemoPred: a web server for predicting the hemolytic activity of peptides. Future medicinal chemistry. 2017;9(3):275-91.

35. Shoombuatong W, Schaduangrat N, Nantasenamat C. Towards understanding aromatase inhibitory activity via QSAR modeling. EXCLI journal. 2018;17:688.

36. Win TS, Schaduangrat N, Prachayasittikul V, Nantasenamat C, Shoombuatong W. PAAP: a web server for predicting antihypertensive activity of peptides. Future medicinal chemistry. 2018;10(15):1749-67.

37. Shoombuatong W, Prathipati P, Owasirikul W, Worachartcheewan A, Simeon S, Anuwongcharoen N, et al. Towards the revival of interpretable QSAR models. Advances in QSAR modeling: Springer; 2017. p. 3-55.

38. Shoombuatong W, Prathipati P, Prachayasittikul V, Schaduangrat N, Ahmad Malik A, Pratiwi R, et al. Towards predicting the cytochrome P450 modulation: from QSAR to proteochemometric modeling. Current drug metabolism. 2017;18(6):540-55.

39. Shoombuatong W, Mekha P, Chaijaruwanich J. Sequence based human leukocyte antigen gene prediction using informative physicochemical properties. International journal of data mining and bioinformatics. 2015;13(3):211-24.

40. Shoombuatong W, Prachayasittikul V, Anuwongcharoen N, Songtawee N, Monnor T, Prachayasittikul S, et al. Navigating the chemical space of dipeptidyl peptidase-4 inhibitors. Drug design, development and therapy. 2015;9:4515.

41. Shoombuatong W, Prachayasittikul V, Prachayasittikul V, Nantasenamat C. Prediction of aromatase inhibitory activity using the efficient linear method (ELM). EXCLI journal. 2015;14:452.

42. Golbraikh A, Muratov E, Fourches D, Tropsha A. Data set modelability by QSAR. Journal of chemical information and modeling. 2014;54(1):1-4.

43. Fourches D, Muratov E, Tropsha A. Trust, but verify II: A practical guide to chemogenomics data curation. Journal of chemical information and modeling. 2016;56(7):1243-52.

44. Shoombuatong W, Hongjaisee S, Barin F, Chaijaruwanich J, Samleerat T. HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees. Computers in biology and medicine. 2012;42(9):885-9.

45. Simeon S, Shoombuatong W, Anuwongcharoen N, Preeyanon L, Prachayasittikul V, Wikberg JE, et al. osFP: a web server for predicting the oligomeric states of fluorescent proteins. Journal of Cheminformatics. 2016;8(1):72.

46. Simeon S, Li H, Win TS, Malik AA, Kandhro AH, Piacham T, et al. PepBio: predicting the bioactivity of host defense peptides. RSC Advances. 2017;7(56):35119-34.

47. Shoombuatong W, Mekha P, Waiyamai K, Cheevadhanarak S, Chaijaruwanicha J. Prediction of human leukocyte antigen gene using k-nearest neighbour classifier based on spectrum kernel. ScienceAsia. 2013;39(1):42-9.

48. Shoombuatong W, Traisathit P, Prasitwattanaseree S, Tayapiwatana C, Cutler R, Chaijaruwanich J. Prediction of the disulphide bonding state of cysteines in proteins using Conditional Random Fields. International journal of data mining and bioinformatics. 2011;5(4):449-64.

49. Charoenkwan P, Shoombuatong W, Lee H-C, Chaijaruwanich J, Huang H-L, Ho S-Y. SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. PloS one. 2013;8(9):e72368.

50. Anuwongcharoen N, Shoombuatong W, Tantimongcolwat T, Prachayasittikul V, Nantasenamat C. Exploring the chemical space of influenza neuraminidase inhibitors. PeerJ. 2016;4:e1958.

51. Suvannang N, Preeyanon L, Malik AA, Schaduangrat N, Shoombuatong W, Worachartcheewan A, et al. Probing the origin of estrogen receptor alpha inhibition via large-scale QSAR study. RSC Advances. 2018;8(21):11344-56.

52. Simeon S, Anuwongcharoen N, Shoombuatong W, Malik AA, Prachayasittikul V, Wikberg JE, et al. Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking. PeerJ. 2016;4:e2322.

53. Shoombuatong W, Prathipati P, Prachayasittikul V, Schaduangrat N, Malik AA, Pratiwi R, et al. Towards Predicting the Cytochrome P450 Modulation: From QSAR to Proteochemometric Modeling. Current drug metabolism. 2017;18(6):540-55.

54. Calle ML, Urrea V. Letter to the editor: stability of random forest importance measures. Briefings in bioinformatics. 2010;12(1):86-9.

55. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

56. Hasan MM, Yang S, Zhou Y, Mollah MNH. SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. Molecular BioSystems. 2016;12(3):786-95.

57. Breiman L. Classification and regression trees: Routledge; 2017.

58. Worachartcheewan A, Shoombuatong W, Pidetcha P, Nopnithipat W, Prachayasittikul V, Nantasenamat C. Predicting metabolic syndrome using the random forest method. The Scientific World Journal. 2015;2015.

59. Chou K-C, Shen H-B. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nature protocols. 2008;3(2):153.

60. Chou K-C. Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of theoretical biology. 2011;273(1):236-47.

61. Chou K-C, Shen H-B. Recent progress in protein subcellular location prediction. Analytical biochemistry. 2007;370(1):1.

62. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics. 2011;12(1):77.

63. Liaw C, Tung C-W, Ho S-Y. Prediction and analysis of antibody amyloidogenesis from sequences. PloS one. 2013;8(1):e53235.

64. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321-57.

65. Huang H-L. Propensity scores for prediction and characterization of bioluminescent proteins from sequences. PloS one. 2014;9(5):e97158.

molecules

MDPI

Article

ACPred: A Computational Tool for the Prediction and Analysis of Anticancer Peptides

Nalini Schaduangrat [1], Chanin Nantasenamat [1], Virapong Prachayasittikul [2] and Watshara Shoombuatong [1,*]

[1] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand; nalini.schaduangrat@gmail.com (N.S.); chanin.nan@mahidol.edu (C.N.)
[2] Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand; virapong.pra@mahidol.ac.th
* Correspondence: watshara.sho@mahidol.ac.th; Tel.: +66-2-441-4371 (ext. 2715)

Received: 1 April 2019; Accepted: 17 May 2019; Published: 22 May 2019

Abstract: Anticancer peptides (ACPs) have emerged as a new class of therapeutic agent for cancer treatment due to their lower toxicity as well as greater efficacy, selectivity and specificity when compared to conventional small molecule drugs. However, the experimental identification of ACPs still remains a time-consuming and expensive endeavor. Therefore, it is desirable to develop and improve upon existing computational models for predicting and characterizing ACPs. In this study, we present a bioinformatics tool called the ACPred, which is an interpretable tool for the prediction and characterization of the anticancer activities of peptides. ACPred was developed by utilizing powerful machine learning models (support vector machine and random forest) and various classes of peptide features. It was observed by a jackknife cross-validation test that ACPred can achieve an overall accuracy of 95.61% in identifying ACPs. In addition, analysis revealed the following distinguishing characteristics that ACPs possess: (i) hydrophobic residue enhances the cationic properties of α-helical ACPs resulting in better cell penetration; (ii) the amphipathic nature of the α-helical structure plays a crucial role in its mechanism of cytotoxicity; and (iii) the formation of disulfide bridges on β-sheets is vital for structural maintenance which correlates with its ability to kill cancer cells. Finally, for the convenience of experimental scientists, the ACPred web server was established and made freely available online.

Keywords: anticancer peptide; therapeutic peptides; support vector machine; random forest; machine learning; classification

1. Introduction

Cancer remains one of the leading causes of death worldwide, accounting for an estimated 18.1 million new cases and resulting in 9.6 million deaths in 2018 [1]. Of these, the most common cancer types causing mortality worldwide include lung, colorectal, stomach, liver and breast cancer, respectively [1,2]. The global burden of cancer has risen dramatically in the last few decades thereby making the prevention of cancer the most significant and challenging public health concern of the 21st century [3]. In spite of the many advanced clinical methods available for the treatment of cancer (e.g., chemotherapy, radiation therapy, hormonal therapy, etc.) their recurrence rate still remains high. In addition, anti-cancer drugs are known to exert their therapeutic effects by killing off normal cells and tissues, which exacerbates the immunodeficiency of patients. Thus, the discovery and development of novel anti-cancer drugs are crucial for decreasing premature death as well as for increasing the survival of many populations. In this regard, peptide-based therapeutics have gained much attention as a new drug class since they are considered to be relatively safe, highly selective,

possess good tolerability, lower production cost, easy to modify and synthesize while also exhibiting promising pharmacological profiles [4–7]. In the last decade, more than 7000 naturally occurring peptides have been discovered and they are found to exhibit a wide range of bioactivities including tumor homing, antihypertensive, antiparasitic, antiviral, antiangiogenic, antibiofilm, antimicrobial and anticancer [8–10]. Currently, 60 peptide-based drugs have been FDA-approved [11] with preclinical and clinical studies in the pipeline for another 150 peptides [6].

Anticancer peptides (ACPs) are a subset of antimicrobial peptides (AMPs) that are have been shown to exhibit anticancer activities. These peptides are typically made up of short (5–30 amino acids) cationic residues that mostly adopts the -helical structure (e.g., BMAP-27, BMAP-28, Cecropin A and B, LL-37, Magainins etc.) or folds into a β-sheet (e.g., Defensins, Lactoferricin and Tachylepsin I) while extended linear forms (e.g., Tritrpticin and Indolicidin) have also been reported [12,13], as shown in Figure 1.

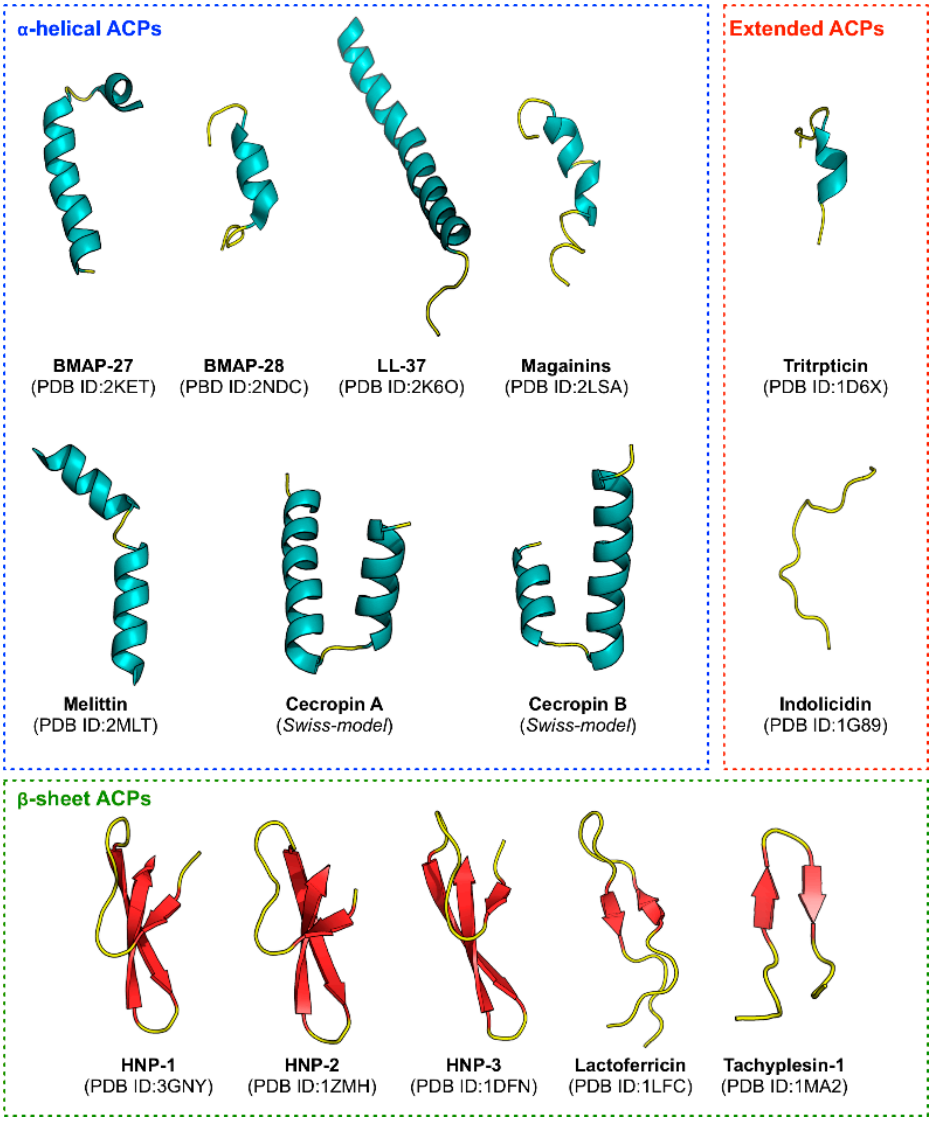

Figure 1. Overview of the structural diversity of three classes of anticancer peptides. Each structure is labeled by its common name followed by the Protein Data Bank (PDB ID) in parenthesis on the subsequent line. In cases where the PDB ID was not available, the SWISS-MODEL server (available at: https://swissmodel.expasy.org/) was used to construct the structure.

In comparison to normal cells, the cancer cell membrane exhibits different properties (i.e., larger surface area due to the higher number of microvili, high negative charge of the cell membrane, higher membrane fluidity, etc.) [14–17]. Thus, the inherent cationic property of ACPs is responsible for the electrostatic interactions between ACPs and the anionic cell membrane of cancer cells, and hence allow the selective killing of cancer cells [18]. Previously, many studies have reported that ACPs display several advantages over small molecule drugs. For instance, they exhibit a shorter half-life that consequently leads to decreased probability of resistance, lowered toxicity (not harming normal cells as much), higher specificity, good solubility as well as good tumor penetration capabilities thereby indicating their great potential in cancer therapy [19–22].

Accurate prediction of ACPs is of great importance for the exploration of their mechanism of action and for the development of therapeutic ACPs. As the experimental identification and development of novel ACPs is a very time-consuming and labor-intensive process, therefore bioinformatics tools are needed for effective analysis of the available big data on existing peptides so as to allow the identification of novel ACPs while also shedding light about their mechanism of action.

Recently, many sequence-based computational methods, for example, AntiCP [23], Hajisharifi et al.'s method [24], ACPP [25], iACP [26], Li and Wang's method [27], iACP-GAEnsC [28], MLACP [29], SAP [30] and TargetACP [31], have been developed using a wide range of machine learning methods and peptide features as summarized in Table 1. In 2014, Hajisharifi et al. [24] established a benchmark dataset spanning 138 experimentally confirmed ACPs (positive dataset) and 227 non-AMPs (negative dataset). In their method, support vector machine (SVM) with pseudo amino acid composition were used for model building and an accuracy of 92.68% were attained. As noticed in Table 1, iACP [26] was the first computational method based on informative features. In this method, Chen et al. utilized a feature selection technique for obtaining an informative feature set subsequently followed by encoding them for constructing an ACP predictor known as iACP [26] via the use of SVM. This method yielded better prediction accuracy (95.06%) than the two previous methods [23,24]. Thusfar, state-of-the-art ACP predictors includes iACP-GAEnsC [28] and TargetACP [31] in which both afforded high prediction accuracies of 96.45% and 98.78%, respectively. The iACP-GAEnsC [28] method follows the concept of evolutionary intelligent genetic algorithm-based ensemble model for improving the true classification rate. iACP-GAEnsC was developed using various classification algorithms (e.g., random forest, *k*-nearest neighbor, support vector machine, generalized neural network and probabilistic neural network) and three types of peptide features (e.g., pseudo g-gap dipeptide composition, amphiphilic pseudo amino acid composition and reduced amino acid alphabet). More recently, Kabir et al. developed the TargetACP [31] for solving the problem of class imbalance presented in the previously described benchmark dataset [24] by integrating sequential and evolutionary-profile information as input features and using SVM as the classifier. In addition, the authors utilized the synthetic minority oversampling technique (SMOTE) to solve the imbalance phenomenon between minority (ACPs) and majority (non-ACPs) samples.

Table 1. Summary of existing methods for predicting anticancer peptides.

Method (Year)	Classifier [a]	Sequence Features [b]	Interpretable	Web Server
AntiCP (2013)	SVM	AAC, DPC, BP	No	✓
Hajisharifi et al. (2014)	SVM	PseACC, LAK	No	
ACPP (2015)	SVM	PRM	No	✓[c]
iACP (2016)	SVM	g-gap DPC	**Yes**	✓
Li and Wang (2016)	SVM	AAC, RACC, acACS	No	
iACP-GAEnsC (2017)	Ensemble method	Pse-g-gap DPC, Am-PseAAC, RACC	No	
MLACP (2017)	RF	AAC, ATC, DPC, PCP	**Yes**	✓
SAP (2018)	SVM	g-gap DPC	No	
TargetACP (2018)	SVM	CPSR, SAAC, PsePSSM	No	
ACPred (this study)	SVM	AAC, DPC, PCP, PseAAC, Am-PseAAC	**Yes**	✓

[a] RF: random forest, SVM: support vector machine. [b] AAC: amino acid composition, ATC: atomic composition, acACS: auto covariance of the average chemical shift, Am-PseAAC: amphiphilic pseudo amino acid composition, BP: binary profile, CPDR: composite protein sequence representation, DPC: dipeptide composition, g-gap DPC: G-Gap dipeptide composition, LAK: local alignment kernel, PCP: Physicochemical properties, PseACC: Pseudo amino acid composition, Pse-g-gap DPC: Pseudo G-Gap dipeptide composition, PRM: protein relatedness measure, RACC: reduce amino acid composition, SAAC: split amino acid composition. [c] The web server is not accessible.

Although, iACP-GAEnsC and TargetACP yielded encouraging performances with reasonably high prediction accuracies, the overall utility of these two methods is limited in terms of interpretability and practical utility thereby warranting further improvements [32]. One of the major limitations of iACP-GAEnsC and TargetACP was that there was no web server provided for these two methods. Therefore, their utility is limited to researchers with informatics background who can develop in-house prediction models. In addition, the underlying mechanisms of the investigated bioactivity for these two methods affords limited interpretability for experimental scientists. Owing to the heterogeneous computational architecture and design of the study as well as differences in the benchmark dataset used, it is not easy to identify and assess which methods and features provided the greatest contribution to good prediction. Motivated by aforementioned issues presented in previous models [32–36], in order to establish an interpretable ACP predictor we propose a systematic effort for the prediction and analysis of anticancer activities of peptides called the ACPred. Firstly, ACPred was built using various types of peptide features and two powerful machine learning algorithms including random forest (RF) and support vector machine (SVM). Secondly, insights into the ACP mechanism of action were attained via the identified important interpretable features and subsequent in-depth analysis pertaining to the underlying biophysical and biochemical properties of anticancer activities of peptides [33,34,36–38]. Thirdly, if-then interpretable rules were extracted using the RF model (IR-ACP). Our prediction results on the benchmark dataset showed that the proposed ACPred achieved an accuracy and MCC of 95.61% and 0.91, respectively, which outperforms most of the existing sequence-based computational methods including Hajisharifi et al.'s method [24], iACP [26], Li and Wang's method [27], iACP-GAEnsC [28], MLACP [29] and SAP [30]. As previously noted, although iACP-GAEnsC and TargetACP boasts good predictive performances but their interpretability and practical utility prompted the development of ACPred. Thus, ACPred is a user-friendly and publicly accessible web server that allow robust predictions to be made without the need to develop in-house prediction models.

2. Materials and Methods

The overall framework of ACPred is illustrated in Figure 2. It can be seen that four major steps are involved in the development of this method as follows: (i) data preparation, (ii) model construction (feature extraction and feature combination), (iii) feature importance analysis and (iv) web server construction. The first step prepares the benchmark dataset, the second step extracts and combines peptide features and the third step analyzes the important features governing ACP prediction via the RF model. Finally, the fourth step involves the construction of a publicly accessible web server using the best model.

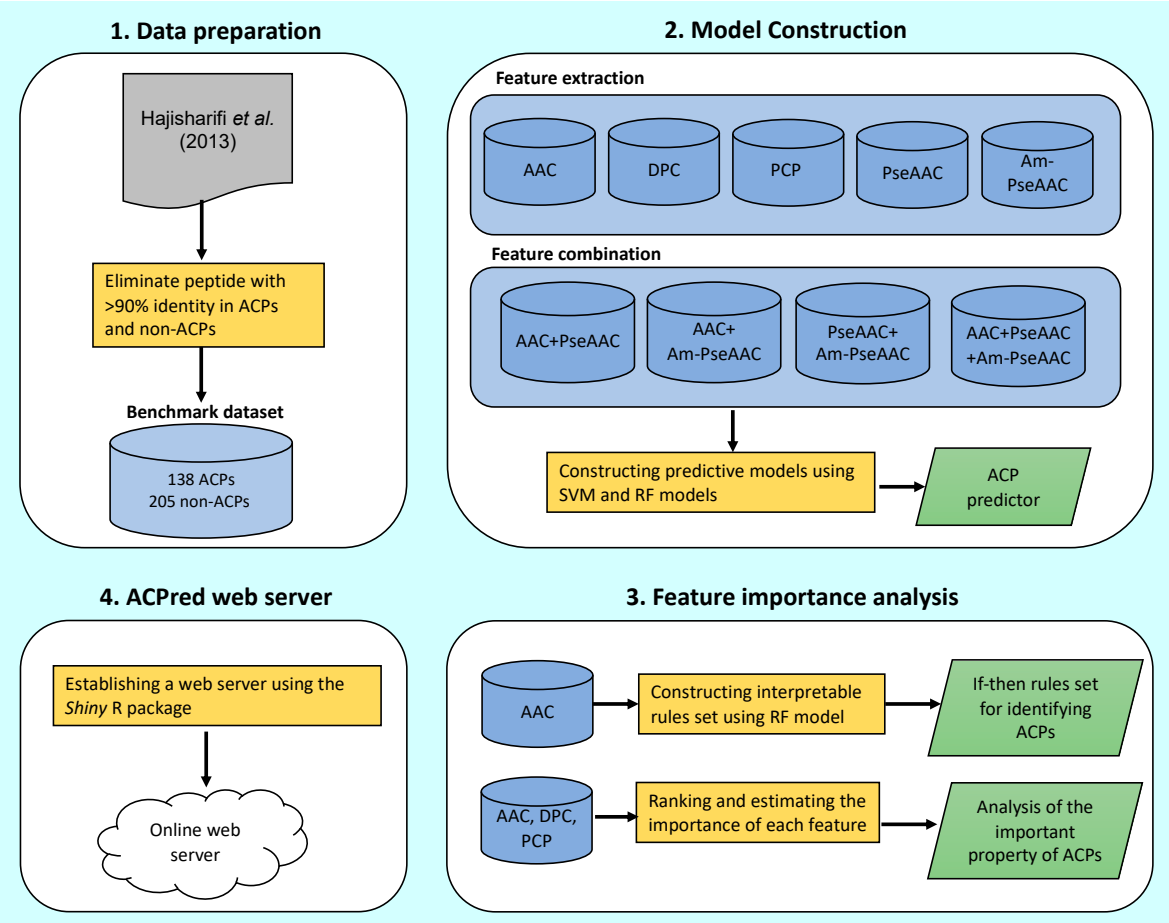

Figure 2. Schematic framework of ACPred.

2.1. Benchmark Dataset

A high-quality benchmark dataset can ensure the reliability of the proposed method. Therefore, the investigated dataset was retrieved from the work of Hajisharifi et al. [24], which contains 192 ACPs and 215 non-ACP. As to guarantee the quality of the benchmark dataset, the following steps were considered. Firstly, in order to avoid a dataset containing redundant peptides, the CD-HIT program was used to eliminate ACPs and non-ACPs with >90% identity. Secondly, ACPs and non-ACP containing special characters, such as X and U, were removed. After such screening procedures, a set of 138 ACPs and 205 non-ACPs was attained for prediction and analysis of ACPs.

2.2. Feature Representation

In order to build an effective prediction model, it is necessary to represent each peptide as numerical vectors that can encompass the perspective of its biophysical and biochemical properties. For peptide or protein sequences, the most classical and interpretable features are amino acid composition (AAC), dipeptide composition (DPC) and physicochemical property (PCP).

According to the classical definition, AAC and DPC are the proportions of each amino acid and dipeptide in a peptide sequence **P** that are expressed as fixed lengths of 20 and 400, respectively. Thus, in terms of AAC and DPC features, a peptide **P** can be expressed by vectors with 20D and 400D (dimension) spaces, respectively, as formulated by the following two equations:

$$\mathbf{P} = [aa_1, aa_2, \ldots, aa_{20}]^T \quad (1)$$

$$\mathbf{P} = [dp_1, dp_2, \ldots, dp_{400}]^T \quad (2)$$

where T is the transpose operator, while aa_1, aa_2, ... , aa_{20} and dp_1, dp_2, ... , dp_{400} are occurrence frequencies of the 20 and 400 native amino acids and dipeptides, respectively, in a peptide sequence **P**. PCP is one of the most intuitive features associated with biophysical and biochemical reactions. In fact, there are 544 PCPs of amino acids extracted from the amino acid index database (AAindex) [39], which is a collection of published literature as well as different biochemical and biophysical properties of amino acids. Each physicochemical property consists of a set of 20 numerical values for amino acids. After removing 13 PCPs with not applicable (NA) as their amino acid indices, a total of 531 PCPs were further used in this study.

As mentioned in the study of [40–42], the sequence order information of AAC and DPC would be lost. To deal with such a dilemma, the pseudo amino acid composition (PseAAC) and amphiphilic pseudo amino acid composition (Am-PseAAC) approaches were proposed. According to Chou's PseAAC [41], the general form of PseAAC for a peptide **P** is formulated by: where the subscript Ω is an integer to reflect the feature's dimension. The value of Ω and the component of Ψ_u, where $u = 1, 2, \ldots, \Omega$ depends on the protein or peptide sequences. In this study, the parameters of PseAAC, i.e., the discrete correlation factor λ and weight of the sequence information ϖ, were set by default in the protr R package [43]. The dimension of PseAAC feature is $20 + \lambda \times \varpi$. Since the hydrophobic and hydrophilic properties of proteins play an important role in its folding and interaction of proteins, Am-PseAAC was introduced by Chou [40]. The dimension of Am-PseAAC feature is $20 + 2\lambda$ whereby, the first 20 components are the 20 basic AAC ($p_1, p_2, \ldots, p_{20}$) and the next 2λ ones are a set of correlation factors that reveal the physicochemical properties such as hydrophobicity and hydrophilicity along a protein or peptide sequence as formulated by:

$$\mathbf{P} = [\Psi_1, \Psi_2, \ldots, \Psi_u, \ldots, \Psi_\Omega]^T \quad (3)$$

where the subscript Ω is an integer to reflect the feature's dimension. The value of Ω and the component of Ψ_u, where $u = 1, 2, \ldots, \Omega$ depend on the protein or peptide sequences. In this study, the parameters of PseAAC, i.e., the discrete correlation factor λ and weight of the sequence information ϖ, were set by defaults using the protr R package [43]. The dimension of PseAAC feature is $20 + \lambda \times \varpi$. Since hydrophobic and hydrophilic properties play an important role in the folding and interaction of proteins, Am-PseAAC was introduced by Chou [40] for representing such features. The dimension of Am-PseAAC feature is $20 + 2\lambda$ whereby the first 20 components are the 20 basic AAC ($p_1, p_2, \ldots, p_{20}$) while the next 2λ ones are a set of correlation factors that pertains to physicochemical properties namely hydrophobicity and hydrophilicity along a protein or peptide sequence as formulated by:

$$\mathbf{P} = [p_1, p_2, \ldots, p_{20}, p_{20+\lambda}, p_{20+\lambda+1}, \ldots p_{20+2\lambda}]^T \quad (4)$$

In this study, the five aforementioned features of peptide sequences were calculated using the protr package in the R software [43].

2.3. Dataset Modelability

The prediction performances of QSAR models are directly influenced by the dataset and the types of feature used. Recently, the concept of dataset modelability has been proposed by Golbraikh et al. [32,35,44,45] for evaluating the feasibility of developing a robust and predictive model for a given dataset of interest. In this study, the MODI score for each type of peptide features and their combinations were calculated using an in-house developed R code. Prediction models are recommended for further analysis if the MODI score is greater than 0.65; otherwise, prediction models are not recommended. Further details pertaining to the calculation of the MODI score can be found elsewhere [44].

2.4. Classifier Selection

In this study, we employed two successful machine learning algorithms namely random forest (RF) and support vector machine (SVM). Previously, these approaches have been successfully used in the prediction of various functions and properties of peptides and proteins [33,36,38,46–51] as well as other biological or chemical entities [35,45,52–55]. Herein, the basic concepts and associated parameter optimizations for the two classifiers are briefly described hereafter.

RF models were constructed according to the described original RF algorithm [56]. These models were developed by growing many weak classification and regression tree (CART) classifiers where each classifier is generated using a random vector sampled independently from the input vector as to enhance the prediction performance of CART [56,57]. In the RF method, the out-of-bag (OOB) approach is used for evaluating the feature importance as follows: (1) two-thirds of the training data was utilized to construct the predictive classifier while the remaining was used for evaluating the performance of such classifier and (2) the feature importance of each feature can be evaluated by measuring the decrease in the prediction performance. It should be noted that the performance evaluation of the model can be either accuracy or Gini index. Herein, the RF classifier was established using the *randomForest* package in the R software [58]. To enhance the performance of the RF model, two parameters namely, *ntree* (i.e., the number of trees used for constructing the RF classifier) and *mtry* (i.e., the number of random candidate features) were determined using the *caret* package in the R software [59] via a 5-fold cross-validation (5-fold CV) approach. The search space of *ntree* and *mtry* were [100, 500] and [1, 10] using steps of 100 and 1, respectively.

SVM is a supervised learning model based on the principles of structure risk minimization and kernel method as proposed by Vapnik [60–62]. SVM model can deal with the problem of over-fitting arising with the use of a small training dataset by mapping the input samples to a higher dimension space and then searching for the maximum-margin hyperplane for constructing the classifier [63,64]. Previously, SVM models were used in many applications because of their predictive performance capabilities when applied to classification, prediction, and regression problems. In the optimization process, the regularization parameter $C \in \{0.25, 0.50, 1, 2, 4\}$ was determined with a 5-fold CV approach using the *caret* package in the R software [59].

2.5. Identification of Important Features

The analysis and identification of feature importance can provide a better understanding on the underlying biophysical and biochemical properties governing anticancer activities of peptides. Herein, the efficient and effective built-in feature importance estimator of the RF method was used to reveal and characterize differences between ACPs and non-ACPs. As mentioned above, the RF method provides two measures for ranking feature importance, i.e., the mean decrease of the Gini index (MDGI) and the mean decrease of prediction accuracy. Since Calle and Urrea [65] demonstrated that the MDGI provided more robust results when compared with the mean decrease of prediction accuracy, we utilized the MDGI to rank the importance of interpretable features that included AAC, DPC and PCP. Until now, these three features have been used to characterize many peptides and proteins such as for predicting HIV-1 CRF01-AE co-receptor usage [46], predicting protein crystallization [51,66], predicting the oligomeric states of fluorescent proteins [38], predicting the bioactivity of host defense peptides [47], prediction of human leukocyte antigen gene [37,49], predicting antifreeze proteins [33], predicting the hemolytic activity of peptides [34] and predicting the antihypertensive activity of peptides [36].

The MDGI is an impurity measure that corresponds to the ability of each feature in discriminating the sample classes. The Gini index can be defined as follows:

$$1 - \sum_{c=1}^{2} p^2(c|t) \quad (5)$$

where denotes the estimated class probability for node t in a tree classifier and c is the class label (i.e., either ACPs or non-ACPs). Features with the largest MDGI value is considered to be an important feature as it significantly contributes to the prediction performance.

2.6. Rule Extraction

This work presented an interpretable rule extraction of ACPs (IR-ACP) based on the RF method cooperating with AAC features for determining how the features work in combination within the learning process. A set of rules from an individual tree is derived from the root to the leaves. In this study, only 100 decision trees were used to extract if-then interpretable rules for explaining the prediction results by means of the *RRF*, *inTrees* and *xtable* packages in the R software [56,57,67]. More details of the rule extraction process can be found in previous related works [67,68].

2.7. Performance Evaluation

Three testing methods consisting of (i) sub-sampling test (2-, 5- or 10-fold cross-validation), (ii) jackknife test or also known as leave-one-out cross-validation (LOO-CV), and (iii) independent (or external) testing dataset are often used to evaluate the prediction performance. The sub-sampling test is one of the popular cross-validation methods for assess the predictive capability of a model. As elucidated in [41,69] and demonstrated by Equation (50) of [70], among the three cross-validation methods, the jackknife test is considered as one of the most rigorous and objective methods for cross-validation in statistics and can provide a unique result for a given benchmark dataset. Thus, in this study, the 5-fold cross-validation (5-fold CV) and jackknife test were used to evaluate the prediction performance of our models. The former set (5-fold CV) makes use of data from the training set where the data set is separated into five subsets. Practically, one subset from a total of five subsets is left out as the testing set while the remaining are used for training the model. This process is repeated iteratively until all data samples have had the chance to be left out as the testing set. During the jackknifing process, each sample in the dataset (N samples) is left out as an independent sample (one sample) and the remaining sequences (N-1 samples) are used for training the model. This process is repeated iteratively until all samples have had the chance to be left out as the independent sample.

In order to evaluate the prediction ability of the model, the following sets of four metrics are used:

$$Ac = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{6}$$

$$Sn = \frac{TP}{(TP + FN)} \tag{7}$$

$$Sp = \frac{TN}{(TN + FP)} \tag{8}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{9}$$

where *Ac*, *Sn*, *Sp* and *MCC* represent accuracy, sensitivity, specificity and Matthews coefficient correlation, respectively. *TP*, *TN*, *FP*, and *FN* represent the instances of true positive, true negative, false positive and false negative, respectively. Moreover, in order to evaluate the prediction performance of models using threshold-independent parameters, receiver operating characteristic (*ROC*) curves were plotted using the pROC package in the R software [71]. The area under the ROC curve (*auAUC*) was used to measure the prediction performance, where *AUC* values of 0.5 and 1 are indicative of perfect and random models, respectively.

2.8. Reproducible Research

To ensure the reproducibility of models proposed herein, all R codes and the benchmark dataset used in the construction of the predictive models, graphical figures and the ACPred web server are available on GitHub at https://github.com/Shoombuatong2527/acpred and https://github.com/chaninlab/acpred-webserver.

3. Results and Discussion

3.1. Prediction Performance

In this study, effort is directed toward the determination of feature types that are beneficial for the prediction of APCs. This was performed by comparing the performance of different feature types as constructed using SVM and RF by means of 5-fold CV and jackknife test. The five basic features consisting of AAC, DPC, PCP, PseAAC and Am-PseAAC were considered as well as their combinations: AAC + PseACC, AAC + Am-PseACC, PseACC + Am-PseACC and AAC + PseACC + Am-PseACC, which were also utilized to investigate the complementarity of the five aforementioned features.

Prior to the development of the predictive model, the assessment of modelability of the benchmark dataset was performed by computing the MODI index. This index helps modelers to estimate the feasibility of obtaining robust and reliable predictive models. For the binary classification problem, if the value of MODI index is greater than 0.65, the feature is considered to be reliable for classification modelling; otherwise, the feature is not recommended for classification modelling.

As shown in Table A1, all feature types and their combinations from the benchmark dataset met these criteria. Thus, it could be stated that our proposed features are reliable and efficient for constructing ACP predictor.

Tables 2 and 3 shows the performance comparison amongst different feature types with RF and SVM on the benchmark dataset via 5-fold CV and jackknife test, respectively. In addition, Figure 3 shows the receiver operating characteristic (*ROC*) curve of the five considered feature sets obtained from RF (top and bottom left) and SVM (top and bottom right) as evaluated by 5-fold CV (top left and right) and jackknife test (bottom left and right). As seen in Tables 2 and 3, SVM model with the Am-PseAAC feature afforded the highest accuracy of 95.03% as evaluated by the jackknife test. Meanwhile, RF model with PseACC feature and SVM model with AAC feature performed well with the second and third the highest accuracy of 93.28% and 92.98%, respectively. These results showed that ACC, PseACC and Am-PseAAC were effective for ACP prediction. Since using a combination of various features might yield better prediction performance, four combinations of the three effective features were also considered. The highest accuracy and *MCC* of 95.61% and 0.91, respectively, was achieved by using SVM model cooperating with the combination feature of AAC and Am-PseAAC, while using the combination of the three effective features, i.e., ACC, PseACC and Am-PseAAC, performed slightly worse than the combination feature of AAC and Am-PseAAC with an accuracy and *MCC* of 94.74% and 0.89, respectively. In addition, our prediction results were well consistent with previous studies [27–29] and related studies [33,34,36,72].

Table 2. Performance comparison of SVM and RF with various types of sequence features over five-fold cross-validation.

Feature	Classifier	*Ac (%)*	*Sn (%)*	*Sp (%)*	*MCC*	*auROC*
AAC	SVM	92.69	83.94	98.54	0.850	0.977
	RF	91.23	92.80	90.32	0.817	0.958
DPC	SVM	83.92	100.00	78.85	0.687	0.942
	RF	87.14	91.89	84.85	0.733	0.944
PCP	SVM	84.80	63.50	99.02	0.698	0.938
	RF	83.63	90.10	80.91	0.661	0.872
PseAAC	SVM	92.98	84.67	98.54	0.856	0.990
	RF	92.11	93.65	91.20	0.835	0.959
Am-PseAAC	SVM	95.03	87.59	100.00	0.899	0.995
	RF	92.40	100.00	88.75	0.848	0.974
AAC + PseAAC	SVM	93.57	85.40	99.02	0.869	0.991
	RF	92.98	96.69	90.95	0.855	0.964
AAC + Am-PseAAC	SVM	95.32	89.05	99.51	0.904	0.994
	RF	93.28	97.50	90.99	0.862	0.969
PseAAC + Am-PseAAC	SVM	94.44	86.86	99.51	0.887	0.994
	RF	92.69	98.28	89.82	0.851	0.967
AAC + PseAAC + Am-PseAAC	SVM	94.15	86.13	99.51	0.881	0.993
	RF	92.98	98.29	90.22	0.857	0.972

Table 3. Performance comparison of SVM and RF with various types of sequence features over jackknife test.

Feature	Classifier	*Ac (%)*	*Sn (%)*	*Sp (%)*	*MCC*	*auROC*
AAC	SVM	92.98	83.94	99.02	0.857	0.978
	RF	91.23	92.80	90.32	0.817	0.959
DPC	SVM	85.09	98.86	80.32	0.706	0.941
	RF	86.84	89.66	85.40	0.725	0.947
PCP	SVM	84.80	63.50	99.02	0.698	0.937
	RF	83.63	88.57	81.44	0.659	0.868
PseAAC	SVM	92.98	84.67	98.54	0.856	0.990
	RF	93.28	97.50	90.99	0.862	0.960
Am-PseAAC	SVM	95.03	87.59	100.00	0.899	0.995
	RF	92.40	99.12	89.08	0.847	0.969
AAC + PseAAC	SVM	93.57	85.40	99.02	0.869	0.990
	RF	93.28	98.31	90.63	0.863	0.962
AAC + Am-PseAAC	SVM	95.61	89.78	99.51	0.910	0.994
	RF	93.57	98.32	91.03	0.869	0.967
PseAAC + Am-PseAAC	SVM	93.86	85.40	99.51	0.875	0.992
	RF	93.57	99.15	90.67	0.870	0.959
AAC + PseAAC + Am-PseAAC	SVM	94.74	87.59	99.51	0.893	0.994
	RF	92.98	99.13	89.87	0.858	0.973

Furthermore, observations pertaining to the performance comparisons from Tables 2 and 3 and Figure 3, it can be briefly summarized as follows. Tables 2 and 3 shows that both RF and SVM models afforded improved prediction performances when using the Am-PseACC feature while the best prediction performance on the benchmark dataset via both 5-fold CV and jackknife test were achieved by the SVM model trained using the combination of AAC and Am-PseAAC features.

For the convenience of subsequent descriptions, we will refer to this method as ACPred as it represents the best model that will be used for further comparisons with other tools.

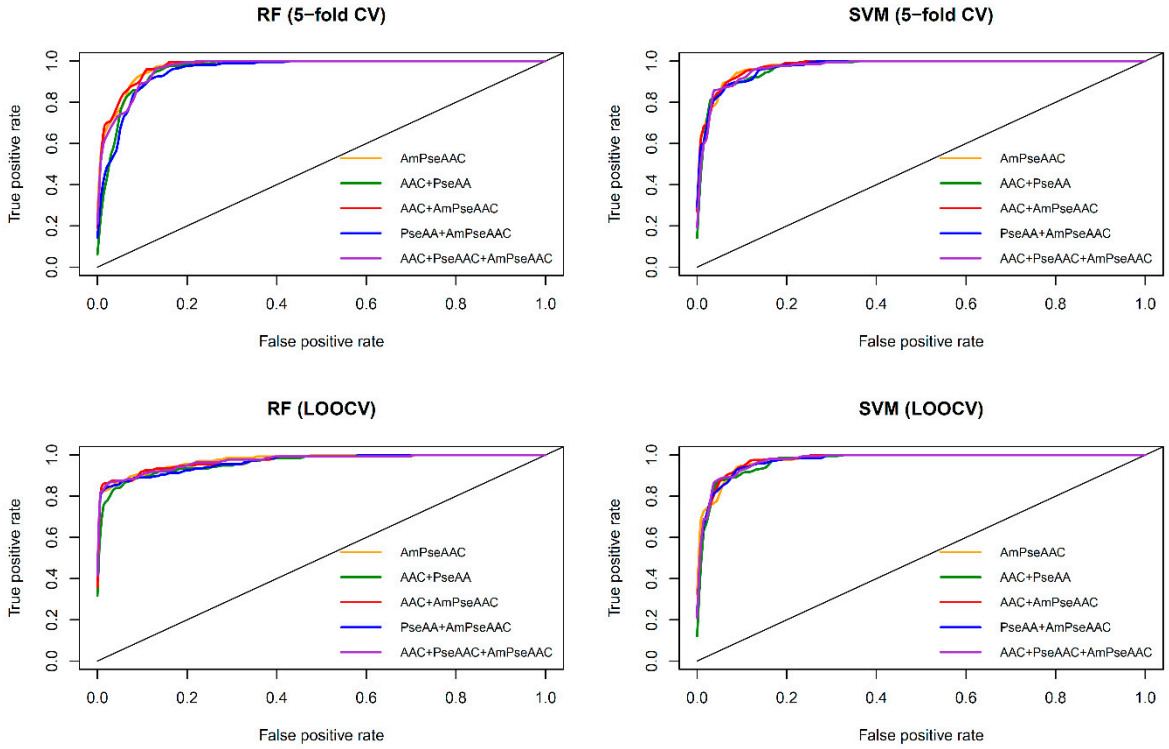

Figure 3. ROC curve of RF (top and bottom left) and SVM (top and bottom right) models as assessed by 5-fold cross-validation (top left and right) and jackknife test or leave-one-out cross-validation (bottom left and right).

3.2. Comparison with Other Methods

To indicate the effectiveness of the proposed method, we compared the performance of our selected model (named ACPred) with other popular ACP predictors. Since Hajisharifi et al. [24] were the first group to establish the benchmark dataset, and iACP [26], iACP-GAEnsC [28] and TargetACP [31] were state-of-the-art ACP predictors that provided prediction results as assessed by the jackknife test, therefore these predictors were used for performance comparisons. Table 4 lists the performance comparisons of ACPred with the three ACP predictors. As noticed from Table 4, prediction results were obtained from two different experimental designs: (i) the prediction model was performed on the benchmark dataset as derived from the work of [24] as well as those of Hajisharifi et al.'s method, iACP, iACP-GAEnsC, TargetACP and ACPred, (ii) the prediction model was built using the balanced dataset including TargetACP and ACPred. The reported results from aforementioned methods as summarized in Table 4 was obtained directly from the work of iACP-GAEnsC [28].

Table 4. Performance comparison of the proposed ACPred model with existing methods.

Method [a]	*Ac (%)*	*Sn (%)*	*Sp (%)*	*MCC*
Hajisharifi et al. [b]	92.68	89.70	85.18	0.78
iACP [b]	95.06	89.86	98.54	0.90
iACP-GAEnsC [b]	96.45	95.36	97.54	0.91
TargetACP [b]	96.22	94.20	97.57	0.92
TargetACP [c]	98.78	99.02	98.54	0.97
ACPred [b]	95.61	89.78	99.51	0.91
ACPred-modified [c]	97.56	96.08	99.02	0.95

[a] Results were reported from the work of TargetACP. [b] Results were performed on the benchmark dataset consisting of 138 ACPs and 205 non-ACPs. [c] Results were performed on the balanced dataset consisting of 205 ACPs and 205 non-ACPs by using the SMOTE technique on the benchmark dataset.

As for performance comparisons on the benchmark dataset, it was found that ACPred was comparable with that of iACP-GAEnsC and TargetACP as indicated by four statistical parameters. Moreover, ACPred yielded a greater prediction performance than the method proposed by Hajisharifi et al. and iACP. To the best of our knowledge, iACP-GAEnsC is the only method that utilizes the synthetic minority oversampling technique (SMOTE) [73] technique for constructing a balanced dataset from the original benchmark dataset [24]. From the perspectives of machine learning, it is not fair to directly compare prediction results between TargetACP and ACPred because TargetACP was trained and tested on the balanced dataset consisting of 205 ACPs and 205 non-ACPs. In order to perform a fair comparison, the herein proposed ACPred method was applied on the balanced dataset (called ACPred-modified), which was generated by means of the SMOTE technique. Herein, the oversampling method was used to add synthetic samples for the minority class (i.e., ACPs) presented in the benchmark dataset by setting the parameters of the number of nearest neighbours (k), the number of extra cases (oversampling) to add to the minority class (perc.over) and the number of cases to reduce (undersampling) from the majority class (perc.under) were set to 9, 50 and 300, respectively. As seen in Table 4, ACPred-modified was found to outperform Hajisharifi et al.'s method, iACP and iACP-GAEnsC with an improvement of 2%–4% and 3%–7% on *Ac* and *MCC*, respectively. However, it was observed that ACPred-modified performed slightly worse than TargetACP by approximately 2%–3% where ACPred-modified and TargetACP provided 97.56% *Ac*/0.95 *MCC* and 98.78% *Ac*/0.97 *MCC*, respectively. Nevertheless, TargetACP was constructed with a sophisticated design thus, it is not easy to identify and assess which features offer the most contribution to the prediction improvement. On the other hand, ACPred was designed in a systematic manner for prediction and characterization of anticancer activities of peptides. In addition, a user-friendly web server ACPred was developed to facilitate high-throughput prediction of ACP. Therefore, the proposed ACPred model could become a practical tool for predicting and interpreting the anticancer activity of peptides, or at least as a complementary tool to existing methods in the field.

Thinking to the future, some possible improvements that could enhance the prediction performance for predicting anticancer activity of peptides are described hereafter. Firstly, it is worthy to explore the separation of peptides according to their sequence lengths (e.g., peptides having 4, 5 or 6 amino acids, peptides having 7–12 amino acids, and peptides having more than 12 amino acids) followed by the development of separate predictive models for each of the five sequence range [36,74]. Secondly, extract 5, 10 and 15 amino acids from the N or C terminus for the development of predictive QSAR models, which has successfully been demonstrated by the group of P.S. Raghava [23] Thirdly, explore the modelability of the dataset using the approach of Golbraikh et al. [44] such that robust models could be developed. Fourthly, the currently employed non-anticancer peptides [24] are not based on true experimentally-determined inactive peptides but are random peptides obtained from the UniProt database therefore, if such true inactive peptides could be determined they may aid in the development of robust models.

Aside from the aforementioned methodological improvements that have been witnessed from the literature in the development of anticancer peptide classifiers, however, the underlying origin of anticancer activity of investigated peptides as rationalized by peptide features is an area that deserves more attention. Current predictive models used for the prediction of anticancer activity of peptides are primarily based on sequence order-independent descriptors that does not consider the order of amino acids in a peptide sequence (e.g., AAC and DPC), which hampers the identification of motifs from peptides. Such limitation has been addressed by the group of KC Chou [41,75,76] in which they developed PseAAC and Am-PseAAC descriptors that take into account sequence order information based on physicochemical properties. However, the interpretability of these descriptors is limited and thus, existing approaches for rationalizing biological interpretation from these predictive models are resorted back to the sequence order-independent descriptors such as AAC and DPC. Analysis of the literature revealed that interpretable descriptors such as the z-scale descriptors [77,78] have been successful for predicting robust and interpretable models, however the flaw of this approach

is that the peptides used in the development of the model should be of equal length so as to conduct sequence alignment. Methodological tweaks addressing such limitation have been made via the use of autocorrelation such that sequence alignment is not needed and would allow the models to be built on peptides of varied lengths. In spite of this, this method is not applicable in this study field in which the prepared dataset removes peptides having high similarity of greater than 90% therefore, rendering the removal of a large number of peptides that is normally needed when using *z*-scale descriptors. As such, there is ample room for the discovery of new and interpretable descriptors for the development of predictive models of peptides, which is an important area for further driving the field of anticancer peptide prediction.

3.3. Biological Space

The identification of feature importance from AAC and DPC descriptors can provide a better understanding on the biochemical and biophysical properties of anticancer activities of peptides. Previously, AAC and DPC features have been analyzed as to further gain insights on how to characterize therapeutic peptides [29,34,36,47] and protein functions [33,51,79]. In this study, the value of MDGI was adopted to rank and estimate the importance of each AAC and DPC feature. Such information is derived from the analysis of the benchmark dataset that consists of 138 ACPs and 205 non-ACPs. Table 5 lists the percentage values of the 20 amino acids for both ACPs and non-ACPs along with amino acid composition difference between the two classes as well as their MDGI values. In addition, a heatmap showing the feature importance for DPC features is shown in Figure 4. From Table 5, it can be observed that the ten informative amino acids with the highest MDGI values are Lys, Cys, Gln, Arg, Ile, Leu, His, Pro, Glu and Gly (29.54, 19.71, 15.89, 11.42, 10.13, 7.84, 7.80, 7.37, 6.88 and 6.56, respectively). Meanwhile, Figure 4 shows that the 20 top-ranked dipeptides according to their MDGI value are KK, KI, CY, IK, AK, KV, IP, LK, TC, SC, CG, LF, FK, GL, FL, LR, PP, LG, IG and AQ. Interestingly, among the 5 top-ranked informative amino acids, 3 of these are polar (i.e., Lys, Gln and Arg), while 2 are non-polar (i.e., Cys and Ile) and hydrophobic residues. In addition, Lys and Arg are also positively-charged residues, which may support the anti-cancer cell penetrating properties of peptides (i.e., that is required for targeting tumors with specificity and low toxicity) [80]. Although the percentage compositions differentiating ACPs from non-ACPs as presented in Table 5 highlight that ACPs a have higher percentage of polar residues and lower percentage of basic residues as compared to non-ACPs, cell surface binding and internalization are critical for specific targeting of cells with anti-cancer activities. According to the membrane of cancer cells presenting anionic molecules such as phosphatidyl serine (PS), heparin sulfate and O-glycosylated mucins, the ACPs that contain positive charge are critical for endocytosis and selectively killing cancer cells. While hydrogen bonds that appear on ACPs were designed for improving the solubilization of hydrophobic molecules on ACPs. However, positively charged amino acids alone are not enough to completely neutralize cancer cells but rather aid in rapid internalization of cancer cells for selective killing with low toxicity to normal cells. Moreover, hydrogen bonded interactions to encapsulate anti-cancer drugs has been utilized for increasing the oral efficiency and drug-controlled release. Therefore, only hydrogen bond on ACPs are not adequate to resist cancer cells. Thus, the combination between crucial amino acids and physicochemical properties could provide a synergistic effect that enhances the efficiency of ACPs [81]. As such, results from the analysis of the top ranked informative AAC are discussed below.

Table 5. Amino acid compositions (%) of anticancer and non-anticancer peptides along with their difference as well as MDGI values. The rank of each amino acid amongst their 20 amino acids are shown in parenthesis for AAC difference and MDGI.

Amino Acid	ACP (%)	Non-ACP (%)	Difference	*p*-Value	MDGI
A-Ala	7.623	11.005	−3.383 (7)	<0.05	6.41 (11)
C-Cys	3.906	8.015	−4.109 (5)	<0.05	19.71 (2)
D-Asp	2.417	3.418	−1.002 (15)	<0.05	3.85 (15)
E-Glu	1.707	3.523	−1.816 (12)	<0.05	6.88 (9)
F-Phe	6.823	2.41	4.413 (2)	<0.05	6.12 (12)
G-Gly	1.975	4.123	−2.148 (11)	<0.05	6.56 (10)
H-His	1.536	5.798	−4.262 (4)	<0.05	7.80 (7)
I-Ile	10.072	6.98	3.092 (9)	<0.05	10.13 (5)
K-Lys	2.542	1.651	0.892 (17)	0.086	29.54 (1)
L-Leu	8.099	3.739	4.36 (3)	<0.05	7.84 (6)
M-Met	9.831	13.888	−4.057 (6)	<0.05	4.34 (14)
N-Asn	11.497	3.964	7.533 (1)	<0.05	4.98 (13)
P-Pro	0.905	2.224	−1.319 (13)	<0.05	7.37 (8)
Q-Gln	5.385	2.711	2.674 (10)	<0.05	15.89 (3)
R-Arg	4.211	7.495	−3.283 (8)	<0.05	11.42 (4)
S-Ser	6.537	5.832	0.705 (19)	0.245	3.50 (16)
T-Thr	3.781	4.704	−0.923 (16)	0.098	3.13 (17)
V-Val	2.258	1.560	0.698 (20)	0.083	2.96 (18)
W-Trp	2.244	1.423	0.821 (18)	<0.05	2.51 (20)
Y-Tyr	6.65	5.539	1.111 (14)	0.091	2.90 (19)

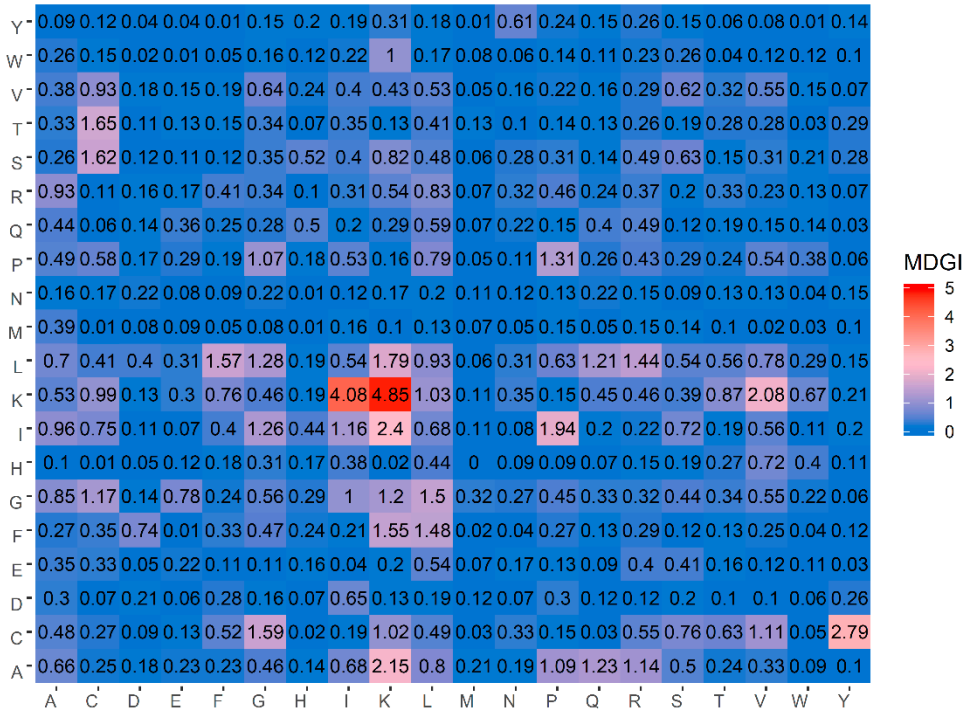

Figure 4. Heat map of the mean decrease of Gini index of dipeptide compositions.

To roughly analyze the characteristics of ACPs, a few studies have shed light on important amino acids and dipeptides commonly found in ACPs by using simple composition analysis approaches without the use of experimental methods. For instance, Tyagi et al. [23] reported results of residue preference at 10 N-terminus and 10 C-terminus by using the sequence logos. Their analysis revealed that Leu and Lys were typically found at the N-terminus while Cys, Leu and Lys were typically found at the C-terminus. Chen et al. [26] also confirmed that amino acids including Lys, Ile, Cys, Glu, and Gly were abundant in ACPs as compared to non-ACPs. Recently, Manavalan et al. [29] revealed that the

three top-ranked informative dipeptides consisted of KK, AK and KL. As can be noticed in Table 5 and Figure 4, our analysis results derived from the MDGI values of AAC and DPC were quite consistent with the three aforementioned studies.

As mentioned above, amongst the five top-ranked informative amino acids, Lys had the highest MDGI value of 29.54. Such feature with the highest MDGI is considered to be the most important factor governing the anticancer activity. As a basic residue, Lys is highly conserved in the composition of therapeutic peptides as it enhances the formation of electrostatic interactions between the peptide and the plasma membrane owing to its cationicity [82]. The role of cationicity in ACPs have been previously investigated by the study of Gopal et al. [83]. The authors modified a synthetic AMP known as HPA3NT3 (FKRLKKLFKKIWNWK), which was derived from Helicobacter pylori by substituting Arg and Asn with Lys at positions 3 and 13, respectively. Gopal et al. reported that the Lys-modified peptide increased the peptide selectivity toward negatively-charged membrane surface, which could allow for endocytosis and specificity of the APC in killing cancer cells. In addition, Wang et al. [84] has shown that the cell-penetrating ability is critical for specifically targeting cationic cancer cells. The authors observed that a Lys-rich peptide, L-K6 (IKKILSKIKKLLK-NH2) could internalize into MCF-7 cells without significant cell surface disruption. Interestingly, the authors also found that the internalized L-K6 could induce MCF-7 cell death without significant cytoskeleton disruption and mitochondrial impairment. Moreover, these findings indicated that the negatively-charged phosphatidylserine (PS) reported to be abundantly exposed on cancer cell surfaces [85,86] might contribute to the preferential binding of L-K6 to MCF-7 cancer cells. Furthermore, defensins are a large subfamily of natural cationic human AMP showing anticancer properties. Their roles in the immunomodulation of the innate and adaptive immune system have extensively been studied in previous works [87,88]. Defensins are categorized as either α-defensins or β-defensins, depending on their sequence homology and Cys connectivity (Figure 1), they are found to be rich in cationic residues (e.g., Lys and Arg) that enhances their immunomodulatory effects [89]. Papo et al. [90] constructed a short 15 amino acid peptide consisting of D- or L- amino acids (D-K_6L_9; LKLLKKLLKKLLKLL-NH2) and observed that D-K_6L_9 specifically targets and lyses cancer cells as well as inhibiting the growth of primary and metastatic tumors in in vivo models. The ability of D-K_6L_9 in selectively targeting cancer cells can be attributed to its cationicity whereby the peptide can bind to highly enriched surface-exposed PS in cancer cells as compared to non-cancer cells [91]. Similarly, Arg, which is a positively-charged residue that differs from Lys by its side chain, was shown to be the fourth informative amino acid form our analysis thus, reiterating the importance of the cationic property of APCs.

The second most important amino acid was Cys that afforded an MDGI value of 19.71. Cys is a highly reactive amino acid in which it contains a thiol group that is oxidized to form a dimer thus creating a disulfide bridge between two spatially adjacent Cys residues. The significance of disulfide bridges are seen in their strong, hydrophobic nature which are extremely important for the overall structural fold of the peptide, while also increasing the peptide stability against proteolytic degradation [92]. Almost a decade ago, GO-201 ($[R]_9$CQCRRKNYGQLDIFP) achieved in facilitating the inhibition of mucin-1 cytoplasmic domain (MUC1-CD) [93]. Thus far, many studies have established the function of the GO-201 peptide whereby its anticancer activity was demonstrated both in vitro and in vivo against human breast cancer, prostate cancer, chronic myelogenous leukemia and pancreatic cancer [93–96]. From these aforementioned studies, the researchers were able to determine key motifs responsible for the anticancer activity of GO-201 as CxC, thus reiterating the crucial role of Cys disulfide bridges [97]. Furthermore, Tyuryaeva et al. [98] explored the molecular mechanism of Cys-induced cytotoxicity and the resulting apoptosis of cancer cells by designing and testing peptide sequences containing the key CxC ($[R]_9$KCGCFF; named DIL) or CxxC ($[R]_9$FF**C**PH**C**YQ; named DOL) motifs. The results obtained from this study suggests that the disulfide oxidoreductase is responsible for the effective cytotoxtic action of Cys. Therefore, it can be inferred that peptides containing the CxC or CxxC motifs form disulfide bridges that pilot a partial loss of protein function and the rapid onset of apoptosis. It can therefore be anticipated that such peptide sequences could help to create a potentially

valuable new class of ACPs. Similarly, Schroeder et al. [99] ascertained a novel peptide GKAKCCK having pronounced antimicrobial activity. The authors discovered that the efficiency of this peptide was critically dependent on the two continuous cysteines and that upon mutating these residues to Ala or Ser, its activity was completely abrogated. Hence, given the well-established connection between antimicrobial and anticancer peptides [13,100], it is likely that the GKAKCCK peptide is able to induce antitumorigenic effects.

3.4. Mechanistic Interpretation of Informative PCP

The physicochemical properties of amino acids play an essential role as effective features for identifying and characterizing protein or peptide functions from their primary sequences [33,34,36,51,79]. It is well known that PCPs [39], such as molecular volume, exposure or accessible surface, polarity (hydrophobicity/hydrophilicity), charge/pK, hydrogen-bonding potential and so forth are correlated with the structure and function of the amino acid sequence [101]. Herein, we have obtained the ten top-ranked informative PCPs corresponding to their MDGI values as shown in Table 6.

Table 6. Ten top-ranked physiocochemical properties from the AAindex having the highest MDGI values.

Rank	AAindex	Categorized Property	Description	MDGI
1	ARGP820101	Hydrophobicity	Hydrophobicity index (Argos et al., 1982)	1.51
2	ARGP820102	Hydrophobicity	Signal sequence helical potential (Argos et al., 1982)	1.40
3	BHAR880101	Hydrophobicity	Average flexibility indices (Bhaskaran-Ponnuswamy, 1988)	1.08
4	ARGP820103	Hydrophobicity	Membrane-buried preference parameters (Argos et al., 1982)	1.04
5	BEGF750102	Beta propensity	Conformational parameter of beta-structure (Beghin-Dirkx, 1975)	1.00
6	BEGF750101	Alpha and turn propensities	Conformational parameter of inner helix (Beghin-Dirkx, 1975)	0.92
7	BIGC670101	Physicochemical properties	Residue volume (Bigelow, 1967)	0.91
8	BEGF750103	Alpha and turn propensities	Conformational parameter of beta-turn (Beghin-Dirkx, 1975)	0.87
9	BIOV880102	Hydrophobicity	Information value for accessibility; average fraction 23% (Biou et al., 1988)	0.85
10	ISOY800107	Hydrophobicity	Normalized relative frequency of double bend (Isogai et al., 1980)	0.83

The most important PCP is seen to be 'hydrophobicity' with an MDGI value of 1.51, which corresponds to the AAindex ARGP820101. In addition, from the analysis, 4 of the 10 top-ranked informative PCPs are related to α-helix properties of peptides consisting of ARGP820102, BEGF750101, ISOY800106 and CHOP780201 with corresponding MDGI values of 1.40, 0.92, 0.64 and 0.61, respectively. Furthermore, 3 of 10 top-ranked informative PCPs including the AAindices of BEGF750102, BEGF750103 and CHOP780101 are related to β-sheet properties with MDGI values of 1.00, 0.87 and 0.83, respectively. Consequently, it could be inferred that the hydrophobicity of a peptide along with its structural orientation (i.e., α-helix or β-sheet) are significant in determining the anticancer properties of peptides. The important PCPs of ACPs are analyzed and discussed below.

3.4.1. Hydrophobic Residues on α-Helical Structure Enhances Cell Penetrating Properties of Anticancer Peptides

It is commonly known that the hydrophobicity property is critical for the 3D structure of a protein. Many descriptors have been developed for measuring the hydrophobicity of proteins, depending on the presence of hydrophobic residues at specific locations in the peptide sequence. These indices are usually based on the 3D crystal structure of proteins coupled with the physicochemical properties of their side chains [102,103]. For instance, Argos et al. [104] utilized the hydrophobic index as measured by Nozaki and Tanford [105] for the 20 canonical amino acids for determining the role of hydrophobicity on the structure of amino acids. As mentioned above, the property of ARGP820101 with an MDGI value of 1.51 pertained to the hydrophobicity property. This result suggests that the presence of hydrophobic amino acid residues on the α-helix might be important for governing the anticancer activity of peptides. Moreover, the property of ARGP820103 (the fourth most important PCPs) is

also associated with hydrophobicity property. Many previous studies have described the need for symbiosis between hydrophobicity and cationicity of peptides in order to enhance its cell penetrating property and its ability to reach target cell as to exert its effects [106–109]. In addition, researchers have made efforts to assess the influence of hydrophobicity on anticancer effects. For instance, Huang et al. [108] systematically altered the hydrophobicity of a 26-residue α-helical peptide (peptide **P**; KWKSFLKTFKSAKKTVLHTALKAISS) by replacing Ala residues with the more hydrophobic Leu residues as to increase its hydrophobicity or by changing Leu residues to Ala residues as to decrease its hydrophobicity. On the basis of these results, the authors observed a correlation between higher hydrophobic variants whereby the Leu-substituted peptides (A12L and A20L) exhibited greater anticancer activity against cancer cells with higher IC_{50} values than peptide P. Thus, peptides with higher hydrophobicity are assumed to penetrate deeper into the hydrophobic core of the cell membrane, thereby causing pores or channels on the cancer cell membrane and exhibiting greater anticancer activity. From these results, it can be inferred that the presence of hydrophobic residues on the α-helical structure of ACPs can enhance its anticancer properties.

3.4.2. Peptides Forming an Amphipathic α-Helix Contributes to Anticancer Activity

Interestingly, four of the ten top-ranked informative PCPs consisting of ARGP820102, BEGF750101, ISOY800106 and CHOP780201 described helical properties of peptides such as helical potential [104], inner helix [110], frequency of helix end [111] and frequency of α-helix [112], respectively, as shown in Table 6. This supports the notion that the helical structure plays a principal role in governing the anticancer property of peptides. In addition, the α-helical conformation as adopted by many ACPs in biological membranes, is now regarded as a key determinant of anticancer activity [113]. BMAP-27 (GRFKRFRKKFKKLFKKLSPVIPLLHL) and BMAP-28 (GGLRSLGRKILRAWKKY GPIIVPIIRI) are well-known peptides derived from bovine sources having a cationic NH_2 terminal that forms an amphipathic α-helix. They exert their cytotoxic activity against human leukemia cells by enhancing membrane permeability, which allows for an influx of Ca++ ions, thereby leading to apoptosis via DNA fragmentation [114]. In addition, the structures of Cecropin A (KWKLFKKIEKVGQNIRDGIIKAGPAVAVVGQATQIAK) and Cecropin B (KWKVFKKIEKMGRNIRNGIVKAGPAIAVLGEAKAL), which were first derived from insects, primarily consisted of two α-helix [115,116] while its amphipathic N-terminal (i.e., that is capable of interacting with anionic membrane components) is responsible for mediating the cytotoxic activity against cancer cells. On the other hand, the C-terminal is hydrophobic and it is postulated to facilitate peptide entry into the membrane thereby enabling oligomerization to occur, which subsequently leads to leakage of cell components due to pore formation and eventual cell death [117]. In addition, Srisailam et al. [118] designed a custom lytic peptide by modifying the cecropin B (CB) to contain two identical hydrophobic segments on both the N and the C terminals, thus disrupting its normal α-helical structure. The resulting peptide CB-3, was unable to effectively lyse cell membranes on cancer cells as well as bacterial cells as compared to the control peptide, therefore highlighting the significance of the peptide structure on its function. Interestingly, Moore et al. [119] uncovered that CB exhibited anticancer activities when tested against colon adenocarcinoma cells in vivo and cytotoxicity against multidrug-resistant breast and ovarian cancer cell lines in vitro.

Similarly, Magainins isolated from the skin of *Xenopus laevis* is comprised of a α-helical secondary structure with separate cationic and hydrophobic faces. Magainin-2 (GIGKFLHSAKKFGKAFVGEIMNS) causes lysis of both hematopoietic and solid tumor cell lines as while observing selective cytotoxic activity against several human bladder cell lines [120]. In addition, it was perceived using florescence spectrometry that magainins could lyse tumor cells by forming ion-conducting α-helical channels in the cancer cell membrane [121]. Melittin (GIGAVLKVLTTGLPALISWIKRKRQQ) is another well-known ACP that is derived from the venom of European honeybee (*Apis mellifera*) [122]. The N-terminal region of melittin is largely hydrophobic whereas the C-terminal contains positively-charged amino acid residues.

Melittin form channels in lipid bilayers and is lytic against both cancer cells and normal cells [123]. Self-association of amphipathic α-helical monomers of melittin is largely suggested to be responsible for perturbations of membrane integrity thereby leading to cellular lysis via the barrel stave mechanism [124]. Furthermore, the only human cathelicidin-derived cationic ACP to date, LL-37 (LLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES) is present throughout the body and assumes a α-helical conformation that is comprised of both cationic and hydrophobic faces oriented in a parallel manner with lipid membranes. As such these peptides are suggested to engage in a carpet-like mechanism for its cytotoxicity [125]. However, unlike other ACPs, LL-37 is toxic to eukaryotic cells at slightly higher concentrations and thus lacks the selectivity of an ACP. Nevertheless, Okumura et al. [126] showed that a C-terminal fragment of LL-37 (hCAP18) was able to induce mitochondrial depolarization and apoptosis in human oral squamous carcinoma cells but not in healthy human gingival fibroblasts and human keratinocyte cells. In summary, it is evident that the amphipathicity associated with certain α-helical structures is pivotal for the elicitation of peptide anticancer activities.

3.4.3. Peptides Forming β-Sheet Are Vital to Peptide Structure and Contribute to Anticancer Activity

As seen in Table 6, 3 out of the 20 top-ranked informative PCPs pertain to β-sheet structure (i.e., BEGF750102, BEGF750103 and CHOP78010) which correspond to properties such as conformational parameter of β-structure, conformational parameter of β-turn and normalized frequency of β-turn [110], respectively. Although, most of the well-known APCs are α-helical structures, some β-sheet ACPs have also been extensively studied. Many studies showed that defensins contain a group of closely related Cys-Arg rich cationic peptides. Out of these, human α- and β-defensins remain the most studied, comprising of six conserved Cys residues that form three intramolecular disulfide bridges between the N-terminal and the C-terminal regions of the peptide [127,128]. In the studies of [129,130], human neutrophil peptides (HNPs) -1 (ACYCRIPACIAGERRYGTCIYQGRLWAFCC), -2 (CYCRIPACIAGERRYGTCIYQGRLWAFCC) and -3 (DCYCRIPACIAGERRYGTCIY QGRLWAFCC) were shown to possess cytotoxic effects against several human and mouse tumor cells such as human B-lymphoma cells, human oral squamous carcinoma cells, and mouse teratocarcinoma cells. HNPs were also shown to kill cancer cells through membrane binding mediated by electrostatic interactions, followed by rapid collapse of the membrane potential and loss of membrane integrity [130,131]. In addition, Mader et al. [100,132] reported that bovine lactoferricin (LfcinB; FKCRRWQWRMKKLGAPSITCVRRAF) isolated from cow's milk, represents a cationic ACP with amphipathic β-sheet configuration displaying anticancer activity against leukemia cells and various other carcinomas [100,132]. Moreover, LfcinB not only exerts its action by inducing apoptosis through direct disruption of the mitochondrial membrane, but is also capable of lysing the membrane itself, depending on the cancer cell type [133,134].

Furthermore, isolated from the hemocytes of horseshoe crab (*Tachypleus tridentatus*), the ACP tachyplesin I (KWCFRVCYRGICYRRCR) is arranged in two anti-parallel β-sheets that are held in place by two disulfide bonds [135]. Chen et al. [136] discovered that tachyplesin I killed cancer cells in a unique way by binding to hyaluronan on human prostate carcinoma cells, as well as to the C1q complement component in human serum, leading to the activation of the classical immune complement pathway. As a result, complement-mediated lysis of tachyplesin I-coated cancer cells was achieved. In a separate study, Adamia et al. [137] observed that the binding activity of tachyplesin I to C1q was dependent on its secondary structure which consists of two disulfide bonds formed by Cys residues at positions 3–16 and 7–12. Previously it has also been demonstrated that, the deletion of Cys residues from tachyplesin I altered its β-sheet structure to that of a linear form thus, disrupting its activity [138]. Therefore, from the aforementioned studies, it can be postulated that β-sheet is vital for maintaining the stability of peptide structures. As noticed in Table 5, our result of informative amino acids is consistent with the above-mentioned studies, where Cys is the second most important amino acid having a MDGI value of 19.71.

Nevertheless, further research is necessary to determine the key amino acids that could transform non-ACPs into ACPs and thus, allow the generation of a library of potentially novel ACPs. The design of effective ACPs from non-ACPs could possibly be based on the improvement of physicochemical properties, like hydrophobicity, positive charge, replacing the amino acid used or adding a functional motif with better activity towards the targeted cancer cells. In that regard, Tada et al. [139] revealed that the replacement of amino acid His to Arg in epidermal growth factor receptor (EGFR) binding peptide increased its anti-tumor activity via stronger binding affinity to the EGFR on cancer cells. Furthermore, the utilization of a motif containing the amino acids Trp, Met and Trp (WMW) was able to sensitize apoptosis and reduce the migratory effect on cancer cells which in turn could synergize the anti-cancer activity of peptides [140]. Therefore, it is of paramount importance to explore this area further.

3.5. Interpretable Rules Acquisition Obtained from the RF Model

Interpretable rules crucial for discriminating ACPs from non-ACPs were derived from the RF model built using AAC features on the benchmark dataset known as IR-ACP. The main advantages of these constructed rules are two folds: (i) to demonstrate which amino acid or which combination of amino acids are effective for ACP prediction, and (ii) to simply discriminate ACPs from non-ACPs without the need to go through the mathematical and computational details. Tables 7 and A2 shows the twenty interpretable rules that were important for ACPs and eleven that were important for non-ACPs.

Table 7. Eight if-then rules for the prediction of anticancer peptides using random forest and amino acid composition.

No.	Rule	Covered Samples	Misclassified Sample	Ac (%)
1	G > 0.041 and I > 0.0615 and L ≤ 0.1715 and K > 0.0385 and M ≤ 0.027	48	1	97.92
2	R ≤ 0.0515 and Q ≤ 0.026 and K > 0.094	48	1	97.92
3	C > 0.1145 and P ≤ 0.073	35	0	100.00
4	L ≤ 0.093 and F > 0.0715 and S ≤ 0.152	33	4	87.88
5	A ≤ 0.0145 and Q ≤ 0.026	26	4	84.62
6	R ≤ 0.0665 and E ≤ 0.044 and H > 0.052	21	4	80.95
7	I > 0.108 and K > 0.055	42	6	85.71
8	E ≤ 0.0545 and G > 0.0365 and K > 0.108 and M ≤ 0.04	49	1	97.96

Table 7 presents eight interpretable rules that were important for identifying ACPs and each rule comprises a different number of criteria. If a query peptide meets all of the criteria in at least one of the eight rules, then it will be predicted as an ACP. As observed in Table 7, four rules afforded a prediction accuracy larger than 95%, namely rules #1, #2, #3 and #8, while the remaining rules provided acceptable results with accuracy ranging between 80.79% and 87.88%. These results indicated that the constructed rules are robust, both in terms of their accuracy and interpretability for predicting and characterizing ACPs.

3.6. ACPred Web Server

In order to make the prediction model presented herein a practical tool that can be widely used by the scientific community, we have constructed a web server called the ACPred using the best model as described in previous sections. The web interface has been established using the Shiny package under the R programming environment. The web server is freely accessible at http://codes.bio/acpred/. Screenshots of the ACPred web server is shown in Figure 5 in which panel A shows the web server prior to submission of input data and panel B shown the web server after the prediction has been made.

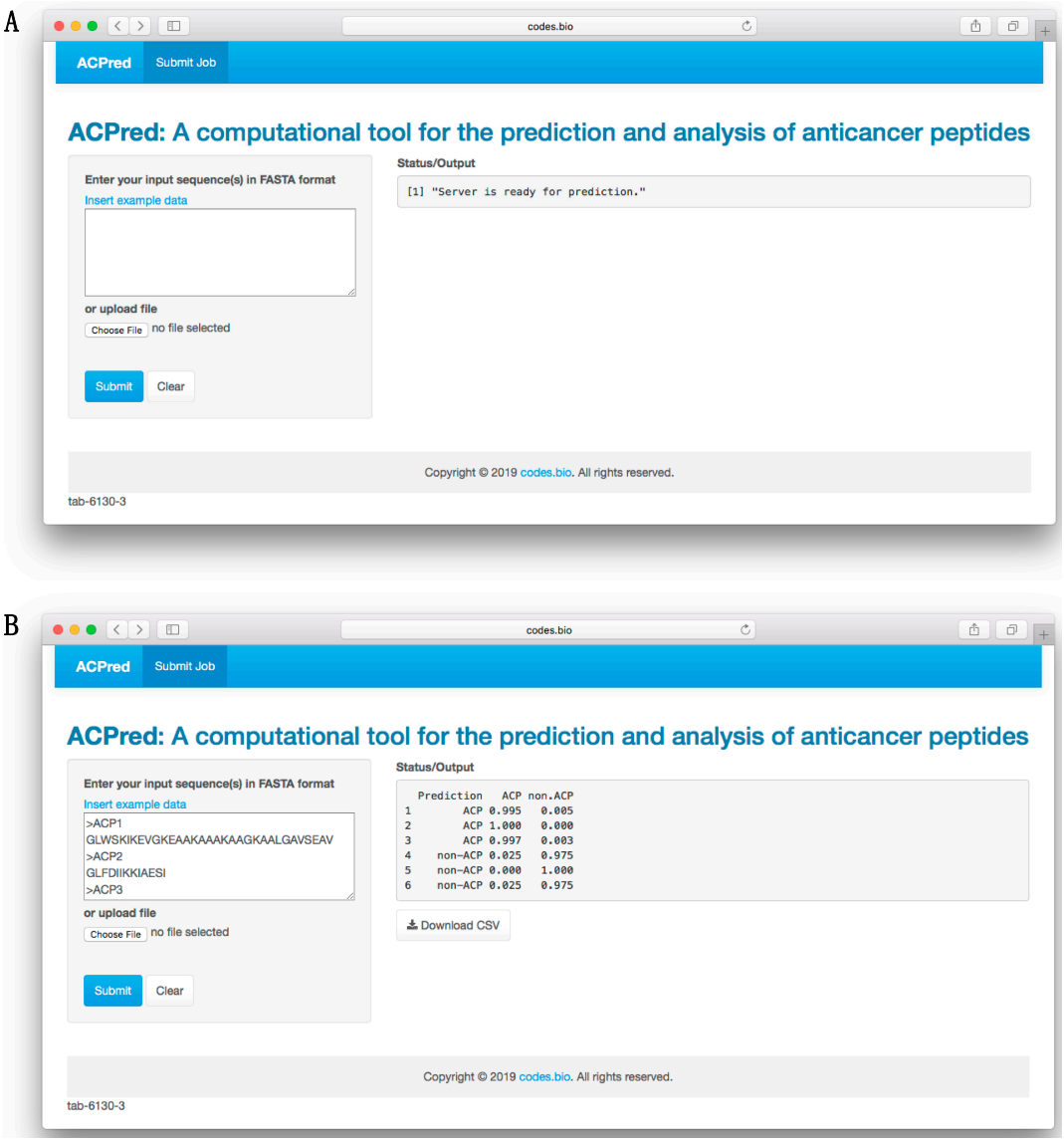

Figure 5. Screenshots of the ACPred web server before (**A**) and after (**B**) submission of sequence data for prediction.

Briefly, a step-by-step guide on using the web server is given below:

- Step 1. Enter the following URL into the web browser, http://codes.bio/acpred/.
- Step 2. Either enter the query peptide sequence into the Input box or upload the sequence file by clicking on the "Choose file" button (i.e., found below the "Enter your input sequence(s) in FASTA format heading").
- Step 3. Click on the "Submit" button to initiate the prediction process.
- Step 4. Prediction results are automatically displayed in the grey box found below the "Status/Output" heading. Typically, it takes a few seconds for the server to process the task. Users can also download the prediction results as a CSV file by pressing on the "Download CSV button".

Additionally, users can also run a local copy of ACPred on their own computer by using the following one-line code in an R environment:

shiny::runGitHub('acpred-webserver', 'chaninlab')

However, prior to running the aforementioned code, it is recommended that users first install the prerequisite R packages. This can be performed by using the following code:

install.packages(c('shiny', 'shinyjs', 'shinythemes', 'protr', 'seqinr', 'caret', 'markdown'))

4. Conclusions

There is increasing evidence indicating that ACPs could be a potential, valuable therapeutic agent for cancer treatment. With the avalanche of peptide sequences generated from many sources, there is significant demand for computational methods that can support the timely discrimination of ACPs from other peptides, as well as offer insights from analysis using only their primary sequences. In this study, we have developed a the ACPred model in a systematic manner by taking advantage of two powerful machine learning methods consisting of support vector machine (SVM) and random forest (RF) for the prediction and characterization of anticancer activities of peptides. The best model for further development of ACPred was achieved by using the SVM method coupled with the combination of amino acid composition and pseudo amino acid composition features. By comparing with the state-of-the-art method, the TargetACP, it was found that the performance of ACPred was comparable to that of TargetACP (95.61% vs. 96.22% accuracies) as assessed by the rigorous jackknife test. Moreover, the feature importance analysis and interpretable rules extraction were carried out by using the RF model to identify which feature was the most important in predicting ACPs while also providing a better understanding of the anticancer activities of investigated peptides. Results from model interpretation revealed that the hydrophobic residues on the α-helical structure, Cys residues on the β-sheet structure and the formation of amphipathic α-helix played a crucial role in the anticancer activity of the investigated peptides. Finally, a web server named ACPred was established and made freely available online at https://codes.bio/acpred/.

Author Contributions: W.S. conceived, designed, performed and analyzed the experiments. N.S. and C.N. analyzed the data. W.S., N.S. and V.P. drafted the manuscript. W.S. and C.N. contributed the code for constructing the web server. C.N. vetted the manuscript. All authors read and approved the manuscript.

Funding: This work is supported by the Office of Higher Education Commission and the Thailand Research Fund (No. MRG6180226); the New Researcher Grant (A31/2561) from Mahidol University; and the Center of Excellence on Medical Biotechnology (CEMB), the S&T Postgraduate Education and Research Development Office (PERDO) and the Office of Higher Education Commission (OHEC), Thailand.

Acknowledgments: We thank the reviewers for their great comments.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Summary of MODI index as derived from various types of peptide features on the benchmark dataset.

Feature	MODI
AAC	0.868
DPC	0.836
PCP	0.722
PseAAC	0.859
Am-PseAAC	0.859
AAC + PseAAC	0.859
AAC + Am-PseAAC	0.859
PseAAC + Am-PseAAC	0.856
AAC + PseAAC + Am-PseAAC	0.856

Table A2. Twenty if-then rules for discriminating ACP from non-ACP using random forest and amino acid composition.

% Covered Samples	Rule	Prediction Result
14.33	R > 0.017 and C ≤ 0.1145 and H ≤ 0.0645 and L > 0.1605 and K ≤ 0.1035	non-ACP
13.74	G > 0.041 and I > 0.0615 and L ≤ 0.1715 and K > 0.0385 and M ≤ 0.027	ACP
9.65	E > 0.058 and G ≤ 0.068	non-ACP
7.89	R ≤ 0.0515 and Q ≤ 0.026 and K > 0.094	ACP
7.89	C ≤ 0.1145 and H ≤ 0.0645 and K ≤ 0.1055 and W > 0.028	non-ACP
5.26	C > 0.1145 and P ≤ 0.073	ACP
4.97	N > 0.019 and K ≤ 0.1705 and M > 0.037	non-ACP
4.09	Q > 0.098 and K ≤ 0.0575	non-ACP
3.22	C > 0.015 and C ≤ 0.043	non-ACP
3.80	L ≤ 0.093 and F > 0.0715 and S ≤ 0.152	ACP
6.43	R > 0.013 and R ≤ 0.1345 and C ≤ 0.1205 and H ≤ 0.0465 and F ≤ 0.1345	non-ACP
2.63	A ≤ 0.0145 and Q ≤ 0.026	ACP
3.22	I ≤ 0.0335 and P > 0.074	non-ACP
2.34	R ≤ 0.0665 and E ≤ 0.044 and H > 0.052	ACP
1.75	N ≤ 0.0565 and H ≤ 0.0415 and K≤ 0.0335	non-ACP
1.46	I > 0.108 and K > 0.055	ACP
1.17	E ≤ 0.0545 and G > 0.0365 and K > 0.108 and M ≤ 0.04	ACP
2.05	I ≤ 0.0665 and F ≤ 0.006	non-ACP
2.05	S ≤ 0.04	non-ACP
2.05	Else	ACP

References

1. WHO Cancer. Available online: https://www.who.int/news-room/fact-sheets/detail/cancer (accessed on 30 January 2019).
2. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [CrossRef] [PubMed]
3. Ferlay, J.; Colombet, M.; Soerjomataram, I.; Mathers, C.; Parkin, D.M.; Pineros, M.; Znaor, A.; Bray, F. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int. J. Cancer* **2018**, *144*, 1941–1953. [CrossRef]
4. Craik, D.J.; Fairlie, D.P.; Liras, S.; Price, D. The future of peptide-based drugs. *Chem. Biol. Drug Des.* **2013**, *81*, 136–147. [CrossRef] [PubMed]
5. Vlieghe, P.; Lisowski, V.; Martinez, J.; Khrestchatisky, M. Synthetic therapeutic peptides: Science and market. *Drug Discov. Today* **2010**, *15*, 40–56. [CrossRef]
6. Lau, J.L.; Dunn, M.K. Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorg. Med. Chem.* **2018**, *26*, 2700–2707. [CrossRef] [PubMed]
7. Fosgerau, K.; Hoffmann, T. Peptide therapeutics: Current status and future directions. *Drug Discov. Today* **2015**, *20*, 122–128. [CrossRef]
8. O'Brien-Simpson, N.M.; Hoffmann, R.; Chia, C.S.B.; Wade, J.D. Editorial: Antimicrobial and Anticancer Peptides. *Front. Chem.* **2018**, *6*, 13. [CrossRef]
9. Jin, G.; Weinberg, A. Human antimicrobial peptides and cancer. *Semin. Cell Dev. Biol.* **2018**, *88*, 156–162. [CrossRef] [PubMed]
10. Karpinski, T.M.; Adamczak, A. Anticancer Activity of Bacterial Proteins and Peptides. *Pharmaceutics* **2018**, *10*, 54. [CrossRef]
11. Usmani, S.S.; Bedi, G.; Samuel, J.S.; Singh, S.; Kalra, S.; Kumar, P.; Ahuja, A.A.; Sharma, M.; Gautam, A.; Raghava, G.P.S. THPdb: Database of FDA-approved peptide and protein therapeutics. *PLoS ONE* **2017**, *12*, e0181748. [CrossRef]
12. Deslouches, B.; Di, Y.P. Antimicrobial peptides with selective antitumor mechanisms: Prospect for anticancer applications. *Oncotarget* **2017**, *8*, 46635–46651. [CrossRef]
13. Gaspar, D.; Veiga, A.S.; Castanho, M.A. From antimicrobial to anticancer peptides. A review. *Front. Microbiol.* **2013**, *4*, 294. [CrossRef]
14. Ran, S.; Downes, A.; Thorpe, P.E. Increased exposure of anionic phospholipids on the surface of tumor blood vessels. *Cancer Res.* **2002**, *62*, 6132–6140.
15. Yoon, W.H.; Park, H.D.; Lim, K.; Hwang, B.D. Effect of *O*-glycosylated mucin on invasion and metastasis of HM7 human colon cancer cells. *Biochem. Biophys. Res. Commun.* **1996**, *222*, 694–699. [CrossRef]
16. Sok, M.; Sentjurc, M.; Schara, M. Membrane fluidity characteristics of human lung cancer. *Cancer Lett.* **1999**, *139*, 215–220. [CrossRef]

17. Chaudhary, J.; Munshi, M. Scanning electron microscopic analysis of breast aspirates. *Cytopathology* **1995**, *6*, 162–167. [CrossRef]
18. Dobrzynska, I.; Szachowicz-Petelska, B.; Sulkowski, S.; Figaszewski, Z. Changes in electric charge and phospholipids composition in human colorectal cancer cells. *Mol. Cell. Biochem.* **2005**, *276*, 113–119. [CrossRef]
19. Domalaon, R.; Findlay, B.; Ogunsina, M.; Arthur, G.; Schweizer, F. Ultrashort cationic lipopeptides and lipopeptoids: Evaluation and mechanistic insights against epithelial cancer cells. *Peptides* **2016**, *84*, 58–67. [CrossRef]
20. Gaspar, D.; Freire, J.M.; Pacheco, T.R.; Barata, J.T.; Castanho, M.A. Apoptotic human neutrophil peptide-1 anti-tumor activity revealed by cellular biomechanics. *Biochim. Biophys. Acta* **2015**, *1853*, 308–316. [CrossRef]
21. Figueiredo, C.R.; Matsuo, A.L.; Massaoka, M.H.; Polonelli, L.; Travassos, L.R. Anti-tumor activities of peptides corresponding to conserved complementary determining regions from different immunoglobulins. *Peptides* **2014**, *59*, 14–19. [CrossRef]
22. Riedl, S.; Zweytick, D.; Lohner, K. Membrane-active host defense peptides–challenges and perspectives for the development of novel anticancer drugs. *Chem. Phys. Lipids* **2011**, *164*, 766–781. [CrossRef] [PubMed]
23. Tyagi, A.; Kapoor, P.; Kumar, R.; Chaudhary, K.; Gautam, A.; Raghava, G. In silico models for designing and discovering novel anticancer peptides. *Sci. Rep.* **2013**, *3*, 2984. [CrossRef]
24. Hajisharifi, Z.; Piryaiee, M.; Beigi, M.M.; Behbahani, M.; Mohabatkar, H. Predicting anticancer peptides with Chou' s pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.* **2014**, *341*, 34–40. [CrossRef] [PubMed]
25. Vijayakumar, S.; Lakshmi, P. ACPP: A web server for prediction and design of anti-cancer peptides. *Int. J. Pept. Res. Ther.* **2015**, *21*, 99–106. [CrossRef]
26. Chen, W.; Ding, H.; Feng, P.; Lin, H.; Chou, K.-C. iACP: A sequence-based tool for identifying anticancer peptides. *Oncotarget* **2016**, *7*, 16895. [CrossRef] [PubMed]
27. Li, F.-M.; Wang, X.-Q. Identifying anticancer peptides by using improved hybrid compositions. *Sci. Rep.* **2016**, *6*, 33910. [CrossRef]
28. Akbar, S.; Hayat, M.; Iqbal, M.; Jan, M.A. iACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space. *Artif. Intell. Med.* **2017**, *79*, 62–70. [CrossRef]
29. Manavalan, B.; Basith, S.; Shin, T.H.; Choi, S.; Kim, M.O.; Lee, G. MLACP: Machine-learning-based prediction of anticancer peptides. *Oncotarget* **2017**, *8*, 77121. [CrossRef]
30. Xu, L.; Liang, G.; Wang, L.; Liao, C. A Novel Hybrid Sequence-Based Model for Identifying Anticancer Peptides. *Genes* **2018**, *9*, 158. [CrossRef]
31. Kabir, M.; Arif, M.; Ahmad, S.; Ali, Z.; Swati, Z.N.K.; Yu, D.-J. Intelligent computational method for discrimination of anticancer peptides by incorporating sequential and evolutionary profiles information. *Chemom. Intell. Lab. Syst.* **2018**, *182*, 158–165. [CrossRef]
32. Shoombuatong, W.; Schaduangrat, N.; Nantasenamat, C. Unraveling the bioactivity of anticancer peptides as deduced from machine learning. *EXCLI J.* **2018**, *17*, 734.
33. Pratiwi, R.; Malik, A.A.; Schaduangrat, N.; Prachayasittikul, V.; Wikberg, J.E.; Nantasenamat, C.; Shoombuatong, W. CryoProtect: A Web Server for Classifying Antifreeze Proteins from Nonantifreeze Proteins. *J. Chem.* **2017**, *2017*. [CrossRef]
34. Win, T.S.; Malik, A.A.; Prachayasittikul, V.; S Wikberg, J.E.; Nantasenamat, C.; Shoombuatong, W. HemoPred: A web server for predicting the hemolytic activity of peptides. *Future Med. Chem.* **2017**, *9*, 275–291. [CrossRef]
35. Shoombuatong, W.; Schaduangrat, N.; Nantasenamat, C. Towards understanding aromatase inhibitory activity via QSAR modeling. *EXCLI J.* **2018**, *17*, 688.
36. Win, T.S.; Schaduangrat, N.; Prachayasittikul, V.; Nantasenamat, C.; Shoombuatong, W. PAAP: A web server for predicting antihypertensive activity of peptides. *Future Med. Chem.* **2018**, *10*, 1749–1767. [CrossRef]
37. Shoombuatong, W.; Mekha, P.; Chaijaruwanich, J. Sequence based human leukocyte antigen gene prediction using informative physicochemical properties. *Int. J. Data Min. Bioinform.* **2015**, *13*, 211–224. [CrossRef]
38. Simeon, S.; Shoombuatong, W.; Anuwongcharoen, N.; Preeyanon, L.; Prachayasittikul, V.; Wikberg, J.E.; Nantasenamat, C. osFP: A web server for predicting the oligomeric states of fluorescent proteins. *J. Cheminform.* **2016**, *8*, 72. [CrossRef]

39. Kawashima, S.; Kanehisa, M. AAindex: Amino acid index database. *Nucleic Acids Res.* **2000**, *28*, 374. [CrossRef]
40. Chou, K.-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2004**, *21*, 10–19. [CrossRef]
41. Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **2011**, *273*, 236–247. [CrossRef]
42. Shen, H.-B.; Chou, K.-C. PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.* **2008**, *373*, 386–388. [CrossRef]
43. Xiao, N.; Cao, D.-S.; Zhu, M.-F.; Xu, Q.-S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* **2015**, *31*, 1857–1859. [CrossRef] [PubMed]
44. Golbraikh, A.; Muratov, E.; Fourches, D.; Tropsha, A. Data set modelability by QSAR. *J. Chem. Inf. Model.* **2014**, *54*, 1–4. [CrossRef]
45. Shoombuatong, W.; Prathipati, P.; Prachayasittikul, V.; Schaduangrat, N.; Malik, A.A.; Pratiwi, R.; Wanwimolruk, S.; Wikberg, J.E.; Gleeson, M.P.; Spjuth, O. Towards Predicting the Cytochrome P450 Modulation: From QSAR to Proteochemometric Modeling. *Curr. Drug Metab.* **2017**, *18*, 540–555. [CrossRef]
46. Shoombuatong, W.; Hongjaisee, S.; Barin, F.; Chaijaruwanich, J.; Samleerat, T. HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees. *Comput. Biol. Med.* **2012**, *42*, 885–889. [CrossRef]
47. Simeon, S.; Li, H.; Win, T.S.; Malik, A.A.; Kandhro, A.H.; Piacham, T.; Shoombuatong, W.; Nuchnoi, P.; Wikberg, J.E.; Gleeson, M.P. PepBio: Predicting the bioactivity of host defense peptides. *RSC Adv.* **2017**, *7*, 35119–35134. [CrossRef]
48. Shoombuatong, W.; Prachayasittikul, V.; Prachayasittikul, V.; Nantasenamat, C. Prediction of aromatase inhibitory activity using the efficient linear method (ELM). *EXCLI J.* **2015**, *14*, 452.
49. Shoombuatong, W.; Mekha, P.; Waiyamai, K.; Cheevadhanarak, S.; Chaijaruwanicha, J. Prediction of human leukocyte antigen gene using k-nearest neighbour classifier based on spectrum kernel. *ScienceAsia* **2013**, *39*, 42–49. [CrossRef]
50. Shoombuatong, W.; Traisathit, P.; Prasitwattanaseree, S.; Tayapiwatana, C.; Cutler, R.; Chaijaruwanich, J. Prediction of the disulphide bonding state of cysteines in proteins using Conditional Random Fields. *Int. J. Data Min. Bioinform.* **2011**, *5*, 449–464. [CrossRef]
51. Charoenkwan, P.; Shoombuatong, W.; Lee, H.-C.; Chaijaruwanich, J.; Huang, H.-L.; Ho, S.-Y. SCMCRYS: Predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. *PLoS ONE* **2013**, *8*, e72368. [CrossRef] [PubMed]
52. Anuwongcharoen, N.; Shoombuatong, W.; Tantimongcolwat, T.; Prachayasittikul, V.; Nantasenamat, C. Exploring the chemical space of influenza neuraminidase inhibitors. *PeerJ* **2016**, *4*, e1958. [CrossRef]
53. Shoombuatong, W.; Prachayasittikul, V.; Anuwongcharoen, N.; Songtawee, N.; Monnor, T.; Prachayasittikul, S.; Prachayasittikul, V.; Nantasenamat, C. Navigating the chemical space of dipeptidyl peptidase-4 inhibitors. *Drug Des. Dev. Ther.* **2015**, *9*, 4515.
54. Suvannang, N.; Preeyanon, L.; Malik, A.A.; Schaduangrat, N.; Shoombuatong, W.; Worachartcheewan, A.; Tantimongcolwat, T.; Nantasenamat, C. Probing the origin of estrogen receptor alpha inhibition via large-scale QSAR study. *RSC Adv.* **2018**, *8*, 11344–11356. [CrossRef]
55. Simeon, S.; Anuwongcharoen, N.; Shoombuatong, W.; Malik, A.A.; Prachayasittikul, V.; Wikberg, J.E.; Nantasenamat, C. Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking. *PeerJ* **2016**, *4*, e2322. [CrossRef]
56. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
57. Breiman, L. *Classification and Regression Trees*; Routledge: London, UK, 2017.
58. Breiman, L. randomForest: Breiman and Cutler's Random Forests for Classification and Regression. R Package Version. 2006. Available online: http://stat-www.berkeley.edu/users/breiman/RandomForests (accessed on 30 January 2019).
59. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
60. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
61. Drucker, H.; Burges, C.J.; Kaufman, L.; Smola, A.J.; Vapnik, V. Support Vector Regression Machines. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1997; pp. 155–161.
62. Joachims, T. Svmlight: Support Vector Machine. SVM-Light Support. Vector Machine. University of Dortmund, 1999, 19. Available online: http://svmlight.joachims.org/ (accessed on 30 January 2019).

63. Lameski, P.; Zdravevski, E.; Mingov, R.; Kulakov, A. SVM parameter tuning with grid search and its impact on reduction of model over-fitting. In *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*; Springer: Berlin, Germany, 2015; pp. 464–474.
64. Tung, C.-W.; Ziehm, M.; Kämper, A.; Kohlbacher, O.; Ho, S.-Y. POPISK: T-cell reactivity prediction using support vector machines and string kernels. *BMC Bioinform.* **2011**, *12*, 446. [CrossRef]
65. Calle, M.L.; Urrea, V. Letter to the editor: Stability of random forest importance measures. *Brief. Bioinform.* **2010**, *12*, 86–89. [CrossRef]
66. Shoombuatong, W.; Huang, H.-L.; Chaijaruwanich, J.; Charoenkwan, P.; Lee, H.-C.; Ho, S.-Y. Predicting Protein Crystallization Using a Simple Scoring Card Method. In Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Singapore, 16–19 April 2013; pp. 23–30.
67. Hasan, M.M.; Yang, S.; Zhou, Y.; Mollah, M.N.H. SuccinSite: A computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. *Mol. Biosyst.* **2016**, *12*, 786–795. [CrossRef]
68. Worachartcheewan, A.; Shoombuatong, W.; Pidetcha, P.; Nopnithipat, W.; Prachayasittikul, V.; Nantasenamat, C. Predicting metabolic syndrome using the random forest method. *Sci. World J.* **2015**, *2015*, 58150. [CrossRef] [PubMed]
69. Chou, K.-C.; Shen, H.-B. Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* **2008**, *3*, 153. [CrossRef] [PubMed]
70. Chou, K.-C.; Shen, H.-B. Recent progress in protein subcellular location prediction. *Anal. Biochem.* **2007**, *370*, 1. [CrossRef]
71. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.-C.; Müller, M. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **2011**, *12*, 77. [CrossRef]
72. Liaw, C.; Tung, C.-W.; Ho, S.-Y. Prediction and analysis of antibody amyloidogenesis from sequences. *PLoS ONE* **2013**, *8*, e53235. [CrossRef]
73. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
74. Kumar, R.; Chaudhary, K.; Chauhan, J.S.; Nagpal, G.; Kumar, R.; Sharma, M.; Raghava, G.P. An in silico platform for predicting, screening and designing of antihypertensive peptides. *Sci. Rep.* **2015**, *5*, 12512. [CrossRef] [PubMed]
75. Shen, H.-B.; Chou, K.-C. Gneg-mPLoc: A top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. *J. Theor. Biol.* **2010**, *264*, 326–333. [CrossRef]
76. Chou, K.-C.; Shen, H.-B. Plant-mPLoc: A top-down strategy to augment the power for predicting plant protein subcellular localization. *PLoS ONE* **2010**, *5*, e11335. [CrossRef]
77. Hellberg, S.; Sjoestroem, M.; Skagerberg, B.; Wold, S. Peptide quantitative structure-activity relationships, a multivariate approach. *J. Med. Chem.* **1987**, *30*, 1126–1135. [CrossRef]
78. Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* **1998**, *41*, 2481–2491. [CrossRef]
79. Huang, H.-L. Propensity scores for prediction and characterization of bioluminescent proteins from sequences. *PLoS ONE* **2014**, *9*, e97158. [CrossRef] [PubMed]
80. Borrelli, A.; Tornesello, A.L.; Tornesello, M.L.; Buonaguro, F.M. Cell Penetrating Peptides as Molecular Carriers for Anti-Cancer Agents. *Molecules* **2018**, *23*, 295. [CrossRef]
81. Benabdelouahab, Y.; Muñoz-Moreno, L.; Frik, M.; de la Cueva-Alique, I.; El Amrani, M.A.; Contel, M.; Bajo, A.M.; Cuenca, T.; Royo, E. Hydrogen Bonding and Anticancer Properties of Water-Soluble Chiral p-Cymene RuII Compounds with Amino-Oxime Ligands. *Eur. J. Inorg. Chem.* **2015**, *2015*, 2295–2307. [CrossRef]
82. Carlberg, C. Current understanding of the function of the nuclear vitamin D receptor in response to its natural and synthetic ligands. *Recent Results Cancer Res.* **2003**, *164*, 29–42.
83. Gopal, R.; Park, S.C.; Ha, K.J.; Cho, S.J.; Kim, S.W.; Song, P.I.; Nah, J.W.; Park, Y.; Hahm, K.S. Effect of Leucine and Lysine substitution on the antimicrobial activity and evaluation of the mechanism of the HPA3NT3 analog peptide. *J. Pept. Sci.* **2009**, *15*, 589–594. [CrossRef]
84. Wang, C.; Dong, S.; Zhang, L.; Zhao, Y.; Huang, L.; Gong, X.; Wang, H.; Shang, D. Cell surface binding, uptaking and anticancer activity of L-K6, a lysine/leucine-rich peptide, on human breast cancer MCF-7 cells. *Sci. Rep.* **2017**, *7*, 8293. [CrossRef] [PubMed]

85. Riedl, S.; Rinner, B.; Asslaber, M.; Schaider, H.; Walzer, S.; Novak, A.; Lohner, K.; Zweytick, D. In search of a novel target—Phosphatidylserine exposed by non-apoptotic tumor cells and metastases of malignancies with poor treatment efficacy. *Biochim. Biophys. Acta* **2011**, *1808*, 2638–2645. [CrossRef] [PubMed]
86. Wang, C.; Chen, Y.W.; Zhang, L.; Gong, X.G.; Zhou, Y.; Shang, D.J. Melanoma cell surface-expressed phosphatidylserine as a therapeutic target for cationic anticancer peptide, temporin-1CEa. *J. Drug Target.* **2016**, *24*, 548–556. [CrossRef] [PubMed]
87. Yang, D.; Biragyn, A.; Hoover, D.M.; Lubkowski, J.; Oppenheim, J.J. Multiple roles of antimicrobial defensins, cathelicidins, and eosinophil-derived neurotoxin in host defense. *Annu. Rev. Immunol.* **2004**, *22*, 181–215. [CrossRef] [PubMed]
88. Yang, D.; Oppenheim, J.J. Antimicrobial proteins act as "alarmins" in joint immune defense. *Arthritis Rheum.* **2004**, *50*, 3401–3403. [CrossRef]
89. Zou, G.; de Leeuw, E.; Li, C.; Pazgier, M.; Li, C.; Zeng, P.; Lu, W.Y.; Lubkowski, J.; Lu, W. Toward understanding the cationicity of defensins. Arg and Lys versus their noncoded analogs. *J. Biol. Chem.* **2007**, *282*, 19653–19665. [CrossRef]
90. Papo, N.; Seger, D.; Makovitzki, A.; Kalchenko, V.; Eshhar, Z.; Degani, H.; Shai, Y. Inhibition of tumor growth and elimination of multiple metastases in human prostate and breast xenografts by systemic inoculation of a host defense-like lytic peptide. *Cancer Res.* **2006**, *66*, 5371–5378. [CrossRef]
91. Zwaal, R.F.; Comfurius, P.; Bevers, E.M. Surface exposure of phosphatidylserine in pathological cells. *Cell. Mol. Life Sci.* **2005**, *62*, 971–988. [CrossRef]
92. Tanabe, H.; Ayabe, T.; Maemoto, A.; Ishikawa, C.; Inaba, Y.; Sato, R.; Moriichi, K.; Okamoto, K.; Watari, J.; Kono, T.; et al. Denatured human alpha-defensin attenuates the bactericidal activity and the stability against enzymatic digestion. *Biochem. Biophys. Res. Commun.* **2007**, *358*, 349–355. [CrossRef]
93. Raina, D.; Ahmad, R.; Joshi, M.D.; Yin, L.; Wu, Z.; Kawano, T.; Vasir, B.; Avigan, D.; Kharbanda, S.; Kufe, D. Direct targeting of the mucin 1 oncoprotein blocks survival and tumorigenicity of human breast carcinoma cells. *Cancer Res.* **2009**, *69*, 5133–5141. [CrossRef]
94. Joshi, M.D.; Ahmad, R.; Yin, L.; Raina, D.; Rajabi, H.; Bubley, G.; Kharbanda, S.; Kufe, D. MUC1 oncoprotein is a druggable target in human prostate cancer cells. *Mol. Cancer Ther.* **2009**, *8*, 3056–3065. [CrossRef]
95. Yin, L.; Ahmad, R.; Kosugi, M.; Kawano, T.; Avigan, D.; Stone, R.; Kharbanda, S.; Kufe, D. Terminal differentiation of chronic myelogenous leukemia cells is induced by targeting of the MUC1-C oncoprotein. *Cancer Biol. Ther.* **2010**, *10*, 483–491. [CrossRef]
96. Banerjee, S.; Mujumdar, N.; Dudeja, V.; Mackenzie, T.; Krosch, T.K.; Sangwan, V.; Vickers, S.M.; Saluja, A.K. MUC1c regulates cell survival in pancreatic cancer by preventing lysosomal permeabilization. *PLoS ONE* **2012**, *7*, e43020. [CrossRef]
97. Raina, D.; Ahmad, R.; Rajabi, H.; Panchamoorthy, G.; Kharbanda, S.; Kufe, D. Targeting cysteine-mediated dimerization of the MUC1-C oncoprotein in human cancer cells. *Int. J. Oncol.* **2012**, *40*, 1643–1649.
98. Tyuryaeva, I.I.; Lyublinskaya, O.G.; Podkorytov, I.S.; Skrynnikov, N.R. Origin of anti-tumor activity of the cysteine-containing GO peptides and further optimization of their cytotoxic properties. *Sci. Rep.* **2017**, *7*, 40217. [CrossRef]
99. Schroeder, B.O.; Wu, Z.; Nuding, S.; Groscurth, S.; Marcinowski, M.; Beisner, J.; Buchner, J.; Schaller, M.; Stange, E.F.; Wehkamp, J. Reduction of disulphide bonds unmasks potent antimicrobial activity of human beta-defensin 1. *Nature* **2011**, *469*, 419–423. [CrossRef]
100. Hoskin, D.W.; Ramamoorthy, A. Studies on anticancer activities of antimicrobial peptides. *Biochim. Biophys. Acta* **2008**, *1778*, 357–375. [CrossRef]
101. Tsai, C.S. *Biomacromolecules: Introduction to Structure, Function and Informatics*; John Wiley & Sons: Hoboken, NJ, USA, 2007.
102. Roseman, M.A. Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds. *J. Mol. Biol.* **1988**, *200*, 513–522. [CrossRef]
103. Kyte, J.; Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol Biol* **1982**, *157*, 105–132. [CrossRef]
104. Argos, P.; Rao, J.K.; Hargrave, P.A. Structural prediction of membrane-bound proteins. *Eur. J. Biochem.* **1982**, *128*, 565–575. [CrossRef]
105. Nozaki, Y.; Tanford, C. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *J. Biol. Chem.* **1971**, *246*, 2211–2217.

106. Mader, J.S.; Hoskin, D.W. Cationic antimicrobial peptides as novel cytotoxic agents for cancer treatment. *Expert Opin. Investig. Drugs* **2006**, *15*, 933–946. [CrossRef]
107. Chen, C.; Yang, C.; Chen, Y.; Wang, F.; Mu, Q.; Zhang, J.; Li, Z.; Pan, F.; Xu, H.; Lu, J.R. Surface Physical Activity and Hydrophobicity of Designed Helical Peptide Amphiphiles Control Their Bioactivity and Cell Selectivity. *ACS Appl. Mater. Interfaces* **2016**, *8*, 26501–26510. [CrossRef]
108. Huang, Y.B.; Wang, X.F.; Wang, H.Y.; Liu, Y.; Chen, Y. Studies on mechanism of action of anticancer peptides by modulation of hydrophobicity within a defined structural framework. *Mol. Cancer Ther.* **2011**, *10*, 416–426. [CrossRef]
109. Park, J.H.; Kim, H.A.; Cho, S.H.; Lee, M. Characterization of hydrophobic anti-cancer drug-loaded amphiphilic peptides as a gene carrier. *J. Cell. Biochem.* **2012**, *113*, 1645–1653. [CrossRef]
110. Chou, P.Y.; Fasman, G.D. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry* **1974**, *13*, 211–222. [CrossRef]
111. Isogai, Y.; Nemethy, G.; Rackovsky, S.; Leach, S.J.; Scheraga, H.A. Characterization of multiple bends in proteins. *Biopolymers* **1980**, *19*, 1183–1210. [CrossRef] [PubMed]
112. Chou, P.Y.; Fasman, G.D. Empirical predictions of protein conformation. *Annu. Rev. Biochem.* **1978**, *47*, 251–276. [CrossRef] [PubMed]
113. Vermeer, L.S.; Lan, Y.; Abbate, V.; Ruh, E.; Bui, T.T.; Wilkinson, L.J.; Kanno, T.; Jumagulova, E.; Kozlowska, J.; Patel, J.; et al. Conformational flexibility determines selectivity and antibacterial, antiplasmodial, and anticancer potency of cationic alpha-helical peptides. *J. Biol. Chem.* **2012**, *287*, 34120–34133. [CrossRef] [PubMed]
114. Risso, A.; Zanetti, M.; Gennaro, R. Cytotoxicity and apoptosis mediated by two peptides of innate immunity. *Cell. Immunol.* **1998**, *189*, 107–115. [CrossRef]
115. Holak, T.A.; Engstrom, A.; Kraulis, P.J.; Lindeberg, G.; Bennich, H.; Jones, T.A.; Gronenborn, A.M.; Clore, G.M. The solution conformation of the antibacterial peptide cecropin A: A nuclear magnetic resonance and dynamical simulated annealing study. *Biochemistry* **1988**, *27*, 7620–7629. [CrossRef]
116. Hung, S.C.; Wang, W.; Chan, S.I.; Chen, H.M. Membrane lysis by the antibacterial peptides cecropins B1 and B3: A spin-label electron spin resonance study on phospholipid bilayers. *Biophys. J.* **1999**, *77*, 3120–3133. [CrossRef]
117. Merrifield, R.B.; Merrifield, E.L.; Juvvadi, P.; Andreu, D.; Boman, H.G. Design and synthesis of antimicrobial peptides. *Ciba Found. Symp.* **1994**, *186*, 5–20, discussion 20-6.
118. Srisailam, S.; Kumar, T.K.; Arunkumar, A.I.; Leung, K.W.; Yu, C.; Chen, H.M. Crumpled structure of the custom hydrophobic lytic peptide cecropin B3. *Eur. J. Biochem.* **2001**, *268*, 4278–4284. [CrossRef]
119. Moore, A.J.; Devine, D.A.; Bibby, M.C. Preliminary experimental anticancer activity of cecropins. *Pept. Res.* **1994**, *7*, 265–269. [PubMed]
120. Lehmann, J.; Retz, M.; Sidhu, S.S.; Suttmann, H.; Sell, M.; Paulsen, F.; Harder, J.; Unteregger, G.; Stockle, M. Antitumor activity of the antimicrobial peptide magainin II against bladder cancer cell lines. *Eur. Urol.* **2006**, *50*, 141–147. [CrossRef]
121. Cruciani, R.A.; Barker, J.L.; Zasloff, M.; Chen, H.C.; Colamonici, O. Antibiotic magainins exert cytolytic activity against transformed cell lines through channel formation. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 3792–3796. [CrossRef] [PubMed]
122. Gauldie, J.; Hanson, J.M.; Shipolini, R.A.; Vernon, C.A. The structures of some peptides from bee venom. *Eur. J. Biochem.* **1978**, *83*, 405–410. [CrossRef] [PubMed]
123. Tosteson, M.T.; Alvarez, O.; Hubbell, W.; Bieganski, R.M.; Attenbach, C.; Caporales, L.H.; Levy, J.J.; Nutt, R.F.; Rosenblatt, M.; Tosteson, D.C. Primary structure of peptides and ion channels. Role of amino acid side chains in voltage gating of melittin channels. *Biophys. J.* **1990**, *58*, 1367–1375. [CrossRef]
124. Sui, S.F.; Wu, H.; Guo, Y.; Chen, K.S. Conformational changes of melittin upon insertion into phospholipid monolayer and vesicle. *J. Biochem.* **1994**, *116*, 482–487. [CrossRef]
125. Henzler Wildman, K.A.; Lee, D.K.; Ramamoorthy, A. Mechanism of lipid bilayer disruption by the human antimicrobial peptide, LL-37. *Biochemistry* **2003**, *42*, 6545–6558. [CrossRef]
126. Okumura, K.; Itoh, A.; Isogai, E.; Hirose, K.; Hosokawa, Y.; Abiko, Y.; Shibata, T.; Hirata, M.; Isogai, H. C-terminal domain of human CAP18 antimicrobial peptide induces apoptosis in oral squamous cell carcinoma SAS-H1 cells. *Cancer Lett.* **2004**, *212*, 185–194. [CrossRef]

127. Dhople, V.; Krukemeyer, A.; Ramamoorthy, A. The human beta-defensin-3, an antibacterial peptide with multiple biological functions. *Biochim. Biophys. Acta* **2006**, *1758*, 1499–1512. [CrossRef]
128. Selsted, M.E.; Harwig, S.S. Determination of the disulfide array in the human defensin HNP-2. A covalently cyclized peptide. *J. Biol. Chem.* **1989**, *264*, 4003–4007.
129. Lichtenstein, A.; Ganz, T.; Selsted, M.E.; Lehrer, R.I. In vitro tumor cell cytolysis mediated by peptide defensins of human and rabbit granulocytes. *Blood* **1986**, *68*, 1407–1410.
130. Lichtenstein, A.K.; Ganz, T.; Nguyen, T.M.; Selsted, M.E.; Lehrer, R.I. Mechanism of target cytolysis by peptide defensins. Target cell metabolic activities, possibly involving endocytosis, are crucial for expression of cytotoxicity. *J. Immunol.* **1988**, *140*, 2686–2694.
131. Lichtenstein, A. Mechanism of mammalian cell lysis mediated by peptide defensins. Evidence for an initial alteration of the plasma membrane. *J. Clin. Investig.* **1991**, *88*, 93–100. [CrossRef]
132. Mader, J.S.; Salsman, J.; Conrad, D.M.; Hoskin, D.W. Bovine lactoferricin selectively induces apoptosis in human leukemia and carcinoma cell lines. *Mol. Cancer Ther.* **2005**, *4*, 612–624. [CrossRef]
133. Furlong, S.J.; Ridgway, N.D.; Hoskin, D.W. Modulation of ceramide metabolism in T-leukemia cell lines potentiates apoptosis induced by the cationic antimicrobial peptide bovine lactoferricin. *Int. J. Oncol.* **2008**, *32*, 537–544. [CrossRef]
134. Eliassen, L.T.; Berge, G.; Leknessund, A.; Wikman, M.; Lindin, I.; Lokke, C.; Ponthan, F.; Johnsen, J.I.; Sveinbjornsson, B.; Kogner, P.; et al. The antimicrobial peptide, lactoferricin B, is cytotoxic to neuroblastoma cells in vitro and inhibits xenograft growth in vivo. *Int. J. Cancer* **2006**, *119*, 493–500. [CrossRef]
135. Nakamura, T.; Furunaka, H.; Miyata, T.; Tokunaga, F.; Muta, T.; Iwanaga, S.; Niwa, M.; Takao, T.; Shimonishi, Y. Tachyplesin, a class of antimicrobial peptide from the hemocytes of the horseshoe crab (*Tachypleus tridentatus*). Isolation and chemical structure. *J. Biol. Chem.* **1988**, *263*, 16709–16713.
136. Chen, J.; Xu, X.M.; Underhill, C.B.; Yang, S.; Wang, L.; Chen, Y.; Hong, S.; Creswell, K.; Zhang, L. Tachyplesin activates the classic complement pathway to kill tumor cells. *Cancer Res.* **2005**, *65*, 4614–4622. [CrossRef]
137. Adamia, S.; Maxwell, C.A.; Pilarski, L.M. Hyaluronan and hyaluronan synthases: Potential therapeutic targets in cancer. *Curr. Drug Targets Cardiovasc. Haematol. Disord.* **2005**, *5*, 3–14. [CrossRef]
138. Rao, A.G. Conformation and antimicrobial activity of linear derivatives of tachyplesin lacking disulfide bonds. *Arch. Biochem. Biophys.* **1999**, *361*, 127–134. [CrossRef]
139. Tada, N.; Horibe, T.; Haramoto, M.; Ohara, K.; Kohno, M.; Kawakami, K. A single replacement of histidine to arginine in EGFR-lytic hybrid peptide demonstrates the improved anticancer activity. *Biochem. Biophys. Res. Commun.* **2011**, *407*, 383–388. [CrossRef] [PubMed]
140. Barras, D.; Chevalier, N.; Zoete, V.; Dempsey, R.; Lapouge, K.; Olayioye, M.A.; Michielin, O.; Widmann, C. A WXW motif is required for the anticancer activity of the TAT-RasGAP317–326 peptide. *J. Biol. Chem.* **2014**, *289*, 23701–23711. [CrossRef] [PubMed]

Sample Availability: Not applicable.

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

EXCLI Journal 2018;17:688-708 – ISSN 1611-2156
Received: June 08, 2018, accepted: July 10, 2018, published: July 20, 2018

Review article:

TOWARDS UNDERSTANDING AROMATASE INHIBITORY ACTIVITY VIA QSAR MODELING

Watshara Shoombuatong, Nalini Schaduangrat, Chanin Nantasenamat*

Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand

* Corresponding author: E-mail: chanin.nan@mahidol.edu (C.N.); Phone: +66 2 441 4371; Fax: +66 2 441 4380

http://dx.doi.org/10.17179/excli2018-1417

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/).

ABSTRACT

Aromatase is a rate-limiting enzyme for estrogen biosynthesis that is overproduced in breast cancer tissue. To block the growth of breast tumors, aromatase inhibitors (AIs) are employed to bind and inhibit aromatase in order to lower the amount of estrogen produced in the body. Although a number of synthetic aromatase inhibitors have been released for clinical use in the treatment of hormone-receptor positive breast cancer, these inhibitors may lead to undesirable side effects (e.g. increased rash, diarrhea and vomiting; effects on the bone, brain and heart) and therefore, the search for novel AIs continues. Over the past decades, there has been an intense effort in employing medicinal chemistry and quantitative structure-activity relationship (QSAR) to shed light on the mechanistic basis of aromatase inhibition. To the best of our knowledge, this article constitutes the first comprehensive review of all QSAR studies of both steroidal and non-steroidal AIs that have been published in the field. Herein, we summarize the experimental setup of these studies as well as summarizing the key features that are pertinent for robust aromatase inhibition.

Keywords: aromatase, aromatase inhibitors, breast cancer, estrogen, QSAR, structure-activity relationship, data mining

INTRODUCTION

Breast cancer is one of the leading causes of death worldwide, with a greater prevalence in developed countries and a rapidly growing health concern in developing countries. It is also the most frequently occurring cancer found in women with an estimated 1.5 million new cases resulting in 570,000 deaths in 2015 (WHO, 2015). In addition, the prevalence of breast cancer in Asia is the highest among the world population (59 % of world population), out of which new cases account for 39 % with 44 % of cases resulting in deaths. In comparison, the prevalence of breast cancer in the continents of North America and Africa represent 5 % and 15 % of the world statistics, respectively (American Cancer Society, 2015).

Estrogen is the primary female sex hormone that acts as a double-edged sword where on one side it regulates important physiological functions for sustaining life (i.e. regulating the menstrual cycle, modulating bone density, maintenance of vessels and skin etc.) while on the other side, it is implicated in the development of breast cancer. Estrogen biosynthesis is catalyzed by aromatase, which converts androstenedione, a 19-carbon (C_{19}) steroid hormone, to estrone (E1) via a three-

EXCLI Journal 2018;17:688-708 – ISSN 1611-2156
Received: June 08, 2018, accepted: July 10, 2018, published: July 20, 2018

step A-ring aromatization. Aromatase also catalyzes the oxidation of testosterone, which is also then converted to estradiol (E2) (Ahmad and Shagufta, 2015) (Figure 1).

A common treatment for early-stage, hormone-sensitive breast cancer is surgery followed by radiotherapy. Furthermore, adjuvant endocrine therapy is given with or without chemotherapy depending on the tumor stage. In pre-menopausal women, most of the estrogen are made in the ovaries with the uptake of androstenedione from the circulation (Nelson and Bulun, 2001). Ovaries can convert androstenedione to estrone via the catalytic activity of aromatase, which is then transported to breast cells. However, in post-menopausal women, the main site of estrogen production are the breasts. As for the latter, the level of estrogens produced in the breast are comparable to that produced in the ovaries by pre-menopausal women, which is four to six times higher than those found in serum.

Approximately 60 % of pre-menopausal and 75 % of post-menopausal cancers are hormone-dependent (Russo et al., 2003), implying that endogenous estrogens are essentially required for proliferation. Many drugs that are used for the treatment of estrogen receptor-positive breast cancer are mechanistically based on the interference of either the estrogen production or the estrogen action.

Aromatase, also known as estrogen synthase or CYP19A1, is part of the cytochrome P450 family. It is consisting of 503 amino-acid residues spanning twelve α-helices and ten β-strands, inside which a heme cofactor is coordinated by a cysteine residue at position 437 (Ghosh et al., 2009). Aromatase is the major producer of estrogen in post-menopausal women and it catalyzes the rate-limiting step for converting androgens to estrogens (Simpson, 1994). As aromatase catalyzes the biosynthesis of estrogen from androgens, thus the inhibition of aromatase activity has be-

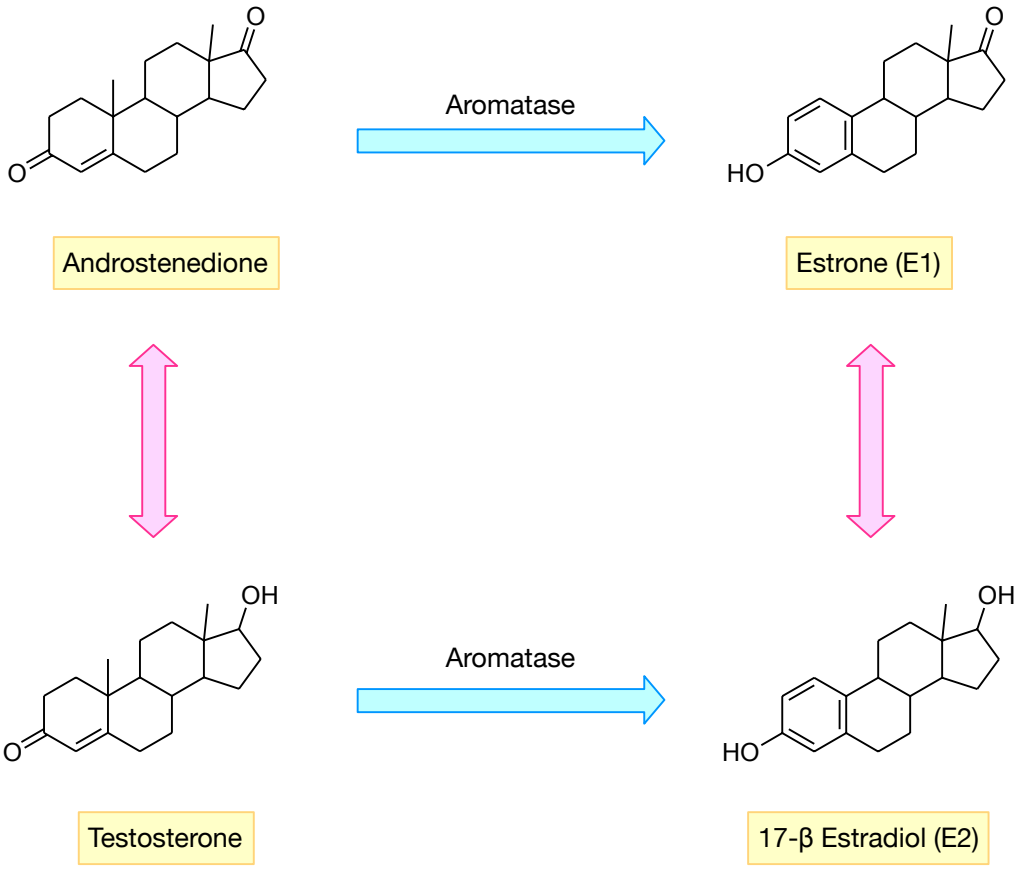

Figure 1: Summary of estrogen biosynthesis pathway as mediated by aromatase.

come the standard treatment for hormone-dependent breast cancers in women (Eisen et al., 2008). Aromatase is located in the plasma membrane of the endoplasmic reticulum of estrogen producing cells and plays a role in development, reproduction, sexual differentiation and behavior as well as in bone and lipid metabolism, brain functions and diseases such as breast and testicular tumors. Hence, in order to block the estrogen production, it is necessary to inhibit the aromatase enzyme that is responsible for its synthesis by using aromatase inhibitors (AIs). AIs constitute the front-line therapy for estrogen-dependent breast cancers. For this reason, inhibiting this terminal step in the estradiol biosynthesis pathway is considered to be a specific and therefore, a preferable strategy.

AROMATASE INHIBITORS

To date, there are three generations of FDA-approved AIs available for inhibiting the activity of aromatase. The first-generation of AIs includes aminoglutethimide, which is marketed in the late 1970s (Santen et al., 1978, 1982; Santen and Misbin, 1981; Graves and Salhanick, 1979) (Figure 2), a derivative of the sedative agent glutethimide that was initially introduced as an anticonvulsant. However, due to its adverse effects, such as high toxicity and lack of selectivity (Demers et al., 1987; Hughes and Burley, 1970), this AI was found to interfere with other CYP450 enzymes involved in cortisol and aldosterone biosynthesis (Santen et al., 1980). Thus, aminoglutethimide was withdrawn from the market. In addition, testolactone was the first-generation steroidal AI that was used to treat advanced-stage breast cancers, albeit with weak potency (Avendaño and Menéndez, 2008). Nevertheless, these first-generation AIs served as the prototype for future generations with an emphasis on developing more potent drugs with higher selectivity and reduced toxicity. Continuing on to the second-generation, fadrozole, which contains an imidazole group (Bonnefoi et al., 1996), is more selective and potent than aminoglutethimide. Nevertheless, it still displayed effects on aldosterone, progesterone and corticosterone biosynthesis. Formestane (Brueggemeier et al., 2005), a steroidal analogue, was the first selective AI to reach clinical trials in the 1990s. It was demonstrated to be effective and was well tolerated (Dowsett and Lloyd, 1990). However, formestane exhibited poor oral bioactivity and had to be administered bi-weekly and thus, lost popularity with the discovery of the newer, more effective third-generation AIs (DrugBank, 2013).

Finally, the third-generation of AIs, are comprised of triazole derivatives, anastrozole and letrozole and one steroidal analogue, exemestane (Dutta and Pant, 2008). These AIs displayed improved efficacy and lower toxicity as compared with the estrogen antagonist, tamoxifen, in both early and advanced breast cancer (Thürlimann et al., 2004). For this reason, the last generation of AIs has been recommended by the FDA as first-line drugs for the therapy of breast carcinoma. Anastrozole and letrozole, are non-steroidal derivatives and competitive inhibitors of androstenedione. Both contain a triazole group that can interact with the prosthetic heme group of aromatase. Exemestane is a steroidal analog of androstenedione thus, permanently binding to the enzyme and deactivating its catalytic activity (Coombes et al., 2007).

Initial attempts to clarify the interaction mechanism of aromatase and its inhibitors have relied on the use of homology-derived models (Loge et al., 2005). Such studies have focused on clinically used AIs such as fadrozole, letrozole and exemestane as well as other natural products such as ligands, flavonoids and coumestrol (Karkola and Wähälä, 2009; Paoletta et al., 2008; Awasthi et al., 2015; Worachartcheewan et al., 2014b; Nantasenamat et al., 2014).

Since the crystal structure of human placental aromatase has been solved by Ghosh et al. (2009), the availability of structural details on the active site of aromatase helps in understanding of the binding characteristics of AIs as well as the evaluation of key reactions needed in the mechanism of aromatase.

EXCLI Journal 2018;17:688-708 – ISSN 1611-2156
Received: June 08, 2018, accepted: July 10, 2018, published: July 20, 2018

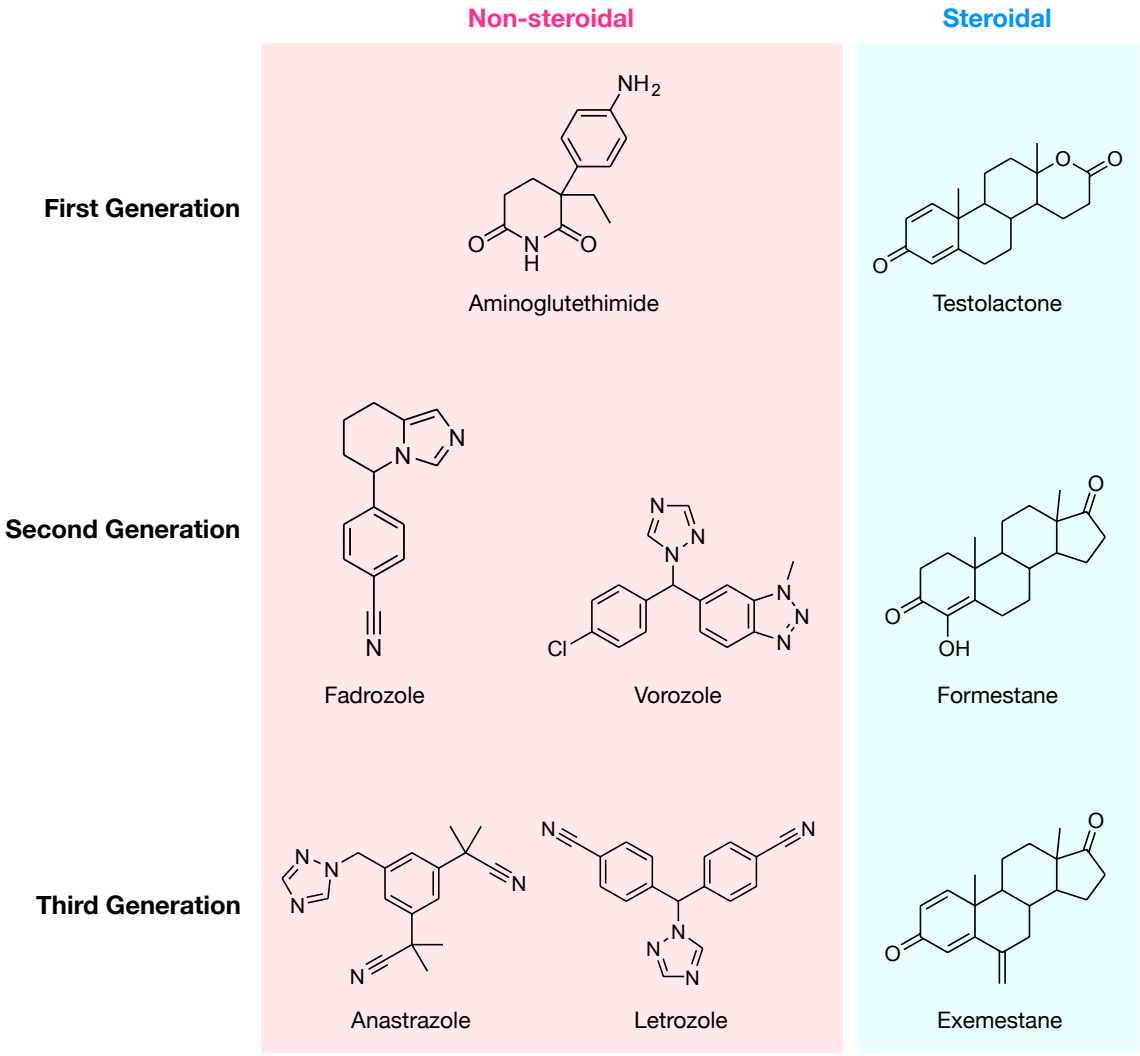

Figure 2: Chemical structures of the three generations of FDA-approved aromatase inhibitors.

This has opened up a plethora of opportunities by enabling the understanding of the molecular basis for the specificity of the aromatase enzyme and its unique catalytic mechanisms, which is imperative for the development of the next-generation of AIs.

CONCEPTS OF QSAR MODELING

Quantitative structure-activity relationship (QSAR) (Nantasenamat et al. 2009, 2010) is a ligand-based approach that seeks to discern the mathematical relationship between chemical structures (i.e. as described by various types of molecular descriptors) and the investigated biological activity through the use of statistical and machine learning techniques.

Historically, the work of Cros (1863), Crum Brown and Fraser (1868) and especially that of Muir et al. (1949) laid the foundations for the subsequent birth of QSAR as formally introduced by Hansch and Fujita (1964) in their landmark work investigating substituent effects of various compounds against various biological activities (i.e. benzoic acids against mosquito larvae, phenols against gram-positive and gram-negative bacteria, phenyl ethyl phosphate insecticides against houseflies, thyroxine derivatives against rodents, diethylaminoethyl benzoates against guinea pigs and carcinogenic compounds against mice) by utilizing substituent constants as descriptors. Ever since, QSAR had been an integral part of computational drug discovery efforts as it had

been utilized to probe the underlying mechanistic basis of various biological activities (Nantasenamat and Prachayasittikul, 2015). Recently, Fujita and Winkler (2016) had shared their perspectives on the two QSAR worlds consisting of (i) classical QSAR and (ii) modern QSAR.

The early years of *classical QSAR* entails investigation on the structure-activity relationship of a congeneric set of compounds (i.e. compounds sharing a common chemical scaffold or chemotype) through the use of a few molecular descriptors. Classical QSAR methodology (Hansch et al., 1963) assumes that the biological activity of investigated chemicals can be explained by simple and interpretable physicochemical properties. These physicochemical properties encode structural features that are considered to be statistically important and that can provide useful insights and understanding pertaining to the interaction being studied. Typically, classical QSAR models are built using partial least-squares (PLS) and multiple linear regression (MLR). It should be noted that this approach does not take into consideration the 3D structure of the receptor-ligand interaction. Thus, this had inspired the development of a 3D-QSAR technique by Cramer et al. (1988) that essentially involves the alignment of a congeneric set of compounds (i.e. compounds sharing a common scaffold or chemotype) and followed by the computation of molecular fields (steric and electrostatic). Furthermore, modifications to the CoMFA concept known as comparative molecular similarity indices analysis (CoMSIA) was proposed by Klebe et al. (1994) to extend CoMFA via the utilization of Gaussian potentials as the basis for calculating similarity and thus, expand its applicability (Kubinyi, 1997).

Over the years, advancements in computation has given rise to *modern QSAR* in which an extensive list of molecular descriptors as well as a wide range of machine learning algorithms can be applied in studying the structure-activity relationship of large sets of heterogeneous and chemically diverse set of compounds. On one end, modern QSAR makes it possible to harness the big data of bioactivity information accumulated over the years for model development while on the other end, the resulting models are often complex and not readily comprehensible to bench scientists. The need for simple and interpretable QSAR models along with best practices has been discussed in a recent book chapter (Shoombuatong et al., 2017). Briefly, desirable characteristics of robust QSAR models have been set forth by the Organisation for Economic Co-operation and Development (OECD) as to encourage the utilization of QSAR models for regulatory purposes. These main OECD principles for the development of robust QSAR models are summarized in Table 1.

The typical workflow for the development of QSAR models is depicted in Figure 3. First, the QSAR modeling process starts by the compilation of a data set that entails collecting information pertaining to the compound name along with their SMILES notation, bioactivity values (e.g. IC_{50}, EC_{50}, K_i, % activity, etc.) as well as reference to the original paper. Second, the data set is subjected to data pre-processing as to ensure the completeness of the data and that there are no missing information or misspellings. Third, chemical structures are drawn and subjected to structure standardization as to remove salts, ensure appropriate charge of functional moieties, select appropriate tautomeric structures, etc. Fourth, molecular descriptors are computed as to provide quantitative description of chemical structures and this is followed by feature selection as to remove useless and/or collinear variables. Fifth, the curated data set is employed for model construction via the use of machine learning algorithms and this entails data splitting, data balancing, data validation, model validation and performance assessment. Finally, the resulting model is subjected to scrutiny on the feature importance as to identify key features contributing to the origin of the biological activity. Summary and guidelines pertaining to the best practices for QSAR model development has been described previously (Tropsha, 2010).

EXCLI Journal 2018;17:688-708 – ISSN 1611-2156
Received: June 08, 2018, accepted: July 10, 2018, published: July 20, 2018

Table 1: The OECD principle guidelines for developing and validating QSAR model.

#	OECD principle	Description
1	Defined endpoint	To ensure that all endpoint values, within a given dataset, are consistent
2	Unambiguous algorithm	To ensure the ability of transparency and reproducibility in the proposed QSAR model
3	Defined domain of applicability	To define how robust, significant and validated QSAR model could be
4	Appropriate measures of goodness-of-fit, robustness and predictivity	To simplify the overall criteria of model validation: the internal performance of a model and the predictivity or predictive power of model
5	Mechanistic interpretation	To ensure that there are assessments of the possibility of a mechanistic interpretation

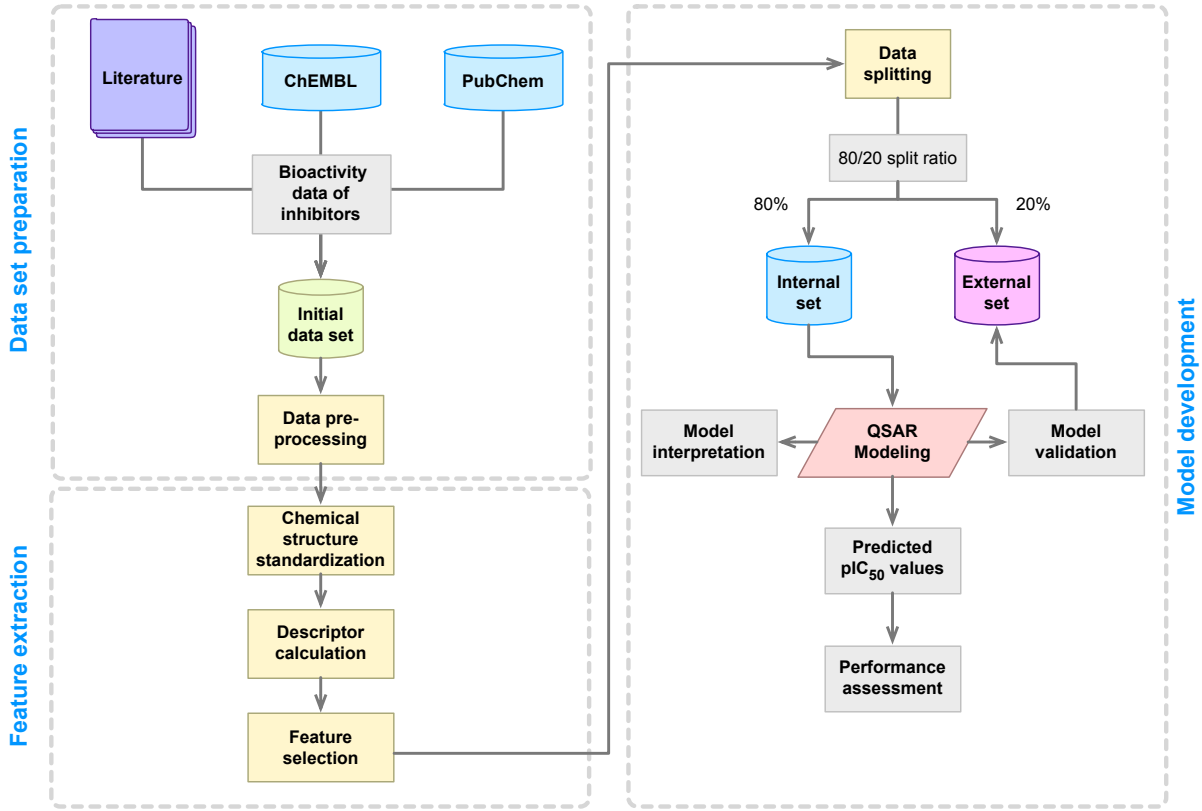

Figure 3: General workflow of QSAR model development.

Since then, the breadth of available molecular descriptors have expanded to encompass a wide range of descriptors spanning one to several dimensions. Such descriptors may account for the general features of a molecule or may consider the fine details of a molecule down to its atomic constitution. A summary of common molecular descriptors (along with its description) used in QSAR models of AIs is provided in Table 2.

Table 2: Summary of common classes of molecular descriptors.

Descriptor class	Description	References
Molecular field	Steric and electrostatic properties of a molecule as derived from the superimposition of molecules in CoMFA analysis	Cramer et al., 1988
Molecular similarity indices	Descriptors as used in CoMSIA that are computed from Steric and Electrostatic ALignment (SEAL) similarity fields as to generate steric, electrostatic, hydrophobic, and hydrogen bonding descriptors	Kubinyi et al., 1998
Molecular surface	As implemented in CoMSA, descriptors are derived from Coulomb electrostatic potential on the molecular surface	Polanski and Gieleciak, 2003
Multivariate image analysis	Pixels derived from 2D image of chemical structures	Barigye et al., 2018
Physicochemical	Pertains to various 1D-3D chemical and physical properties of a molecule	Todeschini and Consonni, 2000
Pharmacophore mapping	A 4D-QSAR approach coupled to self-organizing map that entails the incorporation of conformational freedom into 3D-QSAR model	Bak and Polanski, 2007
Quantum chemical	Electronic properties of a molecule as derived from low-energy conformer as computed by quantum mechanical calculation	Karelson et al., 1996
SMILES	Atomic and bond constituents of a molecule	Worachartcheewan et al., 2014a
Spectral	Based on ^{13}C NMR spectroscopic data of a compound that essentially pertains to electrostatic and electronic properties as derived from frequencies of quantum mechanical properties of a nuclear magnetic moment	Beger et al., 2004

Machine learning

Machine learning is an implementation of artificial intelligence in which computers can automatically learn from data sets by extracting important patterns and making decisions or predictions. A summary of common machine learning algorithms that are used for QSAR modeling along with their strengths and weaknesses are provided in Table 3.

The concepts and in-depth treatment of machine learning is beyond the scope of this review and readers are directed to a previous comprehensive treatment of the topic (Shoombuatong et al., 2017; van Westen et al., 2011). Herein, we cover common machine learning algorithm that have been used in the study of AI activity.

The simplest learning algorithm is multiple linear regression (MLR) (Aiken et al., 2003), which is an extension of the simple linear regression and is used to explain the relationship between a series of features, X=(x_1, x_2, x_2,..., x_N), and output values, Y=(y_1, y_2, y_2,..., y_N), as follows:

EXCLI Journal 2018;17:688-708 – ISSN 1611-2156
Received: June 08, 2018, accepted: July 10, 2018, published: July 20, 2018

$$y_i = \sum_{i=1}^{N} \beta_i x_{ij} + \beta_0 \quad (1)$$

where y_i is the output value, x_{ij} is represented a data for i^{th} compound and j^{th} descriptor of i^{th} compound and β_i is the coefficient parameter.

Partial least squares regression (PLS). This method is a well-known method for constructing predictive models when features or descriptors have inter-correlated latent variables (Helland, 1988; Helland, 2001). PLS is closely related to principal component analysis (PCA) that consists of matrix decomposition into a matrix of eigenvectors and a matrix of its loadings factors. Given a dataset $X^{N\times M}$ with N rows and M columns, the general approach can be written as follows:

$$X^{N\times M} = X^{N\times A} P^{A\times M} + X^{N\times M} \quad (2)$$

This is equivalent to a reduction of an *M*-dimensional variable space to an *A*-dimensional space. The variables in dimension *A* are also called latent variables.

The matrix T contains orthogonal column vectors, also called score vectors, that represents the latent variables.

Artificial neural network (ANN). This method is a computation-based method inspired by networks of biological neurons in the human brain (Puri et al., 2016). Basically, there are 3 different layers in the architecture of ANN: input layer (the input (*X*) is fed into the model through this layer), hidden layers (in general, there can be more than one hidden layers which utilizes some method to operate *X* and deliver to an output layer) and output layer (the data after processing is made available at the output layer).

Support vector machine (SVM). This statistical learning approach is based on the principles of structure-risk minimization and kernel method as proposed by Cortes and Vapnik (1995), which are used to construct a maximum-margin-separating hyperplane. The main advantage of SVM model is to seek the best compromise between the computational cost and the prediction error as to obtain the optimum generalization ability. SVM can be categorized as support vector classification (SVC) and support vector regression (SVR) (Cortes and Vapnik, 1995; Smola and Schölkopf, 2004). The principle idea of this method is to transform an input space with *m*-dimen-

Table 3: Summary of the strength and weakness of the machine-learning algorithms for performing QSAR modeling discussed in this review.

Factor	MLR	PLS	ANN	SVM	DT	RF
Non-linear			✓	✓	✓	✓
Classification and regression	✓	✓	✓	✓	✓	✓
Prediction error	High	Medium	Low	Low	Medium	Low
Computational cost	Low	Medium	Medium	Medium	Medium	High
Memory requirements	Low	Medium	Low	Low	Medium	Medium
Overfitting	✓		✓		✓	
Dimension reduction	✓	✓			✓	✓
Easy to interpret	✓	✓			✓	✓

MLR: multiple linear regression, PLS: partial least squares regression, ANN: artificial neural network, SVM: support vector machine, DT: decision tree, RF: random forest

sional vector into a feature space with n-dimensional vector where $m < n$, and select a separating hyperplane giving the largest distance between the two classes.

Decision tree (DT). This machine learning technique is used for finding and describing a dataset (X,Y) with tree representation or structure (Safavian and Landgrebe, 1991). The tree is composed of a root node, a set of internal nodes, and a set of terminal nodes (leaves). This method is one of well-known built-in feature selector. The main purpose of using DT is to achieve a more concise and transparency of the model to identify the relationship between X and Y variables.

Random forest (RF). This ensemble learning method essentially integrates many classification and regression trees (CART). Breiman (2001) developed the RF method by growing many weak decision trees for enhancing the prediction performance of CART. The last decade has witnessed the significant achievement of RF model in applications of drug developments and related works (Win et al., 2017; Worachartcheewan et al., 2015; Pratiwi et al., 2017; Simeon et al., 2016a; Phanus-umporn et al., 2018; Suvannang et al., 2018). RF model takes advantage of two efficient machine learning techniques: bagging and random feature selection.

QSAR MODELS OF AROMATASE INHIBITORY ACTIVITY

The utilization of QSAR in aromatase research has only scratched the surface of the possible benefits that can be attained. Several classes of aromatase inhibitors have been created with only a few notable classes that have made it to the pre-clinical and clinical testing.

Thus, it is worthwhile to elucidate the physicochemical profiles of effective aromatase inhibitors in comparison with ineffective ones as such knowledge can aid in the optimization of existing compound classes or development of novel classes from available scaffolds and functional group fragments. Particularly, questions such as "What molecular descriptors are crucial for highly potent compound? How big should a potent aromatase inhibitor be? Which functional groups are most commonly found in potent compounds?" could be answered through QSAR efforts.

In 1997, Lipinski published a landmark paper on the Rule of 5 (Lipinski et al., 2001), which has been widely used in the pharmaceutical industries as general guidelines for drug development efforts. The Rule of 5 considers ADMET issues that are critical towards the success of the identified compounds of interest as it may help reduce pre-clinical and clinical failures. A similar approach may be applied to the aromatase system where several Rules may be developed for the identification of potent aromatase inhibitors.

The earliest QSAR study performed for AIs was published in 1994 by Nagy et al. (1994) whereby MLR analysis was conducted on models built with 5 quantum chemical descriptors for 24 compounds assessed by LOO-CV procedure. From the results obtained, the authors were able to discover 2 candidate AIs for further pharmacophore studies.

Furthermore, as can be seen from Figure 4 (top-left), in the years from 1994-2000 only five additional articles (Oprea and García, 1996; Recanatini, 1996; Sulea et al., 1997; Recanatini and Cavalli, 1998; Cavalli et al., 2000) on QSAR of AIs were published, which utilized mainly molecular field descriptors and LOO-CV. For example, Cavalli et al. (2000) quantitatively compared 3D-QSAR models of the cytochrome P450 active site via CoMFA modeling and homology modeling techniques. Once models were built, two non-steroidal AIs were docked into each model and the resulting interaction energies were recorded. The authors noted that although each technique had its drawbacks, both could be used together as a mutual validation technique for ligand-based and target-based 3D models of ligand-target interactions.

In addition, Sulea et al. (1997) described van der Waals envelopes as a steric potential field in a 3D-QSAR CoMFA based modeling of ligand-receptor interactions that was performed on 78 steroidal AIs and evaluated

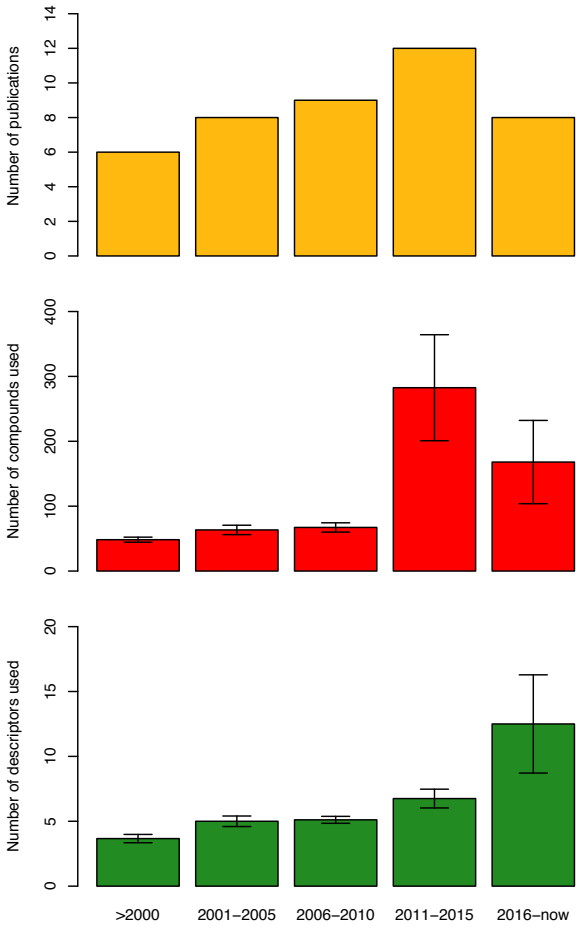

Figure 4: Overview on the number of publications (top), number of compounds (middle) and the number of descriptors (bottom) extracted from articles describing QSAR models of AIs.

LOO-CV procedure. The authors were able to prove that the van der Waals envelopes intersection volumes (INVOL) could be used as an alternative replacement for the more commonly used Lennard-Jones 6-12 potential for the identification of relevant features governing biological activities within CoMFA and 3D-QSAR based models.

Similarly, Oprea and Garcia (1996) analyzed the data of 50 steroidal AIs using CoMFA models coupled with chemometric based Generating Optimal Linear PLS Estimation (GOLPE) models validated using both the LOO-CV and the external test procedures. The authors concluded that using CoMFA, differences in aromatase inhibition among the C6-substituted steroids were shown to be consistent with known, potent inhibitors of aromatase, included in the model. In addition, when direct alignment comparisons were made, these compounds exhibited distinct features that overlapped with the steric and electrostatic fields obtained in the CoMFA model.

Over the course of the next five years (2001-2005) (Gironés and Carbó-Dorca, 2002; Beger and Wilkes, 2002; Beger et al., 2001; Polanski and Gieleciak, 2003; Beger et al., 2004; Leonetti et al., 2004; Cavalli et al., 2005), it can be seen that studies employed higher number of descriptors as well as made use of more descriptor types (e.g. molecular fields, spectral, molecular surface and quantum chemical) were observed in seven publications where the only CV method applied on the datasets was the LOO-CV (Figure 4). Beger et al. (2004) developed a technique which was similar to QSAR modeling, which they called the minimum deviation of structurally assigned spectra analysis (MiDSASA). This method was based on minimum chemical shift differences on substructure fragments instead of relying on substructure fragments as a whole for model production as is typical in SAR modeling. The authors used this MiDSASA template on 50 steroids binding the aromatase enzyme based on the average activity of the four nearest compounds, resulting in a correlation of 0.71. The authors further suggested that models made using the minimum deviation concept can be applied to other chemoinformatic data analyses such as metabolite concentrations in metabolic pathways for metabolomics research.

In addition, Beger et al. (2001) built quantitative spectroscopic data-activity relationship (QSDAR) models for 50 steroidal AIs developed based on data collected via simulated ^{13}C nuclear magnetic resonance (NMR). The models were based on comparative spectral analysis (CoSA) and comparative structurally assigned spectral analysis (CoSASA). From the PLS analysis, the CoSA models exhibited R^2 of 0.78 and Q^2 of 0.71 while the CoSASA based models provided R^2 of 0.75 and Q^2 of 0.66.

Similarly, Polanski and Gieleciak (2003) used CoMSA to analyse the 3D-QSAR models built for 50 steroidal AIs. The authors aimed to predict regions that are important for the binding activity of the ligand with the enzyme. Using uniformative variable elimination as coupled to partial least squares (UVE-PLS) or modified iterative UVE procedure (IVE-PLS), the authors were able to determine that the 3D-QSAR models generated ($Q^2 = 0.96$) outperformed those reported at the time using CoMFA, CoSA or CoSASA.

Furthermore, the number of articles on QSAR of AIs were seen to increase rapidly for the years 2006-2010 (Figure 4) with the publication of ten articles in the time period (Bak and Polanski, 2007; Nagar et al., 2008; Castellano et al., 2008; Mittal et al., 2009; Gueto et al., 2009; Nagar and Saha, 2010a, b; Roy and Roy, 2010a, b; Dai et al., 2010). Most of the QSAR models in this time frame were built utilizing physicochemical descriptors as compared to other techniques in the previous years.

Additionally, the validation methods for AIs QSAR publications in the abovementioned years were tied between LOO-CV only and LOO-CV in conjunction with external validations (Figure 5). For example, Bak and Polanski (2007) conducted a 4D-QSAR study based on the self-organizing map (SOM), which is an unsupervised method based on the Kohonen neural network coupled with the IVE-PLS analysis. The use of this combined 4D-QSAR and IVE-PLS method provided a very stable and predictive modeling technique. The method enabled the authors to identify molecular motifs contributing to the aromatase enzyme binding activity. Gueto et al. (2009) employed structure-based drug design approach using receptor-independent CoMFA maps that were generated from LeapFrog calculations.

A robust model as verified by the bootstrapping method produced statistically significant results via cross-validated analysis, which consisted of 45 and 10 molecules in the training and test sets, respectively. Using this model, the authors were able to predict the activity of novel AI molecules which had more potency than previously reported compounds. Roy and Roy (2010a) performed a 3D-QSAR study on a diverse set of compounds using the crystal structure of aromatase whereby the dataset was divided into training (n=87) and testing (n=29) set by clustering techniques. All the QSAR models were subjected to multiple validation methods such as internal validation, external validation and Y-randomization. The authors concluded that in order to exhibit ideal aromatase inhibitory activity, the compound should contain at least one or two hydrogen bond acceptor groups (such as NO_2 and CN) with optimal hydrophobicity.

Additionally, the increase in popularity of QSAR models for predicting AIs was greatly observed in 2011-2015 (Narayana et al., 2012; Kishore et al., 2013; Nantasenamat et al., 2013a, b, 2014; Dai et al., 2014; Xie et al., 2014; Worachartcheewan et al., 2014a, b; Shoombuatong et al., 2015; Awasthi et al., 2015; Xie et al., 2015; Kumar et al., 2016) whereby the number of publications increased to thirteen, with an even more dramatic rise in the number of compounds used for calculating descriptors using LOO-CV and external validation (Figure 5). Worachartcheewan et al. (2014b) investigated the QSAR of coumarins as potential AIs using 7 quantum chemical descriptors. MLR was used for the analysis of models, which were shown to achieve good predictive performance as verified by LOO-CV affording Q^2 of 0.9239 and $RMSE_{CV}$ of 0.1304 while an external validation confirmed its robustness with Q^2_{Ext} of 0.7268 and $RMSE_{Ext}$ of 0.2927.

Moreover, Nantasenamat et al. (2013b) explored a set of 54 letrozole analogs as AIs in a QSAR study using MLR, ANN and SVM methods. The QSAR model was developed using a set of descriptors giving rise to important physicochemical properties (i.e. number of rings, ALogP and HOMO-LUMO) which were further used for predicting AI activity. The authors observed a strong correlation among the predicted pIC_{50} values with

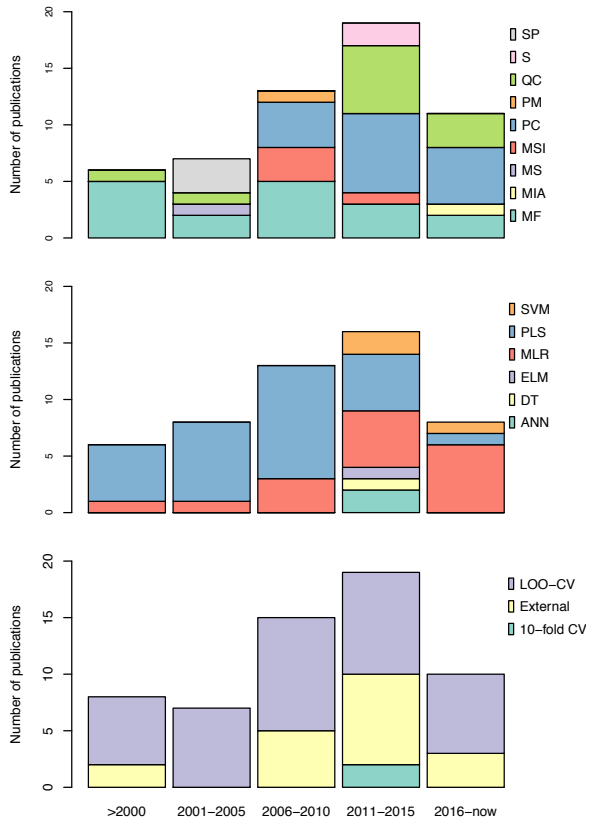

Figure 5: Overview of the types of descriptors (top), machine learning algorithms (middle) and validation methods (bottom) extracted from articles describing QSAR models of AIs.

(Abbreviations: SP, S, QC, PM, PC, MSI, MS, MIA and MF represents spectral, SMILES, quantum chemical, pharmacophore mapping, physicochemical, molecular similarity indice, molecular surface, multivariate image analysis and molecular field, respectively. SVM, PLS, MLR, ELM, DT and ANN represents support vector machine, partial least square, multiple linear regression, efficient linear model, decision tree and artificial neural network, respectively. LOO-CV, external and 10-fold CV represents leave-one-out cross-validation, external test and 10-fold cross-validation, respectively)

their experimental values, displaying correlation coefficient Q^2 values in the range of 0.72–0.83 while the external test set (Q^2_{Ext}) afforded values in the range of 0.65–0.66. Furthermore, Worachartcheewan et al. (2014a) employed the bioactivity data on pIC_{50} of 973 AIs for constructing QSAR models using CORelation And Logic (CORAL) software (http://www.insilico.eu/coral) where the molecular structures are represented by their simplified molecular input line entry system (SMILES) notation and thus eliminating the need to geometrically optimize molecular structures. The Monte Carlo optimization of correlation was used for predicting the aromatase inhibitory activity. Results obtained from rigorous dataset splits and CV techniques indicated that models were reliable with R^2 and Q^2 in ranges of 0.6271–0.7083 and 0.6218–0.7024, respectively. Similarly, Nantasenamat et al. (2014) conducted the first large-scale QSAR study on a non-redundant set of 63 flavonoids using MLR, ANN, SVM and DT methods. The models obtained showed good predictive performance with Q values in the range of 0.8014–0.9870 and 0.8966–0.9943 evaluated by LOO-CV and external test, respectively. Furthermore, in another study conducted by our group Shoombuatong et al. (2015), proposed the simple and interpretable efficient linear method (ELM) for constructing a highly predictive QSAR model. The results indicated that a robust performance was achieved using the ELM method with 10-fold CV MCC values of 0.64 and 0.56 for steroidal and non-steroidal AIs, respectively. In addition, Xie et al. (2014) constructed 3D QSAR models in order to elucidate the steroidal AIs with lower side effects using CoMFA and CoMSIA methods. The models produced were reliable and predictive good statistical results for CoMFA: Q^2 = 0.636, R^2 = 0.988, Q^2_{Ext} = 0.658 and CoMSIA: Q^2 = 0.843, R^2 = 0.989, Q^2_{Ext} = 0.601.

The current trend (2016–2018; Figure 4) shows that eight articles (Song et al., 2016; Ghodsi and Hemmateenejad, 2016; Adhikari et al., 2017a; Prachayasittikul et al., 2017; Pingaew et al., 2018; Lee and Barron, 2018; Barigye et al., 2018) have already been published in comparison to a total of 13 publications for the previous 5 years. Thus, it is promising that the number of publication regarding AIs utilizing QSAR models for prediction will continue to grow.

To further aid in that growth process, the number of compounds used as the data set has seen a steady rise with the number of descriptors for generating QSAR models saw a dramatic increase as compared to previous years. As for the types of descriptors, the trend has moved towards modern QSAR with the utilization of physicochemical properties and quantum chemical structures to build the models. In addition, the main validation techniques remain the same as previous years whereby LOO-CV and external validation were mainly used. Ghodsi and Hemmateenejad (2016) conducted QSAR studies on a series of diarylalkylimidazole and diarylalkyltriazole derivatives previously evaluated as being potent AIs using 870 quantum chemical descriptors (such as dipole moment and energies of HOMO and LUMO orbitals, hydration energies, and lipophilicity) that were analyzed using MLR. The models were validated with the LOO-CV and the authors concluded that lipophilicity was an important factor for the strong binding to aromatase. In addition, the HOMO orbital shape and its imidazole ring distribution was also considered as important. More recently, Adhikari et al. (2017a) performed QSAR studies using various techniques (2D-QSAR, 3D-QSAR and HQSAR) on 67 non-steroidal letrozole-based analogs with promising AI activity. Stepwise multiple linear regression (SMLR) was used to build the models after which, the models were validated with the LOO for internal validation. The results from the 2D-QSAR study suggested the importance of the nitrogen atoms in their electrotopological state thereby inferring that their orientation may modulate the inhibition. The authors noted that these results were further validated with the 3D-QSAR analysis while the HQSAR model inferred the importance of the *p*-cyanophenyl moiety in regulating AI. Additionally, Lee and Barron (2018) conducted 3D-QSAR studies on the bioactivity (IC_{50}) of 124 compounds exhibiting AI activity (steroidal and heterocyclic). Multiple linear regression combined with genetic algorithm (GA-MLR) was used to build the models which was then validated via the LOO and external validation methods. Furthermore, Prachayasittikul et al. (2017) investigated the aromatase inhibitory potency of a series of 2-amino (chloro)-3-chloro-1,4-naphthoquinone derivatives by constructing QSAR models using the IC_{50} values. The models were evaluated based on MLR and LOO-CV which indicated good predictive performance (Q^2 = 0.9783 and $RMSE_{CV}$ = 0.0748) of the constructed model. Therefore, 1,4-naphthoquinone derivatives can be seen as promising compounds needed further evaluations as AIs. The most recent article published by Barigye et al. (2018) reported the first practical application of Discrete Fourier Transformation (DFT) based Multiple Image Analysis (MAI) derived 2D-QSAR model for the classification of an aforementioned set of 973 novel AIs as compiled from the literature (Nantasenamat et al., 2013a).

INSIGHTS FROM QSAR MODELS

Model interpretation is the process by which the underlying features contributing the most to the investigated biological activity are deduced as to help guide the design of novel and robust drugs. The interpretability of a QSAR model is contingent upon the types of descriptors and machine learning algorithms used. As summarized in Table 4, it can be observed that prior to 2010, MLR and PLS models, also known as white-box approaches, were the most popular and yet simple learning algorithms used for QSAR modeling of AIs.

Although these two models are interpretable but they did not perform well on highly complexed data. On the other hand, a black-box approach like ANN and SVM can provide higher accuracy in the same case but they cannot provide details pertaining to how the factors exert its influence on the biological activity of investigated compounds. Analysis of key features for aromatase inhibition from

Table 4: Summary of machine learning algorithm used in QSAR modeling for predicting and analyzing aromatase inhibitor.

Year	Number of compounds	Type of descriptors[a]	Number of descriptors	ML algorithm[b]	Validation[c]	References
1994	24	QC	5	MLR	LOO-CV	Nagy et al., 1994
1996	29	MF	3	PLS	LOO-CV	Recanatini, 1996
1996	50	MF	2	PLS	LOO-CV, External	Oprea and García, 1996
1997	78	MF	4	PLS	LOO-CV	Sulea et al., 1997
1998	60	MF	6	PLS	LOO-CV, External	Recanatini and Cavalli, 1998
2000	49	MF	2	PLS	LOO-CV	Cavalli et al., 2000
2001	50	SP	5	PLS	LOO-CV	Beger et al., 2001
2002	50	QC	6	PLS	LOO-CV	Gironés and Carbó-Dorca, 2002
2002	50	SP	9	PLS	LOO-CV	Beger and Wilkes, 2002
2003	50	MS	5	PLS	LOO-CV	Polanski and Gieleciak, 2003
2004	35	MF	3	PLS	LOO-CV	Leonetti et al., 2004
2004	50	SP	5	MLR,PLS	LOO-CV	Beger et al., 2004
2005	70	MF	5	PLS	LOO-CV	Cavalli et al., 2005
2007	152	PM	2	PLS	LOO-CV	Bak and Polanski, 2007
2008	128	MF	5	PLS	LOO-CV, External	Castellano et al., 2008
2008	33	MF, MSI	4	MLR, PLS	LOO-CV	Nagar et al., 2008
2009	30	MF, MSI	3	PLS	LOO-CV	Mittal et al., 2009
2009	66	MF	7	PLS	LOO-CV, External	Gueto et al., 2009
2010	32	PC	7	PLS	LOO-CV	Dai et al., 2010
2010	59	PC	5	PLS	LOO-CV	Roy and Roy, 2010b
2010	116	PC	4	PLS	LOO-CV, External	Roy and Roy, 2010a
2010	52	MF, MSI	5	MLR, PLS	LOO-CV, External	Nagar and Saha, 2010a
2010	89	PC	6	MLR, PLS	LOO-CV, External	Nagar and Saha, 2010b
2012	39	PC	3	MLR	LOO-CV, External	Narayana et al., 2012
2013	54	PC, QC	3	MLR, ANN, SVM	LOO-CV, External	Nantasenamat et al., 2013b
2013	973	PC, QC	13	DT	10-fold CV	Nantasenamat et al., 2013a
2013	73	QC	5	PLS	External	Kishore et al., 2013
2014	34	PC, QC	7	MLR	LOO-CV, External	Worachartcheewan et al., 2014b
2014	973	S	7	MLR	LOO-CV, External	Worachartcheewan et al., 2014a
2014	63	PC, QC	6	MLR, ANN, SVM	LOO-CV	Nantasenamat et al., 2014
2014	14	PC	5	PLS	LOO-CV	Dai et al., 2014
2015	45	MF	6	PLS	LOO-CV, External	Awasthi et al., 2015
2015	84	MF, MSI	7	PLS	LOO-CV, External	Xie et al., 2015

Year	Number of compounds	Type of descriptors[a]	Number of descriptors	ML algorithm[b]	Validation[c]	References
2015	973	PC, QC	15	ELM	10-fold CV	Shoombuatong et al., 2015
2015	66	MF, S	4	PLS	LOO-CV, External	Xie et al., 2014
2016	46	MF	5	PLS	LOO-CV	Kumar et al., 2016
2016	76	PC, QC	9	MLR	LOO-CV	Ghodsi and Hemmateenejad, 2016
2016	13	PC	4	MLR	LOO-CV	Song et al., 2016
2017	11	PC, QC	4	MLR	LOO-CV	Prachayasittikul et al., 2017
2017	67	MF	5	MLR	LOO-CV, External	Adhikari et al., 2017a
2018	124	PC, QC	9	MLR	LOO-CV, External	Lee and Barron, 2018
2018	34	PC	4	MLR	LOO-CV	Pingaew et al., 2018
2018	973	MIA	60	SVM	External	Barigye et al., 2018

[a]MF: molecular field, MIA: Multivariate image analysis, MS: Molecular surface, MSI: molecular similarity indices, PC: physicochemical, PM: Pharmacophore mapping, QC: quantum chemical, S: SMILES, SP: Spectral
[b]ANN: artificial neural network, MLR: multiple linear regression, PLS: partial least squares regression, SVM: support vector machine
[c]LOO-CV: leave-one-out cross-validation, 10-fold CV: 10-fold cross-validation

selected QSAR works employing descriptors pertaining to quantum chemical and physicochemical properties are performed hereafter (Table 5). Nantasenamat et al. (2013a) performed a large-scale QSAR modeling of a set of steroidal and non-steroidal AIs and revealed that the most important features from PCA analysis were found to be nHAcc, TPSA and LUMO for non-steroidal and Q_m, TPSA and nHAcc and ALogP for steroidal AIs. In addition, fragment analysis provided complementary insights by suggesting that the presence of the azole ring in non-steroidal inhibitors (i.e. known to coordinate with the heme iron) and the presence of carbonyl group in the C3 position of steroidal inhibitors were important for aromatase inhibition.

In addition, using the same set of data, Shoombuatong et al. (2015) used the ELM model to deduce the most important features associated with AI. It was observed that the top four most informative descriptors for the steroidal dataset were C-025 (atom centered fragments; R--CR--R), ESpm14u and ESpm13r (connectivity or bonding between atoms) and MATS6p (involved in polarizability of molecules).

As for the non-steroidal dataset, the most important feature was determined to be molecular graph, polarizability and electronegativity of the compound. Therefore, the authors concluded that the polarizability of the compounds along with a suitable shape may be the determining factors needed for both types of AIs for reaching its intended target. Additionally, Worachartcheewan et al. (2014a) conducted a large-scale study on AIs using SMILES-based descriptors and discovered that the most notable features were the presence of cyclic rings (i.e. found in steroidal inhibitors) and the presence in the molecular structure of oxygen atoms together with double bonds that are disconnected in the structure (++++O---B2==) (i.e. analogous to the ketone groups present in the natural substrate, androstenedione) are important in increasing aromatase inhibitory activity. Furthermore, Ghodsi and Hemmateenejad (2016) conducted QSAR on AIs based on long-chained diarylalkylimidazole and diarylalkyltriazole (non-steroidal) molecule skeletons in which they determined important features to include

EXCLI Journal 2018;17:688-708 – ISSN 1611-2156
Received: June 08, 2018, accepted: July 10, 2018, published: July 20, 2018

Table 5: Summary of key features for aromatase inhibition as deduced from QSAR modeling. Example descriptors are shown in the parenthesis.

Steroidal	Non-steroidal
• Number of cyclic rings • Lipophilic • Polar (TPSA, MATS6p)	• Nitrogen-containing descriptors (G(N···N)) • Polarizability (HOMO, HOMO-LUMO) • Hydrogen bond acceptors (nHAcc)

TPSA: an empirical measure of the polar surface area of a molecule, MATS6p: Moran autocorrelation of lag 6 as weighted by polarizability, G(N···N): Sum of geometric distances between N···N, HOMO: the highest energy molecular orbitals, HOMO-LUMO: the energetic difference between the HOMO and LUMO states, nHAcc: the number of hydrogen bond acceptors present in a molecule

geometrical distances of N and N atoms as well as that of O and O atoms (i.e. nitrogen atoms of azole rings as well as oxygen atoms from steroidal ketones), length of the bridge carbon chain (i.e. methylene spacer separating the azole ring and the phenol ring), number of triple bonds (i.e. triple bond in the nitrile or CN that is an integral part of FDA-approved AIs), HOMO energy (i.e. localization of HOMO orbital predominantly in the imidazolyl ring), etc. Furthermore, Nantasenamat et al. (2014) studied flavonoids with aromatase inhibitory activity, and found that active compounds were found to exhibit smaller size, higher degree of rigidity, lower polarity and charge distribution, and afforded lower electron-withdrawing tendency and higher chemical reactivity than those of the inactive class.

As for the analysis of 3D-QSAR models utilizing descriptors based on molecular fields, Castellano et al. (2008) revealed that the aligned molecules showed the presence of three major regions in which two were pertinent for aromatase inhibition (i.e. one important region afforded both electrostatic and hydrogen bonds while the second important region was occupied by the characteristic azole moiety) whereas the other region was not important for the activity. Adhikari et al. (2017a) performed an extensive study employing a wide range of QSAR models including 2D and 3D QSAR as well as molecular docking to also confirm the importance of the electrostatic property of the nitrogen-containing azole moiety, *p*-cyanophenyl moiety, *p*-nitrophenyl, hydro-phobicity as well as the appropriate size and shape of AIs were crucial for aromatase inhibitory activity. Xie et al. (2015) performed both CoMFA and CoMSIA studies and both studies further supported the importance of bulky steric groups as well as the importance of electrostatic properties pertaining to the presence of azole nitrogen atoms.

CONCLUSION

In spite of extensive research (i.e. medicinal chemistry and QSAR work) in the quest of novel and potent aromatase inhibitors, there has been only a few review articles on the topic (Adhikari et al. 2017b; Yadav et al. 2015). Briefly, Yadav et al. (2015) carried out a review focusing on molecular modeling as well as QSAR of steroidal AIs whereas Adhikari et al. (2017b) based their review on QSAR studies of non-steroidal AIs. Herein, we have performed an extensive review on the mechanistic insights of pertinent features as derived from all previous QSAR models of both steroidal and non-steroidal AIs. Moreover, this review also summarized the experimental setup of all QSAR studies such that a comparative and holistic analysis could be deduced and used for providing a glimpse on the current state-of-the-art in the field as well as serving as the basis for planning future studies to further gain insights on aromatase inhibition. For example, it is anticipated that insights gained from QSAR models alone provides one aspect where it may be beneficial to also call upon complementary methodologies

such as structure-based and systems-based approaches to facilitate and augment the ligand-based QSAR approach. In fact, there have been a few studies employing a multitude of ligand, structure and systems-based approaches in studying aromatase inhibition (Simeon et al. 2016b); Suvannang et al. 2011) and future works along this line is expected to be of great benefit to the scientific community.

Conflict of interests

The authors have declared that no competing interests exist.

Acknowledgements

This work is supported by the Office of Higher Education Commission and the Thailand Research Fund (No. MRG6180226) and the New Researcher Grant (A31/2561) from Mahidol University. This work is also supported by the Center of Excellence on Medical Biotechnology (CEMB), S&T Postgraduate Education and Research Development Office (PERDO), Office of Higher Education Commission (OHEC), Thailand. Partial support from the annual budget grant (B.E. 2557-2559) of Mahidol University is also acknowledged.

REFERENCES

Adhikari N, Amin SA, Jha T, Gayen S. Integrating regression and classification-based QSARs with molecular docking analyses to explore the structure-antiaromatase activity relationships of letrozole-based analogs. Can J Chem. 2017a;95:1285–95.

Adhikari N, Amin SA, Saha A, Jha T. Combating breast cancer with non-steroidal aromatase inhibitors (NSAIs): Understanding the chemico-biological interactions through comparative SAR/QSAR study. Eur J Med Chem. 2017b;137:365–438.

Ahmad I, Shagufta. Recent developments in steroidal and nonsteroidal aromatase inhibitors for the chemoprevention of estrogen-dependent breast cancer. Eur J Med Chem. 2015;102:375–86.

Aiken LS, West SG, Pitts SC. Multiple linear regression. In: Schinka JA, Velicer WF (eds.): Handbook of psychology: Research methods in psychology, Vol. 2 (pp 483-507). Hoboken, NJ: Wiley, 2003.

American Cancer Society. Global cancer facts & figures. 3rd ed. Atlanta, GA: American Cancer Society, 2015.

Avendaño C, Menéndez JC. Anticancer drugs that inhibit hormone action. In: Avendaño C, Menéndez JC. Medicinal chemistry of anticancer drugs (pp 53-91). Amsterdam: Elsevier, 2008.

Awasthi M, Singh S, Pandey VP, Dwivedi UN. Molecular docking and 3D-QSAR-based virtual screening of flavonoids as potential aromatase inhibitors against estrogen-dependent breast cancer. J Biomol Struct Dyn. 2015;33:804–19.

Bak A, Polanski J. Modeling robust QSAR 3: SOM-4D-QSAR with iterative variable elimination IVE-PLS: application to steroid, azo dye, and benzoic acid series. J Chem Inf Model. 2007;47:1469–80.

Barigye SJ, Freitas MP, Ausina P, Zancan P, Sola-Penna M, Castillo-Garit JA. Discrete Fourier transform-based multivariate image analysis: application to modeling of aromatase inhibitory activity. ACS Comb Sci. 2018;20:75–81.

Beger RD, Buzatu DA, Wilkes JG, Lay JO. ^{13}C NMR Quantitative Spectrometric Data-Activity Relationship (QSDAR) models of steroids binding the aromatase enzyme. J Chem Inf Comput Sci. 2001;41:1360–6.

Beger RD, Wilkes JG. Comparative structural connectivity spectra analysis (CoSCoSA) models of steroids binding to the aromatase enzyme. J Mol Recognit. 2002;15:154–62.

Beger RD, Harris S, Xie Q. Models of steroid binding based on the minimum deviation of structurally assigned 13C NMR spectra analysis (MiDSASA). J Chem Inf Comput Sci. 2004;44:1489–96.

Bonnefoi HR, Smith IE, Dowsett M, Trunet PF, Houston SJ, da Luz RJ, et al. Therapeutic effects of the aromatase inhibitor fadrozole hydrochloride in advanced breast cancer. Br J Cancer. 1996;73:539–42.

Breiman L. Random forests. Mach Learn. 2001;45:5–32.

Brueggemeier RW, Hackett JC, Diaz-Cruz ES. Aromatase inhibitors in the treatment of breast cancer. Endocr Rev. 2005;26:331–45.

Castellano S, Stefancich G, Ragno R, Schewe K, Santoriello M, Caroli A, et al. CYP19 (aromatase): exploring the scaffold flexibility for novel selective inhibitors. Bioorg Med Chem. 2008;16:8349–58.

Cavalli A, Greco G, Novellino E, Recanatini M. Linking CoMFA and protein homology models of enzyme-

inhibitor interactions: an application to non-steroidal aromatase inhibitors. Bioorg Med Chem. 2000;8: 2771–80.

Cavalli A, Bisi A, Bertucci C, Rosini C, Paluszcak A, Gobbi S, et al. Enantioselective nonsteroidal aromatase inhibitors identified through a multidisciplinary medicinal chemistry approach. J Med Chem. 2005;48:7282–9.

Coombes RC, Kilburn LS, Snowdon CF, Paridaens R, Coleman RE, Jones SE, et al. Survival and safety of exemestane versus tamoxifen after 2-3 years' tamoxifen treatment (Intergroup Exemestane Study): a randomised controlled trial. Lancet. 2007;369:559–70.

Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.

Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc. 1988;110:5959–67.

Cros AFA. Action de l'alcohol amyliquesur l'organisme. Strasbourg: University of Strasbourg, 1863.

Crum Brown A, Fraser TR. On the connection between chemical constitution and physiological action; with special reference to the physiological action of the salts of the ammonium bases derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. J Anat Physiol. 1868;2:224–42.

Dai Y, Wang Q, Zhang X, Jia S, Zheng H, Feng D, et al. Molecular docking and QSAR study on steroidal compounds as aromatase inhibitors. Eur J Med Chem. 2010;45:5612–20.

Dai Y, Xiao Y, Wang Q, Wei S, Zhang X, Ma Z, et al. Syntheses and QSAR studies of benzylimidazole derivatives and benzylcarbazole as potential aromatase inhibitors. Asian J Chem. 2014;26: 2381-8.

Demers LM, Boucher AE, Santen RJ. Aminoglutethimide therapy in breast cancer: relationship of blood levels to drug-related side effects. Clin Physiol Biochem. 1987;5:287–91.

Dowsett M, Lloyd P. Comparison of the pharmacokinetics and pharmacodynamics of unformulated and formulated 4-hydroxyandrostenedione taken orally by healthy men. Cancer Chemother Pharmacol. 1990;27: 67–71.

DrugBank. Formestane [Internet]. 2013 [cited 2018 Jun 2] Available from: https://www.drugbank.ca/drugs/DB08905.

Dutta U, Pant K. Aromatase inhibitors: past, present and future in breast cancer therapy. Med Oncol. 2008; 25:113–24.

Eisen A, Trudeau M, Shelley W, Messersmith H, Pritchard KI. Aromatase inhibitors in adjuvant therapy for hormone receptor positive breast cancer: a systematic review. Cancer Treat Rev. 2008;34:157–74.

Fujita T, Winkler DA. Understanding the roles of the "two QSARS". J Chem Inf Model. 2016;56:269–74.

Ghodsi R, Hemmateenejad B. QSAR study of diarylalkylimidazole and diarylalkyltriazole aromatase inhibitors. Med Chem Res. 2016;25:834–42.

Ghosh D, Griswold J, Erman M, Pangborn W. Structural basis for androgen specificity and oestrogen synthesis in human aromatase. Nature. 2009;457:219–23.

Gironés X, Carbó-Dorca R. Molecular quantum similarity-based QSARs for binding affinities of several steroid sets. J Chem Inf Comput Sci. 2002;42:1185–93.

Graves PE, Salhanick HA. Stereoselective inhibition of aromatase by enantiomers of aminoglutethimide. Endocrinology. 1979;105:52–7.

Gueto C, Torres J, Vivas-Reyes R. CoMFA, LeapFrog and blind docking studies on sulfonanilide derivatives acting as selective aromatase expression regulators. Eur J Med Chem. 2009;44:3445–51.

Hansch C, Fujita T. *p* -σ-π analysis. A method for the correlation of biological activity and chemical structure. J Am Chem Soc. 1964;86:1616–26.

Hansch C, Muir RM, Fujita T, Maloney PP, Geiger F, Streich M. The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. J Am Chem Soc. 1963;85:2817–24.

Helland IS. On the structure of partial least squares regression. Comun Stat Simul C. 1988;17:581–607.

Helland IS. Some theoretical aspects of partial least squares regression. Chemometr Intell Lab Sys. 2001; 58:97–107.

Hughes SW, Burley DM. Aminoglutethimide: a "side-effect" turned to therapeutic advantage. Postgrad Med J. 1970;46:409–16.

Karelson M, Lobanov VS, Katritzky AR. Quantum-chemical descriptors in QSAR/QSPR studies. Chem Rev. 1996;96:1027–44.

Karkola S, Wähälä K. The binding of lignans, flavonoids and coumestrol to CYP450 aromatase: a molecular modelling study. Mol Cell Endocrinol. 2009;301: 235–44.

Kishore DP, Rana A, Jain UK, Rao PM. Pharmacophore-based 3D-QSAR studies of aromatase inhibitors. Asian J Chem. 2013;25: 10588-94.

Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem. 1994;37:4130–46.

Kubinyi H. QSAR and 3D QSAR in drug design, Part 1: Methodology. Drug Discov Today. 1997;2:457–67.

Kubinyi H, Hamprecht FA, Mietzner T. Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. J Med Chem. 1998;41:2553–64.

Kumar SP, Jha PC, Jasrai YT, Pandya HA. The effect of various atomic partial charge schemes to elucidate consensus activity-correlating molecular regions: a test case of diverse QSAR models. J Biomol Struct Dyn. 2016;34:540–59.

Lee S, Barron MG. 3D-QSAR study of steroidal and azaheterocyclic human aromatase inhibitors using quantitative profile of protein-ligand interactions. J Cheminform. 2018;10(1):2.

Leonetti F, Favia A, Rao A, Aliano R, Paluszcak A, Hartmann RW, et al. Design, synthesis, and 3D QSAR of novel potent and selective aromatase inhibitors. J Med Chem. 2004;47:6792–803.

Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 2001;46: 3–26.

Loge C, Le Borgne M, Marchand P, Robert J-M, Le Baut G, Palzer M, et al. Three-dimensional model of cytochrome P450 human aromatase. J Enzyme Inhib Med Chem. 2005;20:581–5.

Mittal RR, McKinnon RA, Sorich MJ. The effect of molecular fields, lattice spacing and analysis options on comfa predictive ability. QSAR Comb Sci. 2009; 28:637–44.

Muir RM, Hansch CH, Gallup AH. Growth regulation by organic compounds. Plant Physiol. 1949;24:359–66.

Nagar S, Islam MA, Das S, Mukherjee A, Saha A. Pharmacophore mapping of flavone derivatives for aromatase inhibition. Mol Divers. 2008;12:65–76.

Nagar S, Saha A. Modeling of diarylalkyl-imidazole and diarylalkyl-triazole derivatives as potent aromatase inhibitors for treatment of hormone-dependent cancer. J Comput Chem. 2010a;31:2342–53.

Nagar S, Saha A. Exploring benzcyclo derivatives as potent aromatase inhibitors using ligand-based modeling studies. Eur J Med Chem. 2010b;45:4307–15.

Nagy PI, Tokarski J, Hopfinger AJ. Molecular shape and QSAR analyses of a family of substituted dichlorodiphenyl aromatase inhibitors. J Chem Inf Comput Sci. 1994;34:1190–7.

Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. A practical overview of quantitative structure-activity relationship. EXCLI J. 2009;8: 74–88.

Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V. Advances in computational methods to predict the biological activity of compounds. Expert Opin Drug Discov. 2010;5:633–54.

Nantasenamat C, Li H, Mandi P, Worachartcheewan A, Monnor T, Isarankura-Na-Ayudhya C, et al. Exploring the chemical space of aromatase inhibitors. Mol Divers. 2013a;17:661–77.

Nantasenamat C, Worachartcheewan A, Prachayasittikul S, Isarankura-Na-Ayudhya C, Prachayasittikul V. QSAR modeling of aromatase inhibitory activity of 1-substituted 1,2,3-triazole analogs of letrozole. Eur J Med Chem. 2013b;69:99–114.

Nantasenamat C, Worachartcheewan A, Mandi P, Monnor T, Isarankura-Na-Ayudhya C, Prachayasittikul V. QSAR modeling of aromatase inhibition by flavonoids using machine learning approaches. Chem Pap. 2014;68:697–713.

Nantasenamat C, Prachayasittikul V. Maximizing computational tools for successful drug discovery. Expert Opin Drug Discov. 2015;10:321–9.

Narayana BL, Pran Kishore D, Balakumar C, Rao KV, Kaur R, Rao AR, et al. Molecular modeling evaluation of non-steroidal aromatase inhibitors. Chem Biol Drug Des. 2012;79:674–82.

Nelson LR, Bulun SE. Estrogen production and action. J Am Acad Dermatol. 2001;45:S116-24.

Oprea TI, García AE. Three-dimensional quantitative structure-activity relationships of steroid aromatase inhibitors. J Comput Aided Mol Des. 1996;10:186–200.

Paoletta S, Steventon GB, Wildeboer D, Ehrman TM, Hylands PJ, Barlow DJ. Screening of herbal constituents for aromatase inhibitory activity. Bioorg Med Chem. 2008;16:8466–70.

Phanus-umporn C, Shoombuatong W, Prachayasittikul V, Anuwongcharoen N, Nantasenamat C. Correction: Privileged substructures for anti-sickling activity *via* cheminformatic analysis. RSC Adv. 2018;8:8233.

Pingaew R, Mandi P, Prachayasittikul V, Prachayasittikul S, Ruchirawat S, Prachayasittikul V. Synthesis, molecular docking, and QSAR study of sulfonamide-based indoles as aromatase inhibitors. Eur J Med Chem. 2018;143:1604–15.

Polanski J, Gieleciak R. The comparative molecular surface analysis (CoMSA) with modified uniformative variable elimination-PLS (UVE-PLS) method: application to the steroids binding the aromatase enzyme. J Chem Inf Comput Sci. 2003;43:656–66.

Prachayasittikul V, Pingaew R, Worachartcheewan A, Sitthimonchai S, Nantasenamat C, Prachayasittikul S, et al. Aromatase inhibitory activity of 1,4-naphthoquinone derivatives and QSAR study. EXCLI J. 2017;16: 714–26.

Pratiwi R, Malik AA, Schaduangrat N, Prachayasittikul V, Wikberg JES, Nantasenamat C, et al. CryoProtect: A Web server for classifying antifreeze proteins from nonantifreeze proteins. J Chem. 2017;2017: 9861752.

Puri M, Solanki A, Padawer T, Tipparaju SM, Moreno WA, Pathak Y. Introduction to artificial neural network (ANN) as a predictive tool for drug design, discovery, delivery, and disposition. In: Puri M, Pathak Y, Sutariya V, Tipparaju S, Moreno W: Artificial neural network for drug design, delivery and disposition (pp 3-13). New York: Academic Press, 2016.

Recanatini M. Comparative molecular field analysis of non-steroidal aromatase inhibitors related to fadrozole. J Comput Aided Mol Des. 1996;10:74–82.

Recanatini M, Cavalli A. Comparative molecular field analysis of non-steroidal aromatase inhibitors: an extended model for two different structural classes. Bioorg Med Chem. 1998;6:377–88.

Roy PP, Roy K. Docking and 3D-QSAR studies of diverse classes of human aromatase (CYP19) inhibitors. J Mol Model. 2010a;16:1597–616.

Roy PP, Roy K. Molecular docking and QSAR studies of aromatase inhibitor androstenedione derivatives. J Pharm Pharmacol. 2010b;62:1717–28.

Russo J, Hasan Lareef M, Balogh G, Guo S, Russo IH. Estrogen and its metabolites are carcinogenic agents in human breast epithelial cells. J Steroid Biochem Mol Biol. 2003;87:1–25.

Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern. 1991;21:660–74.

Santen RJ, Misbin RI. Aminoglutethimide: review of pharmacology and clinical use. Pharmacotherapy. 1981;1:95–120.

Santen RJ, Santner S, Davis B, Veldhuis J, Samojlik E, Ruby E. Aminoglutethimide inhibits extraglandular estrogen production in postmenopausal women with breast carcinoma. J Clin Endocrinol Metab. 1978;47: 1257–65.

Santen RJ, Samojlik E, Wells SA. Resistance of the ovary to blockade of aromatization with aminoglutethimide. J Clin Endocrinol Metab. 1980;51:473–7.

Santen RJ, Worgul TJ, Lipton A, Harvey H, Boucher A, Samojlik E, et al. Aminoglutethimide as treatment of postmenopausal women with advanced breast carcinoma. Ann Intern Med. 1982;96:94–101.

Shoombuatong W, Prachayasittikul V, Prachayasittikul V, Nantasenamat C. Prediction of aromatase inhibitory activity using the efficient linear method (ELM). EXCLI J. 2015;14:452–64.

Shoombuatong W, Prathipati P, Owasirikul W, Worachartcheewan A, Simeon S, Anuwongcharoen N, et al. Towards the revival of interpretable QSAR models. In: Roy K (ed.): Advances in QSAR modeling (pp 3-55). Cham: Springer International Publ., 2017.

Simeon S, Anuwongcharoen N, Shoombuatong W, Malik AA, Prachayasittikul V, Wikberg JES, et al. Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking. PeerJ 2016a;4:e2322.

Simeon S, Spjuth O, Lapins M, Nabu S, Anuwongcharoen N, Prachayasittikul V, et al. Origin of aromatase inhibitory activity via proteochemometric modeling. PeerJ. 2016b;4:e1979.

Simpson ER. Aromatase cytochrome P450, the enzyme responsible for estrogen biosynthesis. Endocr Rev. 1994;15:342–55.

Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput. 2004;14:199–222.

Song Z, Liu Y, Dai Z, Liu W, Zhao K, Zhang T, et al. Synthesis and aromatase inhibitory evaluation of 4-N-nitrophenyl substituted amino-4H-1,2,4-triazole derivatives. Bioorg Med Chem. 2016;24:4723–30.

Sulea T, Oprea TI, Muresan S, Chan SL. A different method for steric field evaluation in comfa improves model robustness. J Chem Inf Comput Sci. 1997;37: 1162–70.

Suvannang N, Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V. Molecular docking of aromatase inhibitors. Molecules. 2011;16:3597–617.

Suvannang N, Preeyanon L, Malik AA, Schaduangrat N, Shoombuatong W, Worachartcheewan A, et al. Probing the origin of estrogen receptor alpha inhibition*via* large-scale QSAR study. RSC Adv. 2018;8: 11344–56.

Thürlimann B, Hess D, Köberle D, Senn I, Ballabeni P, Pagani O, et al. Anastrozole ('Arimidex') versus tamoxifen as first-line therapy in postmenopausal women with advanced breast cancer: results of the double-blind cross-over SAKK trial 21/95--a sub-study of the TARGET (Tamoxifen or "Arimidex" Randomized Group Efficacy and Tolerability) trial. Breast Cancer Res Treat. 2004;85:247–54.

Todeschini R, Consonni V. Handbook of molecular descriptors. Weinheim: Wiley-VCH, 2000.

Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform. 2010; 29:476–88.

van Westen GJP, Wegner JK, IJzerman AP, van Vlijmen HWT, Bender A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. Med Chem Commun. 2011;2:16–30.

WHO. Breast Cancer [Internet]. Cancer 2015 [cited 2018 Jun 1] Available from: http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/.

Win TS, Malik AA, Prachayasittikul V, Wikberg JES, Nantasenamat C, Shoombuatong W. HemoPred: a web server for predicting the hemolytic activity of peptides. Future Med Chem. 2017;9:275–91.

Worachartcheewan A, Mandi P, Prachayasittikul V, Toropova AP, Toropov AA, Nantasenamat C. Large-scale QSAR study of aromatase inhibitors using SMILES-based descriptors. Chemometr Intell Lab. 2014a;138:120–6.

Worachartcheewan A, Suvannang N, Prachayasittikul S, Prachayasittikul V, Nantasenamat C. Probing the origins of aromatase inhibitory activity of disubstituted coumarins via QSAR and molecular docking. EXCLI J. 2014b;13:1259–74.

Worachartcheewan A, Shoombuatong W, Pidetcha P, Nopnithipat W, Prachayasittikul V, Nantasenamat C. Predicting metabolic syndrome using the random forest method. Sci World J. 2015;2015:581501.

Xie H, Qiu K, Xie X. 3D QSAR studies, pharmacophore modeling and virtual screening on a series of steroidal aromatase inhibitors. Int J Mol Sci. 2014;15: 20927–47.

Xie H, Qiu K, Xie X. Pharmacophore modeling, virtual screening, and 3D-QSAR studies on a series of non-steroidal aromatase inhibitors. Med Chem Res. 2015; 24:1901–15.

Yadav MR, Barmade MA, Tamboli RS, Murumkar PR. Developing steroidal aromatase inhibitors-an effective armament to win the battle against breast cancer. Eur J Med Chem. 2015;105: 1–38.

EXCLI Journal 2018;17:734-752 – ISSN 1611-2156
Received: June 20, 2018, accepted: July 10, 2018, published: July 25, 2018

Review article:

UNRAVELING THE BIOACTIVITY OF ANTICANCER PEPTIDES AS DEDUCED FROM MACHINE LEARNING

Watshara Shoombuatong, Nalini Schaduangrat, Chanin Nantasenamat*

Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand

* Corresponding author: E-mail: chanin.nan@mahidol.edu (C.N.); Phone: +66 2 441 4371; Fax: +66 2 441 4380

http://dx.doi.org/10.17179/excli2018-1447

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/).

ABSTRACT

Cancer imposes a global health burden as it represents one of the leading causes of morbidity and mortality while also giving rise to significant economic burden owing to the associated expenditures for its monitoring and treatment. In spite of advancements in cancer therapy, the low success rate and recurrence of tumor has necessitated the ongoing search for new therapeutic agents. Aside from drugs based on small molecules and protein-based biopharmaceuticals, there has been an intense effort geared towards the development of peptide-based therapeutics owing to its favorable and intrinsic properties of being relatively small, highly selective, potent, safe and low in production costs. In spite of these advantages, there are several inherent weaknesses that are in need of attention in the design and development of therapeutic peptides. An abundance of data on bioactive and therapeutic peptides have been accumulated over the years and the burgeoning area of artificial intelligence has set the stage for the lucrative utilization of machine learning to make sense of these large and high-dimensional data. This review summarizes the current state-of-the-art on the application of machine learning for studying the bioactivity of anticancer peptides along with future outlook of the field. Data and R codes used in the analysis herein are available on GitHub at https://github.com/Shoombuatong2527/anticancer-peptides-review.

Keywords: cancer, anticancer, antitumor, anticancer peptides, host defense peptides, bioactivity, machine learning, QSAR

INTRODUCTION

Cancer is now regarded as the second leading cause of death, and remains a major cause of morbidity throughout the world (Arnold et al., 2015) with lung, liver, colorectal, stomach and breast cancer representing the most common types of cancers occurring worldwide (WHO, 2018a). Estimates from GLOBOCAN indicate that about 14.1 million new cancer cases encompassed approximately 8.8 million deaths in 2015 (Ferlay et al., 2015; WHO, 2018a). In addition, the main mechanisms by which cancers are formed include abnormal, uncontrollable cell growth that leads to the formation of tumors which can then undergo angiogenesis and continue to become metastatic (Felício et al., 2017). Despite recent advances in cancer treatments, such as radiation therapy, targeted therapy or chemotherapeutic agents (Thundimadathil 2012), they have a relatively low success rate and present a risk of recurrence. For instance, the process of killing cancer using chemotherapeutic agents is often associated with deleterious effects, including damages to normal

cells and tissues, and lead to chemical resistance whereby adaptation mutations of cancer cells may occur (Hoskin and Ramamoorthy 2008). Therefore, the discovery and development of a new class of anticancer drugs has become crucial. Furthermore, the situation has become worse due to the fact that many new cancers arise from bacterial and viral etiological agents (Vedham et al., 2014). This fact coupled with the increase in antimicrobial resistance (AMR), especially the multidrug resistant variants, has raised concern. To this effect, the WHO has emphasized an urgency for the discovery of new therapeutic agents (WHO, 2018b).

Peptide therapeutics have attracted great interest for development as drug candidates as they are regarded to be safe, efficacious, highly selective with good tolerability as well as exhibit attractive pharmacological profiles (Craik et al., 2013; Vlieghe et al., 2010; Lau and Dunn, 2018; Fosgerau and Hoffmann, 2015). A summary on the strengths and weaknesses of therapeutic peptides are provided in Figure 1. Owing to their intrinsically smaller size as compared to protein-based biopharmaceuticals, peptides are therefore more economical to produce due to lesser production complexity (Fosgerau and Hoffmann, 2015) while at the same time possess more agility in their pharmacokinetics. The aforementioned properties are distinguishing features that set them apart from small molecules-based drugs and protein-based therapeutics. Thus far, there are more than 7,000 naturally occurring peptides in existence that have been shown to afford a wide range of bioactivities (e.g. tumor homing, antihypertensive, antiparasitic, antiviral, antiangiogenic, antibiofilm, antimicrobial, anticancer, etc.) that can consequently be applied to target various diseases such as cancer, diabetes, cardiovascular diseases, etc. (O'Brien-Simpson et al., 2018; Jin and Weinberg 2018; Karpiński and Adamczak 2018). As of now, 60 peptide-based drugs have been FDA-approved (Usmani et al., 2017) while another 150 peptides (Lau and Dunn 2018) are currently in the pipeline of preclinical and clinical studies.

The breakthrough discovery of cecropin, the first antimicrobial peptide (AMP) (i.e. isolated from injecting silk moth, *Hyalophora cecropia*, with bacteria) was reported by Steiner et al. (1981). In another landmark study conducted by Zasloff (1987), AMPs from the African clawed frog, *Xenopus laevis*, were isolated and characterized for their role in the immune defense and is known as magainins. Since then, thousands of AMPs have been found in almost all living organisms such as plants, bacteria, fungi, animals etc. (Li et al., 2012). Over the past decade, the use of AMPs as therapeutic agents for treating diseases have increased constantly.

The process of understanding the importance of AMPs might be useful for the discovery of new and resistance-free therapies for infectious as well as non-infectious diseases. Antimicrobial peptides constitute a mechanism of immune defense of the innate immune system with low antigenicity (Iwasaki et al., 2009; Pasupuleti et al., 2012) that can be found in numerous eukaryotic organisms of different species (Reddy et al., 2004). More recently, research on AMPs have elucidated that these peptides also provide anticancer activity and thus termed anticancer peptides (ACPs). ACPs have been found to exhibit short time-frame of interaction (i.e. decreases the probability of resistance), low toxicity (i.e. not devoid of side effects as it may harm normal cells), specificity, good solubility as well as good tumor penetration thereby indicating the great potential for the use of ACPs in cancer therapy (Domalaon et al., 2016; Gaspar et al., 2015; Figueiredo et al., 2014; Riedl et al., 2011).

Since they are not traditional drugs, the clinical development of therapeutic peptides face numerous challenges owing to their weaknesses as summarized in Figure 1. Stability of peptides (i.e. lack of correlation between *in vitro* experiments and its efficacy in *in vivo* models) is a challenging issue. In spite of this, promising results have been sparingly been demonstrated in some animal studies (Makobongo et al., 2012; Deslouches et al., 2007; Berge et al., 2010; Camilio et al., 2014;

EXCLI Journal 2018;17:734-752 – ISSN 1611-2156
Received: June 20, 2018, accepted: July 10, 2018, published: July 25, 2018

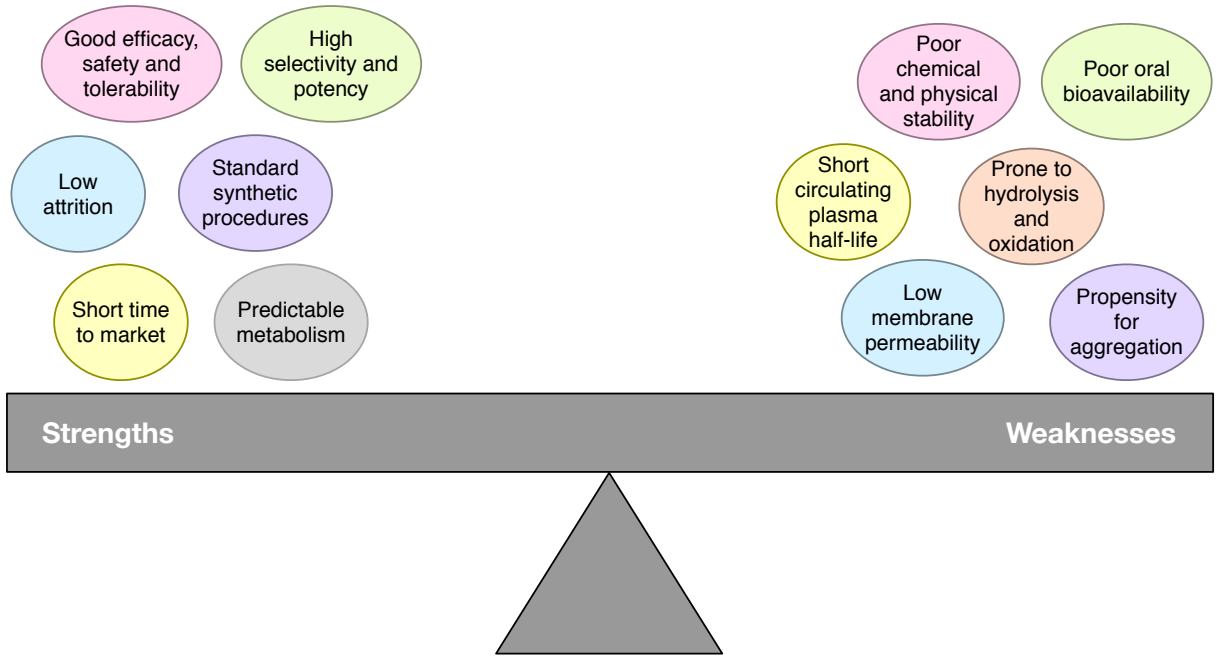

Figure 1: Strengths and weaknesses of therapeutic peptides. Concepts summarized from Fosgerau and Hoffmann, 2015.

Makovitzki et al., 2009) in which good efficacy of peptides were able to establish *in vitro* stability with bioavailability in animal models. Another major drawback of therapeutic peptides is their poor oral bioavailability. This can be addressed by conjugating the peptide with a delivery system that allows it to bypass the digestive system and thus, enhance the pharmacokinetic properties of such peptides. Several studies have been conducted on the modifications and/or conjugations (e.g. substitution with non-canonical amino acid, peptide-peptide hybridization, target or polymer modification, PEGylation etc.) of therapeutic peptides (Narayana et al., 2015; Braunstein et al., 2004; Papo and Shai, 2003; Hu et al., 2016; Spinks et al., 2017; Kelly et al., 2016; Li et al., 2016). Another concern is the short half-life of peptides. However, it should be noted that it is this particular characteristic of therapeutic peptides that allows it to escape resistance unlike other oncogenic therapies. However, research on improving the half-life of peptides without compromising their potency is currently an active area of research (Hao et al., 2015; Podust et al., 2013; Schellenberger et al., 2009; Garay et al., 2012; Penchala et al., 2015). Despite some limitations, no other class of peptides have been able to surpass the multi-functionality of bioactive/therapeutic peptides and thus, these peptides possess high potential for use in many avenues of clinical applications.

The post-genomic era has brought about the birth of several omics (e.g. peptidomics, proteomics, glycomics, transcriptomics, interactomics, etc.) in our attempts to understand the fundamentals of life and how we can contribute to sustainability and the improvement of the quality of life (i.e. development of new diagnostics, therapeutics, etc.). These data are amassing at an exponential rate with no slowing down in hindsight, which sets the stage for the utilization of machine learning in making sense of these data and translating them into useful and actionable insights. There have been extensive reports on the utilization of machine learning approaches for correlating the sequences of therapeutic peptides with their biological activity (Shi et al., 1998; Nagarajan et al., 2006; Alam and Khan, 2014; Mohseni Bababdani and Mousavi, 2013; Tong et al., 2014; Li et al., 2017). A review of the literature indicated that there are currently no review articles concerning the use of machine learning and quantitative

structure-activity relationship (QSAR) as applied to therapeutic peptides. However, there are a few review articles examining the use of QSAR for studying the biological activity of peptides at the general level particularly with emphasis on food protein-derived bioactive peptides (Nongonierma and FitzGerald 2016), peptides in general (as well as proteins and nucleic acids) (Zhou et al., 2008), peptides in general (as well as chemical molecules and proteins) (Du et al., 2008). In a series of recent articles, Lee et al. (2016, 2017, 2018) examined another facet on the use of machine learning (i.e. particularly support vector machine) together with targeted experiments (i.e. killing assays and small-angle X-ray scattering (SAXS) experiments) to explore the membrane activity in undiscovered peptide sequence space in which the aim was not on the antimicrobial activity but on the membrane curvature that is necessary for the activity and the subsequent relationship to sequence homology.

To the best of our knowledge, this review article represents the first systematic review on the utilization of machine learning for studying the bioactivity of anticancer peptides. It is hoped that this review would help contribute to further growth and expansion of the field by providing readers with the current state-of-the-art of the field as well as expected future trends and outlook.

ANTICANCER PEPTIDES

ACPs are small peptides that usually contain 5 to 50 amino acid residues while possessing high hydrophobicity and a positive net charge (i.e. cationic in nature) (Melo et al., 2011). Thus, ACPs can interact with anionic cell membrane components of cancer cells and then selectively kill cancer cells. Additionally, ACPs can interfere with cancer cells by causing apoptosis mediated via mitochondrial disruption (Chen et al., 2001), triggering necrosis via cell lysis (Papo et al., 2006), stimulate the immune system of the host and prevent tumor angiogenesis (Al-Benna et al., 2011). Being a subset of AMPs, the characteristics of ACPs are very similar. However, the physicochemical properties that drive some AMPs to possess anticancer activity is still unclear and more research is needed to understand these differences and help drive specific designs of ACPs. There have been a number of AMPs encountered in nature that possess anticancer activity, such as Aurein 1.2 (GLFDIIKKIAESF) a peptide isolated from a frog species (*Litoria aurea*), represents an AMP with antibacterial activity which was also highly active towards 55 different cancer cell lines *in vitro*, without any significant cytotoxicity (Rozek et al., 2000; Dennison et al., 2007; Giacometti et al., 2007). In addition, the human neutrophil peptide-1 (HNP-1, ACYCRIPACIAGERRYGTCIYQGALWAFCC), represents an intrinsic AMP found in the innate immune system that plays a fundamental role in the defense against pathogens. The full mechanism of action of this peptide against cancer cells has not yet been established, but the activity has already been confirmed for different cancer cell lines, with very low cytotoxicity against healthy cells (McKeown et al., 2006; Gaspar et al., 2015). Furthermore, in terms of their structure, ACPs are mainly categorized as adopting either an α-helix (i.e cecropin, magainin, melittin, and buforin II) or β-sheet (i.e defensins (HNP-1, HNP-2 and HNP-3), lactoferricin B and tachyplesin) conformation due to their inability of fold into a well-defined structure in solution (Hoskin and Ramamoorthy, 2008).

In the more recent years, a lot of focus has been placed on research into ACPs with the increase in AMP databases. One such database, the antimicrobial peptide database (APD3) (Wang et al., 2016) (Available at http://aps.unmc.edu/AP/main.php) recorded as of May 10, 2018, a total of 2,981 AMPs, out of which, 215 have been classified as ACPs from various sources (animals, plants, bacteria, fungi and synthetic) (Figure 2). It should however, be noted that the different categories of the peptides (i.e. antibacterial, antiviral, antiparasitic, anticancer etc.) will

contain peptides that overlap due to some exhibiting dual properties. In addition, upon further analysis of the peptide length determining anticancer activity, it was observed that (Figure 3), out of the 214 ACPs in the database (1 peptide "AP02769" contained a non-canonical amino acid and was excluded from the analysis) 73 (34.11 %) and 60 (28.04 %) were 21-30 and 11-20 amino acids in length, respectively. Furthermore, peptides of length between 21-30 amino acids exhibiting antibacterial, antifungal, antiparasitic and antiviral activities were observed at 746 (29.83 %), 358 (33.58 %), 35 (33.98 %) and 58 (32.22 %), respectively. Therefore, the most optimal peptide length for AMPs, especially for ACPs is 21-30 and hence, it is of great value to optimize the peptide length. Moreover, upon comparison of the most frequently observed amino acid residues constituting each category of AMPs (Figure 4), it can be seen that for ACP functioning, G (Gly at 10.88 %), K (Lys at 10.25 %) and L (Leu at 11.23 %) are the most predominant. Keeping with this tread, the most frequently observed amino acid for all categories of AMPs was Gly which was found at 10.98 %, 10.88 %, 10.79 %, 10.77 % and 11.82 % for ABPs, ACPs, AFPs, APPs and AVPs, respectively. Lysine was also abundantly found in most AMP categories, as indicated in Figure 4, as it is a positively charged residue which could provide improvement in the cell and tissue penetrating properties of peptides (Li and Cho, 2012). In addition, the hydrophobic residue Leucine was also predominant in all AMPs (10.88 %, 11.28 %, 10.66 %, 10.13 % and 8.11 % for ABPs, ACPs, AFPs, APPs and AVPs, respectively) which infers its importance in the structure and function of proteins (Jayaraj et al., 2009) especially since therapeutic peptides usually contain around 50 % hydrophobic residues (Mansour et al., 2014). Furthermore, the anticancer activity of human AMPs have only been evaluated for six peptide classes (Wang et al., 2016) such as HNP-1, HNP-2, HNP-3, hBD-1, LL-37, and granulysin whose structures are shown in Figure 5.

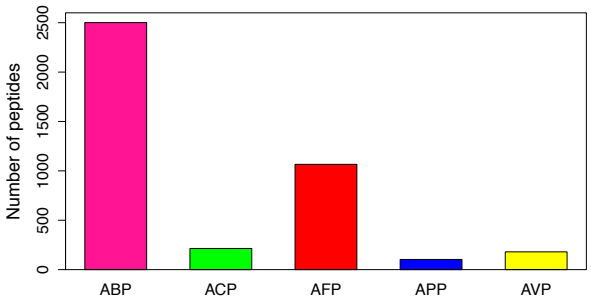

Figure 2: Bar plot of the number of antibacterial peptides (ABP), anticancer peptides (ACP), antifungal peptides (AFP), antiparasitic peptides (APP) and antiviral peptides (AVP). Data is collected from the antimicrobial peptide database (APD3) (Wang et al., 2016).

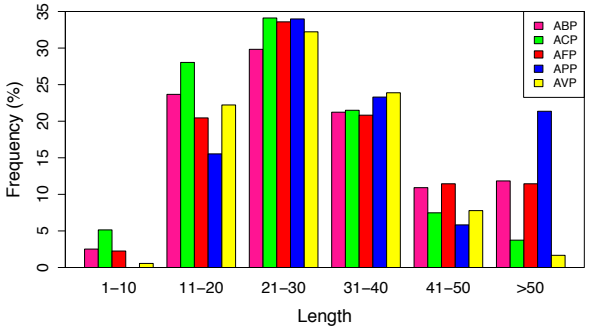

Figure 3: Bar plot showing the peptide length distribution in percentage for antibacterial peptides (ABP), anticancer peptides (ACP), antifungal peptides (AFP), antiparasitic peptides (APP) and antiviral peptides (AVP) collected from the Antimicrobial Peptide Database (APD3) (Wang et al., 2016).

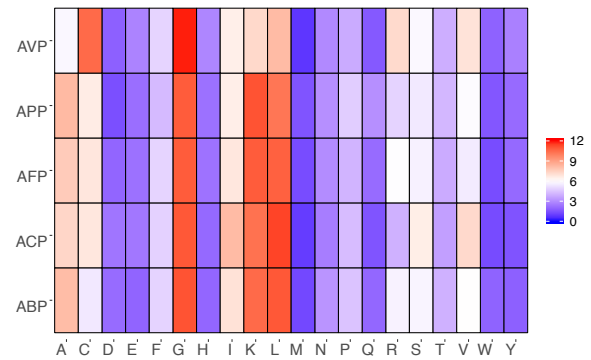

Figure 4: Heat map showing the amino acid compositions in percentage for antibacterial peptides (ABP), anticancer peptides (ACP), antifungal peptides (AFP), antiparasitic peptides (APP) and antiviral peptides (AVP). Data was collected from the Antimicrobial Peptide Database (APD3) (Wang et al., 2016).

EXCLI Journal 2018;17:734-752 – ISSN 1611-2156
Received: June 20, 2018, accepted: July 10, 2018, published: July 25, 2018

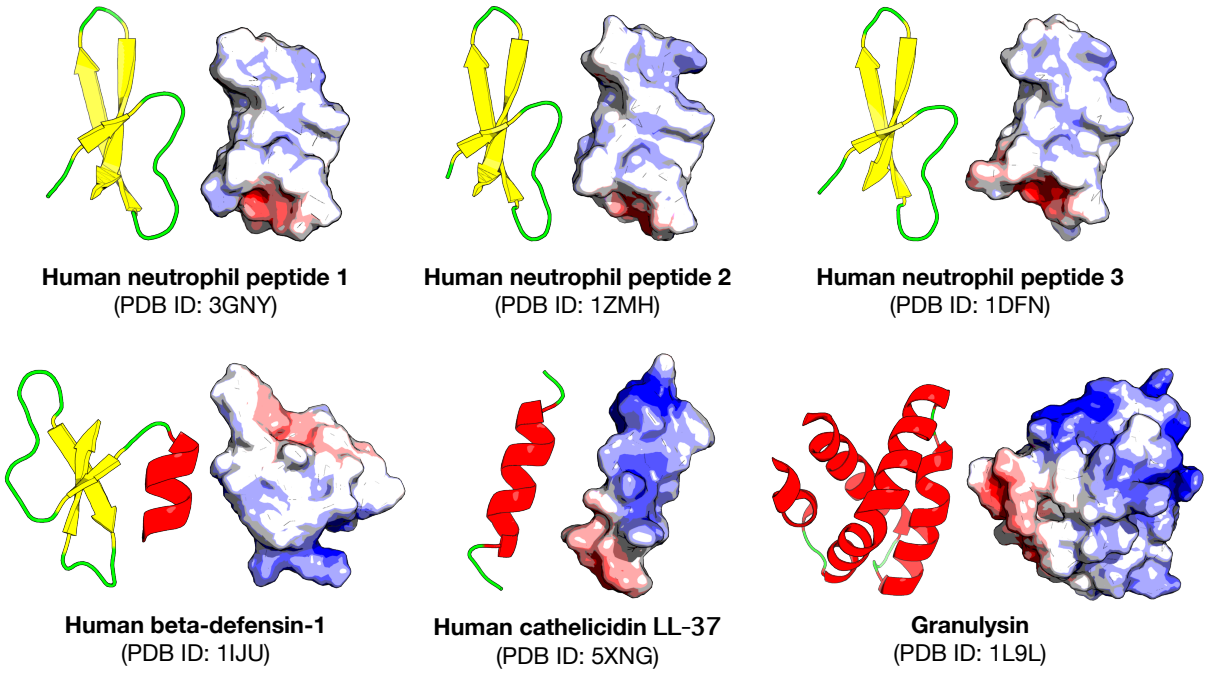

Figure 5: Structures of human-derived anticancer peptides

From thousands of available AMPs and many more that can be synthetically created, only a few have managed to reach clinical trials. Presently, only ten therapeutic peptides to treat various tumor types are currently being evaluated in various phases of preclinical and clinical trials (Felício et al., 2017). This may be due to challenging developmental processes for turning these peptides into potent pharmaceutical drugs (e.g. cost of synthesis, peptide size, charge, and solubility) (Tørfoss et al., 2012). However, with the increase in ACP research, more peptides may reach clinical trials in the future. With the help of synthetic approaches, peptide sequences could be altered so as to enhance their anticancer properties. But, the effect of these structural modifications on the physicochemical properties will need to be elucidated. Recently, these types of studies have increasingly made use of computational approaches (Prada-Gracia et al., 2016; Maccari et al., 2015; Kliger, 2010; Tyagi et al., 2013; Simeon et al., 2017). In addition, several databases exist which have pooled the data of existing sequences that pertain to bioactive or therapeutic peptides. Some of the selected databases are described in Table 1, where, out of all of the individual databases, only one named CancerPPD (Tyagi et al., 2015) is available for ACPs. Five of the databases, are the most comprehensive databases for AMPs that have been combined from various organisms. In addition, it is noteworthy that only one database, THPdb (Usmani et al., 2017) exists whereby data from FDA-approved peptides and proteins are available. Besides the databases mentioned in Table 1, there is another database, TumorHoPe (Kapoor et al., 2012), a database that provides information regarding experimentally characterized tumor targeting/homing peptides. These peptides recognize tumor tissues and tumor-associated micro-environments, including tumor metastasis. Thus, they can be used to deliver drugs selectively in tumors. In addition, a database catering to cell-penetrating peptides, CPPsite (Gautam et al., 2012) that could also be advantageous for recognizing tumors as they exhibit similar properties such as short length (10-30 amino acids), are cationic or amphipathic (containing Arg and Lys residues), and

EXCLI Journal 2018;17:734-752 – ISSN 1611-2156
Received: June 20, 2018, accepted: July 10, 2018, published: July 25, 2018

Table 1: List of selected major databases available for bioactive and therapeutic peptides.

Dataset	Description	No. of entries	URL	Reference
APD3	Database of natural AMPs with defined sequence and activity	2,981	http://aps.unmc.edu/AP/	Wang et al., 2016
BIOPEP	Database of biologically active peptide sequences	3,681	http://www.uwm.edu.pl/biochemia/index.php/en/biopep	Minkiewicz et al., 2008
CAMP	Database of sequences, structures and family-specific signatures of prokaryotic and eukaryotic AMPs	10,247	http://www.camp.bicnirrh.res.in/	Waghu et al., 2016
CancerPPD	Database of experimentally verified anticancer peptides (ACPs) and proteins	3,491	http://crdd.osdd.net/raghava/cancerppd/	Tyagi et al., 2015
DRAMP	Database created with the objective of providing a useful resource for sequence- and structure-activity studies on AMPs	17,349	http://dramp.cpu-bioinfor.org	Fan et al., 2016
LAMP	Tool to aid the discovery and design of AMPs as new antimicrobial agents	5,547	http://biotechlab.fudan.edu.cn/database/lamp	Zhao et al., 2013
PeptideDB	Database of naturally occurring signalling peptides such as cytokines, growth factors, AMPs, peptide hormones etc.	20,027	http://www.peptides.be/	Liu et al., 2008
SATPdb	Database of structurally annotated therapeutic peptides with unique, experimentally validated sequences	19,192	http://crdd.osdd.net/raghava/satpdb/	Singh et al., 2016
THPdb	Database of FDA approved therapeutic peptides and proteins	852	http://crdd.osdd.net/raghava/thpdb/index.html	Usmani et al., 2017

high lipophilicity. More recently, a database dedicated to compiling structural information of bioactive peptides named StraPep (Wang et al., 2018), which currently displays structures for 3,791 peptides as well as provides detailed information for each one (i.e. experimental structure, secondary structure, post-translational modification, etc.).

MACHINE LEARNING

Machine learning is a natural outgrowth of the integration of computer science, mathematics and statistics that allows software application to become accurate in prediction without prior known information (Nasrabadi 2007). The basic application of machine learning is to build algorithms that can formulate a data (a matrix X_{ij}, where each row $\boldsymbol{x}$=(x_{i1}, x_{i2}, x_{i3},..., x_{ii}) is a sample composed of j features) with its proper form and use a prediction model to elucidate an output.

The application of machine learning for correlating the relationship that exists between structures of biological and chemical entities (i.e. peptides and proteins for the for-

mer while small molecules for the latter) with their observed or experimentally measured biological activity gives rise to an exciting field of research known as quantitative structure-activity relationship (QSAR). The formulation of a QSAR model entails the generation of quantitative and/or qualitative description of the biological or chemical entities (i.e. known as descriptors) and their subsequent correlation with the biological activity (e.g. IC_{50}, EC_{50}, % activity, etc.) through the use of machine learning algorithms.

Details on the best practices for the development of QSAR models is beyond the scope of this review and readers are directed to previous literature (Nantasenamat and Prachayasittikul, 2015; Tropsha, 2010; Shoombuatong et al., 2017a, b). Briefly, characteristics of a robust QSAR model is best summarized by the OECD principles (OECD, 2014) as outlined in Table 2. In a nutshell, it can be clear that a robust QSAR model should be properly prepared and curated, afford good performance as well as being interpretable so as to facilitate the utilization of the model for gaining insights into the underlying biological activity.

MODEL SET-UP FOR PREDICTING ANTICANCER PEPTIDES

Based on the prior knowledge of peptide sequence analysis, anticancer peptide prediction should be tackled in two associated ways: discriminating ACPs from non-ACPs and then predicting the anticancer activity of such ACPs. Due to the limitation of the experimental approach (e.g. slow and laborious process, expensive, difficulty in peptide purification etc.) for identifying the anticancer activity, computational tools for discriminating ACPs from non-ACPs is an essential way for saving the time-consuming and expensive cost.

Typically, the computational tool construction based on machine learning algorithm consists of four main elements, e.g. data collection, feature representation, model construction and model evaluation (Shoombuatong et al., 2012, 2015a, b, 2016, 2017a, b; Win et al., 2017; Pratiwi et al., 2017; Nantasenamat et al., 2015). In the point of view of machine learning, the use of reliable dataset plays a crucial role to obtain an efficient and generalized model. Previously, there have been many datasets that were used for developing various prediction models as shown in Table 3. Meanwhile, the remaining important elements are listed in Tables 4 and 5. In the following section, a comprehensive summary of previous works in this field are highlighted.

MACHINE LEARNING MODELS FOR THE PREDICTION OF ANTICANCER PEPTIDE

Previously, a variety of computational approaches, including AntiCP (Tyagi et al., 2013), Hajisharifi et al.,'s method (2014), ACPP (Vijayakumar and Ptv, 2015), iACP (Chen et al., 2016), Feng et al.,'s method (Li and Wang, 2016), iACP-GAEnsC (Akbar et al., 2017), Fazlullah et al.,'s method (Khan et al., 2017) and SAP (Xu et al., 2018), have been developed, which will be discussed in the following section. Almost all of the existing methods were developed by using support vector machine (SVM) cooperating with various types of peptide features, except for iACP-GAEnsC (Akbar et al., 2017) that was based on the ensemble approach. The overview of their datasets, type of features, machine learning algorithms and validation methods are shown in Table 5. Meanwhile, Table 6 lists the performance comparison among the existing methods as evaluated by 5-fold CV, 10-fold CV and jackknife test.

Tyagi et al. (2013) first addressed this problem by using SVM-based predictor named AntiCP, in which TY1, TY2 and TY3 datasets were implemented with AAC, DPC and binary profile. The research however, does not specifically state the type of kernel function used. SVM model with ACC/DPC yielded prediction accuracies of 85.52 %/ 85.29 % and 75.70 %/75.20 % respectively, evaluated

EXCLI Journal 2018;17:734-752 – ISSN 1611-2156
Received: June 20, 2018, accepted: July 10, 2018, published: July 25, 2018

Table 2: Summary of the OECD principles for the development of robust QSAR models.

#	OECD principle	Description
1	Defined endpoint	To ensure that the dataset is of high quality; particularly that all endpoint values are free from error
2	Unambiguous algorithm	To ensure the transparency and reproducibility of the QSAR model
3	Defined domain of applicability	To define the boundaries for which the QSAR model is capable of making predictions for query compounds such that they are not too structurally different than those used to train the model
4	Appropriate measures of goodness-of-fit, robustness and predictivity	To rigorously evaluate the performance of the QSAR model
5	Mechanistic interpretation	To ensure that the model can be mechanistically interpreted

Table 3: Summary of all datasets used in this research for evaluating anticancer peptide prediction.

Dataset	Sequence identity (%)[a]	No. of ACP	No. of Non-ACP	Total number	Reference
TY1	100	225	2,250	2,475	Tyagi et al., 2013
TY2	100	225	1,372	1,597	Tyagi et al., 2013
TY3	100	225	225	450	Tyagi et al., 2013
TY_IND	100	50	50	100	Tyagi et al., 2013
ZOH	90	138	206	344	Hajisharifi et al., 2014
SA_TRAIN	100	217	3,979	4,196	Vijayakumar and Ptv, 2015
SA_IND	100	40	40	80	Vijayakumar and Ptv, 2015
SA_RAND	100	-	2,000	2,000	Vijayakumar and Ptv, 2015
WC_IND	100	150	150	300	Chen et al., 2016
LEE	100	422	422	844	Manavalan et al., 2017

[a] Peptides having more than 90 % or 100 % pairwise sequence identity were removed from the dataset.

Table 4: Summary of all peptide features and their feature groups in this research.

Feature name	CS	ATC	PCP	PseCOM	SM
Amino acid composition (AAC)	✓				
Atomic composition (ATC)			✓		
Auto covariance of the average chemical shift (acACS)		✓			
Amphiphilic pseudo amino acid composition (Am-PseAAC)				✓	
Binary profile (BP)	✓				
Dipeptide composition (DPC)	✓				
G-Gap dipeptide composition (g-gap DPC)	✓				
Local alignment kernel					✓
Physicochemical properties (PCP)			✓		
Pseudo amino acid composition (PseACC)				✓	
Pseudo G-Gap dipeptide composition (Pse-g-gap DPC)				✓	
Protein relatedness measure (PRM)			✓		
Reduce amino acid composition (RACC)			✓		
Split amino acid composition (SAAC)	✓				

CS: Composition, ATC: Autocorrelation, PCP: Physicochemical properties, PseCOM: Pseudo Composition, SM: Similarity measure

Table 5: Summary of existing methods for predicting anticancer peptides.

Method	Classifier[a]	Sequence feature (No.)[b]	Testing method[c]	Web-server	Reference
AntiCP	SVM	AAC, DPC, BP (200)	10-fold CV and independent test	✓	Tyagi et al., 2013
Hajishar-ifi et al.,'s method	SVM	PseACC, LAK (>200)	5-fold CV and in-dependent test		Hajisharifi et al., 2014
ACPP	SVM, AdaBoost	PRM (60)	10-fold CV and independent test	✓[d]	Vijaya-kumar and Ptv, 2015
iACP	SVM	g-gap DPC (400)	5-fold CV, Jack-knife and inde-pendent test	✓	Chen et al., 2016
Li and Wang's method	SVM	AAC, RACC, acACS (80)	5-fold CV, Jack-knife		Li and Wang, 2016
Khan et al.,'s method	SVM, *k*-NN	SAAC, DPC, PseAAC (552)	Jackknife		Khan et al., 2017
iACP-GAEnsC	Ensemble, SVM, *k*-NN, PNN, RF, GRNN	Pse-g-gap DPC, Am-Pse-AAC, RACC (588)	Jackknife		Akbar et al., 2017
MLACP	SVM, RF	AAC, ATC, DPC, PCP (436)	10-fold CV	✓	Manavalan et al., 2017
SAP	SVM, RF, LibD3C	g-gap DPC (400)	5-fold CV		Xu et al., 2018

[a] *k*-NN: *k*-nearest Neighbor, GRNN: generalized neural network, LibD3C: hybrid model of ensemble pruning, PNN: probabilistic neural network, RF: random forest, SVM: support vector machine.
[b] AAC: amino acid composition, ATC: atomic composition , acACS: auto covariance of the average chemical shift, Am-PseAAC: amphiphilic pseudo amino acid composition, BP: binary profile, DPC: dipeptide composition, g-gap DPC: G-Gap dipeptide composition, LAK: local alignment kernel, PCP: Physicochemical properties, PseACC: Pseudo amino acid composition, Pse-g-gap DPC: Pseudo G-Gap dipeptide composition, PRM: protein relatedness measure, RACC: reduce amino acid composition, SAAC: split amino acid composition.
[c] 5-fold CV: 5-fold cross-validation, 10-fold CV: 10-fold cross-validation
[d] The webserver version is currently unavailable.

Table 6: Performance benchmark comparing various computational methods evaluated by 5- and 10-fold cross-validation and jackknife test.

Method	Testing method	Benchmarking dataset	Accuracy (%)	MCC	Reference
AntiCP	10-fold CV	TY3	91.44	0.83	Tyagi et al., 2013
Hajishaifi et al.,'s method	5-fold CV	ZOH	89.70	0.78	Hajisharifi et al., 2014
ACPP	10-fold CV	SA_TRAIN	97.70	0.92	Vijayakumar and Ptv 2015
iACP	5-fold CV	ZOH	94.77	0.89	Chen et al., 2016
Li and Wang's method	Jackknife	ZOH	93.61	0.88	Li and Wang 2016
Khan et al.,'s method	Jackknife	ZOH	93.31	0.86	Khan et al., 2017
iACP-GAEnsC	Jackknife	ZOH	96.45	0.91	Akbar et al., 2017
MLACP	10-fold CV	LEE	96.40	0.89	Manavalan et al., 2017
SAP	5-fold CV	ZOH	91.86	0.83	Xu et al., 2018

5-fold CV: 5-fold cross-validation, 10-fold CV: 10-fold cross-validation

by a 10-fold CV method on TY1 and TY2 datasets. These results revealed that the importance of ACC feature for enhancing ACP prediction was not quite different from DPC feature. But, when binary (NT10) based models were applied, where NT10 was the first 10 residues and each amino acid was represented by (20*10)-dimensional vector, the accuracy improved to 91.44 %. Finally, SVM based on the NT10 models performed well with 89 % accuracy and 0.78 MCC on TY_IND dataset. Finally, a web-server (Available at http://crdd.osdd.net/raghava/anticp/) was developed to help experimental scientists in predicting minimum mutations required for improving anticancer potency, virtual screening of peptides for discovering novel anticancer peptides and scanning natural proteins for identification of ACPs.

Hajisharifi et al., (2014) took advantage of PseAAC feature and local alignment kernels for improving the prediction performance of the model. In the study, the benchmark ZOH dataset was firstly created by collecting data from the antimicrobial peptide database (APD2) (Wang et al., 2009, available at http://aps.unmc.edu/AP/.) The ZOH dataset consisted of 192 ACPs and 215 non-ACPs and then, to prevent an overestimation of prediction results due to highly similar sequences, peptides with more than 90 % similarity were removed from the initial ZOH dataset using CD-HIT (Li and Godzik, 2006). Finally, a total of 138 ACPs and 206 non-ACPs were gained as summarized in Table 2. SVM model conjunction with PseACC feature showed the values of accuracy, sensitivity, specificity and MCC of 83.82 %, 81.84 %, 85.36 % and 0.66, respectively, evaluated by a 5-fold CV procedure. Meanwhile, using a local alignment kernel yielded better prediction results than PseACC feature with improvements of > 6 % and 10 % on both Ac and MCC, respectively.

Only one year later, Vijayakumar and Ptv (2015) utilized two powerful SVM and AdaBoost models cooperating with the protein relatedness measure (PRM) parameters called ACPP. The PRM feature represents each peptide with the degree distribution of amino acids deviating from a theoretical protein/peptide. To build a prediction model, SVM model with radial basis function (RBF) kernel and the tuning *cost* and *gamma* parameters of 2 and 0.0078, respectively, were used, while AdaBoost model based on the linear combination of simple weak classifiers with the tuning number of 10 iterations was applied. In this study, ACPP was evaluated with a 10-fold cross-validation method and independent test. SVM and AdaBoost were first carried out on the imbalanced dataset containing 217 ACPs and 3,979 non-ACPs as summarized in Table 2. The prediction results showed that SVM and AdaBoost yielded MCC values as low as 0.59 and 0.57, while, the balanced dataset (217 peptide sequences on both ACPs and non-ACPs), yielded increased accuracies for SVM and AdaBoost of 0.92 and 0.88, respectively. Based on these results, the authors stated that the PRM feature adopted to classify ACPs from non-ACPs was effective. Although, in this study, a web-server was established at http://acpp.bicpu.edu.in/predict.php, however, it is currently unavailable.

In 2016, there were two different research groups that made efforts to develop ACP predictors, i.e. iACP (Chen et al., 2016) and Feng et al.,'s method (Li and Wang, 2016). Chen et al. (2016) proposed an approach to take advantage of SVM model in conjunction with g-gap dipeptide compositions (g-gap DPC), where g = 0, 1, 2, 3 or 4 and g =0 is DPC, as well as working together with ANOVA (analysis of variance). Herein, SVM model with radial basis function (RBF) kernel and their optimal parameter of *cost* = 2 and *gamma* = 0.125 were used. The ANOVA approach via the incremental feature selection (IFS) was used for selecting informative features among g-gap DPCs. The process of determining the optimal number of features was conducted according to the following steps: (1) the feature with the highest F-score was selected as the input of SVM and the prediction performance assessed with 5-fold CV was calculated to

evaluate the performance of this feature; (2) the feature with the second highest F-score was then combined with the first feature to form a new feature subset and the prediction performance with the criteria was still used to estimate the performance of the new feature subset; (3) this process was done when the prediction performance of 400 features were calculated. The highest accuracy of 94.77 % can be achieved by using g=1 and the 126 top-ranked informative features. Li and Wang (2016) attempted to improve the prediction performance by using SVM model with hybrid composition, i.e. AAC, auto covariance of the average chemical shift (acACS) and reduced amino acid composition (RAAC). The parameters of RBF kernel used were tuned using the grid search method. Initial prediction results for their model using AAC on the ZOH dataset showed the value of accuracy and MCC of 91.86 % and 0.83, respectively. The second and third highest accuracies were obtained from using RACC (84.01 %) and asACS (82.56 %), respectively. Meanwhile, the combination features of AAC, RAAC and acACS performed best with 93.61 % accuracy and 0.87 MCC. The authors of this paper suggested that these combination features were helpful to the prediction of ACPs.

In 2017, three different ACP predictors were developed with various types of machine learning algorithms and peptide features, i.e. Khan et al.,'s method (2017), iACP-GAEnsC (Akbar et al., 2017) and MLACP (Manavalan et al., 2017). Khan et al. (2017) utilized SVM and *k*-nearest Neighbor (*k*-NN) models with a variety of peptide features, i.e. split amino acid composition (SAAC), DPC and PseAAC, to find the suitable feature for discriminating ACPs from non-ACPs. The total number of feature spaces of SAAC, DPC and PseAAC were 400, 62 and 60, respectively. To build prediction models, authors used RBF kernel to create SVM model, while euclidian distance was used to compute the distance among the peptide sequences. The optimum parameters of these two models were obtained during the training phase. The performance comparison evaluated by jackknife test demonstrated that SVM and *k*-NN models using SAAC outperformed the other two features with an accuracy of 93.31 % and 90.17 %. Akbar et al. (2017) examined the ability of a variety of machine learning algorithms, i.e. SVM, random forest (RF), *k*-nearest Neighbor (*k*-NN), generalized neural network (GRNN), and probabilistic neural network (PNN). In this study, each peptide was represented by three different feature extraction schemes using RAAP, Pse-g-Gap dipeptide composition (Pse-g-gap DPC) and amphiphilic PseAAC (Am-PseACC). Finally, the evolution genetic algorithm was used to measure the diversity and optimum outcome or prediction results of the different methods called iACP-GAEnsC. Initial prediction results showed that using Am-PseACC with jackknife test achieved accuracies of 93.60 %, 90.41 %, 91.28 %, 86.33 % and 93.89 % for SVM, *k*-NN, PNN, RF, GRNN and GAEnsC, respectively. Their best accuracy of 94.45 % was achieved by using an ensemble approach with the merging of SVM, *k*-NN, PNN, RF and GRNN associated with a hybrid feature of RAAP, Pse-g-gap DPC and Am-PseACC. Manavalan et al. (2017) developed machine learning-based methods (SVM and RF), named SVMACP, RFACP and MLACP using a combination of features, including ACC, DPC, PCP and ATC. The number of dimensions for ACC, DPC, PCP and ATC features were 20, 400, 11, 5, respectively. For each model, authors optimized the RF (*ntree* and *mtry*) and SVM (*cost* and *gamma*) parameters by using 10-fold CV on the TY3 dataset. In the case of using a single feature, RFACP and SVMACP yielded accuracies ranging from 81.4 %-86.8 % and 75.9 %-85.8 %, respectively. The best accuracy and MCC of 87.2 and 0.70, respectively, was achieved by using RF model with the combination feature of ACC, DPC, PCP and ATC.

Recently, Xu et al. (2018) developed the MRMD method to select important features from g-gap DPC. The selected, informative feature was used as an input feature to train the

the SVM model called SAP. The paper does not specifically state the type of kernel function used. For a 5-fold CV, SAP using all 400 features yielded 91.86 % accuracy and 0.83 MCC, while using selected features offered a 90.70 % accuracy and 0.81 MCC. Furthermore, SAP was also compared with RF and LibD3D, where LibD3D is a selective ensemble model. The overall accuracy comparison showed that SAP (91.78 %) was quite comparable with RF (91.88 %) and LibD3D (89.24 %) models.

The aforementioned articles showed promising results in the use of various types of machine learning algorithms and peptides features as summarized in Tables 5 and 6. As seen in Table 3, the ZOH is known as the valid benchmark dataset used for developing various prediction models (Hajisharifi et al., 2014; Chen et al., 2016; Xu et al., 2018; Khan et al., 2017; Akbar et al., 2017). Amongst these methods, iACP (Chen et al., 2016) and iACP-GAEnsC (Akbar et al., 2017) showed their best predictive accuracies of 94.77 % and 96.45 % as evaluated by 5-fold CV and jackknife test procedures, respectively. In addition, iACP revealed its efficiency by carrying out an independent WC_IND data achieving an accuracy and MCC of 92.67 % and 0.85, respectively. Considering that the independent test is the most rigorous cross-validation method, it might be stated that iACP (Chen et al., 2016) was superior to other prediction methods as demonstrated in Table 5. Amongst the existing methods, some of them (Tyagi et al., 2013; Chen et al., 2016; Manavalan et al., 2017) determined the important amino acids and dipeptide that were enriched in anticancer peptides using componential analysis (Manavalan et al., 2017; Tyagi et al., 2013) and F-score (Chen et al., 2016).

BIOLOGICAL INSIGHTS FROM PREDICTIVE MODELS

Feature importance analysis from existing models (Tyagi et al., 2013; Chen et al., 2016; Manavalan et al., 2017 indicated that in general anticancer peptides are abundant in Cys, Glu, Phe, Gly, lle, Lys and Phe when compared to non-anticancer peptides (Chen et al., 2016). Particularly, Tyagi et al. (2013) reported that Gly, Leu, Ala and Phe were preferential residues at the N-terminus of anticancer peptides while Val, Cys, Leu and Lys were likely to be found at the C-terminus. Furthermore, Manavalan et al. (2017) revealed that the 10 top-ranking features in anticancer peptides were comprised of dipeptides rich in positively charged and aromatic residues (e.g. KK, AK, KL, AL, KA, KW, LA, LK, FA and LF). Moreover, it should also be noted that desirable trait for anticancer peptides is their cell penetrating ability such that they can specifically neutralize their target while maintaining low toxicity.

LIMITATIONS OF CURRENT MACHINE LEARNING MODELS

The use of machine learning algorithm is one of the important factors in the steady growth of the field of anticancer drug discovery and development. Most of the reported anticancer peptide prediction methods were mainly developed in order to enhance the prediction accuracy by taking advantage of the complexity of prediction methods and the number of feature types. Overall, most research articles showed encouraging results with having satisfied accuracies of more than 90 %. Nevertheless, there is still room for development to improve the existing methods as useful and interpretable models for facilitating experimental scientists and related researchers as demonstrated by a series of recent publications (Shoombuatong et al., 2012, 2015a, b, 2016, 2017a, b; Win et al., 2017; Pratiwi et al., 2017; Nantasenamat et al., 2015) and summarized in comprehensive reviews (Nantasenamat et al., 2015; Shoombuatong et al., 2017a, b).

In addition, the most commonly used benchmark dataset ZOH, (Hajisharifi et al., 2014) consisted of 138 ACPs and 206 non-ACPs in which only ACPs were derived from the experimental verification method. It could be stated that existing methods developed by

ZOH dataset might not be completely suited to accurately filter experimentally verified non-ACPs from ACPs. Furthermore, peptide features were intrinsically heterogeneous, noisy and multi-dimensional, but only a few existing methods (Chen et al., 2016; Xu et al., 2018) took advantage of feature selection techniques to qualify and rank the importance and the contributions of the features for the model performance. Thus, these method has utilized only partial information of the biological activity of ACP. It could be stated that the role of different types of peptide features contributing to the biological activity of anticancer peptide are still poorly understood. Additionally, a variety of methods were used to evaluate the prediction performance of ACP predictions as listed in Table 5, including N-fold cross-validation, where N is 5 or 10, jackknife test and independent test. The independent test is an effective way to test the performance of a model in real-world applications and verify the generalization of a model, but only few existing methods (Tyagi et al., 2013; Vijayakumar and Ptv, 2015; Chen et al., 2016; Li and Wang, 2016) were assessed with this method. Finally, according to the fifth principle of OEC which states that, it is necessary and significant of an interpretable QSAR model to provide important factors that can enhance the biological activity of peptides or compounds. Amongst the existing methods, some of them (Tyagi et al., 2013; Chen et al., 2016; Manavalan et al., 2017) provided the results of feature importance analysis by using componential analysis (Manavalan et al., 2017; Tyagi et al., 2013) and F-score (Chen et al., 2016). However, they did not clearly mention which features contributed most to prediction performance. Moreover, the SVM model was not straight-forward enough to interpret the underlying biological implications of anticancer peptides.

CONCLUSION

The success story of therapeutic peptides is starting to gain moment with more than 60 approved by the FDA and more than 150 peptides have reached pre-clinical and clinical stages. In addition, a literature review have indicated that there is a large volume of on-going studies being carried out in the field. In spite of the large sum of papers on the utilization of machine learning approaches for the development of QSAR models of bioactive and therapeutic peptides, however there are few review articles that examine the field in a systematic manner. It is the intent of this review article to fill this gap by providing readers with the current advancements pertaining to the current state-of-the-art on the prediction of anticancer peptides via the use of machine learning approaches.

A survey of existing QSAR models against anticancer peptides suggested that almost all provided reasonably high prediction accuracies and in spite of this, there are limiting factors that may hinder their full potential for application as follows:

(i) Absence of experimentally verified non-anticancer peptides.
(ii) Inclusion of trivial and non-informative features during the model building process.
(iii) Lack of comprehensive evaluation method and failure to make use of interpretable learning methods.

In efforts to augment the robustness of the predictive model, herein are recommendations:

(i) Increase the size of the peptide dataset by combining all data sources together as to capture as much as possible of the pattern of dataset for alleviating uncertainties in the prediction system.
(ii) Familiarize oneself with the background and details of the descriptors being used such that the resulting models could be interpreted in a meaningful manner as to gain biological insights for guiding further experiments.
(iii) Use interpretable learning algorithms as to allow the interpretation of important features responsible for the biological activity.

(iv) Ensure that the model is externally validated on an independent test set as well as defining the applicability domain of the model as to verify the ability of the model for extrapolation to future unknown data.
(v) Ensure the reproducibility of constructed models such that interested users could extend the model further.
(vi) If possible, constructed models should be made publicly available in the form of webservers so as to facilitate easy access to the model's prediction capability.

Conflict of interests

The authors declare that no competing interests exist.

Acknowledgements

This work is supported by the Office of Higher Education Commission and the Thailand Research Fund (No. MRG6180226); the New Researcher Grant (A31/2561) from Mahidol University; and the Center of Excellence on Medical Biotechnology (CEMB), S&T Postgraduate Education and Research Development Office (PERDO), Office of Higher Education Commission (OHEC), Thailand. Partial support from the annual budget grant (B.E. 2557-2559) of Mahidol University is also acknowledged.

REFERENCES

Akbar S, Hayat M, Iqbal M, Jan MA. iACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space. Artif Intell Med. 2017;79:62–70.

Alam S, Khan F. QSAR and docking studies on xanthone derivatives for anticancer activity targeting DNA topoisomerase IIα. Drug Des Devel Ther. 2014;8:183–95.

Al-Benna S, Shai Y, Jacobsen F, Steinstraesser L. Oncolytic activities of host defense peptides. Int J Mol Sci. 2011;12:8027–51.

Arnold M, Karim-Kos HE, Coebergh JW, Byrnes G, Antilla A, Ferlay J, et al. Recent trends in incidence of five common cancers in 26 European countries since 1988: Analysis of the European Cancer Observatory. Eur J Cancer. 2015;51:1164–87.

Berge G, Eliassen LT, Camilio KA, Bartnes K, Sveinbjørnsson B, Rekdal O. Therapeutic vaccination against a murine lymphoma by intratumoral injection of a cationic anticancer peptide. Cancer Immunol Immunother. 2010;59:1285–94.

Braunstein A, Papo N, Shai Y. *In vitro* activity and potency of an intravenously injected antimicrobial peptide and its DL amino acid analog in mice infected with bacteria. Antimicrob Agents Chemother. 2004;48:3127–9.

Camilio KA, Berge G, Ravuri CS, Rekdal O, Sveinbjørnsson B. Complete regression and systemic protective immune responses obtained in B16 melanomas after treatment with LTX-315. Cancer Immunol Immunother. 2014;63:601–13.

Chen W, Ding H, Feng P, Lin H, Chou K-C. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget. 2016;7:16895–909.

Chen Y, Xu X, Hong S, Chen J, Liu N, Underhill CB, et al. RGD-Tachyplesin inhibits tumor growth. Cancer Res. 2001;61:2434–8.

Craik DJ, Fairlie DP, Liras S, Price D. The future of peptide-based drugs. Chem Biol Drug Des. 2013;81:136–47.

Dennison SR, Harris F, Phoenix DA. The interactions of aurein 1.2 with cancer cell membranes. Biophys Chem. 2007;127:78–83.

Deslouches B, Gonzalez IA, DeAlmeida D, Islam K, Steele C, Montelaro RC, et al. De novo-derived cationic antimicrobial peptide activity in a murine model of Pseudomonas aeruginosa bacteraemia. J Antimicrob Chemother. 2007;60:669–72.

Domalaon R, Findlay B, Ogunsina M, Arthur G, Schweizer F. Ultrashort cationic lipopeptides and lipopeptoids: Evaluation and mechanistic insights against epithelial cancer cells. Peptides. 2016;84:58–67.

Du Q-S, Huang R-B, Chou K-C. Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design. Curr Protein Pept Sci. 2008;9:248–60.

Fan L, Sun J, Zhou M, Zhou J, Lao X, Zheng H, et al. DRAMP: a comprehensive data repository of antimicrobial peptides. Sci Rep. 2016;6:24482.

Felício MR, Silva ON, Gonçalves S, Santos NC, Franco OL. Peptides with dual antimicrobial and anticancer activities. Front Chem. 2017;5:5.

Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136:E359-86.

Figueiredo CR, Matsuo AL, Massaoka MH, Polonelli L, Travassos LR. Anti-tumor activities of peptides corresponding to conserved complementary determining regions from different immunoglobulins. Peptides. 2014;59:14–9.

Fosgerau K, Hoffmann T. Peptide therapeutics: current status and future directions. Drug Discov Today. 2015; 20:122–8.

Garay RP, El-Gewely R, Armstrong JK, Garratty G, Richette P. Antibodies against polyethylene glycol in healthy subjects and in patients treated with PEG-conjugated agents. Expert Opin Drug Deliv. 2012;9:1319–23.

Gaspar D, Freire JM, Pacheco TR, Barata JT, Castanho MARB. Apoptotic human neutrophil peptide-1 anti-tumor activity revealed by cellular biomechanics. Biochim Biophys Acta. 2015;1853:308–16.

Gautam A, Singh H, Tyagi A, Chaudhary K, Kumar R, Kapoor P, et al. CPPsite: a curated database of cell penetrating peptides. Database. 2012;2012: bas015.

Giacometti A, Cirioni O, Riva A, Kamysz W, Silvestri C, Nadolski P, et al. *In vitro* activity of aurein 1.2 alone and in combination with antibiotics against gram-positive nosocomial cocci. Antimicrob Agents Chemother. 2007;51:1494–6.

Hajisharifi Z, Piryaiee M, Mohammad Beigi M, Behbahani M, Mohabatkar H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. J Theor Biol. 2014;341:34–40.

Hao X, Yan Q, Zhao J, Wang W, Huang Y, Chen Y. TAT modification of alpha-helical anticancer peptides to improve specificity and efficacy. PLoS ONE. 2015; 10:e0138911.

Hoskin DW, Ramamoorthy A. Studies on anticancer activities of antimicrobial peptides. Biochim Biophys Acta. 2008;1778:357–75.

Hu C, Chen X, Zhao W, Chen Y, Huang Y. Design and modification of anticancer peptides. Drug Des. 2016;5: 138.

Iwasaki T, Ishibashi J, Tanaka H, Sato M, Asaoka A, Taylor D, et al. Selective cancer cell cytotoxicity of enantiomeric 9-mer peptides derived from beetle defensins depends on negatively charged phosphatidylserine on the cell surface. Peptides. 2009;30:660–8.

Jayaraj V, Suhanya R, Vijayasarathy M, Anandagopu P, Rajasekaran E. Role of large hydrophobic residues in proteins. Bioinformation. 2009;3:409–12.

Jin G, Weinberg A. Human antimicrobial peptides and cancer. Semin Cell Dev Biol. 2018; In Press. DOI: 10.1016/j.semcdb.2018.04.006.

Kapoor P, Singh H, Gautam A, Chaudhary K, Kumar R, Raghava GPS. TumorHoPe: a database of tumor homing peptides. PLoS ONE. 2012;7:e35187.

Karpiński TM, Adamczak A. Anticancer activity of bacterial proteins and peptides. Pharmaceutics. 2018; 10(2):54.

Kelly GJ, Kia AF-A, Hassan F, O'Grady S, Morgan MP, Creaven BS, et al. Polymeric prodrug combination to exploit the therapeutic potential of antimicrobial peptides against cancer cells. Org Biomol Chem. 2016; 14:9278–86.

Khan F, Akbar S, Basit A, Khan I, Akhlaq H. Identification of anticancer peptides using optimal feature space of chou's split amino acid composition and support vector machine. In: Proceedings of the 2017 4th International Conference on Biomedical and Bioinformatics Engineering - ICBBE 2017;2017:91–6.

Kliger Y. Computational approaches to therapeutic peptide discovery. Biopolymers. 2010;94:701–10.

Lau JL, Dunn MK. Therapeutic peptides: Historical perspectives, current development trends, and future directions. Bioorg Med Chem. 2018;26:2700–7.

Lee EY, Fulan BM, Wong GCL, Ferguson AL. Mapping membrane activity in undiscovered peptide sequence space using machine learning. Proc Natl Acad Sci USA. 2016;113:13588–93.

Lee EY, Lee MW, Fulan BM, Ferguson AL, Wong GCL. What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning? Interface Focus. 2017;7:20160153.

Lee EY, Wong GCL, Ferguson AL. Machine learning-enabled discovery and design of membrane-active peptides. Bioorg Med Chem. 2018;26: 2708-18.

Li F-M, Wang X-Q. Identifying anticancer peptides by using improved hybrid compositions. Sci Rep. 2016;6: 33910.

Li H, Anuwongcharoen N, Malik AA, Prachayasittikul V, Wikberg JES, Nantasenamat C. Roles of D-amino acids on the bioactivity of host defense peptides. Int J Mol Sci. 2016;17(7):1023.

Li H, Schaduangrat N, Simeon S, Nantasenamat C. Computational study on the origin of the cancer immunotherapeutic potential of B and T cell epitope peptides. Mol Biosyst. 2017;13:2310–22.

Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.

Li Y, Xiang Q, Zhang Q, Huang Y, Su Z. Overview on the recent study of antimicrobial peptides: origins, functions, relative mechanisms and application. Peptides. 2012;37:207–15.

Li ZJ, Cho CH. Peptides as targeting probes against tumor vasculature for diagnosis and drug delivery. J Transl Med. 2012;10(Suppl 1):S1.

Liu F, Baggerman G, Schoofs L, Wets G. The construction of a bioactive peptide database in Metazoa. J Proteome Res. 2008;7:4119–31.

Maccari G, Di Luca M, Nifosì R. In silico design of antimicrobial peptides. Methods Mol Biol. 2015;1268: 195–219.

Makobongo MO, Gancz H, Carpenter BM, McDaniel DP, Merrell DS. The oligo-acyl lysyl antimicrobial peptide C_{12}K-2β_{12} exhibits a dual mechanism of action and demonstrates strong *in vivo* efficacy against *Helicobacter pylori*. Antimicrob Agents Chemother. 2012; 56:378–90.

Makovitzki A, Fink A, Shai Y. Suppression of human solid tumor growth in mice by intratumor and systemic inoculation of histidine-rich and pH-dependent host defense-like lytic peptides. Cancer Res. 2009;69: 3458–63.

Manavalan B, Basith S, Shin TH, Choi S, Kim MO, Lee G. MLACP: machine-learning-based prediction of anticancer peptides. Oncotarget. 2017;8:77121–36.

Mansour SC, Pena OM, Hancock REW. Host defense peptides: front-line immunomodulators. Trends Immunol. 2014;35:443–50.

McKeown STW, Lundy FT, Nelson J, Lockhart D, Irwin CR, Cowan CG, et al. The cytotoxic effects of human neutrophil peptide-1 (HNP1 and lactoferrin on oral squamous cell carcinoma (OSCC *in vitro*. Oral Oncol. 2006;42:685–90.

Melo MN, Ferre R, Feliu L, Bardají E, Planas M, Castanho MARB. Prediction of antibacterial activity from physicochemical properties of antimicrobial peptides. PLoS ONE. 2011;6:e28549.

Minkiewicz P, Dziuba J, Iwaniak A, Dziuba M, Darewicz M. BIOPEP database and other programs for processing bioactive peptide sequences. J AOAC Int. 2008;91:965–80.

Mohseni Bababdani B, Mousavi M. Gravitational search algorithm: A new feature selection method for QSAR study of anticancer potency of imidazo[4,5-b]pyridine derivatives. Chemometr Intell Lab Syst. 2013;122:1–11.

Nagarajan V, Kaushik N, Murali B, Zhang C, Lakhera S, Elasri MO, et al. A Fourier transformation based method to mine peptide space for antimicrobial activity. BMC Bioinformatics. 2006;7(Suppl 2):S2.

Nantasenamat C, Prachayasittikul V. Maximizing computational tools for successful drug discovery. Expert Opin Drug Discov. 2015;10:321–9.

Nantasenamat C, Worachartcheewan A, Jamsak S, Preeyanon L, Shoombuatong W, Simeon S, et al. AutoWeka: toward an automated data mining software for QSAR and QSPR studies. Methods Mol Biol. 2015; 1260:119–47.

Narayana JL, Huang H-N, Wu C-J, Chen J-Y. Efficacy of the antimicrobial peptide TP4 against *Helicobacter pylori* infection: *in vitro* membrane perturbation via micellization and *in vivo* suppression of host immune responses in a mouse model. Oncotarget. 2015;6: 12936–54.

Nasrabadi NM. Pattern recognition and machine learning. J Electron Imaging. 2007;16:049901.

Nongonierma AB, FitzGerald RJ. Learnings from quantitative structure–activity relationship (QSAR studies with respect to food protein-derived bioactive peptides: a review. RSC Adv. 2016;6:75400–13.

O'Brien-Simpson NM, Hoffmann R, Chia CSB, Wade JD. Antimicrobial and anticancer peptides (Editorial). Front Chem. 2018;6:13.

OECD (Organization For Economic Co-operation and Development). Guidance Document on the Validation of (Quantitative Structure-Activity Relationship [(QSAR] Models. Paris: OECD, 2014 (OECD Series on Testing and Assessment, No. 69).

Papo N, Shai Y. New lytic peptides based on the D,L-amphipathic helix motif preferentially kill tumor cells compared to normal cells. Biochemistry. 2003;42: 9346–54.

Papo N, Seger D, Makovitzki A, Kalchenko V, Eshhar Z, Degani H, et al. Inhibition of tumor growth and elimination of multiple metastases in human prostate and breast xenografts by systemic inoculation of a host defense-like lytic peptide. Cancer Res. 2006;66:5371–8.

Pasupuleti M, Schmidtchen A, Malmsten M. Antimicrobial peptides: key components of the innate immune system. Crit Rev Biotechnol. 2012;32:143–71.

Penchala SC, Miller MR, Pal A, Dong J, Madadi NR, Xie J, et al. A biomimetic approach for enhancing the *in vivo* half-life of peptides. Nat Chem Biol. 2015;11:793–8.

Podust VN, Sim B-C, Kothari D, Henthorn L, Gu C, Wang C, et al. Extension of *in vivo* half-life of biologically active peptides via chemical conjugation to XTEN protein polymer. Protein Eng Des Sel. 2013;26:743–53.

Prada-Gracia D, Huerta-Yépez S, Moreno-Vargas LM. Application of computational methods for anticancer drug discovery, design, and optimization. Bol Méd Hosp Infant México. 2016;73:411–23.

Pratiwi R, Malik AA, Schaduangrat N, Prachayasittikul V, Wikberg JES, Nantasenamat C, et al. CryoProtect: a web server for classifying antifreeze proteins from nonantifreeze proteins. J Chem. 2017;2017:9861752.

Reddy KVR, Yedery RD, Aranha C. Antimicrobial peptides: premises and promises. Int J Antimicrob Agents. 2004;24:536–47.

Riedl S, Zweytick D, Lohner K. Membrane-active host defense peptides--challenges and perspectives for the development of novel anticancer drugs. Chem Phys Lipids. 2011;164:766–81.

Rozek T, Wegener KL, Bowie JH, Olver IN, Carver JA, Wallace JC, et al. The antibiotic and anticancer active aurein peptides from the Australian bell frogs Litoria aurea and Litoria raniformis. Eur J Biochem. 2000;267:5330–41.

Schellenberger V, Wang C-W, Geething NC, Spink BJ, Campbell A, To W, et al. A recombinant polypeptide extends the *in vivo* half-life of peptides and proteins in a tunable manner. Nat Biotechnol. 2009;27:1186–90.

Shi LM, Fan Y, Myers TG, O'Connor PM, Paull KD, Friend SH, et al. Mining the NCI anticancer drug discovery databases: genetic function approximation for the QSAR study of anticancer ellipticine analogues. J Chem Inf Comput Sci. 1998;38:189–99.

Shoombuatong W, Hongjaisee S, Barin F, Chaijaruwanich J, Samleerat T. HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees. Comput Biol Med. 2012;42:885–9.

Shoombuatong W, Prachayasittikul V, Anuwongcharoen N, Songtawee N, Monnor T, Prachayasittikul S, et al. Navigating the chemical space of dipeptidyl peptidase-4 inhibitors. Drug Des Devel Ther. 2015a;9:4515–49.

Shoombuatong W, Prachayasittikul V, Prachayasittikul V, Nantasenamat C. Prediction of aromatase inhibitory activity using the efficient linear method (ELM). EXCLI J. 2015b;14:452–64.

Shoombuatong W, Nabu S, Simeon S, Prachayasittikul V, Lapins M, Wikberg JES, et al. Extending proteochemometric modeling for unraveling the sorption behavior of compound–soil interaction. Chemometr Intell Lab Syst. 2016;151:219–27.

Shoombuatong W, Prathipati P, Owasirikul W, Worachartcheewan A, Simeon S, Anuwongcharoen N, et al. Towards the revival of interpretable QSAR models. In: Kunal R (ed): Advances in QSAR modeling (pp 3-55). Basel: Springer International Publ., 2017a. (Challenges and Advances in Computational Chemistry and Physics, Vol. 24).

Shoombuatong W, Prathipati P, Prachayasittikul V, Schaduangrat N, Malik AA, Pratiwi R, et al. Towards predicting the cytochrome P450 modulation: from QSAR to proteochemometric modeling. Curr Drug Metab. 2017b;18:540–55.

Simeon S, Li H, Win TS, Malik AA, Kandhro AH, Piacham T, et al. PepBio: predicting the bioactivity of host defense peptides. RSC Adv. 2017;7:35119–34.

Singh S, Chaudhary K, Dhanda SK, Bhalla S, Usmani SS, Gautam A, et al. SATPdb: a database of structurally annotated therapeutic peptides. Nucleic Acids Res. 2016;44:1119-26.

Spinks CB, Zidan AS, Khan MA, Habib MJ, Faustino PJ. Pharmaceutical characterization of novel tenofovir liposomal formulations for enhanced oral drug delivery: *in vitro* pharmaceutics and Caco-2 permeability investigations. Clin Pharmacol. 2017;9:29–38.

Steiner H, Hultmark D, Engström A, Bennich H, Boman HG. Sequence and specificity of two antibacterial proteins involved in insect immunity. Nature. 1981;292:246–8.

Thundimadathil J. Cancer treatment using peptides: current therapies and future prospects. J Amino Acids. 2012;2012:967347.

Tong J, Zhao X, Zhong L. QSAR studies of imidazo[4,5-b]pyridine derivatives as anticancer drugs using RASMS method. Med Chem Res. 2014;23:4883–92.

Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform. 2010; 29:476–88.

Tyagi A, Kapoor P, Kumar R, Chaudhary K, Gautam A, Raghava GPS. In silico models for designing and discovering novel anticancer peptides. Sci Rep. 2013; 3:2984.

Tyagi A, Tuknait A, Anand P, Gupta S, Sharma M, Mathur D, et al. CancerPPD: a database of anticancer peptides and proteins. Nucleic Acids Res. 2015;43: D837-43.

Tørfoss V, Ausbacher D, Cavalcanti-Jacobsen C de A, Hansen T, Brandsdal B-O, Havelkova M, et al. Synthesis of anticancer heptapeptides containing a unique lipophilic $\beta^{2,2}$-amino acid building block. J Pept Sci. 2012;18:170–6.

Usmani SS, Bedi G, Samuel JS, Singh S, Kalra S, Kumar P, et al. THPdb: Database of FDA-approved peptide and protein therapeutics. PLoS ONE. 2017;12: e0181748.

Vedham V, Divi RL, Starks VL, Verma M. Multiple infections and cancer: implications in epidemiology. Technol. Cancer Res Treat. 2014;13:177–94.

Vijayakumar S, Ptv L. ACPP: A web server for prediction and design of anti-cancer peptides. Int J Pept Res Ther. 2015;21:99–106.

Vlieghe P, Lisowski V, Martinez J, Khrestchatisky M. Synthetic therapeutic peptides: science and market. Drug Discov Today. 2010;15:40–56.

Waghu FH, Barai RS, Gurung P, Idicula-Thomas S. CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. Nucleic Acids Res. 2016;44:D1094-7.

Wang G, Li X, Wang Z. APD2: the updated antimicrobial peptide database and its application in peptide design. Nucleic Acids Res. 2009;37:933-7.

Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education. Nucleic Acids Res. 2016;44:1087-93.

Wang J, Yin T, Xiao X, He D, Xue Z, Jiang X, et al. StraPep: a structure database of bioactive peptides. Database. 2018;2018:bay038.

WHO. Cancer. 2018a. Available from: http://www.who.int/news-room/fact-sheets/detail/cancer.

WHO. Antimicrobial resistance. 2018b. Available from: http://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance.

Win TS, Malik AA, Prachayasittikul V, Wikberg JES, Nantasenamat C, Shoombuatong W. HemoPred: a web server for predicting the hemolytic activity of peptides. Future Med Chem. 2017;9:275–91.

Xu L, Liang G, Wang L, Liao C. A novel hybrid sequence-based model for identifying anticancer peptides. Genes. 2018;9(3).

Zasloff M. Magainins, a class of antimicrobial peptides from Xenopus skin: isolation, characterization of two active forms, and partial cDNA sequence of a precursor. Proc Natl Acad Sci USA. 1987;84:5449–53.

Zhao X, Wu H, Lu H, Li G, Huang Q. LAMP: A database linking antimicrobial peptides. PLoS ONE. 2013; 8:e66557.

Zhou P, Tian F, Wu Y, Li Z, Shang Z. Quantitative Sequence-Activity Model (QSAM): applying QSAR strategy to model and predict bioactivity and function of peptides, proteins and nucleic acids. Curr Comput Aid Drug Des. 2008;4:311–21.

Computational Biology and Chemistry 80 (2019) 419–432

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac

Research Article

HIVCoR: A sequence-based tool for predicting HIV-1 CRF01_AE coreceptor usage

Sayamon Hongjaisee[a,b], Chanin Nantasenamat[c], Tanawan Samleerat Carraway[b,*], Watshara Shoombuatong[c,*]

[a] Research Institute for Health Sciences, Chiang Mai University, Chiangmai 50200, Thailand
[b] Faculty of Associated Medical Sciences, Chiang Mai University, Chiangmai 50200, Thailand
[c] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand

ARTICLE INFO

Keywords:
Coreceptor usage
CRF01_AE
Genotypic assays
Machine learning
Random forest
Support vector machine

ABSTRACT

Determination of HIV-1 coreceptor usage is strongly recommended before starting the coreceptor-specific inhibitors for HIV treatment. Currently, the genotypic assays are the most interesting tools due to they are more feasible than phenotypic assays. However, most of prediction models were developed and validated by data set of HIV-1 subtype B and C. The present study aims to develop a powerful and reliable model to accurately predict HIV-1 coreceptor usage for CRF01_AE subtype called HIVCoR. HIVCoR utilized random forest and support vector machine as the prediction model, together with amino acid compositions, pseudo amino acid compositions and relative synonymous codon usage frequencies as the input feature. The overall success rate of 93.79% was achieved from the external validation test on the objective benchmark dataset. Comparison results indicated that HIVCoR was superior to other bioinformatics tools and genotypic predictors. For the convenience of experimental scientists, a user-friendly webserver has been established at http://codes.bio/hivcor/.

1. Introduction

The first step of Human Immunodeficiency Virus (HIV) infection is the binding of envelope glycoprotein or gp120 with primary cellular receptor on the host cell, which is CD4. Not only CD4, HIV needs another molecule to enter the host cells, these molecules are called as coreceptors. The main coreceptors used by HIV-1 are CCR5 and CXCR4. Based on coreceptor usage, HIV-1 variants that use only CCR5 as a coreceptor to enter a cell are termed R5 viruses, those that only use CXCR4 as a coreceptor are termed X4 viruses, and those that use both CCR5 and CXCR4 as a coreceptor are termed dual/mixed viruses (Berger et al., 1998). In early infection, R5 viruses are preferentially transmitted and generally predominated, whereas X4 viruses commonly emerge in late stage of disease and have been associated with a subsequent increase in the rate of CD4 decline, an accelerated disease progression to AIDS, and a poor prognosis for survival (Koot et al., 1993; Richman and Bozzette, 1994; Zhu et al., 1993). Since coreceptor antagonists have been launched as an antiretroviral drug for treatment of HIV-1 infection, it has increased interest in determination of HIV-1 coreceptor usage which is very strongly recommended before starting this drug regimen.

Currently, the determination of HIV-1 coreceptor usage can be performed by various phenotypic and genotypic assays. The determination of HIV-1 coreceptor usage using phenotypic assay is based on the direct observation of viral coreceptor usage by construction of recombinant/pseudotyped viruses containing the HIV envelope fragment isolated from the patients. These viruses were inoculated into cell lines containing the specific set of coreceptors on their surfaces, the viruses were classified as R5, X4, or dual/mixed viruses depending on their tropisms. In case of genotypic assay, it based on the direct sequencing of V3 region, which is a major domain for coreceptor binding (Deng et al., 1996) and this V3 amino acid sequence is used as a template to predict the coreceptor usage by many genotypic predictors. Each genotypic predictor was developed by the use of various types of protein features on V3 amino acid sequence and machine learning algorithms. Although the phenotypic assay can provide the exact details HIV-1 coreceptor usage, the experimental techniques is time-consuming and cost-expensive. Thus, it is highly desirable to develop genotypic predictors based on machine learning algorithms. In fact, several computational methods have been proposed to discriminate X4 from R5 as summarized in Table 1. Many research groups have made efforts to develop their genotypic predictors for precisely predicting the coreceptor usage in non-B subtypes. For example, the subtype-specific rules were developed to predict the coreceptor usage in HIV-1 subtype D and

* Corresponding author.
E-mail addresses: tanawan.s@cmu.ac.th (T.S. Carraway), watshara.sho@mahidol.ac.th (W. Shoombuatong).

https://doi.org/10.1016/j.compbiolchem.2019.05.006
Received 19 December 2018; Received in revised form 9 April 2019; Accepted 9 May 2019
Available online 20 May 2019
1476-9271/ © 2019 Elsevier Ltd. All rights reserved.

Table 1
Summary of existing methods for prediction of HIV-1 coreceptor usage from sequences.

Year	No. of sequence [a]	Subtypes	Classifier[b]	Available for subtypes[c]	Web server	Ref.
2001	$181^{X4/R5}$, $2^{NSI/SI}$	B	NN	NI		Resch et al. (2001)
2003	$181^{X4/R5}$, $12^{NSI/SI}$	B	Cut-off value	B, C	✓	Jensen et al. (2003)
2007	514	B	SVM	NI		Sander et al. (2007)
2007	651 $^{X4/R5}$, $43^{NSI/SI}$	B, C, and non-B and C	RF	NI		Xu et al. (2007)
2007	$514^{X4/R5}$	B	SVM	NI	✓	Lengauer et al. (2007)
2008	$1425^{X4/R5}$	A, B, C, CRF02_AG, O, D, AI, AG, CRF01_AE, G, and others	SVM with string kernels	NI	✓	Boisvert et al. (2008)
2008	$149^{X4/R5}$	NI	SVM	NI		Lamers et al. (2008)
2012	$273^{X4/R5}$	CRF01_AE	SVM/LMT	CRF01_AE		Shoombuatong et al. (2012)
2013	$1356^{X4/R5}$	A, B, C, D, CRF01_AE, and CRF02_AG	Bulk2clonal	A, B, C, D, CRF01_AE, and CRF02_AG	✓	Cashin et al. (2013), Cashin et al. (2015)
2018	$182^{X4/R5}$	A	RF	A	✓	Lochel et al. (2018)
This study	$666^{X4/R5}$	CRF01_AE	RF	CRF01_AE	✓	This study

[a] NSI/SI dataset was experimentally determined viral tropism based on syncytium formation induced by HIV-1 which was classified as non-syncytium inducing and syncytium inducing viruses, respectively. X4/R5 dataset was experimentally determined viral tropism based on coreceptor usage which was classified as X4 and R5 viruses, respectively.
[b] NN: neural networks, RF: random forest, SVM: support vector machine, SVM/LMT: support vector machine and logistic model trees.

CRF01_AE (Raymond et al., 2011, 2013), while other computational methods were also used to increase the prediction accuracy (Resch et al., 2001; Sander et al., 2007; Xu et al., 2007; Boisvert et al., 2008; Lamers et al., 2008; Shoombuatong et al., 2012; Cashin et al., 2013; Jensen et al., 2003; Cashin et al., 2015; Lochel et al., 2018; Lengauer et al., 2007. As seen in Table 1, almost genotypic predictors were developed and validated by datasets of HIV-1 subtype B and C. Many studies reported that the aforementioned genotypic predictors not performed well in non-B subtypes, including CRF01_AE subtype (Raymond et al., 2011, 2013; Bartlett et al., 2014; Delgado et al., 2012; Garrido et al., 2008; Raymond et al., 2009).

To the best of our knowledge, there are few existing computational models developed for the coreceptor usage prediction of CRF01_AE subtype (Boisvert et al., 2008; Shoombuatong et al., 2012; Cashin et al., 2013; Lengauer et al., 2007). The prediction model SVM/LMT was developed from our group by utilizing the cooperation between support vector machine (SVM) and logistic model tree (LMT) as the prediction engine and the amino acid information on the 9 positions as the input feature. Although, SVM/LMT model yielded quite encouraging results for the prediction of HIV-1 coreceptor usage for CRF01_AE subtype, there was still room for further improvement. First, the training dataset for training SVM/LMT model contained many redundant samples with high similarity. Thus, the SVM/LMT model could provide misleading results with overestimated accuracy. Second, the SVM method is not straight-forward for extracting important features and providing underlying biomedical implications. Third, unfortunately, to the best of our knowledge, no webserver was specifically developed for predicting the coreceptor usage in HIV-1 CRF01_AE subtype.

According to comprehensive reviews (Shoombuatong et al., 2017a, 2017b) and a series of publications (Anuwongcharoen et al., 2016; Pratiwi et al., 2017; Shoombuatong et al., 2015a, 2015b; Simeon et al., 2016a; Simeon et al., 2016b; Win et al., 2017), to establish a really useful and interpretable computational method, the author needs to consider the following procedures: (i) construct or collect a reliable dataset that is experimentally validated sequences for training and validating the model; (ii) represent protein sequences that can truly reflect their intrinsic properties to be predicted; (iii) identify feature importance to improve interpretability and identify the relationship between protein sequences and some biochemical and physicochemical properties; (iv) introduce or develop a powerful and reliable model; (v) evaluate the internal and external predictive power of the model by performing a rigorous cross-validation; (vi) develop a user-friendly webserver for users that can easily get their desired result without needing to go through the mathematical and statistical details.

2. Materials and methods

Briefly, HIVCoR is a RF-based computational approach, which was established using two classes of protein descriptors and single class of DNA descriptor. A summary of the workflow of HIVCoR for predicting HIV-1 CRF01_AE coreceptor usages is shown in Fig. 1.

2.1. Data sets

Benchmark dataset: Amino acid sequences of the V3 region of HIV-1 subtype CRF01_AE were obtained from HIV database (http://www.hiv.lanl.gov). The setting for V3 region of CRF01_AE subtype and CCR5 or CXCR4 coreceptors was set as search criteria. To prepare a reliable dataset, the following steps were performed: amino acid sequences of the V3 region of HIV-1 subtype CRF01_AE with the experimentally confirmed their phenotype by either phenotypic or genotypic assay were included and the amino acid sequences containing special letters, such as X and U, were excluded. After performing such step, we obtained 666 sequences with experimentally confirmed obtained 562 and 104 sequences showing R5 and X4 phenotypes, respectively. As mentioned in (Shoombuatong et al., 2017a, 2017b; Pratiwi et al., 2017; Win et al., 2017; Charoenkwan et al., 2013), a predictor trained and tested with redundant samples might yield overestimated accuracy. We thus removed the sequence which shown the same amino acid sequence. Finally, we obtained a benchmark dataset containing 150 and 40 sequences showing R5 and X4 phenotypes, respectively

Independent dataset: Assessing the reliability of a model for prospective predictions of new samples must be considered an integral part of model development (Shoombuatong et al., 2017a, 2017b). Therefore, a total of 30 V3 amino acid sequences of HIV-1 subtype CRF01_AE with the experimentally confirmed their phenotype by phenotypic assay (26 and 4 sequences showing R5 X4 phenotype, respectively) were used. The 30 V3 amino acid sequences were not included in the benchmark dataset, training dataset, feature selection and parameter optimization of the model. They consisted of 17 sequences from our previous study (unpublished data) and 13 sequences from Hongjaisee et al. (2017). This study was approved by Ethical committees of Faculty of Associated Medical Sciences, Chiang Mai University (Reference number: AMSEC-61EM-005).

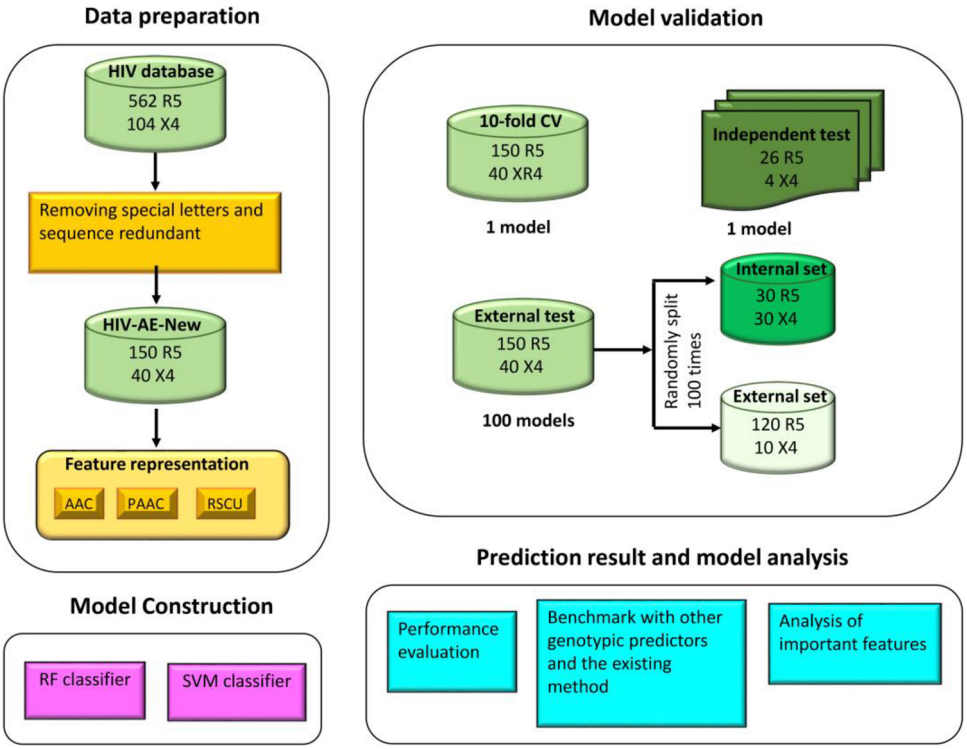

Fig. 1. Schematic workflow of HIVCoR web server for predicting HIV-1 CRF01_AE coreceptor usage.

2.2. Feature extraction

Formulating a given protein or DNA sequence **P** with an effective mathematical expression is one of the important steps to develop a sequence-based predictor for the prediction of HIV-1 CRF01_AE coreceptor usage. In this study, two types of protein features, i.e. amino acid composition (AAC) and pseudo amino acid composition (PseAAC), and one type of DNA features, i.e. relative synonymous codon usage frequency (RSCU), were explored for constructing the HIVCoR predictor.

AAC is the counting of occurrences in a protein sequence which is expressed as a fixed length of 20 components (20-dimensional vector), reflecting the normalized occurrence frequencies p_i of the 20 natural amino acids. Finally, the AAC feature of a protein sequence can be expressed as a vector $\mathbf{P}_{\mathrm{AAC}}$ in a 20-dimensional space:

$$\mathbf{P}_{\mathrm{AAC}} = [p_1, \ldots, p_{20}]^T \quad (1)$$

where T is a transpose operator.

As shown in Eq. (1), the AAC feature only provides the composition formation of protein sequence and its sequence-order or position-specific information would be totally lost. From a machine learning point of view, using AAC feature to represent a protein sequence might considerably limit the prediction quality. To overcome the problem, PseAAC was proposed by incorporating protein sequential information into the traditional AAC feature with the global and local protein structural properties (Chou, 2011). Therefore, in this study, the PseAAC feature was extracted by using the *protr* R package (Xiao et al., 2015). When generating PseAAC feature of a protein sequence ($\mathbf{P}_{\mathrm{PseAA}}$), this feature is defined in 20 + λ components (20 + λ-dimensional vector) where λ is a lambda parameter as formulated below:

$$\mathbf{P}_{\mathrm{PseAAC}} = [p_1, \ldots, p_{20}, p_{20+1}, \ldots, p_{20+\lambda}, \ldots]^T \quad (2)$$

where the first 20 components $[p_1,\ldots,p_{20}]$ are the same as those in the AAC feature and p_{20+1} is the pseudo amino acid component related to the 1st rank of the sequence order correlation, p_{20+2} is the pseudo amino acid component related to the 2nd rank of the sequence order correlation and so forth. In this study, the following parameter set was applied: (1) all of three default properties, i.e. hydrophobicity, hydrophilicity and side chain mass; (2) lambda parameter (λ) and weighting factor (w) were set by using default values, i.e. λ = 30 and w = 30.

RSCU demonstrates the ratio value between the usage frequency of single codon in a gene sample and an expected frequency in the synonymous codon family. Practically, given a DNA sequence, it can be represented with 64 codons according to the genetic code. However, in the view of machine learning, unique codons, i.e. UGG (Trp, W) and AUG (Met, M) and three stop codons, i.e. UGA, UAA, and UAG, must be considered to remove due to their non-informative in learning process of model. Finally, the RSCU feature of a DNA sequence can be expressed as a vector $\mathbf{P}_{\mathrm{RSCU}}$ in a 59-dimensional space:

$$\mathbf{P}_{\mathrm{RSCU}} = [p_1, \ldots, p_{59}]^T \quad (3)$$

2.3. Data separation

The benchmark dataset was highly imbalanced as the ratio of X4 to R5 phenotypes was 1:3.75. From a machine learning point of view, a predictor, if trained and tested with such an imbalanced dataset, might perform poorly on the minority class (Pratiwi et al., 2017; Phanus-umporn et al., 2018). To alleviate this data imbalance issue, an internal dataset was obtained by using an under-sampling technique to randomly selecting 30 and 30 sequences showing R5 and X4 phenotypes from the initial dataset of 150 and 40 sequences, respectively. The remaining sequences of 120 and 10 sequences showing R5 and X4 phenotypes, respectively, were used as external dataset. The more details of internal and external dataset are summarized in Table 2.

2.4. Classifier selection

After the representation of HIV-1 CRF01_AE coreceptor usage, the third step is to develop a sequence-based predictor with a machine learning method. SVM and RF are well-known as the most efficient and powerful methods in protein, peptide and other related problems, such

Table 2
Summary of all datasets used in this study.

Dataset	Phenotype	Original	Data-preprocessing	Internal set	External set
Benchmark	R5	562	150	30	120
	X4	104	40	30	10
Independent	R5	26	–	–	–
	X4	4	–	–	–

Table 3
Performance comparisons between RF and SVM models using different types of features assessed by a 10-fold cross-validation procedure.

Feature	Model	Ac (%)	Sn (%)	Sp (%)	MCC	AUC
AAC	SVM	93.86	95.24	93.55	0.82	0.98
	RF	94.07	92.68	94.36	0.81	0.98
PseAAC	SVM	95.61	95.00	95.79	0.86	0.99
	RF	95.29	100.00	94.38	0.86	1.00
RSCU	SVM	92.36	96.15	91.53	0.79	0.96
	RF	94.49	92.86	94.85	0.83	0.98

AAC: amino acid composition; PseAAC: pseudo amino acid composition; RSCU: relative synonymous codon usage frequency; SVM: support vector machine; RF: random forest; Ac: accuracy; Sn: sensitivity; Sp: specificity; MCC: Matthews coefficient correlation; AUC: area under the ROC curve.

as predicting HIV-1 CRF01_AE coreceptor usages (Shoombuatong et al., 2012), predicting protein crystallizations (Charoenkwan et al., 2013), predicting the oligomeric states of fluorescent proteins (Simeon et al., 2016b), predicting antifreeze proteins (Pratiwi et al., 2017), predicting the hemolytic activity of peptides (Win et al., 2017), prediction of the disulfide bonding state of cysteines (Shoombuatong et al., 2011), predicting aromatase inhibitory activity (Shoombuatong et al., 2015b), prediction of human leukocyte antigen gene (Shoombuatong et al., 2013, 2015c), predicting metabolic syndrome (Worachartcheewan et al., 2015a), predicting influenza neuraminidase inhibitors (Anuwongcharoen et al., 2016), predicting hepatitis C virus NS5B polymerase inhibitory activity (Worachartcheewan et al., 2015b), predicting human acetylcholinesterase inhibitions (Simeon et al., 2016a), predicting anti-sickling activities (Phanus-umporn et al., 2018), and predicting dipeptidyl peptidase-4 inhibitors (Shoombuatong et al., 2015a). Herein, the basic concepts and parameter optimization for the SVM and RF models are briefly described as follows.

SVM is a machine-learning algorithm based on the Vapnik-Chervonenkis theory of statistical learning (Vapnik, 2003; Vapnik, 1999). In practice, SVM can be used for both classification and regression. The principle idea of this method is to transform an input space with m-dimensional vector into a feature space with n-dimensional vector, where $m < n$, and select a separating hyper plane with the largest distance between two classes. In this work, each sample on a dataset has a corresponding label $y \in \{-1, 1\}$ where $+1$ and -1 represent X4 and R5, respectively. Many studies reported that SVM can perform well on small sample size due to its excellent learning and best generalization abilities. In this study, the *kernlab* R package (Karatzoglou et al., 2004) was used to implement SVM model. To obtain an optimal SVM model, the regularization parameter C and kernel parameter γ were tuned by using grid search method with a cross-validation technique, of which the search space for C and γ are $\{2^{-8}, 2^{-7}, ..., 2^{8}\}$ and $\{2^{-8}, 2^{-7}, ..., 2^{8}\}$ with steps of 2 and 2, respectively.

RF is an ensemble method that integrates many classification and regression trees (CART). Breiman et al. (Breiman, 2001) developed the RF method by growing many weak decision trees for enhancing the prediction performance of CART. RF model takes advantage of two powerful machine-learning techniques: bagging and random feature selections. Each decision tree is created as follows: (1) given a dataset S consisting of N samples and F features, in bagging, each decision tree is trained by a dataset S^{bag}, where S^{bag} is derived from randomly selecting M sample from S ($M < N$), while the remaining samples are used as testing sample to evaluate the performance; (2) for each node of a decision tree, the RF model randomly selects *mtry* features from F features and uses them to determine the best possible split evaluated with the Gini index; (3) each decision tree is grown not pruning. Herein, the RF model was created using the *randomForest* R package (Breiman, 2017; Liaw and Wiener, 2002). To obtain an optimal best RF model, two parameters namely *ntree* (i.e. the number of tree used for constructing the RF classifier) and *mtry* (i.e. the number of random candidate features) were tuned in an optimization procedure using the *caret* R package, of which the search space for *ntree* and *mtry* are {100, 200, …, 500} and {1, 2, …, 5} with steps of 100 and 1, respectively.

2.5. Feature importance

Analysis and identification of feature importance play an important role in understanding of the biological and physicochemical properties of HIV-1 CRF01_AE coreceptor. Since the RF method can provide interpretable rules, it is helpful to reveal the different characteristic between X4 and R5. During the RF model construction, each feature is measured its ability in discriminating the sample classes by using Gini index. The Gini index can be defined as

$$1 - \sum_{c=1}^{2} p^2(c|t) \tag{4}$$

where $p(c|t)$ denotes the estimated class probability for node t in a decision tree where c is the class label (i.e. either X4 or R5). The Gini

Table 4
Performance comparisons between RF and SVM models using different types of features assessed by an external test.

Feature	Model	Internal test			External test		
		Ac (%)	MCC	AUC	Ac (%)	MCC	AUC
AAC	SVM	88.17 ± 4.64	0.77 ± 0.09	0.96 ± 0.02	86.10 ± 3.80	0.54 ± 0.06	0.97 ± 0.01
	RF	89.83 ± 3.31	0.80 ± 0.07	0.96 ± 0.02	89.64 ± 2.66	0.58 ± 0.07	0.97 ± 0.01
PseAAC	SVM	93.55 ± 2.90	0.87 ± 0.06	0.99 ± 0.01	92.23 ± 2.63	0.66 ± 0.08	0.99 ± 0.01
	RF	93.45 ± 4.36	0.87 ± 0.05	0.99 ± 0.01	93.79 ± 2.86	0.71 ± 0.09	0.99 ± 0.01
RSCU	SVM	83.98 ± 5.51	0.68 ± 0.11	0.93 ± 0.04	81.97 ± 3.76	0.51 ± 0.06	0.93 ± 0.03
	RF	90.48 ± 4.70	0.81 ± 0.09	0.96 ± 0.03	92.31 ± 2.75	0.70 ± 0.07	0.97 ± 0.01

AAC: amino acid composition; PseAAC: pseudo amino acid composition; RSCU: relative synonymous codon usage frequency; SVM: support vector machine; RF: random forest; Ac: accuracy; MCC: Matthews coefficient correlation; AUC: area under the ROC curve.

Table 5
Performances comparisons between HIVCoR and SVM/LMT models by 10-fold cross-validation and external test on the benchmark dataset.

Validation	Model	Ac (%)	Sn (%)	Sp (%)	MCC	AUC
10-fold CV	SVM/LMT	88.43	90.32	88.11	0.65	0.96
	HIVCoR	95.29	94.38	100.00	0.86	1.00
Internal test	SVM/LMT	73.13 ± 5.52	72.89 ± 6.03	73.83 ± 6.06	0.46 ± 0.11	0.85 ± 0.05
	HIVCoR	93.67 ± 2.93	93.23 ± 3.79	94.30 ± 3.35	0.87 ± 0.06	0.99 ± 0.01
External test	SVM/LMT	74.24 ± 5.14	75.95 ± 11.56	73.99 ± 6.62	0.36 ± 0.07	0.86 ± 0.04
	HIVCoR	93.80 ± 3.34	93.90 ± 8.63	93.79 ± 3.65	0.71 ± 0.10	0.99 ± 0.01

Ac: accuracy; Sn: sensitivity; Sp: specificity; MCC: Matthews coefficient correlation; AUC: area under the ROC curve.

importance or the mean decrease of Gini index (MDGI) is a robust quantity that is used to estimate feature importance. Features with the largest MDGI value are considered to be important as they significantly contribute to the prediction performance. In this study, 10 RF models were constructed by varying the *mtry* parameter settings from 2 to 20 (*mtry = 2, 3, 5, 7, 9, 11, 13, 15, 17, 20)* and fixing the *ntree* parameter with 100 (Pratiwi et al., 2017; Win et al., 2017; Phanus-umporn et al., 2018). The use of multiple RFs might increase the reliability for identifying the feature importance. Furthermore, we employed a multivariate technique named principal component analysis (PCA) to exploring different pattern from amino acid sequences showing R5 and X4 phenotypes (Anuwongcharoen et al., 2016; Shoombuatong et al., 2015a). In this study, PCA was performed on a set of AAC feature using the FactoMineR (Lê et al., 2008) package of the R statistical language.

2.6. Evaluation of the prediction performance

After the construction of prediction models, one of important steps is to objectively identify or evaluate its prediction performances. To provide more easier-to-understand method to measure the prediction performance, four standard statistical parameters namely accuracy (Ac), sensitivity (Sn), specificity (Sp) and Matthews coefficient correlation (MCC) were adopt. Equations describing the details of the four statistical parameters are given:

$$Ac = \frac{TP + TN}{(TP + TN + FP + FN)} \times 100 \quad (5)$$

$$Sn = \frac{TP}{(TP + FN)} \times 100 \quad (6)$$

$$Sp = \frac{TN}{(TN + FP)} \times 100 \quad (7)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

where TP is the number of correctly predicted peptides in the positive dataset (X4) and TN is the number of correctly predicted peptides in the negative dataset (R5). Meanwhile, FP represents the number of incorrectly predicted peptides in the positive dataset and FN represents the number of incorrectly predicted peptides in the negative dataset. Furthermore, the receiver operating characteristic (ROC) curves were used to demonstrate the prediction ability of the proposed model. Herein, the area under the ROC curve (AUC) (Robin et al., 2011) was used to measure the prediction performance, where AUC values of 0.5 and 1 are indicative of random and perfect models, respectively.

After selection of good metrics to measure the prediction ability of the model, the next points modelers need to consider the standard validation method. In statistical predictions, the following three methods are often used to identify the empirical predictive model for its robustness in practical applications: *N*-fold cross-validation, jackknife test and external validation. To fairly compare the existing approaches, a 10-fold cross-validation (10-fold CV) was carried out. In the 10-fold CV procedure, the training data set is randomly divided into 10 subsets. Using 9 subsets were used to build the model and the remaining subset was used to validate the model. This procedure is repeated *N* times, until each sample has been used once as the testing set. In the present study, the internal and external sets were constructed with random procedure. With the bias of single random sampling, we repeated this construction 100 times. In consideration of the impact of the random sampling procedure, we repeated this construction 100 times. The former set was used as the training set and subjected to 5-fold CV in which the data set was partitioned into 5 folds and 1 fold was left out as the testing set while the remaining were used for training the model. This process was repeated iteratively until all the folds have had the chance to be left out as the testing set and the mean and standard deviation values of the statistical parameters are reported.

2.7. Reproducible research

To ensure the reproducibility of the models proposed herein, all R codes and the benchmark dataset used in the construction of the model, graphical figures and the HIVCoR webserver are available on GitHub at https://github.com/Shoombuatong2527/HIVCoR.

3. Results and discussions

To compare our prediction results with previous genotypic predictors and the existing method (Shoombuatong et al., 2012), the 10-fold CV procedure was performed, while the external test was carried out to identify the ranges of external predictive ability of a model. As described in the evaluation of the prediction performance section, the internal and external sets were created for 100 iterations to protect the impact of random splitting procedure, where internal and external sets consisted of 60 and 130 sequences, respectively, as summarized in Table 2. The final prediction results of the external test were represented with the mean and standard deviation values of the statistical parameters. The detailed prediction results are shown in Tables 3–5.

3.1. Prediction performance

In this study, we made an effort to determine the models that were more efficient and applicable to discriminating X4 from R5 by performing RF and SVM models with various types of sequential features, i.e. AAC, PseAAC and RSCU, over both 10-fold CV and external test. Tables 3 and 4 list the prediction results of performance comparisons over both 10-fold CV and external test, respectively. Table 3 shows that the best Ac and MCC of 95.61% and 0.86 is achieved by using SVM model cooperating with PseAAC feature, while the second best Ac and MCC of 95.29% and 0.86 is achieved by using RF model cooperating with the same feature. In case of DNA sequential feature represented with RSCU, this feature can provide the third best performance with Ac and MCC of 94.49% and 0.83, respectively, by cooperating with the RF model.

Interestingly, Table 4 shows that the RF and SVM model cooperating with PseAAC still performed well the first and second best performance with Ac/MCC of 93.79 ± 2.86%/0.71 ± 0.09 and 92.23 ± 2.63%/

Table 6
Performance comparison of various genotypic predictors for the prediction of the coreceptor usage of HIV-1 CRF01_AE subtype on the independent dataset.

No.	V3 amino acid sequences															
1	C	T	R	P	S	N	N	T	R	K	S	I	N	I	G	P
2	C	I	R	P	N	N	N	T	R	T	S	V	G	L	G	P
3	C	T	R	P	S	N	N	T	R	T	S	M	T	I	G	P
4	C	T	R	P	S	N	N	T	R	T	S	I	T	I	G	P
5	C	T	R	P	S	N	N	T	R	A	S	I	S	I	G	P
6	C	T	R	P	S	N	N	T	R	K	G	L	H	I	G	P
7	C	T	R	P	S	N	N	R	R	T	S	T	P	I	G	P
8	C	T	R	P	S	N	N	T	I	T	G	I	T	L	G	P
9	C	T	R	P	S	N	N	T	R	T	S	V	R	I	G	P
10	C	T	R	P	S	N	N	T	R	T	S	V	H	I	G	P
11	C	T	R	P	S	N	N	T	S	T	S	M	T	M	G	P
12	C	T	R	P	S	T	N	V	R	T	S	T	P	M	G	P
13	C	T	R	P	S	N	N	T	R	T	R	V	T	M	G	P
14	C	T	R	P	G	N	N	T	R	I	G	V	T	R	G	P
15	C	S	R	P	S	N	N	V	R	T	S	M	R	I	G	P
16	C	T	R	P	S	N/K	N	T/I	R	T	S	I	H/R	I/M	G	P
17	C	T	R	P	S	N	N	T/I	R	K	S	T/I	P	I	G	P
18	C	T	R	P	Y	N	N	R	R	T	S	T	P	I	G	P
19	C	T	R	P	S	N	K	R	R	T	S	T	P	I	G	P
20	C	T	R	P	S	N	N	R	R	T	R	T	P	I	G	P
21	C	T	R	P	S	N	N	R	R	T	S	V	P	I	G	P
22	C	T	R	P	S	N	N	R	R	T	S	F	P	I	G	P
23	C	T	R	P	S	N	N	R	R	T	S	T	P	I	G	P
24	C	T	R	P	S	N	N	R	R	T	S	T	P	I	G	P
25	C	T	R	P	S	N	N	R	R	T	S	T	P	I	G	P
26	C	T	R	P	S	N	N	R	R	T	S	T	P	I	G	P
27	C	T	R	P	Y	N	N	R	R	T	S	T	P	I	G	P
28	C	T	R	P	S	N	N	R	R	T	S	T	P	I	G	P
29	C	T	R	P	S	N	N	R	R	T	S	T	P	I	G	P
30	C	T	R	P	S	N	N	R	R	T	S	T	P	I	G	P
Total R5																
Total X4																
Concordance (%)																
Sensitivity (%)																
Specificity (%)																
ROC																
Kappa (*k*)																

No.	V3 amino acid sequences															
1	G	R	A	W	Y	A	T	G	Q	I	I	G	D	I	R	Q	A
2	G	Q	V	F	F	R	T	G	D	I	I	G	D	I	R	Q	A
3	G	Q	V	F	Y	R	T	G	D	I	T	G	D	I	R	K	A
4	G	R	V	F	Y	R	T	G	D	I	V	G	D	I	R	K	A
5	G	R	V	Y	Y	K	P	G	D	I	I	G	N	I	R	K	A
6	G	Q	V	F	Y	R	T	G	D	I	I	G	D	I	R	K	A
7	G	Q	A	F	Y	H	Y	G	D	I	I	G	N	I	R	S	A
8	G	Q	V	F	Y	R	T	G	D	I	K	G	N	I	R	K	A
9	G	Q	V	F	Y	K	T	G	D	I	K	G	D	I	R	K	A

(*continued on next page*)

Table 6 (*continued*)

No.	V3 amino acid sequences																
10	G	Q	A	F	Y	K	T	G	G	I	I	G	D	I	R	Q	A
11	G	Q	V	F	Y	R	T	R	E	T	Y	G	D	I	R	K	A
12	G	Q	A	F	Y	R	T	G	D	I	I	G	D	I	R	N	A
13	G	R	V	Y	Y	K	T	G	E	I	V	G	D	I	R	R	A
14	G	H	V	Y	Y	R	T	G	D	I	I	G	D	I	R	K	A
15	G	Q	V	F	Y	R	T	G	S	I	T	G	D	I	R	K	A
16	G	K/Q	V	F	Y	K/R	T	G	D	I	I	G	D	I	R	K/N	A
17	G	Q	V	F	Y	R	T	G	E	I	I	G	N/D	I	R	K	A
18	G	Q	A	F	Y	H	Y	G	D	I	I	G	N	I	R	S	A
19	G	Q	A	F	Y	H	Y	G	D	I	I	G	N	I	R	S	A
20	G	Q	A	F	Y	H	Y	G	D	I	I	G	N	I	R	S	A
21	G	Q	A	F	Y	H	Y	G	D	I	I	G	N	I	R	S	A
22	G	Q	A	F	Y	H	Y	G	D	I	I	G	N	I	R	S	A
23	G	R	A	F	Y	H	Y	G	D	I	I	G	N	I	R	S	A
24	G	Q	A	F	Y	H	Y	G	D	I	T	G	N	I	R	S	A
25	G	Q	A	F	Y	H	Y	G	D	I	V	G	N	I	R	S	A
26	G	Q	A	F	Y	H	Y	G	D	I	I	G	N	I	R	R	A
27	G	Q	A	F	Y	H	Y	G	D	I	I	G	N	I	R	R	A
28	G	R	A	F	Y	H	Y	G	D	I	I	G	N	I	R	R	A
29	G	Q	A	F	Y	H	Y	G	D	I	V	G	N	I	R	R	A
30	G	Q	A	F	Y	H	Y	G	D	I	T	G	N	I	R	R	A
Total R5																	
Total X4																	
Concordance (%)																	
Sensitivity (%)																	
Specificity (%)																	
ROC																	
Kappa (*k*)																	

No.	V3 amino acid sequences		Mean Log RLU in U373 cells		PHENO	Genotypic predictors											
			CCR5	CXCR4		11/25 rule	Net charge	G2p 1%	G2p 2.5%	G2p 5%	G2p 10%	G2p 15%	G2p 20%	PSSM (X4R5)	PSSM (SINSI)	Raymond-'s rule	SVM/LMT	HIVCoR
1	H	C	4.05	2.1	R5	R5	X4	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5
2	H	C	5.87	3.08	R5	R5	R5	R5	R5	R5	X4	X4	X4	R5	R5	R5	R5	R5
3	Y	C	5.15	2.25	R5	R5	R5	R5	R5	R5	R5	R5	X4	R5	R5	R5	R5	R5
4	Y	C	4.31	2.28	R5	R5	R5	R5	R5	R5	X4	X4	X4	R5	R5	R5	R5	R5
5	Y	C	4.79	2.33	R5	R5	X4	R5	R5	X4	X4	X4	X4	X4	X4	R5	R5	R5
6	G	C	5.99	2.18	R5	R5	X4	R5	R5	R5	X4	X4	X4	R5	R5	R5	R5	R5
7	Y	C	6.08	2.14	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5
8	Y	C	5.15	2.24	R5	R5	R5	R5	R5	R5	R5	X4	X4	X4	X4	R5	R5	R5
9	Y	C	4.49	2.12	R5	R5	X4	R5	X4	X4	X4	X4	X4	X4	X4	R5	R5	R5
10	Y	C	5.15	2.07	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5
11	F	C	4.18	2.23	R5	R5	R5	R5	R5	X4	X4	X4	X4	X4	X4	R5	R5	R5
12	Y	C	5.75	2.31	R5	R5	R5	R5	R5	R5	R5	X4	X4	R5	R5	R5	R5	R5
13	Y	C	3.66	3.74	R5X4	X4	X4	X4	X4	X4	X4	X4	X4	X4	X4	X4	X4	X4
14	Y	C	4.06	3.98	R5X4	R5	X4	R5	X4	X4	X4	X4	X4	X4	X4	R5	R5	X4
15	Y	C	5.34	4.72	R5X4	R5	X4	R5	R5	X4	X4	X4	X4	X4	X4	X4	R5	R5
16	H	C	2.26	3.8	X4	R5	X4	R5	R5	X4	X4	X4	X4	X4	X4	X4	R5	R5
17	Y	C	5.86	2.34	R5	R5	X4	R5	R5	R5	R5	X4	X4	X4	R5	X4	R5	R5

(*continued on next page*)

Table 6 (*continued*)

No.	V3 amino acid sequences		Mean Log RLU in U373 cells		PHENO	Genotypic predictors												
			CCR5	CXCR4		11/25 rule	Net charge	G2p 1%	G2p 2.5%	G2p 5%	G2p 10%	G2p 15%	G2p 20%	PSSM (X4R5)	PSSM (SINSI)	Raymond-'s rule	SVM/LMT	HIVCoR
18	Y	C	6.23	2.34	R5	R5	R5	R5	R5	R5	R5	R5	R5	X4	R5	R5	X4	R5
19	Y	C	5.71	2.41	R5	R5	X4	R5	R5	R5	X4	X4	X4	X4	R5	X4	R5	R5
20	Y	C	4.76	3.41	R5	X4	X4	R5	X4	X4	X4	X4	X4	X4	X4	X4	R5	R5
21	Y	C	6.07	2.64	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5
22	Y	C	6.25	2.38	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5
23	Y	C	5.14	2.46	R5	R5	X4	R5	R5	R5	R5	R5	R5	X4	R5	X4	R5	R5
24	Y	C	6.25	2.46	R5	R5	R5	R5	R5	R5	R5	R5	R5	X4	R5	R5	R5	R5
25	Y	C	6.32	2.38	R5	R5	R5	R5	R5	R5	R5	R5	R5	X4	R5	R5	R5	R5
26	Y	C	5.85	2.28	R5	R5	X4	R5	R5	R5	X4	X4	X4	R5	R5	X4	R5	R5
27	Y	C	5.87	2.22	R5	R5	X4	R5	X4	X4	X4	X4	X4	X4	R5	X4	X4	R5
28	Y	C	4.92	2.41	R5	R5	X4	R5	R5	X4	X4	X4	X4	X4	R5	X4	R5	R5
29	Y	C	6.16	2.82	R5	R5	X4	R5	R5	R5	X4	X4	X4	X4	R5	X4	R5	R5
30	Y	C	6.06	2.6	R5	R5	X4	R5	R5	R5	X4	X4	X4	R5	R5	X4	R5	R5
Total R5					26	28	13	29	25	20	13	10	9	12	21	18	27	28
Total X4					4	2	17	1	5	10	17	20	21	18	9	12	3	2
Concordance (%)						87	57	90	83	80	57	47	43	53	83	67	83	93
Sensitivity (%)						25	100	25	50	100	100	100	100	100	100	75	25	50
Specificity (%)						96	50	100	89	77	50	39	35	46	81	65	92	100
ROC						0.61	0.75	0.63	0.7	0.89	0.75	0.7	0.67	0.73	0.9	0.7	0.59	0.75
Kappa (*k*)						0.27	0.21	0.37	0.35	0.47	0.21	0.14	0.12	0.19	0.53	0.22	0.19	0.63

Based on phenotypic assay, R5X4 viruses have the ability to use both of coreceptors. Based on prediction, the predictors should be predicted as X4 viruses. Concordance: number of samples concordant between genotype and phenotype for R5 and X4/R5X4, divided by the total number of samples; Sn: sensitivity; Sp: specificity; ROC: Receiver Operating Characteristic curve; *k*: kappa coefficient; RLU: relative light unit.

Table 7
Summary of amino acid composition of amino acid sequences showing R5 and X4 phenotypes.

Amino acid	Composition of R5 (%)	Composition of X4 (%)	Difference	*p*-value	MDGI
Arg (R)	12.024	14.983	−2.959	0.000	12.498
Tyr (Y)	5.202	6.063	−0.861	0.004	4.968
Asn (N)	6.117	4.111	2.005	0.000	3.302
Lys (K)	2.630	4.948	−2.318	0.000	3.195
Asp (D)	4.821	3.415	1.406	0.000	2.523
Phe (F)	3.106	3.415	−0.309	0.728	1.983
Ser (S)	5.583	3.554	2.029	0.000	1.870
His (H)	0.610	0.767	−0.157	0.521	1.214
Ile (I)	12.538	9.617	2.921	0.000	1.202
Gly (G)	11.643	11.429	0.214	0.453	1.159
Val (V)	2.820	4.530	−1.709	0.000	0.863
Leu (L)	0.400	0.767	−0.366	0.084	0.749
Gln (Q)	3.182	1.672	1.510	0.000	0.683
Glu (E)	0.743	1.394	−0.651	0.034	0.654
Met (M)	0.610	1.463	−0.854	0.007	0.459
Ala (A)	3.620	3.902	−0.282	0.386	0.445
Thr (T)	12.405	12.125	0.279	0.724	0.068
Trp (W)	0.171	0.418	−0.247	0.143	0.054
Pro (P)	6.059	5.784	0.275	0.228	0.045
Cys (C)	5.716	5.645	0.072	0.585	−0.101

The total number of amino acids in the benchmark dataset of R5 and X4 phenotypes are 5248 and 1435, respectively; *p*-value: *t*-Student's test; MDGI: mean decrease of Gini index. The amino acid with the largest value of MDGI is the most important.

0.66 ± 0.08, respectively, on the external test. The prediction results showed that PseAAC is high beneficial feature to the prediction of HIV-1 CRF01_AE coreceptor usage. From Table 4, we could find that RF model outperforms than SVM model with an improvement of > 5% on MCC. As shown in Tables 3 and 4, we could find that RF model cooperating PseAAC feature achieves better performances that other models on the benchmark dataset over 10-fold CV and external test. From this point forward, the best model for the prediction of HIV-1 CRF01_AE coreceptor usage that was developed by the RF model with PseAAC feature will be referred to as HIVCoR.

3.2. Comparison the HIVCoR to SVM/LMT

We also compared the HIVCoR with the previous model developed by our group, called SVM/LMT (Shoombuatong et al., 2012). This approach was constructed by using SVM model cooperating with feature sector named LMT model, called SVM/LMT model. It is not fair to directly compare the results of HIVCoR with SVM/LMT because SVM/LMT was trained and tested on an out-of-date dataset. To perform an equal comparison, a modified SVM/LMT model was implemented using the same dataset and validation. In this study, SVM/LMT was performed both 10-fold CV and external test on the benchmark dataset. The results of performance comparisons are listed in Table 5.

As seen in Table 5, SVM/LMT yields the value of Ac, Sn and Sp of 88.43%, 90.32% and 88.11%. This result was lower than the prediction results reported by our previous work (Shoombuatong et al., 2012) because, in this work, SVM/LMT was trained and tested with redundant samples. We can notice that HIVCoR achieved better performances than SVM/LMT concerning all statistical parameters over both 10-fold CV and external test. Remarkably, HIVCoR outperformed than SVM/LMT having 6.86% and 21% improvement for Ac and MCC, respective, assessed by the external test. Considering that the external test is the most rigorous validation and the new benchmark dataset is more reliable, we thus claimed that the proposed HIVCoR model is superior to SVM/LMT.

3.3. Comparison the HIVCoR to other genotypic predictors

To demonstrate the more advantages of the proposed model, we made a comparison between HIVCoR and other genotypic predictors on the independent dataset. According to data available, 30 amino acid sequences were set as reference which they have already confirmed their phenotype by phenotypic assay. Table 6 shows the V3 amino acid sequences, phenotypic results and predicted coreceptor usage obtained from HIVCoR and other genotypic predictors. To measure the prediction performance of different genotypic predictors, percent of concordance, sensitivity for detecting CXCR4 viruses, specificity for detecting CCR5 viruses, ROC, and kappa coefficient were used.

From Table 6, the simplest method (11/25 rule; X4 will be predicted if any positively charged amino acids is presented at position 11 or 25) (De Jong et al., 1992) has concordance of 87%, sensitivity of 25%, and specificity of 96%. We found that 3 of 4 X4 viruses were mis-predicted as R5, thus the presence of positively charged amino acids at position 11 or 25 did not strongly support for CXCR4 phenotype. Using net charge rule, based on subtracting the number of negatively charged amino acids (D and E) from the number of positively charged ones (K, H, and R), (Fouchier et al., 1992) it predicted 13 of 26 R5 viruses (50% specificity for detecting R5 viruses) whereas it predicted 4 of 4 X4 viruses. Although, this rule has sensitivity for detecting X4 viruses of 100% but it also overestimated for X4 phenotype which contrast with phenotypic results. In addition, Geno2pheno$_{[coreceptor]}$ at FPR 10%, which is currently recommended, was low specificity for detecting R5 viruses (50%) even the sensitivity for detecting X4 viruses were 100%. It suggested that Geno2pheno$_{[coreceptor]}$ at FPR 10% lacked specificity for CRF01_AE subtype, as previously reported (Raymond et al., 2013). Even though, a high false positive rate of Geno2pheno$_{[coreceptor]}$ allow to detect more X4 viruses, but they showed low concordance, specificity, and kappa coefficient. For WebPSSM tool, it predicted more X4 viruses whereas phenotypic assay determined strictly R5 viruses. From all predictors, we found that they mis-predicted and overestimated for X4 phenotype, especially using Geno2pheno$_{[coreceptor]}$ and WebPSSM as same as reported in previous study (Mulinge et al., 2013). In addition, a specific rule for coreceptor usage prediction of CRF01_AE subtype was also used. It is a subtype CRF01_AE-specific algorithm based on the 11/25 rule, net charge rule and mutation of the N-linked glycosylation site (N301) for the determination of coreceptor usage of CRF01_AE (Raymond et al., 2013). As shown in Table 6, this rule gave the prediction result of 75% sensitivity and 65% specificity, indicating that this rule tended to predict the coreceptor usage as X4 viruses, because the presence of positively charged amino acids in the V3 region. However, the N-linked glycosylation site in V3 region may need further for predicting CRF01_AE CXCR4 coreceptor usage.

Using SVM/LMT tool, it showed high performance in concordance (83%) and specificity for detecting R5 viruses (92%), but the sensitivity for detecting X4 viruses was very low (25%). Interestingly, HIVCoR which is a new predictive model based on random forest method gave a highest concordance (93%, $k = 0.63$), compared to the phenotype results. It also showed high specificity for detecting R5 viruses (100%), even the sensitivity for detecting X4 viruses were 50%. Notably, we found that 2 of 4 X4 viruses were mis-predicted as R5. According to their phenotype results, one of them has a coreceptor usage as dual/mixed viruses (R5X4) and another has a coreceptor usage as CXCR4, but it has a low infectivity in U373.CXCR4 cells and contain mixed amino acids at various positions in its V3 region as shown in Table 6. This may lead the predictors voted to R5 whereas phenotypic assay voted to X4. Also, we cannot rule out that the prediction based on V3 amino acid sequences may impact the reliability of tools because genotypic assay is based on only V3 region whereas phenotypic assay takes the whole envelope region. However, HIVCoR showed a high performance for prediction of HIV-1 coreceptor usage in this testing data set which is better than a previous model (SVM/LMT) and other tools. Thus, it can be used as alternative predictor for determination of HIV-1 coreceptor

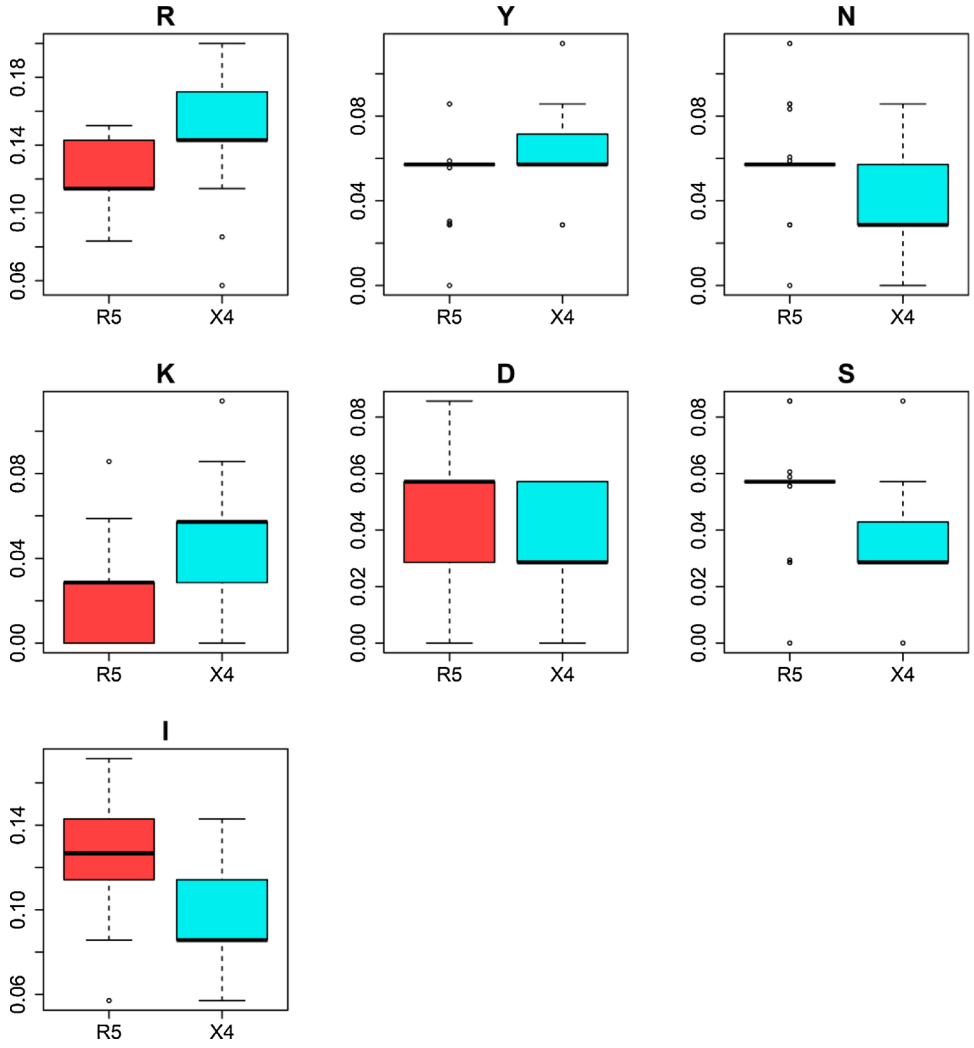

Fig. 2. Box plots of seven amino acid importance derived from the values of MDGI, where red and blue represent R5 and X4, where MDGI is the mean decrease of Gini index.

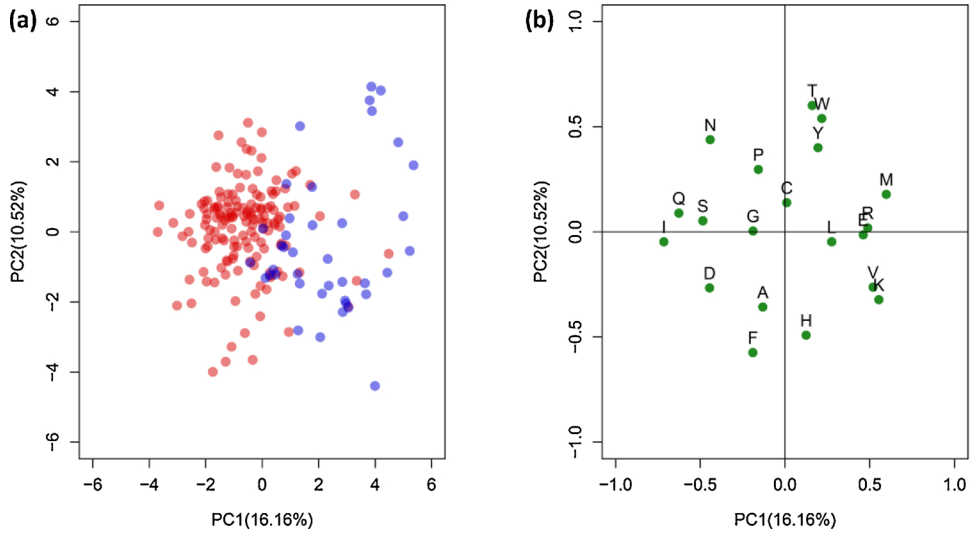

Fig. 3. PCA scores (a) and loadings (b) plots of amino acid sequences of the V3 region of HIV-1 subtype CRF01_AE. R5 and X4 are represented by red and blue circles, respectively, in the scores plot.

Fig. 4. Sequence logo representations of amino acid sequences showing R5 (a) and X4 (b) phenotypes.

usage which is specifically for CRF01_AE subtype.

According to data available, the testing data set was V3 amino acid sequences of HIV-1 subtype CRF01_AE with known phenotype and represented for CRF01_AE in Thailand. Actually, HIV-1 CRF01_AE subtype is not only circulated in Southeast Asian and East Asian countries but also found in Europe, America, and Africa regions. The phylogenetic analysis of HIV-1 CRF01_AE from different countries around the globe showed that a high level of phylogenetic mixing was found between Thai and European strains, Japanese strains, North American strains, and strains from sampled Asian countries, excluding China and Vietnam, whereas African sequences seem to form a single cluster (Angelis et al., 2015). Thus, the data of HIV-1 CRF01_AE strain circulating in Europe, Africa, and other countries in Asia should be also tested to obtain the better performance of predictive model. Moreover, more data set is further required to assess the performance of predictive model, especially for data set of X4 viruses.

3.4. Mechanistic interpretation of feature importance

The analysis of feature importance can provide a better understanding of the important characteristics governing amino acid sequences of the V3 region of HIV-1 subtype CRF01_AE. Since RF model is the efficient built-in ability of feature importance estimator, it is helpful to provide better understanding of the different characteristic between R5 and X4. In this study, we thus utilized the RF model with the mean decrease of the Gini index to rank the important features. Furthermore, PCA and box plot were applied in exploring the different characteristic of amino acid sequences showing R5 and X4 phenotypes. Table 7 shows the values of MDGI of 20 amino acids by using the RF model with the benchmark dataset consisting with 150 R5 and 40 X4. The total number of amino acid used in the used dataset of R5 and X4 is 5248 and 1435, respectively. Meanwhile, this table also presents the amino acid compositions and *p*-value of 20 amino acids.

As seen in Table 7, the most important amino acid is Arg with an MDGI value of 12.498, while the second and third most important amino acids are Tyr and Asn with MDGI values of 4.968 and 3.302, respective. The top-ten informative amino acid are Arg, Tyr, Asn, Lys, Asp, Phe, Ser, His, Ile and Gly having an MDGI value larger than 1.1. Interestingly, 7 out of 10 top-ranked informative amino acid were significant in the differentiation of R5 from X4, i.e. Arg, Tyr, Asn, Lys, Asp, Ser and Ile ($p < 0.005$). Table 7 and Fig. 2 reveal that Asn, Asp, Ser and Ile are significantly abundant in R5, while Arg, Tyr and Lys are significantly abundant in X4.

Among the HIV-1 gp120 domains, the V3 region is recognized as a major domain for coreceptor interaction (Deng et al., 1996). The V3 is an approximately 35 amino acids long, glycosylated, highly variable, and disulfide-bonded structure. It is composed of three regions: base, flexible stem, and conserved-hairpin crown (tip) (Hartley et al., 2005). Coreceptor interaction has been proposed that V3 base binds to the N-terminus of coreceptor whereas the V3 tip interacts directly with the second extracellular loop of coreceptor (Huang et al., 2005). In fact, the V3 amino acid positions 11, 24, and 25 have been shown to be involved in binding with different coreceptors and the introduction of positively charged amino acids at these positions strongly associated with X4 phenotype (De Jong et al., 1992; Fouchier et al., 1992; Chen et al., 2013; Svicher et al., 2011). It has been explained that the interaction facilitates via electrostatic attraction between positively charged in V3 and negatively charged on extracellular surface of CXCR4 coreceptor that contains sulfated tyrosines (Brelot et al., 1999). In addition, other V3 amino acid positions also significantly correlated with CCR5 or CXCR4 coreceptor usage, e.g. V3 determinants correlated with CXCR4 usage was 5Y, 7K, 9K, 13T, 14V, 19V, and 32K whereas V3 determinants correlated with CCR5 usage was 11S, 25D, 19A, and 27I (Svicher et al., 2011).

We also took an advantage of PCA to data exploration and pattern recognition. As seen in Fig. 3, the percentage of variance explained by the first two principal components (PCs) was 26.68, where red and blue points represent R5 and X4, respectively. Fig. 3(a) shows the score plot representing the scores of the first PC versus the scores of the second PC, while Fig. 3(b) shows the loading plot representing the coefficients of each feature for the first PC versus the coefficients for the second PC. Analyzing the score and loading plots, we could find that Asn, Asp, Ser, Ile and Gln significantly contributed to R5 at level of $p < 0.005$, while Lys, Val and Met significantly contributed to X4 at level of $p < 0.05$.

In addition, we also made an effort to analyze whether certain types of residues are preferred at specific positions in R5 and X4. As suggested

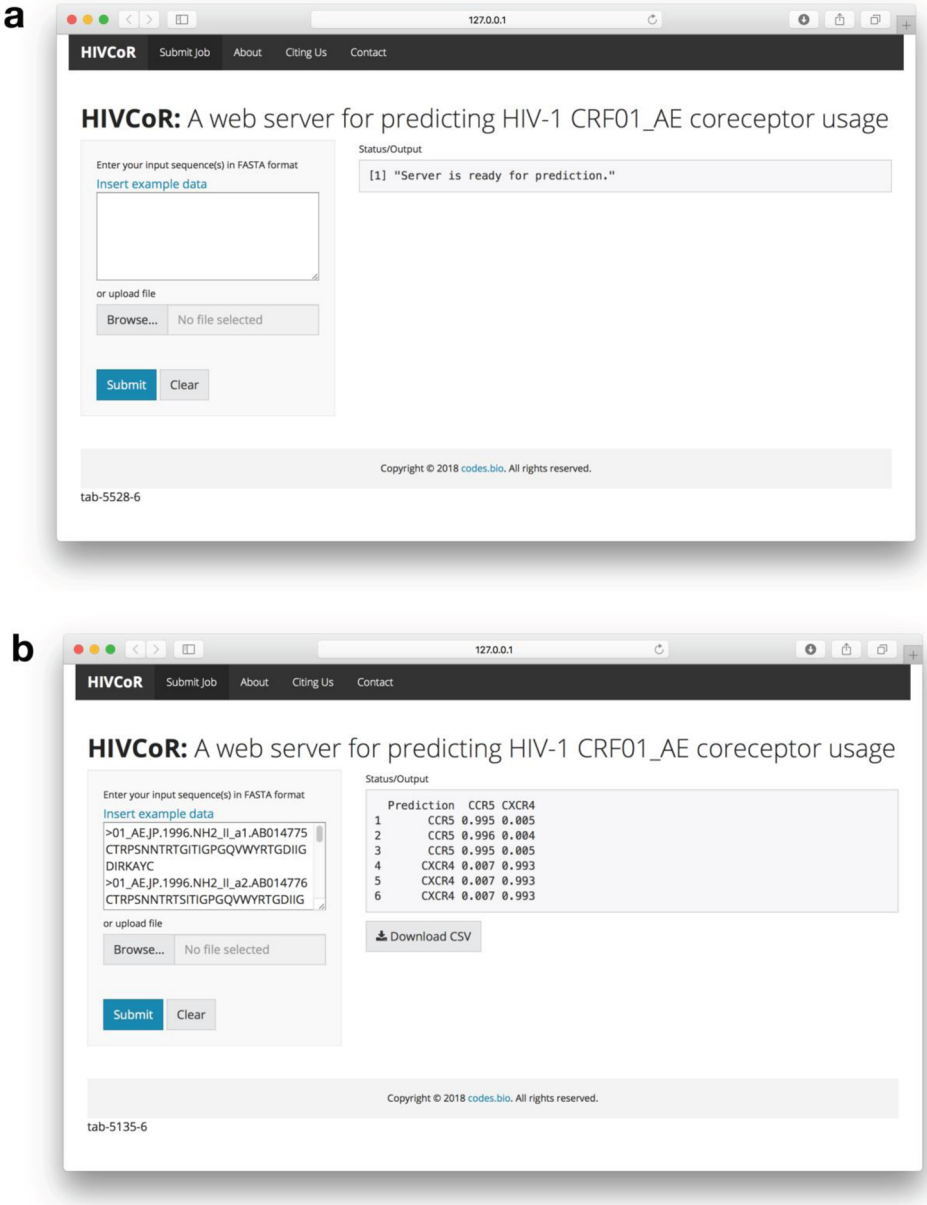

Fig. 5. Screenshots of the HIVCoR web server before (a) and after (b) submission of amino acid sequence(s).

in the work of Shoombuatong et al. (2012), nine positions (P5, P7, P11, P13, P14, P18, P19, P27, and P32) on the V3 amino acid sequences played an important role in discriminating X4 from R5. From Fig. 4, several observations from benchmark dataset can be made as follows; Asn (P13), Arg (P13 and P18), Ala (P19), Thr (P27) and Gln (P32) were more abundant in R5, while Gln (P18), Val (P19), Ile (P27) and Lys (P32) were more abundant in X4. Interestingly, 5 out of 6 important amino acid in R5 were hydrophilic residues, while 2 out of 4 important amino acid in X4 were hydrophobic residues. Earlier study showed that a hydrophobic patch located in the V3 tip region (position 12–20) has been shown to be critical for envelope trimer stability. A single-residue change to a hydrophilic or even to a neutral residue in this hydrophobic patch destabilized the Env trimer and results in an increase of gp120 shedding, a decrease of chemokine receptor binding, a decrease of Env-mediated membrane fusion and a reduction in virus infectivity (Bowder et al., 2018; Xiang et al., 2010). Moreover, this hydrophobic patch also involved in the binding of neutralizing antibodies (NAbs) (de Taeye et al., 2018). However, the interaction between coreceptor and V3 loop is usually driven by electrostatics, between a highly positive V3 loop and a highly negative coreceptor (Brelot et al., 1999; Morikis et al., 2007). The presence of charged amino acids in the V3 loop plays more important role in the coreceptor interaction.

In this study, there are newly-found informative position that have never been reported in our previous study (Shoombuatong et al., 2012), i.e. positions 29. In this position, it was noticed that Asp was more abundant in R5. The amino acid characteristics at these positions may play an important role in determining viral coreceptor usage and distinguishing between CCR5 and CXCR4 usage, especially in HIV-1 CRF01_AE subtype. However, these residues need to be further confirmed about their functions in the binding to coreceptors.

3.5. HIVCoR webserver

To facilitate high-through prediction of HIV-1 coreceptor usage and for convenience of the experimental scientists, a user-friendly and freely accessible web server called HIVCoR was created using the Shiny

package under the R programming environment. Screenshots of the HIVCoR webserver is shown in Fig. 5. Furthermore, a step-by-step guide on how to use the HIVCoR webserver is given as follows:

Step 1. Open the webserver at http://codes.bio/hivcor/. Then you will see the top page of HIVCoR on your computer screen as shown in Fig. 5(a).

Step 2. Type or copy/paste the query amino acid sequence with the FASTA format into the Input box. User can also upload the file of amino acid sequences by clicking on the *Choose file* button.

Step 3. After pressing the *Submit* button, the webserver will automatically initiate the modeling process and make the prediction results, as shown in Fig. 5(b). Typically, it will take a few seconds for the web server to process a task. Users can also download the prediction results as a CSV file by pressing on the *Download CSV* button.

4. Conclusions

Computational prediction of HIV-1 coreceptor usage is strongly recommended before starting the coreceptor-specific inhibitors in HIV treatment and save a lot of cost and time. In this study, we have made an effort to develop a new effective genotypic model for predicting HIV-1 coreceptor usage in CRF01_AE subtype, called HIVCoR. HIVCoR are constructed by using random forest model cooperating with pseudo amino acid composition. The prediction results for both 10-fold cross-validation and external validation test on the benchmark dataset revealed the superiority of HIVCoR over other bioinformatics tools and genotypic predictors. To facilitate the studies of experimental scientist, a webserver for HIVCoR has been established at http://codes.bio/hivcor/ by which users can easily obtain their desired result without needing to go through the mathematical and statistical details. Owning to its high success rate and user-friendliness, it is anticipated that HIVCoR will become one of useful tools used for HIV treatment.

Conflict of interest

No conflict of interest.

Acknowledgements

This work was co-supported by the Office of Higher Education Commission and the Thailand Research Fund (No. MRG6180226) and the AMS Research Grant 2017 from Faculty of Associated Medical Sciences, Chiang Mai University, Thailand. Authors would like to thank Dr. Woottichai Khamduang from Faculty of Associated Medical Sciences, Chiang Mai University for his support with statistical analysis.

References

Angelis, K., Albert, J., Mamais, I., Magiorkinis, G., Hatzakis, A., Hamouda, O., et al., 2015. Global dispersal pattern of HIV type 1 subtype CRF01_AE: a genetic trace of human mobility related to heterosexual sexual activities centralized in southeast Asia. J. Infect. Dis. 211 (11), 1735–1744.

Anuwongcharoen, N., Shoombuatong, W., Tantimongcolwat, T., Prachayasittikul, V., Nantasenamat, C., 2016. Exploring the chemical space of influenza neuraminidase inhibitors. PeerJ 4, e1958.

Bartlett, A.D., Macartney, M.J., Conibear, T.C., Feyertag, F., Smith, C.J., Johnson, M.A., et al., 2014. The utility of different bioinformatics algorithms for genotypic HIV-1 tropism testing in a large clinical cohort with multiple subtypes. AIDS 28 (11), 1611–1617.

Berger, E.A., Doms, R.W., Fenyo, E.M., Korber, B.T., Littman, D.R., Moore, J.P., et al., 1998. A new classification for HIV-1. Nature 391 (6664), 240.

Boisvert, S., Marchand, M., Laviolette, F., Corbeil, J., 2008. HIV-1 coreceptor usage prediction without multiple alignments: an application of string kernels. Retrovirology 5, 110.

Bowder, D., Hollingsead, H., Durst, K., Hu, D., Wei, W., Wiggins, J., et al., 2018. Contribution of the gp120 V3 loop to envelope glycoprotein trimer stability in primate immunodeficiency viruses. Virology 521, 158–168.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Breiman, L., 2017. Classification and Regression Trees. Routledge.

Brelot, A., Heveker, N., Adema, K., Hosie, M.J., Willett, B., Alizon, M., 1999. Effect of mutations in the second extracellular loop of CXCR4 on its utilization by human and feline immunodeficiency viruses. J. Virol. 73 (4), 2576–2586.

Cashin, K., Gray, L.R., Jakobsen, M.R., Sterjovski, J., Churchill, M.J., Gorry, P.R., 2013. CoRSeqV3-C: a novel HIV-1 subtype C specific V3 sequence based coreceptor usage prediction algorithm. Retrovirology 10, 24.

Cashin, K., Gray, L.R., Harvey, K.L., Perez-Bercoff, D., Lee, G.Q., Sterjovski, J., et al., 2015. Reliable genotypic tropism tests for the major HIV-1 subtypes. Sci. Rep. 5, 8543.

Charoenkwan, P., Shoombuatong, W., Lee, H.-C., Chaijaruwanich, J., Huang, H.-L., Ho, S.-Y., 2013. SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. PLoS One 8 (9), e72368.

Chen, M., Svicher, V., Artese, A., Costa, G., Alteri, C., Ortuso, F., et al., 2013. Detecting and understanding genetic and structural features in HIV-1 B subtype V3 underlying HIV-1 co-receptor usage. Bioinformatics 29 (4), 451–460.

Chou, K.-C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273 (1), 236–247.

De Jong, J.J., De Ronde, A., Keulen, W., Tersmette, M., Goudsmit, J., 1992. Minimal requirements for the human immunodeficiency virus type 1 V3 domain to support the syncytium-inducing phenotype: analysis by single amino acid substitution. J. Virol. 66 (11), 6777–6780.

de Taeye, S.W., de la Pena, A.T., Vecchione, A., Scutigliani, E., Sliepen, K., Burger, J.A., et al., 2018. Stabilization of the gp120 V3 loop through hydrophobic interactions reduces the immunodominant V3-directed non-neutralizing response to HIV-1 envelope trimers. J. Biol. Chem. 293 (5), 1688–1701.

Delgado, E., Fernandez-Garcia, A., Vega, Y., Cuevas, T., Pinilla, M., Garcia, V., et al., 2012. Evaluation of genotypic tropism prediction tests compared with in vitro co-receptor usage in HIV-1 primary isolates of diverse subtypes. J. Antimicrob. Chemother. 67 (1), 25–31.

Deng, H., Liu, R., Ellmeier, W., Choe, S., Unutmaz, D., Burkhart, M., et al., 1996. Identification of a major co-receptor for primary isolates of HIV-1. Nature 381 (6584), 661–666.

Fouchier, R.A., Groenink, M., Kootstra, N.A., Tersmette, M., Huisman, H.G., Miedema, F., et al., 1992. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J. Virol. 66 (5), 3183–3187.

Garrido, C., Roulet, V., Chueca, N., Poveda, E., Aguilera, A., Skrabal, K., Garrido, C., Roulet, V., Chueca, N., Poveda, E., Aguilera, A., Skrabal, K., 2008. Evaluation of eight different bioinformatics tools to predict viral tropism in different human immunodeficiency virus type 1 subtypes. J. Clin. Microbiol. 46 (3), 887–891.

Hartley, O., Klasse, P.J., Sattentau, Q.J., Moore, J.P., 2005. V3: HIV's switch-hitter. AIDS Res. Hum. Retroviruses 21 (2), 171–189.

Hongjaisee, S., Braibant, M., Barin, F., Ngo-Giang-Huong, N., Sirirungsi, W., Samleerat, T., 2017. Effect of amino acid substitutions within the V3 region of HIV-1 CRF01_AE on interaction with CCR5-coreceptor. AIDS Res. Hum. Retroviruses 33 (9), 946–951.

Huang, C.C., Tang, M., Zhang, M.Y., Majeed, S., Montabana, E., Stanfield, R.L., et al., 2005. Structure of a V3-containing HIV-1 gp120 core. Science 310 (5750), 1025–1028.

Jensen, M.A., Li, F.S., van't Wout, A.B., Nickle, D.C., Shriner, D., He, H.X., et al., 2003. Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. J. Virol. 77 (24), 13376–13388.

Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. Kernlab – an S4 package for kernel methods in R. J. Stat. Softw. 11 (9), 1–20.

Koot, M., Keet, I.P., Vos, A.H., de Goede, R.E., Roos, M.T., Coutinho, R.A., et al., 1993. Prognostic value of HIV-1 syncytium-inducing phenotype for rate of CD4 + cell depletion and progression to AIDS. Ann. Intern. Med. 118 (9), 681–688.

Lamers, S.L., Salemi, M., McGrath, M.S., Fogel, G.B., 2019. Prediction of R5, X4, and R5X4 HIV -1 coreceptor usage with evolved neural networks. IEEE/ACM Trans. Comput. Biol Bioinform 5 (2), 291–300.

Lê, S., Josse, J., Husson, F., 2008. FactoMineR: an R package for multivariate analysis. J. Stat. Softw. 25 (1), 1–18.

Lengauer, T., Sander, O., Sierra, S., Thielen, A., Kaiser, R., 2007. Bioinformatics prediction of HIV coreceptor usage. Nat. Biotechnol. 25 (12), 1407–1410.

Liaw, A., Wiener, M., 2002. Classification and regression by random. Forest R News 2 (3), 18–22.

Lochel, H.F., Riemenschneider, M., Frishman, D., Heider, D., 2018. SCOTCH: subtype A coreceptor tropism classification in HIV-1. Bioinformatics.

Morikis, D., Rizos, A.K., Spandidos, D.A., Krambovitis, E., 2007. Electrostatic modeling of peptides derived from the V3-loop of HIV-1 gp120: implications of the interaction with chemokine receptor CCR5. Int. J. Mol. Med. 19 (3), 343–351.

Mulinge, M., Lemaire, M., Servais, J.Y., Rybicki, A., Struck, D., da Silva, E.S., et al., 2013. HIV-1 tropism determination using a phenotypic env recombinant viral assay highlights overestimation of CXCR4-usage by genotypic prediction algorithms for CRRF01_AE and CRF02_AG. PLoS One 8 (5), e60566.

Phanus-umporn, C., Shoombuatong, W., Prachayasittikul, V., Anuwongcharoen, N., Nantasenamat, C., 2018. Privileged substructures for anti-sickling activity via cheminformatic analysis. RSC Adv. 8 (11), 5920–5935.

Pratiwi, R., Malik, A.A., Schaduangrat, N., Prachayasittikul, V., Wikberg, J.E., Nantasenamat, C., Shoombuatong, W., 2017. CryoProtect: a web server for classifying antifreeze proteins from nonantifreeze proteins. J. Chem. 2017.

Raymond, S., Delobel, P., Mavigner, M., Cazabat, M., Souyris, C., Encinas, S., et al., 2009. Genotypic prediction of human immunodeficiency virus type 1 CRF02-AG tropism. J. Clin. Microbiol. 47 (7), 2292–2294.

Raymond, S., Delobel, P., Chaix, M.L., Cazabat, M., Encinas, S., Bruel, P., et al., 2011. Genotypic prediction of HIV-1 subtype D tropism. Retrovirology 8, 56.

Raymond, S., Delobel, P., Rogez, S., Encinas, S., Bruel, P., Pasquier, C., et al., 2013.

Genotypic prediction of HIV-1 CRF01-AE tropism. J. Clin. Microbiol. 51 (2), 564–570.
Resch, W., Hoffman, N., Swanstrom, R., 2001. Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks. Virology 288 (1), 51–62.
Richman, D.D., Bozzette, S.A., 1994. The impact of the syncytium-inducing phenotype of human immunodeficiency virus on disease progression. J. Infect. Dis. 169 (5), 968–974.
Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves BMC. Bioinform 12 (1), 77.
Sander, O., Sing, T., Sommer, I., Low, A.J., Cheung, P.K., Harrigan, P.R., et al., 2007. Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. PLoS Comput. Biol. 3 (3), e58.
Shoombuatong, W., Traisathit, P., Prasitwattanaseree, S., Tayapiwatana, C., Cutler, R., Chaijaruwanich, J., 2011. Prediction of the disulphide bonding state of cysteines in proteins using Conditional Random Fields. Int. J. Data Mining Bioinform. 5 (4), 449–464.
Shoombuatong, W., Hongjaisee, S., Barin, F., Chaijaruwanich, J., Samleerat, T., 2012. HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees. Comput. Biol. Med. 42 (9), 885–889.
Shoombuatong, W., Mekha, P., Waiyamai, K., Cheevadhanarak, S., Chaijaruwanicha, J., 2013. Prediction of human leukocyte antigen gene using k-nearest neighbour classifier based on spectrum kernel. ScienceAsia 39, 42–49.
Shoombuatong, W., Prachayasittikul, V., Anuwongcharoen, N., Songtawee, N., Monnor, T., Prachayasittikul, S., et al., 2015a. Navigating the chemical space of dipeptidyl peptidase-4 inhibitors. Drug Des. Dev. Therapy 9, 4515.
Shoombuatong, W., Prachayasittikul, V., Prachayasittikul, V., Nantasenamat, C., 2015b. Prediction of aromatase inhibitory activity using the efficient linear method (ELM). EXCLI J. 14, 452.
Shoombuatong, W., Mekha, P., Chaijaruwanich, J., 2015c. Sequence based human leukocyte antigen gene prediction using informative physicochemical properties. Int. J. Data Mining Bioinform. 13 (3), 211–224.
Shoombuatong, W., Prathipati, P., Owasirikul, W., Worachartcheewan, A., Simeon, S., Anuwongcharoen, N., et al., 2017a. Towards the Revival of Interpretable QSAR Models. Advances in QSAR Modeling. Springer, pp. 3–55.
Shoombuatong, W., Prathipati, P., Prachayasittikul, V., Schaduangrat, N., Ahmad Malik, A., Pratiwi, R., et al., 2017b. Towards predicting the cytochrome P450 modulation: from QSAR to proteochemometric modeling. Curr. Drug Metabol. 18 (6), 540–555.
Simeon, S., Anuwongcharoen, N., Shoombuatong, W., Malik, A.A., Prachayasittikul, V., Wikberg, J.E., et al., 2016a. Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking. PeerJ 4, e2322.
Simeon, S., Shoombuatong, W., Anuwongcharoen, N., Preeyanon, L., Prachayasittikul, V., Wikberg, J.E., et al., 2016b. osFP: a web server for predicting the oligomeric states of fluorescent proteins. J. Cheminform. 8 (1), 72.
Svicher, V., Alteri, C., Artese, A., Zhang, J.M., Costa, G., Mercurio, F., et al., 2011. Identification and structural characterization of novel genetic elements in the HIV-1 V3 loop regulating coreceptor usage. Antivir. Ther. 16 (7), 1035–1045.
Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Trans. Neur. Netw. 10 (5), 988–999.
Vapnik, V., 2003. An Overview of Statistical Learning Theory, vol. 190. Nato Science Series Subseries III Computer and Systems Sciences, pp. 1–28.
Win, T.S., Malik, A.A., Prachayasittikul, V., Wikberg, J.E., S Nantasenamat, C., Shoombuatong, W., 2017. HemoPred: a web server for predicting the hemolytic activity of peptides. Future Med. Chem. 9 (3), 275–291.
Worachartcheewan, A., Shoombuatong, W., Pidetcha, P., Nopnithipat, W., Prachayasittikul, V., Nantasenamat, C., 2015a. Predicting metabolic syndrome using the random forest method. Scient. World J. 2015.
Worachartcheewan, A., Prachayasittikul, V., Anuwongcharoen, N., Shoombuatong, W., Prachayasittikul, V., Nantasenamat, C., 2015b. On the origins of hepatitis C virus NS5B polymerase inhibitory activity using machine learning approaches. Curr. Top. Med. Chem. 15 (18), 1814–1826.
Xiang, S.H., Finzi, A., Pacheco, B., Alexander, K., Yuan, W., Rizzuto, C., et al., 2010. A V3 loop-dependent gp120 element disrupted by CD4 binding stabilizes the human immunodeficiency virus envelope glycoprotein trimer. J. Virol. 84 (7), 3147–3161.
Xiao, N., Cao, D.-S., Zhu, M.-F., Xu, Q.-S., 2015. protr/ProtrWeb R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics 31 (11), 1857–1859.
Xu, S., Huang, X., Xu, H., Zhang, C., 2007. Improved prediction of coreceptor usage and phenotype of HIV-1 based on combined features of V3 loop sequence using random forest. J. Microbiol. 45 (5), 441–446.
Zhu, T., Mo, H., Wang, N., Nam, D.S., Cao, Y., Koup, R.A., et al., 1993. Genotypic and phenotypic characterization of HIV-1 patients with primary infection. Science 261 (5125), 1179–1181.

Bioinformatics, 2020, 1–7
doi: 10.1093/bioinformatics/btaa160
Advance Access Publication Date: 7 March 2020
Original Paper

OXFORD

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/doi/10.1093/bioinformatics/btaa160/5799076 by Ajou University Medical Library user on 16 April 2020

Sequence analysis

HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation

Md. Mehedi Hasan[1,2], Nalini Schaduangrat[3], Shaherin Basith[4], Gwang Lee[4,5,*], Watshara Shoombuatong[3,*] and Balachandran Manavalan[4,*]

[1]Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan, [2]Japan Society for the Promotion of Science, Chiyoda-ku, Tokyo 102-0083, Japan, [3]Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand, [4]Department of Physiology, Ajou University School of Medicine, Suwon 16499, Republic of Korea and [5]Department of Molecular Science and Technology, Ajou University, Suwon 16499, Republic of Korea

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson

Received on January 13, 2020; revised on February 19, 2020; editorial decision on March 2, 2020; accepted on March 3, 2020

Abstract

Motivation: Therapeutic peptides failing at clinical trials could be attributed to their toxicity profiles like hemolytic activity, which hamper further progress of peptides as drug candidates. The accurate prediction of hemolytic peptides (HLPs) and its activity from the given peptides is one of the challenging tasks in immunoinformatics, which is essential for drug development and basic research. Although there are a few computational methods that have been proposed for this aspect, none of them are able to identify HLPs and their activities simultaneously.

Results: In this study, we proposed a two-layer prediction framework, called HLPpred-Fuse, that can accurately and automatically predict both hemolytic peptides (HLPs or non-HLPs) as well as HLPs activity (high and low). More specifically, feature representation learning scheme was utilized to generate 54 probabilistic features by integrating six different machine learning classifiers and nine different sequence-based encodings. Consequently, the 54 probabilistic features were fused to provide sufficiently converged sequence information which was used as an input to extremely randomized tree for the development of two final prediction models which independently identify HLP and its activity. Performance comparisons over empirical cross-validation analysis, independent test and case study against state-of-the-art methods demonstrate that HLPpred-Fuse consistently outperformed these methods in the identification of hemolytic activity.

Availability and implementation: For the convenience of experimental scientists, a web-based tool has been established at http://thegleelab.org/HLPpred-Fuse.

Contact: glee@ajou.ac.kr or watshara.sho@mahidol.ac.th or bala@ajou.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In the last few decades, peptide therapeutics has gained much interest due to their perceived properties of safety, efficacy, high selectivity coupled with good tolerability and attractive pharmacological profile (Fosgerau and Hoffmann, 2015; Lau and Dunn, 2018; Vlieghe *et al.*, 2010). However, their short half-life and limited oral bioavailability make the development of effective peptide drugs that reach the market extremely challenging (Gautam *et al.*, 2014). To date, over 60 peptide drugs have been approved in Europe, Japan and the USA with over 150 in active clinical development, and an additional 260 have been tested in human clinical trials (Lau and Dunn, 2018).

So far, the foremost reason for a number of potential therapeutic peptides failing at clinical trials could be attributed to their toxicity profiles (Eddershaw *et al.*, 2000). Particularly, the undesirable toxicity arising through their hemolytic activity toward eukaryotic cells (Gautam *et al.*, 2014). Previously, drug-induced hemolysis has been considered as an undesirable drawback arising from chemotherapeutic bioactivities, further leading to reduction in the antimicrobial activity (Jeswani *et al.*, 2015). The term hemolysis indicates the immature destruction of red blood cells (before reaching their expected lifespan of 120 days) (Win *et al.*, 2017). The consequence of such action could be seen in the reduction of hemoglobin levels in the blood which further exacerbates its decreased oxygen-carrying capacity.

© The Author(s) 2020. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Thus, owing to the aforementioned ramifications, the hemolytic activity of peptides is an undesirable property for constructing therapeutic peptides and the timely identification of such hemolytic activity is deemed to be an essential step in the development of peptide-based drugs.

Over the past decades, a plethora of articles describing therapeutic peptides (i.e. cell-penetrating peptides, antihypertensive, anticancer, antimicrobial and antiviral peptides) and their hemolytic potential (i.e. the hemolytic property of peptides under consideration as potential therapeutic peptides) have been published (Manavalan *et al.*, 2017, 2018, 2019b, c; Rao *et al.*, 2019; Schaduangrat *et al.*, 2019a, b; Wei *et al.*, 2018). Presently, there is much scarcity in information pertaining to the design of therapeutic peptides without the exposition of its hemolytic activity. A systematic analysis using computational approaches to discern hemolytic peptides (HLP) and non-hemolytic peptides (non-HLP) is therefore pertinent for demarcating the peptide features responsible for low/high hemolytic activity, which could be taken into account while designing therapeutic peptides.

Accurate identification of hemolytic activities of peptides is of great importance for the drug development and basic research. Due to the limitation of experimental approaches (i.e. time-consuming, laborious and expensive), it is highly desirable to develop a computational model which can discriminate HLPs from non-HLPs. In fact, only two computational models, i.e. HemoPI (Chaudhary *et al.*, 2016) and HemoPred (Win *et al.*, 2017), have previously sought to unravel this aspect. Chaudhary *et al.* (2016) proposed the first computational model, named HemoPI, for predicting hemolytic activities of peptides by employing three datasets (HemePI-1, HemePI-2 and HemePI-3). HemoPI was developed using support vector machine (SVM) in conjunction with motif information and position-specific compositional information. In 2017, Win *et al.* (2017) developed the second method, called HemoPred, by using random forest (RF) with hybrid features [a linear combination of amino acid composition (AAC) and dipeptide composition (DPC)]. Although the aforementioned methods stimulated discovery of potential HLPs, the existing methods may need further improvement due to the use of single machine learning (ML) classifier and lack of more informative features. Moreover, none of the abovementioned methods can automatically predict hemolytic peptides (HLPs or non-HLPs) and HLPs activity (high or low) simultaneously from the given peptide sequences. Hence, their practical usage is quite limited.

Motivated by these considerations, we thus developed the first two-layer approach, called HLPpred-Fuse, which is highly efficient and reliable in predicting both hemolytic peptides (HLPs or non-HLPs) and their activity (high and low). In brief, HLPpred-Fuse could be established according to the following steps: (i) feature representation scheme was employed to generate 54 probabilistic features derived from 54 prediction models that utilized nine different types of peptide features, i.e. AAC, DPC, amino acid index (AAI), binary profile (BPF), composition–transition–distribution (CTD), conjoint triad (CTF), quasi-sequence order (QSO), grouped dipeptide composition (GDPC), grouped tripeptide composition (GTPC) and six ML classifiers, i.e. SVM (Vapnik, 2013), RF (Breiman, 2001), gradient boosting (GB) (Friedman, 2002), extremely randomized tree (ERT) (Geurts *et al.*, 2006), k-nearest neighbor (KNN) (Weinberger *et al.*, 2006) and AdaBoost (AB) (Freund and Schapire, 1996); and (ii) 54 probabilistic features were fused and used as an input for ERT to develop a final predictor separately for both layers. The systematic workflow for the prediction of HLPs and their activities is depicted in Figure 1. Basically, HLPpred-Fuse is composed of a two-layer framework, where the first layer is used for predicting whether a given sequence is HLP or non-HLP, while the second layer is used for identifying the predicted HLPs activity. Cross-validation (CV) analysis, independent evaluation and case study showed that our proposed HLPpred-Fuse outperformed existing predictors in identifying HLPs and HLPs activity, respectively. To the best of our knowledge, this is the first study that employs diverse sequence-based information and multiple classifiers for constructing the two-layer framework in HLP prediction, which is potentially useful for assisting HLP research.

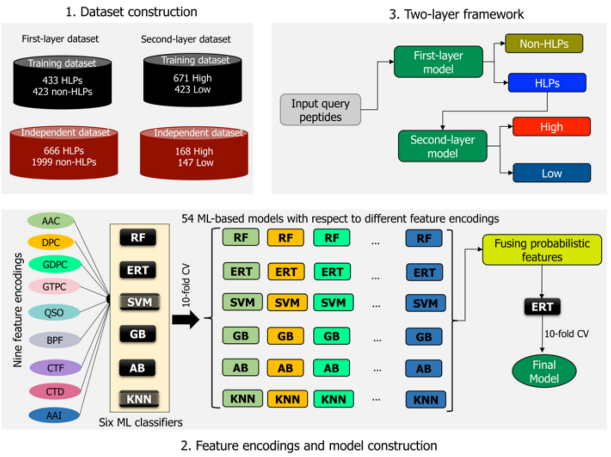

Fig. 1. Development and flowchart of HLPpred-Fuse. It involves the following steps: (i) two separate datasets were constructed for the first- and second-layer prediction model development; (ii) nine feature encodings were utilized and used as an input for six different classifiers resulting in their corresponding prediction models. Subsequently, the predicted probability of positives was fused and considered as the fused features and inputted to ERT for development of the final prediction model; (iii) construction of a two-layer framework

2 Materials and methods

2.1 Dataset construction

The previous two methods, HemoPI (Chaudhary *et al.*, 2016) and HemoPred (Win *et al.*, 2017), were trained and evaluated using the same dataset, which was constructed by Chaudhary *et al.* (2016). Basically, the authors have constructed three datasets, namely HemoPI-1 (Dataset_1), HemoPI-2 (Dataset_2) and HemoPI-3 (Dataset_3) whereby the first two datasets contained the same positive samples, but different negative samples. Specifically, Dataset_1 and Dataset_2, respectively considered random peptides and low hemolytic activity peptides as negative samples. However, Dataset_3 comprises of high and low HLPs as positive and negative samples. In this study, we employed the same datasets and excluded <6 amino acid residue peptides due to the difficulty in capturing meaningful sequence information from shorter peptides. The number of positive and negative samples of each dataset in training and independent datasets is shown in Supplementary Table S1. Here, Dataset_1 and Dataset_3, respectively were used in the development of first- and second-layer prediction models among the three datasets. It should be noted that for Dataset_1, the independent positive sample was supplemented with our own data collection from the Hemolytik database (Gautam *et al.*, 2014) and the negative samples were replaced with independent negative samples reported in PEPred-SUITE method (Wei *et al.*, 2019).

2.2 Feature encoding schemes

In this study, we used nine different feature encoding schemes to encode the peptide sequences into feature vector with fixed length (Chen *et al.*, 2018, 2019), which included AAC, AAI, BPF, DPC, CTD, CTF, QSO, GDPC and GTPC. In brief, each encoding definition is explained in the following subsections:

2.2.1 AAC

AAC computes frequency of all 20 natural amino acid type in a peptide sequence. AAC can be computed as follows:

$$f(a) = \frac{M(a)}{L}, \quad a \in \{A,\ C,\ D,\ E,\ F \ldots, Y\} \tag{1}$$

where M(a) is the total number of amino acid type a, while L is the length of the peptide.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/doi/10.1093/bioinformatics/btaa160/5799076 by Ajou University Medical Library user on 16 April 2020

2.2.2 DPC

DPC provides 400 feature vectors and it can be computed as follows:

$$f(a,\ b) = \frac{M_{ab}}{L-1},\ a,\ b \in \{A,\ C,\ D,\ E,\ F \ldots, Y\} \tag{2}$$

where M_{ab} is the total number of dipeptides denoted by amino acid types a and b.

2.2.3 GDPC

GDPC is another variation of DPC encoding. In GDPC encoding, 20 standard amino acids are classified into five groups according to their physicochemical properties (PCPs) (Lee *et al.*, 2011). The five groups include aliphatic (g1: IMLVAG), aromatic (g2: WYF), positive charge (g3: HRK), negative charge (g4: ED) and uncharged (g5: QNPCTS). GDPC provides 25 descriptors which can be computed as follows:

$$f(a,\ b) = \frac{M_{ab}}{L-1},\ a,\ b \in \{g1,\ g2,\ g3,\ g4,\ g5\} \tag{3}$$

where M_{ab} is the total number of dipeptides denoted by amino acid types a and b

2.2.4 GTPC

GTPC provides 125 descriptors that can be computed as follows:

$$f(a,\ b,\ c) = \frac{M_{abc}}{L-1},\ a,\ b,\ c \in \{g1,\ g2,\ g3,\ g4,\ g5\} \tag{4}$$

where M_{abc} is the total number of tripeptides denoted by amino acid types a, b and c.

2.2.5 BPF

In BPF encoding, each amino acid residues are represented by a 20-dimensional binary vector. As we are dealing with diverse length, only five residues at both terminus (N- and C-) were considered that provides 200 vectors.

2.2.6 AAI

Saha *et al.* (2012) identified eight superior AAIs (accessions TSAJ990101, MIYS990104, LIFS790101, MAXF760101, CEDJ970104, BIOV880101, NAKH920108 and BLAM930101) based on clustering technique from the original 566 AAIs. Here, we used those indexes and generated 160 descriptors (8 indexes × 20 amino acids).

2.2.7 QSO

QSO generates 100-dimensional vectors that considers sequence order effect, which is dependent on the PCP distance among amino acids (Chou, 2000). It can be computed as follows:

$$\begin{cases} X_a = \dfrac{N_a}{\sum_{a=1}^{20} N_a + w\sum_{d=1}^{30}\tau_d},\ a = 1,2,3,\ldots,20 \\ X_d = \dfrac{w\tau_{d-20}}{\sum_{a=1}^{20} N_a + w\sum_{d=1}^{30}\tau_d},\ d = 21,22,\ldots,50 \\ \tau_d = \sum_{j=1}^{L-d}(dist_{j,\ j+d})^2,\ d = 1,2,3,\ldots,30 \end{cases} \tag{5}$$

where N_a represents normalized frequency of amino acid a. $dist_{j,j+\mathrm{d}}$ represents the distance among two amino acids at position j and $j+\mathrm{d}$. Notably, chemical distance matrix (Grantham, 1974) and Schneider–Wrede physicochemical distance matrix (Schneider and Wrede, 1994) were used to compute the distance. w represents weight factor with a value of 0.1.

2.2.8 CTD

CTD encodes amino acid PCP distribution patterns. As shown in Supplementary Table S2, 20 standard amino acids are grouped into three classes according to seven PCPs. Composition (C) is represented as a 21-dimensional vector, which can be computed as follows:

$$C_a = \frac{M_a}{L},\ a = 1,2,3 \tag{6}$$

where M_a represents total number of amino acid type (class) a.

Transition (T) encoded as 21-dimensional vector, which represents the occurrence with which a type a residue is followed by type b residue or vice versa. It can be calculated as follows:

$$\begin{cases} T_{ab} = \dfrac{M_{ab} + M_{ba}}{L-1} \\ T_{bc} = \dfrac{M_{bc} + M_{cb}}{L-1} \\ T_{ac} = \dfrac{M_{ac} + M_{ca}}{L-1} \end{cases} \tag{7}$$

where M_{ab} represents the number of dipeptide ab in the sequence.

Distribution (D) descriptors consists of five values for the three classes based on the fractions of the entire sequence, where the first residue of a given group is located, and where 25, 50, 75 and 100% of incidents are contained. D encoded as 105-dimensional vector. In total, CTD encoded as 147-dimensional feature vector.

2.2.9 CTF

CTF encoding considers the properties of one amino acid and its neighboring amino acids by regarding any three continuous amino acids as a single component (triad). Triads comprised of three amino acids belonging to the same group are treated as identical. Here, 20 standard amino acids categorized into seven groups: {*A*, *V*, *G*}, {*F*, *L*, *I*, *P*}, {*T*, *S*, *Y*, *M*}, {*H*, *Q*, *N*, *W*}, {*R*, *K*}, {*D*, *E*}, {*C*} according to their PCPs. Finally, CTF considers all sets of triads within a given sequence and count their frequencies, which gives 343 ($7 \times 7 \times 7$) dimensional vector.

2.3 Machine learning classifiers

In the point of ML, the prediction performance not only depend on feature encodings but also choice of the classifiers. In this study, six different ML classifiers were employed, namely SVM (Vapnik, 2013), RF (Breiman, 2001), GB (Friedman, 2002), ERT (Geurts *et al.*, 2006), KNN (Weinberger *et al.*, 2006) and AB (Freund and Schapire, 1996). All the classification models were developed using Scikit-learn version 0.18 (Pedregosa *et al.*, 2011). A comprehensive description of how each classifier was implemented in the current study has been defined previously (Basith *et al.*, 2019; Charoenkwan *et al.*, 2019; Laengsri *et al.*, 2019a, b; Manavalan *et al.*, 2018, 2019a–d; Schaduangrat *et al.*, 2019a,b; Shoombuatong *et al.*, 2019). In brief, grid search was employed to optimize hyper-parameters for each classifier, whose search range is presented in Supplementary Table S3. In this study, we carried out ten randomized 10-fold CV. A detailed description of CV has been elaborated in previous studies (Chen *et al.*, 2019; Hasan *et al.*, 2019; Lv *et al.*, 2019; Yang *et al.*, 2019). For each 10-fold CV, we searched different combinations of hyper-parameters for each ML classifier separately and identified their best hyper-parameter based on the highest accuracy (ACC). Finally, we obtained 10 best hyper-parameters due to 10 randomized 10-fold CV, where the median hyper-parameter was selected, and their corresponding model was developed and considered as the final prediction model.

Furthermore, we applied five commonly used measures (Basith *et al.*, 2020; Schaduangrat *et al.*, 2019a,b; Shoombuatong *et al.*, 2019; Su *et al.*, 2019), including specificity (Sp), sensitivity (Sn), ACC, balanced accuracy (BACC) and Matthew's correlation coefficient (MCC) to quantify the performance of the proposed model and compared it with the existing predictors. Please refer to Supplementary Information for metrics definitions.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/doi/10.1093/bioinformatics/btaa160/5799076 by Ajou University Medical Library user on 16 April 2020

3 Results and discussion

3.1 Construction of the first-layer model to predict hemolytic peptide

We utilized nine different sequence-based encodings, including AAC, DPC, GPDC, GTPC, BPF, AAI, QSO, CTD and CTF, and assessed its ability to predict HLPs or non-HLPs by integrating them with six different conventional ML classifiers (SVM, RF, GB, ERT, KNN and AB). In total, we generated 54 single feature (SF)-based models and their prediction results are listed in Figure 2A and Supplementary Table S4. Performance comparisons showed that SVM, RF, GB, ERT, KNN and AB models yielded ACC values in the ranges of 91.0–95.9%, 92.5–97.0%, 92.6–97.6%, 93.2–97.6%, 88.9–95.4% and 90.0–95.2%, respectively, with respect to the nine different encodings. As shown in Figure 2A and Supplementary Table S4, it could be noted that four out of six classifiers consisting of ERT, RF, GB and AB models utilizing QSO encoding achieved an ACC value in the range of 95.2–97.6%, while the remaining two classifiers (SVM and KNN models) using AAI encoding gave a comparable ACC value in the range of 95.4–95.9%. This observation indicates that QSO and AAI encodings have a slightly upper hand in the prediction of HLP when compared with other encodings. Based on the best model for each classifier from Figure 2A, we observed that RF, ERT and GB achieved a similar performance, thus indicating that these classifiers are remarkably better than other three classifiers (97.0–97.6% versus 95.2–95.9%).

Furthermore, we examined remaining models and noted that these models also achieved an excellent performance with ACC values in the range of 90.0–96.9, indicating their potential in discriminating HLPs from non-HLPs. Instead of making an effort to determine the best model among 54 SF-based models, we fused their advantages into a single model. Practically, there are different ways to integrate multiple prediction models, such as ensemble method (Manavalan *et al.*, 2019a; Wang *et al.*, 2018, 2019; Zhang *et al.*, 2020) and meta-predictor (Boopathi *et al.*, 2019; Manavalan *et al.*, 2019d; Qiang *et al.*, 2018; Schaduangrat *et al.*, 2019a,b; Wei *et al.*, 2018). Herein, we employed a meta-predictor approach to develop the final model. Specifically, the predicted probability from the abovementioned 54 SF-based models (9 encodings × 6 ML classifiers) were concatenated and considered as new 54-dimensional (54D) feature vector representing multiview information. As seen in Supplementary Table S4, since the average ACC value derived from ERT models from nine encodings is better than other classifiers, we inputted 54D vector to ERT classifier and developed the final predictor called HLPred-Fuse. It should be noted that we have also carried out the feature optimization procedure on the 54D feature vector using F-score and sequential feature search as employed in previous studies (Boopathi *et al.*, 2019; Zhang *et al.*, 2019). However, the prediction result after the feature optimization procedure did not provide any improvements in the performance. Hence, in this study, we considered the original 54D feature vector as the final feature vector.

3.2 Comparison of HLPpred-Fuse with SF-based models and existing predictors

The prediction result showed that the proposed HLPpred-Fuse achieved MCC, ACC, Sn and Sp of 0.967, 98.4%. 99.1% and 97.7%, respectively. In order to determine whether our approach could improve the performance and outperform the existing methods, we first compared the results of HLPpred-Fuse with the best models from 54 SF-based models as depicted in Figure 2B. As noticed in Figure 2B, HLPpred-Fuse outperformed ERT–QSO and GB–QSO models with improvements of 3.1% and 2.6% on Sp and Sn, respectively. Overall, HLPpred-Fuse achieved a more balanced performance ($|Sn-Sp| = 0.7$) when compared with the above two models ($|Sn-Sp| > 3.0$), thereby demonstrating that fusing multiview information is essential for more balanced and improved prediction. Secondly, we compared HLPpred-Fuse with the state-of-the-art predictors, including HemoPI (Chaudhary *et al.*, 2016) and HemoPred (Win *et al.*, 2017). Previously reported predictors were also trained with the same training dataset as employed in this study. Specifically, MCC, Sn, Sp and ACC of HLPpred-Fuse are 4.6–7.7%, 2.4–5.3% 2.6–2.9% and 2.5–4.1% higher than previous predictors as shown in Figure 2C, further indicating that HLPpred-Fuse has the capability to identify HLPs more accurately than the existing predictors.

3.3 Performance evaluation of hemolytic peptide prediction methods based on the independent dataset

To check the robustness of HLPpred-Fuse, we performed an independent evaluation with test data set containing 663 HLPs and 1999 non-HLPs. Consequently, HLPpred-Fuse was considered to compare its performance with the abovementioned 54 SF-based predictors and existing predictors i.e. HemoPI and HemoPred. Particularly, HemoPI has two prediction models that uses different feature encodings which were included in our comparison analysis. Due to an imbalanced independent dataset, BACC was utilized to assess the predictive ability of the model (Basith *et al.*, 2020). Results showed that HLPpred-Fuse achieved MCC, BACC, Sn and Sp of 0.823, 90.5%, 84.5% and 96.4%, respectively indicating that the proposed predictor outperforms all 54 SF-based models by providing more than 2.0% increase in BACC value and 4.1% in MCC value as summarized in Supplementary Table S5 and Figure 3A.

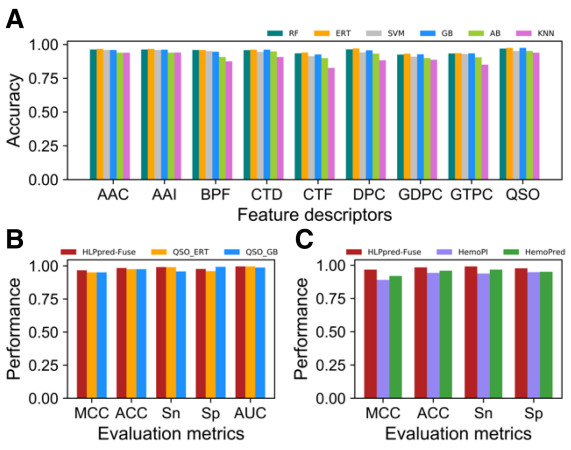

Fig. 2. First-layer prediction performances based on Dataset_1 training dataset. (**A**) Performances of six different classifiers utilizing nine different encodings are considered as single feature-based models. (**B**) Performance of HLPpred-Fuse with top two single feature-based models. (**C**) Performances comparison of HLPpred-Fuse with the state-of-the art predictors

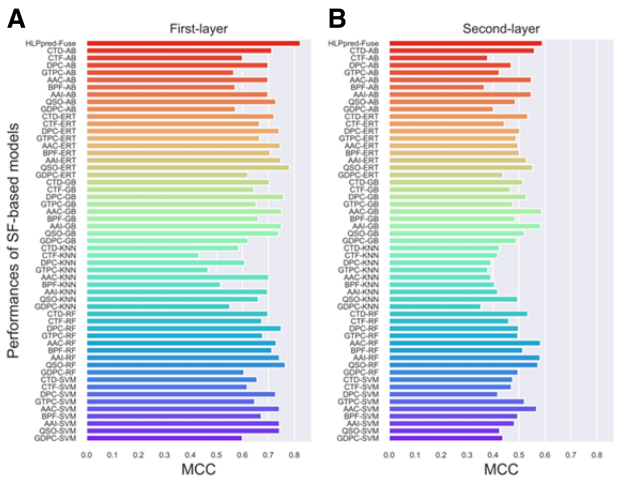

Fig. 3. Performances on independent evaluation. (**A**) Performances of HLPpred-Fuse along with the single feature-based models in classifying HLPs from non-HLPs. (**B**) Performances of HLPpred-Fuse along with the single feature-based models on classifying high from low HLP activities

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/doi/10.1093/bioinformatics/btaa160/5799076 by Ajou University Medical Library user on 16 April 2020

Table 1. Comparison of HLPpred-Fuse with the state-of-the-art predictors on Dataset_1 independent dataset

Methods	MCC	BACC	Sn	Sp	AUC	*P*-value
HLPpred-Fuse	0.823	0.905	0.845	0.964	0.967	—
HemoPI (Model_1)	0.754	0.873	0.804	0.943	0.952	0.040
HemoPI (Model_2)	0.727	0.867	0.807	0.927	0.943	0.003
HemoPred[a]	0.340	0.709	0.652	0.728	NA	NA

[a]HemoPred provides only class label not predicted probability value. Therefore, AUC value cannot be computed and it is mentioned as not available (NA). Area under ROC curve is provided in Supplementary Figure S1A.

Table 2. Results of the proposed predictor and the state-of-the-art predictors on Dataset_3 independent dataset

Methods	MCC	BACC	Sn	Sp	AUC[a]	*P*-value
HLPpred-Fuse	0.585	0.792	0.821	0.762	0.869	—
HemoPI	0.577	0.780	0.887	0.673	0.842	0.404

[a]Area under ROC curve is provided in Supplementary Figure S1B.

Notably, a similar comparison was also performed for the second layer (Fig. 3B) which is discussed later. Interestingly, HLPpred-Fuse showed an improvement of ∼1.0 (in terms of ACC) during CV; however, BACC was significantly improved during independent test, thus suggesting that fusing multiview information is crucial to achieve consistent performance. Furthermore, performance comparison of HLPpred-Fuse with the existing predictors revealed that HLPpred-Fuse achieved great improvements with all five metrics (Table 1). Remarkably, MCC, ACC, Sn and Sp of the proposed HLPpred-Fuse are 6.9–48.2%, 3.2–19.6%, 4.1–19.2% and 2.1–23.6% higher than the existing predictors. Furthermore, the significant differences calculated between HLPpred-Fuse and other methods by using two-tailed *t*-test (Hanley and McNeil, 1982) showed that HLPpred-Fuse significantly outperformed the existing methods with a P-value < 0.05, thus indicating that HLPpred-Fuse is more robust in discriminating real HLPs from non-HLPs than the existing predictors.

3.4 Construction of the second-layer model to predict hemolytic activity and its comparison with the existing predictors

In order to predict the activity of HLP (high or low), the same procedure of the first-layer model was applied on Dataset_3 to develop the second-layer prediction. Based on the results from Figure 4A and Supplementary Table S6, SVM, RF, GB, ERT, KNN and AB models provide ACC values in the ranges of 73.2–77.9%, 76.5–79.9%, 75.9–80.4%, 76.3–80.6%, 72.0–77.7% and 70.5–75.4%, respectively with respect to the nine different encodings. From Figure 4A and Supplementary Table S6, several observations can be made as follows: firstly, we found that the lowest ACC of the first-layer model was 90%, while the best model of the second layer yielded an ACC of 80.6%. Secondly, there are three classifiers (ERT, RF and SVM models) using AAI encoding which achieved ACC in the range of 78.0–80.6%, whereas two classifiers (AB and KNN models) using DPC encoding achieved ACC in the range of 75.5–77.7%, and GB achieved ACC of 80.4% using GTPC encoding. This observation suggests that AAI, GTPC and DPC are slightly more beneficial to the activity of HLPs as compared to other encodings. Based on the best model for each classifier from Figure 4A, we observed that RF, ERT and GB models achieved similar ACC in the range of 79.9–80.6%, thus suggesting that these classifiers are remarkably better when compared with the other three classifiers (75.5–78.0% ACC). Instead of selecting the best model from Figure 4A, the predicted probability as high activity of HLPs was generated and concatenated using the above-mentioned 54 SF-based models and considered as

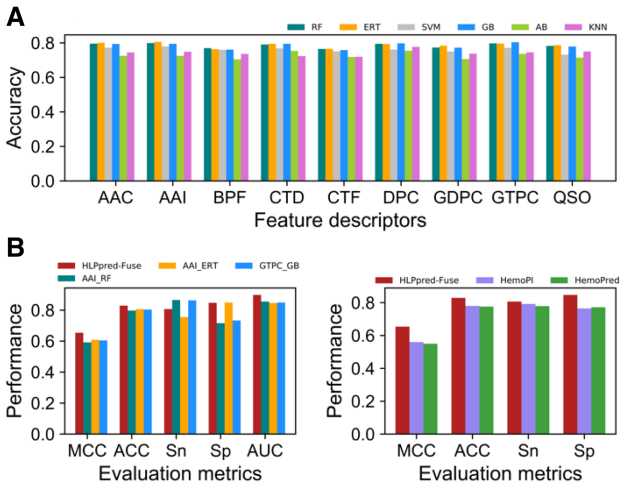

Fig. 4. Second-layer (activity) prediction performances based on Dataset_3 training dataset. (**A**) Performances of six different classifiers utilizing nine different encodings are considered as single feature-based models. (**B**) Performances of HLPpred-Fuse with top three single feature-based models. (**C**) Performance comparison of HLPpred-Fuse with the state-of-the-art predictors

the new 54D feature vector used in this aspect. Subsequently, the 54D feature vector was used as an input in the ERT model for developing the final predictor. As seen in Figure 4B, HLPpred-Fuse yielded MCC, ACC, Sn, Sp and AUC of 0.654, 82.9%, 80.7%, 84.7% and 0.898, respectively. In our performance comparison, HLPpred-Fuse was compared with both 54 SF-based models and the existing predictors, HemoPI (Chaudhary *et al.*, 2016) and HemoPred (Win *et al.*, 2017) as described below.

Firstly, we compared HLPpred-Fuse with 54 SF-based models and the performance comparison showed that the prediction results of HLPpred-Fuse were approximately 4.5–6.0% MCC and 2.25–2.90% ACC higher than those of the three best SF-based models. Our statistical comparative analysis indicated that HLPpred-Fuse significantly outperformed the 54 SF-based models at the level of P-value < 0.05. Secondly, we further compared HLPpred-Fuse with the existing predictors, HemoPI (Chaudhary *et al.*, 2016) and HemoPred (Win *et al.*, 2017). To objectively evaluate the performance of the proposed HLPpred-Fuse and fairly compare it with the existing predictors, the same datasets that have been used in the previous studies were used herein. We noted that MCC, Sn, Sp and ACC of HLPpred-Fuse are 9.4–10.4%, 1.5–2.8%, 7.5–8.2% and 4.9–5.3% higher than previous predictors Figure 4C, further emphasizing that HLPpred-Fuse has the capability to identify hemolytic activity more accurately than the existing predictors.

3.5 Performance evaluation of hemolytic activity prediction methods based on the independent dataset

To check the robustness of the proposed method, HLPpred-Fuse was performed on the independent test dataset containing 168 high activity and 148 low activity peptides and its performances were compared with the SF-based predictors and the existing predictors. The results showed that HLPpred-Fuse yielded MCC, BACC, Sn and Sp of 0.585, 79.2%, 82.1% and 76.2%, respectively. Performance comparison between HLPpred-Fuse with the SF-based models showed that HLPpred-Fuse achieved better performance than 54 SF-based models with improvements of >0.80% and >0.20% on BACC and MCC values, respectively, as depicted in Figure 3B and Supplementary Table S7. Since HemoPred could not provide the prediction result on the activity of HLPs, only HemoPI was used to compare with our proposed HLPpred-Fuse. As shown in Table 2, improvements of 0.80% and 1.20% for MCC and ACC, respectively were observed for HLPpred-Fuse when compared with HemoPI. In summary, it is evident that HLPpred-Fuse has better discriminative

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/doi/10.1093/bioinformatics/btaa160/5799076 by Ajou University Medical Library user on 16 April 2020

power on the activity of peptides (high and low) than the existing predictor.

It is worth mentioning that the above protocol was applied on Dataset_2 and their corresponding model was developed and evaluated on an independent dataset. Result shows that our framework significantly outperformed the existing predictors both on training (Supplementary Table S8) and independent datasets (Supplementary Table S9).

3.6 Case study

We established an additional case study based on experimentally validated 157 non-HLPs (negative samples) for further validation of our proposed method to correctly discriminate non-HLPs from HLPs. Notably, these samples are linear peptides with no modifications which were extracted from the Hemolytik database (Gautam *et al.*, 2014). Practically, this type of analysis acts as a checkpoint to verify whether the prediction programs accurately identify negatives or decrease the number of false positives, which will be more crucial for reducing experimental efforts. Notably, HLPpred-Fuse is the first bioinformatics tool based on a two-layer model that can filter out non-HLPs and then determine the activity of HLPs simultaneously. In our experiments, HLPpred-Fuse can correctly identify 139 cases. While using only second-layer model to check whether the actual negatives were predicted as high or low activity, HLPpred-Fuse can correctly identify 134 cases, while HemoPI can correctly predict 115 cases. It should be noted that Hemopred was excluded from this analysis due to the usage of only first-layer model in their webserver. The above analysis clearly demonstrated the practical applicability of HLPpred-Fuse over other methods.

3.7 Feature representation in HLPpred-Fuse

To understand the feature distribution, we adopted UMAP plots that reduces a given actual feature representation into a two-dimensional space. In the first layer, we selected only the top five feature encodings (QSO, DPC, AAI, AAC and CTD) having the highest ACC values resulting from the ERT classifier and compared them with the probabilistic features. As seen in Supplementary Figure S2A–E, we observed that the selected top five feature encodings were able to capture partial discrimination between HLPs and non-HLPs. However, two separate clusters of HLPs and non-HLPs (Supplementary Fig. S2F) resulted while using probabilistic features. This might explain why the proposed HLPpred-Fuse outperformed other conventional models. Unlike the result of the first layer, the top five feature encodings from the second layer shows a clear inhomogeneity in the distribution of high and low activity within the UMAP plot space as shown in Supplementary Figure S3A–E. Meanwhile, probabilistic features appear to be more clustered with few overlaps when compared with other encodings (Supplementary Fig. S3F), again indicating the necessity of reduced information which has high capability to discriminate high and low activity peptides. Overall, extracting model-generated information (based on multiple feature encodings and classifiers) have a high capability to capture the differences between two classes, when compared with the actual feature dimensions of each encoding. Consequently, HLPpred-Fuse showed a superior performance over 54 SF-based models as implemented in this study.

4 Conclusion

Accurate identification of the hemolytic activity of peptides may help to expedite peptide-based drug discovery in the search for novel and efficient therapeutic peptides. Therefore, in this study, we proposed a two-layer prediction framework, called HLPpred-Fuse, for improving the prediction of both HLPs and its activity from unknown peptides. To build an efficient prediction model, a novel feature representation learning scheme based on six different ML algorithms in conjunction with nine feature encodings covering comprehensive sequence information, was applied to generate 54 informative probabilistic features. Subsequently, these multiview probabilistic features were fused to construct the final prediction model. Rigorous CV, independent test and case study demonstrated that HLPpred-Fuse significantly outperformed existing methods and other conventional ML algorithms, thus indicating that the proposed model is more effective and promising for the prediction of the hemolytic activity of peptides. It is anticipated that our developed predictor, HLPpred-Fuse will serve as a useful, high-throughput and cost-effective tool for large-scale analysis of therapeutic peptides and also in the timely identification of HLPs. Furthermore, it would be expected that integrating other feature encodings (Chen *et al.*, 2018, 2019) and ML algorithms (Cao *et al.*, 2017; Chan *et al.*, 2019; Conover *et al.*, 2019; Hou *et al.*, 2019a, b) might further improve the performance.

Acknowledgements

The authors would like to thank Dr Rajiv Gandhi Govindaraj for his critical comments during manuscript preparation.

Funding

This work is supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Science and ICT (MSIT) [2018R1D1A1B07049572, 2019R1I1A1A01062260 and 2016M3C7A1904392], a grant from the Korea Basic Science Institute (KBSI) National Research Facilities & Equipment Center (NFEC) funded by the Korea government (Ministry of Education) [2019R1A6C1010003] and the TRF Research Grant for New Scholar [MRG6180226].

Conflict of Interest: none declared.

References

Basith,S. *et al.* (2019) SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Mol. Ther. Nucleic Acids*, **18**, 131–141.

Basith,S. *et al.* (2020) Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. *Med. Res. Rev.* doi: 10.1002/med.21658.

Boopathi,V. *et al.* (2019) mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides. *Int. J. Mol. Sci.*, **20**. doi: 10.3390/ijms20081964.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Cao,R. *et al.* (2017) QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. *Bioinformatics*, **33**, 586–588.

Chan,L. *et al.* (2019) Survey of AI in cybersecurity for information technology management. In: *2019 IEEE Technology & Engineering Management Conference (TEMSCON)*. IEEE, pp. 1–8.

Charoenkwan,P. *et al.* (2019) iQSP: a sequence-based tool for the prediction and analysis of quorum sensing peptides via Chou's 5-steps rule and informative physicochemical properties. *Int. J. Mol. Sci.*, **21**. doi: 10.3390/ijms21010075.

Chaudhary,K. *et al.* (2016) A web server and mobile app for computing hemolytic potency of peptides. *Sci. Rep.*, **6**, 22843.

Chen,W. *et al.* (2019) iRNA-m7G: identifying N7-methylguanosine sites by fusing multiple features. *Mol. Ther. Nucleic Acids*, **18**, 269–274.

Chen,Z. *et al.* (2018) iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, **34**, 2499–2502.

Chen,Z. *et al.* (2019) iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief. Bioinform.*, **10**. doi: 10.1093/bib/bbz041.

Chou,K.C. (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun.*, **278**, 477–483.

Conover,M. *et al.* (2019) AngularQA: protein model quality assessment with LSTM networks. *Comput. Math. Biophys.*, **7**, 1–9.

Eddershaw,P.J. *et al.* (2000) ADME/PK as part of a rational approach to drug discovery. *Drug Discov. Today*, **5**, 409–414.

Fosgerau,K. and Hoffmann,T. (2015) Peptide therapeutics: current status and future directions. *Drug Discov. Today*, **20**, 122–128.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/doi/10.1093/bioinformatics/btaa160/5799076 by Ajou University Medical Library user on 16 April 2020

Freund,Y. and Schapire,R.E. (1996) Experiments with a new boosting algorithm. In: *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, Icml. Citeseer. pp. 148–156.

Friedman,J.H. (2002) Stochastic gradient boosting. *Comput. Stat. Data Anal.*, **38**, 367–378.

Gautam,A. *et al.* (2014) Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides. *Nucleic Acids Res.*, **42**, D444–D449.

Geurts,P. *et al.* (2006) Extremely randomized trees. *Mach. Learn.*, **63**, 3–42.

Grantham,R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862–864.

Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Hasan,M.M. *et al.* (2019) Prediction of S-nitrosylation sites by integrating support vector machine and random forest. *Mol. Omics*, **15**, 451–458.

Hou,J. *et al.* (2019a) Deep convolutional neural networks for predicting the quality of single protein structural models. *bioRxiv* 590620.

Hou,J. *et al.* (2019b) Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. *Proteins*, **87**, 1165–1178.

Jeswani,G. *et al.* (2015) Recent approaches for reducing hemolytic activity of chemotherapeutic agents. *J. Control Release*, **211**, 10–21.

Laengsri,V. *et al.* (2019a) TargetAntiAngio: a sequence-based tool for the prediction and analysis of anti-angiogenic peptides. *Int. J. Mol. Sci.*, **20**. doi: 10.3390/ijms20122950.

Laengsri,V. *et al.* (2019b) ThalPred: a web-based prediction tool for discriminating thalassemia trait and iron deficiency anemia. *BMC Med. Inform. Decis. Mak.*, **19**, 212.

Lau,J.L. and Dunn,M.K. (2018) Therapeutic peptides: historical perspectives, current development trends, and future directions. *Bioorg. Med. Chem.*, **26**, 2700–2707.

Lee,T.-Y. *et al.* (2011) Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences. *Bioinformatics*, **27**, 1780–1787.

Lv,H. *et al.* (2019) Evaluation of different computational methods on 5-methylcytosine sites identification. *Brief. Bioinform.* doi: 10.1093/bib/bbz048.

Manavalan,B. *et al.* (2017) MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget*, **8**, 77121–77136.

Manavalan,B. *et al.* (2018) Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *J. Proteome Res.*, **17**, 2715–2726.

Manavalan,B. *et al.* (2019a) 4mCpred-EL: an ensemble learning framework for identification of DNA N(4)-methylcytosine sites in the mouse genome. *Cells*, **8**. doi: 10.3390/cells8111332.

Manavalan,B. *et al.* (2019b) AtbPpred: a robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees. *Comput. Struct. Biotechnol. J.*, **17**, 972–981.

Manavalan,B. *et al.* (2019c) mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics*, **35**, 2757–2765.

Manavalan,B. *et al.* (2019d) Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol. Ther. Nucleic Acids*, **16**, 733–744.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Qiang,X. *et al.* (2018) CPPred-FL: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning. *Brief. Bioinform.* doi: 10.1093/bib/bby091.

Rao,B. *et al.* (2019) ACPred-Fuse: fusing multi-view information improves the prediction of anticancer peptides. *Brief. Bioinform.* doi: 10.1093/bib/bbz088.

Saha,I. *et al.* (2012) Fuzzy clustering of physicochemical and biochemical properties of amino acids. *Amino Acids*, **43**, 583–594.

Schaduangrat,N. *et al.* (2019a) ACPred: a computational tool for the prediction and analysis of anticancer peptides. *Molecules*, **24**, E1973.

Schaduangrat,N. *et al.* (2019b) Meta-iAVP: a sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation. *Int. J. Mol. Sci.*, **20**, E5743.

Schneider,G. and Wrede,P. (1994) The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. *Biophys. J.*, **66**, 335–344.

Shoombuatong,W. *et al.* (2019) THPep: a machine learning-based approach for predicting tumor homing peptides. *Comput. Biol. Chem.*, **80**, 441–451.

Su,R. *et al.* (2019) Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. *Brief. Bioinform.* doi: 10.1093/bib/bby124.

Vapnik,V. (2013) *The Nature of Statistical Learning Theory*. Springer Science & Business Media, New York.

Vlieghe,P. *et al.* (2010) Synthetic therapeutic peptides: science and market. *Drug Discov. Today*, **15**, 40–56.

Wang,J. *et al.* (2018) Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors. *Bioinformatics*, **34**, 2546–2555.

Wang,J. *et al.* (2019) Bastion3: a two-layer ensemble predictor of type III secreted effectors. *Bioinformatics*, **35**, 2017–2028.

Wei,L. *et al.* (2018) ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics*, **34**, 4007–4016.

Wei,L. *et al.* (2019) PEPred-Suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning. *Bioinformatics*, **35**, 4272–4280.

Weinberger,K.Q. *et al.* (2006) Distance metric learning for large margin nearest neighbor classification. In: Advances in Neural Information Processing Systems, pp. 1473–1480.

Win,T.S. *et al.* (2017) HemoPred: a web server for predicting the hemolytic activity of peptides. *Future Med. Chem.*, **9**, 275–291.

Yang,H. *et al.* (2019) A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae. *Brief. Bioinform.* doi: 10.1093/bib/bbz123.

Zhang,M. *et al.* (2019) MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters. *Bioinformatics*, **35**, 2957–2965.

Zhang,Y. *et al.* (2020) PeNGaRoo, a combined gradient boosting and ensemble learning framework for predicting non-classical secreted proteins. *Bioinformatics*, **36**, 704–712.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/doi/10.1093/bioinformatics/btaa160/5799076 by Ajou University Medical Library user on 16 April 2020

Computational and Structural Biotechnology Journal 18 (2020) 906–912

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

i4mC-Mouse: Improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes

Md. Mehedi Hasan [a,b], Balachandran Manavalan [c], Watshara Shoombuatong [d], Mst. Shamima Khatun [a], Hiroyuki Kurata [a,e,*]

[a] Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan
[b] Japan Society for the Promotion of Science, 5-3-1 Kojimachi, Chiyoda-ku, Tokyo 102-0083, Japan
[c] Department of Physiology, Ajou University School of Medicine, Suwon 443380, Republic of Korea
[d] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[e] Biomedical Informatics R&D Center, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

ARTICLE INFO

Article history:
Received 7 January 2020
Received in revised form 31 March 2020
Accepted 3 April 2020
Available online 8 April 2020

Keywords:
Mouse genome
Sequence analysis
Sequence encoding
Machine learning

ABSTRACT

*N*4-methylcytosine (4mC) is one of the most important DNA modifications and involved in regulating cell differentiations and gene expressions. The accurate identification of 4mC sites is necessary to understand various biological functions. In this work, we developed a new computational predictor called i4mC-Mouse to identify 4mC sites in the mouse genome. Herein, six encoding schemes of k-space nucleotide composition (KSNC), k-mer nucleotide composition (Kmer), mono nucleotide binary encoding (MBE), dinucleotide binary encoding, electron–ion interaction pseudo potentials (EIIP) and dinucleotide physicochemical composition were explored that cover different characteristics of DNA sequence information. Subsequently, we built six RF-based encoding models and then linearly combined their probability scores to construct the final predictor. Among the six RF-based models, the Kmer, KSNC, MBE, and EIIP encodings are sufficient, which contributed to 10%, 45%, 25%, and 20% of the prediction performance, respectively. On the independent test the i4mC-Mouse predicted the 4mC sites with accuracy and MCC of 0.816 and 0.633, respectively, which were approximately 2.5% and 5% higher than those of the existing method (4mCpred-EL). For experimental biologists, a freely available web application was implemented at http://kurata14.bio.kyutech.ac.jp/i4mC-Mouse/.

© 2020 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

In both prokaryotes and eukaryotes, N4-methylcytosine (4mC), 5-Methylcytosine (5mC), and N6-methyladenine (6 mA) alterations can regulate various functions including genomic imprinting, cell developmental, and gene expressions, and play crucial roles in the genomic diversity [1,2]. The 5mC modification is a common type of methylation alteration and well-explored that exemplifies an important role in biological developments [3,4] that are associated by the various diseases such as diabetes, neurological, and cancer [5,6]. The 4mC modification is also an effective methylation that defends the self-DNA from being degraded by restriction enzymes.

Until now, many experimental methodologies, such as mass spectrometry, methylation-precise PCR, and Single Molecule of Real-Time (SMRT) sequencing [7–10], have been efficiently used to identify the epigenetic 4mC sites. The exact dataset of modifications of 4mC sites is still limited due to the shortage of experimental identification approaches. Moreover, the aforementioned experimental approaches are labor-intensive and expensive works. Thus, computational tools are required for analysis of the accessible big data on the genome of mouse so as to allow the identification of novel 4mC sites, while shedding light on their mechanism [11,12]. Several computational approaches have been proposed by using the recently constructed database named MethSMRT [13] to predict 4mC sites from seven different species, i.e. *E. coli*, *G. subterraneus*, *A. thaliana*, *D. melanogaster*, *C. elegans*, *G. pickeringii,* and *Rosaceae genome*. [11,12,14–16]. To the best of author's knowledge, only one predictor is available for the 4mC sites in the mouse genome, named 4mCpred-EL [11]. This method implemented multiple encodings and machine learning (ML) algorithms, which was

* Corresponding author at: Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan.
E-mail address: kurata@bio.kyutech.ac.jp (H. Kurata).

https://doi.org/10.1016/j.csbj.2020.04.001
2001-0370/© 2020 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

applied to the dataset derived from the MethSMRT. Although the 4mCpred-EL yielded encouraging results, there is still room for further enhancement, probably because the employed feature information is not sufficient to capture the discriminative information between the two classes.

Motivated by the aforementioned problems, in this work, we have implemented a computational tool called i4mC-Mouse for the identification of 4mCs in the genome of mouse. A workflow of the proposed i4mC-Mouse is summarized in Fig. 1. Initially, six probabilities of 4mC sites were predicted by using a random forest (RF) classifier in conjunction with the k-mer nucleotide (NT) arrangement (Kmer), k-space NT composition (KSNC), NT mono binary encoding (MBE), dinucleotide binary encoding (DBE), electron–ion pseudopotentials (EIIP), and dinucleotide physicochemical composition (DPC). Secondly, to select the successive feature vectors, the Wilcoxon rank sum test (WR) was accessed. Finally, the four (Kmer, KSNC, MBE and EIIP) models evaluated the probability scores of 4mC sites and these scores were linearly combined to develop the i4mC-Mouse. Our results on independent test showed that i4mC-Mouse outperformed the existing predictor 4mCpred-EL. Finally, for the convenience of experimental scientists, our proposed model was implemented as a web application.

2. Materials and methods

2.1. Dataset construction

To develop a sequence-based predictor of 4mCs, a reliable dataset is necessary. To make a fair comparison, we used the previous dataset [11], which was collected from MethSMRT [13]. The DNA sequence windows are set to 41 base pairs (bp) having "C" at the center. To yield a high-quality dataset, we considered the sequences with a modQV score of $\geq$20 and excluded the remaining sequences. It is worth mentioning that the previous study applied a CD-HIT of 80% [17] and excluded the sequences that share 80% sequence identity. To develop a more reliable model and avoid an overestimation of prediction model, we applied CD-HIT of 70% and excluded the sequences that showed greater than 70% sequence identity. After such screening procedures, we finally obtained the benchmark dataset containing 906 positive samples, which are 74 samples lower than those of the 4mCPred-EL. A subset of 906 non-4mCs were randomly extracted from the non-4mCs. After obtaining the balanced dataset consisting of 906 4mCs and 906 non-4mCs, we divided them into the training and independent sets, such as 80% samples (746 4mCs and 746 non-4mCs) and 20% samples (160 4mCs and 160 non-4mCs), respectively.

2.2. Feature encoding

The next crucial step is to represent a DNA sequence as fixed-length feature vectors [18,19]. Six encoding methods of Kmer, KSNC, MBE, DBE, EIIP and DPC were used. The potential capability of these encodings employed in many domains has already been mentioned in our previous studies [20,21].

Kmer: This encoding has been extensively used in different prediction tasks [15,22,23]. In this study, a DNA sample with L length is articulated as $D = d_1, d_2, d_3, \ldots d_L$, d_i is one of the NTs (A, C, G, T, N). Considering tri-, and tetra -nucleotides, the Kmer scheme gener-

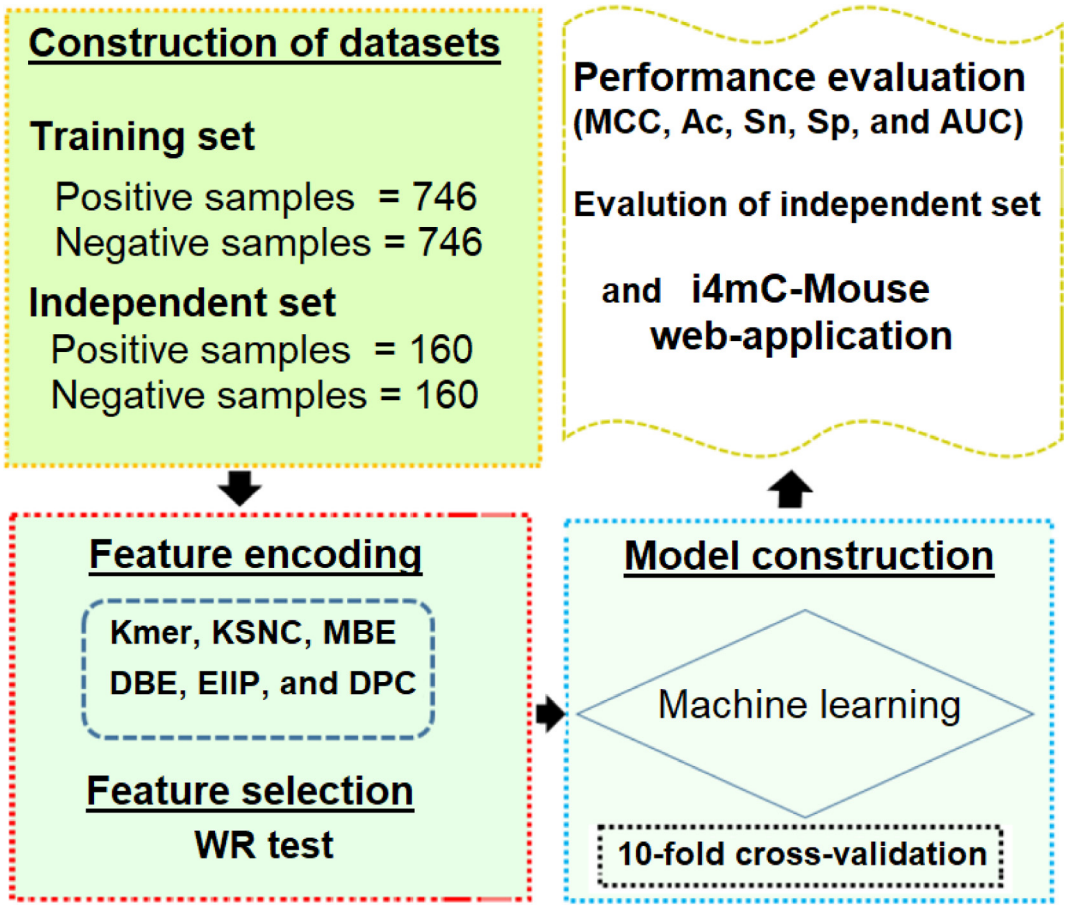

Fig. 1. A computational framework of the i4mC-Mouse. It includes three steps: (i) dataset construction; (ii) selection of six different encoding schemes that convert DNA sequences into numerical feature vectors; and (iii) model evaluation and construction using a CV test. Then, construction of a webserver for the final prediction model (i4mC-Mouse).

ated a 750 (=$5^3 + 5^4$) dimensional (D) feature vector. Here the letter 'N' signifies a non-standard nucleotide.

KSNC: This encoding signifies the frequency NTs information by using the pair-wise similarity searches [23] and widely used in bioinformatics tasks [24–26]. The NT (A, C, G, T, N) pairs (nc_i, where i = 1, 2,. . .,25) were encoded and standardized as

$$\text{NTpair} = \frac{F(nci)}{w - d - 1} \tag{1}$$

where $F(nc_i)$ is the sum of nc_i privileged 4mC sites. The w and d are the sequence length and space length between NTs, respectively. For a range of *dmax* is 0 to 3, the KSNC signifies a 100-D feature vector.

MBE: The MBE exactly depicts the NT for the sequence of curated samples at each position, where A, T, G, C, and N are represented by (1,0,0,0,0), (0,1,0,0,0), (0,0,1,0,0), (0,0,0,1,0), and (0,0,0,0,1), respectively. In MBE, for a length of NT sequence, a $w \times 5$-D vector was generated.

DBE: In the DBE scheme, the possible 16 dinucleotides are encoded as 0/1 (four-dimensional vector) [11]. For instance, AT (0,0,0,1), AA (0,0,0,0), GG (1,1,1,1), and AC (0,0,1,0) are encoded [27,28]. All N pair dinucleotides are regarded as zero. For a sequence of 4mC or non-4mC with a DBE, a 160 $\{(w-1) \times 4\} - D$ vector was generated.

EIIP: To encode the electron–ion energies in the DNA, Nair and Sreedharan developed EIIP [29]. In this study, EIIP values were encoded as follows: A (0.1260), C (0.1340), G (0.0806), T (0.1335), and N (0.0000). The EIIP scheme transformed a sequence into a w-D feature vector.

DPC: Fifteen types of DPC were collected from the recent publications [20,21]. The physicochemical properties are encoded as a 375 (25 dinucleotides $\times$ 15 physicochemical properties)-D vector.

2.3. Feature selection

Inclusion of non-informative and noisy feature might cause unsatisfied prediction performances [30,31]. In fact, there are several feature selection and ranking approaches, such as Chi-square, mRMR, and WR test. In this work, the WR feature selection method was used [32].

2.4. Machine learning classifier

The computational model employed herein was constructed by using the RF algorithm [33]. The RF classifier is widely used in various biological problems [34–40]. The RF classifier is a collaborative model consisting of many regression and classification trees, and the prediction performances are enhanced by increasing the number of weak CART classifiers. In this study, the RF package 'randomForest' (https://cran.r-project.org/) was used.

It is crucial to compare the proposed RF-based models with other commonly used ML-based models, i.e. Naive Bayes (NB) [41,42], SVM [37,43], k-nearest neighbor (KNN), and AdaBoost (AB). The NB and AB classifiers were performed in R programming (https://www.r-project.org/), while the KNN classifier was implemented in our house PERL program. The SVM^{light} was used to build the SVM algorithm [38]. Notably, all these classifiers are extensively applied to various prediction problems [44–48].

2.5. Combined model

To increase the prediction performance of the proposed model, we linearly combined the probability scores of the six, single encoding-based models, as given by:

$$\text{Combined } (s) = \sum_{i=1}^{6} w_i x_i(s), \ \sum_{i=1}^{6} w_i = 1 \tag{2}$$

where Combined (s) specifies the combination of the 6 scores evaluated by the single encoding scheme-employing MLs, w_i characterizes the weight of the i-th encoding model and xi(s) specifies the ML scores of sample s based on the i-th encoding model. These weight values were adjusted based on the AUC values via 10-fold cross-validation (CV) tests.

2.6. Evaluation metrics

Four statistical metrics: Matthews correlation coefficient (MCC), accuracy (Ac), sensitivity (Sn), and specificity (Sp) were used to evaluate the performance of the predictors as follows [39,49–52]:

$$\text{MCC} = \frac{\text{n(TP)} \times \text{n(TN)} - \text{n(FP)} \times \text{n(FN)}}{\sqrt{[\text{n(TN)} + \text{n(FN)}] \times [\text{n(TP)} + \text{n(FP)}] \times [\text{n(TN)} + \text{n(FP)}] \times [\text{n(TP)} + \text{n(FN)}]}} \tag{3}$$

$$\text{Ac} = \frac{\text{n(TP)} + \text{n(TN)}}{\text{n(TP)} + (\text{FN}) + \text{n(FP)} + \text{n(TN)}} \tag{4}$$

$$\text{Sn} = \frac{\text{n(TP)}}{\text{n(TP)} + \text{n(FN)}} \tag{5}$$

$$\text{Sp} = \frac{\text{n(TN)}}{\text{n(TN)} + \text{n(FP)}} \tag{6}$$

where n(TP) and n(TN) specify the numbers correctly predicted samples of 4mCs and non-4mCs, respectively. n(FP) and n(FN) specify the numbers incorrectly predicted samples of 4mCs and non-4mCs, respectively.

3. Results and discussion

3.1. Nucleotide preference analysis

We aim to develop a computational model for discriminating 4mC samples from non-4mC ones. Therefore, we sought to determine the composition of sequence preferences between the 4mC and non-4mC samples by using the pLogo software [53]. The pLogo examines the statistically significant differences in position-specific NTs ($p < 0.05$). As seen in Fig. 2, the C base was overrepresented compared to the other bases in the 4mC samples and the A base was under-represented compared to the other bases, while the G and T bases were observed at both the over- and underrepresented positions. In summary, the over- and under-represented A and C bases were considerably varied between the 4mC and non-4mC samples, suggesting the importance of position-specific preferences of nucleotide base pairs, which is consistent with the previous study [11].

3.2. Performance evaluation of i4mC-Mouse

First, the training dataset was converted into feature vectors by using six schemes (Kmer, KSNC, MBE, DBE, EIIP, and DPC) and individually inputted to a RF classifier. Second, we evaluated the successive feature vectors for the six, single encoding models by 10-fold CV tests. To reduce the feature dimension and improve the prediction performance, we carried out the WR test approach to select an optimal feature set on each encoding and compared its performance with the control. As shown in Table S1, the feature selection improved the performance on the three encodings (Kmer (160D), KSNC (80D) and DPC (110D)), while the remaining three

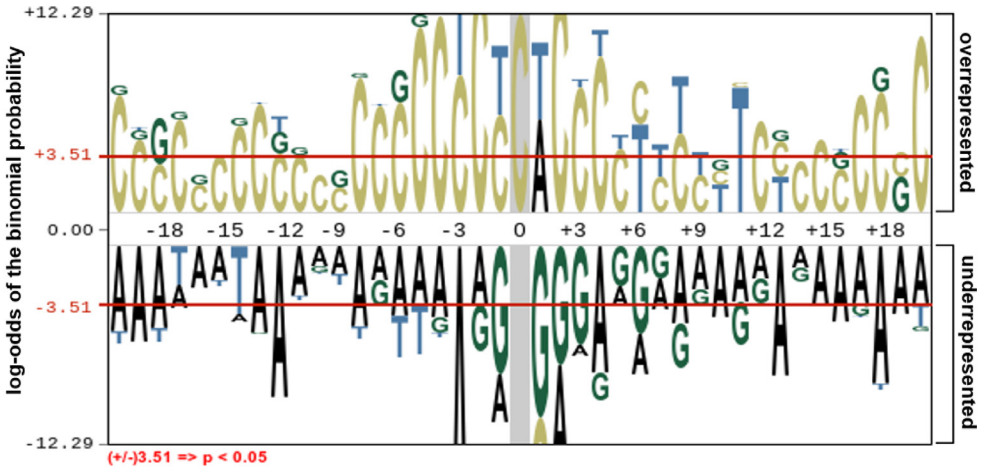

Fig. 2. Sequence logo representation of 4mC samples. The 20 upstream and 20 downstream DNA residues surrounding the mouse 4mC site were analyzed.

encodings (MBE, DBE and EIIP) did not outperform their controls. Therefore, we used three optimal feature set-based models for the subsequent analysis. Fig. 3 and Table 1 show the prediction performances of the six, single encoding-based models and the combined model (i4mC-Mouse). The six, single encoding-based models of Kmer, KSNC, MBE, DBE, EIIP and DPC provided AUCs of 0.869, 0.882, 0.851, 0.814, 0.840 and 0.822, respectively. In terms of Ac and MCC, the KSNC encoding outperformed the other encodings, where the AUC of the KSNC was approximately ~1–7% higher than the AUCs of the other encodings.

In the combined model, a linear regression model was used to integrate the six RF probability scores, as mentioned in the method section, where the weight coefficients of the Kmer, KSNC, MBE, DBE, EIIP, and DPC schemes are 0.10, 0.45, 0.25, 0.00, 0.20 and 0.00, respectively. Notably, our approach excluded the two models (DBE and DPC) by assigning weight 0.00 and considered the remaining four models. The contribution of Kmer, KSNC, MBE and EIIP are 10%, 45%, 25%, and 20%, respectively, in the final prediction. As noticed in Table 1, at a Sp control of 90.42%, the i4mC-Mouse yielded MCC, Ac, Sn, and Sp of 0.651, 79.30% 68.31%, and 90.42% respectively. To show the advantage of our approach, we computed the statistically significant differences between the i4mC-Mouse and each single encoding-based model using two-tailed *t*-test [54]. The i4mC-Mouse outperformed the five models at a *p*-value of <0.05, except the KSNC model at a *p*-value of 0.063.

Table 1
Prediction performances of the i4mC-Mouse model and the single encoding-based RF models.

Methods	MCC	Ac (%)	Sn (%)	Sp (%)	AUC	P-value
Kmer	0.566	74.81	59.53	90.10	0.869	0.011
KSNC	0.602	76.90	63.42	90.30	0.882	0.063
MBE	0.486	71.20	53.81	88.61	0.851	0.006
DBE	0.432	69.13	48.11	90.10	0.814	0.001
EIIP	0.473	70.80	52.31	89.21	0.840	0.001
DPC	0.428	69.21	49.91	88.52	0.822	0.001
i4mC-Mouse	0.651	79.30	68.31	90.20	0.904	–

* i4mC-Mouse specifies the linear arrangement of the RF scores for Kmer, KSNC, MBE, DBE, EIIP, and DPC encodings and their weight values are 0.10, 0.45, 0.25, 0.00, 0.20, and 0.00, respectively.

3.3. Effect of ML algorithms on prediction performances of the combined model

We applied the above procedure (the construction of six encoding-based models and combined models) to other commonly used four classifiers (NB, SVM, AB and KNN) and compared their

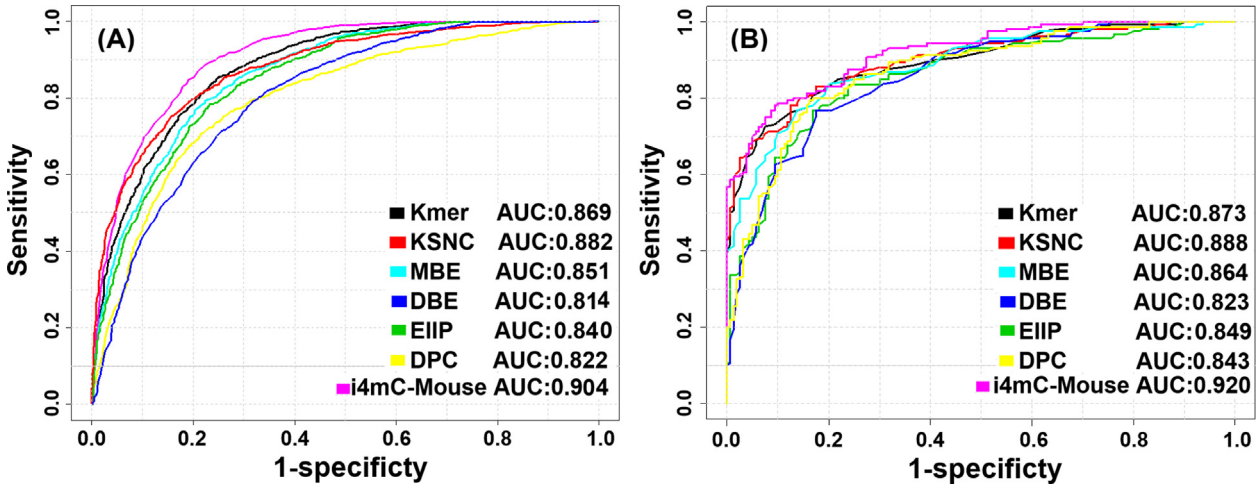

Fig. 3. Performance comparisons of single encoding-based models and i4mC-Mouse. The ROC curves were evaluated on the training dataset by a 10-fold CV test (A) and independent dataset (B).

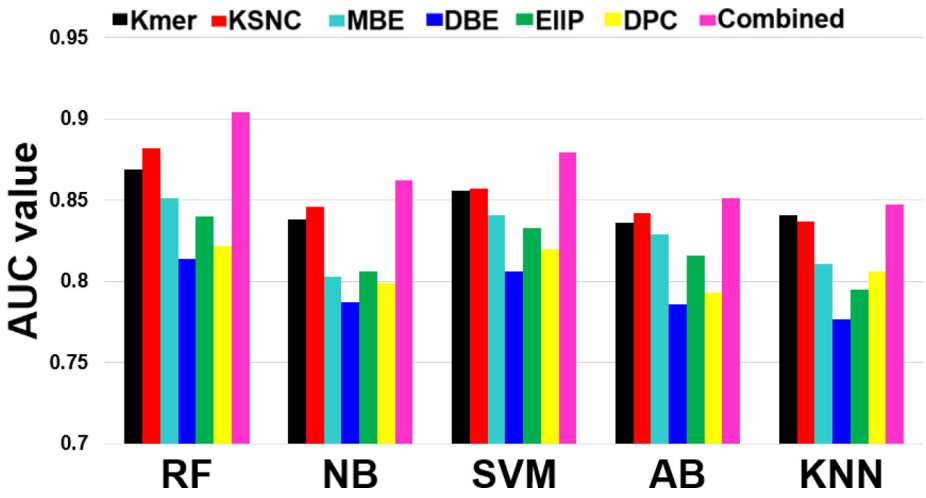

Fig. 4. Effect of different ML algorithms on the AUC values of the six single encoding-based models and i4mC-Mouse. The performances were evaluated on the training datasets by a 10-fold CV test.

performances with the RF-based models. Instead of selecting default ML parameters, 10-fold CV was employed to optimize their respective ML parameters on each encoding-based classifier. Finally, an optimal model was obtained for each classifier, whose performances are shown in Fig. 4. We noted that the combined model for each classifier performed better than the individual encoding-based model, indicating the integration of multiple information is effective in achieving the best performance. Furthermore, comparison among the combined models with five different classifiers showed that the RF achieved the best performance, while the SVM was comparable to the RF model. Specifically, AUCs of the RF (i.e. i4mC-Mouse) were ~1–5% higher than those of any other combined models, demonstrating that the RF model is the most suitable for the i4mC prediction.

3.4. Comparison of i4mC-Mouse with 4mCpred-EL on the independent dataset

We compared the proposed i4mC-Mouse with the existing method (4mCpred-EL) on the same independent dataset consisting of 160 4mCs and 160 non-4mCs, as shown in Table 2. We directly submitted to the independent dataset to the 4mCpred-EL web server. The 4mCpred-EL yielded 79.10% Ac, 75.72% Sn, 82.51% Sp, 0.584 MCC, and 0.881 AUC, while the i4mC-Mouse provided 81.61% Ac, 80.71% Sn, 82.52% Sp, 0.633 MCC, and 0.920 AUC. The i4mC-Mouse outperformed the 4mCpred-EL with increased ratios of >3%, >5% and >5% on Ac, Sn and MCC, respectively. The better performance of the i4mC-Mouse would be due to the followings: selection of an appropriate classifier, a linear combination of single encoding-based models, and reduction of dataset redundancy.

3.5. i4mC-Mouse web server

A user-friendly and freely accessible web application was established for the prediction of mouse genome at http://kurata14.bio.kyutech.ac.jp/i4mC-Mouse/. The manuals are as follows: (i) select the exact 41 bp DNA 4mC genome (ii) browse or enter the query sequences from users' own file (FASTA format) to the input page, where a sample is shown our server page, (iii) push the 'Submit' button. The server completes the query tasks with the probability scores within one min.

Table 2
Comparison between the i4mC-Mouse and 4mCpred-EL.

Method	MCC	Ac (%)	Sn (%)	Sp (%)	AUC
4mCpred-EL	0.584	79.10	75.72	82.51	0.881
i4mC-Mouse	0.633	81.61	80.71	82.52	0.920

The performances were evaluated on the independent dataset.

4. Conclusions

4mC plays an important role in the DNA modifications and is involved in regulating cell differentiations and gene expression levels. Therefore, accurate identification of 4mC sites is an essential step to understand the exact biological functions. To date, several computational prediction tools have been developed to identify 4mC sites from different species [11,12,14–16,20,55,56], but only one method is available for mouse species. In this study, we have developed a new computational model, called i4mC-Mouse, for improving the prediction of 4mCs in the mouse genome. We employed six encoding schemes of Kmer, KSNC, MBE, DBE, EIIP and DPC to cover various aspects of DNA sequences and optimized the successive features via the WR feature selection method. The final constructed i4mC-Mouse was a linear combination of the predicted probabilities by four, single encoding-based RF-models, where the Kmer, KSNC, MBE and EIIP encodings contributed to 10%, 45%, 25%, and 20%, respectively. On the independent test the i4mC-Mouse outperformed the existing method (4mCpred-EL). The i4mC-Mouse is demonstrated to be the most accurate predictor. Finally, a freely available web application was implemented.

5. Author statement

MH and HK conceived the project. MMH and KMS collected and analyzed the datasets. MMH drafted the manuscript. HK, MMH, MB, SW and KMS thoroughly revised the manuscript. All authors approved and read the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the Grant-in-Aid for JSPS Research Fellow (19F19377) from Japan Society for the Promotion of Science (JSPS), partially supported from Japan Society for the Promotion of Science by Grant-in-Aid for Scientific Research (B) (19H04208) and by the developing key technologies for discovering and manufacturing pharmaceuticals used for next-generation treatments and diagnoses both from the Ministry of Economy, Trade and Industry, Japan (METI) and from Japan Agency for Medical Research and Development (AMED).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.04.001.

References

[1] Rathi P, Maurer S, Summerer D. Selective recognition of N4-methylcytosine in DNA by engineered transcription-activator-like effectors. Philos Trans R Soc London Ser B, Biol Sci 2018;373(1748).
[2] Jeltsch A, Jurkowska RZ. New concepts in DNA methylation. Trends Biochem Sci 2014;39(7):310–8.
[3] Akalin A, Garrett-Bakelman FE, Kormaksson M, Busuttil J, Zhang L, Khrebtukova I, et al. Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. PLoS Genet 2012;8(6):e1002781.
[4] Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 2008;9(6):465–76.
[5] Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 2012;13(7):484–92.
[6] Ling C, Groop L. Epigenetics: a molecular link between environmental factors and type 2 diabetes. Diabetes 2009;58(12):2718–25.
[7] Doherty R, Couldrey C. Exploring genome wide bisulfite sequencing for DNA methylation analysis in livestock: a technical assessment. Front Genet 2014;5:126.
[8] Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 2010;7(6):461–5.
[9] Boch J, Bonas U. Xanthomonas AvrBs3 family-type III effectors: discovery and function. Annu Rev Phytopathol 2010;48:419–36.
[10] Buryanov YI, Shevchuk TV. DNA methyltransferases and structural-functional specificity of eukaryotic DNA modification. Biochem Biokhimiia 2005;70 (7):730–42.
[11] Manavalan B, Basith S, Shin TH, Lee DY, Wei L, Lee G. 4mCpred-EL: an ensemble learning framework for identification of DNA N(4)-methylcytosine sites in the mouse genome. Cells 2019;8(11).
[12] He W, Jia C, Zou Q. 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. Bioinformatics 2019;35(4):593–601.
[13] Ye P, Luan Y, Chen K, Liu Y, Xiao C, Xie Z. MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. Nucleic Acids Res 2017;45(D1):D85–9.
[14] Wei L, Luan S, Nagai LAE, Su R, Zou Q. Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. Bioinformatics 2019;35(8):1326–33.
[15] Hasan MM, Manavalan B, Khatun MS, Kurata H. i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome. Int J Biol Macromol 2019.
[16] Chen W, Yang H, Feng P, Ding H, Lin H. iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics 2017;33(22):3518–23.
[17] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28(23):3150–2.
[18] Xu ZC, Feng PM, Yang H, Qiu WR, Chen W, Lin H. iRNAD: a computational tool for identifying D modification sites in RNA sequence. Bioinformatics 2019.
[19] Yang H, Lv H, Ding H, Chen W, Lin H. iRNA-2OM: a sequence-based predictor for identifying 2'-O-methylation sites in Homo sapiens. J Comput Biol 2018;25 (11):1266–77.
[20] Manavalan B, Basith S, Shin TH, Wei L, Lee G. Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. Mol Ther Nucleic Acids 2019;16:733–44.
[21] Manavalan B, Shin TH, Lee G. DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. Oncotarget 2018;9(2):1944.
[22] Yang H, Yang W, Dao FY, Lv H, Ding H, Chen W, et al. A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae. Brief Bioinf 2019.
[23] Zhou Y, Zeng P, Li YH, Zhang Z, Cui Q. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Res 2016;44(10):e91.
[24] Hasan MM, Guo D, Kurata H. Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information. Mol BioSyst 2017;13(12):2545–50.
[25] Hasan MM, Yang S, Zhou Y, Mollah MN. SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. Mol BioSyst 2016;12(3):786–95.
[26] Khatun MS, Hasan MM, Kurata H. PreAIP: computational prediction of anti-inflammatory peptides by integrating multiple complementary features. Front Genet 2019.
[27] Hasan MM, Manavalan B, Shoombuatong W, Khatun MS, Kurata H. i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation. Plant Mol Biol 2020.
[28] Lv H, Zhang ZM, Li SH, Tan JX, Chen W, Lin H. Evaluation of different computational methods on 5-methylcytosine sites identification. Brief Bioinf 2019.
[29] Nair AS, Sreenadhan SP. A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). Bioinformation 2006;1(6):197–202.
[30] Feng CQ, Zhang ZY, Zhu XJ, Lin Y, Chen W, Tang H, et al. iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. Bioinformatics 2019;35(9):1469–77.
[31] Dao FY, Lv H, Wang F, Feng CQ, Ding H, Chen W, et al. Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. Bioinformatics 2019;35(12):2075–83.
[32] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27(8):1226–38.
[33] Liaw A. Wiener: classification and regression by random forest. R news 2002;2:18–22.
[34] Boopathi V, Subramaniyam S, Malik A, Lee G, Manavalan B, Yang DC. mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides. Int J Mol Sci 2019;20(8).
[35] Manavalan B, Shin TH, Kim MO, Lee G. PIP-EL: a new ensemble learning method for improved proinflammatory peptide predictions. Front Immunol 2018;9:1783.
[36] Manavalan B, Basith S, Shin TH, Wei L, Lee G. mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics 2018.
[37] Hasan MM, Zhou Y, Lu X, Li J, Song J, Zhang Z. Computational identification of protein pupylation sites by using profile-based composition of k-spaced amino acid pairs. PLoS ONE 2015;10(6):e0129635.
[38] Khatun S, Hasan M, Kurata H. Efficient computational model for identification of antitubercular peptides by integrating amino acid patterns and properties. FEBS Lett 2019;593(21):3029–39.
[39] Hasan MM, Khatun MS, Kurata H. Large-scale assessment of bioinformatics tools for lysine succinylation sites. Cells 2019;8(2).
[40] Schaduangrat N, Nantasenamat C, Prachayasittikul V, Shoombuatong W. ACPred: a computational tool for the prediction and analysis of anticancer peptides. Molecules 2019;24(10).
[41] Hasan MM, Rashid MM, Khatun MS, Kurata H. Computational identification of microbial phosphorylation sites by the enhanced characteristics of sequence information. Sci Rep 2019;9(1):8258.
[42] Feng PM, Ding H, Chen W, Lin H. Naive Bayes classifier with feature selection to identify phage virion proteins. Comput Math Methods Med 2013;2013:530696.
[43] Lai HY, Zhang ZY, Su ZD, Su W, Ding H, Chen W, et al. iProEP: a computational predictor for predicting promoter. Mol Ther Nucleic acids 2019;17:337–46.
[44] Manavalan B, Basith S, Shin TH, Wei L, Lee G. AtbPpred: a robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees. Comput Struct Biotechnol J 2019;17:972–81.
[45] Hasan MM, Manavalan B, Khatun MS, Kurata H. Prediction of S-nitrosylation sites by integrating support vector machines and random forest. Mol Omics 2019;15(6):451–8.
[46] Charoenkwan P, Yana J, Schaduangrat N, Nantasenamat C, Hasan MM, Shoombuatong W. iBitter-SCM: identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. Genomics 2020. https://doi.org/10.1016/j.ygeno.2020.03.019.
[47] Shoombuatong W, Schaduangrat N, Pratiwi R, Nantasenamat C. THPep: a machine learning-based approach for predicting tumor homing peptides. Comput Biol Chem 2019;80:441–51.
[48] Win TS, Schaduangrat N, Prachayasittikul V, Nantasenamat C, Shoombuatong W. PAAP: a web server for predicting antihypertensive activity of peptides. Future Med Chem 2018;10(15):1749–67.
[49] Yang W, Zhu XJ, Huang J, Ding H, Lin H. A brief survey of machine learning methods in protein sub-Golgi localization. Curr Bioinform 2019;14:234–40.
[50] Ding H, Yang W, Tang H, Feng PM, Huang J, Chen W, et al. PHYPred: a tool for identifying bacteriophage enzymes and hydrolases. Virologica Sinica 2016;31 (4):350–2.

[51] Basith S, Manavalan B, Hwan Shin T, Lee G. Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. Med Res Rev 2020.
[52] Hasan MM, Schaduangrat N, Basith S, Lee G, Shoombuatong W, Manavalan B. HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. Bioinformatics 2020.
[53] O'Shea JP, Chou MF, Quader SA, Ryan JK, Church GM, Schwartz D. pLogo: a probabilistic approach to visualizing sequence motifs. Nat Methods 2013;10(12):1211–2.
[54] Raju TN, Gosset WS, Silverman WA. Two "students" of science. Pediatrics 2005;116(3):732–5.
[55] Zeng F, Fang G, Yao L. A deep neural network for identifying DNA N4-methylcytosine sites. Front Genet 2020;11:209.
[56] Lv H, Dao FY, Zhang D, Guan ZX, Yang H, Su W, et al. iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. ISCIENCE 2020. https://doi.org/10.1016/j.isci.2020.100991.

Plant Molecular Biology
https://doi.org/10.1007/s11103-020-00988-y

i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation

Md. Mehedi Hasan[1,2] · Balachandran Manavalan[3] · Watshara Shoombuatong[4] · Mst. Shamima Khatun[1] · Hiroyuki Kurata[1,5]

Received: 6 February 2020 / Accepted: 29 February 2020
© Springer Nature B.V. 2020

Abstract

DNA N^6-methyladenine (6 mA) is one of the most vital epigenetic modifications and involved in controlling the various gene expression levels. With the avalanche of DNA sequences generated in numerous databases, the accurate identification of 6 mA plays an essential role for understanding molecular mechanisms. Because the experimental approaches are time-consuming and costly, it is desirable to develop a computation model for rapidly and accurately identifying 6 mA. To the best of our knowledge, we first proposed a computational model named i6mA-Fuse to predict 6 mA sites from the Rosaceae genomes, especially in *Rosa chinensis* and *Fragaria vesca*. We implemented the five encoding schemes, i.e., mononucleotide binary, dinucleotide binary, k-space spectral nucleotide, k-mer, and electron–ion interaction pseudo potential compositions, to build the five, single-encoding random forest (RF) models. The i6mA-Fuse uses a linear regression model to combine the predicted probability scores of the five, single encoding-based RF models. The resultant species-specific i6mA-Fuse achieved remarkably high performances with AUCs of 0.982 and 0.978 and with MCCs of 0.869 and 0.858 on the independent datasets of *Rosa chinensis* and *Fragaria vesca*, respectively. In the *F. vesca*-specific i6mA-Fuse, the MBE and EIIP contributed to 75% and 25% of the total prediction; in the *R. chinensis*-specific i6mA-Fuse, Kmer, MBE, and EIIP contribute to 15%, 65%, and 20% of the total prediction. To assist high-throughput prediction for DNA 6 mA identification, the i6mA-Fuse is publicly accessible at https://kurata14.bio.kyutech.ac.jp/i6mA-Fuse/.

Key message

The existing prediction models are not suitable to identify 6mA in the Rosaceae genome because the existing algorithms are species-specific. Thus, a novel predictor is desired to be established to identify 6mA sites in the Rosaceae genome. To the best of our knowledge, we first propose a computation model named i6mA-Fuse (Identification of N6-MethylAdenine sites by Fusing multiple feature representation) to predict 6mA sites from the Rosaceae genomes, especially in *Rosa chinensis* and *Fragaria vesca*.

Keywords DNA 6 mA · Sequence analysis · Feature encoding · Machine learning

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11103-020-00988-y) contains supplementary material, which is available to authorized users.

✉ Hiroyuki Kurata
kurata@bio.kyutech.ac.jp

[1] Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

[2] Japan Society for the Promotion of Science, 5-3-1 Kojimachi, Chiyoda-ku, Tokyo 102-0083, Japan

[3] Department of Physiology, Ajou University School of Medicine, Suwon 443380, Korea

[4] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand

[5] Biomedical Informatics R&D Center, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

Published online: 05 March 2020

Introduction

DNA N^6 methyladenine (6 mA) regulates various biological functions including genomic imprinting, cell developmental, and chromosome solidity and magnifies genomic diversity in both prokaryotes and eukaryotes (Xiong et al. 2019; Zhang et al. 2018). The 6 mA plays an important role to examine the host DNA and defenses the host genome via the several modification systems (Du et al. 2019), which is evenly circulated across the genome (Liu et al. 2018b; O'Brown and Greer 2016). However, the biological functions and epigenetic modifications of 6 mA still remain unclear. Genomic 6 mA distributions are essential for revealing potential biological functions of DNA. Recently different experimental methods have been accomplished to identify 6 mA, including liquid chromatography coupled with real-time sequencing for single-molecule, and methyladenine—precise PCR (McIntyre et al. 2019; Zhang et al. 2015), but these methods were time-consuming, laborious, and expensive. The rapid development in machine learning (ML)-based algorithms have driven computational chemistry to an unprecedented revolution with the explosive growth of biological sequences in the next generation sequencing era (Chen et al. 2015; Chou 2019; Liu et al. 2016; Sun et al. 2020). Therefore, ML-based methods can be used as an alternative to experimental efforts.

Nowadays, several species-specific ML-based approaches have been established for the identification of 6 mA sites, including rice(Basith et al. 2019; Chen et al. 2019a; Huang et al. 2020; Yu and Dai 2019), and *Mus musculus* (Feng et al. 2019). Although several computational methods have been proposed for 6 mA prediction in some species (Qianfei Huang et al. 2020; Wang and Yan 2018), none of them were developed to specifically identify 6mAs in the Rosaceae genomes. Particularly, the existing prediction models are not suitable to identify 6 mA in the Rosaceae genome because the existing algorithms are species-specific. Thus, a novel predictor is desired to be established to identify 6 mA sites in the Rosaceae genome.

To the best of our knowledge, we first propose a computation model named i6mA-Fuse (Identification of N6-MethylAdenine sites by Fusing multiple feature representation) to predict 6 mA sites from the Rosaceae genomes, especially in *Rosa chinensis* and *Fragaria vesca*. An overall framework is depicted in Fig. 1. Firstly, the five feature vectors were respectively generated by the five encoding schemes of the k-mer composition (Kmer), k-space spectral nucleotide composition (KSNC), mononucleotide binary encoding (MBE), dinucleotide binary encoding (DBE), and electron–ion interaction pseudopotentials (EIIP). Subsequently, a random forest (RF) classifier was used to build the five, single encoding-employing models. Finally, the predicted probability scores of an appropriate encoding-based models were combined through a linear regression to make a final prediction. As far as we know, the i6mA-Fuse is the first computational predictor of 6mAs within the Rosaceae genomes.

Materials and methods

An outline of i6mA-Fuse is presented in Fig. 1. Four key phases are discussed as follows: (i) datasets construction, (ii) feature extraction, (iii) probability scores calculation, and (iv) final model construction.

Datasets construction

A high quality dataset could guarantee the reliability and robustness of the proposed model (Xu et al. 2019). In this

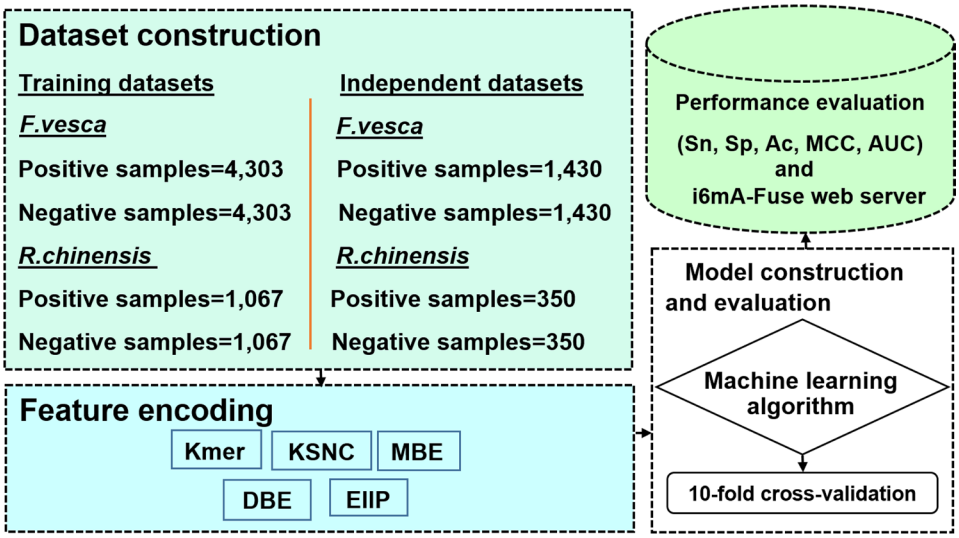

Fig. 1 An overall framework of i6mA-Fuse. It involves three steps: (i) dataset construction based on MDR database; (ii) employing five different encoding schemes for converting nucleotides into numerical feature vectors; and (iii) model construction and evaluation using cross-validation. Subsequently, a webserver was constructed based on the proposed model (i6mA-Fuse), where it predicts putative 6 mA sites from the submitted query sequences

study, the positive samples (6mAs) were extracted from the reliable MDR database for *F. vesca* and *R. chinensis* (Liu et al. 2019). Each sample had a length of 41 base pairs with adenine nucleotide ("A") at the center, whose modification score was ≥ 20. To avoid any similarity bias, we utilized the CD-HIT program (Fu et al. 2012) to exclude highly similar samples by setting a cutoff threshold of 0.7. After such a screening procedure, we obtained the positive samples of 5733 and 1417 for *R. chinensis* and *F. vesca*, respectively. Meanwhile, the standard process was applied to collect negative samples as described in the previous studies (Basith et al. 2019; Lv et al. 2019a). Herein, we obtained strict and objective datasets consisting of 5733 positive and negative samples for *R. chinensis*, and 1417 positive and negative samples for *F. vesca.*

To further validate the predictive ability of the proposed model, the aforementioned datasets were randomly divided into the training and independent datasets at a 3 to 1 ratio. Finally, we obtained 4303/1067 positive and negative samples for *F. vesca/R. chinensis* as training datasets, while 1430/350 positive and negative samples for *F. vesca/ R. chinensis* were treated as independent datasets. To confirm the reproducibility of models, all the curated dataset used in this study are available at https://kurata14.bio.kyutech.ac.jp/i6mA-Fuse/help.php.

Feature extraction

One of the vital procedures is to express DNA sequences with an operative mathematical expression that can accurately reproduce the intrinsic correlation with the anticipated objective (Yang et al. 2019a). In this study, the five encoding schemes consisting of Kmer, KSNC, MBE, DBE, and EIIP were used for constructing the i6mA-Fuse predictor. The Kmer scheme encodes the occurrence of nucleotide frequency in a DNA sequence (Liu et al. 2018a; Manavalan et al. 2018c). The DNA sequence is expressed by Kmer as: $F = f_1, f_2, f_3, \ldots, f_L$, where L is the positive/negative dataset length. The F_i is a nucleotide of A, C, G, and T. Therefore, mono-, di-, tri-, and tetra–nucleotides were encoded and combined to form a 340 ($4^1+4^2+4^3+4^4$) dimensional feature vector. The KSNC scheme encodes the DNA nucleobase information of the curated samples using the frequency wise pair similarity search (Charoenkwan et al. 2013; Zhou et al. 2016). A space of nucleobase frequency pairs is encoded and normalized as

$$\text{Frequencypair} = \frac{N(nf_i)}{w - S - 1} \quad (1)$$

where $N(nf_i)$ is the sum of nf_i inside DNA samples with w length of sample and S is the space between two nucleotides. The KSNC generates a $4 \times 4 \times (Smax + 1)$ dimensional vector for a sequence, where *Smax* was equal to 3. The MBE scheme exactly encodes the nucleotide at each position as a binary vector, where the A, G, T, and C are encoded as (1,0,0,0), (0,0,1,0), (0,1,0,0), and (0,0,0,1), respectively. The MBE encodes the sequence with length of w, a 164-dimensional vector was generated. The DPE scheme encodes 16 potential di-nucleotide as 0 or 1 as 4-dimensional vector (Manavalan et al. 2019a). For example, AA, AC, AT, and GG were coded to (0,0,0,0), (1,1,1,1), (0,0,0,1), and (0,0,1,0), respectively. Using the DPE, a DNA sample is transformed to a 160-dimension feature. The EIIP scheme expresses the electron–ion energies beside with the curated sequences, which is extensively used in bioinformatics research (Basith et al. 2019; Jia et al. 2018). The EIIP indexes of {A, C, G, T} were set to {0.1260, 0.1340, 0.0806, 0.1335} that generates a w-dimensional feature vector for a sequence.

Machine learning algorithms

This study utilized an ensemble method named RF model to develop i6mA-Fuse predictor (Liaw and Wiener 2002; Schaduangrat et al. 2019; Shoombuatong et al. 2019; Su et al. 2019; Win et al. 2017). Typically, when training data of size T with Q features is given, RF builds Q subsets of the data by the bootstrap sampling, and then randomly assigns Q features to each node to optimize the trees based on the by Gini impurity. We used 'randomForest' implemented in R (https://cran.r-project) with a default cut-off tree number of 1000 to evaluate the optimum performance. This package has been successfully applied to many protein and peptide prediction problems (Hasan et al. 2019b; Lv et al. 2019b; Manavalan et al. 2018b; Zhou et al. 2016). In order to prove the effectiveness of the proposed methods, we compared our RF predictor with the well-known five ML algorithms, i.e. SVM, Adaboost (AB), Naïve Bayes (NB), artificial neural network (ANN), k-nearest neighbor (KNN). In this study, the ANN and NB models (Frank et al. 2004) were implemented from the WEKA software, while the KNN model was developed by using an in-house developed PERL language. For the SVM model, we used the SVM^{light} package with default parameters (Hasan et al. 2019c; Khatun et al. 2019b).

Fusion model

To improve the prediction performance, the RF probability scores estimated by the Kmer, KSNC, MBE, DPE, and EIIP encoding schemes were linearly combined using the following formula:

$$\text{Combined} = w_1 * \text{Kmer} + w_2 * \text{KSNC} + w_3 * \text{MBE} + w_4 * \text{DBE} + w_5 * \text{EIIP} \quad (2)$$

where w_1, w_2, w_3, w_4, w_4, and w_5 are the weight values exhibiting the contribution of each encoding, where the summation of w_1, w_2, w_3, w_4, w_4, and w_5 is 1. Herein, the linear fusion models of the five RF scores estimated by using the five encoding schemes are referred as Combined model. To enhance the predictive performance of the proposed model, each weight coefficient was adjusted in the range of 0–1 with an interval of 0.05 using a grid-search strategy.

Hybrid model

We investigated the effect of the hybrid feature (H) on 6 mA site prediction. The five encoding feature vectors (F) of Kmer, KSNC, MBE, DPE, and EIIP were combined as follows:

$$H = (F(\text{Kmer}), F(\text{KSNC}), F(\text{MBE}), F(\text{DPE}), F(\text{EIIP})) \quad (3)$$

where H is the sequential combination of five different feature vectors with 1406 dimensions.

Meta-predictor

We generated a meta-classifier to check its potential in 6 mA site prediction. In brief, the probability scores of 30 prediction models (6 encodings × 5 ML classifier) were considered as a new feature vector as follows:

$$P_{met} = (P(M(1), \text{E}(1)), \ldots P(M(i), \text{E}(j)), \ldots, P(M(s), E(t))) \quad (4)$$

where P_{met} is the new feature vector, $P(M(s), \text{E}(j))$ the expected probability by each ML of $M(s)$ with encoding scheme $E(j)$, i the index of the ML, j the index of the encoding scheme, s the total of ML classifiers, and t the total of encodings.

Performance assessment metrics

To assess the performances of i6mA-Fuse, we used four standard measurements consisting of accuracy (Ac), sensitivity (Sn), specificity (Sp), and the Matthews correlation coefficient (MCC) (Basith et al. 2020; Ding et al. 2016; Yang et al. 2019b):

$$\text{Ac} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (5)$$

$$\text{Sn} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (6)$$

$$\text{Sp} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \quad (7)$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (8)$$

where TP and TN describe the number of positive samples correctly predicted and the number of negative samples correctly identified, respectively. Meanwhile, FN and FP indicate the number of positive samples falsely identified as negative ones and the number of negative samples falsely identified as positive ones, respectively. Furthermore, in order to assess the prediction performance of algorithms using threshold-independent parameters, AUC values were calculated by using the ROC curve.

Results and discussion

Nucleotide preference of *F. vesca* and *R. chinensis*

The position-specific preferences of nucleotide compositions were analyzed by the two-sample logo software (Vacic et al. 2006) as depicted in Fig. 2. Figure 2a, b display the nucleotide preferences on the DNA sequences having a length of 41 base pairs and the A base at the center for *F. vesca* and *R. chinensis*, respectively. We examined the DNA preferences of the sequences adjacent the A bases of *F. vesca* and *R. chinensis*, while the enriched nucleotides indicate a statistical significance at a level of $p < 0.05$ (two-sample t-test). In case of *F. vesca*. Figure 2a shows that the A base is enriched at positions 12, 13, 15, 17–20, 25, 28, 29, 32, and 33, while the G base is more enriched at positions 13, 20, 22, 23, 26, and 29 than other nucleotides. The T base was significantly depleted at position 13, 17–20, 23, 25, 26, 28, 29, and 38. In case of *R. chinensis*, the G base was enriched at positions 1, 3, 7–11, 13,14, 16, 22–24, 26, 29, 32, 34, 37, and 38, while the C base was more enriched at positions 2, 5–8, 10, 11, 14, 15, 19, 24, 27, 30, 35, 36, and 38–40 than other nucleotides. As seen in Fig. 2b, the T base was significantly depleted at position 1–4, 9–20, 22, 23, 25, 26, 28, 29, and 32–39. Enrichment and depletion of nucleotides at a specific position might be significant information for discriminating positives from and negatives on both the *F. vesca* and *R. chinensis* samples. Thus, in the present study, the mentioned significant position-specific preferences of nucleotide compositions are used as input features to develop the i6mA-Fuse.

Performance comparison by cross-validation test

We carried out a series of comparative simulations using RF models with the five feature vectors of Kmer, KSNC, MBE, DBE, and EIIP encodings, and evaluated their

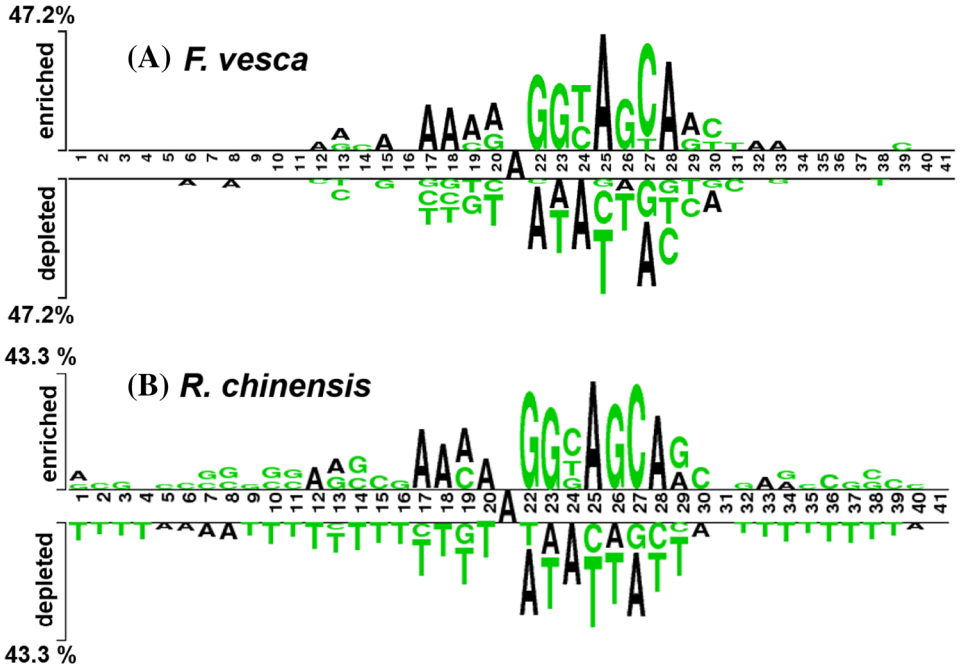

Fig. 2 Nucleotide preferences of the surrounding positive samples compared to negative samples. **a** *F. vesca*. **b** *R. chinensis*. The level of Y-axis is dissimilar due to the different datasets. Only nucleotides that are significantly enriched or depleted (*t*-test, $P < 0.05$) nearby the centered positive and negative samples are shown

performances on the training dataset by tenfold cross-validation. The cross-validation results are listed in Table 1 and Fig. 3. The previous studies suggested several ways to integrate multiple prediction models for improving the performances, including meta-predictor (Boopathi et al. 2019; Manavalan et al. 2019b, c, d), ensemble approach, and linear regression(Hasan et al. 2019a, d; Khatun et al. 2019a). Herein, the i6mA-Fuse linearly combined the five probability scores evaluated by the five, single encoding-employing RF models.

For *F. vesca*, MBE encoding achieved the highest prediction results with Ac = 0.925, Sn = 0.891, Sp = 0.958, MCC = 0.971, and AUC = 0.858. Meanwhile, the second highest prediction result was obtained by EIIP encoding, which gave Ac = 0.915, Sn = 0.879, Sp = 0.950, MCC = 0.839, and AUC = 0.963. These two encoding schemes yielded Ac values of 0.915–0.925, which were 11.6–24.6% higher than Ac values of the other three encodings. To construct the i6mA-Fuse, the optimal weight coefficients of Kmer, KSNC, MBE, DBE, and EIIP were 0.00, 0.00, 0.75, 0.00, and 0.25. While the three encodings of

Table 1 Cross-validation results of the proposed predictors and other five encodings

Genome	Method[a]	MCC	Ac	Sn	Sp	AUC	[b]P-value
F. vesca	Kmer	0.529	0.703	0.542	0.864	0.846	< 0.001
	KSNC	0.398	0.674	0.492	0.855	0.804	< 0.001
	MBE	0.858	0.925	0.891	0.958	0.971	0.293
	DBE	0.648	0.804	0.704	0.904	0.926	< 0.001
	EIIP	0.839	0.915	0.879	0.950	0.963	0.068
	i6mA-Fuse	0.873	0.934	0.908	0.957	0.981	–
R. chinensis	Kmer	0.559	0.734	0.548	0.921	0.877	< 0.001
	KSNC	0.477	0.721	0.528	0.912	0.844	< 0.001
	MBE	0.839	0.912	0.872	0.952	0.956	0.244
	DBE	0.608	0.793	0.684	0.902	0.902	< 0.001
	EIIP	0.816	0.900	0.866	0.933	0.945	0.028
	[a]i6mA-Fuse	0.851	0.916	0.881	0.950	0.968	–

[a]The i6mA-Fuse is constructed by the linear combination of the probability scores derived from RF models in conjunction with Kmer, KSNC, MBE, DBE, and EIIP encodings and their weight coefficients (*F. vesca* and *R. chinensis*) are (0.00 and 0.15), (0.00, 0.00), (0.75, 0.65), (0.00, 0.00) and (0.25, 0.20), respectively

[b]P-value was calculated with the AUC values and two-sample *t*-test

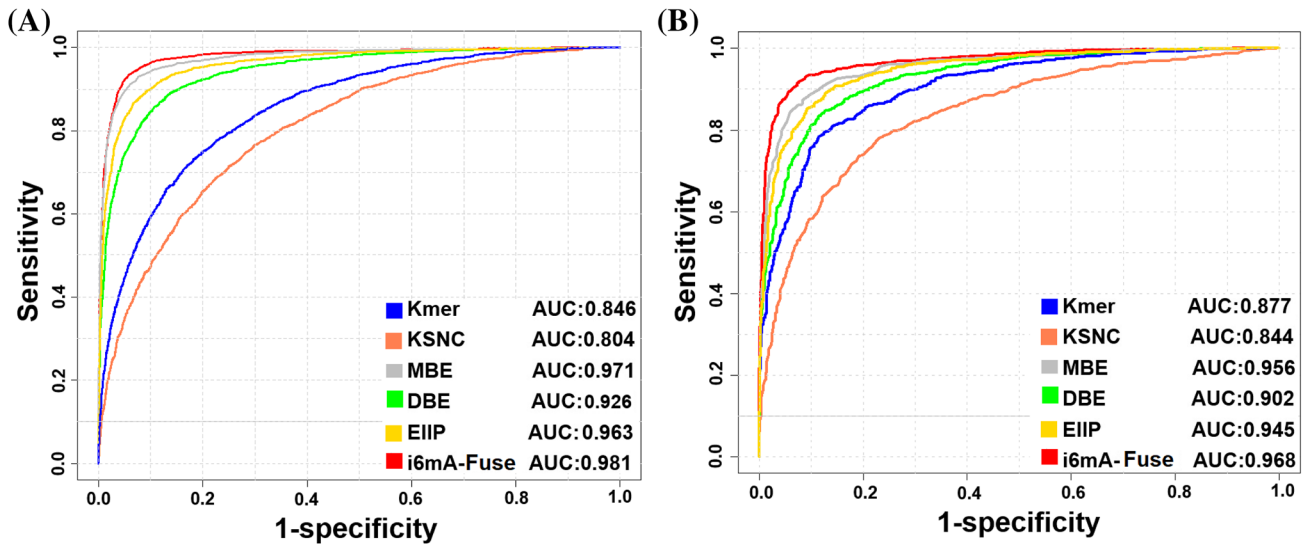

Fig. 3 ROC curves of i6mA-Fuse and the single encoding-employing models as evaluated by10-fold cross-validation. **a** *F. vesca*. **b** *R. chinensis*

Kmer, KSNC, and DBE did not contribute to any prediction performance, the MBE and EIIP contributed to 75% and 25% of the total prediction. The ROC curves of the i6mA-Fuse and the single encoding-employing models, evaluated by10-fold cross-validation, are presented in Fig. 3a. The i6mA-Fuse yielded AUC = 0.981, MCC = 0.873, Ac = 0.934, Sn = 0.908, and Sp = 0.957, surpassing all the single-employing models (Table 1). According to a *P*-value of 0.05, the i6mA-Fuse significantly outperformed the three models employing the single encoding of Kmer, KSNC, and DBE.

For *R. chinensis*, MBE achieved the best performance with an AUC value of 0.956 and Ac of 0.912 for all the single encoding-employing models. Meanwhile, the second highest prediction with an AUC of 0.945 and Ac of 0.900 was obtained by using EIIP encoding. To construct the i6mA-Fuse, we combined the RF scores with optimal weight coefficients of Kmer, KSNC, MBE, DBE, and EIIP of 0.15, 0.00, 0.65, 0.00, and 0.20, respectively, indicating that Kmer, MBE, and EIIP contribute to 15%, 65%, and 20% of the total prediction, respectively. The i6mA-Fuse yielded a peak AUC value of 0.968, while indicating MCC = 0.851, Ac = 0.916, Sn = 0.881, and Sp = 0.950 (Table 1). The i6mA-Fuse significantly outperformed the four models employing the single encoding of Kmer, KSNC, DBE, and EIIP (*P*-value of 0.05 by two-sample *t*-test).

Performance comparison among different ML algorithms by cross-validation test

To validate the effectiveness of the RF classifier in the i6mA-Fuse, we compared its performance with the five ML classifiers of SVM, AB, NB, ANN, and KNN on the training dataset. To make a fair comparison, we implemented the five ML classifiers in the same manners as the RF classifier. Figure S1 shows that the RF model provided better results than the other ML classifiers, while the prediction results of the SVM model were comparable to RF model in both two genomes. In *F. vesca*, the i6mA-Fuse achieved ~ 2–6% higher AUCs than any other ML-based combined models (Figure S1A). Meanwhile, the AUC values of the i6mA-Fuse were ~ 3–6% higher than those of the other combined models in *R. chinensis* (Figure S1B), thus demonstrating the superiority of RF.

Performance comparison of i6mA-Fuse with hybrid model and meta-predictor

We compared the linear regression model, employed by the i6mA-Fuse, with the two different models namely a hybrid model and meta-predictor. First, we concatenated the five feature encoding vectors of Kmer, KSNC, MBE, DPE, and EIIP and obtained 1406-dimensional feature vector. We generated the hybrid model for both species (*F. vesca* and *R. chinensis*), inputted these features to six different classifiers (RF, SVM, AB, NB, ANN, and KNN), and evaluated their performances by tenfold CV on the training datasets. In case of *F. vesca*, the hybrid model of RF, SVM, AB, NB, ANN, and KNN achieved AUC values of 0.978, 0.963, 0.957, 0.944, 0.919 and 0.933, respectively (Figure S2). Similarly, for *R. chinensis*, the hybrid model of RF, SVM, AB, NB, ANN, and KNN algorithms achieved AUCs of 0.958, 0.962, 0.927, 0.942, 0.918 and 0.922, respectively (Figure S2). Furthermore, we constructed the meta-predictor as described elsewhere (Manavalan et al. 2018a, 2019d; Wei et al. 2019). In *F. vesca* the i6mA-Fuse achieved higher AUCs than the hybrid model and meta-predictor (Figures S2 & S3). In *R.*

chinensis, the i6mA-Fuse showed ~ 1.0–8.0% higher AUCs than the hybrid model and meta-predictor (Figures S2 & 3).

Performance of i6mA-Fuse by independent test

The i6mA-Fuse was evaluated by independent tests for two genomes. We compared the prediction performance of the i6mA-Fuse with that of the five, single encoding-employing models (Kmer, KSNC, MBE, DBE, and EIIP) by using an independent dataset of *F. vesca*. MBE encoding gave higher performance than any other single–encoding employing models for two genomes, as shown in Fig. 4a, b. Moreover, the ROC curves displayed that the i6mA-Fuse achieved better than all the single encoding-employing methods. The i6mA-Fuse achieved outstanding performances (MCC = 0.858, Ac = 0.929, Sn = 0.915, Sp = 0.943, and AUC = 0.978) and (MCC = 0.869, Ac = 0.937, Sn = 0.928, Sp = 0.948, and AUC = 0.982) for *R. chinensis* and *F. vesca*, respectively, on the independent sets (Table 2).

Validation of i6mA-Fuse with other species datasets

To further examine the generalization of i6mA-Fuse, the proposed i6mA-Fuse was applied to identify 6 mA sites in other species, i.e. rice and mouse genomes. We collected the rice genome dataset from SDM6A (Basith et al. 2019), which contains 221 positive and 221 negative samples, and the mouse genome dataset from iDNA6mA-PseKNC (Feng et al. 2019), where we randomly selected 200 positive and 200 negative samples. The prediction performances of AUC, MCC, Sp, Sn, and Ac were shown in Table S1. The *R. chinensis*- and *F. vesca*-specific i6mA-Fuse yielded AUC values of 0.870 and 0.928 for the rice genome, respectively. For the mouse genome, they provided AUC values of 0.748 and 0.769, respectively. The i6mA-Fuse can be applicable to the rice genome, especially the *F. vesca*–specific i6mA-Fuse (Table S1). It suggests that the sequences surrounding 6 mA sites of Rosaceae genomes have common characteristic patterns to rice genomes.

Conclusions

The accurate prediction of 6 mA sites is one of the challenging tasks in bioinformatics. Because the experimental approaches are time-consuming and costly, it is desirable to develop a computational model for rapidly and accurately identifying 6 mA sites. Although several computational methods have been proposed in some species (Basith et al. 2019; Chen et al. 2019a; Feng et al. 2019; Lv et al. 2019a; Yu and Dai 2019), none of them were developed to specifically identify 6 mA sites in the Rosaceae genomes. In this study, we developed the first species-specific predictor named i6mA-Fuse for identifying 6 mA sites of the Rosaceae genomes, especially in *R. chinensis* and *F. vesca*. We constructed the random forest (RF) models with the five encoding schemes of Kmer, KSNC, MBE, DPE, and EIIP, and then combined the predicted probability scores of the five models through a linear regression. The resultant species-specific i6mA-Fuse achieved remarkably high

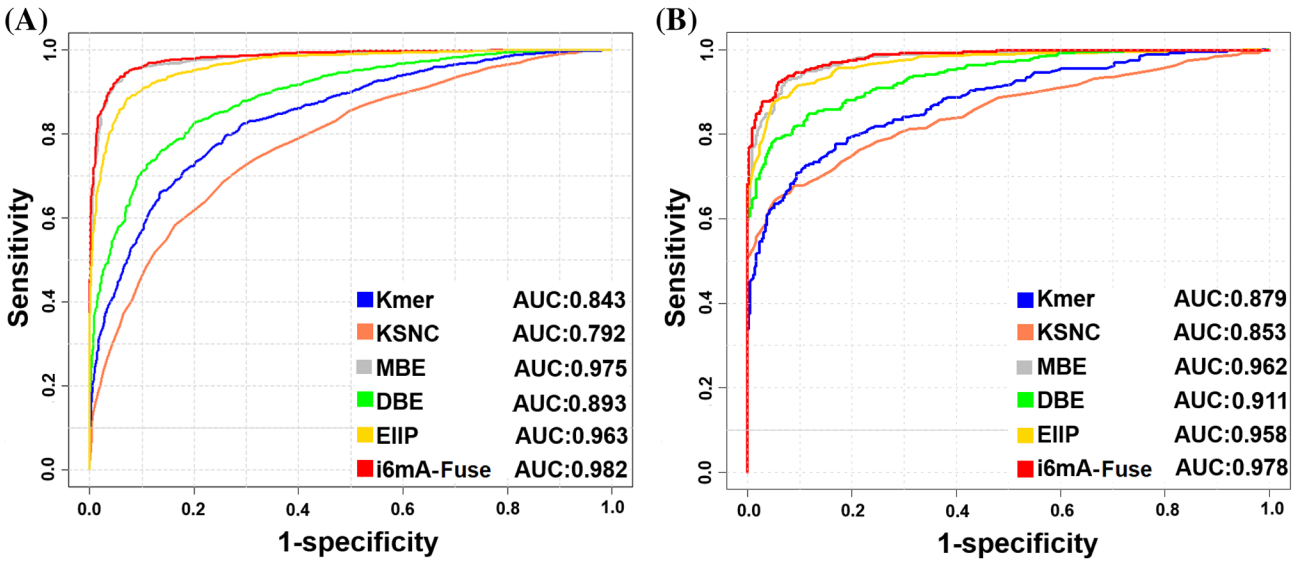

Fig. 4 ROC curves of i6mA-Fuse and the single encoding-employing models as evaluated by independent test. **a** *F. vesca*. **b** *R. chinensis*

Table 2 Independent test of the proposed predictors and other five encodings

Genome	Method	MCC	Ac	Sn	Sp	AUC
F. vesca	Kmer	0.532	0.761	0.655	0.857	0.843
	KSNC	0.434	0.708	0.582	0.836	0.792
	MBE	0.862	0.932	0.917	0.947	0.975
	DBE	0.643	0.810	0.704	0.893	0.893
	EIIP	0.842	0.916	0.878	0.948	0.963
	i6mA-Fuse	0.869	0.937	0.928	0.948	0.982
R. chinensis	Kmer	0.613	0.737	0.732	0.874	0.879
	KSNC	0.565	0.782	0.687	0.877	0.853
	MBE	0.835	0.918	0.901	0.935	0.962
	DBE	0.644	0.803	0.713	0.893	0.911
	EIIP	0.819	0.909	0.884	0.934	0.958
	[a]i6mA-Fuse	0.858	0.929	0.915	0.943	0.978

[a]The i6mA-Fuse is constructed by the linear combination of the probability scores derived from RF models in conjunction with Kmer, KSNC, MBE, DBE, and EIIP encodings and their weight coefficients (*F. vesca* and *R. chinensis*) are (0.00 and 0.15), (0.00, 0.00), (0.75, 0.65), (0.00, 0.00) and (0.25, 0.20), respectively

performances with AUCs of 0.982 and 0.978 and with MCCs of 0.869 and 0.858 on the independent datasets of *Rosa chinensis* and *Fragaria vesca*, respectively. In the *F. vesca*-specific i6mA-Fuse, the MBE and EIIP contributed to 75% and 25% of the total prediction; in the *R. chinensis*-specific i6mA-Fuse, Kmer, MBE, and EIIP contribute to 15%, 65%, and 20% of the total prediction. Interestingly, the i6mA-Fuse can be applicable to rice genome. To show the superiority of the linear regression, we compared it with the two combination methods of the hybrid model and meta-classifier. To further improve the prediction performance, we may use recently proposed integration methods (Chen et al. 2019b; Li et al. 2019; Zhang et al. 2019) and various modes of Chou's pseudo information (Chen et al. 2016; Chou 2011; Liu et al. 2015).To assist high-throughput identification for DNA 6 mA sites, the i6mA-Fuse is publicly accessible at https://kurata14.bio.kyutech.ac.jp/i6mA-Fuse/.

Acknowledgements This work is supported by the Grant-in-Aid for JSPS Research Fellow (19F19377) from Japan Society for the Promotion of Science (JSPS), partially supported from Japan Society for the Promotion of Science by Grant-in-Aid for Scientific Research (B) (19H04208) and by the developing key technologies for discovering and manufacturing pharmaceuticals used for next-generation treatments and diagnoses both from the Ministry of Economy, Trade and Industry, Japan (METI) and from Japan Agency for Medical Research and Development (AMED).

Author contributions MMH and HK conceived the project. MMH and KMS collected and analyzed the datasets. MMH drafted the manuscript. HK, MMH, MB, SW and KMS thoroughly revised the manuscript. All authors approved and read the final manuscript.

References

Basith S, Manavalan B, Shin TH, Lee G (2019) SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. Mol Ther Nucleic Acids 18:131–141. https://doi.org/10.1016/j.omtn.2019.08.011

Basith S, Manavalan B, Shin TH, Lee G (2020) Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. Med Res Rev. https://doi.org/10.1002/med.21658

Boopathi V, Subramaniyam S, Malik A, Lee G, Manavalan B, Yang DC (2019) mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides. Int J Mol Sci. https://doi.org/10.3390/ijms20081964

Charoenkwan P, Shoombuatong W, Lee HC, Chaijaruwanich J, Huang HL, Ho SY (2013) SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. PLoS ONE 8:e72368. https://doi.org/10.1371/journal.pone.0072368

Chen W, Lin H, Chou KC (2015) Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol BioSyst 11:2620–2634. https://doi.org/10.1039/c5mb00155b

Chen W, Lv H, Nie F, Lin H (2019a) i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics 35:2796–2800. https://doi.org/10.1093/bioinformatics/btz015

Chen W, Tang H, Ye J, Lin H, Chou KC (2016) iRNA-PseU: identifying RNA pseudouridine sites. Mol Ther Nucleic Acids 5:e332. https://doi.org/10.1038/mtna.2016.37

Chen Z et al (2019b) Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. Brief Bioinform. https://doi.org/10.1093/bib/bbz112

Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273:236–247. https://doi.org/10.1016/j.jtbi.2010.12.024

Chou KC (2019) Advance in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs. Curr Med Chem. https://doi.org/10.2174/0929867326666190507082559

Ding H, Yang W, Tang H, Feng PM, Huang J, Chen W, Lin H (2016) PHYPred: a tool for identifying bacteriophage enzymes and

hydrolases. Virol Sin 31:350–352. https://doi.org/10.1007/s12250-016-3740-6
Du K et al (2019) Epigenetically modified N(6)-methyladenine inhibits DNA replication by human DNA polymerase eta. DNA Repair 78:81–90. https://doi.org/10.1016/j.dnarep.2019.03.015
Feng P, Yang H, Ding H, Lin H, Chen W, Chou KC (2019) iDNA6mA-PseKNC: identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics 111:96–102. https://doi.org/10.1016/j.ygeno.2018.01.005
Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics 20:2479–2481. https://doi.org/10.1093/bioinformatics/bth261
Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152. https://doi.org/10.1093/bioinformatics/bts565
Hasan MM, Khatun MS, Kurata H (2019a) Large-scale assessment of bioinformatics tools for lysine succinylation sites. Cells. https://doi.org/10.3390/cells8020095
Hasan MM, Manavalan B, Khatun MS, Kurata H (2019b) i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome. Int J Biol Macromol. https://doi.org/10.1016/j.ijbiomac.2019.12.009
Hasan MM, Manavalan B, Khatun MS, Kurata H (2019c) Prediction of S-nitrosylation sites by integrating support vector machines and random forest. Mol Omics 15:451–458. https://doi.org/10.1039/c9mo00098d
Hasan MM, Rashid MM, Khatun MS, Kurata H (2019d) Computational identification of microbial phosphorylation sites by the enhanced characteristics of sequence information. Sci Rep 9:8258. https://doi.org/10.1038/s41598-019-44548-x
Huang Q, Zhang J, Wei L, Guo F, Zou Q (2020) 6mA-RicePred: a method for identifying DNA N(6)-methyladenine sites in the rice genome based on feature fusion. Front Plant Sci 11:4. https://doi.org/10.3389/fpls.2020.00004
Jia C, Yang Q, Zou Q (2018) NucPosPred: predicting species-specific genomic nucleosome positioning via four different modes of general PseKNC. J Theor Biol 450:15–21. https://doi.org/10.1016/j.jtbi.2018.04.025
Khatun MS, Hasan MM, Kurata H (2019a) PreAIP: computational prediction of anti-inflammatory peptides by integrating multiple complementary features. Front Genet 10:129. https://doi.org/10.3389/fgene.2019.00129
Khatun S, Hasan M, Kurata H (2019b) Efficient computational model for identification of antitubercular peptides by integrating amino acid patterns and properties. FEBS Lett 593:3029–3039. https://doi.org/10.1002/1873-3468.13536
Li F et al (2019) DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz721
Liaw A, Wiener M (2002) Classification and regression by random forest. R News 2:18–22
Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC (2015) Pse-in-One: a web server for generating various modes of pseudo components of DNA RNA, and protein sequences. Nucleic Acids Res 43:W65–71. https://doi.org/10.1093/nar/gkv458
Liu B, Fang L, Long R, Lan X, Chou KC (2016) iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics 32:362–369. https://doi.org/10.1093/bioinformatics/btv604
Liu B, Li K, Huang DS, Chou KC (2018a) iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach. Bioinformatics 34:3835–3842. https://doi.org/10.1093/bioinformatics/bty458
Liu X, Lai W, Zhang N, Wang H (2018b) Predominance of N(6)-methyladenine-specific DNA fragments enriched by multiple immunoprecipitation. Anal Chem 90:5546–5551. https://doi.org/10.1021/acs.analchem.8b01087
Liu ZY et al (2019) MDR: an integrative DNA N6-methyladenine and N4-methylcytosine modification database for Rosaceae. Hortic Res 6:78. https://doi.org/10.1038/s41438-019-0160-4
Lv H et al (2019a) iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice. Front Genet 10:793. https://doi.org/10.3389/fgene.2019.00793
Lv Z, Jin S, Ding H, Zou Q (2019b) A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features. Front Bioeng Biotechnol 7:215. https://doi.org/10.3389/fbioe.2019.00215
Manavalan B, Shin TH, Lee G (2018) DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. Oncotarget 9:1944–1956. https://doi.org/10.18632/oncotarget.23099
Manavalan B, Shin TH, Kim MO, Lee G (2018) AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest. Front Pharmacol 9:276. https://doi.org/10.3389/fphar.2018.00276
Manavalan B, Basith S, Shin TH, Wei L, Lee G (2018a) mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty1047
Manavalan B, Basith S, Shin TH, Lee DY, Wei L, Lee G (2019a) 4mCpred-EL: an ensemble learning framework for identification of DNA N(4)-methylcytosine sites in the mouse. Genome Cells. https://doi.org/10.3390/cells8111332
Manavalan B, Basith S, Shin TH, Wei L, Lee G (2019b) AtbPpred: a robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees. Comput Struct Biotechnol J 17:972–981. https://doi.org/10.1016/j.csbj.2019.06.024
Manavalan B, Basith S, Shin TH, Wei L, Lee G (2019c) mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics 35:2757–2765. https://doi.org/10.1093/bioinformatics/bty1047
Manavalan B, Basith S, Shin TH, Wei L, Lee G (2019d) Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. Mol Ther Nucleic Acids 16:733–744. https://doi.org/10.1016/j.omtn.2019.04.019
McIntyre ABR, Alexander N, Grigorev K, Bezdan D, Sichtig H, Chiu CY, Mason CE (2019) Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. Nat Commun 10:579. https://doi.org/10.1038/s41467-019-08289-9
O'Brown ZK, Greer EL (2016) N6-methyladenine: a conserved and dynamic DNA mark. Adv Exp Med Biol 945:213–246. https://doi.org/10.1007/978-3-319-43624-1_10
Qianfei Huang F, Zhang Z, Wei L, Guo F, Zou Q (2020) 6mA-RicePred: a method for identifying DNA N6-methyladenine sites in the rice genome based on feature fusion. Front Plant Sci. https://doi.org/10.3389/fpls.2020.00004
Schaduangrat N, Nantasenamat C, Prachayasittikul V, Shoombuatong W (2019) Meta-iAVP: a sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation. Int J Mol Sci. https://doi.org/10.3390/ijms20225743
Shoombuatong W, Schaduangrat N, Pratiwi R, Nantasenamat C (2019) THPep: a machine learning-based approach for predicting tumor homing peptides. Comput Biol Chem 80:441–451. https://doi.org/10.1016/j.compbiolchem.2019.05.008
Su R, Hu J, Zou Q, Manavalan B, Wei L (2019) Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. Brief Bioinform. https://doi.org/10.1093/bib/bby124

Sun S, Wang C, Ding H, Zou Q (2020) Machine learning and its applications in plant molecular studies. Brief Funct Genom 19:40–48. https://doi.org/10.1093/bfgp/elz036

Vacic V, Iakoucheva LM, Radivojac P (2006) Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics 22:1536–1537. https://doi.org/10.1093/bioinformatics/btl151

Wang X, Yan R (2018) RFAthM6A: a new tool for predicting m(6)A sites in *Arabidopsis thaliana*. Plant Mol Biol 96:327–337. https://doi.org/10.1007/s11103-018-0698-9

Wei L, Su R, Luan S, Liao Z, Manavalan B, Zou Q, Shi X (2019) Iterative feature representations improve N4-methylcytosine site prediction. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz408

Win TS, Malik AA, Prachayasittikul V, Wikberg SJE, Nantasenamat C, Shoombuatong W (2017) HemoPred: a web server for predicting the hemolytic activity of peptides. Future Med Chem 9:275–291. https://doi.org/10.4155/fmc-2016-0188

Xiong J, Ye TT, Ma CJ, Cheng QY, Yuan BF, Feng YQ (2019) N 6-Hydroxymethyladenine: a hydroxylation derivative of N6-methyladenine in genomic DNA of mammals. Nucleic Acids Res 47:1268–1277. https://doi.org/10.1093/nar/gky1218

Xu ZC, Feng PM, Yang H, Qiu WR, Chen W, Lin H (2019) iRNAD: a computational tool for identifying D modification sites in RNA sequence. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz358

Yang W, Zhu XJ, Huang J, Ding H, Lin H (2019) A brief survey of machine learning methods in protein sub-Golgi localization. Curr Bioinform 14:234–240

Yang H, Yang W, Dao FY, Lv H, Ding H, Chen W, Lin H (2019) A comparison and assessment of computational method for identifying recombination hotspots in *Saccharomyces cerevisiae*. Brief Bioinform. https://doi.org/10.1093/bib/bbz123

Yu H, Dai Z (2019) SNNRice6mA: a deep learning method for predicting DNA N6-methyladenine sites in rice genome. Front Genet 10:1071. https://doi.org/10.3389/fgene.2019.01071

Zhang G et al (2015) N6-methyladenine DNA modification in Drosophila. Cell 161:893–906. https://doi.org/10.1016/j.cell.2015.04.018

Zhang Q et al (2018) N(6)-methyladenine DNA methylation in Japonica and Indica rice genomes and its association with gene expression, plant development, and stress responses. Mol Plant 11:1492–1508. https://doi.org/10.1016/j.molp.2018.11.005

Zhang Y et al (2019) PeNGaRoo, a combined gradient boosting and ensemble learning framework for predicting non-classical secreted proteins. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz629

Zhou Y, Zeng P, Li YH, Zhang Z, Cui Q (2016) SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Res 44:e91. https://doi.org/10.1093/nar/gkw104

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

iBitter-SCM: Identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides

Phasit Charoenkwan[a], Janchai Yana[b], Nalini Schaduangrat[c], Chanin Nantasenamat[c], Md. Mehedi Hasan[d], Watshara Shoombuatong[c,*]

[a] Modern Management and Information Technology, College of Arts, Media and Technology, Chiang Mai University, Chiang Mai 50200, Thailand
[b] Department of Chemistry, Faculty of Science and Technology, Chiang Mai Rajabhat University, 50300, Thailand
[c] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[d] Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

ARTICLE INFO

Keywords:
Bitter peptide
Scoring card method
Propensity score
Interpretable model
Machine learning
Classification

ABSTRACT

In general, hydrolyzed proteins, plant-derived alkaloids and toxins displays unpleasant bitter taste. Thus, the perception of bitter taste plays a crucial role in protecting animals from poisonous plants and environmental toxins. Therapeutic peptides have attracted great attention as a new drug class. The successful identification and characterization of bitter peptides are essential for drug development and nutritional research. Owing to the large volume of peptides generated in the post-genomic era, there is an urgent need to develop computational methods for rapidly and effectively discriminating bitter peptides from non-bitter peptides. To the best of our knowledge, there is yet no computational model for predicting and analyzing bitter peptides using sequence information. In this study, we present for the first time a computational model called the iBitter-SCM that can predict the bitterness of peptides directly from their amino acid sequence without any dependence on their functional domain or structural information. iBitter-SCM is a simple and effective method that was built using the scoring card method (SCM) with estimated propensity scores of amino acids and dipeptides. Our benchmarking results demonstrated that iBitter-SCM achieved an accuracy and Matthews coefficient correlation of 84.38% and 0.688, respectively, on the independent dataset. Rigorous independent test indicated that iBitter-SCM was superior to those of other widely used machine-learning classifiers (e.g. k-nearest neighbor, naive Bayes, decision tree and random forest) owing to its simplicity, interpretability and implementation. Furthermore, the analysis of estimated propensity scores of amino acids and dipeptides were performed to provide a better understanding of the biophysical and biochemical properties of bitter peptides. For the convenience of experimental scientists, a web server is provided publicly at http://camt.pythonanywhere.com/iBitter-SCM. It is anticipated that iBitter-SCM can serve as an important tool to facilitate the high-throughput prediction and de novo design of bitter peptides.

1. Introduction

Abnormal bitterness might be associated with dietary danger. In general, hydrolyzed proteins, plant-derived alkaloids and toxins exhibit unpleasant bitter taste. Thus, the bitter taste perception plays a crucial role in protecting animals from poisonous plants and environmental toxins [1]. The taste perception of humans can be categorized into four well-known groups: sweet, bitter, sour and salty, in addition to two controversial groups, i.e. fat taste and amino acid taste [2]. Although, abnormal or extreme bitterness tends to be associated with dietary danger, a number of diverse plant-derived food produce bitterness such as cucumber, pumpkin, zucchini, squash, lettuce, spinach and kale. In addition, many bitter compounds are important drugs or drug candidates encompassing ions, alkaloids, polyphenols, glucosinolates and peptides. Proteolytic hydrolysis of peptides and proteins have been known to make foods unfavorable [3,4]. In this process, caseins are digested into peptides containing bulky hydrophobic groups at their C-terminal region [3]. Hence, the hydrophobic property of the amino acid side chain at the C-terminus can be attributed to its bitterness. The successful identification and characterization of bitter peptides is essential for drug development and nutritional research.

High-throughput experimental approaches for identifying bitter peptides are time-consuming and costly, thus the development of accurate and fast computational methods is in great demand. Particularly,

* Corresponding author.
E-mail address: watshara.sho@mahidol.ac.th (W. Shoombuatong).

https://doi.org/10.1016/j.ygeno.2020.03.019
Received 21 February 2020; Received in revised form 19 March 2020; Accepted 22 March 2020
0888-7543/ © 2020 Elsevier Inc. All rights reserved.

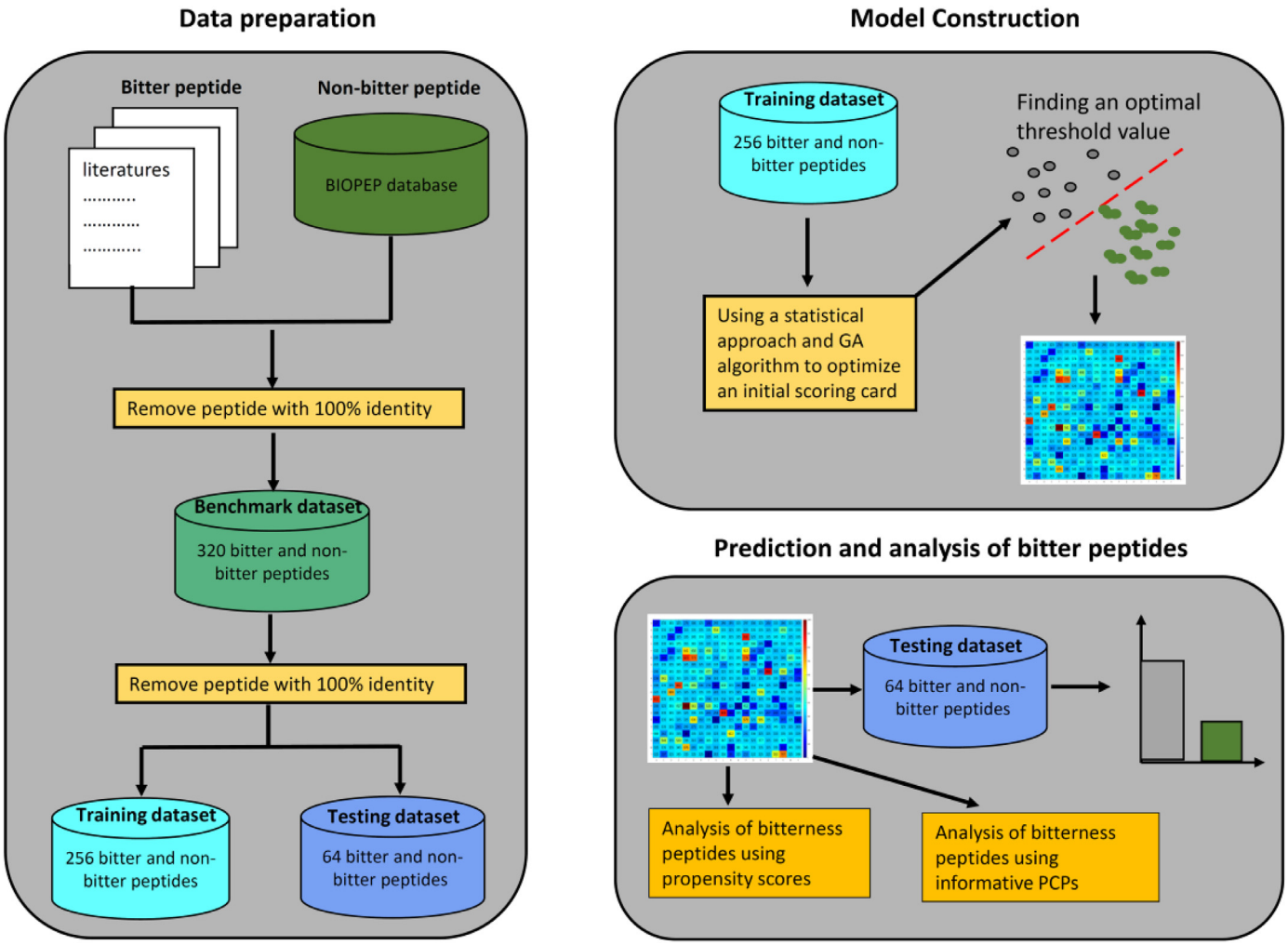

Fig. 1. Schematic framework of iBitter-SCM for predicting and analyzing bitter peptides.

such computational approach is based on quantitative structure–-activity relationship (QSAR) modeling. QSAR is a ligand-based approach that seeks to discern the mathematical relationship between various types of descriptors (X) and their investigated biological activity (Y) through the use of machine learning (ML) models [5]. As mentioned in the Organization for Economic Co-operation and Development (OECD) guideline [6–8], the development of robust QSAR models entails the following characteristics: (i) a defined endpoint; (ii) an unambiguous algorithm; (iii) a defined domain of applicability; (iv) appropriate measures of goodness-of-fit, robustness, and predictive ability; and (v) a mechanistic interpretation.

As of now, several QSAR models have been published on the prediction of bitterness of peptides [9–14] by primarily relying on information pertaining to the contribution from the primary structure of therapeutic peptides to their observed potency [10,15]. For example, Yin et al. [11] developed 28 QSAR models to predict the bitterness of dipeptides. They proposed the quantitative multidimensional amino acid descriptors E (E1–E5) whereby E1, E2, E3, E4 and E5 represents hydrophobicity, steric properties or side chain bulk/molecular size, preferences for amino acids to occur in α-helices, composition and the net charge, respectively. In their QSAR model development, the analysis is based on support vector regression (SVR) in which they were used for constructing models for analyzing 48 angiotensin-converting enzyme (ACE) inhibitors dipeptides, 55 ACE inhibitors tripeptides and 48 bitter dipeptides. They reported that their optimal QSAR model yielded leave-one-out cross-validation and root mean squared error of 0.886, − 0.985 and 0.021, − 0.250, respectively. In 2013, Soltani et al. [16] analyzed 229 experimental bitterness values of 224 peptides and 5 amino acids, particularly the bitter thresholds ($\log(1/T)$) in which T is the bitter-tasting threshold concentration (M). Three well-known ML methods (e.g. multiple linear regression (MLR), support vector machine (SVM) and artificial neural network (ANN)) were used for model building. Particularly, the set of 224 peptides were described by a set of 3D consisting of 1295 descriptors. Recently, Xu and Chung [17] proposed QSAR models for predicting bitter peptides by integrating fourteen amino acid descriptors. Their cross-validation dataset contained 48 dipeptides, 52 tripeptides and 23 tetrapeptides along with their bitter taste thresholds. Cross-validation results for di-, tri- and tetrapeptides were 0.941 ± 0.001, 0.742 ± 0.004 and 0.956 ± 0.002, respectively.

In 2016, Huang et al. [18] proposed an open-access tool known as BitterX for identifying human bitter taste receptors, TAS2Rs, for small molecules. BitterX is the first online available tool for predicting ligands for bitter taste receptors in human and vice versa. Three different algorithms, namely sequential minimal optimization (SMO), logistic regression (LR) and random forest (RF), were used to implement the prediction model in discriminating bitter from non-bitter compounds. In order to avoid overfitting, the authors randomly divided the original dataset into training (70%) and hold-out test (30%) sets. Accuracies on the training and hold-out test sets were 0.93 and 0.83, respectively. Recently, Dagan-Wiener et al. [19] developed an ML classifier named BitterPredict for predicting whether a compound is bitter or not on the basis of their chemical structures. In BitterPredict, non-bitter molecules were gathered from the literature and used as the negative set while

Huang et al. [18] used chemical structures which were not described as bitter from the Available Chemicals Directory (ACD, http://www.accelrys.com). In the hold-out test, BitterPredict yielded in excess of 80% accuracy.

To the best of our knowledge, no method has yet been developed for predicting and characterizing bitter peptides on the basis of their sequence information. In the present work, we propose a novel computational model named iBitter-SCM for predicting and analyzing bitter peptides. A dataset containing 320 bitter and non-bitter peptides that was manually gathered from the literature [3,9–14,17,20–26] and from BIOPEP [27], respectively, were used for developing the iBitter-SCM predictor proposed herein. In the iBitter-SCM, we employed a newly developed scoring card method (SCM) [28–30] for estimating the propensity scores of amino acids and dipeptides. Fig. 1 shows the workflow of iBitter-SCM which was used in predicting and analyzing bitter peptides.To prove the effectiveness of the proposed model, we compared iBitter-SCM with those of other widely used ML models consisting of SVM, RF, k-nearest neighbor (KNN), naive Bayes (NB) and decision tree (DT). Comparative results using rigorous independent test indicated that iBitter-SCM was superior to those of other widely used ML classifiers on the basis of its simplicity, interpretability and implementation. Furthermore, the estimated propensity scores of amino acids and dipeptides were used to unravel the origin of biochemical and biophysical properties of investigated bitter peptides. Finally, for the convenience of experimental scientists, the iBitter-SCM web server was established and made freely available online at http://camt.pythonanywhere.com/iBitter-SCM. It is anticipated that iBitter-SCM can serve as an important tool to facilitate the high-throughput prediction and de novo design of bitter peptides.

2. Materials and methods

2.1. Benchmark datasets

In order to prepare a high quality benchmark dataset that can guarantee the capability and reliability of the proposed model, the following procedures [24,30–40] were performed. Firstly, only peptides with experimentally confirmed bitterness were included. In the present study, we manually collected experimentally validated bitter peptides from various literatures [3,9–14,17,20–26]. Secondly, peptides containing ambiguous residues, such as X, B, U and Z, were excluded. Thirdly, duplicate peptide sequences were removed. After performing such pre-processing, all these peptides were unique and considered as the positive dataset consisting of 320 bitter peptides. In the case of the negative dataset, very few experimentally validated non-bitter peptides were obtained due to less scientific significance [41]. Thus, the standard process from previous studies [42–44] for the construction of a negative dataset was used by randomly generating 320 peptides from BIOPEP [27]. Because, the BIOPEP database is well-known and universally accepted database. These random peptides were considered as non-bitter peptides. Finally, the benchmark dataset named BTP640 containing 320 bitter and non-bitter peptides was obtained. To avoid overestimation in the prediction model, BTP640 dataset was randomly divided into training (named BTP-CV) and independent sets (named BTP-TS) with a ratio of 8:2. Finally, the BTP-CV dataset consisted of 256 bitter and non-bitter peptides, while the BTP-TS dataset consisted of 64 bitter and non-bitter peptides.

2.2. Feature representation

Given a protein sequence (**P**), it can be represented as:

$$\mathbf{P} = p_1p_2p_3\cdots p_{1N} \quad (1)$$

where p_i and N denote the i^{th} residue in the protein **P** and the peptide length, respectively. Note that the residue p_i is in the set of natural amino acid, i.e. A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y. It is well-known that DPC is one of the most interpretable and effective features for predicting and analyzing various types of protein and peptide functions [23,24,28–30,38–40,45–48]. In general, DPC are the proportion of dipeptides in a peptide sequence P that are expressed as a fixed length of 400. Thus, a protein or peptide sequence **P** can be expressed by a vector with 400D (dimension), as formulated by:

$$\mathbf{P} = [dp_1, dp_2, ..., dp_{400}]^T \quad (2)$$

where T is the transposed operator, while dp_1, dp_2…, dp_{400} is occurrence frequencies of the 400 native dipeptides, respectively, in a protein or peptide sequence P.

2.3. Scoring card method

SCM method is a general-purpose approach for predicting and analyzing protein and peptide functions from their amino acid sequence without any dependence on their functional domain or structural information by estimating propensity scores of 400 dipeptides and 20 amino acids to be the protein and peptide with the investigated function [28–30]. More specifically, this method is an efficient and simple approach that performs well not only for predicting proteins using an efficient function, but also for gaining insight into the characteristic of proteins based on the estimated propensity scores of amino acids and dipeptides [24,30,38–40].

In the present study, we employed SCM method to develop a predictor for discriminating bitter peptides from non-bitter peptides and to characterize the biophysical and biochemical properties of bitter peptides. In iBitter-SCM, six main steps are involved in the development of the model as follows: (i) preparing the training (BTP-CV) and independent (BTP-TS) datasets, (ii) calculating the initial dipeptide propensity score (init-DPS) using a statistical approach, (iii) optimizing init-DPS to obtain the augmented dipeptide propensity score (opti-DPS) using a genetic algorithm (GA), (iv) estimating amino acid propensity scores using a statistical approach [46], (v) discriminating bitter peptides from non-bitter peptides using a weighted-sum with opti-DPS, and (vi) bitter peptides characterization using the propensity scores of amino acids and dipeptides. More details for the SCM method can be found in our previous studies [28–30].

2.4. Characterization of the bitter taste of peptides

To analyze the biophysical and biochemical properties of the bitterness on peptide sequences, we estimated the propensity scores of amino acid and dipeptides using the SCM method [23,24,28–30,38–40]. The values of propensity scores represent their influence on each amino acid and dipeptide on the biological, functional and structural properties of bitter peptides. Furthermore, the informative physicochemical property (PCP) of amino acids extracted from the amino acid index database (AAindex) [49] were identified using the proposed iBitter-SCM. PCPs are one of the most intuitive features associated with biophysical and biochemical reactions. Previously, our studies utilized these features for predicting and analyzing various functions of proteins and peptides from primary sequences [23,24,28–30,38–40]. More details of the determination of informative PCPs can be found in our previous studies [28–30].

2.5. Performance evaluation

In order to evaluate the prediction ability of the model, we used four metrics widely used for the problem of binary classification problems, i.e. accuracy (Ac), sensitivity (Sn), specificity (Sp) and Matthews coefficient correlation (MCC) [6–8,31,50–56]. To further investigate the prediction performance of the proposed model using threshold-independent parameters, receiver operating characteristic (ROC) curves were plotted. The area under the ROC curve (auAUC) was utilized to

Table 1
Comparison of BLAST tool using various *E*-value cut-offs.

E-value	Ac (%)	Sn (%)	Sp (%)
0.1	65.08	34.92	95.24
0.01	56.35	15.87	96.83
0.001	53.97	11.11	96.83
0.0001	50.79	4.76	96.83

assess the prediction performance, where AUC values of 0.5 and 1 are indicative of random and perfect models, respectively.

2.6. Reproducible research

To guarantee and ensure the reproducibility of the proposed model, herein, all codes and datasets used in the construction of the predictive models are available on GitHub at https://github.com/Shoombuatong2527/Benchmark-datasets.

3. Results

3.1. Prediction results using BLAST

BLAST is a commonly-used tool for identifying homologous sequences based on the sequence similarity concept [57]. To construct a BLAST-based predictor, the BTP-CV dataset was employed as the BLASTP database [58], while the BTP-TS dataset were considered as the BLASTP query sequences. Table 1 lists the performance comparison of BLASTP with *E*-values ranging from 0.1 to 0.0001. The highest accuracy of 65.08% was observed at an E-value cut-off set to 0.1, while the lowest accuracy of 50.79% was observed at the E-value cut-off set to 0.0001. Based on the prediction results from Table 1, it could be noticed that the capability of similarity-based search methods is insufficient for predicting bitter peptides. Therefore, it is desirable to develop an intelligent computational model based on ML methods for the prediction of bitter peptides.

3.2. Prediction performance of iBitter-SCM

In this study, we made an effort to develop a sequence-based predictor named iBitter-SCM to discriminate bitter peptides from non-bitter peptides by utilizing the SCM model as the prediction model using propensity scores of dipeptides as the input feature. Due to the non-deterministic characteristic of the GA algorithm [28,30,59], ten SCM models in conjunction with ten different opti-DPS were performed to generate ten experiments based on the BTP-CV dataset. In this section, a comparative study was carried out to determine the optimal opti-DPS that is most capable for bitter peptide prediction through the execution of performance comparisons among ten SCM models. Table S1 and Table 2 list the details of performance comparisons of the ten SCM models over 10-fold CV and independent test, respectively.

We observed that the three top-ranked opti-DPS having the highest performances according to their prediction results (Ac, MCC, auROC) were the opti-DPS from experiments #7 (87.30%, 0.753, 0.906), #5 (87.11%, 0.751, 0.903) and #2 (87.30%, 0.750, 0.903), respectively. In case of the independent test, it could be noticed that the three top-ranked opti-DPS having the highest test prediction results are the opti-DPS from experiments #5 (84.38%, 0.688, 0.904), #9 (84.38%, 0.688, 0.890) and #7 (83.59%, 0.672, 0.860), respectively. As could be seen in Table S1 and Table 2, the opti-DPS from experiments #5 and #7 are found in the three top-ranked experiments over both 10-fold CV and independent test at ranks (2, 1) and (1, 3), respectively. Meanwhile, the opti-DPS from experiment #5 achieved very comparable results to experiment #7 (87.11% vs 87.30%). Furthermore, due to the fact that the independent test is the most rigorous cross-validation method to demonstrate the robustness and reliability of the model, it could be noted that the opti-DPS from experiment #5 significantly contributes to bitter peptide prediction. For convenience, the best predictor based on SCM method in conjunction with the opti-DPS from experiment #5 will be referred to as iBitter-SCM. Meanwhile, this optimal opti-DPS was also used for providing a good understanding of the biochemical and biophysical properties of bitter peptides, as depicted in Fig. 2.

Table 2
Comparison of ten SCM models with ten different optimized dipeptide propensity scores (opti-DPS) over independent test.

#Exp.	Fitness score	Threshold	Ac (%)	Sn (%)	Sp (%)	MCC	auROC
1	0.901	334	82.81	81.25	84.38	0.657	0.896
2	0.909	332	82.81	84.38	81.25	0.657	0.865
3	0.904	343	82.81	76.56	89.06	0.661	0.881
4	0.911	331	82.81	85.94	79.69	0.658	0.871
5	**0.909**	**333**	**84.38**	**84.38**	**84.38**	**0.688**	**0.904**
6	0.908	334	82.81	84.38	81.25	0.657	0.872
7	0.911	334	83.59	82.81	84.38	0.672	0.860
8	0.908	333	82.81	84.38	81.25	0.657	0.884
9	0.901	333	84.38	85.94	82.81	0.688	0.890
10	0.907	333	82.81	85.94	79.69	0.658	0.893
Mean	0.907	334.000	83.20	83.59	82.81	0.665	0.882
STD.	0.004	3.300	0.66	2.88	2.85	0.013	0.014

The experiment #5 having the optimal prediction performances over both 10-fold CV and independent test is used for further analysis.

3.3. Contribution and effectiveness of the estimated propensities of dipeptides

In our feature representation, the init-DPS is calculated using normalized dipeptide composition, while the opti-DPS is generated by optimizing init-DPS with a GA algorithm as described in the *Scoring card method* section. As mentioned above, the optimal opti-DPS derived from experiment #5 significantly contributes to bitter peptide prediction. In this section, to investigate the effectiveness of this opti-DPS, we compare its performance with init-DPS as assessed by 10-fold CV and independent test, shown in Table 3.

As shown in Table 3, the Ac and MCC values of the optimal opti-DPS are 87.11% and 0.751, which are 2% and 4% higher than init-DPS over 10-fold CV. Furthermore, the optimal opti-DPS outperforms the init-DPS with improvements of 3%, 2%, 8% and 6% on Ac, Sn, Sp and MCC, respectively. Moreover, a histogram is also used to demonstrate the discriminative power of optimal opti-DPS and init-DPS for bitter peptide prediction as depicted in Fig. 3(a) and (b), respectively, where blue and red lines, represent the score distribution of bitter and non-bitter peptides, respectively. In practice, more overlap between the red and blue bars demonstrate that the feature is less capable in bitter peptide prediction. Thus, it could be clearly seen in Fig. 3 that the opti-DPS has a great ability for discriminating bitter peptides from non-bitter peptides.

3.4. Comparison of iBitter-SCM with conventional classifiers

In order to prove the effectiveness of the proposed model, comparisons for its performance with related and existing methods were conducted [60,61]. To the best of our knowledge, iBitter-SCM is the first sequence-based model for the prediction and characterization of bitter peptides. Therefore, in this study, we selected widely used ML models to compare our proposed model iBitter-SCM on the same benchmark dataset, i.e. BTP-CV and BTP-TS datasets. Here, we selected KNN, NB, DT, SVM and RF models, and these five ML models were implemented with the same feature representation and cross-validation methods using the Scikit-Learn package [62]. This package has been successfully applied to many protein and peptide prediction problems [32–37,63].

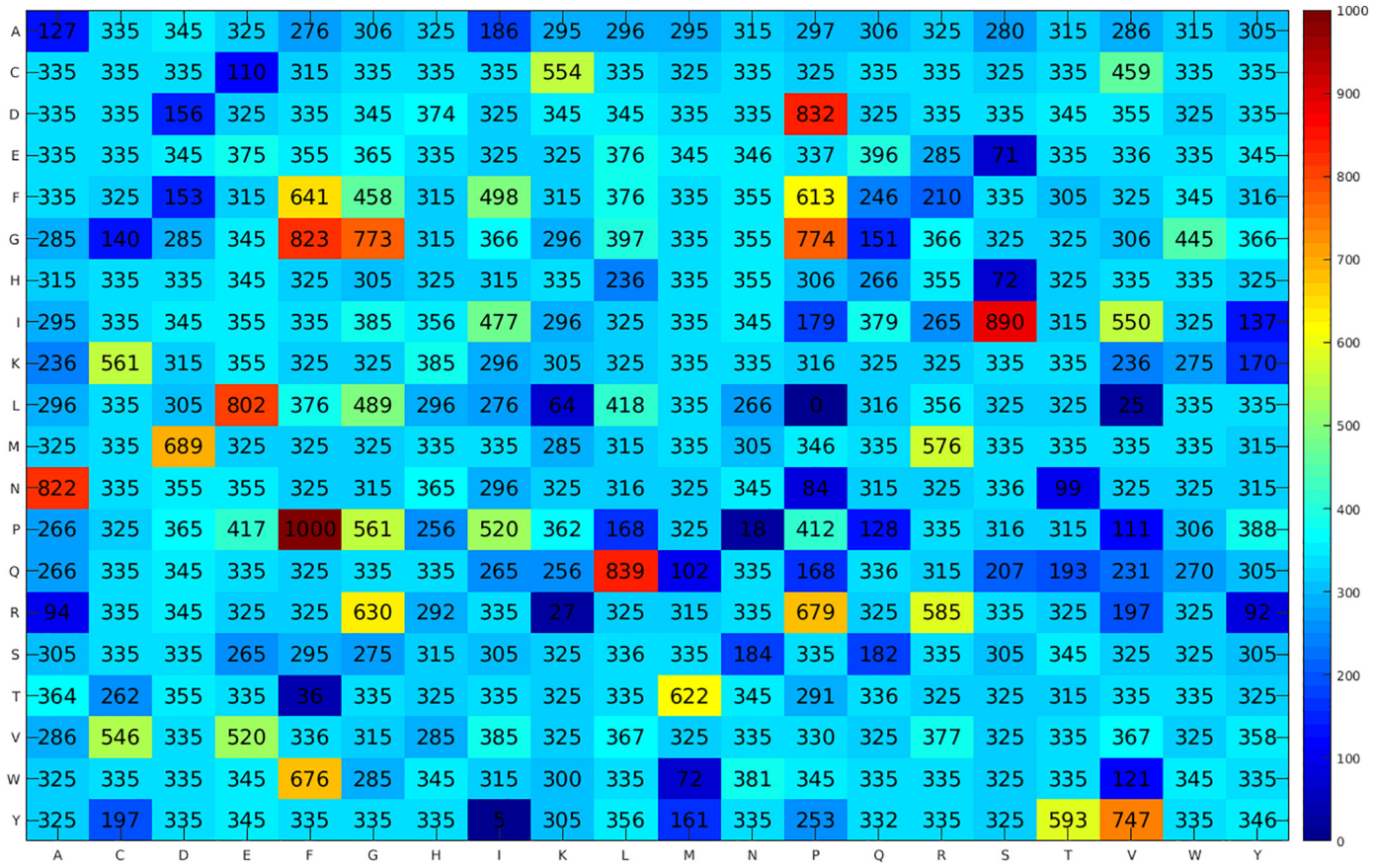

Fig. 2. Heatmap of dipeptide propensity scores obtained from the proposed iBitter-SCM.

Table 3
Comparison of optimized (opti-DPS) and initial (init-DPS) dipeptide propensity scores assessed by 10-fold cross-validation and independent test.

Method	10-fold CV		Independent test			
	Ac (%)	MCC	Ac (%)	Sn (%)	Sp (%)	MCC
Init-DPS	85.17	0.716	81.25	86.46	76.04	0.628
opti-DPS	87.11	0.751	84.38	84.38	84.38	0.688

Tables 4 lists the details of the comparisons between iBitter-SCM and the five ML classifiers over 10-fold CV and independent test. From Table 4 and Fig. 4A, we can find that the highest prediction results (Ac, MCC, auROC) of (87.11%, 0.751 and 0.903) is achieved by iBitter-SCM, while SVM and RF models perform well with the second and third highest prediction results of (77.54%, 0.560, 0.859) and (76.18, 0.537, 0.858), respectively. For the independent test, Table 4 and Fig. 4B show that iBitter-SCM and SVM models outperform the rest of the ML classifiers concerning most metrics, i.e. Ac, Sp and MCC. Although, the RF model yielded higher Sn and auROC than iBitter-SCM, their Ac and MCC values are lower than that of iBitter-SCM. Considering that the independent test is the most rigorous cross-validation method and a high value of MCC is preferable, it could be stated that the iBitter-SCM is more effective and stable than the RF model. Based on the details of comparison from Table 4 and Fig. 4, these results indicate that the iBitter-SCM is promising and holds potential in providing an accurate and reliable result for unseen peptides when compared to the widely used ML models developed in this study.

3.5. Identification of peptides having high bitterness intensities

One of the most important benefits of iBitter-SCM is to facilitate researchers by easily identifying their desired protein with a very simple weighted-sum function (S(P)) [38,39]. In this study, iBitter-SCM was used to calculate the bitter scores (BS) on the BTP-CV dataset for the discovery of peptides having high bitterness intensity. Table 5 lists the top twenty peptides having the highest bitterness intensities along with their sequences, bitter scores and log(1/T). As noticed in Table 5, all top ten peptides have bitter scores of greater than 331, where P is classified as bitter peptide when its bitter score is greater than 331; otherwise, P is classified as a non-bitter peptide. Nine out of the top-ten potential peptides having the highest bitter scores (bitter score, *log*(1/T)) are PF (1000, 2.8 [64]), PFP (806.50, 3.4 [9,16]), GF (823.00, 2.83, [9,16]), PFP (806.50, 3.4, [9,16]), GPFF (805.00, 3.8, [9,16]), LE (805.00, 3.8, [9,16]), GP (774.00, 1.79, [9,16]), GGP (773.50, 2.04, [9,16]) and RPFF (773.33, 4.4, [9,16]).

3.6. Analysis of bitter peptides using propensity scores of amino acids and dipeptides

The characterization and analysis of feature importance can provide a unique understanding of bitter peptides. In this study, the propensity score of amino acids and dipeptides were used to provide an understanding of the biochemical and biophysical properties of bitter peptides. Table 6 and Fig. 2 list the propensity scores of amino acids and dipeptides, respectively, derived from the BTP-CV dataset consisting of 256 bitter and non-bitter peptides. The propensity scores of dipeptides were obtained from the optimal opti-DPS from experiment #5, while the propensity scores of amino acids were calculated from such opti-DPS using a straightforward statistical approach [23,24,28–30,38–40].

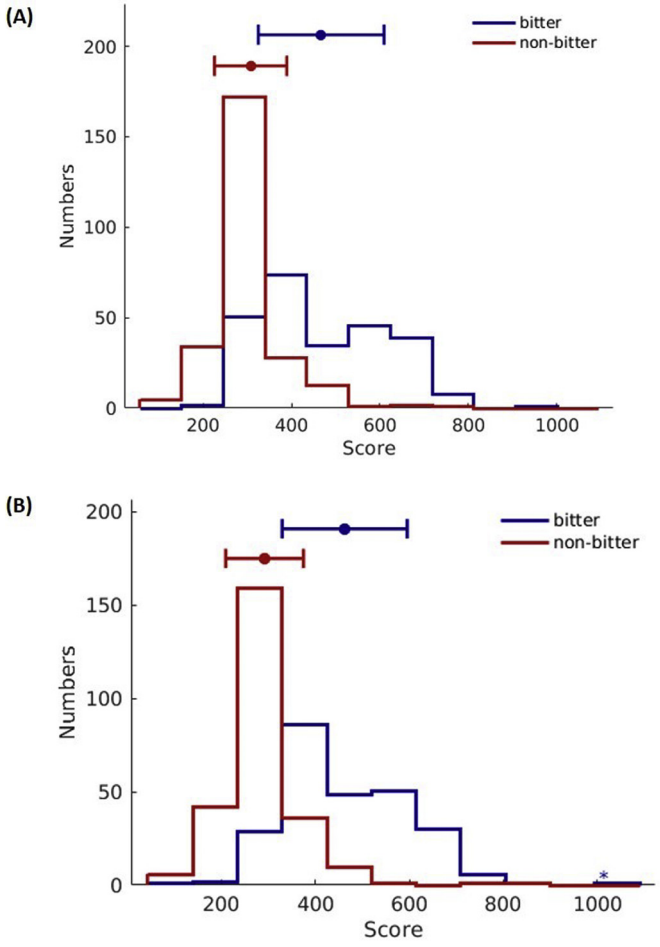

Fig. 3. The histogram of scores of bitter and non-bitter peptides derived from iBitter-SCM on the benchmark dataset by using initial (a) and optimized (b) dipeptide propensity scores, respectively.

Table 4
Performance comparison of iBitter-SCM with conventional classifiers as evaluated via 10-fold cross-validation and independent test.

Dataset	Classifier	Ac (%)	Sn (%)	Sp (%)	MCC	auROC
BTP-CV	SVM	77.54	83.26	71.89	0.560	0.859
	RF	76.18	86.35	66.02	0.537	0.858
	NB	74.03	83.22	64.75	0.493	0.789
	KNN	73.63	85.52	61.62	0.489	0.736
	DT	74.42	85.58	63.32	0.485	0.764
	iBitter-SCM	87.11	91.31	82.82	0.751	0.903
BTP-TS	SVM	84.38	82.81	85.94	0.688	0.862
	RF	83.59	90.63	76.56	0.679	0.916
	NB	76.56	89.06	64.06	0.549	0.855
	KNN	83.59	85.94	81.25	0.673	0.836
	DT	78.91	85.94	71.88	0.584	0.789
	iBitter-SCM	84.38	84.38	84.38	0.688	0.904

Amino acids and dipeptides with the highest propensity scores are the most important elements in bitter peptides.

As noticed in Table 6, the top-five amino acids having the highest propensity scores are Gly, Phe, Pro, Glu and Asp, while the top-five amino acids having the lowest propensity scores are Tyr, Ser, Lys, Ala and Gln. Based on the propensity scores of dipeptides as shown in Fig. 2 and Table S2, the ten top-ranked dipeptides represented in bitter peptides include PF, IS, QL, DP, GF, NA, LE, GP, GG and YV. Meanwhile, the t ten top-ranked dipeptides represented in non-bitter peptides include LP, YI, PN, LV, RK, TF, LK, ES, HS and WM. In general, the set of hydrophobic amino acid consists of Gly, Phe, Pro, Ile, Cys, Val, Leu, Met, Trp, Tyr and Ala (ranked at 1, 2, 3, 6, 8, 9, 10, 11, 12, 16, 19, respectively), and the other is hydrophilic set [23]. As seen in Table 6, seven out of the ten top-ranked informative amino acid containing Gly, Phe, Pro, Ile, Cys, Val and Leu are hydrophobic amino acids. Interestingly, the scores of Phe, and Pro are ranked at 2 and 3, respectively. It could be stated that Phe and Pro are important in bitter peptides. Results from the analysis of the top-five amino acids having the highest and lowest propensity scores are discussed below.

Many studies have reported that Phe and Pro (ranked at 2 and 3, respectively) play an important role in enhancing bitterness. For example, Ishibashi et al. [22] studied the role of Phe in bitter peptides by synthesizing oligopeptides containing Phe residue and evaluated their taste with threshold concentration (TH·V) and ratio of bitterness to caffeine (Rcaf), where peptides with lowest and largest values of TH·V and Rcaf, respectively, produced the strongest bitterness. This study reported the comparative study of taste peptides by fixing Phe at the C– and N-terminal region. As shown in Table II-IV of the previous study [22], the dipeptide (i.e. FG, FV, FL, FI and FP) and tripeptide (FIV, FPP, FPL and FPP) containing Phe at the N-terminus produced the bitter taste. Meanwhile, the dipeptide and tripeptide containing Phe at the C-terminus exhibiting the bitter taste are (GF, AF, VF, LF, IF, FF, YF, PF, SF, TF, KF, RF, DF and EF) and (VIF, LPF, RPF, PPF and FPF). Among various dipeptides and tripeptides, FPF exhibited the smallest TH·V (mM) and largest Rcaf of 0.4 and 2.5, respectively. Remarkably, RPFF and RRPFF having FF at C-terminus produced bitterness at a higher rate of 25 and 50 times, respectively, than caffeine as shown in Table III of the previous study [22].

In the same group of the previous study [22], they also investigated the role of Pro residue in bitter peptides [26]. They synthesized oligopeptides containing Pro and Gly residues and evaluated their taste with TH·V and Rcaf. Their experimental results showed that Pro residue exhibited both sweet and bitter with 10 mM in TH·V, while the di- and triproline showed a perceptible bitterness with 4.5 and 2.0 mM in TH·V, respectively. Meanwhile, other peptides containing Pro and Gly (GP, PPG, PGG, PGP, GPG and GGP) produced great bitterness. In this study, they further determined the taste of peptides containing Pro with some hydrophobic (Phe and Tyr), basic (Arg) and small (Gly) amino acids. Several observations from this study can be made as follows: (i) in the case of peptides containing Pro with some hydrophobic amino acids, a strong bitter taste of up to 0.4 and 0.3 mM in TH·V was found in PFP/FPF and YPF, respectively; and in the case of peptides containing Pro with some basic amino acids, a strong bitter taste of up to 0.8 mM in TH·V was found in RP, RPG and GRP. As seen in Table 6, Ala (ranked at 19) has the lowest propensity score of 303.18, while its percentages of bitter peptide, non-bitter peptide and their differences are 2.901, 8.288 and −5.387, respectively. These results suggested that Ala is not an important element for bitterness.

3.7. Analysis of bitter peptides using informative physicochemical properties

Previously, many studies have reported that the degrees of bitterness of peptides are affected by the biochemical and biophysical properties of peptides, such as hydrophobicity [3,22,25,64], side chain [20,26,65,66] and molecular structure [67]. To analyze the important biochemical and biophysical properties of bitter peptides, Pearson correlation coefficient (R) between the propensity scores of amino acids and the PCPs in AAindex were used to identify informative PCPs that were useful in bitter peptides. As seen in Table 7, the three selected PCPs are PONP800104 (R = 0.495), MEIH800103 (R = 0.403) and COWR900101 (R = 0.396), respectively, according to the important biochemical and biophysical properties for the bitterness of peptides as mentioned above. Additional details of the top twenty informative PCPs having the highest R values are listed in Table S3. The analysis results cooperating with the propensity scores of amino acids and the three selected PCPs are discussed below.

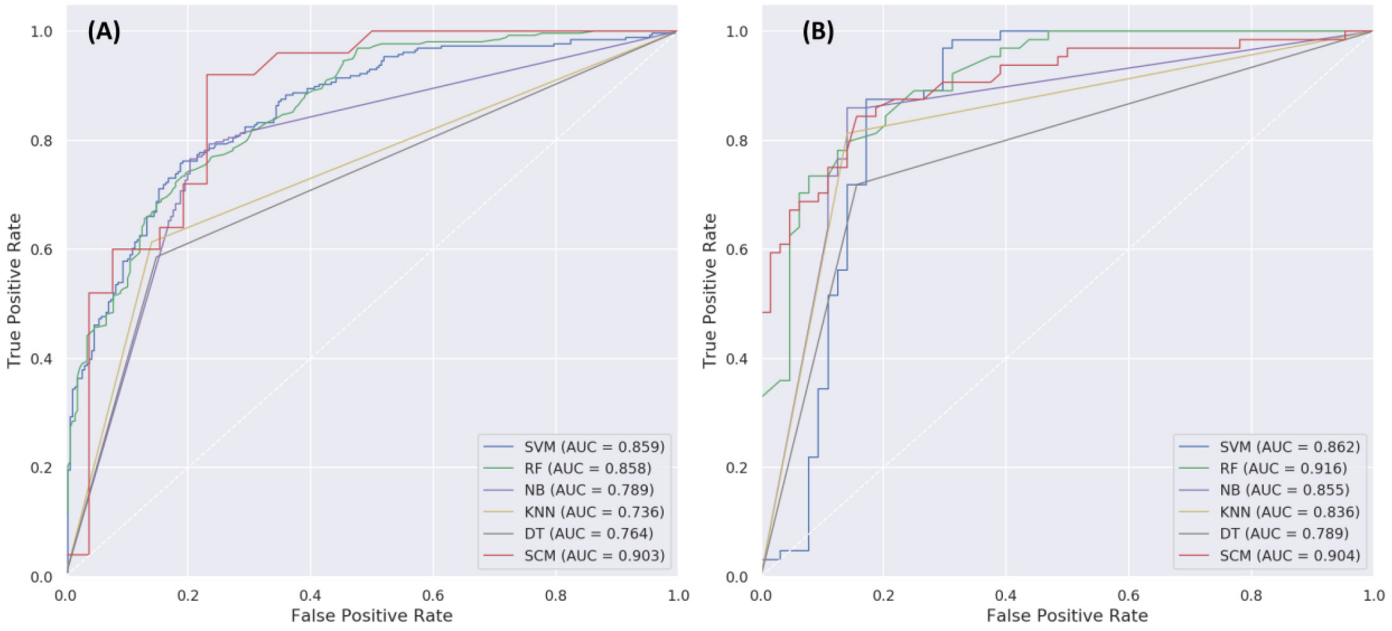

Fig. 4. ROC curves and AUC values of the proposed iBitter-SCM and other conventional methods over 10-fold cross-validation (A) and independent test (B).

Table 5
Top twenty peptides having high bitter tasting potency according to their bitter scores (BS).

Peptides	BS	*log*(1/T)	Reference
PF	1000.00	2.8	[64]
RPF	839.50	2.83	[17]
GF	823.00	2.36	[9,16]
PFP	806.50	3.4	[9,16]
GPFF	805.00	3.8	[9,16]
LE	802.00	2.52	[9,16]
GP	774.00	1.79	[9,16]
GGP	773.50	2.04	[9,16]
RPFF	773.33	4.4	[9,16]
GG	773.00	tasteless	[67]
GGFF	745.67	2.85	[9,16]
GFF	732.00	3.23	[9,16]
GPPF	728.67	2.52	[9,16]
RPFG	712.33	3.41	[9,16]
RGP	702.00	1.9	[9,16]
LGGGG	702.00	1.90	[21]
RGFF	698.00	3.8	[9,16]
RPGGFF	695.40	4.04	[9,16]
GGFFGG	693.6	3.7	[9,16]
RPFFRPFF	692.8571	5	[9,16]

Table 6
The propensity scores of twenty amino acids to be bitter peptides (Score) along with amino acid compositions (%) of bitter (BTP) and non-bitter (non-BTP) peptides.

Amino acid	BTP (%)	Non-BTP (%)	*P*-value	Difference	Score
G-Gly	15.986	6.736	0.000	9.250(1)	389.25(1)
F-Phe	13.157	5.269	0.000	7.888(2)	380.00(2)
P-Pro	16.390	17.048	0.684	−0.658(11)	352.90(3)
E-Glu	5.135	1.615	0.000	3.520(3)	345.53(4)
D-Asp	2.278	1.234	0.122	1.044(6)	344.75(5)
I-Ile	7.286	6.499	0.504	0.787(7)	342.98(6)
R-Arg	5.506	3.985	0.148	1.521(5)	338.65(7)
C-Cys	0.000	0.488	0.055	−0.488(9)	336.98(8)
V-Val	6.999	4.305	0.016	2.694(4)	335.23(9)
L-Leu	9.451	9.972	0.736	−0.521(10)	334.90(10)
M-Met	0.203	2.061	0.000	−1.859(16)	334.33(11)
W-Trp	1.897	2.221	0.688	−0.324(8)	328.03(12)
T-Thr	0.598	1.879	0.013	−1.280(12)	325.28(13)
N-Asn	1.715	3.078	0.042	−1.363(13)	321.58(14)
H-His	0.679	3.477	0.000	−2.799(17)	318.23(15)
Y-Tyr	5.074	6.677	0.201	−1.603(14)	317.20(16)
S-Ser	0.856	2.696	0.001	−1.841(15)	312.35(17)
K-Lys	2.251	6.425	0.000	−4.173(18)	309.50(18)
A-Ala	2.901	8.288	0.000	−5.387(20)	303.18(19)
Q-Gln	1.640	6.047	0.000	−4.408(19)	302.30(20)

3.7.1. Importance of hydrophobic amino acid residue for the manifestation of bitterness

The PONP800104 described as “Surrounding hydrophobicity in alpha-helix [68]” is ranked at 3 with a Pearson's correlation $R = 0.495$. Ponnuswamy et al. [68] estimated a new scale of hydrophobic indices of amino acid residues for characteristic hydrophobic environments in globular proteins using the crystal data on 21 proteins. The new scale of amino acid residues provides valuable information with regards to hydrophobic domains, nucleation sites, surface domains, loop sites and the spatial positions of residues in protein molecules. Furthermore, Table S3 shows that there are three other important PCPs relating to hydrophobic properties of amino acids, i.e. WILM950101, EISD860103 and COWR900101 described as hydrophobicity coefficient, direct of hydrophobic moment and hydrophobicity index, respectively.

It is well-known that hydrophobic amino acid residues, such as Phe [66], Pro [26], Ile [69], Arg [26], Val [65], Leu [20] and Tyr [66], play an important role in the manifestation of bitterness of peptides [70]. Tamura et al. mentioned the mechanism of action for the production of bitterness in peptides with the hydrophobic amino acid residue [71]. They have reported that bitter peptides possess two main active sites, i.e. binding (BU) and stimulating (SU) units, indicating that the intensity of the bitterness is determined at a hydrophobicity recognition zone located on the wall of a pocket of the receptor when both BU and SU coexisted at the bottom of the receptor [70]. Matoba and Hata synthesized various dipeptides to elucidate the relationship between bitter peptides and their chemical structures [3]. In fact, Gly was reported to produce a sweet taste, while the peptides of diGly, triGly and tetraGly was also tasteless. However, the change in taste of Gly could be found by connecting itself with others hydrophobic amino acids (Pro, Val, Ile, Leu, Phe and Tyr). They demonstrated this evidence by forming dipeptides and tripeptides with Gly. In their experimental results, the GP, PG, GV, VG, GL, LG, GI, IG, FG, GF, GY and YG dipeptides gave a bitter taste, while the tripeptides of GIG, GLG, GFG, GPG, GYG and GVG

Table 7
The three important physicochemical properties (PCPs) derived from iBitter-SCM.

Amino acid	Score	PONP800104	MEIH800103	COWR900101
G-Gly	389.25(1)	15.36(1)	90(8)	0(11)
F-Phe	380.00(2)	14.08(4)	108(1)	1.74(3)
P-Pro	352.90(3)	11.51(16)	78(15)	0.86(7)
E-Glu	345.53(4)	12.55(11)	72(16)	−0.37(13)
D-Asp	344.75(5)	10.98(20)	71(17)	−0.51(14)
I-Ile	342.98(6)	14.63(2)	105(2)	1.81(1)
R-Arg	338.65(7)	11.28(18)	81(14)	−1.56(18)
C-Cys	336.98(8)	14.49(3)	104(3)	0.84(8)
V-Val	335.23(9)	12.88(9)	94(6)	1.34(5)
L-Leu	334.90(10)	14.01(5)	104(4)	1.8(2)
M-Met	334.33(11)	13.4(7)	100(5)	1.18(6)
W-Trp	328.03(12)	12.06(13)	94(7)	1.46(4)
T-Thr	325.28(13)	13(8)	83(11)	−0.26(12)
N-Asn	321.58(14)	12.24(12)	70(18)	−1.03(17)
H-His	318.23(15)	11.59(15)	90(9)	−2.28(20)
Y-Tyr	317.20(16)	12.64(10)	83(12)	0.51(9)
S-Ser	312.35(17)	11.26(19)	83(13)	−0.64(15)
K-Lys	309.50(18)	11.96(14)	65(20)	−2.03(19)
A-Ala	303.18(19)	13.65(6)	87(10)	0.42(10)
Q-Gln	302.30(20)	11.3(17)	66(19)	−0.96(16)

also exhibited a bitter taste. In general, Pro, Val, Ile, Leu, Phe and Tyr produced light bitterness or tasteless results. On the other hand, peptides coupled with Gly in the carboxyl group side chain produced a strong bitter taste, such as GP > PG or VG > GV or GF > FG or GGF > FGG. Tamura et al. [71] also confirmed that the bitterness could be increased when the hydrophobicity of peptides and compounds increased. Their analysis showed that the VV dipeptide gave an umami taste, while the peptides of triVal and tetraVal exhibited bitterness of 0.2 and 0.5 times stronger, respectively, than that of caffeine. Similarly, the peptides of Phe, diPhe and triPhe produced a bitterness of 0.05, 0.8 and 5 times stronger, respectively than that of caffeine. Furthermore, the peptides composed of a high number of hydrophobic amino acids, such as Gly and Phe, can produce a strong bitterness. These analysis results indicated that the bitterness receptor recognized the hydrophobicity of bitter peptides. This evidence implied that a peptide composed of a high number of hydrophobic amino acids tends to produce a high intensity of bitterness [3,71].

3.7.2. Importance of hydrophobic amino acids located at the C-terminus for determining bitterness intensity

As discussed in the previous section, there were four important physicochemical properties, i.e. PONP800104, WILM950101, EISD860103 and COWR900101, relating to hydrophobic properties of amino acids. It is well-known that hydrophobic amino acids play an important role in binding the bitter taste receptor [3,71]. In addition, many studies also mentioned the effect of the location in conjunction with the hydrophobic properties of amino acids for increasing the bitter properties in peptides.

Ishibashi et al. investigated the assumption that hydrophobic amino acids located at the C-terminus exhibit higher bitterness than those at the N-terminus [20,26,65,66]. They made their efforts to clarify the determinant site of bitterness for the bitter taste receptor. Among one of their comparative studies on the taste of various peptides, Phe residue was fixed at the C- and N-terminus and linked with various amino acids [65,66]. Their analysis results reported that the dipeptides of FG, FF and FV have more intense bitterness when Phe was situated at the C-terminus rather than at the N-terminus. The Rcaf values of FG and FV (C- vs N-terminus) were (0.83 vs 0.17) and (0.33 vs 0.17), respectively. Moreover, the oligopeptides RPFF and RRPFF having F at the C-terminus produced a bitterness which were higher than caffeine at 25 and 50 times, respectively. Furthermore, based on the analysis results derived from QSAR modeling, Xu and Chung [17] mentioned that hydrophobic amino acids located at the C-terminus was the key determinant for bitterness. Their informative variables derived from selecting frequencies and standardized regression coefficients consisted of the descriptor sets describing various characteristics of C– and N-terminal amino acids. The two top-ranked informative variables were N1-T-3 and N2HESH-2, indicating the importance of high hydrophobicity of C-terminal amino acids. Although, Ishibashi et al. [22] suggested that the dipeptide PF exhibited slight bitterness with the Rcaf value of 0.03, the tripeptides of LPF, RPF and PPF showed greater bitterness with the Rcaf values of 2.5, 0.67 and 0.43, respectively. In addition, Ishibashi et al. [67] also attempted to determine the role of Val coupled with various other amino acids. Their results showed that the dipeptides of VL, VI and VF exhibited higher bitterness than LV, IV and FV. These results indicated that the intensity of bitter property of peptides linking the C-terminal Val was lower than those linking the N-terminal Val.

3.7.3. The number of carbon atoms on the amino acid side chain affects the intensity of bitterness

In this section, we focused on the other properties affecting the bitterness of peptides, i.e. side chain skeleton and linear/branch side chain. Ishibashi et al. [20,26,65,66] investigated the effect of various structural side chains of amino acids by synthesizing some oligopeptides containing different alkyl groups of side chains. In general, Gly which is the smallest hydrophobic amino acid having the α-carbon linked with a hydrogen atom. Their analysis results showed that Gly, diGly, and triGly exhibited tasteless properties, while no special taste was observed from tetraGly, pentaGly and hexaGly. In case of Ala and aminobutylric acid (Abu) having methyl (-CH_3) and ethyl (-C_2H_5) side chains, respectively, oligopeptides composed of Gly & Ala, Gly & Abu and triAbu (Abu_3) also operates tasteless, while free Abu exhibited a sweet taste. It is well-known that Gly, Ala and aminobutylric acid (Abu) have a side chain skeleton of less than three carbons. Considering Val that has propyl side chains (-C_3H_7), Val exhibited a mix taste of bitter and sweet. But the bitterness was apparent in the case of Val linked at the C-terminus, such as GGV and GVV. This observation is well consistent with the previous section. Similarly, Pro having a side chain of three carbons exhibited a composite of bitter and sweet taste. Interestingly, the experimental results of Ishibashi et al. [26] showed that almost all Pro-containing peptides could operate bitterness, except in only PG. To confirm the role of the number of carbons in the side chain skeleton, Leu, Ile, Phe and Tyr containing larger than four carbons were investigated [20,65,66]. Almost all peptides containing Gly and Leu [20,65] or Ile [20,65] or Phe [65,66] or Tyr [65,66] produced bitterness. These analysis results confirmed that the number of carbons of the side chain significantly affected to the intensity of bitterness. Meanwhile, the linear and branched side chains were also investigated. Peptides composing synthetic normal Val (n-Val) having a linear propyl chain produced greater bitterness as compared to that of Val having isopropyl or branched chain [65]. Remarkably, most peptides containing normal Leu (n-Leu) having linear butyl chain showed higher intensity of bitter taste than that of Leu and Ile having branched butyl chain [65]. The different intensity of bitter taste of linear and branched side chains of propyl and butyl chains indicate that the linear side chain was more significant than the branched chain.

3.8. iBitter-SCM web server

In order to facilitate the scientific community, we developed a user-friendly web server iBitter-SCM, which is made freely available online at http://camt.pythonanywhere.com/iBitter-SCM. Here, step-by-step guidelines are provided on how to use the iBitter-SCM web server in order to obtain the desired results.

1) Open the web server at http://camt.pythonanywhere.com/iBitter-SCM where the user will see the top screen of iBitter-SCM on the

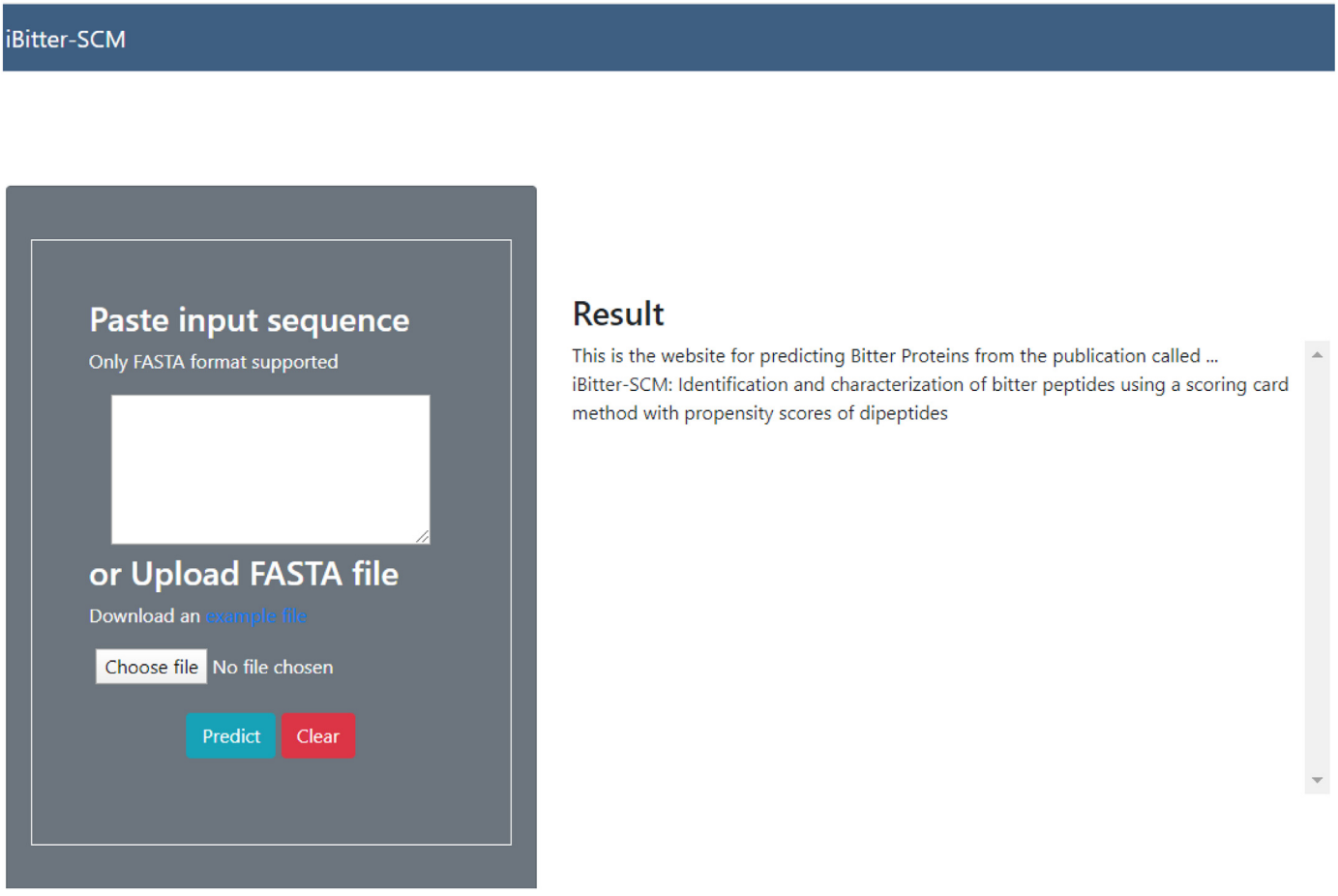

Fig. 5. Screenshots of the iBitter-SCM web server.

user's computer screen, as shown in Fig. 5.

2) Enter the query sequence into the text box or upload a FASTA file by clicking on the “Choose file” button.
3) Click on the “Submit” button in order to start the prediction process and the results are outputted as shown on the right-hand side of the web server.

4. Conclusions

Computational prediction and analysis of bitter peptides can help to accelerate the drug development and nutritional research process. Hence, it is desirable to develop accurate computational methods for rapidly and effectively identifying bitterness of peptides. For this purpose, we have developed a sequence-based predictor called the iBitter-SCM that can predict whether a given peptide sequence is bitter or not bitter. To build an efficient and interpretable predictive model, we applied the scoring card method (SCM) for estimating the propensity scores of amino acids and dipeptides. To the best of our knowledge, this represents the first computational model based on sequence information for predicting and analyzing bitterness of peptides. The iBitter-SCM could achieve an accuracy and Matthews coefficient correlation of 84.38% and 0.688, respectively, on the independent dataset. Comparative analysis revealed that iBitter-SCM was superior to those of other well-known classifiers (e.g. k-nearest neighbor, naive Bayes, decision tree and random forest models) owing to its simplicity, interpretability, and implementation. Investigation of the propensity scores of amino acids and informative physicochemical properties obtained from iBitter-SCM provided pertinent insights into the biochemical and biophysical properties of bitter peptides. Finally, the model was deployed as a web server also named iBitter-SCM that was established and made freely available online at http://camt.pythonanywhere.com/iBitter-SCM. It is anticipated that iBitter-SCM would serve as a useful and efficient tool for the prediction of bitter peptides and in doing so help to facilitate the drug development process.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgments

This work was co-supported by the TRF Research Grant for New Scholar (No. MRG6180226) and College of Arts, Media and Technology, Chiang Mai University, and partially supported by Chiang Mai University and the TRF Research Career Development Grant (No. RSA6280075).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2020.03.019.

References

[1] A. Drewnowski, C. Gomez-Carneros, Bitter taste, phytonutrients, and the consumer: a review, Am. J. Clin. Nutr. 72 (6) (2000) 1424–1435.
[2] E. Adler, M.A. Hoon, K.L. Mueller, J. Chandrashekar, N.J.P. Ryba, C.S. Zuker, A novel family of mammalian taste receptors, Cell 100 (6) (2000) 693–702.
[3] T. Matoba, T. Hata, Relationship between bitterness of peptides and their chemical structures, Agric. Biol. Chem. 36 (8) (1972) 1423–1431.
[4] L. Lemieux, R.E. Simard, Bitter flavour in dairy products. I. A review of the factors

ARTICLE IN PRESS

likely to influence its development, mainly in cheese manufacture, Lait 71 (6) (1991) 599–636, https://doi.org/10.1051/lait:1991647.
[5] C. Hansch, A. Leo, D. Hoekman, Exploring QSAR: Fundamentals and Applications in Chemistry and Biology, American Chemical Society, Washington, DC, 1995.
[6] N. Schaduangrat, S. Lampa, S. Simeon, M.P. Gleeson, O. Spjuth, C. Nantasenamat, Towards reproducible computational drug discovery, J. Cheminforma. 12 (1) (2020) 9.
[7] W. Shoombuatong, et al., Towards the revival of interpretable QSAR models, Advances in QSAR modeling, Springer, 2017, pp. 3–55.
[8] W. Shoombuatong, N. Schaduangrat, C. Nantasenamat, Towards understanding aromatase inhibitory activity via QSAR modeling, EXCLI J. 17 (2018) 688.
[9] H.-O. Kim, E.C. Li-Chan, Quantitative structure− activity relationship study of bitter peptides, J. Agric. Food Chem. 54 (26) (2006) 10102–10111.
[10] J. Wu, R.E. Aluko, Quantitative structure-activity relationship study of bitter di-and tri-peptides including relationship with angiotensin I-converting enzyme inhibitory activity, J. Peptide Sci. 13 (1) (2007) 63–69.
[11] J. Yin, Y. Diao, Z. Wen, Z. Wang, M. Li, Studying peptides biological activities based on multidimensional descriptors (E) using support vector regression, Int. J. Pept. Res. Ther. 16 (2) (2010) 111–121.
[12] J. Tong, S. Liu, P. Zhou, B. Wu, Z. Li, A novel descriptor of amino acids and its application in peptide QSAR, J. Theor. Biol. 253 (1) (2008) 90–97.
[13] Z.-h. Lin, H.-x. Long, Z. Bo, Y.-q. Wang, Y.-z. Wu, New descriptors of amino acids and their application to peptide QSAR study, Peptides 29 (10) (2008) 1798–1805.
[14] G. Liang, L. Yang, L. Kang, H. Mei, Z. Li, Using multidimensional patterns of amino acid attributes for QSAR analysis of peptides, Amino Acids 37 (4) (2009) 583–591.
[15] A.H. Pripp, Y. Ardö, Modelling relationship between angiotensin-(I)-converting enzyme inhibition and the bitter taste of peptides, Food Chem. 102 (3) (2007) 880–888.
[16] S. Soltani, H. Haghaei, A. Shayanfar, J. Vallipour, K. Asadpour Zeynali, A. Jouyban, QSBR study of bitter taste of peptides: application of GA-PLS in combination with MLR, SVM, and ANN approaches, Biomed. Res. Int. 2013 (2013).
[17] B. Xu, H.Y. Chung, Quantitative structure–activity relationship study of bitter Di-, Tri-and tetrapeptides using integrated descriptors, Molecules 24 (15) (2019) 2846.
[18] W. Huang, et al., BitterX: a tool for understanding bitter taste in humans, Sci. Rep. 6 (2016) 23450.
[19] A. Dagan-Wiener, I. Nissim, N.B. Abu, G. Borgonovo, A. Bassoli, M.Y. Niv, Bitter or not? BitterPredict, a tool for predicting taste from chemical structure, Sci. Rep. 7 (1) (2017) 1–13.
[20] N. Ishibashi, Y. Arita, H. Kanehisa, K. Kouge, H. Okai, S. Fukui, Bitterness of leucine-containing peptides, Agric. Biol. Chem. 51 (9) (1987) 2389–2394.
[21] K.H. Ney, Bitterness of peptides: amino acid composition and chain length, Food Taste Chem. 115 (1979) 149–173.
[22] N. Ishibashi, et al., Bitterness of phenylalanine-and tyrosine-containing peptides, Agric. Biol. Chem. 51 (12) (1987) 3309–3313.
[23] H.-L. Huang, Propensity scores for prediction and characterization of bioluminescent proteins from sequences, PLoS One 9 (5) (2014) e97158.
[24] T. Vasylenko, Y.-F. Liou, H.-A. Chen, P. Charoenkwan, H.-L. Huang, S.-Y. Ho, SCMPSP: prediction and characterization of photosynthetic proteins based on a scoring card method, BMC Bioinforma. 16 (1) (2015) S8 BioMed Central.
[25] Y. Nosho, K. Otagiri, I. Shinoda, H. Okai, Studies on a model of bitter peptides including arginine, Proline and phenylalanine residues. II. [1] Bitterness behavior of a Tetrapeptide (Arg-pro-Phe-Phe) and its derivatives, Agric. Biol. Chem. 49 (6) (1985) 1829–1837.
[26] N. Ishibashi, et al., Taste of proline-containing peptides, Agric. Biol. Chem. 52 (1) (1988) 95–98.
[27] P. Minkiewicz, J. Dziuba, A. Iwaniak, M. Dziuba, M. Darewicz, BIOPEP database and other programs for processing bioactive peptide sequences, J. AOAC Int. 91 (4) (2008) 965–980.
[28] P. Charoenkwan, S. Kanthawong, N. Schaduangrat, J. Yana, W. Shoombuatong, PVPred-SCM: improved prediction and analysis of phage virion proteins using a scoring card method, Cells 9 (2) (2020) 353.
[29] H.-L. Huang, et al., Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition, BMC Bioinforma. 13 (17) (2012) S3 BioMed Central.
[30] P. Charoenkwan, W. Shoombuatong, H.-C. Lee, J. Chaijaruwanich, H.-L. Huang, S.-Y. Ho, SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs, PLoS One 8 (9) (2013) e72368.
[31] B. Manavalan, S. Subramaniyam, T.H. Shin, M.O. Kim, G. Lee, Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy, J. Proteome Res. 17 (8) (2018) 2715–2726.
[32] B. Manavalan, T.H. Shin, G. Lee, PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine, Front. Microbiol. 9 (2018) 476.
[33] B. Manavalan, T.H. Shin, M.O. Kim, G. Lee, PIP-EL: a new ensemble learning method for improved proinflammatory peptide predictions, Front. Immunol. 9 (2018) 1783.
[34] B. Manavalan, T.H. Shin, M.O. Kim, G. Lee, AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest, Front. Pharmacol. 9 (2018) 276.
[35] M.M. Hasan, S. Yang, Y. Zhou, M.N.H. Mollah, SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties, Mol. BioSyst. 12 (3) (2016) 786–795.
[36] M. Hasan, M. Khatun, M. Mollah, N. Haque, C. Yong, G. Dianjing, NTyroSite: computational identification of protein nitrotyrosine sites using sequence evolutionary features, Molecules 23 (7) (2018) 1667.
[37] M. Hasan, M. Khatun, H. Kurata, Large-scale assessment of bioinformatics tools for lysine succinylation sites, Cells 8 (2) (2019) 95.
[38] T. Vasylenko, et al., SCMBYK: prediction and characterization of bacterial tyrosine-kinases based on propensity scores of dipeptides, BMC Bioinforma. 17 (19) (2016) 514.
[39] Y.-F. Liou, et al., SCMHBP: prediction and analysis of heme binding proteins using propensity scores of dipeptides, BMC Bioinforma. 15 (16) (2014) S4.
[40] Y.-F. Liou, et al., SCMMTP: identifying and characterizing membrane transport proteins using propensity scores of dipeptides, BMC Genomics 16 (12) (2015) S6.
[41] S. Zheng, et al., E-bitter: bitterant prediction by the consensus voting from the machine-learning methods, Front. Chem. 6 (2018) 82.
[42] A. Gautam, et al., In silico approaches for designing highly effective cell penetrating peptides, J. Transl. Med. 11 (1) (2013) 74.
[43] R. Kumar, et al., An in silico platform for predicting, screening and designing of antihypertensive peptides, Sci. Rep. 5 (2015) 12512.
[44] A. Tyagi, P. Kapoor, R. Kumar, K. Chaudhary, A. Gautam, G. Raghava, In silico models for designing and discovering novel anticancer peptides, Sci. Rep. 3 (2013) 2984.
[45] N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, W. Shoombuatong, ACPred: a computational tool for the prediction and analysis of anticancer peptides, Molecules 24 (10) (2019) 1973.
[46] R. Pratiwi, et al., CryoProtect: a web server for classifying antifreeze proteins from nonantifreeze proteins, J. Chem. 2017 (2017).
[47] T.S. Win, N. Schaduangrat, V. Prachayasittikul, C. Nantasenamat, W. Shoombuatong, PAAP: a web server for predicting antihypertensive activity of peptides, Future Med. Chem. 10 (15) (2018) 1749–1767.
[48] V. Laengsri, C. Nantasenamat, N. Schaduangrat, P. Nuchnoi, V. Prachayasittikul, W. Shoombuatong, TargetAntiAngio: a sequence-based tool for the prediction and analysis of anti-angiogenic peptides, Int. J. Mol. Sci. 20 (12) (2019) 2950.
[49] S. Kawashima, M. Kanehisa, AAindex: amino acid index database, Nucleic Acids Res. 28 (1) (2000) 374.
[50] W. Shoombuatong, et al., Towards predicting the cytochrome P450 modulation: from QSAR to proteochemometric modeling, Curr. Drug Metab. 18 (6) (2017) 540–555.
[51] W. Shoombuatong, N. Schaduangrat, C. Nantasenamat, Unraveling the bioactivity of anticancer peptides as deduced from machine learning, EXCLI J. 17 (2018) 734.
[52] F.-Y. Dao, et al., Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique, Bioinformatics 35 (12) (2019) 2075–2083.
[53] H.-Y. Lai, et al., iProEP: a computational predictor for predicting promoter, Mol. Ther. Nucleic Acids 17 (2019) 337–346.
[54] H. Lin, Z.-Y. Liang, H. Tang, W. Chen, Identifying sigma70 promoters with novel pseudo nucleotide composition, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017.
[55] H. Lv, Z.-M. Zhang, S.-H. Li, J.-X. Tan, W. Chen, H. Lin, Evaluation of different computational methods on 5-methylcytosine sites identification, Brief. Bioinform. (2019) bbz048.
[56] Z.-C. Xu, P.-M. Feng, H. Yang, W.-R. Qiu, W. Chen, H. Lin, iRNAD: a computational tool for identifying D modification sites in RNA sequence, Bioinformatics 35 (23) (2019) 4922–4929.
[57] S.F. Altschul, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 3389–3402.
[58] S. Ashkenazi, R. Snir, Y. Ofran, Assessing the relationship between conservation of function and conservation of sequence using photosynthetic proteins, Bioinformatics 28 (24) (2012) 3203–3210.
[59] P. Charoenkwan, N. Schaduangrat, C. Nantasenamat, T. Piacham, W. Shoombuatong, iQSP: a sequence-based tool for the prediction and analysis of quorum sensing peptides via Chou's 5-steps rule and informative physicochemical properties, Int. J. Mol. Sci. 21 (1) (2020) 75.
[60] M.M. Hasan, B. Manavalan, W. Shoombuatong, M.S. Khatun, H. Kurata, i6mA-fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation, Plant Mol. Biol. (2020) 1–10.
[61] M. Hasan, N. Schaduangrat, S. Basith, G. Lee, W. Shoombuatong, B. Manavalan, HLPpred-fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation, Bioinformatics (2020) btaa160.
[62] A. Abraham, et al., Machine learning for neuroimaging with scikit-learn, Front. Neuroinforma. 8 (2014) 14.
[63] M.M. Hasan, H. Kurata, GPSuc: global prediction of generic and species-specific succinylation Sites by aggregating multiple sequence features, PLoS One 13 (10) (2018).
[64] K. Otagiri, Y. Nosho, I. Shinoda, H. Fukui, H. Okai, Studies on a model of bitter peptides including arginine, proline and phenylalanine residues. I. Bitter taste of di- and tripeptides, and bitterness increase of the model peptides by extension of the peptide chain, Agric. Biol. Chem. 49 (4) (1985) 1019–1026.
[65] N. Ishibashi, et al., Role of the hydrophobic amino acid residue in the bitterness of peptides, Agric. Biol. Chem. 52 (1) (1988) 91–94 1988/01/01.
[66] N. Ishibashi, et al., Bitterness of phenylalanine- and tyrosine-containing peptides, Agric. Biol. Chem. 51 (12) (1987) 3309–3313 1987/12/01.
[67] N. Ishibashi, et al., Role of the hydrophobia amino acid residue in the bitterness of pep tides, Agric. Biol. Chem. 52 (1) (1988) 91–94.
[68] P. Ponnuswamy, M. Prabhakaran, P. Manavalan, Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins, Biochim. Biophys. Acta Protein Struct. Mol. Enzymol. 623 (2) (1980) 301–316.
[69] J. Kirimura, A. Shimizu, A. Kimizuka, T. Ninomiya, N. Katsuya, Contribution of peptides and amino acids to the taste of foods, J. Agric. Food Chem. 17 (4) (1969) 689–695 1969/07/01.
[70] K. Maehashi, L. Huang, Bitter peptides and bitter taste receptors, Cell. Mol. Life Sci. 66 (10) (2009) 1661–1671.
[71] M. Tamura, et al., Mechanism for the bitter tasting potency of peptides using o-aminoacyl sugars as model compounds+, Agric. Biol. Chem. 54 (6) (1990) 1401–1409.

International Journal of
Molecular Sciences

MDPI

Article

iQSP: A Sequence-Based Tool for the Prediction and Analysis of Quorum Sensing Peptides via Chou's 5-Steps Rule and Informative Physicochemical Properties

Phasit Charoenkwan [1], Nalini Schaduangrat [2], Chanin Nantasenamat [2], Theeraphon Piacham [3] and Watshara Shoombuatong [2,*]

1 College of Arts, Media and Technology, Chiang Mai University, Chiang Mai 50200, Thailand; phasit.c@cmu.ac.th
2 Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand; nalini.sch@mahidol.edu (N.S.); chanin.nan@mahidol.edu (C.N.)
3 Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand; theeraphon.pia@mahidol.ac.th
* Correspondence: watshara.sho@mahidol.ac.th; Tel.: +66-2-441-4371 (ext. 2715)

Received: 18 November 2019; Accepted: 18 December 2019; Published: 20 December 2019

Abstract: Understanding of quorum-sensing peptides (QSPs) in their functional mechanism plays an essential role in finding new opportunities to combat bacterial infections by designing drugs. With the avalanche of the newly available peptide sequences in the post-genomic age, it is highly desirable to develop a computational model for efficient, rapid and high-throughput QSP identification purely based on the peptide sequence information alone. Although, few methods have been developed for predicting QSPs, their prediction accuracy and interpretability still requires further improvements. Thus, in this work, we proposed an accurate sequence-based predictor (called iQSP) and a set of interpretable rules (called IR-QSP) for predicting and analyzing QSPs. In iQSP, we utilized a powerful support vector machine (SVM) cooperating with 18 informative features from physicochemical properties (PCPs). Rigorous independent validation test showed that iQSP achieved maximum accuracy and MCC of 93.00% and 0.86, respectively. Furthermore, a set of interpretable rules IR-QSP was extracted by using random forest model and the 18 informative PCPs. Finally, for the convenience of experimental scientists, the iQSP web server was established and made freely available online. It is anticipated that iQSP will become a useful tool or at least as a complementary existing method for predicting and analyzing QSPs.

Keywords: quorum sensing peptides; physicochemical properties; support vector machine; random forest; machine learning; classification

1. Introduction

It is widely known that bacteria can interconnect within its population using cell–cell communication tools. One such tool, quorum sensing (QS) is a molecular mechanism that depends on the population density to trigger cell–cell signaling which changes the behavior of the bacterial community when the population reaches a critical level [1,2]. Genes directing the beneficial activities performed synchronously by a bacterial population are controlled by the QS mechanism (i.e., bioluminescence, sporulation, competence, antibiotic production, biofilm formation, and virulence factor secretion) [3]. Furthermore, in order to orchestrate these collective behaviors, QSPs (quorum sensing peptides or autoinducing (AI) peptides) are secreted by the bacteria to respond to extracellular

signaling molecules. Typically, in a fresh culture, bacteria continually generate the signal starting at a low concentration, which accumulates as the population density increases. Upon reaching a threshold concentration, receptor protein interaction occurs which activates receptor kinase by phosphorylation thus, inducing a coordinated change in gene expression via transcription of target genes in the population [4]. QSPs have been widely studied in Gram-positive bacteria and are shown to be species specific (i.e., *Staphylococcus* spp., *Clostridium* spp., or *Enterococcus* spp) [5]. On the other hand, Gram-negative bacteria (i.e., *Pseudomonas* spp., *Acinetobacter* spp., or *Burkholderia* spp.) reportedly produce a different class of autoinducers known as, acyl-homoserine lactones (AHLs) which comprise of a lactone ring coupled with an aliphatic acyl chain with varying length and modifications [6]. In addition, a wide-ranging variety of other signaling molecules have also been identified [7], including fatty acids used by *Xanthomonas* spp., *Burkholderia* spp., *Xylella* spp. [8], ketones (*Vibrio* spp. and *Legionella* spp. [9]), epinephrine, norepinephrine and AI-3 (enterohemorrhagic bacteria; [10]) or quinolones (*Pseudomonas aeruginosa*; [11]). Additionally, AI-2, a furanosyl borate diester, is used by both Gram-negative and Gram-positive bacteria [12]. Finally, QSPs are well understood at the molecular level in many bacterial species, and have been extensively reviewed [13–17]. An example of experimentally elucidated QSP structures are shown in Figure 1.

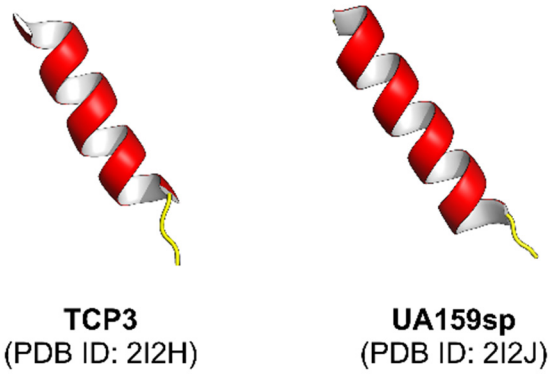

Figure 1. Structures of selected quorum sensing peptides that have been experimentally elucidated, where red and yellow colors represent alpha-helix and loop structures, respectively. Each structure is labelled by a common name followed by the Protein Data Bank identification number (PDB ID) in parenthesis on the subsequent line.

As previously mentioned, QSPs control the production of virulence factors in bacteria including many antibiotic resistant bacteria such as lectin, exotoxin A, pyocyanin, and elastase in *Pseudomonas aeruginosa* [15,18,19] and hemolysins, protein A, enterotoxins, lipases, and fibronectin proteins in *Staphylococcus aureus* [20,21]. These virulence factors allow the bacteria to evade the host immune responses. In addition, the overuse and abuse of antibiotics coupled with increased resistance, has prompted the discovery of various anti-QS agents as alternatives to traditional antibiotics [4]. These anti-QS agents can abolish the QS signaling pathways and prevent the formation or accumulation of virulence factors, therefore reducing bacterial virulence without causing drug-resistance. However, no such drug in currently approved by the FDA. At present, the combination of anti-QS agents with antibiotics provides the most effective strategy to combat bacterial infections [22,23]. As such, many studies have successfully demonstrated the synergistic effects of combining antibiotics with anti-QS agents [24–27]. Thus, it is promising to discover novel therapeutic agents to help boost research in this area.

Although, experimental approaches are known as an objective way to identify the biological activities of QSPs, it is time-consuming and costly. Meanwhile, with the avalanche of the peptide sequences in many free-access databases, the development of fast, efficient, and intelligent computational models for predicting and analyzing QSPs is urgently need to serve clinical application, drug development, and basic research. Currently, there are only two computational methods [28,29]

developed for discriminating QSPs from Non-QSPs using the information of peptide sequences as summarized in Table 1. The first QSP predictor was proposed by Akanksha et al. [29] named QSPpred. This method was developed by using support vector machine (SVM) cooperating with four types of peptide features, i.e., amino acid composition (AAC), dipeptide composition (DPC), binary pattern of N- and C-terminal residues and physicochemical properties (PCP). In 2018, Leyi et al. proposed a second QSP predictor named QSPpred-FL. This predictor was a two-layer prediction framework, where the first layer was used to establish initial features derived from 10 different types of peptide features and generate a new feature vector having 99-dimensional feature vector, while the final QSP predictor (QSPpred-FL) was construct in the second layer by using random forest (RF) with the top four features consisting of g-gap dipeptide composition (GDP), overlapping property features (OVP), composition–transition–distribution (CTD) and adaptive skip dipeptide composition (ASDC). Although, the two above-mentioned methods could provide quite promising prediction results, there is still room for improvement in the performance and interpretability of the predictors. First, QSPpred and QSPpred-FL were constructed with 630D and 913D feature vectors, respectively. The prediction model constructed with high-dimensional feature spaces could be prone to overfitting and might yield overestimated prediction results [30–33]. Second, QSPpred-FL was not evaluated on an independent set therefore, its predictive ability on unknown peptides could not be determined. Third, QSPpred and QSPpred-FL was performed on the benchmark and independent datasets with a single random sampling procedure. Hence, these two methods might tend to find the possible bias of the random sampling process and provided a good predictive result by chance. Finally, the mechanism of QSPpred and QSPpred-FL suffers from their low interpretable ability for experimental and related researchers as it is not easy to identify and assess which features provide crucial contribution to the high prediction results.

Table 1. Summary of existing methods for predicting quorum sensing peptides.

Method	Classifier [a]	Sequence Feature [b]	No. of Features	Independent Test
QSPpred	SVM	AAC, DPC, N5C5Bin, PCP	630	Yes
QSPpred-FL	RF	GDP, OVP, CTD, ASDC	913	No
iQSP (this study)	SVM	PCP	18	Yes

[a] RF: random forest, SVM: support vector machine. [b] AAC: amino acid composition, ASDC: adaptive skip dipeptide composition, DPC: dipeptide composition, CTD: composition–transition–distribution, GDP: g-gap dipeptide composition, NCBin: binary pattern of N- and C-terminal residues, OVP: overlapping property features, PCP: physicochemical properties.

To improve the prediction and interpretability performances, we proposed a systematic effort via Chou's 5-steps rule [34,35] for predicting and analyzing QSPs called the iQSP. The development of iQSP consists of five main parts: (i) Collecting the benchmarking dataset, (ii) representing a peptide sequence with PCPs from the AAindex database, (iii) selecting m informative PCPs from 531 PCPs, (iv) developing the QSP predictor using SVM model with the m informative PCPs, and (v) extracting interpretable rules by using the RF method with the m informative PCPs called IR-QSP. In this study, the m informative PCPs were identified using the SVM model in cooperating with the genetic algorithm utilizing self-assessment-report (GA-SAR). The performance comparisons showed that iQSP achieved an accuracy, MCC and auROC of 93.00%, 0.91% and 0.96%, respectively, as assessed by the rigorous independent validation test, which showed a significant improvement as compared with QSPpred. Amongst a set of interpretable rules, there were seven out of eight interpretable rules that could yield a prediction accuracy of greater than 80%. Finally, iQSP was developed as a user-friendly and publicly accessible web server that allow robust predictions to be made without the need to develop in-house prediction models.

2. Results and Discussion

In this study, we developed a sequence-based model for predicting and analyzing QSPs named iQSP. Firstly, the sequence logo representation and Gini index of amino acids were used for characterizing the informative properties between QSPs and Non-QSPs. Secondly, the GA-SAR algorithm was used to generate ten feature subsets, where each subset consisted of m informative features of PCPs. Thirdly, SVM models were individually constructed using the mentioned feature subsets. After that, the if–then interpretable rules (called IR-QSP) were generated by using the RF method and the optimal feature subset. Finally, iQSP is deployed as a free prediction web server so as to afford easy and rapid classification of query protein sequence as being QSPs and Non-QSPs. The overall framework of the proposed model, iQSP is shown in Figure 2.

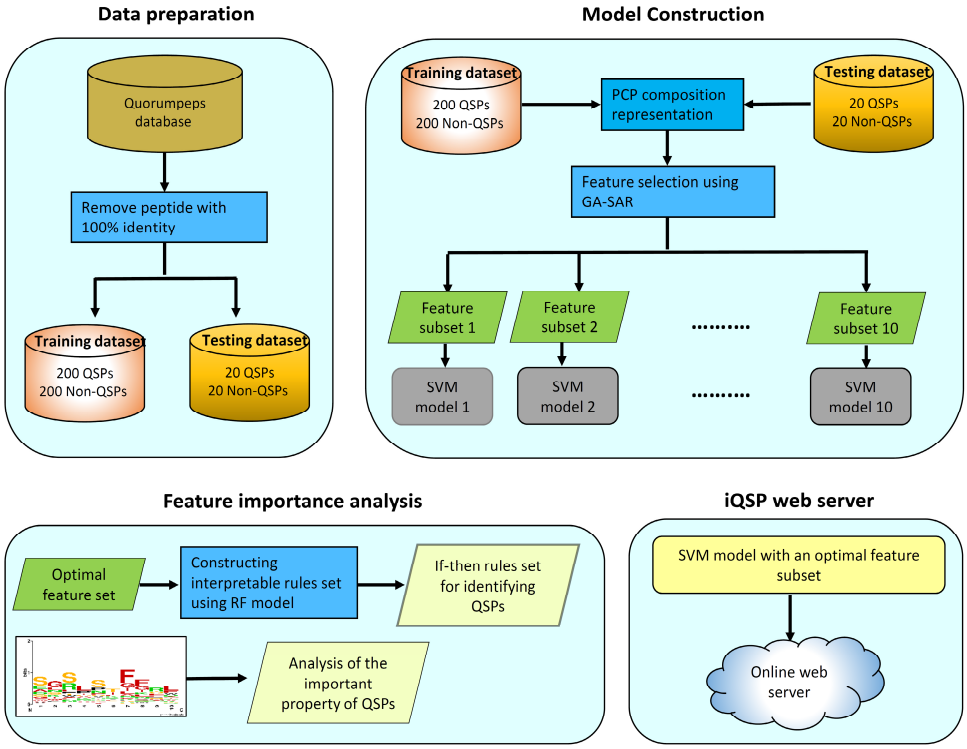

Figure 2. Schematic framework of iQSP. Arrows in the figure represents the direction that data flows from one process to the next process.

2.1. Composition Analysis

The analysis of feature importance can provide valuable information for predicting its function and activity. Previously, AAC has been used for analyzing the inherent characteristics and patterns of many therapeutic peptides [36–41] and protein functions [42–44]. In this study, the mean decrease of Gini index (MDGI) was utilized to rank the importance of each AAC feature. Features with the largest MDGI value are considered to be important as they significantly contribute to the prediction performance. In order to increase the reliability and validity for determining the feature importance, 100 RF models were constructed by varying the mtry parameter settings from 1 to 100 (mtry = 1, 2, 3, . . . , 100) and fixing the ntree parameter with 500 [36,39,42,45]. Finally, the average value of MDGI on 100 runs of feature importance estimations were used in this study. Table 2 lists the percentage values of the twenty amino acids for both QSPs and Non-QSPs along with amino acid compositional differences between the two classes as well as their MDGI values. Furthermore, the sequence logo [46] of the first and last ten residues at the N- and C-terminal regions of both QSPs and Non-QSPs were created to visualize the positional information for each amino acid as shown in Figure 3.

Table 2. Amino acid compositions (%) of quorum sensing (QSP) and non-quorum sensing (Non-QSP) peptides along with their mean decrease of Gini index (MDGI) values.

Amino Acid	QSP (%)	Non-QSP (%)	Difference	*p*-Value	MDGI
F	0.109	0.049	0.059 (1)	0.000	41.94 (1)
K	0.047	0.093	−0.045 (20)	0.000	19.37 (2)
L	0.076	0.106	−0.030 (18)	0.002	17.57 (3)
V	0.043	0.054	−0.011 (14)	0.114	15.56 (4)
W	0.034	0.015	0.019 (4)	0.000	10.99 (5)
I	0.063	0.067	−0.004 (11)	0.629	10.15 (6)
A	0.053	0.084	−0.032 (19)	0.000	9.15 (7)
Y	0.042	0.020	0.022 (3)	0.000	8.69 (8)
R	0.039	0.050	−0.011 (15)	0.128	7.99 (9)
C	0.049	0.061	−0.013 (16)	0.091	7.82 (10)
S	0.079	0.062	0.017 (5)	0.023	6.70 (11)
G	0.078	0.094	−0.016 (17)	0.032	6.56 (12)
P	0.041	0.043	−0.001 (10)	0.846	6.25 (13)
N	0.070	0.043	0.026 (2)	0.009	5.24 (14)
T	0.051	0.041	0.010 (7)	0.098	4.38 (15)
D	0.033	0.029	0.004 (8)	0.411	4.37 (16)
E	0.026	0.030	−0.004 (12)	0.423	3.64 (17)
M	0.028	0.016	0.012 (6)	0.013	2.84 (18)
H	0.010	0.017	−0.007 (13)	0.048	2.71 (19)
Q	0.029	0.026	0.003 (9)	0.484	2.49 (20)

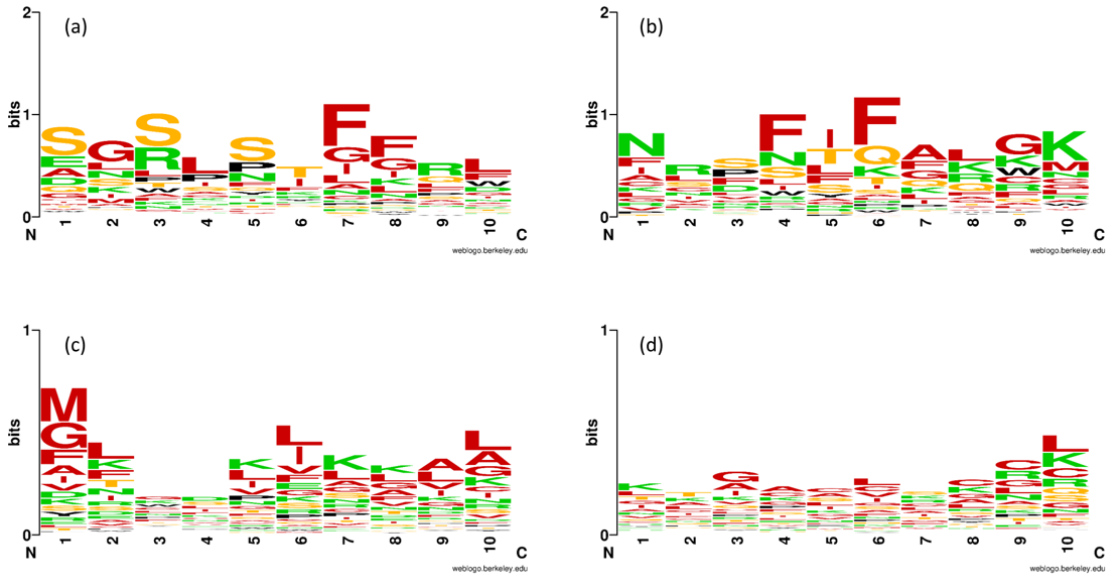

Figure 3. Sequence logo representations, where x- and y-axis represent the first and last 10 residues at N- and C-terminal regions from QSPs (**a**,**b**) and Non-QSPs (**c**,**d**), and proportional to the propensities of amino acids, respectively. Colors are: red for hydrophobic (A, I, L, M, F, V, C, G), green for charged (R, D, E, K), orange for polar (Q, H, S, T), and black for the remaining amino acids (P, Y, W, N).

Table 2 shows that the ten top-ranked important amino acids according to MDGI values are Phe, Lys, Leu, Val, Trp, Ile, Ala, Tyr, Arg, and Cys. There are 6 out of 10 top-ranked important amino acids having an MDGI value larger than 10, i.e., Phe, Lys, Leu, Val, Trp, and Ile. As seen, amongst the ten informative amino acids, the analysis of AAC with the percentage of certain residues for QSPs suggests that Phe, Trp, and Tyr are dominant in QSPs, while Lys, Leu, Ala, Arg, and Cys are dominant in Non-QSP peptides at a significance level of p-value ≤ 0.05. Furthermore, the four sequence logo representations were created to visualize the positional information for each amino acid as shown in Figure 3. The overall stack height of each position indicates its sequence conservation while the size of

the residue represents its propensity. Figure 3a,b reveal that Ser, Gly, Phe, and Arg as well as Phe and Lys are abundant in the first ten residues from the N- and C-terminal regions, respectively, of QSPs. Meanwhile, Figure 3c,d show that Met, Leu, Lys, Ala, and Gly as well as Leu and Cys are abundant in the first ten residues from the N- and C-terminal regions, respectively, of Non-QSPs. Therfore, the information gathered from the analysis of feature importance and the sequence logo illustrations in Table 2 and Figure 3, respectively, can be briefly summarized as follows: (i) Phe, Lys, Leu, Ala, and Arg are crucial amino acid residues that could potentially be used for discriminating QSPs from Non-QSPs; (ii) Phe and Ser as well as Lys and Leu are seen to be favored by QSPs and Non-QSPs, respectively, which are found in the first and last ten residues at the N- and C-terminal regions; and (iii) These observations were in good consistency with the feature importance as estimated using MDGI values where Phe, Lys, Leu and Ser are ranked 1, 2, 3, and 11, respectively. Moreover, our analysis were quite compatible with previous works [29,47].

2.2. Prediction Capabilities of the Different Subset of Physicochemical Properties

To make a fair comparison with the existing methods, a series of comparative experiments were carried out, while the same benchmark (200 QSPs and 200 Non-QSPs) and independent (20 QSPs and 20 Non-QSPs) datasets were used to develop and investigate the efficiency and effectiveness of our proposed QSP predictor. However, previous studies [28,29,48] performed these two datasets using a single random sampling procedure. As elaborated in [39,40,42,43,45,49,50], this procedure might find a possible bias of the random sampling process and provided a good predictive result by chance. Therefore, we repeated this construction procedure with ten independent rounds to alleviate the aforementioned problems [36,39,41–43,45]. In this study, the proposed GA-SAR algorithm was used to identify m informative features from 531 PCPs, where the number of m is in the range of 5–20. We hypothesize that if a feature is selected by GA-SAR, it is considered to be beneficial for QSP prediction [51–53]. Due to the non-deterministic characteristics of the GA-SAR algorithm, ten individual experiments were performed to generate ten different feature subsets. The lists of m informative features in the ten different feature subset are demonstrated in Table A1. In our experimental setting, each feature subset was used as the input feature to construct 10 SVM models based on random sampling with ten independent rounds. Therefore, the final prediction results of 10-fold CV and independent validation tests of each feature set were obtained by averaging the five statistical parameters (Ac, Sn, Sp, MCC, and auROC) on the benchmark and independent datasets, as shown in Tables 3 and 4, respectively.

Table 3. Performance comparisons of SVM models built with various subsets of physicochemical properties evaluated by means of ten-fold cross-validation subjected to ten rounds of random splits.

Subset	# Feature	Ac (%)	Sn (%)	Sp (%)	MCC	AUC
1	14	91.23 ± 1.31	91.17 ± 2.65	88.23 ± 3.03	0.82 ± 0.03	0.95 ± 0.05
2	17	90.69 ± 1.26	92.08 ± 2.82	87.04 ± 2.89	0.82 ± 0.03	0.95 ± 0.05
3	16	91.58 ± 1.77	92.79 ± 2.73	88.45 ± 2.97	0.83 ± 0.04	0.94 ± 0.05
4	17	91.63 ± 1.75	91.49 ± 2.73	89.55 ± 3.71	0.84 ± 0.04	0.92 ± 0.05
5	17	92.01 ± 1.40	91.43 ± 3.64	90.27 ± 2.39	0.84 ± 0.03	0.92 ± 0.08
6	18	91.07 ± 1.77	90.06 ± 2.69	88.79 ± 3.40	0.82 ± 0.04	0.91 ± 0.10
7	15	89.28 ± 1.99	88.78 ± 3.88	86.44 ± 2.67	0.79 ± 0.04	0.93 ± 0.07
8	17	88.24 ± 2.15	85.04 ± 3.61	86.81 ± 3.86	0.76 ± 0.05	0.92 ± 0.07
9	18	90.54 ± 1.26	90.60 ± 4.12	87.57 ± 2.68	0.81 ± 0.03	0.93 ± 0.09
10	17	92.19 ± 1.09	90.12 ± 2.98	92.16 ± 1.96	0.84 ± 0.02	0.93 ± 0.07

Feature represents the number of features used for constructing a model.

Table 4. Performance comparisons of SVM models built with various subsets of physicochemical properties evaluated by means of independent validation test subjected to ten rounds of random splits.

Subset	# Feature	Ac (%)	Sn (%)	Sp (%)	MCC	AUC
1	14	92.50 ± 1.67	95.50 ± 3.69	89.50 ± 4.97	0.85 ± 0.03	0.95 ± 0.03
2	17	91.50 ± 3.58	98.00 ± 3.50	85.00 ± 7.07	0.84 ± 0.07	0.96 ± 0.03
3	16	92.00 ± 2.58	92.50 ± 5.40	91.50 ± 5.80	0.84 ± 0.05	0.95 ± 0.02
4	17	92.25 ± 2.19	93.00 ± 4.83	91.50 ± 4.12	0.85 ± 0.04	0.97 ± 0.01
5	17	92.50 ± 2.36	94.00 ± 5.68	91.00 ± 6.15	0.86 ± 0.05	0.96 ± 0.02
6	18	93.00 ± 1.97	92.50 ± 5.40	93.50 ± 4.12	0.86 ± 0.04	0.96 ± 0.02
7	15	92.00 ± 1.97	94.00 ± 5.16	90.00 ± 7.45	0.85 ± 0.04	0.96 ± 0.02
8	17	91.75 ± 2.90	94.00 ± 5.16	89.50 ± 6.43	0.84 ± 0.06	0.95 ± 0.04
9	18	91.50 ± 3.38	90.50 ± 7.62	92.50 ± 7.91	0.84 ± 0.06	0.97 ± 0.04
10	17	92.50 ± 1.67	95.00 ± 3.33	90.00 ± 4.08	0.85 ± 0.03	0.95 ± 0.04

Feature represents the number of features used for constructing a model.

Amongst the ten different feature subsets, Table 3 shows that the subsets with the five highest Ac over 10-fold CV are subsets 10, 5, 4, 3, and 1 (92.19 ± 1.09%, 92.01 ± 1.40%, 91.63 ± 1.75%, 91.58 ± 1.77%, and 91.23 ± 1.31%, respectively). Meanwhile, the subsets with the five highest Ac over the independent validation test are subset 6, 10, 5, 1, and 4 (93.00 ± 1.97%, 92.50 ± 1.67%, 92.50 ± 2.36%, 92.50 ± 1.67%, and 92.25 ± 2.19%, respectively), as summarized in Table 4. As noticed in Tables 3 and 4, subsets 10, 5, 4, 1, and 6 showed good predictive powers with their ranks (10-fold CV, independent validation test) at (1, 2), (2,3), (3, 5), (5, 4), and (6,1) respectively. Although, subset 6 having a 91.07 ± 1.77% Ac was ranked at 5, its prediction result over independent validation test with 93.00 ± 1.97% Ac outperformed that of other subsets. Due to the fact that the independent test is an effective way to demonstrate the robustness and reliability of the model in real-world applications [36,39–43,45,49,50], it could be stated that subset 6 having eighteen informative PCPs provided a crucial contribution to the prediction performance. This feature subset yielded a prediction performance (Ac/MCC/auROC) over the 10-fold CV and independent validation test of 91.07 ± 1.77%/0.82 ± 0.04/0.91 ± 0.10 and 93.00 ± 1.97%/0.86 ± 0.04/0.96 ± 0.02, respectively.

Furthermore, the performance comparisons between SVM models with 531 PCPs as well as eighteen informative PCPs were conducted to investigate the effectiveness of our selected feature subset, which is illustrated in Figure 4. As noticed in Figure 4c, the values of Ac, Sn, Sp, MCC, and auROC derived from using the eighteen informative PCPs are higher than using the 531 PCPs by 7%, 2%, 11%, 8%, and 3%, respectively. These results demonstrated that the inclusion of numerous redundant and uninformative features caused poor prediction results. For convenience, the best QSP predictor based on SVM model in conjunction with the eighteen informative PCPs will be referred to as iQSP.

2.3. Comparison with Existing Methods

To demonstrate the effectiveness and power of our method, we conducted a comparative study of our final model (named iQSP) with the existing methods. To date, there are only two existings methods developed for the prediction of QSPs, i.e., QSPpred [29] and QSPpred-FL [28,48], performing on the benchmark and independent datasets over 10-fold CV and independent validation test. Table 5 lists the preformance comparisons of iQSP and the existing methods. From Table 5, we can observe that QSPpred-FL yields the highest prediction performance of 94.30% Ac and 0.885 MCC over 10-fold CV, while our proposed model iQSP gave a 91.07 ± 1.77% Ac and 0.82 ± 0.04 MCC. On the other hand, based on the independent validation test, iQSP outperformed that other methods with 93.00 ± 1.97% Ac, 0.86 ± 0.04 MCC and 0.96 ± 0.02 auROC, which was better than the existing QSP predictors [28,29,48]. Although, iQSP achieved slightly better than QSPpred-FL, our proposed model showed significant improvement than QSPpred-FL considering the two objectives: using the less complexity of prediction methods (1 SVM vs. 99 RFs) and a minimum number of features used (18D vs. 913D).

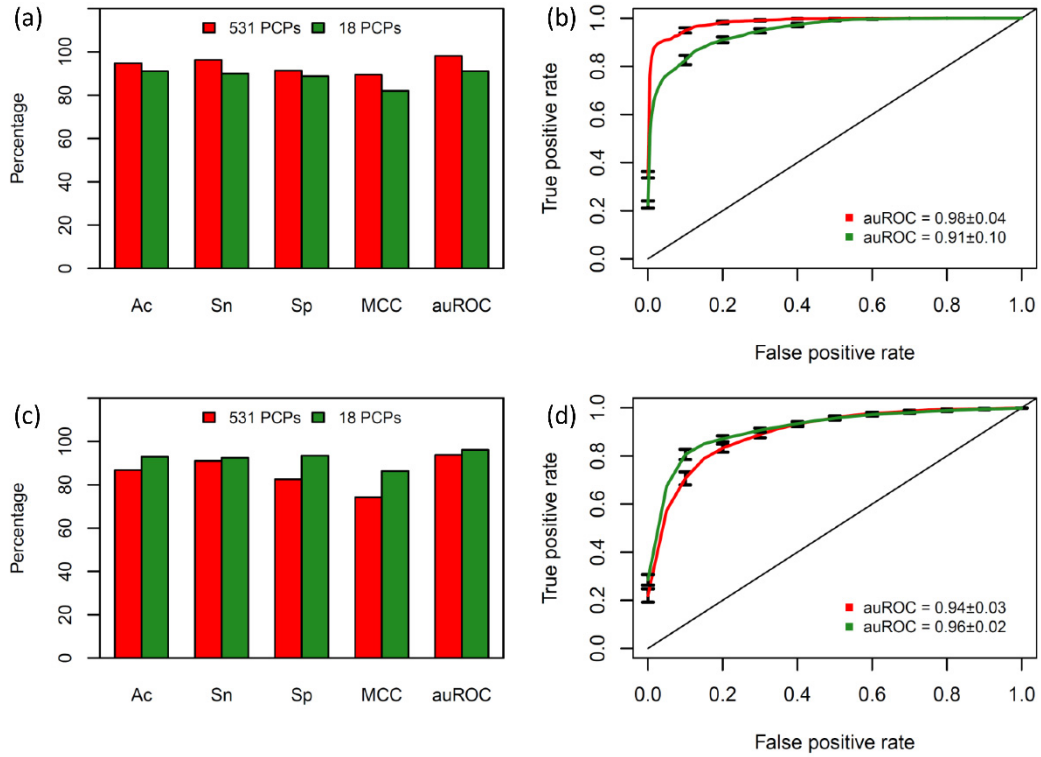

Figure 4. Performance comparisons of SVM models in conjunction with 531 PCPs and the eighteen informative PCPs assessed by 10-fold cross-validation (**a**,**b**) and independent validation test (**c**,**d**).

Table 5. Performance comparisons between iQSP and existing methods assessed by 10-fold cross-validation and independent validation tests.

Method	# Feature	10-Fold CV			Independent Test		
		Ac (%)	MCC	auROC	Ac (%)	MCC	auROC
QSPpred [a]	430	91.25	0.830	0.960	90.00	0.800	0.950
QSPpred-FL [b]	913	94.30	0.885	N/A	92.50	0.860	0.962
iQSP	18	91.07 ± 1.77	0.82 ± 0.04	0.91 ± 0.10	93.00 ± 1.97	0.86 ± 0.04	0.96 ± 0.02

[a] Results were reported from the work of QSPpred. [b] Results were obtained by feeding the peptide sequences in the independent set to the webserver of QSPpred-FL. # Feature represents the number of features used for constructing a model. N/A symbol represents the authors did not provide the result.

To further investigate the power of the proposed iQSP, we compared its performance with other conventional classifiers, i.e., k-nearest neighbor (k-NN), decision tree (rpart), and random forest (RF). The k-NN, rpart and RF classifiers were performed on the same datasets and implemented using the caret R package [54]. Rigorous 10-fold CV and independent validation test with ten independent rounds of these classifiers based on the optimal feature subset are reported in Table 6 and Figure 5. The more details of the parameter optimization of these three classifiers were described in the works [37,38,55–61]. Based the independent validation test, we noticed that the Ac, MCC and auROC values of iQSP were higher than those of other classifiers by >2%, >4%, and >2%, respectively, suggesting that iQSP holds very high potential to provide an accurate and reliable result in unseen peptides when compared to the existing methods and the conventional classifiers developed in this study.

Table 6. Performance comparison of iQSP with other conventional classifiers by using the optimal feature subset. Models were evaluated by means of 10-fold cross-validation and independent validation test subjected to ten rounds of random splits.

Classifier	10-Fold CV			Independent Test		
	Ac (%)	MCC	auROC	Ac (%)	MCC	auROC
k-NN	85.13 ± 0.27	0.72 ± 0.01	0.86 ± 0.00	85.75 ± 1.21	0.73 ± 0.02	0.91 ± 0.03
DT	83.57 ± 2.74	0.67 ± 0.06	0.87 ± 0.03	83.75 ± 3.39	0.68 ± 0.07	0.86 ± 0.05
RF	87.93 ± 0.48	0.76 ± 0.01	0.95 ± 0.01	91.00 ± 3.16	0.82 ± 0.06	0.94 ± 0.02
iQSP	91.07 ± 1.77	0.82 ± 0.04	0.91 ± 0.10	93.00 ± 1.97	0.86 ± 0.04	0.96 ± 0.02

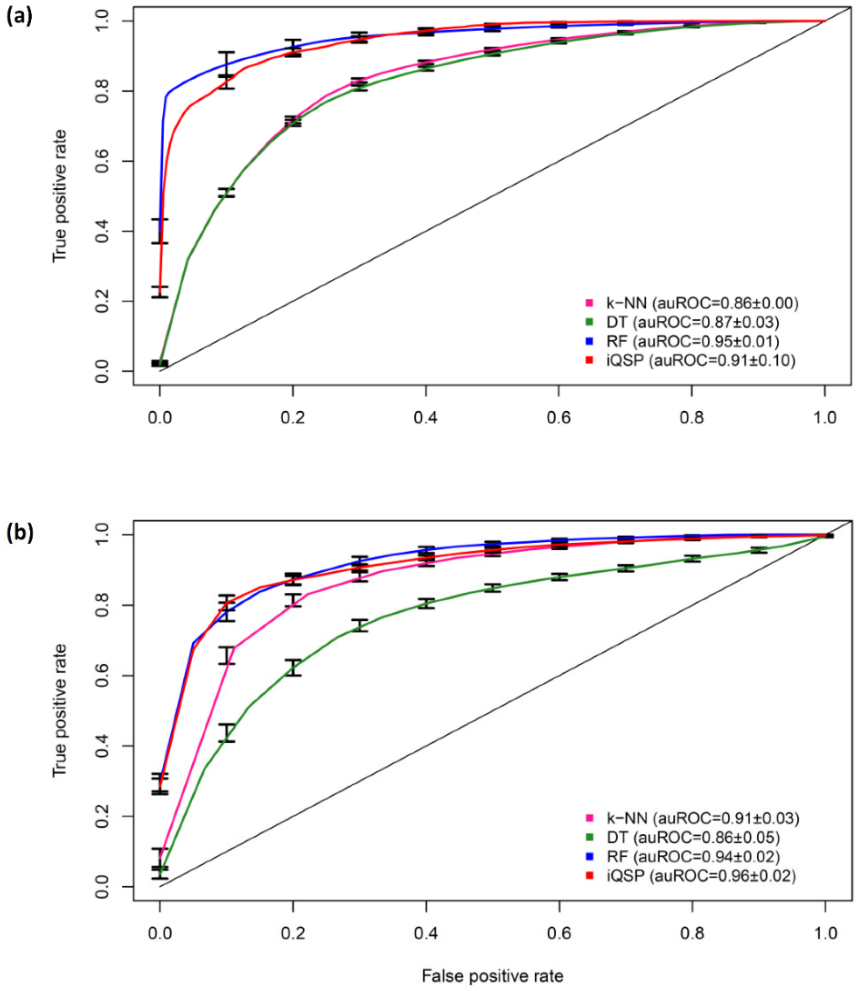

Figure 5. ROC curves of the proposed model iQSP with the conventional classifiers evaluated by 10-fold cross-validation (**a**) and the independent validation test (**b**) with ten independent rounds, where the bar represents the standard deviation of prediction results from ten independent round.

Based on Tables 3–6 and Figure 3, the superior performance of our proposed model iQSP over 10-fold CV and independent validation test might mainly be due to the following reasons: (i) Performing with multiple random sampling procedure to protect against the risk of having good predictive result by chance [39–43,49,50]; (ii) using an efficient feature selection method (GA-SAR) to identify m informative features from 531 PCPs. Using eighteen informative PCPs could provide faster and more cost-effective models, while model developers could gain an insight into the underlying prediction processes [58,62–64]; (iii) selecting a powerful method for QSP prediction. Although, iQSP displayed a superior performance over the existing methods assessed by the rigorous cross-validation methods, there is still room for further improvements, including increasing the size of QSPs by gathering peptide

sequences from various data sources, utilizing an interpretable learning algorithm, such as scoring card method [44,53], improving the interpretation of important features responsible for the biological activity [50,64] and exploring different ML algorithms, such as extreme gradient boosting [65] or deep learning [66].

2.4. Feature Contribution Analysis

Informative features play a vital role for developing an accurate predictor as well as providing a better understanding of QSPs. In this section, the eight informative PCPs utilized to analyze and characterize QSPs are explored. In addition, the MDGI score was used to rank the importance of those informative PCPs. Table 6 summarizes the detailed information of the eighteen informative PCPs and their corresponding MDGI scores. As noticed in Table 7, the top seven important PCPs are QIAN880137, AURR980102, ROBB760113, PRAM820101, GRAR740101, PALJ810111, and PONP800102 having an MDGI value of greater than 10. The most important PCP is the AAindex ID QIAN880137 with an MDGI value of 32.50 denoting 'weights for coil'. In the analysis, two out of the eighteen informative PCPs are related to AAC and composition, i.e., DAYM780101 (MDGI = 8.64) and GRAR740101 (MDGI = 12.26). Akanksha et al. [29] reported that using AAC and DPC as input features yield Ac values as high as 89.00% and 87.50%, respectively, as evaluated by the independent validation test. Furthermore, amongst the eighteen informative PCPs, there are two informative PCPs, i.e., PONP800102 (MDGI = 10.96) and MANP780101 (MDGI = 8.36), related to hydrophobicity. In 2007, Raymond et al. [47] tested their hypothesis that the hydrophobic face of 21-amino-acid signaling peptide might be important for receptor binding by replacing a hydrophobic residue (Phe) with a hydrophilic residue (Gln). In this study, three peptides, i.e., F7Q, F11Q, and F15Q, were synthesized and assessed for their abilities to activate quorum sensing. Their results indicated that the substitution of Phe with Gln significantly affected the activity of the signal peptide in activation of quorum sensing. Additionally, Akanksha et al. [29] also mentioned the importance of Phe in QSPs whereby Phe was dominant in the first and last 5 residues at N- and C-terminal regions. As noticed in Table 2, these results are consistent with our analysis results which show that Phe and Gln are ranked at 1 and 20, respectively.

2.5. Interpretable Rules Acquisition

In this work, the if–then interpretable rules called IR-QSP were constructed by using the RF method in conjunction with the optimal subset consisting of 18 informative PCPs (Table 7), as mentioned above. The main advantages of these constructed rules are twofold: (i) To demonstrate which PCP or which combination of PCPs are effective for QSP prediction, and (ii) to simply discriminate QSPs from Non-QSPs without the need to go through the mathematical and computational details. Tables 8 and A2 list eight interpretable rules that were important for QSP and twelve that were important for Non-QSP. If a query peptide meets all of the criteria in at least one of the eight rules, then it is identified as QSP. As observed in Tables 8 and A2, almost all the rules can yield a prediction accuracy of greater than 80%, except for rule #7. Interestingly, these four rules can achieve an Ac value of greater than 90%, i.e., rules #1, #2, #3, and #4. Thus, these results indicate that the construction of rules are reliable and easy-to-use, both in terms of their accuracy and interpretability for predicting and characterizing QSPs.

Table 7. The eighteen informative physicochemical properties [58] derived from the genetic algorithm utilizing self-assessment-report (GA-SAR) algorithm and their MDGI.

Rank	AAindex ID	MDGI	Description
1	QIAN880137	32.55	Weights for coil at the window position of 4 (Qian-Sejnowski, 1988)
2	AURR980102	16.62	Normalized positional residue frequency at helix termini N′ (Aurora-Rose, 1998)
3	ROBB760113	13.56	Information measure for loop (Robson-Suzuki, 1976)
4	PRAM820101	12.62	Intercept in regression analysis (Prabhakaran-Ponnuswamy, 1982)
5	GRAR740101	12.26	Composition (Grantham, 1974)
6	PALJ810111	11.71	Normalized frequency of beta-sheet in alpha + beta class (Palau et al., 1981)
7	PONP800102	10.96	Average gain in surrounding hydrophobicity (Ponnuswamy et al., 1980)
8	MUNV940103	9.07	Free energy in beta-strand conformation (Munoz-Serrano, 1994)
9	DAYM780101	8.64	Amino acid composition (Dayhoff et al., 1978a)
10	MANP780101	8.36	Average surrounding hydrophobicity (Manavalan-Ponnuswamy, 1978)
11	KUMS000103	8.23	Distribution of amino acid residues in the alpha-helices in thermophilic proteins (Kumar et al., 2000)
12	ROBB760104	8.18	Information measure for C-terminal helix (Robson-Suzuki, 1976)
13	ISOY800107	8.09	Normalized relative frequency of double bend (Isogai et al., 1980)
14	GEIM800101	7.80	Alpha-helix indices (Geisow-Roberts, 1980)
15	PRAM900102	7.59	Relative frequency in alpha-helix (Prabhakaran, 1990)
16	NADH010104	7.20	Hydropathy scale based on self-information values in the two-state model (20% accessibility) (Naderi-Manesh et al., 2001)
17	FUKS010106	6.47	Interior composition of amino acids in intracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001)
18	WIMW960101	5.54	Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water (Wimley-White, 1996)

2.6. iQSP Web Server

In an effort to maximize the full potential usage of the predictive model proposed in this study, the model along with optimal parameter settings were embedded inside an R powered website by means of the Shiny package. The resulting iQSP web server is publicly available at http://codes.bio/iqsp/. The server accepts an input the query peptide sequence in FASTA format that it submits for feature calculation and further fed into the predictive model for prediction of the class label as to whether it is a QSP or Non-QSP. Screenshots of the iQSP web server are shown in Figure 6.

Table 8. Fourteen if–then rules for the prediction of quorum sensing peptides using random forest and the 18 informative physicochemical properties.

No.	Rule	Cover Samples	Misclassified Samples	Ac (%)
1	GRAR740101 ≤ 0.9055 & MANP780101 > 0.7495 & PRAM900102 > 0.848 & QIAN880137 ≤ 0.237	10	0	100.00
2	PONP800102 > −0.751 & PONP800102 ≤ 1.0025 & QIAN880137 ≤ −0.104 & ROBB760104 ≤ 0.3645 & ROBB760104 > −0.5205	61	1	98.36
3	PALJ810111 ≤ 1.369 & QIAN880137 > −0.104 & QIAN880137 ≤ 0.417 & ROBB760113 ≤ 0.5975 & AURR980102 ≤ 0.6955	21	2	90.48
4	GEIM800101 > −0.3135 & GRAR740101 > −0.176 & ISOY800107 ≤ 1.367 & MANP780101 > −0.3325 & PALJ810111 ≤ 1.0905 & QIAN880137 ≤ −0.0985	94	6	93.62
5	PALJ810111 > −0.786 & QIAN880137 > 0.237 & QIAN880137 > 0.403 & ROBB760113 > 0.5975 & AURR980102 ≤ 0.811 & KUMS000103 ≤ 0.793	45	7	84.44
6	GRAR740101 ≤ 0.341 & ISOY800107 ≤ −0.089 & PALJ810111 ≤ 1.2455 & QIAN880137 > −0.009 & ROBB760113 ≤ −0.0285	17	3	82.35
7	GRAR740101 > −0.708 & PRAM900102 ≤ 1.2985 & QIAN880137 > −0.104 & QIAN880137 ≤ 1.105 & AURR980102 ≤ 0.6585 & KUMS000103 ≤ 0.974	94	28	70.21
8	PONP800102 ≤ 1.1095 & PRAM900102 ≤ 0.8295 & QIAN880137 ≤ 0.2625 & QIAN880137 > −0.9055 & ROBB760113 > −0.5875 & ROBB760113 ≤ 1.031	121	15	87.60

The procedure for using the iQSP web server for predicting the class label can be summarized as follows:

Step 1. Proceed to the iQSP web server by going to the URL, http://codes.bio/iqsp/ and wait until the text box below the "Status/Output" (found on the right hand side) returns the message [1] "Server is ready for prediction".

Step 2. Enter the query peptide sequence (in FASTA format) into the text box found below the text "Enter your input sequence(s) in FASTA format". Alternatively, the user can also save the FASTA sequences into a text file and upload this text file to the server by clicking on the "Choose File" found below the text "or upload file".

Step 3. Finally, click on the "Submit" button to start the prediction process. Shortly after, the prediction results will be displayed in the text box found below the "Status/Output". This will return the prediction results as a 4 column verbatim text for which the first, second, third and fourth column corresponds to an arbitrary sequence identification number (1, 2, 3, etc.), predicted class label (i.e., as either QSP or Non-QSP), the probability score of the query peptide being a Non-QSP and the probability score of the query peptide being a QSP, respectively.

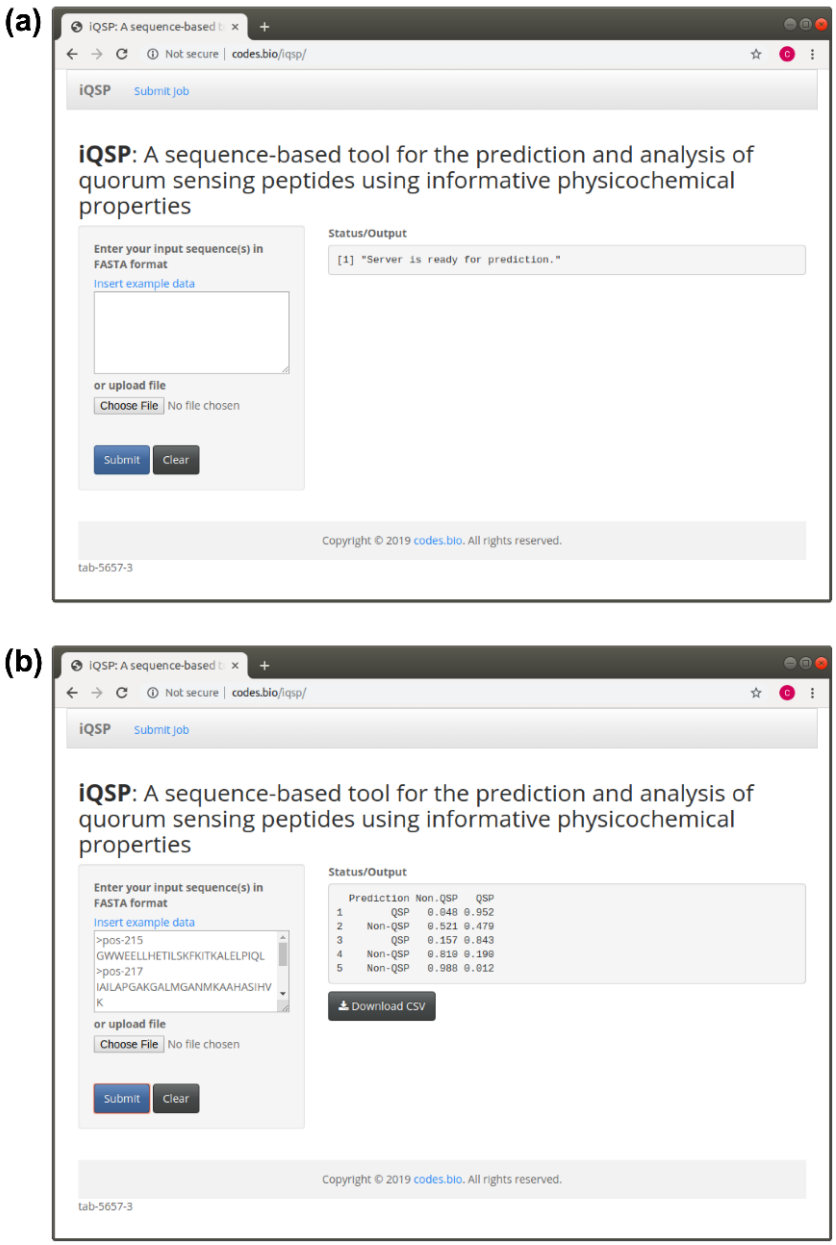

Figure 6. Screenshot of the iQSP web server before (**a**) and after (**b**) submission of the input peptide sequence.

3. Materials and Methods

In order to establish a robust sequence-based tool for modeling the investigated QSPs, we followed Chou's five-step guidelines as mentioned in a series of recent publications [67–72] and summarized in two comprehensive review papers [34,35]: (i) Compilation of a reliable dataset that contains experimentally validated sequences for training and validating the model; (ii) quantifying peptides sequences to describe their physicochemical properties; (iii) developing the prediction model using robust algorithm; (iv) assess the prediction model using standard cross-validation tests; and (v) constructing a user-friendly web-server for obtaining the prediction without the need to understand complex mathematical and statistical details. Furthermore, Figure 2 shows the workflow of iQSP which works in discriminating peptides as QSPs or Non-QSPs.

3.1. Benchmark Dataset

To make a fair comparison with the existing methods, the same benchmark and independent datasets derived from the work [29] were taken to develop and validate the proposed model. In this study, the benchmark dataset consists of 200 QSPs and 200 Non-QSPs, while the independent dataset consists of 20 QSPs and 20 Non-QSPs. The benchmark (S_{TR}) and independent (S_{TS}) datasets used in this study can be summarized by the following formula:

$$S_{TR} = S_{TR}^{+} \cup S_{TR}^{-} \quad (1)$$

$$S_{TS} = S_{TS}^{+} \cup S_{TS}^{-} \quad (2)$$

where S^{+} and S^{-} represent peptide sequences of QSPs and Non-QSPs, respectively, while the symbol $\cup$ represents the union from the set theory. However, these two datasets were constructed with a single random sampling procedure. Therefore, to alleviate the impact of the random sampling procedure, we repeated this construction procedure with ten independent rounds and the final prediction results were obtained by averaging the five statistical parameters.

3.2. Feature Representation

PCP is one of the most intuitive features associated with biophysical and biochemical reactions and is also referred to as an available, easy and interpretable feature. In fact, a total of 531 PCPs without NA values were derived from version 9.0 of the Amino acid index database (AAindex) [58], which is a collection of the published literature pertaining to different physicochemical and biophysical properties of amino acids and pairs of amino acids (http://www.genome.jp/aaindex/). Each PCP consisted of a set of 20 numerical values for amino acids. The PCP feature has been extensively used for the prediction and analysis of various protein [42,52,53] and peptide [36,39–41,45] functions. To utilize PCP features for extracting a peptide sequence, peptide with the length of L amino acid residues is encoded into an L-dimensional vector of 531 PCPs (531D). In this study, the number of L is in the range of 5–20, i.e., 5, 6, 7, . . . , 20. Additional details on how to obtain a minimum number of L can be found in the sub-section of Identification of Informative Physicochemical Properties.

3.3. Support Vector Machine

SVM method is an effective ML algorithm for supervised pattern and has been widely used in various biological problems [36–39,41,43,45,52,73–80]. This method is based on the Vapnik–Chervonenkis theory of statistical learning [81–83]. The basic idea of this method is to map the original feature vectors having p-dimensional vector into a higher Hilbert space with n-dimensional vector, where p < n, and then determine a separate hyper plane with the largest distance between the two classes. In this work, each sample on the benchmark (S_{TR}) and independent (S_{TS}) datasets have a corresponding label (−1 and 1) where +1 and −1 represent QSP and Non-QSP, respectively. In this study, the *kernlab* R package [84] was used to implement the SVM model. To enhance the performance of the SVM model, the regularization parameter C and kernel parameter γ were tuned by using grid search method with a cross-validation technique, of which the search space for C and γ are $[2^{-8}, 2^{8}]$ and $[2^{-8}, 2^{8}]$ with steps of 2 and 2, respectively.

3.4. Performance Evaluation

In statistical predictions, there are three testing methods most often used to assess the predictive ability of the model: (i) Sub-sampling test (2-, 5- or 10-fold cross-validation), (ii) jackknife test also known as leave-one-out cross-validation (LOO-CV), and (iii) independent (or external) validation test. The sub-sampling test is one of the most popular cross-validation methods to assess the predictive capability of a model. As described in [35,85] and investigated by Equation (50) of [86], among those testing methods, the jackknife test is considered as one of the most rigorous that can provide a

unique result for a given benchmark dataset. However, to perform a fair comparison with the existing methods and reduce the computational time, the 10-fold cross-validation (10-fold CV) and independent validation test were used to evaluate the prediction performance of our models. The former set (10-fold CV) makes use of data from the training set where the data set is separated into ten subsets. Practically, one subset from a total of ten subsets is left out as the testing set while the remaining are used for training the model. This process is repeated iteratively until all data samples have had the chance to be left out as the testing set.

In order to evaluate the prediction ability of the model, the following sets of four metrics are used:

$$\mathrm{Ac} = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \tag{3}$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{4}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})} \tag{5}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{6}$$

where Ac, Sn, Sp, and MCC represent accuracy, sensitivity, specificity and Matthews coefficient correlation, respectively. TP, TN, FP, and FN represent the instances of true positive, true negative, false positive, and false negative, respectively. Moreover, in order to evaluate the prediction performance of models using threshold-independent parameters, receiver operating characteristic (ROC) curves were plotted using the pROC package in the R software [87]. The area under the ROC curve (AUC) was used to measure the prediction performance, where AUC values of 0.5 and 1 are indicative of perfect and random models, respectively.

3.5. Identification of Informative Physicochemical Properties

In order to determine a minimal number of informative features while maximizing prediction accuracy of SVM model, the customized implementation of genetic algorithm (GA), called genetic algorithm utilizing self-assessment-report (GA-SAR), was developed to elucidate the mentioned problem by taking an advantage of an inheritable GA [88] to select an informative feature. GA-SAR utilizes a self-assessment-report (SAR) approach to construct a profile used for reporting the usefulness of 531 PCPs based on the assumption that a good feature will be highly correlated with the output variable, but uncorrelated to each other [51,89]. Before starting the GA-SAR process, the usefulness of 531 PCPs are calculated. After that, the profile is applied in a mutation function of GA to create a chance of adding/deleting features. During the process of GA optimization, features having low usefulness scores are deleted, while a minimal number of informative PCPs are specified.

The chromosome of GA-SAR consisted of binary genes indicating an occurrence of each feature in the feature set and the parameter of model genes for tuning the parameters of the classifier. Herein, the binary genes contain 531 genes and two 4-bit for encoding the parameters C (2^{-8}, 2^{-7}, ... , 2^{8}) and γ (2^{-8}, 2^{-7}, ... , 2^{8}) of the SVM model. The identification of informative PCPs using GA-SAR algorithm is described as follows:

(1) Randomly generate 50 chromosomes with randomly assigned values of binary genes to make the number of features (m) equal to our preferred number, where m is in the range from 5 to 20. (Initialization)

(2) Assess the prediction performances for each chromosome over 10-fold CV procedure. (Evaluation)

(3) Implement a tournament selection to prepare a mating pool. (Selecting)

(4) Perform a 20-point crossover on the selected parents. (Crossover)

(5) Apply the SAR mutation operator and if the number of chosen genes are greater than the specified number of features, delete some genes. In contrast, if the number of chosen genes are less than the specified number of features, add some genes. The probability of genes to be added or deleted is referred to as SAR. (Mutation)

(6) Stop if the number of generation is equal to 50; otherwise go to Step 2. (Termination)

3.6. Construction of Interpretable Rules

This work presents an interpretable rule extraction of QSPs (IR-QSP) based on the RF method cooperating with the *m* informative PCPs for determining the biophysical and biochemical properties of QSPs. A set of rules from an individual tree is derived from the root to the leaves. In this study, only 100 decision trees were used to extract the if–then interpretable rules for explaining the prediction results by means of RF method, inTrees and xtable packages in the R software [90–92]. More details of the rule extraction process can be found in previous related works [36,61,93].

3.7. Construction of the iQSP Web Server

The best performing model described in this study is used as the basis for deployment as a web server. Particularly, the underlying machine learning model is encapsulated inside a website by means of the Shiny package in an R programming environment. By default of the Shiny package, the user interface follows a responsive web design principle in which the website can display optimally in various device platform whether it be a mobile phone, tablet or desktop computers of various screen resolutions. The code and data used to operate the iQSP web server is hosted on Digital Ocean and is publicly available at http://codes.bio/iqsp/. In fact, user-friendly and publicly accessible web-servers that can display the findings manipulated by users according to their need might significantly enhance their impacts, driving medicinal chemistry into an unprecedented revolution [34,94–107]. Keeping this point in our mind, we shall make efforts in our future work to provide a web-server to provide such functionality.

4. Conclusions

An accurate tool for predicting and analyzing QSPs is essential for understanding their roles in clinical applications and providing a promising way to combat bacterial infections. Thus, in this study, we developed an efficient and interpretable sequence-based predictor for predicting and analyzing QSPs, called iQSP, by utilizing a set of m informative physicochemical properties (PCPs) in conjunction with a powerful support vector machine (SVM). Performance comparisons for both rigorous cross-validation and independent validation tests demonstrated the superiority of iQSP over the existing QSP predictors. Moreover, feature selection and interpretable rule extraction were carried out to construct easy-to-use if–then rules (IR-QSP) for discriminating QSPs from Non-QSPs by using random forest (RF) model and the m informative PCPs. Finally, to help potential users of iQSP, a web server named iQSP was implemented and made freely available online at https://codes.bio/iqsp/. It is anticipated that iQSP may become a powerful and cost-effective approach for predicting and analyzing peptides on a large scale. Due to various potential application of our systematic approach employed in this study, we could extend this for predicting and analyzing many other types of protein and peptide functions such as *S*-palmitoylation sites in proteins [68], lysine crotonylation sites [108], and phosphotyrosine sites [109]. Furthermore, our method could be integrated with other beneficial peptide features such as pseudo amino acid composition [110–112] or amphiphilic pseudo amino acid composition as proposed [113] by Chou [35,114] for further improving the QSP prediction

Author Contributions: W.S. conceived, designed, performed, and analyzed the experiments. W.S. and P.C. performed the feature selection and the analysis. P.C. contributed the GA-SAR code. N.S., T.P., P.C. and W.S. drafted the manuscript. W.S., P.C. and C.N. contributed the code for constructing the web server. C.N. vetted the manuscript. All authors reviewed and approved the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was co-supported by the TRF Research Grant for New Scholar (No. MRG6180226) and College of Arts, Media and Technology, Chiang Mai University, and partially supported by the TRF Research Career Development Grant (No. RSA6280075) from the Thailand Research Fund the Office of Higher Education Commission and Mahidol University.

Acknowledgments: The authors thank Supicha Kachenton, Chuleeporn Phanus-umporn and Nuttapat Anuwongcharoen for valuable ideas and comments. We also thank the peer reviewers for their valuable comments in helping to improve this manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

Table A1. Ten different feature subsets and their corresponding physicochemical properties in the AAindex database.

Subset	# Feature	AAindex
1	14	FASG760103, ISOY800102, KYTJ820101, LIFS790103, MEIH800102, OOBM850101, PALJ810113, QIAN880137, ROSG850101, TANS770101, AURR980116, VINM940104, NADH010107, FUKS010106
2	17	ANDN920101, GRAR740102, JANJ780103, LIFS790101, NAGK730102, PALJ810102, QIAN880137, RACS820102, ROBB760109, ROBB760111, SUEM840102, TANS770104, TANS770108, AURR980115, VINM940103, BLAM930101, PUNT030102
3	16	EISD860101, GEIM800105, HOPA770101, LIFS790101, MEIH800103, OOBM770101, PALJ810111, QIAN880128, RACS770101, ROBB760107, ROBB790101, AURR980110, VINM940101, KIMC930101, FUKS010106, TSAJ990102,
4	17	DAYM780101, ISOY800108, MEIH800102, NAKH900104, PALJ810111, PALJ810113, PRAM820102, RACS770101, ROBB760111, SNEP660104, AURR980118, VINM940101, PARS000101, NADH010106, FUKS010106, TSAJ990101, GEOR030107
5	17	ANDN920101, FINA910102, GEIM800102, GRAR740102, ISOY800101, LIFS790102, NOZY710101, PALJ810106, QIAN880138, ROBB760107, ROBB760112, ZIMJ680101, AURR980113, MUNV940102, MUNV940105, MONM990201, FUKS010109
6	18	DAYM780101, GEIM800101, GRAR740101, ISOY800107, MANP780101, PALJ810111, PONP800102, PRAM820101, PRAM900102, QIAN880137, ROBB760104, ROBB760113, AURR980102, MUNV940103, WIMW960101, KUMS000103, NADH010104, FUKS010106
7	15	GRAR740101, ISOY800108, LIFS790103, MIYS850101, OOBM850105, PALJ810114, QIAN880137, RACS820109, ROBB760108, ROSG850101, TANS770107, AURR980116, VINM940104, PARS000102, GEOR030108
8	17	ANDN920101, ARGP820103, BIGC670101, BULH740101, BUNA790102, CHOC760104, CHOP780211, CHOP780212, DESM900101, FASG760101, GRAR740101, NAKH900101, RADA880107, ROBB790101, TANS770109, GUOD860101, PONJ960101
9	18	FAUJ880112, FINA910103, GRAR740101, JOND750102, NISK860101, PALJ810108, PONP800104, PRAM820101, QIAN880125, QIAN880134, ROBB760108, WERD780104, AURR980113, VINM940101, VINM940102, PARS000102, FUKS010108, WOLR790101
10	17	ANDN920101, BROC820102, BUNA790101, CHAM830104, CHOP780207, CIDH920101, CIDH920104, DESM900102, FASG760101, FINA910102, JANJ780101, LEWP710101, PALJ810110, QIAN880125, VASM830101, AURR980108, COSI940101

Feature represents the number of features used for constructing a model.

Table A2. Twenty if–then rules for discriminating QSPs from Non-QSPs using random forest model in conjunction with the eighteen informative physicochemical properties.

% Covered Samples	Rule	Prediction Result
3.57	GRAR740101 ≤ 0.9055 & MANP780101 > 0.7495 & PRAM900102 > 0.848 & QIAN880137 ≤ 0.237	QSP
14.03	PONP800102 > −0.751 & PONP800102 ≤ 1.0025 & QIAN880137 ≤ −0.104 & ROBB760104 ≤ 0.3645 & ROBB760104 > −0.5205	QSP
4.08	PALJ810111 ≤ 1.369 & QIAN880137 > −0.104 & QIAN880137 ≤ 0.417 & ROBB760113 ≤ 0.5975 & AURR980102 ≤ 0.6955	QSP
9.95	GEIM800101 > −0.3135 & GRAR740101 > −0.176 & ISOY800107 ≤ 1.367 & MANP780101 > −0.3325 & PALJ810111 ≤ 1.0905 & QIAN880137 ≤ −0.0985	QSP
9.95	PALJ810111 > −0.786 & QIAN880137 > 0.237 & QIAN880137 > 0.403 & ROBB760113 > 0.5975 & AURR980102 ≤ 0.811 & KUMS000103 ≤ 0.793	QSP
2.81	GRAR740101 ≤ 0.341 & ISOY800107 ≤ −0.089 & PALJ810111 ≤ 1.2455 & QIAN880137 > −0.009 & ROBB760113 ≤ −0.0285	QSP
2.81	GRAR740101 > −0.708 & PRAM900102 ≤ 1.2985 & QIAN880137 > −0.104 & QIAN880137 ≤ 1.105 & AURR980102 ≤ 0.6585 & KUMS000103 ≤ 0.974	QSP
1.53	PONP800102 ≤ 1.1095 & PRAM900102 ≤ 0.8295 & QIAN880137 ≤ 0.2625 & QIAN880137 > −0.9055 & ROBB760113 > −0.5875 & ROBB760113 ≤ 1.031	QSP
11.73	GEIM800101 > −0.547 & PONP800102 > 0.9625 & PRAM900102 ≤ 1.026	Non-QSP
3.57	QIAN880137 > −0.0985 & QIAN880137 ≤ 0.482 & ROBB760113 > 0.5975 & AURR980102 > −0.1535	Non-QSP
2.30	QIAN880137 > 0.257 & ROBB760113 > 0.613 & ROBB760113 > 0.7775 & AURR980102 > 0.584	Non-QSP
1.79	DAYM780101 ≤ 0.1035 & MANP780101 > 0.5635 & QIAN880137 ≤ 0.2625 & ROBB760113 > 0.5215	Non-QSP
1.53	GRAR740101 > −0.045 & PALJ810111 > 1.0905 & QIAN880137 ≤ −0.0985	Non-QSP
6.12	MANP780101 > −0.5985 & ROBB760104 > −0.3545 & AURR980102 ≤ 0.5935 & AURR980102 ≤ 0.5125 & KUMS000103 > 0.663	Non-QSP
3.57	PONP800102 > 0.5045 & QIAN880137 > −0.039 & ROBB760104 > −0.1345 & ROBB760113 ≤ 0.5975 & AURR980102 ≤ 1.22 & FUKS010106 ≤ 1.1985	Non-QSP
1.28	QIAN880137 > 0.2625 & QIAN880137 ≤ 0.7795 & ROBB760104 > −0.3885 & AURR980102 ≤ 0.6585 & AURR980102 ≤ 0.515 & NADH010104 > −0.023	Non-QSP
3.06	GEIM800101 ≤ 1.636 & PONP800102 ≤ 0.6385 & PRAM900102 > −0.017 & PRAM900102 > 0.627 & QIAN880137 > −0.009	Non-QSP
2.55	PALJ810111 ≤ 0.436 & PRAM900102 > 0.1605 & QIAN880137 > −0.009 & AURR980102 ≤ 0.6525	Non-QSP
1.53	GEIM800101 ≤ −0.3305 & GRAR740101 > −0.2425 & PONP800102 ≤ 0.98 & QIAN880137 ≤ 0.237	Non-QSP
12.24	Else	Non-QSP

References

1. Fuqua, W.C.; Winans, S.C.; Greenberg, E.P. Quorum sensing in bacteria: The LuxR-LuxI family of cell density-responsive transcriptional regulators. *J. Bacteriol.* **1994**, *176*, 269–275. [CrossRef]
2. Waters, C.M.; Bassler, B.L. Quorum sensing: Cell-to-cell communication in bacteria. *Annu. Rev. Cell Dev. Biol.* **2005**, *21*, 319–346. [CrossRef]
3. Rutherford, S.T.; Bassler, B.L. Bacterial quorum sensing: Its role in virulence and possibilities for its control. *Cold Spring Harb. Perspect. Med.* **2012**, *2*, a012427. [CrossRef]
4. Remy, B.; Mion, S.; Plener, L.; Elias, M.; Chabriere, E.; Daude, D. Interference in bacterial quorum sensing: A biopharmaceutical perspective. *Front. Pharm.* **2018**, *9*, 203. [CrossRef]
5. Monnet, V.; Juillard, V.; Gardan, R. Peptide conversations in Gram-positive bacteria. *Crit. Rev. Microbiol.* **2016**, *42*, 339–351. [CrossRef]
6. Schuster, M.; Sexton, D.J.; Diggle, S.P.; Greenberg, E.P. Acyl-homoserine lactone quorum sensing: From evolution to application. *Annu. Rev. Microbiol.* **2013**, *67*, 43–63. [CrossRef]
7. Hazan, R.; Que, Y.A.; Maura, D.; Strobel, B.; Majcherczyk, P.A.; Hopper, L.R.; Wilbur, D.J.; Hreha, T.N.; Barquera, B.; Rahme, L.G. Auto poisoning of the respiratory chain by a quorum-sensing-regulated molecule favors biofilm formation and antibiotic tolerance. *Curr. Biol.* **2016**, *26*, 195–206. [CrossRef]
8. Zhou, S.; Zhang, A.; Yin, H.; Chu, W. Bacillus sp. QSI-1 modulate quorum sensing signals reduce aeromonas hydrophila level and alter gut microbial community structure in fish. *Front. Cell. Infect. Microbiol.* **2016**, *6*, 184. [CrossRef]
9. Tiaden, A.; Hilbi, H. alpha-Hydroxyketone synthesis and sensing by Legionella and Vibrio. *Sensors* **2012**, *12*, 2899–2919. [CrossRef]
10. Kendall, M.M.; Sperandio, V. Quorum sensing by enteric pathogens. *Curr. Opin. Gastroenterol.* **2007**, *23*, 10–15. [CrossRef]
11. Heeb, S.; Fletcher, M.P.; Chhabra, S.R.; Diggle, S.P.; Williams, P.; Camara, M. Quinolones: From antibiotics to autoinducers. *FEMS Microbiol. Rev.* **2011**, *35*, 247–274. [CrossRef]
12. Chen, X.; Schauder, S.; Potier, N.; Van Dorsselaer, A.; Pelczer, I.; Bassler, B.L.; Hughson, F.M. Structural identification of a bacterial quorum-sensing signal containing boron. *Nature* **2002**, *415*, 545–549. [CrossRef]
13. Mukherjee, S.; Bassler, B.L. Bacterial quorum sensing in complex and dynamically changing environments. *Nat. Rev. Microbiol.* **2019**, *17*, 371–382. [CrossRef]
14. Whiteley, M.; Diggle, S.P.; Greenberg, E.P. Progress in and promise of bacterial quorum sensing research. *Nature* **2017**, *551*, 313–320. [CrossRef]
15. Jiang, Q.; Chen, J.; Yang, C.; Yin, Y.; Yao, K. Quorum Sensing: A prospective therapeutic target for bacterial diseases. *Biomed. Res. Int.* **2019**, *2019*, 2015978. [CrossRef]
16. Papenfort, K.; Bassler, B.L. Quorum sensing signal-response systems in Gram-negative bacteria. *Nat. Rev. Microbiol.* **2016**, *14*, 576–588. [CrossRef]
17. Verbeke, F.; De Craemer, S.; Debunne, N.; Janssens, Y.; Wynendaele, E.; Van de Wiele, C.; De Spiegeleer, B. Peptides as quorum sensing molecules: Measurement techniques and obtained levels in vitro and in vivo. *Front. Neurosci.* **2017**, *11*, 183. [CrossRef]
18. De Kievit, T.R.; Gillis, R.; Marx, S.; Brown, C.; Iglewski, B.H. Quorum-sensing genes in Pseudomonas aeruginosa biofilms: Their role and expression patterns. *Appl. Environ. Microbiol.* **2001**, *67*, 1865–1873. [CrossRef]
19. Lee, J.; Zhang, L. The hierarchy quorum sensing network in Pseudomonas aeruginosa. *Protein Cell* **2015**, *6*, 26–41. [CrossRef]
20. Eickhoff, M.J.; Bassler, B.L. SnapShot: Bacterial quorum sensing. *Cell* **2018**, *174*, 1328–1328.e1. [CrossRef]
21. Gallardo-Garcia, M.M.; Sanchez-Espin, G.; Ivanova-Georgieva, R.; Ruiz-Morales, J.; Rodriguez-Bailon, I.; Vinuela Gonzalez, V.; Garcia-Lopez, M.V. Relationship between pathogenic, clinical, and virulence factors of Staphylococcus aureus in infective endocarditis versus uncomplicated bacteremia: A case-control study. *Eur. J. Clin. Microbiol. Infect. Dis.* **2016**, *35*, 821–828. [CrossRef]
22. Liu, J.; Yu, H.; Huang, Y.; Yu, Z.; Fan, G.; Jin, J.; Liu, X.; Wang, G. Complete genome sequence of a novel bacteriophage infecting Bradyrhizobium diazoefficiens USDA110. *Sci. China Life Sci.* **2018**, *61*, 118–121. [CrossRef]

23. Han, M.; Gu, J.; Gao, G.F.; Liu, W.J. China in action: National strategies to combat against emerging infectious diseases. *Sci. China Life Sci.* **2017**, *60*, 1383–1385. [CrossRef]
24. Vadekeetil, A.; Saini, H.; Chhibber, S.; Harjai, K. Exploiting the antivirulence efficacy of an ajoene-ciprofloxacin combination against Pseudomonas aeruginosa biofilm associated murine acute pyelonephritis. *Biofouling* **2016**, *32*, 371–382. [CrossRef]
25. Bahari, S.; Zeighami, H.; Mirshahabi, H.; Roudashti, S.; Haghi, F. Inhibition of Pseudomonas aeruginosa quorum sensing by subinhibitory concentrations of curcumin with gentamicin and azithromycin. *J. Glob. Antimicrob. Resist.* **2017**, *10*, 21–28. [CrossRef]
26. Furiga, A.; Lajoie, B.; El Hage, S.; Baziard, G.; Roques, C. Impairment of pseudomonas aeruginosa biofilm resistance to antibiotics by combining the drugs with a new quorum-sensing inhibitor. *Antimicrob. Agents Chemother.* **2015**, *60*, 1676–1686. [CrossRef]
27. Brackman, G.; Cos, P.; Maes, L.; Nelis, H.J.; Coenye, T. Quorum sensing inhibitors increase the susceptibility of bacterial biofilms to antibiotics in vitro and in vivo. *Antimicrob. Agents Chemother.* **2011**, *55*, 2655–2661. [CrossRef]
28. Wei, L.; Hu, J.; Li, F.; Song, J.; Su, R.; Zou, Q. Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. *Brief. Bioinform.* **2018**, *10*, 1–14. [CrossRef]
29. Rajput, A.; Gupta, A.K.; Kumar, M. Prediction and analysis of quorum sensing peptides based on sequence features. *PLoS ONE* **2015**, *10*, e0120066. [CrossRef]
30. Chen, W.; Ding, H.; Feng, P.; Lin, H.; Chou, K.-C. iACP: A sequence-based tool for identifying anticancer peptides. *Oncotarget* **2016**, *7*, 16895. [CrossRef]
31. Raiko, T.; Ilin, A.; Karhunen, J. Principal component analysis for large scale problems with lots of missing values. In *Proceedings of European Conference on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 691–698.
32. Subramanian, J.; Simon, R. Overfitting in prediction models—Is it a problem only in high dimensions? *Contemp. Clin. Trials* **2013**, *36*, 636–641. [CrossRef]
33. Xing, E.P.; Jordan, M.I.; Karp, R.M. Feature selection for high-dimensional genomic microarray data. *ICML* **2001**, *1*, 601–608.
34. Chou, K.-C. Advance in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs. *Curr. Med. Chem.* **2019**, *26*, 4918–4943. [CrossRef]
35. Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **2011**, *273*, 236–247. [CrossRef]
36. Schaduangrat, N.; Nantasenamat, C.; Prachayasittikul, V.; Shoombuatong, W. ACPred: A computational tool for the prediction and analysis of anticancer peptides. *Molecules* **2019**, *24*, 1973. [CrossRef]
37. Simeon, S.; Li, H.; Win, T.S.; Malik, A.A.; Kandhro, A.H.; Piacham, T.; Shoombuatong, W.; Nuchnoi, P.; Wikberg, J.E.; Gleeson, M.P. PepBio: Predicting the bioactivity of host defense peptides. *RSC Adv.* **2017**, *7*, 35119–35134. [CrossRef]
38. Simeon, S.; Shoombuatong, W.; Anuwongcharoen, N.; Preeyanon, L.; Prachayasittikul, V.; Wikberg, J.E.; Nantasenamat, C. osFP: A web server for predicting the oligomeric states of fluorescent proteins. *J. Cheminform.* **2016**, *8*, 72. [CrossRef]
39. Win, T.S.; Schaduangrat, N.; Prachayasittikul, V.; Nantasenamat, C.; Shoombuatong, W. PAAP: A web server for predicting antihypertensive activity of peptides. *Future Med. Chem.* **2018**, *10*, 1749–1767. [CrossRef]
40. Laengsri, V.; Nantasenamat, C.; Schaduangrat, N.; Nuchnoi, P.; Prachayasittikul, V.; Shoombuatong, W. TargetAntiAngio: A sequence-based tool for the prediction and analysis of anti-angiogenic peptides. *Int. J. Mol. Sci.* **2019**, *20*, 2950. [CrossRef]
41. Shoombuatong, W.; Schaduangrat, N.; Pratiwi, R.; Nantasenamat, C. THPep: A machine learning-based approach for predicting tumor homing peptides. *Comput. Biol. Chem.* **2019**, *80*, 441–451. [CrossRef]
42. Pratiwi, R.; Malik, A.A.; Schaduangrat, N.; Prachayasittikul, V.; Wikberg, J.E.; Nantasenamat, C.; Shoombuatong, W. CryoProtect: A web server for classifying antifreeze proteins from nonantifreeze proteins. *J. Chem.* **2017**, *2017*, 15. [CrossRef]
43. Hongjaisee, S.; Nantasenamat, C.; Carraway, T.S.; Shoombuatong, W. HIVCoR: A sequence-based tool for predicting HIV-1 CRF01_AE coreceptor usage. *Comput. Biol. Chem.* **2019**, *80*, 419–432. [CrossRef]

44. Charoenkwan, P.; Shoombuatong, W.; Lee, H.-C.; Chaijaruwanich, J.; Huang, H.-L.; Ho, S.-Y. SCMCRYS: Predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. *PLoS ONE* **2013**, *8*, e72368. [CrossRef]
45. Win, T.S.; Malik, A.A.; Prachayasittikul, V.S.; Wikberg, J.E.; Nantasenamat, C.; Shoombuatong, W. HemoPred: A web server for predicting the hemolytic activity of peptides. *Future Med. Chem.* **2017**, *9*, 275–291. [CrossRef]
46. Crooks, G.E.; Hon, G.; Chandonia, J.-M.; Brenner, S.E. WebLogo: A sequence logo generator. *Genome Res.* **2004**, *14*, 1188–1190. [CrossRef]
47. Syvitski, R.T.; Tian, X.-L.; Sampara, K.; Salman, A.; Lee, S.F.; Jakeman, D.L.; Li, Y.-H. Structure-activity analysis of quorum-sensing signaling peptides from Streptococcus mutans. *J. Bacteriol.* **2007**, *189*, 1441–1450. [CrossRef]
48. Wei, L.; Zhou, C.; Su, R.; Zou, Q. PEPred-Suite: Improved and robust prediction of therapeutic peptides using adaptive feature representation learning. *Bioinformatics* **2019**, *35*, 4272–4280. [CrossRef]
49. Shoombuatong, W.; Schaduangrat, N.; Nantasenamat, C. Towards understanding aromatase inhibitory activity via QSAR modeling. *EXCLI J.* **2018**, *17*, 688.
50. Shoombuatong, W.; Schaduangrat, N.; Nantasenamat, C. Unraveling the bioactivity of anticancer peptides as deduced from machine learning. *EXCLI J.* **2018**, *17*, 734.
51. Shoombuatong, W.; Prachayasittikul, V.; Prachayasittikul, V.; Nantasenamat, C. Prediction of aromatase inhibitory activity using the efficient linear method (ELM). *EXCLI J.* **2015**, *14*, 452.
52. Shoombuatong, W.; Mekha, P.; Chaijaruwanich, J. Sequence based human leukocyte antigen gene prediction using informative physicochemical properties. *Int. J. Data Min. Bioinform.* **2015**, *13*, 211–224. [CrossRef]
53. Shoombuatong, W.; Huang, H.-L.; Chaijaruwanich, J.; Charoenkwan, P.; Lee, H.-C.; Ho, S.-Y. Predicting protein crystallization using a simple scoring card method. In Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Singapore, 16–19 April 2013; pp. 23–30.
54. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
55. Anuwongcharoen, N.; Shoombuatong, W.; Tantimongcolwat, T.; Prachayasittikul, V.; Nantasenamat, C. Exploring the chemical space of influenza neuraminidase inhibitors. *PeerJ* **2016**, *4*, e1958. [CrossRef]
56. Prachayasittikul, V.; Worachartcheewan, A.; Shoombuatong, W.; Prachayasittikul, V.; Nantasenamat, C. Classification of P-glycoprotein-interacting compounds using machine learning methods. *EXCLI J.* **2015**, *14*, 958.
57. Shoombuatong, W.; Mekha, P.; Waiyamai, K.; Cheevadhanarak, S.; Chaijaruwanicha, J. Prediction of human leukocyte antigen gene using k-nearest neighbour classifier based on spectrum kernel. *Sci. Asia* **2013**, *39*, 42–49. [CrossRef]
58. Shoombuatong, W.; Prachayasittikul, V.; Anuwongcharoen, N.; Songtawee, N.; Monnor, T.; Prachayasittikul, S.; Prachayasittikul, V.; Nantasenamat, C. Navigating the chemical space of dipeptidyl peptidase-4 inhibitors. *Drug Des. Dev. Ther.* **2015**, *9*, 4515.
59. Shoombuatong, W.; Prathipati, P.; Prachayasittikul, V.; Schaduangrat, N.; Ahmad Malik, A.; Pratiwi, R.; Wanwimolruk, S.; ES Wikberg, J.; Paul Gleeson, M.; Spjuth, O. Towards predicting the cytochrome P450 modulation: From QSAR to proteochemometric modeling. *Curr. Drug Metab.* **2017**, *18*, 540–555. [CrossRef]
60. Worachartcheewan, A.; Prachayasittikul, V.; Anuwongcharoen, N.; Shoombuatong, W.; Prachayasittikul, V.; Nantasenamat, C. On the origins of hepatitis C virus NS5B polymerase inhibitory activity using machine learning approaches. *Curr. Top. Med. Chem.* **2015**, *15*, 1814–1826. [CrossRef]
61. Worachartcheewan, A.; Shoombuatong, W.; Pidetcha, P.; Nopnithipat, W.; Prachayasittikul, V.; Nantasenamat, C. Predicting metabolic syndrome using the random forest method. *Sci. World J.* **2015**, *2015*. [CrossRef]
62. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef]
63. Shoombuatong, W.; Nabu, S.; Simeon, S.; Prachayasittikul, V.; Lapins, M.; Wikberg, J.E.; Nantasenamat, C. Extending proteochemometric modeling for unraveling the sorption behavior of compound–soil interaction. *Chemom. Intell. Lab. Syst.* **2016**, *151*, 219–227. [CrossRef]
64. Shoombuatong, W.; Prathipati, P.; Owasirikul, W.; Worachartcheewan, A.; Simeon, S.; Anuwongcharoen, N.; Wikberg, J.E.; Nantasenamat, C. Towards the revival of interpretable QSAR models. In *Advances in QSAR Modeling*; Springer: Cham, Switzerland, 2017; pp. 3–55.

65. Liu, K.; Chen, W.; Lin, H. XG-PseU: An eXtreme Gradient Boosting based method for identifying pseudouridine sites. *Mol. Genet. Genom.* **2019**. [CrossRef]
66. Veltri, D.; Kamath, U.; Shehu, A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics* **2018**, *34*, 2740–2747. [CrossRef]
67. Awais, M.; Hussain, W.; Khan, Y.D.; Rasool, N.; Khan, S.A.; Chou, K.-C. iPhosH-PseAAC: Identify phosphohistidine sites in proteins by blending statistical moments and position relative features according to the Chou's 5-step rule and general pseudo amino acid composition. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2019**. [CrossRef]
68. Hussain, W.; Khan, Y.D.; Rasool, N.; Khan, S.A.; Chou, K.-C. SPalmitoylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins. *Anal. Biochem.* **2019**, *568*, 14–23. [CrossRef]
69. Hussain, W.; Khan, Y.D.; Rasool, N.; Khan, S.A.; Chou, K.-C. SPrenylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins. *J. Theor. Biol.* **2019**, *468*, 1–11. [CrossRef]
70. Ju, Z.; Wang, S.-Y. Prediction of lysine formylation sites using the composition of k-spaced amino acid pairs via Chou's 5-steps rule and general pseudo components. *Genomics* **2019**. [CrossRef]
71. Le, N.Q.K.; Yapp, E.K.Y.; Ho, Q.-T.; Nagasundaram, N.; Ou, Y.-Y.; Yeh, H.-Y. iEnhancer-5Step: Identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding. *Anal. Biochem.* **2019**, *571*, 53–61. [CrossRef]
72. Le, N.Q.K.; Yapp, E.K.Y.; Ou, Y.-Y.; Yeh, H.-Y. iMotor-CNN: Identifying molecular functions of cytoskeleton motor proteins using 2D convolutional neural network via Chou's 5-step rule. *Anal. Biochem.* **2019**, *575*, 17–26. [CrossRef]
73. Boopathi, V.; Subramaniyam, S.; Malik, A.; Lee, G.; Manavalan, B.; Yang, D.-C. mACPpred: A support vector machine-based meta-predictor for identification of anticancer peptides. *Int. J. Mol. Sci.* **2019**, *20*, 1964. [CrossRef]
74. Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. mAHTPred: A sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* **2018**, *35*, 2757–2765. [CrossRef]
75. Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. Meta-4mCpred: A Sequence-Based Meta-Predictor for Accurate DNA 4mC Site Prediction Using Effective Feature Representation. *Mol. Ther. Nucleic Acids* **2019**, *16*, 733–744. [CrossRef]
76. Manavalan, B.; Govindaraj, R.G.; Shin, T.H.; Kim, M.O.; Lee, G. iBCE-EL: A new ensemble learning framework for improved linear B-cell epitope prediction. *Front. Immunol.* **2018**, *9*, 1695. [CrossRef]
77. Manavalan, B.; Shin, T.H.; Kim, M.O.; Lee, G. PIP-EL: A new ensemble learning method for improved proinflammatory peptide predictions. *Front. Immunol.* **2018**, *9*, 1783. [CrossRef]
78. Shoombuatong, W.; Hongjaisee, S.; Barin, F.; Chaijaruwanich, J.; Samleerat, T. HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees. *Comput. Biol. Med.* **2012**, *42*, 885–889. [CrossRef]
79. Srinivasulu, Y.S.; Wang, J.-R.; Hsu, K.-T.; Tsai, M.-J.; Charoenkwan, P.; Huang, W.-L.; Huang, H.-L.; Ho, S.-Y. Characterizing informative sequence descriptors and predicting binding affinities of heterodimeric protein complexes. *BMC Bioinform.* **2015**, *16*, S14. [CrossRef]
80. Schaduangrat, N.; Nantasenamat, C.; Prachayasittikul, V.; Shoombuatong, W. Meta-iAVP: A sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation. *Int. J. Mol. Sci.* **2019**, *20*, 5743. [CrossRef]
81. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media: Cham, Switzerland, 2013.
82. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
83. Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **1999**, *10*, 988–999. [CrossRef]
84. Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. kernlab-an S4 package for kernel methods in R. *J. Stat. Softw.* **2004**, *11*, 1–20. [CrossRef]
85. Chou, K.-C.; Shen, H.-B. Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* **2008**, *3*, 153. [CrossRef] [PubMed]
86. Chou, K.-C.; Shen, H.-B. Recent progress in protein subcellular location prediction. *Anal. Biochem.* **2007**, *370*, 1. [CrossRef] [PubMed]

87. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.-C.; Müller, M. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **2011**, *12*, 77. [CrossRef] [PubMed]
88. Ho, S.-Y.; Chen, J.-H.; Huang, M.-H. Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications. *IEEE Trans. Syst. ManCybern. Part B (Cybern.)* **2004**, *34*, 609–620. [CrossRef] [PubMed]
89. Hall, M.A. *Correlation-Based Feature Selection for Machine Learning*; University of Waikato: Hamilton, New Zealand, 1999.
90. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
91. Hasan, M.M.; Yang, S.; Zhou, Y.; Mollah, M.N.H. SuccinSite: A computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. *Mol. Biosyst.* **2016**, *12*, 786–795. [CrossRef]
92. Breiman, L. *Classification and Regression Trees*; Routledge: London, UK, 2017.
93. Laengsri, V.; Shoombuatong, W.; Adirojananon, W.; Nantasenamart, C.; Prachayasittikul, V.; Nuchnoi, P. ThalPred: A web-based prediction tool for discriminating thalassemia trait and iron deficiency anemia. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 212.
94. Cheng, X.; Zhao, S.-G.; Lin, W.-Z.; Xiao, X.; Chou, K.-C. pLoc-mAnimal: Predict subcellular localization of animal proteins with both single and multiple sites. *Bioinformatics* **2017**, *33*, 3524–3531. [CrossRef]
95. Cheng, X.; Xiao, X.; Chou, K.-C. pLoc-mPlant: Predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC. *Mol. Biosyst.* **2017**, *13*, 1722–1727. [CrossRef]
96. Cheng, X.; Xiao, X.; Chou, K.-C. pLoc-mVirus: Predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC. *Gene* **2017**, *628*, 315–321. [CrossRef]
97. Liu, B.; Wang, S.; Long, R.; Chou, K.-C. iRSpot-EL: Identify recombination spots with an ensemble learning approach. *Bioinformatics* **2016**, *33*, 35–41. [CrossRef] [PubMed]
98. Cheng, X.; Zhao, S.-G.; Xiao, X.; Chou, K.-C. iATC-mISF: A multi-label classifier for predicting the classes of anatomical therapeutic chemicals. *Bioinformatics* **2016**, *33*, 341–346. [CrossRef] [PubMed]
99. Zhang, C.-J.; Tang, H.; Li, W.-C.; Lin, H.; Chen, W.; Chou, K.-C. iOri-Human: Identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget* **2016**, *7*, 69783. [CrossRef] [PubMed]
100. Qiu, W.-R.; Jiang, S.-Y.; Xu, Z.-C.; Xiao, X.; Chou, K.-C. iRNAm5C-PseDNC: Identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. *Oncotarget* **2017**, *8*, 41178. [CrossRef]
101. Cheng, X.; Zhao, S.-G.; Xiao, X.; Chou, K.-C. iATC-mHyb: A hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals. *Oncotarget* **2017**, *8*, 58494. [CrossRef]
102. Chen, W.; Feng, P.; Yang, H.; Ding, H.; Lin, H.; Chou, K.-C. iRNA-AI: Identifying the adenosine to inosine editing sites in RNA sequences. *Oncotarget* **2017**, *8*, 4208. [CrossRef]
103. Liu, B.; Fang, L.; Long, R.; Lan, X.; Chou, K.-C. iEnhancer-2L: A two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics* **2015**, *32*, 362–369. [CrossRef]
104. Liu, B.; Fang, L.; Liu, F.; Wang, X.; Chen, J.; Chou, K.-C. Identification of real microRNA precursors with a pseudo structure status composition approach. *PLoS ONE* **2015**, *10*, e0121501. [CrossRef]
105. Feng, P.; Ding, H.; Yang, H.; Chen, W.; Lin, H.; Chou, K.-C. iRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. *Mol. Ther. Nucleic Acids* **2017**, *7*, 155–163. [CrossRef]
106. Chen, W.; Tang, H.; Ye, J.; Lin, H.; Chou, K.-C. iRNA-PseU: Identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids* **2016**, *5*, e332.
107. Chou, K.-C. Progresses in predicting post-translational modification. *Int. J. Pept. Res. Ther.* **2019**. [CrossRef]
108. Qiu, W.-R.; Sun, B.-Q.; Xiao, X.; Xu, Z.-C.; Jia, J.-H.; Chou, K.-C. iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. *Genomics* **2018**, *110*, 239–246. [CrossRef]
109. Khan, Y.D.; Rasool, N.; Hussain, W.; Khan, S.A.; Chou, K.-C. iPhosY-PseAAC: Identify phosphotyrosine sites by incorporating sequence statistical moments into PseAAC. *Mol. Biol. Rep.* **2018**, *45*, 2501–2509. [CrossRef] [PubMed]

110. Liu, B.; Liu, F.; Wang, X.; Chen, J.; Fang, L.; Chou, K.-C. Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* **2015**, *43*, W65–W71. [CrossRef] [PubMed]
111. Chou, K.-C.; Shen, H.-B. MemType-2L: A web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun.* **2007**, *360*, 339–345. [CrossRef] [PubMed]
112. Shen, H.-B.; Chou, K.-C. PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.* **2008**, *373*, 386–388. [CrossRef] [PubMed]
113. Chou, K.-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2004**, *21*, 10–19. [CrossRef] [PubMed]
114. Chou, K.-C. Impacts of bioinformatics to medicinal chemistry. *Med. Chem.* **2015**, *11*, 218–234. [CrossRef]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

Analytical Biochemistry 599 (2020) 113747

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

iTTCA-Hybrid: Improved and robust identification of tumor T cell antigens by utilizing hybrid feature representation

Phasit Charoenkwan[a], Chanin Nantasenamat[b], Md Mehedi Hasan[c], Watshara Shoombuatong[b,*]

[a] Modern Management and Information Technology, College of Arts, Media and Technology, Chiang Mai University, Chiang Mai, 50200, Thailand
[b] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok, 10700, Thailand
[c] Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka, 820-8502, Japan

ARTICLE INFO

Keywords:
Cancer immunotherapy
T-cell antigen
T-cell epitope
Random forest
Support vector machine
Machine learning

ABSTRACT

In spite of the repertoire of existing cancer therapies, the ongoing recurrence and new cases of cancer poses a challenging health concern that prompts for novel and effective treatment. Cancer immunotherapy represents a promising venue for treatment by harnessing the body's immune system to combat cancer. Therefore, the identification of tumor T cell antigen represents an exciting area to explore. Computational tools have been instrumental in the identification of tumor T cell antigens and it is highly desirable to attain highly accurate models in a timely fashion from large volumes of peptides generated in the post-genomic era. In this study, we present a reliable, accurate, unbiased and automated sequence-based predictor named iTTCA-Hybrid for identifying tumor T cell antigens. The iTTCA-Hybrid approach proposed herein employs two robust machine learning models (e.g. support vector machine and random forest) constructed using five feature encoding strategies (i.e. amino acid composition, dipeptide composition, pseudo amino acid composition, distribution of amino acid properties in sequences and physicochemical properties derived from the AAindex). Rigorous independent test indicated that the iTTCA-Hybrid approach achieved an accuracy and area under the curve of 73.60% and 0.783, respectively, which corresponds to 4% and 7% performance increase than those of existing methods thereby indicating the superiority of the proposed model. To the best of our knowledge, the iTTCA-Hybrid is the first free web server (Available at http://camt.pythonanywhere.com/iTTCA-Hybrid) for identifying tumor T cell antigens presented by the MHC class I. The proposed web server allows robust predictions to be made without the need to develop in-house prediction models.

1. Introduction

Cancer poses a major global health concern as it represents the leading causes of death by accounting for almost 10 million deaths in 2018 [1]. The dramatic increase in cancer prevalence over the past decades has called for intense efforts in prevention and treatment. Novel therapeutic agents for combating cancer is much sought after owing to the inherently high cancer recurrence despite the availability of an arsenal of therapeutic regimens for cancer treatment [2]. Small-molecule drugs against cancer have demonstrated potent therapeutic properties but the undesirable cytotoxic effects on healthy cells still pose a major impediment that calls for the search for potent yet safe therapeutic agents [3]. Immune therapy represents a promising therapeutic approach as it primes the body's natural defenses against cancer. Unlike deleterious side effects from chemo- and radiation therapies that are known to give rise to drug resistance, immune therapy has been shown to afford high selectivity and efficacy while causing lowered side effects [4]. As such, immunotherapy provides a new avenue of opportunities for the development of potent cancer treatment.

Cancer immunotherapy exert its activity via innate or adaptive immunity. The innate immunity serves as the first-line of defense and elicits non-specific response, for instance, via host defense peptides that target membranes of tumor cells owing to their inherent amphipathic properties [5,6]. On the other hand, the adaptive immunity acts via specific response mediated by B and T cells. In the case of T cells, T cell receptors (TCR) found at the surface encounters tumor antigens that are presented by the major histocompatibility complex (MHC) class I and II molecules found at the surface of antigen presenting cells (and hence its name) [7]. Tumors are immunogenic as they inherently harbor antigenic determinants (known as epitopes) as derived from peptides and proteins. A formidable challenge in the development of effective cancer

* Corresponding author.
E-mail address: watshara.sho@mahidol.ac.th (W. Shoombuatong).

https://doi.org/10.1016/j.ab.2020.113747
Received 16 March 2020; Received in revised form 13 April 2020; Accepted 16 April 2020
Available online 22 April 2020
0003-2697/ © 2020 Elsevier Inc. All rights reserved.

vaccines is the difficulty of epitopes that are capable of inducing a strong and selective immune response. Definitive evidence suggests that peptides derived from tumor-associated antigens can induce a peptide-specific T-cell immune response [8]. Furthermore, cancer neoantigens that are not present in normal cells represents a promising therapeutic opportunity as they represent specific targets and inherently leads to lower risk of autoimmunity and immune tolerance [9–11]. Taken together, the identification of T-cell epitopes in tumor antigens represent a lucrative first step towards the discovery and development of cancer immunotherapeutic peptides.

Although, experimental approaches are known as the most reliable way of characterizing the biological activities of T-cell epitopes in tumor antigens, it is time-consuming and costly. The availability of large volumes of peptides in several public databases, provides ample opportunity for the development of fast and intelligent computational models for identifying T-cell epitopes in tumor antigens, which is urgently needed for the development of cancer vaccines. Over the years, researchers have developed bioinformatics and immunoinformatics tool for facilitating peptide vaccine discovery [6,12–33]. Since epitopes play a crucial role for both clinical and basic science research, their robust characterization and analysis affords huge potential for vaccine design, disease prevention, diagnosis and treatment [34]. Thus far, there are four major approaches that have been employed for predicting epitopes namely sequence-based methods [35–37], structure-based methods [38], hybrid methods and consensus methods. For example, in 2004, Manoj Bhasin and Raghava [35] developed a sequence-based method known as the CTLPRED for predicting cytotoxic T lymphocyte (CTL) epitopes, which represent potential candidates in subunit vaccine design for various diseases by making use of antigenic sequence. The CTLPRED was developed using support vector machine (SVM) and artificial neural network (ANN) as well as quantitative matrix (QM). CTLPRED was performed on a non-redundant dataset of T cell epitopes and non-epitopes containing 1137 experimentally proven MHC class I restricted T cell epitopes. Prediction accuracies for QM-, ANN- and SVM-based methods were 70.0, 72.2 and 75.2%, respectively, as assessed by a leave-one-out cross-validation (LOOCV) test set. Finally, the machine learning methods were used for consensus and combined prediction of CTL epitopes. Finally, an online prediction server CTLPRED was implemented and freely accessible at http://www.imtech.res.in/raghava/ctlpred/. As mentioned above, the first important step for the discovery and development of cancer immunotherapeutic peptides is to identify T-cell epitopes in tumor antigens.

To the best of our knowledge, there is only one study focusing on the prediction of tumor T cell antigens represented in the MHC class I context, which is the TTAgP1.0 [39]. This method was developed via the use of the random forest (RF) model that makes use of two peptide feature types namely the relative frequency (Rfre) and amino acid composition (AAC) of peptides (PEP). Although, TTAgP1.0 has its own merit and reasonable prediction accuracies, there are four major limitations that needs to be addressed so as to improve the utilization of computational tools [39]. Firstly, Lissabet et al. [39] did not provide their dataset used. However, several researchers have mentioned the benefits of data sharing so as to facilitate reproducible research. Secondly, TTAgP1.0 was developed via the use of only two types of peptide features which might not be able to fully capture the significant information of tumor T cell antigens. Thirdly, TTAgP1.0 was implemented and performed on an imbalanced dataset (442 positive and 295 negative samples) that likewise contributes to classification errors and bias. Fourthly, no web server was provided from the study. Hence, its utility and usage is quite limited for the broad scientific community.

To address these aforementioned issues, the present study describes the development of a robust, unbiased and automated sequence-based predictor referred herein as the iTTCA-Hybrid for identifying tumor T cell antigens. Fig. 1 shows the workflow of the iTTCA-Hybrid for identifying tumor T cell antigens as represented in the MHC class I context. Since, previous studies did not publicly share the dataset used in the studies, therefore this study establishes the first publicly available benchmark dataset consisting of 529 tumor and 393 non-tumor T cell antigen so as to ensure the reproducibility of the proposed model. Secondly, this study employs two robust machine learning models (e.g. RF and SVM) in conjunction with five feature encoding strategies. Thirdly, to cope with the class imbalance problem and to improve the predictive power of the proposed model, extracted features were hybridized and the synthetic minority over-sampling technique (SMOTE) method was utilized to overcome possible bias arising from the class imbalance problem. On the basis of these comparative results, iTTCA-Hybrid can achieve significantly improved performance than those of existing methods thereby indicating the effectiveness and robustness of the proposed method. Finally, for the convenience of experimental biologists, the iTTCA-Hybrid web server was established and made freely available online at http://camt.pythonanywhere.com/iTTCA-Hybrid.

2. Materials and methods

2.1. Benchmark datasets

In practice, same training and independent sets should be employed for the development and assessment of the proposed model so as to provide a comprehensive and unbiased comparison [13,27–29,40,41]. However, previous studies [39] did not publicly share the dataset used in the study. In order to make a fair comparison, we generated an up-to-date benchmark dataset that contains the largest number of samples following the procedures described in the previous study [39] as follows. (i) Tumor T cell antigens were obtained from the TANTIGEN [42] and TANTIGEN 2.0 [43] whereby a total 727 MHC class I peptides was collected and considered as positive samples, (ii) Non-tumor T cell antigens were obtained from well-known IEDB database (https://www.iedb.org/). In order to construct the dataset of non-tumor T cell antigens, only T cell antigens that are reported to have no association to any disease were collected and considered as negative samples (containing 367 samples). (iii) Duplicate peptide sequences were removed. After performing such pre-processing, all peptides were unique and constituted the benchmark dataset consisting of 529 positive and 393 negative samples.

In order to examine the efficacy of the proposed model, an independent dataset (named TTCA-IND) was constructed by randomly selecting 20% from the benchmark dataset while treating the remaining samples as the training dataset (named TTCA-CV). Finally, the TTCA-CV dataset consisted of 470 positive and 318 negative samples while the TTCA-IND dataset consisted of 122 positive and 75 negative samples as summarized in Table 1. In order to ensure the reproducibility of the model proposed herein, datasets used in the construction of predictive models are shared publicly and is available on GitHub at https://github.com/Shoombuatong2527/Benchmark-datasets.

2.2. Feature representation

Given a peptide sequence (P) that can be represented as:

$$P = p_1 p_2 p_3 \cdots p_N \quad (1)$$

where p_i and N denote the ith residue in the peptide P and the peptide length, respectively. Note that residue p_i is in the set of natural amino acid, i.e. A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y.

The simplest feature encoding strategies used to represent a peptide sequence are AAC and dipeptide composition (DPC) that are expressed as fixed lengths of 20 and 400, respectively. Thus, a peptide P can be expressed by vectors with 20D and 400D (dimension) spaces for AAC and DPC features, respectively, as formulated by:

$$P = [aa_1, aa_2, \ldots, aa_{20}]^T \quad (2)$$

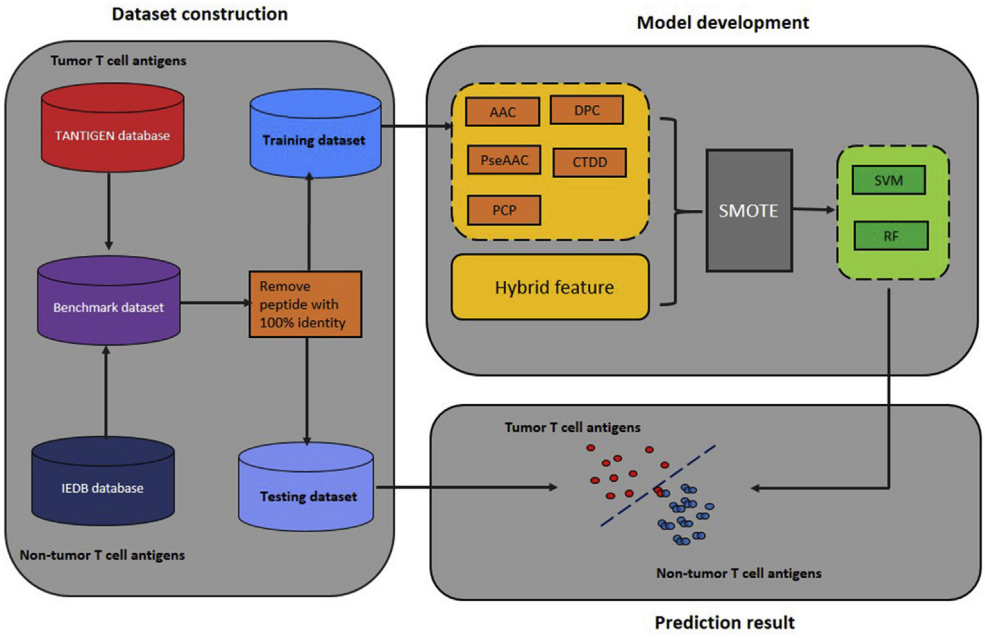

Fig. 1. Schematic framework of iTTCA-Hybrid for predicting tumor T cell antigens.

Table 1
Dataset distribution of peptides used in this study.

Dataset	Original	Non-duplicate peptides	Benchmark dataset	
			Positive	Negative
TTCA-CV	727	592	470	122
TTCA-IND	394	393	318	75

$$P = [dp_1, dp_2, ..., dp_{400}]^T \quad (3)$$

where T is the transposed operator while aa_1, aa_2, ..., aa_{20} and dp_1, dp_2, ..., dp_{400} are the normalized occurrence frequencies of 20 and 400 native amino acids and dipeptides, respectively, in a peptide sequence P.

As for physicochemical (PCP) features, there are a total of 544 PCPs for amino acids as derived from version 9.0 of the Amino acid index database (AAindex) [44]. In this study, PCPs having NA values were excluded. Finally, a set of 531 PCPs (531D) were used to extract peptide sequences for the development of the model. Previously, PCPs are one of the most intuitive features associated with biophysical and biochemical reactions and is also referred to as easy and interpretable features [6,17,18,31,40,45].

To address the sequence-order information, the feature encoding strategy named pseudo amino acid composition (PseAAC) was proposed by Chou [46,47]. According to Chou's PseAAC, the general form of PseAAC for a peptide P is formulated by:

$$P = [\Psi_1, \Psi_2, ..., \Psi_u, ..., \Psi_\Omega]^T \quad (4)$$

where the subscript Ω is an integer to reflect the feature's dimension. The value of Ω and the component of Ψ_u, where u= 1,2, ...,Ω is dependent on the protein or peptide sequences.

Distribution of amino acid properties in sequences (CTDD) are often used to describe the overall composition of the amino acid properties of peptide sequences by using CTD descriptor having three types of descriptors, Composition (C), Transition (T), and Distribution (D) [48]. The CTDD feature is the distribution descriptor that describes the distribution of the first, 25%, 50%, 75%, and 100% amino acids of a particular property within the peptide sequence. Additional details of CTDD feature can be found in Refs. [48]. Previously, several studies have reported that AAC, DPC, PseAAC, PCP and CTDD features are effective features for predicting and analyzing various types of protein and peptide functions [6,13,18,30,33,40,41,49,50]. Thus, in this study, these five different types of sequence features were generated by using the iFeature module in the Python environment [48].

2.3. Synthetic minority over-sampling technique

From the point of view of machine learning, the class imbalanced dataset has a tendency to cause a prediction model to overfit as well as to perform poorly on the minority class. To address this problem, there are two widely used approaches containing undersampling and oversampling methods. The idea of undersampling methods is to create a new subset of the imbalanced dataset by eliminating some of the samples from the majority class, while oversampling methods create a new subset of the imbalanced dataset by creating new ones from the original minority class samples [51]. In this study, we employ the powerful SMOTE algorithm [52] for performing oversampling of the minority class (i.e., negative samples) by taking each minority class sample and introducing synthetic samples as summarized in Table 1.

2.4. Classifier selection

RF models were constructed according to the originally described algorithm [53]. RF is an ensemble-based ML algorithm capable of dealing with binary and multi-class classification problems and has been widely used in various biological problems [6,12,16–18]. To improve the prediction performance of classification and regression tree (CART) [53,54], These models were established by generating a number of CART classifiers. In RF, the out-of-bag (OOB) approach is employed for assessing the feature importance as follows: (i) two-thirds of the training data are used to construct the classifier while the remaining was used for evaluating the performance of the constructed classifier and (ii) the feature importance of each feature can be assessed by measuring the decrease in the prediction performance. It should be noted that the performance evaluation of the model can be either accuracy or Gini index.

SVM is one of the most widely used ML algorithm for dealing with binary classification problem and has been widely used in computational biology [13,27,29–31]. This method is based on Vapnik-

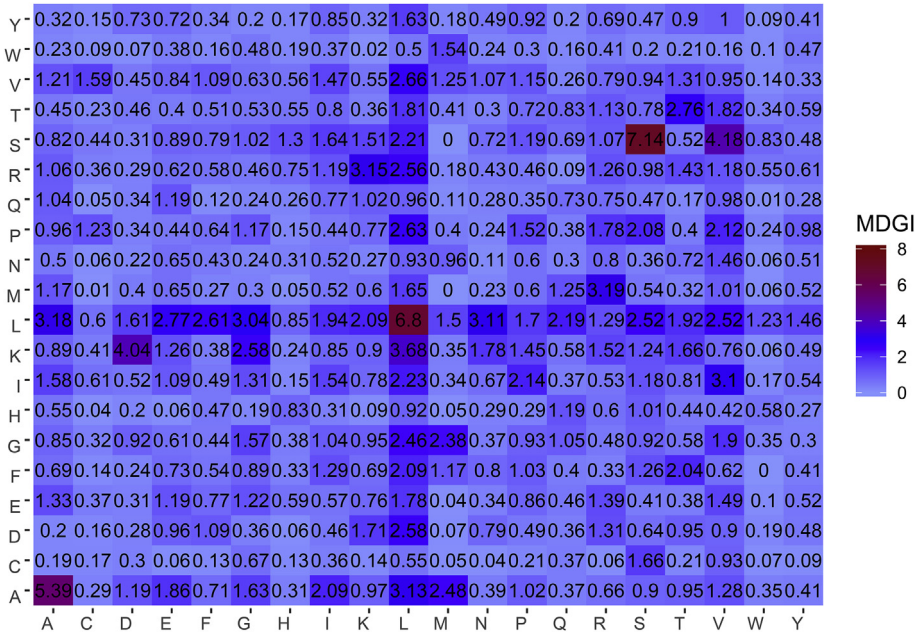

Fig. 2. Heat map of the mean decrease of Gini index (MDGI) of dipeptide compositions. It should be noted that features with the largest value of MDGI are deemed to be the most important.

Chervonenkis's theory of statistical learning [55–57]. This method determines the optimal hyperplane with the largest distance between two classes that minimizes the misclassification rate. To make linear separation on high dimensional samples, SVM employs one of the available well-known kernel functions to transform the sample space having *p*-dimensional feature vector into the feature space with *n*-dimensional feature vector, where $p < n$. In this work, the widely used radial basis function is applied to non-linearly transform the feature space. Herein, the RF and SVM classifiers were applied in the development of predictive models using the scikit-learn package (version 0.22) with default parameters.

2.5. Comparison between iTTCA-Hybrid with TTAgP1.0

To the best of our knowledge, there is only one study focusing on the prediction of tumor T cell antigens represented in the MHC class I context namely the TTAgP1.0 [39]. TTAgP1.0 is constructed using RF and two types of peptide features (i.e., Rfre and AAC). Since, the present study makes use of up-to-date dataset containing 592 tumor and 394 non-tumor T cell antigens, it is therefore not fair to directly compare the results of TTAgP1.0 with that of the proposed iTTCA-Hybrid. In order to make a fair comparison, a modified TTAgP1.0 method (TTAgP1.0-MODI) was established using the RF model with the two aforementioned feature types on our new dataset.

2.6. Performance evaluation

In order to evaluate the prediction ability of the model, we used four widely used metrics for solving the two-class prediction problem as follows:

$$\mathrm{Ac} = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP}+\mathrm{TN} + \mathrm{FP}+\mathrm{FN})} \tag{5}$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{6}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})} \tag{7}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{8}$$

where Ac, Sn, Sp and MCC represents the accuracy, sensitivity, specificity and Matthews coefficient correlation, respectively. Meanwhile, TP is the number of true positives and TN is the number of true negatives, FP is the number of false positives and FP is the number of false negatives. To investigate the prediction performance of the proposed model using threshold-independent parameters, receiver operating characteristic (ROC) curves were plotted. The area under the ROC curve (auAUC) was utilized to assess the prediction performance where AUC values of 0.5 and 1 are indicative of random and perfect models, respectively [13,41,58–62].

3. Results and discussion

3.1. Analysis of amino acids and dipeptides preference

RF model provides two measures for ranking feature importance namely the mean decrease of the Gini index (MDGI) and the mean decrease of prediction accuracy. Since Calle and Urrea [63] demonstrated that the MDGI provided more robust results when compared with that of the mean decrease of prediction accuracy, we therefore use the MDGI to rank the importance of interpretable features that includes AAC and DPC. Features with the largest MDGI value was considered to be an important feature as it significantly contributed to the prediction performance. Supplementary Table S1 lists the percentage values of 20 amino acids for both tumor and non-tumor T cell antigens from the TTCA-CV dataset along with the amino acid composition difference between the two classes as well as their MDGI values. In addition, a heatmap showing the feature importance for DPC is provided in Fig. 2. Features with the largest MDGI value was considered to be an important feature [54]. As seen in Table 1, it can be seen that Glu and Pro are presented in high proportion in the tumor T cell antigens while Ser and Val are present in high proportions in the non-tumor T cell antigens. On the other hand, the top-five informative dipeptides consisted of SS, LL, AA, SV and KD with corresponding MDGI values of 7.138, 6.801, 5.390, 4.181 and 4.044, respectively, as shown in Fig. 2.

Table 2
Comparison of SVM and RF models with various types of features without an oversampling approach (SMOTE) over 10-fold cross-validation.

Feature[a]	Classifier[b]	Ac (%)	Sn (%)	Sp (%)	MCC	AUC
AAC	RF	70.56	88.30	44.35	0.369	0.721
	SVM	63.70	87.66	28.31	0.209	0.618
DPC	RF	66.11	81.28	43.74	0.274	0.686
	SVM	69.55	85.96	45.35	0.348	0.725
PseAAC	RF	66.76	82.55	43.43	0.285	0.689
	SVM	69.28	93.62	33.31	0.344	0.711
CTDD	RF	72.46	87.66	49.99	0.417	0.753
	SVM	70.05	92.55	36.77	0.365	0.689
PCP	RF	62.94	73.62	47.23	0.217	0.631
	SVM	60.78	91.49	15.42	0.120	0.567
DPC + PseAAC	RF	71.06	88.30	45.56	0.386	0.723
	SVM	68.91	87.23	41.85	0.331	0.725
DPC + CTDD	RF	71.82	88.51	47.14	0.400	0.756
	SVM	70.18	96.38	31.45	0.384	0.679
PseAAC + CTDD	RF	73.35	88.51	50.93	0.436	0.752
	SVM	71.32	93.19	38.98	0.399	0.688
DPC + PseAAC + CTDD	RF	72.84	89.36	48.41	0.425	0.759
	SVM	70.30	96.38	31.76	0.386	0.680

[a] AAC: amino acid composition, DPC: dipeptide composition, PseAAC: pseudo amino acid composition, CTDD: distribution of amino acid properties in sequences, PCP: physicochemical properties from AAindex.
[b] RF: random forest, SVM: support vector machine.

3.2. Evaluation of various feature extraction methods without SMOTE

In this section, we carried out comparative experiments via the use of five basic features (i.e., AAC, DPC, PseAAC, CTDD and PCP) and their hybrid features (i.e., DPC + PseAAC, DPC + CTDD, PseAAC + CTDD and DPC + PseAAC + CTDD) to assess their contributions to the prediction of tumor T-cell antigens without the use of the SMOTE algorithm. We hypothesized that using the combinations of various feature types could alleviate the weakness of one another. Each feature was assessed by using two robust ML classifiers namely SVM and RF and evaluated via 10-fold CV and independent test. Although, the jackknife test can provide unique results for any given training dataset but this comes at the cost of its time-consuming nature. Thus, the 10-fold CV and independent tests were used for evaluating the prediction performance of proposed models. Tables 2 and 3 list the results of performance comparisons of the comparative experiments over 10-fold CV and independent test, respectively. Furthermore, Fig. 3A and B plot ROC curves of the three selected optimal models over 10-fold CV and independent test, respectively. Several observations can be made from Tables 2 and 3 as follows.

First, in case of using the basic feature, the RF model making use of the CTDD feature achieved the highest Ac of 72.46% with MCC of 0.417 and AUC of 0.753, while the second highest Ac of 70.56% with MCC of 0.369 and AUC of 0.721 was obtained by RF model built using the AAC feature. The prediction performance of the RF model built with CTDD attained on an independent dataset was 70.56% Ac, 0.355 MCC and 0.739 AUC. It was observed that the CTDD feature could be effectively used for prediction of tumor T-cell antigens with Ac in excess of 70% as assessed by 10-fold CV. Secondly, in the case of using hybrid features, RF models built using hybrid features of PseAAC + CTDD and DPC + PseAAC + CTDD performed well with the first and second highest Ac of 73.35% and 72.74%, respectively. Interestingly, these two hybrid features consistently provided acceptable results with Ac of greater than 70% as evaluated on the independent dataset.

Table 3
Comparison of SVM and RF models with various types of features without an oversampling approach (SMOTE) over independent test.

Feature[a]	Classifier[b]	Ac (%)	Sn (%)	Sp (%)	MCC	AUC
AAC	RF	69.54	81.97	49.33	0.332	0.724
	SVM	65.48	84.43	34.67	0.221	0.693
DPC	RF	59.90	73.77	37.33	0.117	0.658
	SVM	69.54	84.43	45.33	0.326	0.780
PseAAC	RF	67.51	80.33	46.67	0.286	0.721
	SVM	72.59	95.08	36.00	0.404	0.756
CTDD	RF	70.56	82.79	50.67	0.355	0.739
	SVM	75.63	93.44	46.67	0.471	0.786
PCP	RF	63.96	72.95	49.33	0.226	0.637
	SVM	58.88	86.07	14.67	0.010	0.524
DPC + PseAAC	RF	68.02	83.61	42.67	0.289	0.713
	SVM	72.59	86.89	49.33	0.397	0.782
DPC + CTDD	RF	70.05	83.61	48.00	0.340	0.747
	SVM	76.65	97.54	42.67	0.511	0.792
PseAAC + CTDD	RF	71.07	84.43	49.33	0.363	0.756
	SVM	75.13	93.44	45.33	0.460	0.785
DPC + PseAAC + CTDD	RF	72.08	86.07	49.33	0.385	0.757
	SVM	76.65	97.54	42.67	0.511	0.792

[a] AAC: amino acid composition, DPC: dipeptide composition, PseAAC: pseudo amino acid composition, CTDD: distribution of amino acid properties in sequences, PCP: physicochemical properties from AAindex.
[b] RF: random forest, SVM: support vector machine.

3.3. Evaluation of various features extraction methods with SMOTE

In this section, we attempt to demonstrate the utility of the oversampling approach (SMOTE) for enhancing the prediction performance by treating the data imbalance problem. In order to evaluate the impact of the oversampling approach, the same comparative experiments using five basic features and their hybrid features were carried out. Tables 3 and 4 lists the results from performance comparisons as evaluated via 10-fold CV and independent test, respectively. Furthermore, Fig. 3C and D shows the plot of ROC curves for the three selected optimal models as assessed by 10-fold CV and independent test, respectively. Several observations can be made as follows.

By comparing Tables 4 and 5, it could be clearly noticed that prediction results of classifiers were improved by using the oversampling approach. Firstly, in case of using the basic feature, the highest Ac values of 77.55%, 76.70 and 76.17% over 10-fold CV were achieved by using RF with AAC, RF with CTDD and SVM with DPC features, respectively. Secondly, in case of utilizing the hybrid features, RF models built in conjunction with hybrid features of DPC + CTDD and PseAAC + CTDD provided Ac of 78.83% and 77.45%, respectively. Remarkably, using the hybrid feature of PseAAC + CTDD can efficaciously predict tumor T-cell antigens with an Ac of 73.35%, MCC of 0.436 and AUC of 0.752 on the independent dataset. Based on the observation that can be deduced from Tables 2–5, the optimal prediction performance as evaluated by 10-fold CV and independent dataset were achieved by using RF as the predictor together with the use of the hybrid feature of PseAAC + CTDD as the input feature on a balanced dataset derived from the SMOTE method. For convenience of the subsequent description, we will refer to this method as iTTCA-Hybrid.

3.4. Reliability of iTTCA-Hybrid

To avoid the bias of single random sampling, the dataset was subjected to 20 iterations of randomly generated training and independent datasets [12,17,18,49,50]. Thus, 20 prediction models were built and assessed via 10-fold CV and independent tests. Prediction results are provided in Supplementary Table S2. It can be seen from Supplementary Table S2 that iTTCA-Hybrid yields accuracy and MCC of 78.73 ± 1.18 and 0.851 ± 0.012, respectively, as assessed by 10-fold CV. Meanwhile, independent test results of the predictor presented herein afforded accuracy and MCC of 73.65 ± 1.14 and 0.437 ± 0.032, respectively. Prediction results indicated that the proposed iTTCA-Hybrid is accurate, reliable and stable for the identification of tumor T cell antigens.

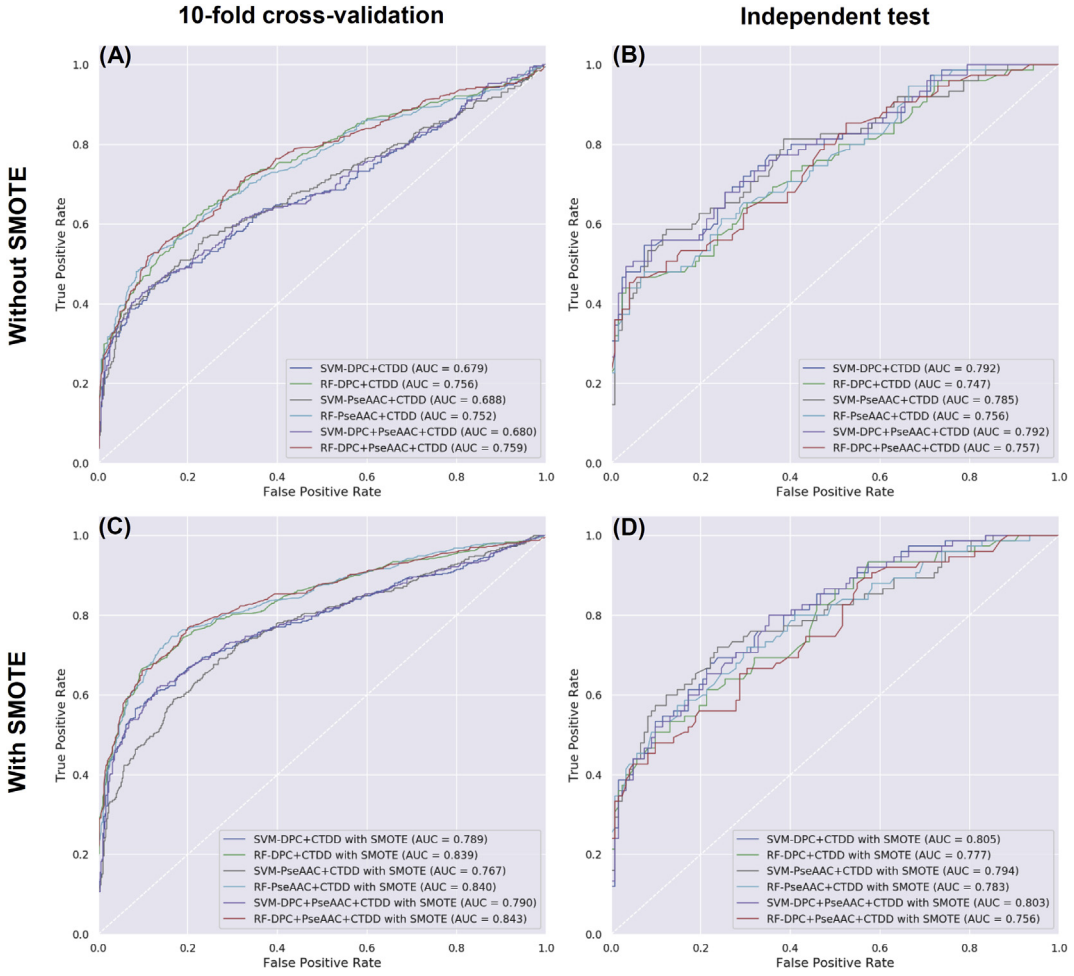

Fig. 3. The ROC curves of RF and SVM models in conjunction with various hybrid features over 10-fold cross-validation (A and C) and independent test (B and D), and with (C and D) and without (A and B) SMOTE method.

Table 4
Comparison of SVM and RF models with various types of features with an oversampling approach (SMOTE) over 10-fold cross-validation.

Feature[a]	Classifier[b]	Ac (%)	Sn (%)	Sp (%)	MCC	AUC
AAC	RF	77.55	86.60	68.51	0.573	0.839
	SVM	70.64	71.91	69.36	0.423	0.765
DPC	RF	73.30	79.36	67.23	0.490	0.830
	SVM	76.17	82.98	69.36	0.544	0.851
PseAAC	RF	75.53	81.28	69.79	0.526	0.824
	SVM	69.89	84.89	54.89	0.417	0.750
CTDD	RF	76.70	84.89	68.51	0.546	0.834
	SVM	70.53	78.94	62.13	0.418	0.768
PCP	RF	67.77	70.00	65.53	0.362	0.733
	SVM	54.89	67.66	42.13	0.103	0.583
DPC + PseAAC	RF	75.85	85.11	66.60	0.537	0.835
	SVM	76.38	84.26	68.51	0.550	0.849
DPC + CTDD	RF	77.45	86.81	68.09	0.566	0.839
	SVM	72.87	83.40	62.34	0.471	0.789
PseAAC + CTDD	RF	78.83	85.53	72.13	0.588	0.840
	SVM	70.43	79.36	61.49	0.417	0.767
DPC + PseAAC + CTDD	RF	77.13	85.11	69.15	0.555	0.843
	SVM	73.62	84.04	63.19	0.486	0.790

[a] AAC: amino acid composition, DPC: dipeptide composition, PseAAC: pseudo amino acid composition, CTDD: distribution of amino acid properties in sequences, PCP: physicochemical properties from AAindex.
[b] RF: random forest, SVM: support vector machine.

Table 5
Comparison of SVM and RF models with various types of features with an oversampling approach (SMOTE) over independent test.

Feature[a]	Classifier[b]	Ac (%)	Sn (%)	Sp (%)	MCC	AUC
AAC	RF	67.01	78.69	48.00	0.279	0.716
	SVM	68.53	72.13	62.67	0.343	0.706
DPC	RF	65.48	77.87	45.33	0.244	0.702
	SVM	69.54	81.97	49.33	0.332	0.775
PseAAC	RF	68.02	77.87	52.00	0.307	0.736
	SVM	67.51	78.69	49.33	0.292	0.761
CTDD	RF	68.53	77.05	54.67	0.323	0.746
	SVM	75.63	82.79	64.00	0.476	0.800
PCP	RF	63.96	71.31	52.00	0.234	0.655
	SVM	56.35	66.39	40.00	0.065	0.541
DPC + PseAAC	RF	69.04	80.33	50.67	0.324	0.737
	SVM	69.04	80.33	50.67	0.324	0.775
DPC + CTDD	RF	72.08	82.79	54.67	0.392	0.777
	SVM	74.11	86.07	54.67	0.434	0.805
PseAAC + CTDD	RF	73.60	82.79	58.67	0.428	0.783
	SVM	75.13	83.61	61.33	0.462	0.794
DPC + PseAAC + CTDD	RF	72.08	81.97	56.00	0.394	0.756
	SVM	73.60	84.43	56.00	0.425	0.803

[a] AAC: amino acid composition, DPC: dipeptide composition, PseAAC: pseudo amino acid composition, CTDD: distribution of amino acid properties in sequences, PCP: physicochemical properties from AAindex.
[b] RF: random forest, SVM: support vector machine.

Table 6
Comparison of iTTCA-Hybrid with other related methods over 10-fold cross-validation and independent test.

Cross-validation	Classifier[a]	Ac (%)	Sn (%)	Sp (%)	MCC	AUC
10-fold CV	TTAgP1.0-MODI[a]	70.81	86.17	48.09	0.375	0.730
	iTTCA-Hybrid[a]	73.35	88.51	50.93	0.436	0.752
	TTAgP1.0-MODI[b]	77.34	83.83	70.85	0.565	0.838
	iTTCA-Hybrid[b]	78.83	85.53	72.13	0.588	0.840
Independent test	TTAgP1.0-MODI[a]	69.54	81.97	49.33	0.332	0.737
	iTTCA-Hybrid[a]	71.07	84.43	49.33	0.363	0.756
	TTAgP1.0-MODI[b]	71.07	78.69	58.67	0.379	0.747
	iTTCA-Hybrid[b]	73.60	82.79	58.67	0.428	0.783

[a] Models were performed on unbalanced data.
[b] Models were performed on balanced data using an oversampling approach (SMOTE).

3.5. Comparison of iTTCA-Hybrid with TTAgP1.0

To further assess the predictive efficiency and effectiveness of the proposed iTTCA-Hybrid, the model was subjected to comparison with existing method (TTAgP1.0). Herein, we developed the TTAgP1.0-MODI to fairly compare the proposed model by performing the same experimental setting and cross-validation methods. Table 6 and Fig. 4 show the performance comparison between iTTCA-Hybrid and TTAgP1.0-MODI as evaluated by 10-fold CV and independent test.

In the case of 10-fold CV test, the Ac, MCC and AUC of iTTCA-Hybrid as performed on the imbalanced dataset afforded performance values of 73.35%, 0.436 and 0.752, respectively, while the TTAgP1.0-MODI model yielded 70.81%, 0.375 and 0.732, respectively. As can be noticed in Table 6 and Fig. 4, the iTTCA-Hybrid model that make use of the SMOTE algorithm for oversampling was found to outperform TTAgP1.0-MODI with improvements of 8%, 10% and 9%, respectively, for Ac, MCC and AUC. As for the independent test (Table 6), similar results to those obtained from the cross-validation test was observed. Particularly, iTTCA-Hybrid was found to achieve significantly improved performance than that of TTAgP1.0-MODI as assessed by the five aforementioned statistical parameters. Interestingly, iTTCA-Hybrid was found to afford higher performance than that of TTAgP1.0-MODI in which Ac, Sp, MCC and AUC demonstrated improvements of 4%, 9%, 9% and 6%, respectively. Taken all together, comparative results demonstrated that the predictor proposed herein was able to achieve significantly improved as compared to those of existing method, TTAgP1.0-MODI.

3.6. iTTCA-hybrid web server

In order to facilitate rapid and easy use of the predictive model presented herein as well as for the benefit of the scientific community, we herein developed a user-friendly web server known as the iTTCA-Hybrid, which is made freely available online at http://camt.pythonanywhere.com/iTTCA-Hybrid. Step-by-step guidelines on how to use the iTTCA-Hybrid web server in order to obtain prediction results are as follows. Firstly, the user accesses the web server at http://camt.pythonanywhere.com/iTTCA-Hybrid where the user will see the top screen of iTTCA-Hybrid as shown in Fig. 5. Secondly, the user enters the query sequence into the text box or uploads a FASTA file by clicking on the “Choose file” button. Thirdly, the user clicks on the “Submit” button in order to start the prediction process. Typically, in this step, it takes a few seconds for the server to process the task. Finally, after finishing the prediction process, results are displayed as shown on the right-hand side of the web server. The user can see examples of FASTA-formatted sequences by clicking on the “example file” button.

4. Discussion

The aim of the present study is to develop an accurate, unbiased and automated sequence-based tumor T cell antigen predictor named herein as the iTTCA-Hybrid. To the best of our knowledge, there is only one study in existence that focuses on the prediction of tumor T cell antigens represented in the MHC class I context namely the TTAgP1.0 [39]. As mentioned above, our comparative analyses revealed that the method proposed herein have been found to afford improved performance and reliability. Herein, we further discuss about the prediction as summarized in the following paragraphs.

The first point is the objectivity of the benchmark dataset. So as to afford a fair comparison, the same benchmark dataset should be used to develop and assess the proposed model in order to provide a comprehensive and unbiased comparison [64–68]. However, previous studies did not publicly share the dataset used for the model development [39]. To address this issue, we compiled a benchmark dataset containing 529 experimentally verified tumor T cell antigens from the TANTIGEN [42] and TANTIGEN 2.0 [43] databases as well as 393 non-tumor T cell antigens collected from IEDB database (https://www.iedb.org/).

The second point pertains to the significance of using various feature types for covering various aspects of sequence information [20,22,23,28]. TTAgP1.0 was developed by making use of only two types of peptide features namely the Rfre and PEP. In the meantime, the proposed iTTCA-Hybrid was built using five basic features and their hybrid features. Table 3 shows that the highest accuracy of 73.35% was achieved from the hybrid feature of PseAAC + CTDD, while TTAgP1.0

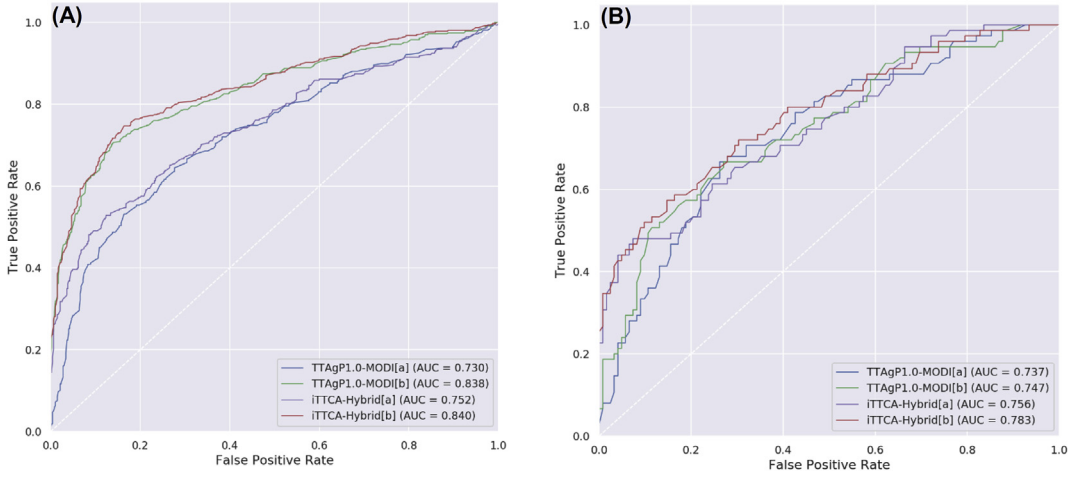

Fig. 4. The ROC curves of iTTCA-Hybrid with TTAgP1.0 over 10-fold cross-validation (A) and independent test (B).

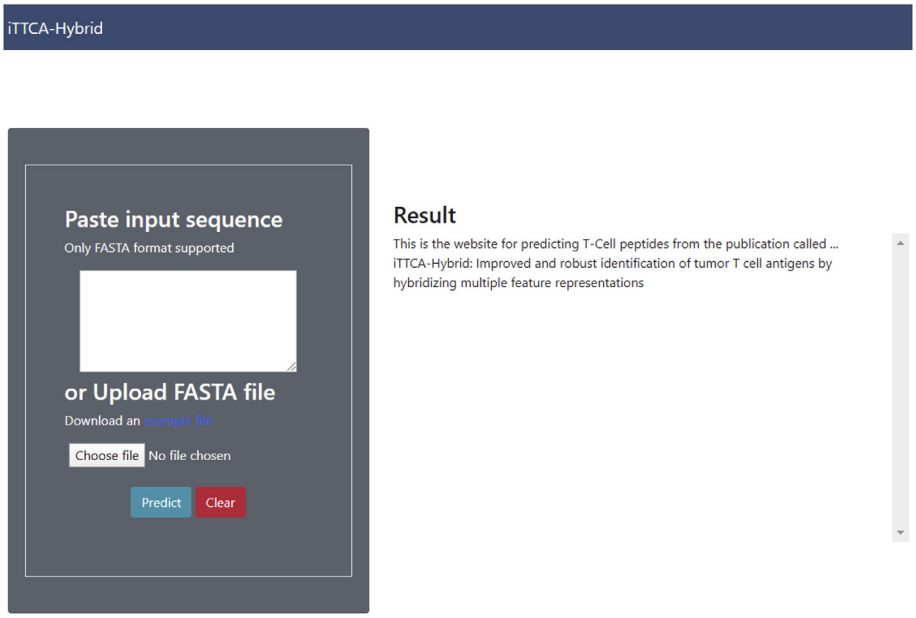

Fig. 5. Screenshots of the iTTCA-Hybrid web server.

afforded lower accuracy of 70.81%. Results indicated that the hybrid feature described herein was more beneficial for identifying tumor T cell antigen than that of Rfre and PEP.

The third point concerns the problem of bias arising from the class imbalance problem. Since, our benchmark dataset was imbalanced where the ratio of non-tumor to tumor T cell antigens was 1 : 1.35. From the machine learning point of view, the class imbalance dataset has the tendency to cause erroneous prediction model that may potentially be overfitting as well as perform poorly on the minority class [6,45]. To overcome this mentioned problem, the SMOTE method was utilized to cope with the potential bias arising from the class imbalance problem [6,45,50–52]. Based on our comparative results, we observed that iTTCA-Hybrid was superior to that of TTAgP1 as evaluated by Ac, Sp, MCC and AUC in which improvements of 4%, 9%, 9% and 6%, respectively, were observed when compared to that of TTAgP1.0. These results indicated that our proposed model that makes use of the SMOTE method were more efficient and reliable in this aspect.

The fourth point deals with the importance of the deploying the predictive model as a publicly web server. The benefit of hosting the model as a web server are many namely facilitating the access of the predictive capabilities of the model by experimental biologists without the need to develop in-house models [19,21,30,69–72]. No web server was provided by previous studies, which is therefore addressed in the present study whereby the iTTCA-Hybrid web server was established and made freely available online at http://camt.pythonanywhere.com/iTTCA-Hybrid. This implemented web server allows fast and easy robust predictions to be made without the need to develop in-house prediction models.

5. Conclusions

In light of the promise of cancer immunotherapy, we have developed a reliable, accurate, unbiased and automated sequence-based tumor T cell antigen predictor named iTTCA-Hybrid with high prediction performance. This is implemented by effectively employing multiple peptide features together with the use of the robustness of random forest. Benchmarking results affirmed that the proposed iTTCA-Hybrid produced significant performance improvements as compared to those of existing methods. Finally, for the convenience of experimental biologists, the iTTCA-Hybrid model was established as freely available web server that can be accessed at http://camt.pythonanywhere.com/iTTCA-Hybrid. It is highly anticipated that the proposed iTTCA-Hybrid represents a useful and robust tool for rational screening and design of cancer vaccines. Although, the predictor presented herein afforded the best performance than those of existing methods, there is still ample room for further improvements. For instance, ongoing work in our lab is geared towards improving the usefulness and applicability for field of cancer immunotherapy via the use of the interpretable scoring card method (SCM) [31,40,41] for shedding insights on the biophysical and biochemical properties of tumor T cell antigens.

Author contributions

W.S. conceived and analyzed the experiments. W.S. and P.C. designed and performed the experiments. P.C. and W.S analyzed the data. P.C contributed the code for constructing iTTCA-Hybrid model and the web server. W.S., C.N., M.H. and P.C. drafted the manuscript. All authors read and approved the manuscript.

Declaration of competing interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was fully supported by the TRF Research Grant for New Scholar (No. MRG6180226) and College of Arts, Media and Technology, Chiang Mai University, and partially supported by the TRF Research Career Development Grant (No. RSA6280075) from the Thailand Research Fund, the Office of Higher Education Commission and Mahidol University.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ab.2020.113747.

References

[1] V.O. Carvalho, Left atrial volume and exercise capacity in adult heart transplant recipients, J. Cardiothorac. Surg. 6 (1) (2011) 9.

[2] S. Simard, et al., Fear of cancer recurrence in adult cancer survivors: a systematic review of quantitative studies, J. Canc. Survivorship 7 (3) (2013) 300–322.
[3] H. Perdry, B.S. Maher, M.-C. Babron, T. McHenry, F. Clerget-Darpoux, M.L. Marazita, An ordered subset approach to including covariates in the transmission disequilibrium test, BMC Proc. 1 (1) (2007) S77 BioMed Central.
[4] J. Couzin-Frankel, Cancer Immunotherapy, American Association for the Advancement of Science, 2013.
[5] H. Li, N. Anuwongcharoen, A.A. Malik, V. Prachayasittikul, J.E. Wikberg, C. Nantasenamat, Roles of d-amino acids on the bioactivity of host defense peptides, Int. J. Mol. Sci. 17 (7) (2016) 1023.
[6] N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, W. Shoombuatong, ACPred: a computational tool for the prediction and analysis of anticancer peptides, Molecules 24 (10) (2019) 1973.
[7] H. Li, N. Schaduangrat, S. Simeon, C. Nantasenamat, Computational study on the origin of the cancer immunotherapeutic potential of B and T cell epitope peptides, Mol. Biosyst. 13 (11) (2017) 2310–2322.
[8] E. Mizukoshi, et al., Comparative analysis of various tumor-associated antigen-specific t-cell responses in patients with hepatocellular carcinoma, Hepatology 53 (4) (2011) 1206–1216.
[9] T.N. Schumacher, R.D. Schreiber, Neoantigens in cancer immunotherapy, Science 348 (6230) (2015) 69–74.
[10] S. Bobisse, P.G. Foukas, G. Coukos, A. Harari, Neoantigen-based cancer immunotherapy, Ann. Transl. Med. 4 (14) (2016).
[11] S.K. Saini, N. Rekers, S.R. Hadrup, Novel tools to assist neoepitope targeting in personalized cancer immunotherapy, Ann. Oncol. 28 (suppl_12) (2017) xii3–xii10.
[12] T.S. Win, A.A. Malik, V. Prachayasittikul, J.E. S Wikberg, C. Nantasenamat, W. Shoombuatong, HemoPred: a web server for predicting the hemolytic activity of peptides, Future Med. Chem. 9 (3) (2017) 275–291.
[13] M. Hasan, N. Schaduangrat, S. Basith, G. Lee, W. Shoombuatong, B. Manavalan, HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation, Bioinformatics (2020) btaa160.
[14] B. Manavalan, R.G. Govindaraj, T.H. Shin, M.O. Kim, G. Lee, iBCE-EL: a new ensemble learning framework for improved linear B-cell epitope prediction, Front. Immunol. 9 (2018) 1695.
[15] S. Basith, B. Manavalan, T.H. Shin, G. Lee, iGHBP: computational identification of growth hormone binding proteins from sequences using extremely randomised tree, Comput. Struct. Biotechnol. J. 16 (2018) 412–420.
[16] N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, W. Shoombuatong, Meta-iAVP: a sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation, Int. J. Mol. Sci. 20 (22) (2019) 5743.
[17] T.S. Win, N. Schaduangrat, V. Prachayasittikul, C. Nantasenamat, W. Shoombuatong, PAAP: a web server for predicting antihypertensive activity of peptides, Future Med. Chem. 10 (15) (2018) 1749–1767.
[18] W. Shoombuatong, N. Schaduangrat, R. Pratiwi, C. Nantasenamat, THPep: a machine learning-based approach for predicting tumor homing peptides, Comput. Biol. Chem. 80 (2019) 441–451.
[19] B. Manavalan, S. Basith, T.H. Shin, S. Choi, M.O. Kim, G. Lee, MLACP: machine-learning-based prediction of anticancer peptides, Oncotarget 8 (44) (2017) 77121.
[20] B. Manavalan, S. Basith, T.H. Shin, D.Y. Lee, L. Wei, G. Lee, 4mCpred-EL: An ensemble learning framework for identification of DNA N4-methylcytosine sites in the mouse genome, Cells 8 (11) (2019) 1332.
[21] B. Manavalan, S. Basith, T.H. Shin, L. Wei, G. Lee, AtbPpred: a robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees, Comput. Struct. Biotechnol. J. 17 (2019) 972–981.
[22] B. Manavalan, S. Basith, T.H. Shin, L. Wei, G. Lee, Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation, Mol. Ther. Nucleic Acids 16 (2019) 733–744.
[23] B. Manavalan, S. Basith, T.H. Shin, L. Wei, G. Lee, mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation, Bioinformatics 35 (16) (2019) 2757–2765.
[24] B. Manavalan, T.H. Shin, M.O. Kim, G. Lee, AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest, Front. Pharmacol. 9 (2018) 276.
[25] B. Manavalan, T.H. Shin, M.O. Kim, G. Lee, PIP-EL: a new ensemble learning method for improved proinflammatory peptide predictions, Front. Immunol. 9 (2018) 1783.
[26] B. Manavalan, T.H. Shin, G. Lee, PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine, Front. Microbiol. 9 (2018) 476.
[27] M.M. Hasan, B. Manavalan, W. Shoombuatong, M.S. Khatun, H. Kurata, i4mC-Mouse: improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes, Comput. Struct. Biotechnol. J. 18 (2020) 906–912.
[28] M.M. Hasan, B. Manavalan, M.S. Khatun, H. Kurata, i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome, Int. J. Biol. Macromol. (2019), https://doi.org/10.1016/j.ijbiomac.2019.12.009.
[29] M.M. Hasan, B. Manavalan, W. Shoombuatong, M.S. Khatun, H. Kurata, i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation, Plant Mol. Biol. (2020) 1–10.
[30] W. Chen, P. Feng, F. Nie, iATP: A Sequence Based Method for Identifying Anti-tubercular Peptides, Medicinal chemistry (Shariqah (United Arab Emirates)), 2019.
[31] P. Charoenkwan, W. Shoombuatong, H.-C. Lee, J. Chaijaruwanich, H.-L. Huang, S.-Y. Ho, SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs, PloS One 8 (9) (2013).
[32] W. Shoombuatong, S. Hongjaisee, F. Barin, J. Chaijaruwanich, T. Samleerat, HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees, Comput. Biol. Med. 42 (9) (2012) 885–889.
[33] S. Hongjaisee, C. Nantasenamat, T.S. Carraway, W. Shoombuatong, HIVCoR: a sequence-based tool for predicting HIV-1 CRF01_AE coreceptor usage, Comput. Biol. Chem. 80 (2019) 419–432.
[34] X. Yang, X. Yu, An introduction to epitope prediction methods and software, Rev. Med. Virol. 19 (2) (2009) 77–96.
[35] M. Bhasin, G. Raghava, Prediction of CTL epitopes using QM, SVM and ANN techniques, Vaccine 22 (23–24) (2004) 3195–3204.
[36] K. Yu, N. Petrovsky, C. Schönbach, J.L. Koh, V. Brusic, Methods for prediction of peptide binding to MHC molecules: a comparative study, Mol. Med. 8 (3) (2002) 137–148.
[37] V. Brusic, G. Rudy, G. Honeyman, J. Hammer, L. Harrison, Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network, Bioinformatics 14 (2) (1998) 121–130.
[38] J.C. Tong, T.W. Tan, S. Ranganathan, Modeling the structure of bound peptide ligands to major histocompatibility complex, Protein Sci. 13 (9) (2004) 2523–2532.
[39] J.F.B. Lissabet, L.H. Belén, J.G. Farias, TTAgP 1.0: a computational tool for the specific prediction of tumor T cell antigens, Comput. Biol. Chem. 83 (2019) 107103.
[40] P. Charoenkwan, S. Kanthawong, N. Schaduangrat, J. Yana, W. Shoombuatong, PVPred-SCM: improved prediction and analysis of phage virion proteins using a scoring card method, Cells 9 (2) (2020) 353.
[41] P. Charoenkwan, J. Yana, N. Schaduangrat, C. Nantasenamat, M.M. Hasan, W. Shoombuatong, iBitter-SCM: identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides, Genomics (2020), https://doi.org/10.1016/j.ygeno.2020.03.019.
[42] L.R. Olsen, S. Tongchusak, H. Lin, E.L. Reinherz, V. Brusic, G.L. Zhang, TANTIGEN: a comprehensive database of tumor T cell antigens, Canc. Immunol. Immunother. 66 (6) (2017) 731–735.
[43] G. Zhang, L. Chitkushev, D.B. Keskin, V. Brusic, TANTIGEN 2.0: an online database and analysis platform for tumor T cell antigens, 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2019, pp. 2228–2231.
[44] S. Kawashima, M. Kanehisa, AAindex: amino acid index database, Nucleic Acids Res. 28 (1) (2000) 374-374.
[45] R. Pratiwi, et al., CryoProtect: a web server for classifying antifreeze proteins from nonantifreeze proteins, J. Chem. 2017 (2017).
[46] K.-C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (3) (2015) 218–234.
[47] K.-C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (1) (2011) 236–247.
[48] Z. Chen, et al., iFeature: a python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 34 (14) (2018) 2499–2502.
[49] V. Laengsri, C. Nantasenamat, N. Schaduangrat, P. Nuchnoi, V. Prachayasittikul, W. Shoombuatong, TargetAntiAngio: a sequence-based tool for the prediction and analysis of anti-angiogenic peptides, Int. J. Mol. Sci. 20 (12) (2019) 2950.
[50] X. He, et al., TargetFreeze: identifying antifreeze proteins via a combination of weights using sequence evolutionary information and pseudo amino acid composition, J. Membr. Biol. 248 (6) (2015) 1005–1014.
[51] E. Ramentol, Y. Caballero, R. Bello, F. Herrera, SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory, Knowl. Inf. Syst. 33 (2) (2012) 245–265.
[52] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[53] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[54] L. Breiman, Classification and Regression Trees, Routledge, 2017.
[55] V. Vapnik, The Nature of Statistical Learning Theory, Springer science & business media, 2013.
[56] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[57] V.N. Vapnik, An overview of statistical learning theory, IEEE Trans. Neural Network. 10 (5) (1999) 988–999.
[58] X.-J. Zhu, C.-Q. Feng, H.-Y. Lai, W. Chen, L. Hao, Predicting protein structural classes for low-similarity sequences by evaluating different features, Knowl. Base Syst. 163 (2019) 787–793.
[59] H. Lin, Z.-Y. Liang, H. Tang, W. Chen, Identifying sigma70 promoters with novel pseudo nucleotide composition, IEEE ACM Trans. Comput. Biol. Bioinf 16 (4) (2017) 1316–1321.
[60] H.-Y. Lai, et al., iProEP: a computational predictor for predicting promoter, Mol. Ther. Nucleic Acids 17 (2019) 337–346.
[61] C.-Q. Feng, et al., iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators, Bioinformatics 35 (9) (2019) 1469–1477.
[62] F.-Y. Dao, et al., Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique, Bioinformatics 35 (12) (2019) 2075–2083.
[63] M.L. Calle, V. Urrea, Letter to the editor: stability of random forest importance measures, Briefings Bioinf. 12 (1) (2010) 86–89.
[64] W. Shoombuatong, et al., Towards predicting the cytochrome P450 modulation: from QSAR to proteochemometric modeling, Curr. Drug Metabol. 18 (6) (2017) 540–555.
[65] N. Schaduangrat, S. Lampa, S. Simeon, M.P. Gleeson, O. Spjuth, C. Nantasenamat, Towards reproducible computational drug discovery, J. Cheminf. 12 (1) (2020) 9.
[66] W. Shoombuatong, et al., Towards the revival of interpretable QSAR models, Advances in QSAR Modeling, Springer, 2017, pp. 3–55.
[67] W. Shoombuatong, N. Schaduangrat, C. Nantasenamat, Towards understanding aromatase inhibitory activity via QSAR modeling, EXCLI J. 17 (2018) 688.
[68] W. Shoombuatong, N. Schaduangrat, C. Nantasenamat, Unraveling the bioactivity of anticancer peptides as deduced from machine learning, EXCLI J. 17 (2018) 734.

[69] M.M. Hasan, M.M. Rashid, M.S. Khatun, H. Kurata, Computational identification of microbial phosphorylation sites by the enhanced characteristics of sequence information, Sci. Rep. 9 (1) (2019) 1–9.
[70] M.M. Hasan, S. Yang, Y. Zhou, M.N.H. Mollah, SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties, Mol. Biosyst. 12 (3) (2016) 786–795.
[71] M.M. Hasan, Y. Zhou, X. Lu, J. Li, J. Song, Z. Zhang, Computational identification of protein pupylation sites by using profile-based composition of k-spaced amino acid pairs, PloS One 10 (6) (2015).
[72] K. Liu, W. Chen, iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications, Bioinformatics (2020), https://doi.org/10.1093/bioinformatics/btaa155.

International Journal of
Molecular Sciences

MDPI

Article

Meta-iAVP: A Sequence-Based Meta-Predictor for Improving the Prediction of Antiviral Peptides Using Effective Feature Representation

Nalini Schaduangrat [1], Chanin Nantasenamat [1], Virapong Prachayasittikul [2] and Watshara Shoombuatong [1,*]

[1] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand; nalini.sch@mahidol.edu (N.S.); chanin.nan@mahidol.edu (C.N.)
[2] Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand; virapong.pra@mahidol.ac.th
* Correspondence: watshara.sho@mahidol.ac.th; Tel.: +66-2441-4371 (ext. 2715)

Received: 24 October 2019; Accepted: 13 November 2019; Published: 15 November 2019

Abstract: In spite of the large-scale production and widespread distribution of vaccines and antiviral drugs, viruses remain a prominent human disease. Recently, the discovery of antiviral peptides (AVPs) has become an influential antiviral agent due to their extraordinary advantages. With the avalanche of newly-found peptide sequences in the post-genomic era, there is a great demand to develop a sequence-based predictor for timely identifying AVPs as this information is very useful for both basic research and drug development. In this study, we propose a novel sequence-based meta-predictor with an effective feature representation, called Meta-iAVP, for the accurate prediction of AVPs from given peptide sequences. Herein, the effective feature representation was extracted from a set of prediction scores derived from various machine learning algorithms and types of features. To the best of our knowledge, the model proposed herein represents the first meta-based approach for the prediction of AVPs. An overall accuracy and Matthews correlation coefficient of 95.20% and 0.90, respectively, was achieved from the independent test set on an objective benchmark dataset. Comparative analysis suggested that Meta-iAVP was superior to that of existing methods and therefore represents a useful tool for AVP prediction. Finally, in an effort to facilitate high-throughput prediction of AVPs, the model was deployed as the Meta-iAVP web server and is made freely available online at http://codes.bio/meta-iavp/ where users can submit query peptide sequences for determining the likelihood of whether or not these peptides are AVPs.

Keywords: therapeutic peptides; antiviral peptide; classification; machine learning; random forest; meta-predictor

1. Introduction

Human morbidity, mortality, and economic productivity continue to be affected by viral infections and their associated diseases. The dominance of sporadic viral outbreaks by zoonotic viruses such as Ebola and Zika in recent years have added to the prevalence of viral species with which humans are already in battle (i.e., human immunodeficiency virus (HIV), rhinoviruses, and influenza viruses). Viruses are successful in causing malaise to humans due to their high genetic variation, different routes of transmission, efficient replication, and the capability to persist in the host cells [1]. Furthermore, according to the global threat list of 2019 as compiled by the WHO, virus infections were seen to dominate [2]. Although, up until recently, trial and error has led to the discovery of 90 antiviral drugs approved for the treatment of 9 virus families (i.e., HIV, hepatitis B virus, hepatitis C virus,

human cytomegalovirus, influenza virus, herpes virus, varicella-zoster virus, respiratory syncytial virus, and human papillomavirus), these drugs cannot begin to cover the >200 viruses discovered thus far [3,4]. In addition, major breakthroughs in combating viral infections by vaccine production have led to remarkable advances in modern medicine such as the eradication and control of disease such as small pox [5] and polio [6], respectively. Nevertheless, the development of new vaccines remains a huge challenge in terms of time and expenses [7]. Unfortunately, the ever-increasing reports of antiviral resistance [8–10] coupled with the emergence and re-emergence of viral epidemics as observed for H1N1 [11], Ebola [12], and Zika [13] viruses, demands the production of new antiviral drugs with broad-spectrum activity [14]. More recently, peptide-based drugs have gained much interest as a new class of drugs due to their ability to be highly selective, relatively safe while also possessing good tolerability and a lower production cost [15]. Besides the advantages of peptide-based drugs, a short half-life, immunogenic potential, and low oral absorption are some of their current limitations [16].

Antiviral peptides (AVPs) is a subset belonging to the group of antimicrobial peptides (AMPs) and in that regard, exhibits antiviral activity. As of 23 September, 2019, the antimicrobial database (APD3) contains a total of 3129 AMPs, out of which, 188 are antiviral peptides [17]. Similarly, another database of antimicrobial peptides, DRAMP 2.0, contains 19,899 entries which consist of general, patent, and clinical AMPs [18,19]. In addition, a database focused solely on antiviral peptides contains 2683 experimentally verified AVPs including 624 modified AVPs [1]. Additionally, there are other databases that focus on the structure and antimicrobial activity of natural and synthetic peptides [20] as well as therapeutic peptides [21–23]. Thus, it is evident that peptide-based research is gaining momentum. In some cases, a given peptide shows more than one activity and is, therefore, called a promiscuous peptide (i.e., showing dual antimicrobial and antiviral effects). In addition, AVPs have been shown to possess cationic and amphipathic characteristics with positive net charges, all of which are essential for these peptides to work as antimicrobials [24]. Moreover, hydrophobicity seems to be a key property for peptides with activity against enveloped viruses [25,26]. To date, Fuzeon™ (Enfuvirtide), a synthetic peptide that blocks viral fusion by binding to gp41 (polypeptide chain) of HIV type-1 envelope protein is the only peptide to have been commercialized [27]. In addition, Bulevirtide™ (Myrcludex B), an anti-Hepatitis B and Hepatitis D peptide targeting sodium taurocholate co-transporting polypeptide (NTCP) of liver cells, has also been studied in a phase IIb clinical trial [28] and is scheduled for phase III trials [29]. The structure of some AVPs that have already been elucidated experimentally, are shown in Figure 1.

Furthermore, there are several mechanisms of action whereby antiviral therapeutic agents can inhibit viral activity (i.e., block the attachment of viruses, prevent fusion of viruses to host cells, interrupt the signaling process of viruses, or inhibit the replication of viruses in host cells) [30]. Currently, some studies have shown that AVPs inhibit the fusion of viruses to host cells [14,31,32]; while others have shown that AVPs interfere with viral replication [33–35] and attachment of the virus to host cells [36–38]. For example, P9, an AVP derived from mouse β-defensin acts against various flu strains (i.e., H1N1, H3N2, H5N1, H7N7, and H7N9) by binding to viral glycoproteins and inhibiting RNA replication through the prevention of viral fusion in the endosome [39]. Additionally, protegrin-1, a cyclical cationic peptide derived from swine white blood cells, showed potent antiviral activity against dengue virus by inhibiting the specific viral protease important for dengue virus replication, named NS2B-NS3pro [40]. Hence, accurately identifying the biological activities of peptides provides great importance for the exploration of the mechanism of action of AVPs and the development of antiviral drugs. However, the experimental approaches are still very slow, inefficient, and expensive. Besides, with the rapid explosion of newly-found peptide sequences in the post-genomic era, the peptide sequences in various database are rapidly increasing day by day. In that regard, bioinformatics-based tools are crucial for efficient analysis of the ever-increasing availability of data. Thus, it is in a great demand to develop a prediction model based on an efficient machine learning algorithm for fast and reliably identifying the biological activities of peptides

according to their primary sequences. This process could further shed light on novel AVPs having potent clinical outcomes.

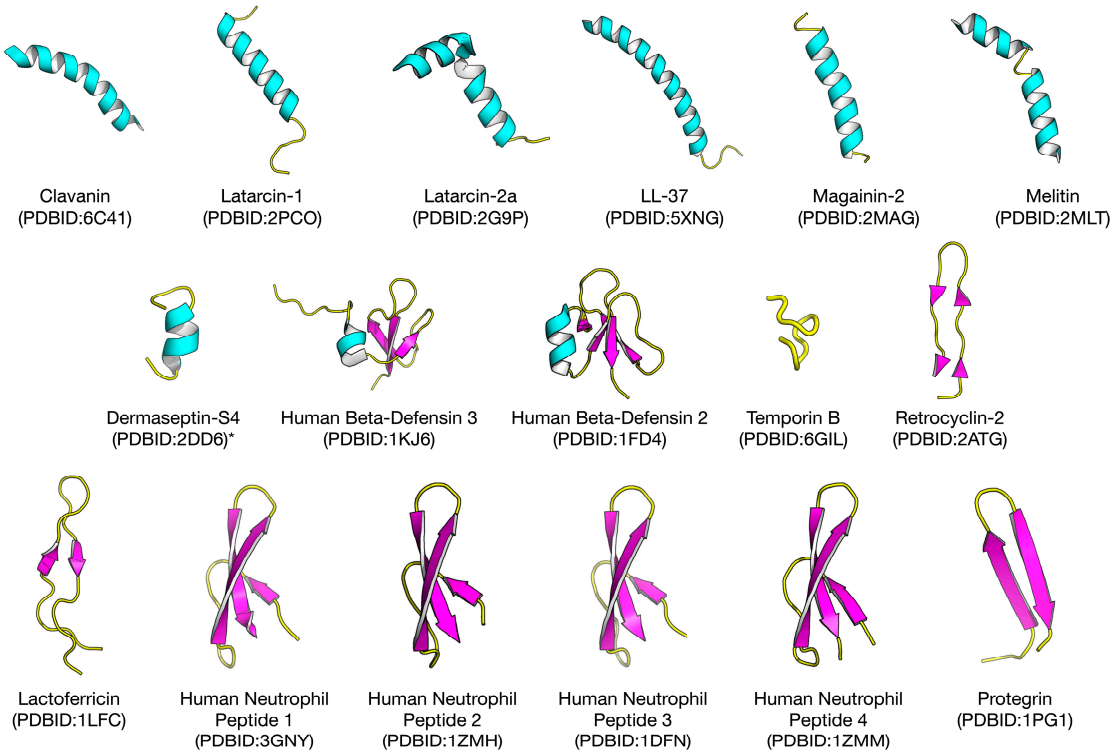

Figure 1. Structures of selected antiviral peptides that have been experimentally elucidated. Each structure is labelled by a common name followed by the Protein Data Bank Identification number (PDBID) in parenthesis on the subsequent line. * Dermaseptin-S4: The structure and available PDBID is that of a truncated peptide, which was experimentally tested to be effective.

Until now, there are four prediction models based on various machine learning (ML) algorithms that have been developed for AVP prediction, i.e., AVPpred [41], Chang et al.'s method [42], Zare et al.'s method [43], and AntiVPP 1.0 [44]. Three of the four prediction models [41,42,44] were performed on the same benchmark datasets, as summarized in Table 1. Initially, Thakur et al. [41] was the first to propose a prediction model for AVP prediction called AVPred as well as established the two benchmark datasets $T^{544p+407n}$ + $V^{60p+45n}$ and $T^{544p+544n}$ + $V^{60p+60n}$. AVPred was constructed by using a support vector machine (SVM)-based model with physicochemical properties from the AAindex database. AVPred provided moderate prediction accuracies on the independent datasets $V^{60p+45n}$ and $V^{60p+60n}$ of 85.7% and 92.5%, respectively. Shortly afterward, Chang et al. [42] utilized a combination feature of amino acid composition (AAC) and aggregation tendencies to develop a random forest (RF) model. Their prediction model achieved higher prediction accuracies as compared to AVPred with 89.5% and 93.3% for $T^{544p+407n}$ + $V^{60p+45n}$ and $T^{544p+544n}$ + $V^{60p+60n}$ datasets, respectively. Recently, Lissabet et al. [44] proposed a computation tool based on RF in conjunction with various physicochemical properties called AntiVPP 1.0. In their experimental setting, AntiVPP 1.0 was developed using one of the two benchmarked datasets, i.e., $T^{544p+544n}$ + $V^{60p+60n}$ and obtained a prediction accuracy of 93.0% which did not show any improvement as compared to AVPpred and Chang et al.'s method. Although, the above-mentioned methods produced promising results, there is still room for improvement in regards to prediction performance. First, the features used for constructing the previous methods did not offer the sequence-order or position-specific information and hence might considerably limit the prediction quality. Second, most of the existing predictors [41,42,44] were

developed using the embodiment of redundant features, causing a decrease in performance. Finally, the accuracy and transferability of the prediction model still require improvement.

Table 1. Summary of existing methods for predicting antiviral peptides.

Method	Classifier [a]	Sequence Feature [b]	Stand-Alone Program	Webserver
AVPpred [41]	SVM	AAindex	–	✓
Chang et al.'s method [42]	RF	AAC, aggregation	–	–
AntiVPP 1.0 [44]	RF	PCP	✓	–
Meta-iAVP (This study)	Meta-predictor	AAC, Am-PseAAC	–	✓

[a] RF: Random forest and SVM: Support vector machine. [b] AAC: Amino acid composition, AAindex: Amino Acid index database, aggregation: Aggregation propensity, Am-PseAAC: Amphiphilic pseudo amino acid composition, and PCP: Physicochemical properties.

Motivated by the aforementioned issues, we proposed a novel sequence-based meta-predictor, called Meta-iAVP, for the prediction of AVPs from given peptide sequences to address the shortcomings of the existing methods. First, the benchmark datasets were collected to construct a model and fairly compare with the previous models. Second, we encoded the peptide sequence with AAC, pseudo amino acid composition (PseAAC), amphiphilic pseudo amino acid composition (Am-PseAAC), dipeptide composition (DPC), and g-gap dipeptide composition (GDC). Third, we fed each feature separately into six different ML algorithms, i.e., RF, SVM, *k*-nearest neighbor (*k*-NN), recursive partitioning and regression trees (rpart), generalized linear model (glm), and extreme gradient boosting (XGBoost), to generate a new feature representation. Subsequently, effective feature representation was used to build a meta-predictor. The performance comparisons on the two benchmark datasets illustrated that Meta-iAVP significantly outperformed other existing AVP predictors. To the best of our knowledge, our proposed model is the first meta-based approach in the prediction of AVPs. We anticipate that Meta-iAVP may serve as a useful computational resource for high-throughput AVP prediction and also facilitate experimental researchers in the discovery of novel AVPs. Finally, for the convenience of experimental scientists, a Meta-iAVP web server was established and made freely available online at http://codes.bio/meta-iavp/.

2. Results

In this study, AVPs and Non-AVPs were predicted by the proposed method Meta-iAVP. Firstly, the importance of each amino acid to antiviral activities of peptides using mean decrease of Gini index (MDGI) and univariate analysis were performed. Secondly, the features that are beneficial for discriminating AVPs from Non-AVPs were determined by conducting performance comparisons between five types of features, i.e., AAC (20D), DPC (400D), GDC (400D), PseAAC (20 + 2λ), and Am- PseAAC (20 + 2λ), and six commonly used ML algorithms. Thirdly, Meta-iAVP based on the meta-predictor was constructed by using the new feature representation as the input feature. Finally, to serve easy and rapid classification of query peptide sequence, Meta-iAVP is exploited as a free prediction web server for discriminating AVPs and Non-AVPs. Figure 2 summarizes the workflow of the computational approach of Meta-iAVP.

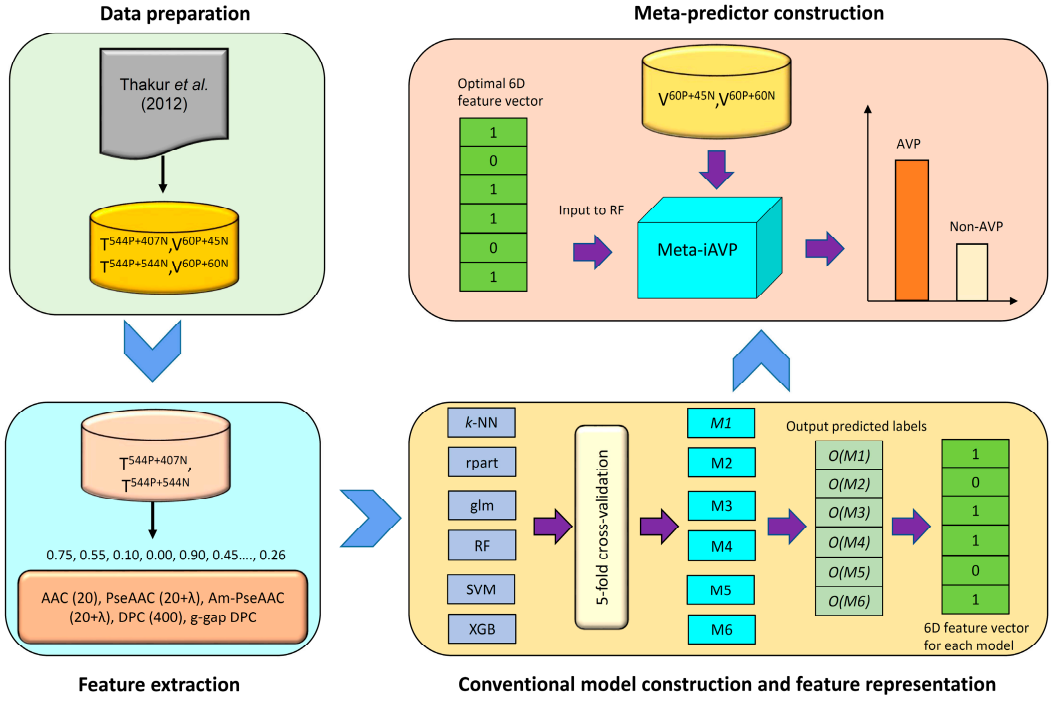

Figure 2. Schematic framework of Meta-iAVP. Overview of the proposed methodology for discriminating AVPs from Non-AVPs involving the following steps: (1) preparing two benchmark datasets; (2) extracting a peptide sequence with five types of features to encode six models; (3) constructing six ML models to generate a six-dimentional feature for each type of feature O(M), where 1 and 0 are represented with AVPs and Non-AVPs, respectively; and (4) establishing the meta-predictor for each benchmark dataset that separates a query peptide into AVPs and Non-AVPs.

2.1. Biological Space of Antiviral Peptides

As previously mentioned, AAC and DPC descriptors allow us to decipher the biochemical and biophysical properties of antiviral peptides. Preceding studies have used the AAC and DPC as to gain further insights on the characterization of therapeutic peptides [45–48] and various protein functions [49–52]. In this study, the value of MDGI was adopted to rank and estimate the importance of each AAC and DPC feature. Tables 2 and 3 list the percentage values of the top 20 amino acids for both AVPs and Non-AVPs as derived from experimental validation and random datasets, respectively. In addition, a heatmap showing the feature importance for DPC is provided in Figure 3. From Tables 2 and 3, it can be observed that the ten informative amino acids with the highest MDGI values are Lys, Thr, Leu, Ile, Ser, Trp, Asn, Arg, Cys, and Glu (49.27, 46.27, 35.06, 34.52, 30.95, 30.93, 30.19, 28.52, 26.33, and 24.87, respectively) and Lys, Pro, Cys, Thr, Ser, Trp, Val, Ala, Gly, and Leu (77.11, 68.87, 57.68, 46.84, 39.57, 36.83, 25.69, 24.40, 24.25, and 23.80, respectively) for the experimental validation and random Non-AVP datasets, respectively. Meanwhile, Figure 3a,b shows that the five top-ranked dipeptides according to their MDGI value are LL, RK, LV, WI, and EI for the experimentally validated dataset ($T^{544p+407n}$dataset) and KR, KK, GP, AS, and SA for the random Non-AVP dataset ($T^{544p+544n}$ dataset), respectively.

Table 2. Amino acid compositions (%) of AVP and Non-AVP along with their mean decrease of Gini index (MDGI) values on $T^{544p+407n}$ dataset.

Amino Acid	AVP (%)	Non-AVP (%)	Difference	*p*-Value	MDGI
K-Lys	0.092	0.078	0.014	<0.05	49.27(1)
T-Thr	0.032	0.055	−0.023	<0.05	46.27(2)
L-Leu	0.119	0.09	0.029	<0.05	35.06(3)
I-Ile	0.068	0.046	0.022	<0.05	34.52(4)
S-Ser	0.054	0.057	−0.003	0.464	30.95(5)
W-Trp	0.049	0.024	0.025	<0.05	30.93(6)
N-Asn	0.04	0.049	−0.009	<0.05	30.19(7)
R-Arg	0.079	0.082	−0.003	0.685	28.52(8)
C-Cys	0.038	0.035	0.003	0.499	26.33(9)
E-Glu	0.062	0.051	0.011	<0.05	24.87(10)
D-Asp	0.038	0.042	−0.004	0.204	22.93(11)
A-Ala	0.074	0.079	−0.005	0.384	21.85(12)
V-Val	0.049	0.062	−0.013	<0.05	21.1(13)
P-Pro	0.033	0.054	−0.021	<0.05	19.73(14)
Q-Gln	0.036	0.036	0	0.916	17.84(15)
G-Gly	0.047	0.059	−0.012	<0.05	17.25(16)
H-His	0.016	0.022	−0.006	<0.05	14.9(17)
F-Phe	0.041	0.038	0.003	0.358	14.49(18)
Y-Tyr	0.021	0.03	−0.009	<0.05	12.09(19)
M-Met	0.011	0.014	−0.003	0.085	6.27(20)

Table 3. Amino acid compositions (%) of AVP and Non-AVP along with their MDGI values on $T^{544p+544n}$ dataset.

Amino Acid	AVP (%)	Non-AVP (%)	Difference	*p*-Value	MDGI
K-Lys	0.092	0.046	0.045	<0.05	77.11(1)
P-Pro	0.033	0.068	−0.035	<0.05	68.87(2)
C-Cys	0.038	0.022	0.015	<0.05	57.68(3)
T-Thr	0.032	0.053	−0.021	<0.05	46.84(4)
S-Ser	0.054	0.083	−0.029	<0.05	39.57(5)
W-Trp	0.049	0.015	0.033	<0.05	36.83(6)
V-Val	0.049	0.069	−0.02	<0.05	25.69(7)
A-Ala	0.074	0.087	−0.013	<0.05	24.40(8)
G-Gly	0.047	0.072	−0.025	<0.05	24.25(9)
L-Leu	0.119	0.117	0.002	0.728	23.80(10)
I-Ile	0.068	0.042	0.026	<0.05	23.42(11)
H-His	0.016	0.021	−0.005	<0.05	23.13(12)
E-Glu	0.062	0.056	0.006	0.108	20.13(13)
Q-Gln	0.036	0.04	−0.004	0.18	18.50(14)
N-Asn	0.04	0.03	0.01	<0.05	18.48(15)
R-Arg	0.079	0.061	0.018	<0.05	17.67(16)
F-Phe	0.041	0.038	0.003	0.321	16.57(17)
D-Asp	0.038	0.038	0	0.982	15.75(18)
Y-Tyr	0.021	0.023	−0.001	0.537	10.57(19)
M-Met	0.011	0.017	−0.006	<0.05	10.33(20)

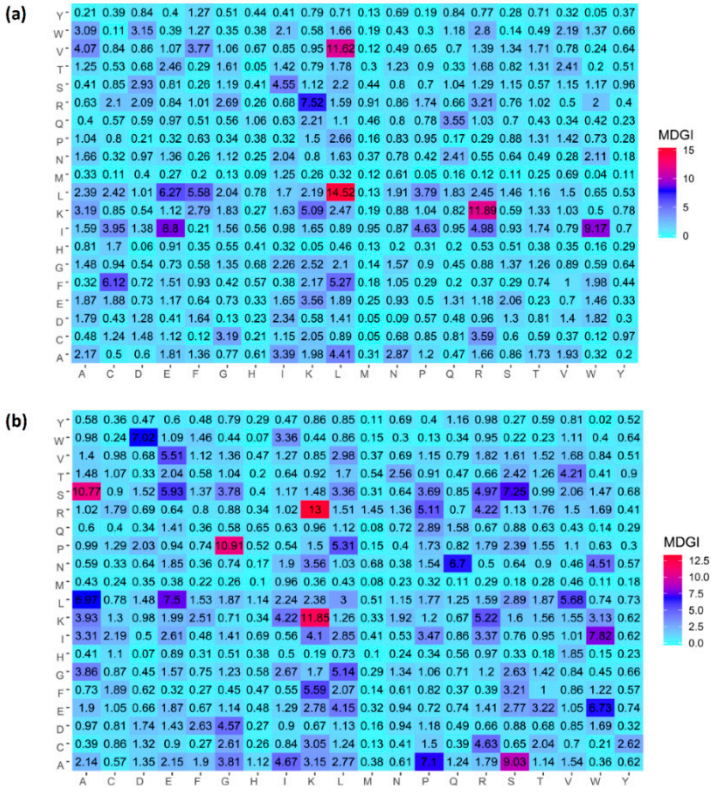

Figure 3. Heat map of the mean decrease of Gini index of dipeptide compositions for the $T^{544p+407n}$ + $V^{60p+45n}$ (**a**) and $T^{544p+544n}$ + $V^{60p+60n}$ (**b**) datasets. It should be noted that features with the largest value of MDGI are deemed to be the most important.

Interestingly, three of the five top-ranked informative amino acids from both Tables 2 and 3, are common and represent polar amino acids (i.e., Lys, Thr, and Ser), while the other amino acids are non-polar and hydrophobic residues (i.e., Leu and Ile for the experimental dataset and Pro for the random Non-AVP dataset). As stated, the top ranked amino acid, Lysine (Lys) was observed in both the experimentally validated dataset as well as the random Non-AVP dataset. Being a basic residue, Lys is abundantly found in the composition of therapeutic peptides due to its ability to enhance the electrostatic properties that facilitate the interaction and insertion of peptides into the anionic cell walls and phospholipid membranes of microorganisms [53]. Thus, the cationic role of Lys is observed in various AMPs which also function as AVPs. For instance, first published in 1986, the study by Daher et al. [54] reported the antiviral role of a cationic peptide, α-defensin which was described as inhibiting a number of viruses including herpes simplex virus types one and two, cytomegalovirus as well as inhibiting the vesicular stomatitis virus with human neutrophil peptide 1 (HNP1) in vitro. Since then, many reports have shown antiviral activity of cationic host-defense peptides such as α-defensins (i.e., HNP-1, HNP-2, HNP-3, and HNP-4), β-defensins (i.e., HBD-2 and HBD-3), and θ-defensin (i.e., Retrocyclin-2), and the use of effective antiviral therapy with cathelicidins (i.e., LL-37), as previously reviewed [36,55–59]. Furthermore, Mandelboim et al. observed that the initiation of lysis via natural killer cells by the P8 epitope of coxsackie viral peptide was pronounced with Lys as compared to other basic amino acid residues such as Arg or His [60]. Hence, the role of Lys in providing cationic properties to a given peptide sequence is fundamental and leads to the enhancement of its antiviral activities.

Threonine (Thr) is another common amino acid observed between the two datasets of Tables 2 and 3. Thr plays an essential role in the phosphorylation of virus-encoded serine/threonine kinases, a unique feature of large DNA viruses [61]. This important phosphorylation usually results in a functional change of the target protein by interfering with its enzymatic activity, cellular location,

and/or association with other proteins [61]. Therefore, a disruption of this property could hinder the efficient spread of the virus. This notion was also elucidated in a study conducted by Santos et al. [62] on a nuclear shuttle protein (NSP)-interacting kinase (NIK1) which acts as a receptor-like kinase identified as a virulence target of the begomovirus NSP. The authors conducted mutagenesis on residues Thr-474 and Thr-468 on the A-loop of the NIK1 and observed that these mutations impaired autophosphorylation and were unable to attain kinase activation. In addition, Hale et al. [63] reported that an Ala substitution of Thr-215 of the NS1 protein phosphorylation mechanism caused a disruption in viral propagation of human influenza A virus. Similarly, Hemonnot et al. [64] conducted mutational analysis of HIV mitogen-activated protein (MAP) kinase extracellular signal-regulated kinase-2 (ERK-2) by substitution of Thr-23 to Ala-23. The resulting electron microscopy and western blot analysis showed that the substitution of a single Thr-23 residue, which provided an essential function in the release of viral particles from the cell surface, was disrupted. Thus, from the aforementioned studies, it is clear that Thr is extremely vital for proper kinase phosphorylation of viral proteins which further allow for efficient viral budding from infected cells.

The third most important amino acid observed from Tables 2 and 3 was Serine (Ser) which plays an essential role in several cellular and metabolic processes [65]. In addition, as previously mentioned, Ser also makes up an important component of virus-encoded serine/threonine kinases [61]. Furthermore, an extensively studied and well-known AMP, lactoferrin, is recognized as a potent inhibitor of various viruses such as human immunodeficiency virus, herpes simplex virus types one and two, human cytomegalovirus, hepatitis C virus, hepatitis B virus, and respiratory syncytial virus. [66]. One such study conducted by Scala et al. [66], examined in detail the structure of lactoferrin-derived peptides and their activity against influenza virus using protein-protein interactions. In addition, all the peptide fragments tested were derived from the Ser418-Pro429 loop which formed a structural conformation that was critical for the resulting peptide activity. The authors noted that the presence of Ser was observed in the top three active peptide fragments. Hence, the presence of Ser in terms of formation of effective peptides for antiviral activity is highly advantageous.

2.2. Performance Comparison of Various Types of Features

To assess the effectiveness of each feature in discriminating AVPs from Non-AVPs, the five-fold CV and independent validation test were conducted for each feature by performing six commonly used ML models. Figures 4 and 5 provide the performance comparisons over the five-repeated five-fold CV and independent test results on $T^{544p+407n}$ + $V^{60p+45n}$ and $T^{544p+544n}$ + $V^{60p+60n}$ datasets, respectively. As seen in Figures 4 and 5, the average Ac over the five-repeated five-fold CV on $T^{544p+407n}$ and $T^{544p+544n}$ datasets are (78.52%, 78.72%, 79.69%, 78.68%, and 77.04%) and (84.91%, 84.88%, 85.28%, 82.19%, and 86.44%) for ACC, PseAAC, Am-PseAAC, DPC, and GDC, respectively. The average Ac of each type of feature was obtained by averaging six Ac values derived from six ML algorithms over the five-repeated five-fold CV and independent validation test. Meanwhile, the performance comparisons on the independent validation datasets $V^{60p+45n}$ and $V^{60p+60n}$ were (80.29%, 83.17%, 79.01%, 79.49%, and 77.41%) and (86.16%, 86.44%, 85.88%, 86.02%, and 84.59%) for ACC, PseAAC, Am-PseAAC, DPC, and GDC, respectively. For performance comparisons among the six ML models, the prediction results showed that average Ac over the five-repeated five-fold CV and independent test results on $T^{544p+407n}$ + $V^{60p+45n}$ and $T^{544p+544n}$ + $V^{60p+60n}$ datasets were (80.55%, 76.36%, 74.46, 86.86%, 85.93%, and 85.65%) and (82.16%, 78.00%, 74.09%, 86.86%, 85.93%, and 85.65%), respectively.

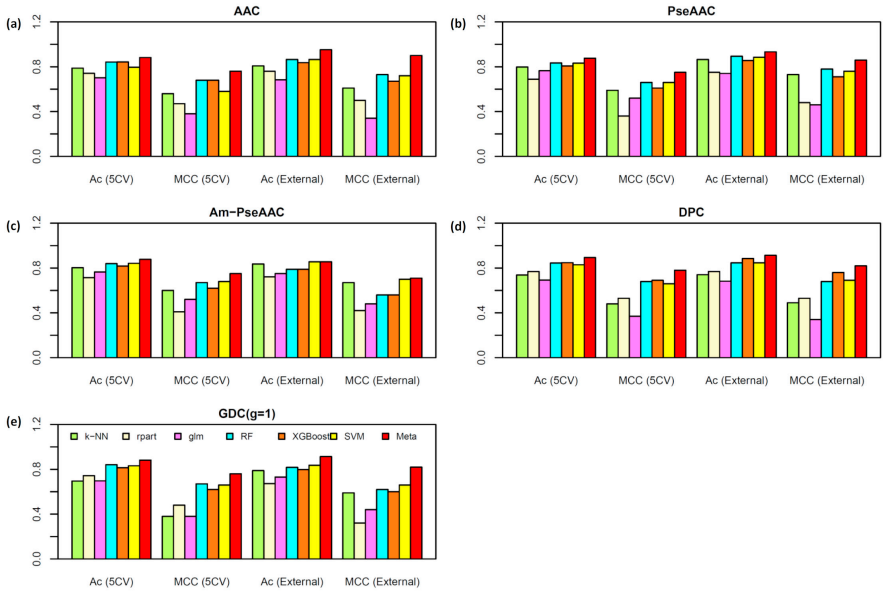

Figure 4. Performance comparisons of AVP predictors based on different six machine learning algorithms types of features, i.e., AAC (**a**), PseAAC (**b**), Am-PseAAC (**c**), DPC (**d**), and GDC (**e**), on the dataset $T^{544p+407n}$ + $V^{60p+45n}$, respectively.

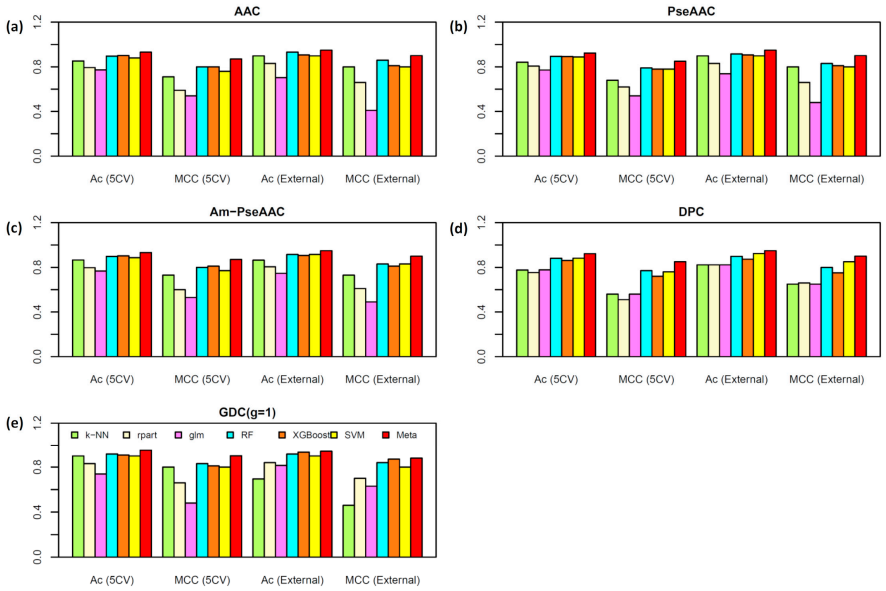

Figure 5. Performance comparisons of AVP predictors based on different six machine learning algorithms types of features, i.e., AAC (**a**), PseAAC (**b**), Am-PseAAC (**c**), DPC (**d**), and GDC (**e**), on the dataset $T^{544p+544n}$ + $V^{60p+60n}$, respectively.

By observing the performance comparisons in Figures 4 and 5, it could be summarized as follows: (i) ACC and DPC features did not afford better performance than other three predictors but they provide more interpretability for discriminating AVPs from Non-AVPs, which is helpful for biologists in designing novel peptides. This observation is quite consistent with previous works [41,42]; (ii) the top three most powerful ML models over the five-repeated five-fold CV and independent test are RF, XGBoost, and SVM; and (iii) these prediction results demonstrate that the three top-ranked important features in discriminating AVPs from Non-AVPs are PseAAC, AAC, and DPC, where AAC and PseAAC are the most beneficial features for discriminating AVPs from Non-AVPs on the benchmark datasets $T^{544p+407n}$ + $V^{60p+45n}$ and $T^{544p+544n}$ + $V^{60p+60n}$, respectively.

2.3. Construction of the Meta-iAVP Model

In general, the meta-predictor utilizes an important pattern from the predicted output derived from different predictors under the assumption that using combined methods will provide substantially accurate prediction results than a single method [67–71]. As described above, AAC and PseAAC are the most important features for discriminating AVPs from Non-AVPs. Thus, to verify the power of these two features in AVP prediction, the six ML models are trained with the AAC and PseAAC features for performing on the benchmark datasets $T^{544p+407n}$ + $V^{60p+45n}$ and $T^{544p+544n}$ + $V^{60p+60n}$, respectively, and their performance comparisons are listed in Table 4. Amongst the six ML models, Table 4 shows that the RF model with the AAC feature performs best with the highest Ac, Sn, Sp, and MCC of 86.54%, 86.54%, 86.36%, and 0.73, respectively, over the independent validation test on $V^{60p+45n}$ dataset. Meanwhile, the RF model with the PseAAC feature shows superiority in discriminating AVPs from Non-AVPs on the dataset $V^{60p+60n}$ with the highest Ac, Sn, Sp, and MCC of 91.53%, 90.00%, 93.10%, and 0.83, respectively. Therefore, the AAC and PseAAC features were used as the initial features for constructing the new feature representation to train the meta-predictor, as summarized in the Section 3.6.

Table 4. Performance comparisons between Meta-iAVP and the six machine learning algorithms as assessed by the five-repeated five-fold cross-validation and independent validation tests.

Dataset	Method [a]	Ac (%)	Sn (%)	Sp (%)	MCC
$T^{544p+407n}$	*k*-NN	78.79	88.24	66.13	0.56
	rpart	74.09	81.03	64.82	0.47
	glm	70.15	82.87	53.27	0.38
	RF	84.22	85.70	82.34	0.68
	XGBoost	84.33	86.69	80.97	0.68
	SVM	79.53	83.81	73.86	0.58
	Meta-predictor	88.17	89.23	86.94	0.76
$T^{544p+544n}$	*k*-NN	84.15	82.53	86.07	0.68
	rpart	80.63	82.37	79.73	0.62
	glm	77.11	77.78	76.78	0.54
	RF	89.44	84.18	94.68	0.79
	XGBoost	89.16	87.48	90.90	0.78
	SVM	88.79	87.13	90.71	0.78
	Meta-predictor	92.31	88.44	96.16	0.85
$V^{60p+45n}$	*k*-NN	80.77	95.00	61.36	0.61
	rpart	75.96	86.67	61.36	0.50
	glm	68.27	86.67	43.18	0.34
	RF	86.54	86.67	86.36	0.73
	XGBoost	83.65	85.00	81.82	0.67
	SVM	86.54	93.33	77.27	0.72
	Meta-predictor	95.19	96.67	93.18	0.90
$V^{60p+60n}$	*k*-NN	89.83	85.00	94.83	0.80
	rpart	83.05	88.33	77.59	0.66
	glm	73.73	78.33	68.97	0.48
	RF	91.53	90.00	93.10	0.83
	XGBoost	90.68	90.00	91.38	0.81
	SVM	89.83	88.33	91.38	0.80
	Meta-predictor	94.92	93.33	96.55	0.90

[a] *k*-NN: *k*-nearest neighbor, rpart: ecursive partitioning and regression trees, glm: Generalized linear model, RF: Random forest, XGBoost: Extreme gradient boosting, and SVM: Support vector machine.

To demonstrate the superiority and capability of our proposed model, we compared the aforementioned prediction results with the meta-predictor. Table 4 shows that the overall Ac and MCC values obtained from the meta-predictor are 4–9% and 9–17%, respectively, which are higher than

those resulting from *k*-NN, rpart, glm, RF, XGBoost, and SVM models on both $V^{60p+45n}$ and $V^{60p+60n}$ datasets. It could be stated that our proposed meta-predictors are justified as the more powerful and highly efficient AVP predictor. For convenience of the subsequent description, we will refer to these two meta-predictors as Meta-iAVP.

2.4. Analysis of new feature representation

As seen in Figure 4, Figure 5 and Table 4, the improved performances of the proposed model was achieved due to the method that takes new feature representation as the input feature and the meta-predictor as the prediction engine. In the previous sub-section, the AAC and PseAAC were mentioned as the optimum features amongst the five popular-used features, thus, these two features were used to compare with the new feature representation. To demonstrate the effectiveness of the new feature representation, the principle component analysis (PCA) approach is used to compare the distribution of AVPs (red circles) and Non-AVPs (blue circles) by representing them with PCA scores as illustrated in Figure 6. In this study, PCA analysis was performed using the FactorMineR R package [72] in R programing environment. To perform PCA analysis, $T^{544p+407n}$ + $V^{60p+45n}$ and $T^{544p+544n}$ + $V^{60p+60n}$ datasets were represented by the first two PCs (PC1 and PC2), where the percentage of variance can be explained by the first two PCs where high percentage values is suggestive of the feature importance for the predictive model. Figure 6a,c depict the distribution of AAC and a new feature representation, respectively, obtained from the dataset $T^{544p+407n}$ + $V^{60p+45n}$, while Figure 6b,d represent the distribution of PseAAC and a new feature representation, respectively, obtained from $T^{544p+544n}$ + $V^{60p+60n}$ dataset. It should be noted that, more overlap between the red and blue circles indicate the feature is less capable in AVP prediction. Remarkably, Figure 6c,d revealed that the new feature representation is efficient and effective as the input feature for discriminating AVPs from Non-AVPs. This might explain why the proposed model, Meta-iAVP, outperformed the other conventional models.

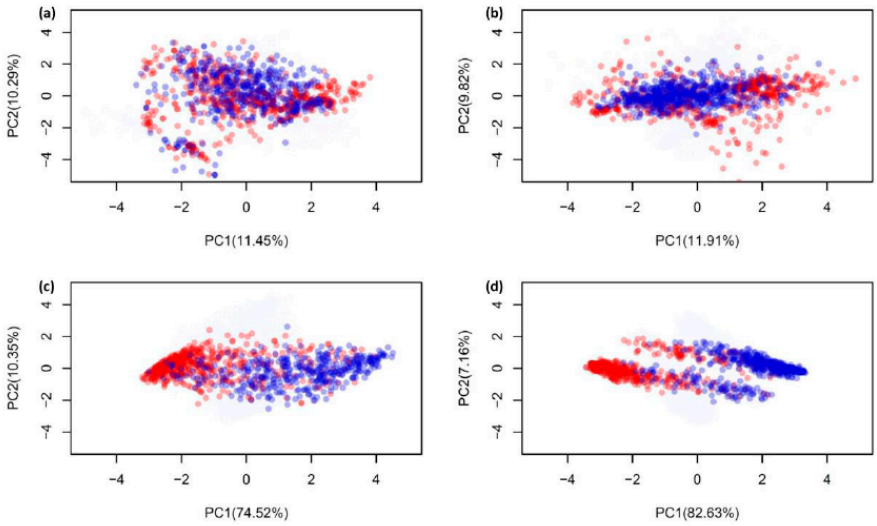

Figure 6. Principle component analysis (PCA) scores plot of the distribution of AVPs and Non-AVPs, where AVPs and Non-AVPs are represented by red and blue circles, respectively. (**a**) and (**c**) represent the distribution of amino acid composition and a new feature representation, respectively, obtained from the dataset $T^{544p+407n}$ + $V^{60p+45n}$, while (**b**) and (**d**) represent the distribution of pseudo amino acid composition and a new feature representation, respectively, obtained from the dataset $T^{544p+544n}$ + $V^{60p+60n}$.

2.5. Comparison of Meta-iAVP with the State-of-Art Predictors

To indicate the effectiveness of Meta-iAVP, we benchmarked it against the three state-of-art AVP predictors namely AVPpred [41], Chang et al.'s method [42], and AntiVPP 1.0 [44]. Among the three

AVP predictors, only AVPpred and Chang et al.'s method provided the prediction results over five-fold CV and independent test results on $T^{544p+407n}$ + $V^{60p+45n}$ and $T^{544p+544n}$ + $V^{60p+60n}$. In view of this, we only performed comparisons between Meta-iAVP with AVPpred and Chang et al.'s method. The overall performance comparisons of Meta-iAVP with other three existing methods over five-fold CV and independent test results on $T^{544p+407n}$ + $V^{60p+45n}$ and $T^{544p+544n}$ + $V^{60p+60n}$ are shown in Table 5. The pioneer work on the benchmark datasets was firstly reported by Thakur et al. [41]. Initially, they provided prediction results (Ac, MCC) on the independent dataset $V^{60p+45n}$ and $V^{60p+60n}$ with (85.70%, 0.71) and (92.50%, 0.85), respectively. Later on, Chang et al. [42] utilized the RF model cooperating with their proposed features to enhance the prediction performance. Their prediction model yielded (89.50%, 0.79) and (93.30%, 0.87) on the independent datasets $V^{60p+45n}$ and $V^{60p+60n}$, respectively, indicating that Chang et al.'s method outperformed AVPpred. Meanwhile, as noticed in Table 5, our proposed model Meta-iAVP achieved the best performances in terms of Ac, Sn, and MCC ($V^{60p+45n}$, $V^{60p+60n}$) of (95.20%, 94.90%), (93.20%, 98.30%), and (0.90, 0.90), respectively. Remarkably, Ac and MCC of Meta-iAVP were approximately 3.3–11.0% and 3.0–11.0% higher than the three state-of-art AVP predictors, thus demonstrating the superiority of our proposed predictor.

Table 5. Performance comparisons between Meta-iAVP and the three existing methods as assessed by the five-repeated five-fold cross-validation and independent validation tests.

Dataset	Method [a]	Ac (%)	Sn (%)	Sp (%)	MCC
$T^{544p+407n}$	AVPpred	85.00	82.20	**88.20**	0.70
	Chang et al.'s method	85.10	86.60	83.00	0.70
	AntiVPP 1.0	-	-	-	-
	Meta-iAVP	**88.20**	**89.20**	86.90	**0.76**
$T^{544p+544n}$	AVPpred	90.00	**89.70**	90.30	0.80
	Chang et al.'s method	91.50	89.00	94.10	0.83
	AntiVPP 1.0	-	-	-	-
	Meta-iAVP	**93.20**	89.00	**97.40**	**0.87**
$V^{60p+45n}$	AVPpred	85.70	88.30	82.20	0.71
	Chang et al.'s method	89.50	91.70	86.70	0.79
	AntiVPP 1.0	-	-	-	-
	Meta-iAVP	**95.20**	**96.70**	**93.20**	**0.90**
$V^{60p+60n}$	AVPpred	92.50	**93.30**	91.70	0.85
	Chang et al.'s method	93.30	91.70	95.00	0.87
	AntiVPP 1.0	93.00	87.00	97.00	0.87
	Meta-iAVP	**94.90**	91.70	**98.30**	**0.90**

[a] Results were reported from the works of AVPpred, Chang et al.'s method, and AntiVPP 1.0. The highest values for each performance measure are shown in bold.

With regard to the performance comparison as discussed in the two previous sub-sections, the consistent performance comparison over five-fold CV and independent validation test demonstrates that the proposed Meta-iAVP could accurately discriminate AVPs from Non-AVPs on unknown peptides. In particular, its high MCC value indicates that this new AVP model could effectively reduce the number of both false positive (FP) and false negative (FN) as well as narrow down experimental efforts. As our proposed model outperformed the other existing methods, it is reasonable due to the following aspects: (i) amongst various types of features employed in this study, PseAAC and Am-PseAAC features are firstly employed in AVP prediction. Many studies reported that these two feature have been successfully implemented to predict many peptides and proteins [15,47,50,73–79]; (ii) the parameters of our proposed model were optimized by using the five-repeated five-fold CV

indicating that our estimated parameters were more stable and accurate [80]; (iii) most of the existing predictors [41,42,44] were developed by using a combination of various types of features causing two outcomes: Information redundancy and the overfitting problem. On the other hand, we used only six-dimensional (6D) feature vectors that provided not only sufficient but also comprehensive information for AVP prediction; and (iv) our final meta-predictor was constructed by taking advantage of feature learning scheme. As seen in Tables 4 and 5, the performance comparisons revealed that our proposed model is more effective and promising for AVP prediction.

2.6. Meta-iAVP web server

In an effort to maximize the utility of the prediction model by the scientific community, we have deployed the predictive model as a web server that is also called the Meta-iAVP (i.e., using the best model as described in previous sections). The web interface of the web server was established using the Shiny package under the R programming environment. The web server is freely accessible at http://codes.bio/meta-iavp/. Screenshots of the Meta-iAVP web server are shown in Figure 7 in which panel A shows the web server prior to submission of input data and panel B shows the web server after the prediction has been made.

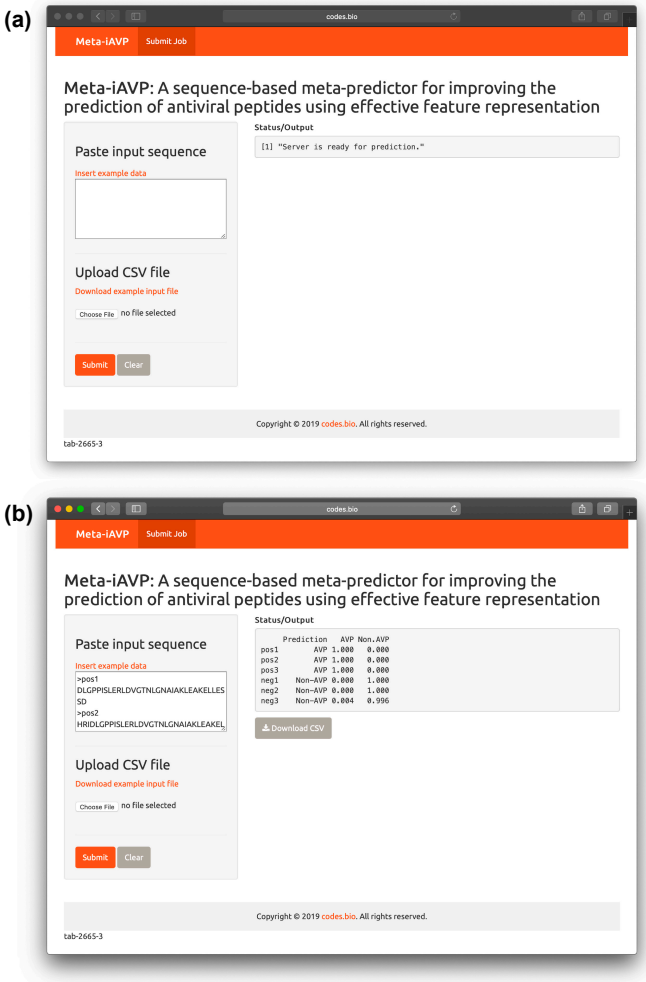

Figure 7. Screenshots of the Meta-iAVP web server before (**a**) and after (**b**) submission of sequence data for prediction. Predictions are shown alongside the probability values for each class predictions. Results are provided as a downloadable CSV file by clicking on the gray button underneath the prediction output.

Briefly, a step-by-step guide on using the web server is given below:

- **Step 1.** Proceed to entering the following URL into the web browser, http://codes.bio/meta-iavp/.
- **Step 2.** Users have the option of either entering the query peptide sequence directly into the Input box or uploading the sequence file by clicking on the "Choose file" button (i.e., found below the "Enter your input sequence(s) in FASTA format heading").
- **Step 3.** Click on the "Submit" button in order to start the prediction process.
- **Step 4.** Once predictions are made, the results output are shown in the grey box found below the "Status/Output" heading. The prediction process requires only a few seconds to process. After predictions are made, the prediction output can be conveniently downloaded as a CSV file by pressing on the "Download CSV button".

3. Materials and Methods

In practice, the prediction of peptide function is quite difficult and hard, particularly in dealing with a complicated biological system. Nevertheless, the development of an accurate prediction method might be deemed rewarding and successful if it could help provide some useful information. Thus, the present study was devoted to develop a new meta-predictor for discriminating AVPs from Non-AVPs in peptide sequences. To establish a really useful computational method for a biological system, we followed Chou's five-step guidelines mentioned in [81–85]: (i) construct or collect a reliable dataset that is experimentally validated sequences for training and validating the model; (ii) represent peptides sequences that can truly reflect their intrinsic properties to be predicted; (iii) develop a powerful algorithm or engine to operate the prediction; (iv) evaluate the prediction method with appropriate and rigorous cross-validation tests; and (v) develop a user-friendly web-server for users that can easily get their desired result without needing to go through the mathematical and statistical details. Below, we describe in detail how to deal with these steps one by one. Furthermore, Figure 2 shows the workflow of Meta-iAVP which works in discriminating peptides as AVPs or Non-AVPs.

3.1. Dataset Preparation

One of the most important steps is to establish a reliable and stringent benchmark dataset to train and test the proposed method. To objectively evaluate the performance of the proposed method and fairly compare it with the existing methods [41,42,44], the same datasets, i.e., $T^{544p+407n}$, $T^{544p+544n}$, $T^{60p+45n}$, and $T^{60p+60n}$, which were obtained from the study by Thakur et al. [41] were taken as the benchmark dataset in this study. For training the prediction model, the two benchmark datasets $T^{544p+407n}$ and $T^{544p+544n}$ that were used in this study can be summarized by the following formula:

$$T^{544p+407n} = T^{544p} \cup T^{407n} \quad (1)$$

$$T^{544p+544n} = T^{544p} \cup T^{407n} \quad (2)$$

where T^{544p} and T^{407n} represent collections of 544 and 407 experimentally validated AVP and Non-AVPs, respectively, while T^{544n} represent a collection of 544 non-experimentally validated Non-AVPs and the symbol $\cup$ represents the union from the set theory. Meanwhile, for assessing the efficient ability in predicting unknown peptides, the independent validation datasets $V^{60p+45n}$ and $V^{60p+60n}$ were used to evaluate the prediction performance from the prediction model constructed by the datasets $T^{544p+407n}$ and $T^{544p+544n}$, respectively, summarized by the following formula:

$$V^{60p+45n} = V^{60p} \cup V^{45n} \quad (3)$$

$$V^{60p+60n} = V^{60p} \cup V^{60nn} \quad (4)$$

where V^{60p} and V^{45n} represent collections of 60 and 45 experimentally validated AVP and Non-AVPs, respectively, while V^{60n} represent a collection of 60 non-experimentally validated Non-AVPs.

3.2. Feature Extraction of Peptides

In development of a sequence-based predictor for predicting the biological activity, the feature extraction process is one of the most crucial aspects where peptide sequences are represented in a way that can afford a comprehensive and proper descriptor of the features reflecting their biological activities. Given a peptide sequence (**P**), it can be represented as:

$$\mathbf{P} = p_1 p_2 p_3 \cdots p_{1N} \tag{5}$$

where p_i and N denote the ith residue in the peptide **P** and the peptide length, respectively. To develop the sequence-based predictor based on machine learning models, five different compositions and properties (i.e., AAC, DPC, PseAAC, Am-PseAAC, and GDC) that cover various aspects of sequence information were used. These five features have been successfully used to predict many peptides and proteins, such as human leukocyte antigen gene [86,87]; protein crystallization [50,88], the oligomeric states of fluorescent proteins [89], the bioactivity of host defense peptides [48], human leukocyte antigen gene [86,87], antifreeze proteins [49], hemolytic activity of peptides [46], antihypertensive activity of peptides [47], and anti-angiogenic activity of peptides [74].

AAC and DPC are the proportions of each amino acid and dipeptide in a peptide sequence P that are expressed as fixed lengths of 20 and 400, respectively. Thus, in terms of AAC and DPC features, a peptide **P** can be expressed by vectors with 20D and 400D (dimension) spaces, respectively, as formulated by:

$$\mathbf{P} = [aa_1, aa_2, \ldots, aa_{20}]^{\mathbf{T}} \tag{6}$$

$$\mathbf{P} = \left[dp_1, dp_2, \ldots, dp_{400}\right]^{\mathbf{T}} \tag{7}$$

where T is the transposed operator, while aa_1, aa_2 ... , aa_{20} and dp_1, dp_2 ... , dp_{400} are occurrence frequencies of the 20 and 400 native amino acids and dipeptides, respectively, in a peptide sequence P. As described, DPC is defined as the fraction of any two adjacent amino acids as a dipeptide pair. It could be stated that the information of non-adjacent amino acids might be lost. Thus, the GDC feature is developed to remedy such problem. This feature represents the number of occurrences of two amino acids that are separated by g gaps (i.e., $g = 0$ represents a DPC feature). In this work, g = 1, 2, 3, 4, and 5 was used.

As mentioned in previous studies [81–83] and shown in Equations (3)–(4), AAC, DPC and g-gap features only provide compositional information of a peptide sequence, but all the sequence-order information may be completely lost. To remedy this limitation, PseAAC and Am-PseAAC approaches were proposed by Chou [80,81]. According to Chou's PseAAC, the general form of PseAAC for a peptide P is formulated by:

$$\mathbf{P} = [\Psi_1, \Psi_2, \ldots, \Psi_u, \ldots, \Psi_\Omega]^{T} \tag{8}$$

where the subscript Ω is an integer to reflect the feature's dimension. The value of Ω and the component of Ψ_u, where $u = 1, 2, \ldots, \Omega$ is dependent on the protein or peptide sequences. In this study, the parameters of PseAAC (i.e., the discrete correlation factor λ and weight of the sequence information ϖ) were estimated by using the optimization procedure as described hereafter. The dimension of PseAAC feature is $20 + \lambda \times \varpi$. Since the hydrophobic and hydrophilic properties of proteins play an important role in the folding and interaction of proteins, Am-PseAAC was introduced by Chou [81]. The dimension of Am-PseAAC feature is $20 + 2\lambda$. The first 20 components are the 20 basic AAC ($p_1, p_2, \ldots, p_{20}$) while the next 2λ ones denote the set of correlation factors that reveal the physicochemical properties such as hydrophobicity and hydrophilicity (as) along a protein or peptide sequence as formulated by:

$$\mathbf{P} = \left[p_1, p_2, \ldots, p_{20}, p_{20+\lambda}, p_{20+\lambda+1}, \cdots p_{20+2\lambda}\right]^{\mathrm{T}} \tag{9}$$

The concrete values of hydrophobicity and hydrophilicity are given in Table A1. In this study, the five aforementioned features of peptide sequences were generated by using the protr package in the R programming environment [90]. The parameters of PseAAC (weight^1 and lamda^1) and Am-PseAAC (weight^2 and lamda^2) were optimized by varying weight and lambda values from 0 to 1 and 1 to 10 with step sizes of 0.1 and 1, respectively, on the whole $T^{544p+407n}$ and $T^{544p+544n}$ datasets as assessed by a 5-fold CV procedure. More details of how to estimate such parameters can be found elsewhere [15,73–75].

3.3. Machine Learning Algorithms

The capability of prediction for the proposed model developed herein is dependent not only on the feature representation process but also on the selection of machine learning algorithms. This study exploited six popular and convenient ML algorithms, namely k-NN, rpart, glm, RF, XGB, and SVM, for discriminating AVPs from Non-AVPs. Previously, these ML algorithms have been extensively utilized in various domains [84,85,91–99]. In this study, the six ML algorithms were implemented using the caret package in the R software [100]. Herein, the b concept and associated parameter optimization for the six ML algorithm are given as follows:

The *k*-NN method is conceptually based on a distance function to measure the similarity between a pair of samples. This method is categorized as an instance-based learning algorithm that has been shown to be very effective for a variety of problem domains [86]. Given a dataset consisting of labeled peptide D, a positive integer k and an unknown peptide P_{new}, the *k*-NN classifier finds the k nearest neighbors of P_{new} in the dataset D, called knn(P_{new}), and returns the dominating class, i.e., AVPs or Non-AVPs, in knn(x) as the prediction result of label of the peptide P_{new}. Optimization of *k*-NN parameter (k) was determined by using the search space to maximize a five-fold CV accuracy on the benchmark datasets $T^{544p+407n}$ and $T^{544p+544n}$ are [5,23] with the step of two.

The rpart method has been developed since the 1980s [101]. This method uses recursive partitioning for classification, regression and survival trees. This method can be used to build classification or regression models using two main steps. Firstly, the single feature which provides the best split for the dataset into two groups is identified. After that, each dataset in further divided into two groups as a sub-group, and so on recursively until a particular stopping criterion is reached, i.e., either reaching a minimum size or on improvement can be made. The second step is to resample a dataset and trim back to full tree.

The glm method is one of the most useful ML algorithms used for classification and regression tasks, because it can be applied to many different types of domains. This method is a flexible generalization of ordinary linear regression that allows the output variables having error distribution models rather than a normal distribution. The glm method attempts to determine the relationship between a set of features and classes by fitting a linear equation to a dataset consisting of labeled peptide D. In the glm analysis, stepwise regression is used to select the most informative feature for improving the prediction performance. For rpart and glm methods, the default caret parameter setting was used [90].

RF was constructed according to the described original RF algorithm [101,102]. This model is an ensemble model consisting of many classification and regression tree (CART) classifiers to perform classification and regression tasks and improves prediction performances of CART classifiers by growing a number of weak CART classifiers. RF utilizes the concepts of bagging and random feature selection. The prediction result of the classification task is obtained by using a simple voting among outputs of all trees to get one final prediction. In regression, a final prediction is the average of prediction results of many decision trees. Herein, the RF classifier was established using the randomForest package in the R software [101]. To enhance the performance of the RF model, two parameters namely ntree (i.e., the number of tree used for constructing the RF classifier) and mtry (i.e., the number of random candidate features) were determined using the caret R package [100] with a five-fold CV

approach. The search space of ntree and mtry are (100,500) and (1,10) with the steps of 100 and 1, respectively.

XGBoost is a meta-algorithm used to construct an ensemble of strong learners from weak learners, typically decision trees, on a modified dataset [103]. XGBoost, proposed by Chen and Guestrin [104] is a boosted tree algorithm, which follows the principle of gradient boosting. In recent years, XGBoost has been used extensively by data scientists and achieves satisfactory results on various biological problems [105]. In this study, the prediction of AVPs can be considered as a binary classification problem. Given a peptide sequence, we used XGBoost to predict its class label (−1 or 1), where +1 and −1 represent AVPs and Non-AVPs, respectively. For achieving the best XGBoost model, five parameters namely eta (i.e., the number of the learning rate), max_depth (i.e., the number of the depth of the tree), colsample_bytree (i.e., the number of features or variables to construct a learner), subsample (i.e., the number of samples or observations to construct a learner), and nrounds (i.e., the maximum number of iterations) were determined using the caret R package [100] with a five-fold CV approach. The search space of eta, max_depth, colsample_bytree, subsample and nrounds are (0.3, 0.4), (1,5), (0.6,0.8), (0.500, 1.000), and (50,250) with the steps of 0.1, 1, 0.2, 0.125, and 50, respectively.

SVM method is a well-known ML algorithm based on the Vapnik-Chervonenkis theory of statistical learning [106–108], which has been widely used in various biological problems [67–71,73,75,82,87,109,110]. The principle idea of this method is to map the original feature vectors having m-dimensional vector into a higher Hilbert space with n-dimensional vector, where m < n, and then determine a separating hyper plane with the largest distance between two classes. In this work, each sample on the benchmark datasets $T^{544p+407n}$ and $T^{544p+544n}$ has a corresponding label (−1 and 1) where +1 and −1 represent AVPs and Non-AVPs, respectively. Many studies reported that SVM can perform well on small sample size due to its excellent learning and best generalization abilities [73,75]. In this study, the *kernlab* R package [111] was used to implement the SVM model. To obtain an optimal SVM model, the regularization parameter C and kernel parameter γ were tuned by using grid search method with a cross-validation technique, of which the search space for C and γ are $(2^{-8},2^{8})$ and $(2^{-8}, 2^{8})$ with steps of two and two, respectively.

3.4. Feature Importance Analysis

In this work, we performed the analysis and identification of feature importance for each type of sequence feature by using the RF method to provide a better understanding of the biophysical and biochemical properties of AVPs. In practice, the RF method provides two measures for ranking feature importance, i.e., the mean decrease of Gini index and the mean decrease of prediction accuracy. Since Calle and Urrea [112] demonstrated that the MDGI provided a more robust result as compared to the mean decrease of prediction accuracy, we utilized the MDGI value to rank the importance of interpretable features including AAC and DPC. The Gini index can be defined as MDGI is an impurity measure that corresponds to the ability of each feature in discriminating the sample classes. The Gini index can be defined as

$$1 - \sum_{c=1}^{2} p^2(c|t) \tag{10}$$

where $\sum_{c=1}^{2} p^2(c|t)$ denotes the estimated class probability for node t in a tree classifier and c is the class label (i.e., either AVP or Non-AVP). Features with the largest MDGI value is considered to be an important feature as it significantly contributes to the prediction performance. Herein, the MDGI values of feature importance for each type of sequence feature is estimated using the randomForest package in the R software [101].

3.5. Performance Evaluation

For the prediction problem, it is essential to determine the success and error rates of a given classifier. In practice, there are three CV methods which are traditional approaches, i.e., sub-sampling test or k-fold cross-validation (k-fold CV), jackknife test, and independent validation test or external test. Among these, the jackknife test is recognized as the least arbitrary and most objective one, as mention by equation 28–32 in Chou [81]. Meanwhile, the external test is considered as one of the most rigorous and objective methods for cross-validation in statistics. In k-fold cross-validation procedure, the training set is randomly separated into k subsets. From the k subsets, a single subset is taken as the testing set to validate the prediction model trained and learned by the remaining k-1 subsets. This process is repeated k times, until each subset had been used as the testing set. During the jackknifing process, a single sample in the whole dataset having N samples is taken as the testing set and the remaining N-1 samples are used for training the model. This process is repeated N times, until each sample has been used as the testing set.

In order to evaluate the prediction ability of the model, the following sets of four metrics are used as follows:

$$\mathrm{Ac} = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \tag{11}$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{12}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})} \tag{13}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{14}$$

where Ac, Sn, Sp, and MCC are called accuracy, sensitivity, specificity and Matthews coefficient correlation, respectively. TP, TN, FP, and FN represent the instances of true positive, true negative, false positive and false negative, respectively. In 2009, Kim [80] demonstrated that the repeated k-fold CV procedure yielded better performances than the non-repeated k-fold CV by reducing the variability of the model. In this study, the five-repeated five-fold CV in conjunction with an independent validation test are used to measure the performance of the model.

3.6. Feature Representation Learning

Previously, feature learning scheme has been successfully implemented to predict many peptides and proteins [68–70]. Therefore, in this study, the same protocol was utilized to generate a new feature representation, as illustrated in Figure 2. The procedures of this scheme are briefly described as follows:

3.6.1. Constructing Initial Features

As mentioned above, each peptide sequence was extracted as a numerical representation based on AAC, PseAAC, Am-PseAAC, DPC, and GDC called initial features. The parameters of PseAAC (weight^1 and lamda^1) and Am-PseAAC (weight^2 and lamda^2) were optimized by varying weight and lambda values from 0 to 1 and 1 to 10 with step sizes of 0.1 and 1, respectively, on the benchmark datasets $T^{544p+407n}$ and $T^{544p+544n}$ as assessed by a five-fold CV procedure. In this study, values of weight^1, weight^2, lamda^1, and lamda^2 as performed on the benchmark datasets $T^{544p+407n}$ and $T^{544p+544n}$ are (0.6, 0.1, 3, and 4) and (0.6, 0.2, 4, and 3), respectively. Meanwhile, the parameter of GDC feature (g-gap) were optimized by choosing from one to five as assessed by a five-fold CV procedure. The optimum values of g on the benchmark datasets $T^{544p+407n}$ and $T^{544p+544n}$ are one and three, respectively.

3.6.2. Constructing a New Feature Representation

Firstly, the initial features for each type of feature were exploited to train six ML models (i.e., *k*-NN, rpart, glm, RF, XGBoost, and SVM) using the two benchmark datasets and five-fold CV for generating the predicted label. Secondly, for each type of feature, the new feature representation O(M) was obtained by concatenating all the predicted labels from the six ML models. In our experiment, the predicted label is represented with either the value of 0 or 1, where 1 and 0 represent the predicted results as AVPs and Non-AVPs, respectively. Finally, for a given peptide sequence P, the sequence P is represented with a new 6D feature vector.

3.6.3. Learning a New Feature for Meta-Predictor Representation

The new feature representations were used as input to train the RF model and subsequently used for formulating the final meta-predictor separately for the two benchmark datasets by means of the five-repeated five-fold CV.

3.7. Development of the Meta-iAVP Web Server

The best predictive model was deployed as a web server by harnessing the Shiny R package to craft the web interface. Firstly, the web server accepts as input the input sequence in FASTA format (i.e., either by from the input text box or from the uploaded FASTA file). Secondly, upon submission of the input sequence by invoking the Submit button, the query sequences are subjected to descriptor calculation and subsequently applied to the predictive model described previously. The resulting prediction of the class labels (i.e., as either AVP or Non-AVP) along with their probability values are displayed in the prediction output box. Results from the prediction process is also provided as a CSV file upon invoking the Download button found directly underneath the output box.

4. Conclusions

Owing to the medical significance and potential utility of AVPs as promising antiviral drug candidates, there is intensive efforts in the development of computational models for rapidly and accurately identifying AVPs on unknown peptides. In this study, we have developed a novel meta-predictor for AVP prediction called the Meta-iAVP. In constructing this meta-predictor, a feature representation learning scheme based on six different ML algorithms and five feature types were applied in model construction. Experimental results demonstrated the superiority of the proposed Meta-iAVP model based on the feature representation learning scheme over models constructed by the aforementioned ML algorithms and features. Furthermore, to confirm the effectiveness of the Meta-iAVP model, we have also performed comparative analyses with other state-of-the-art AVP predictors. It was observed from rigorous five-fold cross-validation and independent validation test that the proposed model was more effective and promising for AVPs prediction. To maximize the convenience of the vast majority of experimental scientists, the model was deployed as a web server that also goes by the same name, Meta-iAVP, which has been made freely available at http://codes.bio/meta-iavp/. It is anticipated that Meta-iAVP will serve as a useful, high throughput and cost-effective tool for large-scale analysis of AVPs that would help contribute to a series of interesting follow-up research studies involving antiviral peptides and other related therapeutic peptides. Although, Meta-iAVP displayed a superior performance over that of existing methods as assessed by rigorous cross-validation methods, there is still room for further improvements. For example, to improve the usefulness and efficacy for drug development and experimental research, we will make an effort to develop a computational model for predicting the inhibition of specific viruses in future studies.

Author Contributions: W.S. conceived, designed, performed, and analyzed the experiments. N.S. and W.S. analyzed the data. W.S., N.S., C.N. and V.P. drafted the manuscript. W.S. and C.N. contributed the code for constructing the web server. C.N. vetted the manuscript. All authors read and approved the manuscript.

Funding: This work is supported by the TRF Research Grant for New Scholar (No. MRG6180226) and the TRF Research Career Development Grant (No. RSA6280075) from the Thailand Research Fund, the Office of Higher Education Commission and Mahidol University.

Acknowledgments: We thank the reviewers for their great comments.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

Table A1. Original values of hydrophobicity and hydrophilicity for 20 amino acids.

Amino Acid	Hydrophobicity [81]	Hydrophilicity [81]
A-Ala	0.62	−0.50
C-Cys	0.29	−1.00
D-Asp	−0.90	3.00
E-Glu	−0.74	3.00
F-Phe	1.19	−2.50
G-Gly	0.48	0.00
H-His	−0.40	−0.50
I-Ile	1.38	−1.80
K-Lys	−1.50	3.00
L-Leu	1.06	−1.80
M-Met	0.64	−1.30
N-Asn	−0.78	0.20
P-Pro	0.12	0.00
Q-Gln	−0.85	0.20
R-Arg	−2.53	3.00
S-Ser	−0.18	0.30
T-Thr	−0.05	−0.40
V-Val	1.08	−1.50
W-Trp	0.81	−3.40
Y-Tyr	0.26	−2.30

References

1. Qureshi, A.; Thakur, N.; Tandon, H.; Kumar, M. AVPdb: A database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Res.* **2014**, *42*, D1147–D1153. [CrossRef] [PubMed]
2. Infectious Diseases Dominate WHO's List of 2019 Health Threats. Available online: https://www.contagionlive.com/news/infectious-diseases-dominate-whos-list-of-2019-health-threats (accessed on 23 January 2019).
3. Woolhouse, M.; Scott, F.; Hudson, Z.; Howey, R.; Chase-Topping, M. Human viruses: Discovery and emergence. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2012**, *367*, 2864–2871. [CrossRef] [PubMed]
4. De Clercq, E.; Li, G. Approved Antiviral Drugs over the Past 50 Years. *Clin. Microbiol. Rev.* **2016**, *29*, 695–747. [CrossRef] [PubMed]
5. Strassburg, M.A. The global eradication of smallpox. *Am. J. Infect. Control.* **1982**, *10*, 53–59. [CrossRef]
6. Bahl, S.; Bhatnagar, P.; Sutter, R.W.; Roesel, S.; Zaffran, M. Global Polio Eradication-Way Ahead. *Indian J. Pediatr.* **2018**, *85*, 124–131. [CrossRef] [PubMed]
7. Mahmoud, A. New vaccines: Challenges of discovery. *Microb. Biotechnol.* **2016**, *9*, 549–552. [CrossRef] [PubMed]
8. Duraffour, S.; Andrei, G.; Topalis, D.; Krecmerova, M.; Crance, J.M.; Garin, D.; Snoeck, R. Mutations conferring resistance to viral DNA polymerase inhibitors in camelpox virus give different drug-susceptibility profiles in vaccinia virus. *J. Virol.* **2012**, *86*, 7310–7325. [CrossRef]

9. Musiime, V.; Kaudha, E.; Kayiwa, J.; Mirembe, G.; Odera, M.; Kizito, H.; Nankya, I.; Ssali, F.; Kityo, C.; Colebunders, R.; et al. Antiretroviral drug resistance profiles and response to second-line therapy among HIV type 1-infected Ugandan children. *AIDS Res. Hum. Retrovir.* **2013**, *29*, 449–455. [CrossRef]
10. Le Page, A.K.; Jager, M.M.; Iwasenko, J.M.; Scott, G.M.; Alain, S.; Rawlinson, W.D. Clinical aspects of cytomegalovirus antiviral resistance in solid organ transplant recipients. *Clin. Infect. Dis.* **2013**, *56*, 1018–1029. [CrossRef]
11. Hui, D.S.C.; Lee, N.; Chan, P.K.S. A clinical approach to the threat of emerging influenza viruses in the Asia-Pacific region. *Respirology* **2017**, *22*, 1300–1312. [CrossRef]
12. Marston, B.J.; Dokubo, E.K.; van Steelandt, A.; Martel, L.; Williams, D.; Hersey, S.; Jambai, A.; Keita, S.; Nyenswah, T.G.; Redd, J.T. Ebola Response Impact on Public Health Programs, West Africa, 2014–2017. *Emerg. Infect. Dis.* **2017**, *23*. [CrossRef] [PubMed]
13. Souza, W.V.; Albuquerque, M.; Vazquez, E.; Bezerra, L.C.A.; Mendes, A.; Lyra, T.M.; Araujo, T.V.B.; Oliveira, A.L.S.; Braga, M.C.; Ximenes, R.A.A.; et al. Microcephaly epidemic related to the Zika virus and living conditions in Recife, Northeast Brazil. *BMC Public Health* **2018**, *18*, 130. [CrossRef] [PubMed]
14. Vigant, F.; Santos, N.C.; Lee, B. Broad-spectrum antivirals against viral fusion. *Nat. Rev. Microbiol.* **2015**, *13*, 426–437. [CrossRef] [PubMed]
15. Schaduangrat, N.; Nantasenamat, C.; Prachayasittikul, V.; Shoombuatong, W. ACPred: A Computational Tool for the Prediction and Analysis of Anticancer Peptides. *Molecules* **2019**, *24*, 1973. [CrossRef] [PubMed]
16. Lau, J.L.; Dunn, M.K. Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorg. Med. Chem.* **2018**, *26*, 2700–2707. [CrossRef]
17. Zhe, W.; Wang, G. APD: The Antimicrobial Peptide Database. *Nucleic Acids Res.* **2004**, *32*, D590–D592.
18. Kang, X.; Dong, F.; Shi, C.; Liu, S.; Sun, J.; Chen, J.; Li, H.; Xu, H.; Lao, X.; Zheng, H. DRAMP 2.0, an updated data repository of antimicrobial peptides. *Sci. Data* **2019**, *6*, 148. [CrossRef]
19. Fan, L.; Sun, J.; Zhou, M.; Zhou, J.; Lao, X.; Zheng, H.; Xu, H. DRAMP: A comprehensive data repository of antimicrobial peptides. *Sci. Rep.* **2016**, *6*, 24482. [CrossRef]
20. Pirtskhalava, M.; Gabrielian, A.; Cruz, P.; Griggs, H.L.; Squires, R.B.; Hurt, D.E.; Grigolava, M.; Chubinidze, M.; Gogoladze, G.; Vishnepolsky, B. DBAASP v. 2: An enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. *Nucleic Acids Res.* **2015**, *44*, D1104–D1112. [CrossRef]
21. Singh, S.; Chaudhary, K.; Dhanda, S.K.; Bhalla, S.; Usmani, S.S.; Gautam, A.; Tuknait, A.; Agrawal, P.; Mathur, D.; Raghava, G.P. SATPdb: A database of structurally annotated therapeutic peptides. *Nucleic Acids Res.* **2015**, *44*, D1119–D1126. [CrossRef]
22. Rajput, A.; Thakur, A.; Sharma, S.; Kumar, M. aBiofilm: A resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance. *Nucleic Acids Res.* **2017**, *46*, D894–D900. [CrossRef]
23. Sharma, D.; Priyadarshini, P.; Vrati, S. Unraveling the web of viroinformatics: Computational tools and databases in virus research. *J. Virol.* **2015**, *89*, 1489–1501. [CrossRef]
24. Bulet, P.; Stocklin, R.; Menin, L. Anti-microbial peptides: From invertebrates to vertebrates. *Immunol. Rev.* **2004**, *198*, 169–184. [CrossRef]
25. Badani, H.; Garry, R.F.; Wimley, W.C. Peptide entry inhibitors of enveloped viruses: The importance of interfacial hydrophobicity. *Biochim. Biophys. Acta.* **2014**, *1838*, 2180–2197. [CrossRef]
26. Wang, C.K.; Shih, L.Y.; Chang, K.Y. Large-Scale Analysis of Antimicrobial Activities in Relation to Amphipathicity and Charge Reveals Novel Characterization of Antimicrobial Peptides. *Molecules* **2017**, *22*, 2037. [CrossRef]
27. Dando, T.M.; Perry, C.M. Enfuvirtide. *Drugs* **2003**, *63*, 2755–2766. [CrossRef]
28. Bogomolov, P.; Alexandrov, A.; Voronkova, N.; Macievich, M.; Kokina, K.; Petrachenkova, M.; Lehr, T.; Lempp, F.A.; Wedemeyer, H.; Haag, M.; et al. Treatment of chronic hepatitis D with the entry inhibitor myrcludex B: First results of a phase Ib/IIa study. *J. Hepatol.* **2016**, *65*, 490–498. [CrossRef]
29. Clinical Progress of the Entry Inhibitor Myrcludex B. Available online: https://www.google.com.hk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=2ahUKEwiD9vaQn-blAhUhCqYKHZlwAWwQFjAAegQIBRAC&url=http%3A%2F%2Fregist2.virology-education.com%2Fpresentations%2F2018%2FHBVCure%2F15_Urban.pdf&usg=AOvVaw0xSMm7DxOyIj7S5qhMFuAC (accessed on 7 November 2018).
30. Castel, G.; Chteoui, M.; Heyd, B.; Tordo, N. Phage display of combinatorial peptide libraries: Application to antiviral research. *Molecules* **2011**, *16*, 3499–3518. [CrossRef]

31. Henriques, S.T.; Craik, D.J. Cyclotides as templates in drug design. *Drug Discov. Today* **2010**, *15*, 57–64. [CrossRef]
32. Nawae, W.; Hannongbua, S.; Ruengjitchatchawalya, M. Molecular dynamics exploration of poration and leaking caused by Kalata B1 in HIV-infected cell membrane compared to host and HIV membranes. *Sci. Rep.* **2017**, *7*, 3638. [CrossRef]
33. Ngai, P.H.; Ng, T.B. Phaseococcin, an antifungal protein with antiproliferative and anti-HIV-1 reverse transcriptase activities from small scarlet runner beans. *Biochem. Cell Biol.* **2005**, *83*, 212–220. [CrossRef]
34. Zhao, Z.; Hong, W.; Zeng, Z.; Wu, Y.; Hu, K.; Tian, X.; Li, W.; Cao, Z. Mucroporin-M1 inhibits hepatitis B virus replication by activating the mitogen-activated protein kinase (MAPK) pathway and down-regulating HNF4alpha in vitro and in vivo. *J. Biol. Chem.* **2012**, *287*, 30181–30190. [CrossRef] [PubMed]
35. Rothan, H.A.; Bahrani, H.; Rahman, N.A.; Yusof, R. Identification of natural antimicrobial agents to treat dengue infection: In vitro analysis of latarcin peptide activity against dengue virus. *BMC Microbiol.* **2014**, *14*, 140. [CrossRef]
36. Quintero-Gil, C.; Parra-Suescun, J.; Lopez-Herrera, A.; Orduz, S. In-silico design and molecular docking evaluation of peptides derivatives from bacteriocins and porcine beta defensin-2 as inhibitors of Hepatitis E virus capsid protein. *Virusdisease* **2017**, *28*, 281–288. [CrossRef]
37. Chiang, A.W.; Wu, W.Y.; Wang, T.; Hwang, M.J. Identification of Entry Factors Involved in Hepatitis C Virus Infection Based on Host-Mimicking Short Linear Motifs. *PLoS Comput. Biol.* **2017**, *13*, e1005368. [CrossRef]
38. Yin, P.; Zhang, L.; Ye, F.; Deng, Y.; Lu, S.; Li, Y.P.; Zhang, L.; Tan, W. A screen for inhibitory peptides of hepatitis C virus identifies a novel entry inhibitor targeting E1 and E2. *Sci. Rep.* **2017**, *7*, 3976. [CrossRef]
39. Nyanguile, O. Peptide antiviral strategies as an alternative to treat lower respiratory viral infections. *Front. Immunol.* **2019**, *10*, 1366. [CrossRef]
40. Rothan, H.A.; Abdulrahman, A.Y.; Sasikumer, P.G.; Othman, S.; Abd Rahman, N.; Yusof, R. Protegrin-1 inhibits dengue NS2B-NS3 serine protease and viral replication in MK2 cells. *BioMed Res. Int.* **2012**, *2012*. [CrossRef]
41. Thakur, N.; Qureshi, A.; Kumar, M. AVPpred: Collection and prediction of highly effective antiviral peptides. *Nucleic Acids Res.* **2012**, *40*, W199–W204. [CrossRef]
42. Chang, K.Y.; Yang, J.-R. Analysis and prediction of highly effective antiviral peptides based on random forests. *PLoS ONE* **2013**, *8*, e70166. [CrossRef]
43. Zare, M.; Mohabatkar, H.; Faramarzi, F.K.; Beigi, M.M.; Behbahani, M. Using Chou's pseudo amino acid composition and machine learning method to predict the antiviral peptides. *Open Bioinform. J.* **2015**, *9*, 16–19. [CrossRef]
44. Lissabet, J.F.B.; Belén, L.H.; Farias, J.G. AntiVPP 1.0: A portable tool for prediction of antiviral peptides. *Comput. Biol. Med.* **2019**, *107*, 127–130. [CrossRef] [PubMed]
45. Manavalan, B.; Basith, S.; Shin, T.H.; Choi, S.; Kim, M.O.; Lee, G. MLACP: Machine-learning-based prediction of anticancer peptides. *Oncotarget* **2017**, *8*, 77121–77136. [CrossRef] [PubMed]
46. Win, T.S.; Malik, A.A.; Prachayasittikul, V.; JE, S.W.; Nantasenamat, C.; Shoombuatong, W. HemoPred: A web server for predicting the hemolytic activity of peptides. *Future Med. Chem.* **2017**, *9*, 275–291. [CrossRef]
47. Win, T.S.; Schaduangrat, N.; Prachayasittikul, V.; Nantasenamat, C.; Shoombuatong, W. PAAP: A web server for predicting antihypertensive activity of peptides. *Future Med. Chem.* **2018**, *10*, 1749–1767. [CrossRef]
48. Simeon, S.; Li, H.; Win, T.S.; Malik, A.A.; Kandhro, A.H.; Piacham, T.; Shoombuatong, W.; Nuchnoi, P.; Wikberg, J.E.; Gleeson, M.P. PepBio: Predicting the bioactivity of host defense peptides. *RSC Adv.* **2017**, *7*, 35119–35134. [CrossRef]
49. Pratiwi, R.; Malik, A.A.; Schaduangrat, N.; Prachayasittikul, V.; Wikberg, J.E.; Nantasenamat, C.; Shoombuatong, W. CryoProtect: A Web Server for Classifying Antifreeze Proteins from Nonantifreeze Proteins. *J. Chem.* **2017**, *2017*, 15. [CrossRef]
50. Charoenkwan, P.; Shoombuatong, W.; Lee, H.-C.; Chaijaruwanich, J.; Huang, H.-L.; Ho, S.-Y. SCMCRYS: Predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. *PLoS ONE* **2013**, *8*, e72368. [CrossRef]
51. Huang, H.L. Propensity scores for prediction and characterization of bioluminescent proteins from sequences. *PLoS ONE* **2014**, *9*, e97158. [CrossRef]

52. Liou, Y.F.; Charoenkwan, P.; Srinivasulu, Y.; Vasylenko, T.; Lai, S.C.; Lee, H.C.; Chen, Y.H.; Huang, H.L.; Ho, S.Y. SCMHBP: Prediction and analysis of heme binding proteins using propensity scores of dipeptides. *BMC Bioinform.* **2014**, *15*, S4. [CrossRef]
53. Oren, Z.; Shai, Y. Mode of action of linear amphipathic alpha-helical antimicrobial peptides. *Biopolymers* **1998**, *47*, 451–463. [CrossRef]
54. Daher, K.A.; Selsted, M.E.; Lehrer, R.I. Direct inactivation of viruses by human granulocyte defensins. *J. Virol.* **1986**, *60*, 1068–1074. [PubMed]
55. Daly, N.L.; Chen, Y.K.; Rosengren, K.J.; Marx, U.C.; Phillips, M.L.; Waring, A.J.; Wang, W.; Lehrer, R.I.; Craik, D.J. Retrocyclin-2: Structural analysis of a potent anti-HIV theta-defensin. *Biochemistry* **2007**, *46*, 9920–9928. [CrossRef] [PubMed]
56. Currie, S.M.; Findlay, E.G.; McHugh, B.J.; Mackellar, A.; Man, T.; Macmillan, D.; Wang, H.; Fitch, P.M.; Schwarze, J.; Davidson, D.J. The human cathelicidin LL-37 has antiviral activity against respiratory syncytial virus. *PLoS ONE* **2013**, *8*, e73659. [CrossRef] [PubMed]
57. Gwyer Findlay, E.; Currie, S.M.; Davidson, D.J. Cationic host defence peptides: Potential as antiviral therapeutics. *BioDrugs* **2013**, *27*, 479–493. [CrossRef] [PubMed]
58. Yasin, B.; Wang, W.; Pang, M.; Cheshenko, N.; Hong, T.; Waring, A.J.; Herold, B.C.; Wagar, E.A.; Lehrer, R.I. Theta defensins protect cells from infection by herpes simplex virus by inhibiting viral adhesion and entry. *J. Virol.* **2004**, *78*, 5147–5156. [CrossRef]
59. Murakami, M.; Lopez-Garcia, B.; Braff, M.; Dorschner, R.A.; Gallo, R.L. Postsecretory processing generates multiple cathelicidins for enhanced topical antimicrobial defense. *J. Immunol.* **2004**, *172*, 3070–3077. [CrossRef]
60. Mandelboim, O.; Wilson, S.B.; Vales-Gomez, M.; Reyburn, H.T.; Strominger, J.L. Self and viral peptides can initiate lysis by autologous natural killer cells. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 4604–4609. [CrossRef]
61. Jacob, T.; Van den Broeke, C.; Favoreel, H.W. Viral serine/threonine protein kinases. *J. Virol.* **2011**, *85*, 1158–1173. [CrossRef]
62. Santos, A.A.; Carvalho, C.M.; Florentino, L.H.; Ramos, H.J.; Fontes, E.P. Conserved threonine residues within the A-loop of the receptor NIK differentially regulate the kinase function required for antiviral signaling. *PLoS ONE* **2009**, *4*, e5781. [CrossRef]
63. Hale, B.G.; Knebel, A.; Botting, C.H.; Galloway, C.S.; Precious, B.L.; Jackson, D.; Elliott, R.M.; Randall, R.E. CDK/ERK-mediated phosphorylation of the human influenza A virus NS1 protein at threonine-215. *Virology* **2009**, *383*, 6–11. [CrossRef] [PubMed]
64. Hemonnot, B.; Cartier, C.; Gay, B.; Rebuffat, S.; Bardy, M.; Devaux, C.; Boyer, V.; Briant, L. The host cell MAP kinase ERK-2 regulates viral assembly and release by phosphorylating the p6gag protein of HIV-1. *J. Biol. Chem.* **2004**, *279*, 32426–32434. [CrossRef] [PubMed]
65. Kalhan, S.C.; Hanson, R.W. Resurgence of serine: An often neglected but indispensable amino Acid. *J. Biol. Chem.* **2012**, *287*, 19786–19791. [CrossRef] [PubMed]
66. Scala, M.C.; Sala, M.; Pietrantoni, A.; Spensiero, A.; Di Micco, S.; Agamennone, M.; Bertamino, A.; Novellino, E.; Bifulco, G.; Gomez-Monterrey, I.M.; et al. Lactoferrin-derived Peptides Active towards Influenza: Identification of Three Potent Tetrapeptide Inhibitors. *Sci. Rep.* **2017**, *7*, 10593. [CrossRef]
67. Manavalan, B.; Govindaraj, R.G.; Shin, T.H.; Kim, M.O.; Lee, G. iBCE-EL: A new ensemble learning framework for improved linear B-cell epitope prediction. *Front. Immunol.* **2018**, *9*, 1695. [CrossRef]
68. Boopathi, V.; Subramaniyam, S.; Malik, A.; Lee, G.; Manavalan, B.; Yang, D.-C. mACPpred: A Support Vector Machine-Based Meta-Predictor for Identification of Anticancer Peptides. *Int. J. Mol. Sci.* **2019**, *20*, 1964. [CrossRef]
69. Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. mAHTPred: A sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* **2018**, *35*, 2757–2765. [CrossRef]
70. Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. Meta-4mCpred: A Sequence-Based Meta-Predictor for Accurate DNA 4mC Site Prediction Using Effective Feature Representation. *Mol. Ther. Nucleic Acids* **2019**, *16*, 733–744. [CrossRef]
71. Manavalan, B.; Shin, T.H.; Kim, M.O.; Lee, G. PIP-EL: A new ensemble learning method for improved proinflammatory peptide predictions. *Front. Immunol.* **2018**, *9*, 1783. [CrossRef]
72. Lê, S.; Josse, J.; Husson, F. FactoMineR: An R package for multivariate analysis. *J. Stat. Softw.* **2008**, *25*, 1–18. [CrossRef]

73. Shoombuatong, W.; Schaduangrat, N.; Pratiwi, R.; Nantasenamat, C. THPep: A machine learning-based approach for predicting tumor homing peptides. *Comput. Biol. Chem.* **2019**, *80*, 441–451. [CrossRef] [PubMed]
74. Laengsri, V.; Nantasenamat, C.; Schaduangrat, N.; Nuchnoi, P.; Prachayasittikul, V.; Shoombuatong, W. TargetAntiAngio: A Sequence-Based Tool for the Prediction and Analysis of Anti-Angiogenic Peptides. *Int. J. Mol. Sci.* **2019**, *20*, 2950. [CrossRef] [PubMed]
75. Hongjaisee, S.; Nantasenamat, C.; Carraway, T.S.; Shoombuatong, W. HIVCoR: A sequence-based tool for predicting HIV-1 CRF01_AE coreceptor usage. *Comput. Biol. Chem.* **2019**, *80*, 419–432. [CrossRef] [PubMed]
76. Kawashima, S.; Kanehisa, M. AAindex: Amino acid index database. *Nucleic Acids Res.* **2000**, *28*, 374. [CrossRef]
77. Akbar, S.; Hayat, M.; Iqbal, M.; Jan, M.A. iACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space. *Artif. Intell. Med.* **2017**, *79*, 62–70. [CrossRef]
78. Hajisharifi, Z.; Piryaiee, M.; Beigi, M.M.; Behbahani, M.; Mohabatkar, H. Predicting anticancer peptides with Chou' s pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.* **2014**, *341*, 34–40. [CrossRef]
79. Chou, K.-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2004**, *21*, 10–19. [CrossRef]
80. Kim, J.-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Comput. Stat. Data Anal.* **2009**, *53*, 3735–3745. [CrossRef]
81. Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **2011**, *273*, 236–247. [CrossRef]
82. Chen, W.; Ding, H.; Feng, P.; Lin, H.; Chou, K.-C. iACP: A sequence-based tool for identifying anticancer peptides. *Oncotarget* **2016**, *7*, 16895. [CrossRef]
83. Chou, K.-C. Impacts of bioinformatics to medicinal chemistry. *Med. Chem.* **2015**, *11*, 218–234. [CrossRef] [PubMed]
84. Shoombuatong, W.; Schaduangrat, N.; Nantasenamat, C. Towards understanding aromatase inhibitory activity via QSAR modeling. *EXCLI J.* **2018**, *17*, 688. [PubMed]
85. Shoombuatong, W.; Schaduangrat, N.; Nantasenamat, C. Unraveling the bioactivity of anticancer peptides as deduced from machine learning. *EXCLI J.* **2018**, *17*, 734. [PubMed]
86. Shoombuatong, W.; Mekha, P.; Waiyamai, K.; Cheevadhanarak, S.; Chaijaruwanicha, J. Prediction of human leukocyte antigen gene using k-nearest neighbour classifier based on spectrum kernel. *ScienceAsia* **2013**, *39*, 42–49. [CrossRef]
87. Shoombuatong, W.; Mekha, P.; Chaijaruwanich, J. Sequence based human leukocyte antigen gene prediction using informative physicochemical properties. *Int. J. Data Min. Bioinform.* **2015**, *13*, 211–224. [CrossRef] [PubMed]
88. Shoombuatong, W.; Huang, H.-L.; Chaijaruwanich, J.; Charoenkwan, P.; Lee, H.-C.; Ho, S.-Y. Predicting protein crystallization using a simple scoring card method. Proceedings of 2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Singapore, 16–19 April 2013; pp. 23–30.
89. Simeon, S.; Shoombuatong, W.; Anuwongcharoen, N.; Preeyanon, L.; Prachayasittikul, V.; Wikberg, J.E.; Nantasenamat, C. osFP: A web server for predicting the oligomeric states of fluorescent proteins. *J. Cheminformatics* **2016**, *8*, 72. [CrossRef]
90. Xiao, N.; Cao, D.-S.; Zhu, M.-F.; Xu, Q.-S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* **2015**, *31*, 1857–1859. [CrossRef]
91. Shoombuatong, W.; Prachayasittikul, V.; Anuwongcharoen, N.; Songtawee, N.; Monnor, T.; Prachayasittikul, S.; Prachayasittikul, V.; Nantasenamat, C. Navigating the chemical space of dipeptidyl peptidase-4 inhibitors. *Drug Des. Dev. Ther.* **2015**, *9*, 4515–4549.
92. Shoombuatong, W.; Prachayasittikul, V.; Prachayasittikul, V.; Nantasenamat, C. Prediction of aromatase inhibitory activity using the efficient linear method (ELM). *EXCLI J.* **2015**, *14*, 452–464.
93. Anuwongcharoen, N.; Shoombuatong, W.; Tantimongcolwat, T.; Prachayasittikul, V.; Nantasenamat, C. Exploring the chemical space of influenza neuraminidase inhibitors. *PeerJ* **2016**, *4*, e1958. [CrossRef]

94. Prachayasittikul, V.; Worachartcheewan, A.; Shoombuatong, W.; Prachayasittikul, V.; Nantasenamat, C. Classification of P-glycoprotein-interacting compounds using machine learning methods. *EXCLI J.* **2015**, *14*, 958. [PubMed]
95. Shoombuatong, W.; Prathipati, P.; Prachayasittikul, V.; Schaduangrat, N.; Ahmad Malik, A.; Pratiwi, R.; Wanwimolruk, S.; ES Wikberg, J.; Paul Gleeson, M.; Spjuth, O. Towards predicting the cytochrome P450 modulation: From QSAR to proteochemometric modeling. *Curr. Drug Metab.* **2017**, *18*, 540–555. [CrossRef] [PubMed]
96. Mandi, P.; Shoombuatong, W.; Phanus-Umporn, C.; Isarankura-Na-Ayudhya, C.; Prachayasittikul, V.; Bülow, L.; Nantasenamat, C. Exploring the origins of structure–oxygen affinity relationship of human haemoglobin allosteric effector. *Mol. Simul.* **2015**, *41*, 1283–1291. [CrossRef]
97. Shoombuatong, W.; Nabu, S.; Simeon, S.; Prachayasittikul, V.; Lapins, M.; Wikberg, J.E.; Nantasenamat, C. Extending proteochemometric modeling for unraveling the sorption behavior of compound–soil interaction. *Chemom. Intell. Lab. Syst.* **2016**, *151*, 219–227. [CrossRef]
98. Nava Lara, R.A.; Aguilera-Mendoza, L.; Brizuela, C.A.; Peña, A.; Del Rio, G. Heterologous Machine Learning for the Identification of Antimicrobial Activity in Human-Targeted Drugs. *Molecules* **2019**, *24*, 1258. [CrossRef]
99. Rajput, A.; Gupta, A.K.; Kumar, M. Prediction and analysis of quorum sensing peptides based on sequence features. *PLoS ONE* **2015**, *10*, e0120066. [CrossRef]
100. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
101. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
102. Breiman, L. *Classification and Regression Trees*; Boca Raton: New York, NY, USA, 2017.
103. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, 1189–1232. [CrossRef]
104. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; Bibliometrics: San Francisco, CA, USA, 2016; pp. 785–794.
105. Wang, H.; Liu, C.; Deng, L. Enhanced prediction of hot spots at protein-protein interfaces using extreme gradient boosting. *Sci. Rep.* **2018**, *8*, 14285. [CrossRef]
106. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 2013.
107. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
108. Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **1999**, *10*, 988–999. [CrossRef] [PubMed]
109. Shoombuatong, W.; Hongjaisee, S.; Barin, F.; Chaijaruwanich, J.; Samleerat, T. HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees. *Comput. Biol. Med.* **2012**, *42*, 885–889. [CrossRef] [PubMed]
110. Rajput, A.; Kumar, M. Anti-flavi: A web platform to predict inhibitors of flaviviruses using QSAR and peptidomimetic approaches. *Front. Microbiol.* **2018**, *9*, 3121. [CrossRef] [PubMed]
111. Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. kernlab-an S4 package for kernel methods in R. *J. Stat. Softw.* **2004**, *11*, 1–20. [CrossRef]
112. Calle, M.L.; Urrea, V. Letter to the editor: Stability of random forest importance measures. *Brief. Bioinform.* **2010**, *12*, 86–89. [CrossRef]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

Research Article

For reprint orders, please contact: reprints@future-science.com

Future Medicinal Chemistry

PAAP: a web server for predicting antihypertensive activity of peptides

Thet Su Win[1,2], Nalini Schaduangrat[1], Virapong Prachayasittikul[3], Chanin Nantasenamat*,[1] & Watshara Shoombuatong**,[1]

[1]Center of Data Mining & Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[2]Department of Medical Technology, University of Medical Technology, Yangon 11012, Myanmar
[3]Department of Clinical Microbiology & Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
*Author for correspondence: chanin.nan@mahidol.edu
**Author for correspondence: watshara.sho@mahidol.ac.th

Aim: Hypertension is associated with development of cardiovascular disease and has become a significant health problem worldwide. Naturally-derived antihypertensive peptides have emerged as promising alternatives to synthetic drugs. **Materials & methods:** This study introduces predictor of antihypertensive activity of peptides constructed using random forest classifier as a function of various combinations of amino acid, dipeptide and pseudoamino acid composition descriptors. **Results:** Classification models were assessed via independent test set that demonstrated accuracy of 84.73%. Feature importance analysis revealed the preference of proline and hydrophobic amino acids at the C-terminal as well as the preference of short peptides for robust activity. **Conclusion:** Model presented herein serves as a useful tool for predicting and analysis of antihypertensive activity of peptides.

First draft submitted: 27 November 2017; Accepted for publication: 25 April 2018; Published online: 24 July 2018

Keywords: angiotensin-converting enzyme inhibitory peptides • antihypertensive peptides • hypertension • machine learning • random forest

Hypertension is the most common metabolic health problem worldwide affecting 25% of the population and its prevalence increases with age. Hypertension not only increases the risk of stroke but can also lead to hypertensive cardiac disorders and associated abnormalities such as multi-infarct dementia and renal failure [1]. Because of its high prevalence and severe consequences, uncontrolled hypertension is regarded as one of the leading causes of death in developed countries [2]. Hence, it is extremely important to actively discover new drugs for the prevention and treatment in order to reduce hypertension-related complications.

Today, a number of drugs are currently used to control hypertension such as thiazide diuretics, calcium channel blockers, angiotensin-converting enzyme (ACE) inhibitors, angiotensin II receptor antagonists (ARBs) and beta-blockers. Despite being effective at controlling hypertension, these drugs are likely to cause notable side effects such as cough, hyperkalemia, fatigue, dizziness, headaches, loss of taste, leg edema and hypotension [3,4]. Moreover, in spite of improved understanding in blood pressure recognition and control, more than half (54%) of the population living with hypertension, remains suboptimally controlled [5]. Therefore, further research is needed to explore cost-effective *de novo* preventive medicines in order to create a more updated treatment guideline for hypertension.

Most antihypertensive agents target the ACE enzyme and angiotensin II receptor type I (AT_1) which are part of the renin–angiotensin system. In our body, local and systemic blood pressure is determined by cardiac output and vascular resistance that are mainly regulated by the renin–angiotensin system. In which, a proteolytic enzyme, called renin, is a major blood pressure regulator secreted by the kidneys in response to decreased blood pressure or low glomerular filtration rate. Renin sequentially cleaves angiotensinogen to angiotensin I which in turn is converted to angiotensin II by ACE. Angiotensin II then drives to increase the blood pressure by inducing vasoconstriction and stimulating aldosterone production by the adrenal gland which consequently increases sodium reabsorption [6].

newlands press

To discover or design a new, specific ACE inhibitor, bioactive peptides have been reported as the therapeutic potential for hypertension and cardiac disorders. These bioactive peptides are so called antihypertensive peptides (AHTPs) extracted from natural food sources, for example, milk and dairy products, meat, fish, wheat, cereal, peas, etc. Since experimental approaches are time-consuming and labor-intensive, it is desirable to develop a computational approach for effective large-scale data analysis to categorize available peptides as being either AHTP or non-AHTP and characterize some important biochemical and biophysical properties that can improve the activity of AHTP.

In 2011, Wang *et al.* [7] constructed a quantitative structure–activity relationship (QSAR) model by using partial least square regression method incorporating G-scale descriptors. The limitation of this study was that the proposed model was developed for predicting the inhibitory activity of tiny peptides, in other words, di- and tripeptides. Recently, Kumar *et al.* [8] developed computational models based on support vector machine (SVM) and amino acid composition (AAC), atomic composition and chemical descriptors as the peptide features, called AHTPIN, for predicting small (consisting of tetra-, penta- and hexapeptides), large and medium peptides. Five prediction models were constructed for individually predict peptide sequences on the five datasets assessed with a leave-one-out cross-validation (LOO-CV). Although, the AHTPIN predictor constructed with SVM model cooperating with atomic descriptor provided reasonably high prediction accuracies of 75.16%, 72.04%, 77.39%, 82.34% and 82.24% for tetra-, penta-, hexa-, small and large peptides datasets, respectively, there is still room for improvement in the performance and reliability of the AHTP predictor: it is more convenient to develop a single AHTP predictor for predicting various size of peptide sequences; rigorous validations must be considered an important part of model development and SVM model is well-recognized as a black-box approach that does not allow itself to interpretability, but the characterization and analysis of AHTPs are important for both basic research and drug development. There is much room for improvement in the prediction accuracy, model reliability and reliability.

Keeping these limitations in mind, in the present study, we made an effort to develop a more efficient and effective computation model for prediction and analysis of AHTPs. The proposed model named 'predictor of antihypertensive activity of peptides' (PAAP) was constructed by using the random forest (RF) algorithm cooperating with amino acid, dipeptide and pseudo-AAC (PseAAC). It was encouraging that the proposed model PAAP obtained the promising performances over LOO-CV, ten-fold cross-validation, leave-group-out cross-validation, external test and independent test. Furthermore, we analyze the feature importance derived from RF model to understand the biochemical and biophysical implications underlying the model. Finally, for convenience of scientific community, a web-server for the proposed model is provided at https://codes.bio/paap/

Materials & methods

Briefly, PAAP is an RF-based computational approach, which was established using the combination of three classes of interpretable protein descriptors. A summary of the workflow of PAAP for predicting and analyzing AHTPs is shown in Figure 1.

Datasets

In order to fairly compare with the previously published method, we have obtained datasets from the previous study of Kumar *et al.* [8] that consisted of three datasets corresponding to various sizes of AHTPs: small (4–6 residues), medium (7–12 residues) and large (greater than 12 residues) peptides. The dataset of small peptides consist of three subdatasets, in other words, tetra-, penta- and hexapeptides. The datasets of tetra-, penta-, hexa-, medium and large peptides contain 153, 270, 199, 268 and 76 AHTPs, respectively. Since there are no experimentally validated non-AHTPs in existence, therefore an equal number of peptides with respect to the five datasets were randomly selected from Swiss-Prot proteins and treated as non-AHTPs. For example, in the case of tetrapeptides, the number of non-AHTPs is 153.

Finally, the total number of AHTPs and non-AHTPs used in this study was 1066 per arm called AHTP2132. Herein, the AHTP2132 dataset was further divided into two subsets consisting of internal (800 AHTPs and 800 non-AHTPs) and external (266 AHTPs and 266 non-AHTPs) sets for model development and predictive capacity, respectively. Furthermore, to establish the ability of our model in filtering AHTPs, our proposed method was rigorously validated against an independent dataset. In the present study, an independent dataset of 203 novel AHTPs, which were not included in the training process, feature selection and parameter optimization, was created by manual compilation from various literatures.

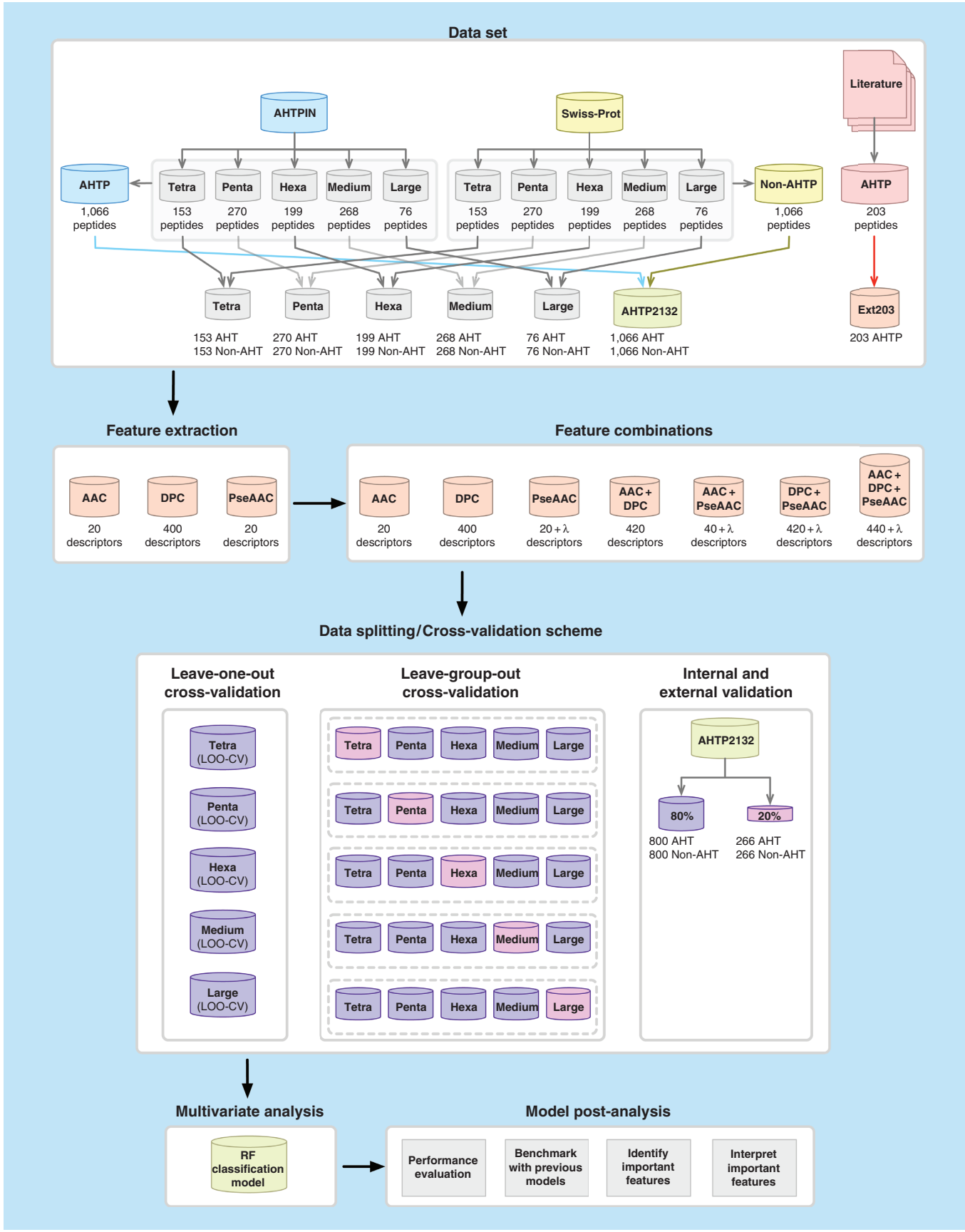

Figure 1. Schematic workflow of predictor of antihypertensive activity of peptides web server for predicting the antihypertensive activity of peptides.
AHT: Antihypertensive; RF: Random forest.

Feature extraction

One of the most challenging problems in computational biology is the development of a sequence-based predictor for rationalizing the pivotal property of biological samples (such as protein, peptide, DNA or RNA). One of the fundamental procedures is to formulate the biological samples with an effective mathematical expression that can accurately reflect the intrinsic correlation with the desired target. In this study, three types of peptide sequential features, in other words, AAC, dipeptide composition (DPC) and PseAAC, were used for constructing the AHTP predictor.

AAC is the proportion of each amino acid in a protein sequence that is expressed as a fixed length of 20 components (20-dimensional vector), reflecting the normalized occurrence frequencies p_i of the 20 natural amino acids. The AAC feature of a peptide sequence can be expressed as a vector $\mathbf{P}_{\mathrm{AAC}}$ in a 20-dimensional space:

$$\mathbf{P}_{\mathrm{AAC}} = [p_1, ., p_{20}]^{\mathrm{T}} \tag{1}$$

where T is a transpose operator.

DPC is the proportion of two consecutive amino acids or dipeptides that has a fixed length of 20 × 20 = 400 components (400-dimensional vector), reflecting the normalized occurrence frequencies pp_i of the 400 dipeptides. DPC encompasses information regarding AAC along the local order of amino acids. The DPC feature of a peptide sequence can be expressed as a vector $\mathbf{P}_{\mathrm{DPC}}$ in a 400-dimensional space:

$$\mathbf{P}_{\mathrm{DPC}} = [pp_1, ., pp_{400}]^{\mathrm{T}} \tag{2}$$

As shown in Equation 1, the AAC feature only provides the composition formation of protein sequence and its sequence-order or position-specific information would be totally lost. Therefore, using AAC to represent a peptide sequence might considerably limit the prediction quality. To remedy this problem, PseAAC was proposed to represent the sample of a protein or peptide by incorporating protein sequential information into the traditional AAC feature [9]. Therefore, in this study, the PseAAC feature was extracted by using the *protr* R package [10]. When generating PseAAC feature of a peptide sequence ($\mathbf{P}_{\mathrm{PseAA}}$), this feature is defined in 20+λ components (20+λ-dimensional vector) where λ is a lambda parameter as formulated below:

$$\mathbf{P}_{\mathrm{PseAAC}} = [p_1, ., p_{20}, p_{20+1}, ., p_{20+\lambda}, .]^{\mathrm{T}} \tag{3}$$

where the first 20 components $[p_1,.,p_{20}]$ are the same as those in the AAC feature and p_{20+1} is the PseAAC component related to the first rank of the sequence order correlation, p_{20+2} is the PseAAC component related to the second rank of the sequence order correlation and so forth. In this study, the following parameter set was applied: all of three default properties, in other words, hydrophobicity, hydrophilicity and side chain mass; the best parameter of lambda parameter (λ) and weighting factor (w) were obtained from using all possible combinations between these two parameters in which the search space for λ and w are {0.1, 0.2,.,1} and {1, 2,.,10} with steps of 0.1 and 1, respectively, for constructing the classification model. A pair of λ and w with the lowest prediction error was then selected.

Classifier selection

The predictive power of the computational models generated herein is dependent on not only the feature representations but also on the classifier used. This study utilizes a popular classifier namely RF as this approach has been successfully used in the prediction of various functions and properties of peptides [11] as well as for other biological or chemical entities [12–17]. Herein, the basic concepts and associated parameter optimization for the RF model is briefly described.

RF is an ensemble classification and regression tree (CART) classifier [11,18,19] where each classifier is generated using a random vector sampled independently from the input vector. Breiman *et al.* [18] developed the RF method by growing many weak CART trees for enhancing the prediction performance of CART. The out-of-bag (OOB) approach is used for evaluating the feature importance as follows: a third of the training dataset is utilized to construct the predictive classifier, while the remaining dataset is used for evaluating the performance of said classifier and the

feature importance of each feature can be evaluated by measuring the decrease of the prediction performance. It should be noted that the performance evaluation of the model can be either accuracy or Gini index. Herein, the RF classifier was established using the randomForest R package. For achieving the best RF model, two parameters namely *ntree* (i.e., the number of tree used for constructing the RF classifier) and *mtry* (i.e., the number of random candidate features) were tuned in an optimization procedure using the *caret* R package, of which the search space for *ntree* and *mtry* are {100,200, .,500} and {1,2,.,10} with steps of 100 and 1, respectively.

Feature importance

The analysis and identification of feature importance can provide a better understanding of the biological and physicochemical properties of antihypertensive activities. The efficient and effective built-in feature importance estimator of the RF method was used to reveal differences between AHPTs and non-AHTPs. The Gini index is an impurity measure that corresponds to the ability of each feature in discriminating the sample classes. The Gini index can be defined as:

$$1-\sum_{c=1}^{2} p^2(c\,|\,t) \tag{4}$$

where $p(c|t)$ denotes the estimated class probability for node t in a tree classifier where c is the class label (i.e., either AHPTs or non-AHTPs). The Gini importance or the mean decrease of Gini index (MDGI) is a robust quantity that is used to estimate feature importance. Features with the largest MDGI value are considered to be important as they significantly contribute to the prediction performance. In order to increase the reliability for identifying the feature importance, 10 RF models were constructed by varying the *mtry* parameter settings from 2 to 20 (*mtry = 2, 3, 5, 7, 9, 11, 13, 15, 17, 20)* and fixing the *ntree* parameter with 100 [11,20]. Finally, the average values of MDGI from 10 runs of feature importance estimations were used in this study.

Evaluation of the prediction performance

Usually, in statistical predictions, the following three methods are often used to identify the empirical predictive model for its robustness in practical applications: *N*-fold cross-validation, jackknife test and external validation [9,11,19]. To fairly compare the existing approaches, a LOO-CV was carried out. In the LOO-CV procedure, the training dataset is randomly divided into *N* samples. From *N* samples, a single sample is used to validate the model trained by using *N*-1 samples. This procedure is repeated *N* times, until each sample has been used once as the testing set. Furthermore, to evaluate the prediction model as being more robust and efficient, two additional procedures consisting of an external validation test and a leave-group-out cross-validation (LGO-CV) [21] were considered and graphically illustrated in Figure 1. In the present study, the internal and external sets were constructed with random procedure. In consideration of the impact of the random sampling procedure, we repeated this construction 20-times. The former set was used as the training set and subjected to ten-fold CV in which the dataset was partitioned into 10 subsets and a subset was left out as the testing set while the remaining were used for training the model. This process was repeated iteratively until all the folds have had the chance to be left out as the testing set and the mean and standard deviation values of the statistical parameters were reported.

Additionally, the validation of the empirical prediction model is crucial for assessing its robustness. The performance of the prediction method is usually measured by four standard statistical parameters namely accuracy (Ac), sensitivity (Sn), specificity (Sp) and Matthews coefficient correlation (MCC). Equations describing the details of the four statistical parameters are shown below:

$$Ac = \frac{TP+TN}{(TP+TN+FP+FN)} \times 100 \tag{5}$$

$$Sn = \frac{TP}{(TP+FN)} \times 100 \tag{6}$$

$$Sp = \frac{TN}{(TN + FP)} \times 100 \tag{7}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{8}$$

where TP is the number of correctly predicted peptides in the positive dataset (AHTPs) and TN is the number of correctly predicted peptides in the negative dataset (non-AHTPs). Meanwhile, FP represents the number of incorrectly predicted peptides in the positive dataset and FN represents the number of incorrectly predicted peptides in the negative dataset.

Developing the web server

The PAAP web server was implemented in an R environment using the Shiny package that facilitates the deployment of R code as an interactive web application. The main R code used for the construction of the predictive model is reliant on the following R packages: *protr*, *seqinr* and *randomForest*. As for the web application, it is reliant on the *shiny*, *shinyjs*, *shinythemes* and *markdown* R packages.

Research reproducibility

To ensure the reproducibility and facilitate the utilization of the proposed models, all R codes and datasets used for computing the prediction results and creation of graphical figures are available on GitHub at https://github.com/Shoombuatong2527/PAAP and https://github.com/chaninlab/PAAP.

Results & discussion

Prediction of antihypertensive activity of peptides

In this study, interpretable RF model and three peptide features (e.g., AAC, DPC and PseACC) were used to establish prediction models which were then evaluated with the LOO-CV procedure. Table 1 shows the performance comparison amongst RF models with various types of peptide features and all their possible combinations, in other words, AAC (20-dimensional), DPC (400-dimensional), PseAAC (20 + λ-dimensional), AAC + DPC (420-dimensional), AAC + C PseAAC (40 + λ-dimensional), PseAAC + DPC (420 + λ-dimensional) and AAC + PseAAC + DPC (440 + λ-dimensional), on five datasets. As mentioned in a previous work [8], the five AHTP predictors were proposed for discriminating AHTPs from non-AHTPs on five datasets. Herein, an effort was made to develop only a single AHTP predictor for discriminating AHTPs from non-AHTPs on the five datasets. Our proposed method was further subjected to rigorous validation procedures consisting of ten-fold CV, external validation test and LGO-CV.

As seen in Table 1, using the PseAAC feature provides the best accuracy of 75.82%, 69.07%, 78.39%, 80.43% and 84.47% for the datasets of tetrapeptide, pentapeptide, hexapeptide, medium peptide and large peptide, respectively. Meanwhile, the second best accuracy of 74.84%, 67.04%, 75.13%, 79.76% and 83.55% for the datasets of tetra-, penta-, hexa-, medium and large peptides, respectively, were achieved from using the AAC feature. As suggested by previous related studies [11,19], the combined use of the three peptide features might increase the prediction performance. Therefore, in the present study, four combinations of peptide features were also used for constructing AHTP predictors. The combination of AAC + DPC performed well with the highest accuracies of 76.14%, 73.15%, 78.39%, 80.30% and 84.87% for the datasets of tetrapeptide, pentapeptide, hexapeptide, medium peptide and large peptide, respectively. This result suggested that using the combination of AAC + DPC had an excellent prediction performance with an overall accuracy of 77.83.

Furthermore, the RF model with the combination of AAC + DPC was also subjected to ten-fold CV, external validation test and LGO-CV to confirm its reliability and true predictive power. In the case of ten-fold CV and external validation test, this method reached a mean accuracy/sensitivity/specificity/MCC of 77.43 ± 0.64%/78.23 ± 0.64%/76.68 ± 0.83%/0.55 ± 0.01 and 78.54 ± 1.48%/76.95 ± 1.97%/80.13 ± 2.40%/0.57 ± 0.03, respectively, while still providing a satisfactory result with an accuracy of 75.44% (72.91% sensitivity and 78.50% specificity) as assessed by the LGO-CV procedure.

Table 1. Performance comparisons of random forest models with various types of features over leave-out-out cross-validation.

Peptide class	Feature	Ac (%)	Sn (%)	Sp (%)	MCC
Tetrapeptide	AAC	74.84	75.00	74.68	0.50
	PseAAC	75.82	75.82	75.82	0.52
	DPC	75.82	75.82	75.82	0.52
	AAC + PseAAC	75.16	75.50	74.84	0.50
	AAC + DPC	76.14	75.97	76.32	0.52
	DPC + PseAAC	70.26	68.24	72.79	0.41
	AAC + DPC + PseAAC	75.49	75.32	75.66	0.51
Pentapeptide	AAC	67.04	66.79	67.29	0.34
	PseAAC	69.07	69.00	69.14	0.38
	DPC	57.78	57.34	58.27	0.16
	AAC + PseAAC	68.70	67.47	70.12	0.38
	AAC + DPC	73.15	72.73	73.58	0.46
	DPC + PseAAC	71.85	71.38	72.35	0.44
	AAC + DPC + PseAAC	70.37	69.78	70.99	0.41
Hexapeptide	AAC	75.13	77.47	73.15	0.50
	PseAAC	78.39	81.56	75.80	0.57
	DPC	76.63	76.50	76.77	0.53
	AAC + PseAAC	78.89	83.63	75.33	0.58
	AAC + DPC	78.39	81.22	76.04	0.57
	DPC + PseAAC	78.14	80.43	76.17	0.56
	AAC + DPC + PseAAC	78.39	80.87	76.28	0.57
Medium peptide	AAC	79.76	80.85	78.74	0.6
	PseAAC	80.43	80.77	80.11	0.61
	DPC	80.16	81.71	78.76	0.60
	AAC + PseAAC	80.71	81.39	80.05	0.61
	AAC + DPC	80.30	80.22	80.38	0.61
	DPC + PseAAC	79.48	79.73	79.25	0.59
	AAC + DPC + PseAAC	80.16	80.33	80.00	0.60
Large peptide	AAC	83.55	84.00	83.12	0.67
	PseAAC	84.87	86.30	83.54	0.70
	DPC	82.89	87.88	79.07	0.66
	AAC + PseAAC	84.87	87.32	82.72	0.70
	AAC + DPC	84.87	88.41	81.93	0.70
	DPC + PseAAC	83.55	86.96	80.72	0.67
	AAC + DPC + PseAAC	85.53	88.57	82.93	0.71

AAC: Amino acid composition; Ac: Accuracy; DPC: dipeptide composition; MCC; Matthews coefficient correlation+; PseAAC: Pseudo amino acid composition; Sn: Sensitivity; Sp: Specificity.

In summary, the RF model incorporated with the combination of AAC + DPC provided an excellent prediction performance as assessed by LOO-CV, ten-fold CV, external validation test and LGO-CV with accuracies of 78.57%, 77.43 ± 0.64%, 78.54 ± 1.48% and 75.44%, respectively. Therefore, this method is justified as an excellent AHTP predictor. For convenience of the subsequence description, we will refer to this method as PAAP.

Comparison with existing AHTP predictor

In order to determine the value of our proposed method ATHpred, it is necessary to compare it with previously published methods. Previously, only one computational model for the prediction of small, medium and large peptides (AHTPIN) [8] has been proposed whereby the antihypertensive activity of peptides was predicted from sequences. The best prediction results of AHTPIN were obtained from using the SVM model incorporated with atomic composition. Table 2 lists the performance comparison between the PAAP and AHTPIN as assessed by the LOO-CV procedure. Furthermore, in order to show that the proposed method is robust and capable

Table 2. Summarize of performance comparisons between predictor of antihypertensive activity of peptides and AHTIN assessed by leave-one-out cross-validation.

Peptide class	Method	Ac (%)	Sn (%)	Sp (%)	MCC
Tetrapeptide	AHTpin	75.16	70.59	79.74	0.51
	PAAP	76.14	75.97	76.32	0.52
Pentapeptide	AHTpin	72.04	72.96	71.11	0.44
	PAAP	73.15	72.73	73.58	0.46
Hexapeptide	AHTpin	77.39	74.87	79.9	0.55
	PAAP	78.39	81.22	76.04	0.57
Medium peptides	AHTpin	82.34	81.25	83.42	0.65
	PAAP	80.30	80.22	80.38	0.61
Large peptides	AHTpin	82.24	84.21	80.26	0.68
	PAAP	84.87	88.41	81.93	0.70
Average	AHTpin	77.83	76.78	78.89	0.57
	PAAP	78.57	79.71	77.65	0.57

MCC: Matthews coefficient correlation; PAAP: Predictor of antihypertensive activity of peptide.

Table 3. Summary of performance comparisons between predictor of antihypertensive activity of peptides and AHTPIN as assessed by independent test.

Peptide class	# peptides	Method	Prediction results	Ac (%)
Tetrapeptide	28	AHTPIN†	26 AHTPs and 2 non-AHTPs	92.86
		PAAP	27 AHTPs and 1 non-AHTP	96.43
Pentapeptide	31	AHTPIN†	21 AHTPs and 10 non-AHTPs	67.74
		PAAP	27 AHTPs and 4 non-AHTPs	87.10
Hexapeptide	24	AHTPIN†	19 AHTPs and 5 non-AHTPs	79.17
		PAAP	19 AHTPs and 5 non-AHTPs	79.17
Medium peptides	86	AHTPIN†	70 AHTPs and 16 non-AHTPs	81.40
		PAAP	69 AHTPs and 17 non-AHTP	80.23
Large peptides	34	AHTPIN†	26 AHTPs and 2 non-AHTPs	76.47
		PAAP	27 AHTPs and 1 non-AHTP	79.41
Total/average	203	AHTPIN†	165 AHTPs and 38 non-AHTPs	81.28
		PAAP	172 AHTPs and 31 non-AHTPs	84.73

†Results were obtained by feeding the peptide sequences on an independent set to the web server of AHTPIN [8], where the independent set consists of 203 peptides were experimental validated AHTPs.
PAAP: Predictor of antihypertensive activity of peptide.

for predicting unknown peptide sequences, PAAP and AHTPIN were used to predict an independent dataset consisting of 203 AHTPs. Our proposed model, PAAP was trained on AHTP2132 dataset and the best parameter of *ntree* and *mtry* of 400 and 10, respectively, while the prediction results of AHTPIN were derived from http://crdd.osdd.net/raghava/ahtpin.

From Table 2, it can be seen that the overall prediction results of PAAP are very comparable with a better performance than AHTPIN concerning all four standard statistical parameters. In the case of the tetra-, hexa- and large peptides datasets, PAAP outperformed AHTPIN with improvements of >7 and >2% on sensitivity and MCC, respectively. Furthermore, to quantify a model's potential for predicting unknown peptides, the PAAP and AHTPIN models were run on an independent dataset. The independent dataset consisted of 203 peptides that is made up of experimentally validated AHTPs. Table 3 shows the performance comparisons between PAAP and AHTPIN as assessed by the independent test set. Interestingly, PAAP performed better than AHTPIN with the highest accuracies of 96.43% (27/28), 87.10% (27/31) and 79.41% (27/34) on tetrapeptides, pentapeptide and large peptide datasets, respectively. Performance comparisons on five datasets revealed that PAAP was able to outperform AHTPIN with average accuracies of 84.73% (172/203) and 81.23% (165/203), respectively. It is anticipated that PAAP may become a useful high-throughput tool for both discriminating AHTPs from non-AHTPs and identifying ACE inhibitory activities of uncharacterized AHTPs. In order to identify ACE inhibitory

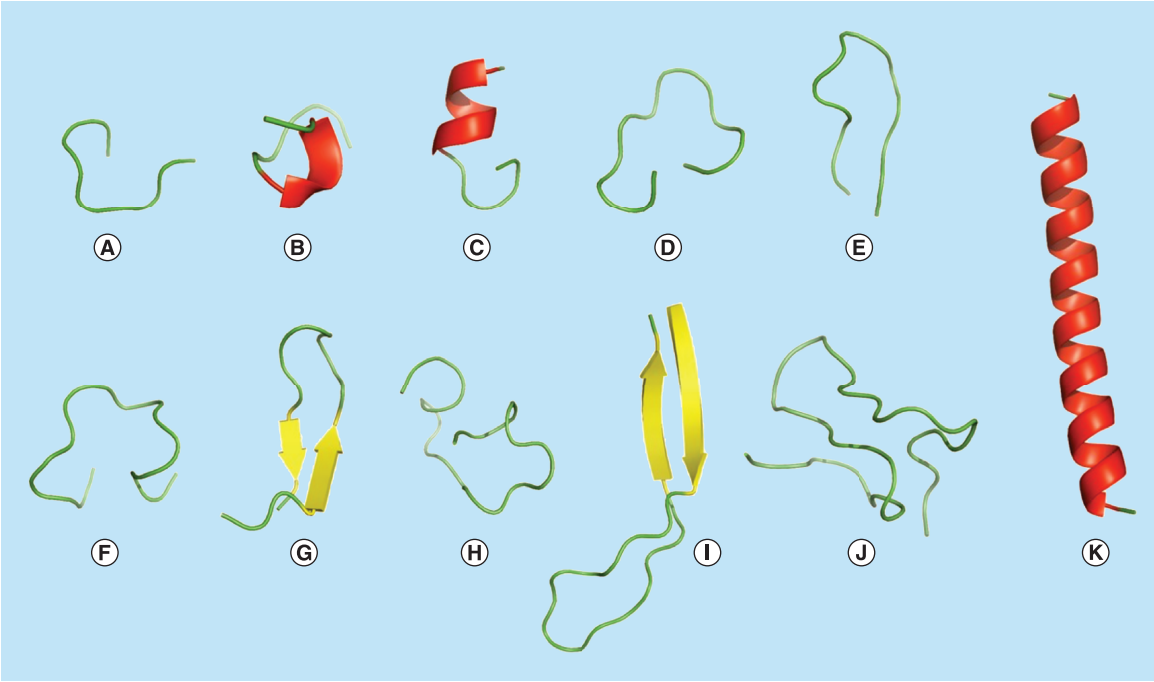

Figure 2 3D structure of different sized antihypertensive activities of peptides. (A–D) are medium-sized AHTPs comprising of 7–12 amino acids and **(E–K)** are large-sized antihypertensive activities of peptides with more than 13 amino acids. Peptide structures were constructed by PEP-FOLD server (http://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-FOLD3/) and visualized in PyMOL version 1.7.6.3 (https://pymol.org/) [27].

activities, the present PAAP model can be applied without major modifications by using appropriate types of peptide features and parameters that are needed for assessing the performance of the model. For instance, Wu *et al.* [22] developed a THP predictor by using partial least squares as the learning method, while 3-*z* and 5-*z* scales were used as peptide features. For performance evaluation, prediction models were subjected to LOO-CV and the performance was quantified with R^2 and Q^2.

Feature importance of AHTP

It is well-known that AHTPs are very short sequences usually composing of less than 30 amino acid residues. As noticed in AHTPs from AHTPDB database [23] and other AHTPs derived from experimental-based literatures, most of AHTPs are composed of 4–14 amino acids. Figure 2 shows the 3D structures of AHTPs with different lengths of amino acids are resemble of α-helix, β-sheet, loop and mixed structures. In this study, the MDGI value and percent composition of amino acids were used to rank the feature importance of AAC and DPC features to provide understanding of the biochemical and biophysical properties of antihypertensive activities of peptides. Meanwhile, sequence logos were used to visualize the information content for each amino acid in those three datasets as shown in Figure 3. The overall height of the stack for each position indicates its sequence conservation, while the size of residue demonstrates its propensity. In order to rank the feature importance using the MDGI values, the average MDGI values on 10 runs as derived from RF (i.e., using different values of the *mtry* parameter) were used that is followed by the analysis of the feature importance. Features with the largest MDGI value represents the most important feature. Figures 3 & 4 presents the average MDGI values of 20 amino acids and 400 dipeptides, respectively. Previously, few studies reported that most of the documented peptides with the highest ACE-inhibitory activity were relatively short peptides [24–26]. Therefore, we also evaluate the MDGI values and percent composition of 20 amino acids on four different dataset, in other words, AHTP2132, tetra-, penta- and hexa- peptides, as summarized in Table 4.

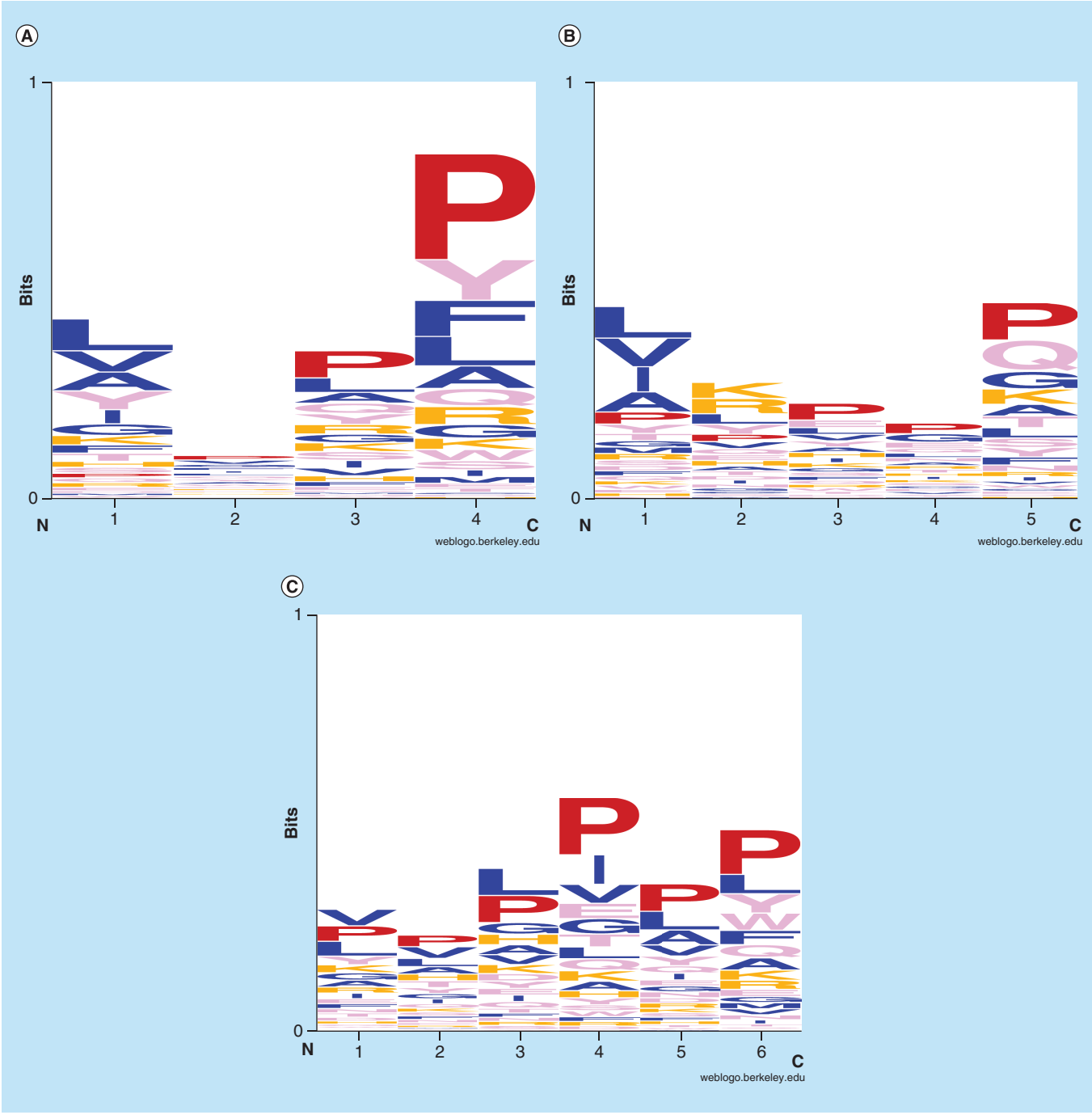

Figure 3. Sequence logo representations of AHTPs. Shown are the sequence logo of tetra- **(A)**, penta- **(B)** and hexa- **(C)** peptides. Colors are: green for positively charged (K, R, H), blue for hydrophobic (A, L, I, M, F, V, C, G), grey for glycine (G) and pink for the remaining amino acids (N, D, E, Q, K, S, T, W, Y).

Important AAC for AHT activity

Previously, many studies made their efforts to utilize the AAC feature for characterizing the biochemical and biophysical properties of various proteins and peptides because of their ability for modulating bioactivities [28–30]. Table 4 lists the 20 MDGI values and percent composition of amino acids for AHTPs. It is better seen the total number of amino acids in tetra-, penta- and hexapeptides datasets are 7685, 612, 1350 and 1194, respectively. The MDGI values and percent composition of amino acids can be analyzed to further gain insight

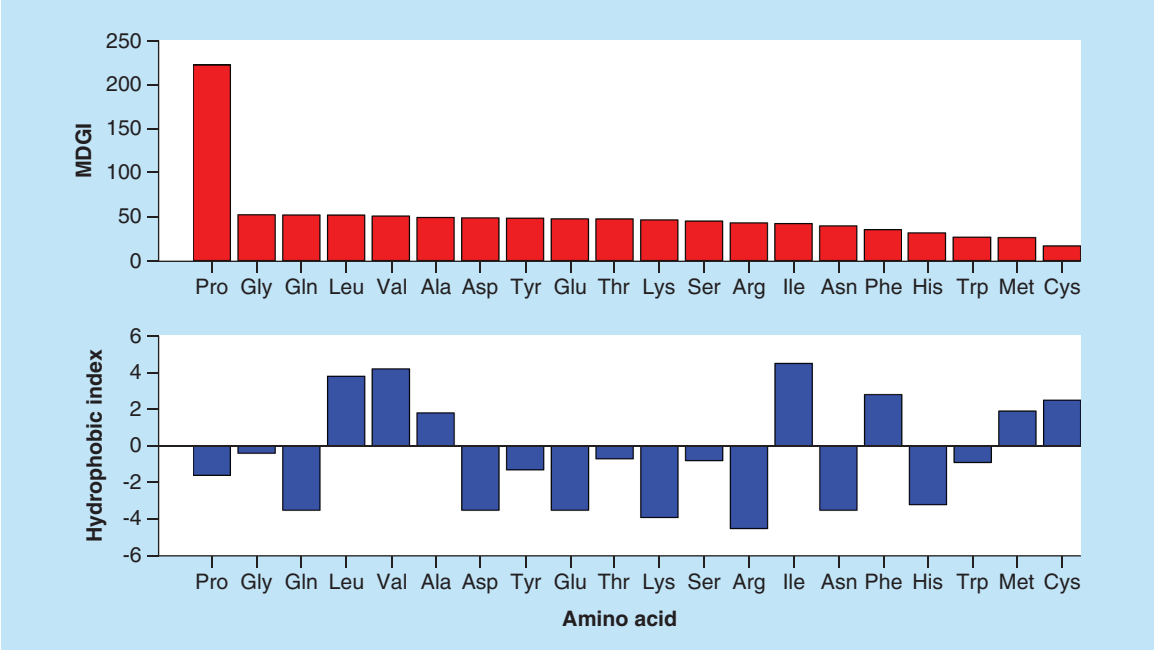

Figure 4. Plot of mean decrease of Gini index and hydrophobic index.

Table 4. Amino acid composition (percent composition; % com) and mean decrease of Gini index values of 20 amino acids as derived from four different datasets (AHTP2132, tetra-, penta- and hexapeptides).

Amino acid	AHTP2132		Tetrapeptide		Pentapeptide		Hexapeptide	
	% com	MDGI	% com	MDGI	% com	MDGI	% com	MDGI
Ala	5.47 (6)	49.14 (6)	7.52 (4)	6.89 (6)	6.15 (6)	13.24 (8)	6.53 (4)	6.89 (15)
Arg	3.80 (14)	42.95 (13)	4.74 (10)	5.92 (11)	5.48(10)	13.77 (7)	3.35 (14)	8.09 (10)
Asn	3.19 (15)	39.53 (15)	1.47 (19)	6.34 (10)	3.19 (15)	9.23 (16)	2.85 (15)	7.87 (12)
Asp	2.38 (17)	48.59 (7)	1.47(17)	5.13 (16)	2.67(17)	12.11(12)	2.26 (18)	9.30 (5)
Cys	0.68 (20)	16.71 (20)	0.65(20)	1.88(20)	1.78(18)	6.96(19)	0.50 (20)	1.98 (20)
Gln	5.76 (5)	51.93 (3)	5.39(9)	7.99(3)	6.81 (3)	15.34(3)	4.52 (9)	7.50 (13)
Glu	4.81 (11)	47.47 (9)	2.61 (15)	5.73 (12)	4.74 (11)	12.41(10)	4.10 (10)	11.22 (2)
Gly	6.22 (4)	52.13 (2)	5.88 (7)	7.56 (5)	6.15 (7)	16.43 (2)	5.36 (7)	8.16 (9)
His	2.85 (16)	31.61 (17)	2.78 (13)	5.42 (14)	3.56 (14)	10.16 (15)	3.69 (12)	5.48 (17)
Leu	8.98 (2)	51.78 (4)	10.62 (2)	6.38 (9)	8.67 (2)	15.13 (4)	10.72 (2)	9.59 (4)
Ile	5.22 (9)	42.18 (14)	5.39 (8)	6.61 (7)	5.56 (8)	12.20 (11)	5.44 (6)	7.38 (14)
Lys	5.32 (7)	46.28 (11)	4.41(11)	5.14 (15)	6.59 (4)	13.82 (5)	4.77 (8)	8.00 (11)
Met	1.51 (19)	26.29 (19)	1.47(18)	3.01 (19)	1.56 (20)	8.55 (18)	1.17 (19)	4.17 (19)
Phe	4.91 (10)	35.27 (16)	6.21(6)	7.61(4)	2.74 (16)	8.63 (17)	3.60 (13)	4.79 (18)
Pro	**16.06 (1)**	**222.00 (1)**	**15.52 (1)**	**24.41(1)**	**12.52 (1)**	**21.50 (1)**	**17.76 (1)**	**45.48 (1)**
Ser	3.81 (13)	44.97 (12)	3.43 (12)	6.38 (8)	3.70 (13)	11.94 (13)	2.35 (17)	10.37(3)
Thr	3.99 (12)	47.45 (10)	1.96 (16)	3.95 (17)	4.30 (12)	12.99 (9)	3.77 (11)	8.59 (7)
Trp	1.64 (18)	26.77 (18)	2.78 (14)	3.73 (18)	1.70 (19)	6.55 (20)	2.76 (16)	6.13 (16)
Tyr	5.27 (8)	48.24 (8)	8.99 (3)	10.88 (2)	5.56 (9)	10.90 (14)	6.11 (5)	9.22 (6)
Val	8.13 (3)	50.66 (5)	6.70 (5)	5.54 (13)	6.59 (5)	13.78 (6)	8.38 (3)	8.20 (8)

The total number of amino acids of antihypertensive peptides in tetra-, penta- and hexapeptides datasets are, respectively, 7685, 612, 1350 and 1194.
MDGI: Mean decrease of Gini index.

into AHTPs characterization. Features with the largest value of MDGI are considered to be the most important, which corresponds to the amino acid Pro with a MDGI value of 221.99. Interestingly, Figure 4 and Table 4 show that MDGI values, percent composition and sequence logo representations for Pro are dominant over other amino acids with a MDGI value/% composition of 24.41/15.52, 21.50/12.52 and 45.48/17.76 for the datasets of tetra-, penta- and hexa- peptides, respectively.

Many recent articles and book chapters have reported that Pro is well-documented as the most favorable amino acid in ACE inhibitory or AHTPs. In 1999, Yamamoto *et al.* [31] demonstrated that Pro can resist degradation by digestive enzymes and may pass from the small intestine into the blood circulation in the sequence of short peptides. This hypothesis was confirmed by the work of Pan *et al.* [32]. This work purified AHT tripeptides Val-Pro-Pro (VPP) and lle-Pro-Pro (IPP) and showed their significant increase in ACE-inhibitory activity at the 5% significance level than other reported ACE inhibitory peptides. Until now, VPP and IPP have been recognized as the best ACE inhibitory tripeptides [25,33–36], which were often found in the primary structure of bovine β-casein (84–86 Val-Pro-Pro and 74–76 lle-Pro-Pro) and κ-casein (108–110 lle-Pro-Pro) [37]. These two tripeptides were reported their abilities to resist *in vivo* degradation along with exert antihypertensive activity through the inhibition of ACE in aorta [22]. In 2010, Lameu *et al.* [35] showed that the Ile-Pro-Pro motif is relevant for the efficiency of the pit viper Bothrops jararaca (Bj-PROs), which was first reported as natural inhibitors of ACE in inducing [Ca2+]i transients. This study showed the partial loss of the biological activity of Bj-PROs, which can be found by the substitutions of lIe or Pro by Ala at the C-terminal region.

The importance of specific amino acids and biochemical properties for AHTPs has been repeatedly suggested. Many studies reported that one of the crucial factors affecting the great activity of ACE-inhibitory or AHTPs are the amino acid sequence of a peptide at C-terminal [31,38]. Experimental and computational studies derived from different ACE-inhibitory peptides have revealed that binding to ACE is strongly influenced by the C-terminal tripeptide sequence of the substrate and the rigid structure of Pro. Manzanares *et al.* [39] and other research groups [40] reported that the superiority of Pro at the C-terminus is probably due to the rigid ring structure of Pro that may lock the carboxyl group into a conformation favorable for interacting with the positively charged residue at the active site of the enzyme. This exceptional conformational property of Pro affects the stability of secondary structures. That is precisely the reason why Pro is preferably located at the end of peptide structures rather than at other positions [24,31,38,40]. In 2015, the study of molecular docking by Arcanjo *et al.* [36] explained that Pro-rich oligopeptide BPP interacts with the C-domain active site of ACE via competitive inhibition for the substrate entry. In this interaction of peptide and ACE enzyme, the cyclic structure of Pro is engaged with the carboxyl group in a favorable conformation that can interact with positively charged residues on the active site of the enzyme. Hence, it might be stated that Pro took part to stabilize the peptide binding onto the active site of ACE.

As mentioned above, the majority of ACE-inhibitory peptides are relatively peptides up to six amino acids in length and Pro at the C-terminus. In addition, many research groups repeatedly reported the important biochemical and biophysical properties at C-terminus that were potent with respect to ACE inhibitions. Most of the documented amino acids and their biochemical properties are usually hydrophobicity and positively charged amino acids at C-terminus [41–44]. Our analysis results show that the important hydrophobic amino acids derived from the MDGI values (rank) on the datasets of tetra-, penta- and hexapeptides are Gly (5, 2 and 9), Leu (9, 4 and 4) and Val (13, 6 and 8), respectively, while their ranks of %composition are (7, 7 and 7), (2, 2 and 2) and (5, 5 and 3) for Gly, Leu and Val, respectively. In 1998, Pihlanto-Leppälä *et al.* [45] determined ACE-inhibitory activities of milk whey and casein proteins. This study reported that peptides with the highest ACE-inhibitory activity was found in casein peptides β-cn f108–113 and α_{s1} -cn f157–164, as they contained Pro residues. Meanwhile, β-1g f15–20 and α-la f105–110 peptides containing hydrophobic amino acid and β- and α- latorphins having Leu at C-terminus were shown to have ACE-inhibitory activities. In addition, a couple of QSAR studies explained in more details that a peptide with hydrophobicity and positively charged amino acids on its C-terminal end is more likely to inhibit the ACE enzyme [22,41,42,46]. Pripp *et al.* [46] utilized QSAR modeling to determine the interpretation of the relationship between physicochemical properties and the activity of ACE-inhibition. Based on the relationship result with R = 0.73, the suggested properties are hydrophobicity and positive charge amino acids at C-terminal region for peptides containing up to six amino acids. In 2006, Wu *et al.* [22] also confirmed that the hydrophobicity and positive charge amino acids were preferred for amino acid at C-terminus, especially in di- and tripeptides. In conclusion, the feature importance analysis of the AAC feature using the MDGI value and the percentage composition of amino acids showed that that Pro residue and hydrophobic amino acid like Gly, Leu and Val at

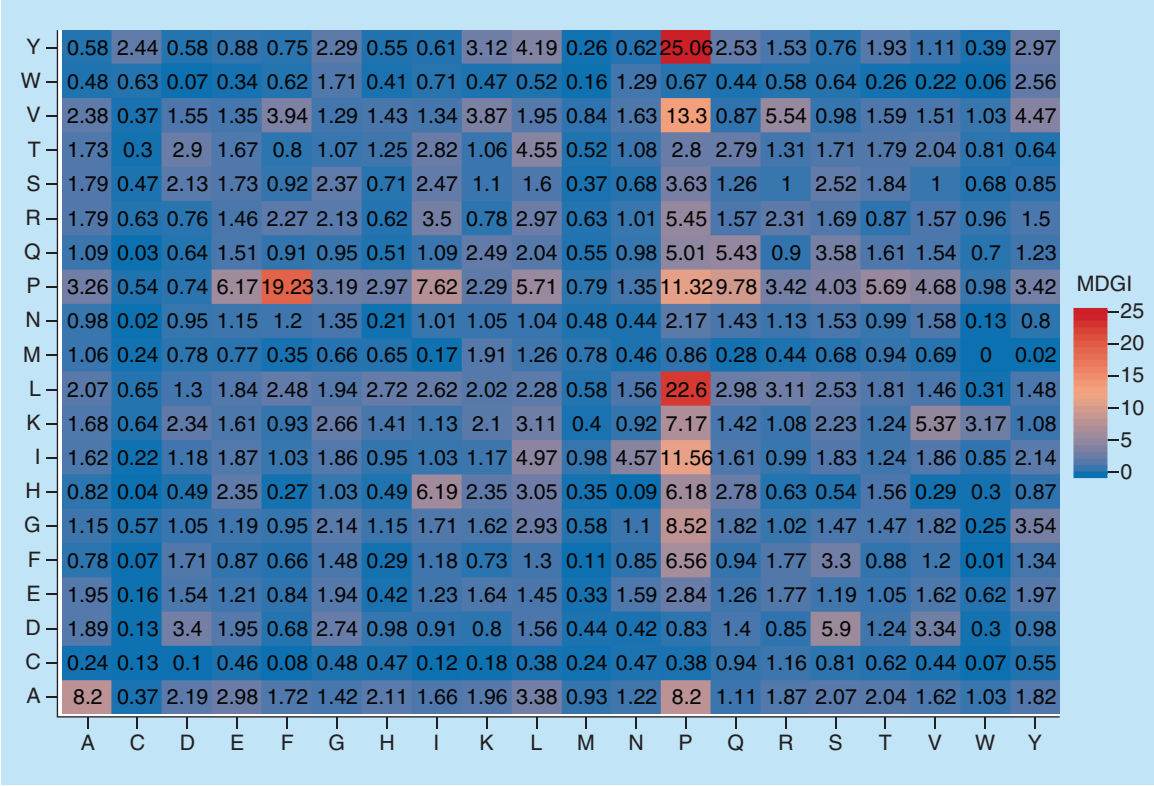

Figure 5. Heat map of the mean decrease of Gini index of dipeptide compositions.

the C-terminal end of peptide containing up to six amino acid have a critical role in the improvement activity of ACE-inhibitory and AHTPs.

Important DPC for AHT activity

The heatmap of feature importance for DPC feature is shown in Figure 5. The top-five informative dipeptides are YP, LP, PF, VP and IP having an MDGI value larger than 10. Interestingly, 9 out of 10 top-ranked informative dipeptides consist of at least one Pro, while PP is important at rank 6. This analysis result reinforces the importance of Pro on the antihypertensive activity of peptides. In fact, Pro containing dipeptides show a significant role in regulating the bioactivity of proteins via the formation of *cis*/*trans* isomers [47] allowing these Pro containing dipeptides further influence on the enzyme's inhibition activity of AHTPs [48].

As seen in Figure 5, 3 out of 5 top-ranked informative dipeptides consist of Pro residue and a large hydrophobic amino acid, in other words, Leu, Val and lIe. In 2006, Wu *el al.* [22] performed a QSAR model to model ACE-inhibitory peptides consisting of 168 dipeptides for predicting their potent ACE inhibitory. This study confirmed our analysis result that Pro is the most favorable amino acid for binding to ACE, while peptide with a highly ACE-inhibition should have a large hydrophobic amino acid. Furthermore, 13 of the top-ranked DPC are made up of two hydrophobic amino acids and while the rest of top DPC are composed of one hydrophobic residue. This finding was supported by the existing literature indicating that the interactions of peptides to the active site of ACE is governed by their hydrophobic property. Hence, hydrophobic residues are essential for the induction of potent inhibitory activity by the AHTP toward ACE enzyme [49,50]. Moreover, many of the ACE inhibitory dipeptides contain hydrophobic side chains that influence the ACE inhibitory activity [22,51,52]. Therefore, it can be inferred that the composition of Pro and hydrophobic residues in a peptide chain plays a key role for antihypertensive activity and dipeptides comprised of a combination of hydrophobic residues are more likely to be antihypertensive.

Previously, many studies on the effects of skim milk in the reduction of blood pressure have been studied, both in terms of intervention studies [53–55] and clinical trials [56–58]. Additionally, increased protein intake has also been associated with attenuated blood pressure [59]. Thus, the bioactive peptides found in milk fermented with *Lactobacillus helveticus* in which the tripeptide Val-Pro-Pro (VPP) have been associated with antihypertensive

properties as well as immunostimulating effects [60]. This in turn, exhibits ACE-inhibitory activity and thus can attenuate blood pressure [61]. Moreover, clinical trials to test the antihypertensive effects of the two main peptides found in sour milk, such as VPP and Ile-Pro-Pro (IPP), were conducted in Japan [62] and Finland [63]. After 8 weeks of the trial whereby patients consumed either VPP, IPP or placebo daily, it was observed that the group that consumed milk (containing VPP or IPP) exhibited a drastic decrease in systolic blood pressure between the 4 and 8 weeks of trial [25,62,63]. Such clinical trials have also been conducted and consequently meta-analysis has been conducted. For example, meta-analysis conducted by Xu *et al.* [64] and Pripp *et al.* [65] made use of data from 12 to 15 clinical trials, respectively, in which both of these studies observed significant decrease in systolic and diastolic blood pressures. Moreover, Siltari *et al.* [48] computationally docked the above mentioned tripeptides, VPP and IPP, in order to elucidate their inhibition capabilities against ACE-1 enzyme. Their findings indicated that the *cis* conformation of these casein-derived tripeptides were more significantly potent in their ACE-1 inhibition activity. This information can be valuable as nonpharmacological treatment in terms of ACE inhibition and blood pressure lowering capacity.

PAAP web server

For convenience of scientific community, a user-friendly and publicly accessible web server for the proposed model PAAP (based on the AHTP2132 dataset) has been established using the Shiny package under the R programming environment. As mentioned above, the best prediction performance was achieved by using the RF model cooperating with the combination feature of AAC + DPC. However, in practice, the length of this combination feature is a 420-dimensional vector, practically resulting in a time-consuming procedure for both constructing the prediction model and predicting the desired peptide sequence. As to reduce the computational time, we thus selected the PseAAC feature, which were found to be comparable with the aforementioned feature, for constructing the PAAP web server. Screenshots of the PAAP web server is shown in Figure 6.

A step-by-step guide on how to use the PAAP web server is given below:

Step 1. Go to the PAAP web server at http://codes.bio/paap/

Step 2. Enter the query sequence into the Input box or upload the sequence file by clicking on the *Choose file* button (i.e., found below the *Enter your input sequence(s) in FASTA format* heading). Finally, press on the *Submit* button to initiate the prediction process.

Step 3. Prediction results are automatically displayed in the grey box found below the *Status/Output* heading. Typically, it takes a few seconds for the server to process the task. Users can also download the prediction results as a CSV file by pressing on the *Download CSV* button.

Additionally, users could also run a local copy of PAAP on their own computer using a one-line code as follows in an R environment:

```
shiny::runGitHub('PAAP', 'chaninlab', subdir = "paap_shiny_server")
```

However, prior to running the aforementioned code, it is recommended that users first install the prerequisite R packages. This can be performed by using the following code:

```
install.packages(c('shiny', 'shinyjs', 'shinythemes', 'protr', 'seqinr', 'randomForest', 'markdown'))
```

Conclusion

Prediction and characterization of antihypertensive or ACE inhibitory peptides are valuable for both basic research and drug development. To the best of our knowledge, there is only one study focusing on predicting AHTPs. Therefore, in this study, a new computational model, called PAAP, was established for predicting/analyzing AHTPs and saving a lot of cost and time. In the development of PAAP, the RF algorithm was used as the prediction engine enhancing with AAC, dipeptide composition and PseAAC composition. The prediction results obtained by rigorous LOO-CV, ten-fold cross-validation and independent dataset test have indicated that our proposed model PAAP is indeed promising and able to complementary role to the existing previous model. Additionally, the feature importance analysis can support the identification of important biochemical and biophysical properties to get insights into the role of the activity of AHTP.

Future perspective

The global increase in hypertension cases coupled with their related risk of stokes and other diseases, warrants the exploration of better therapeutic options. Furthermore, although a number of drugs are currently in use for the

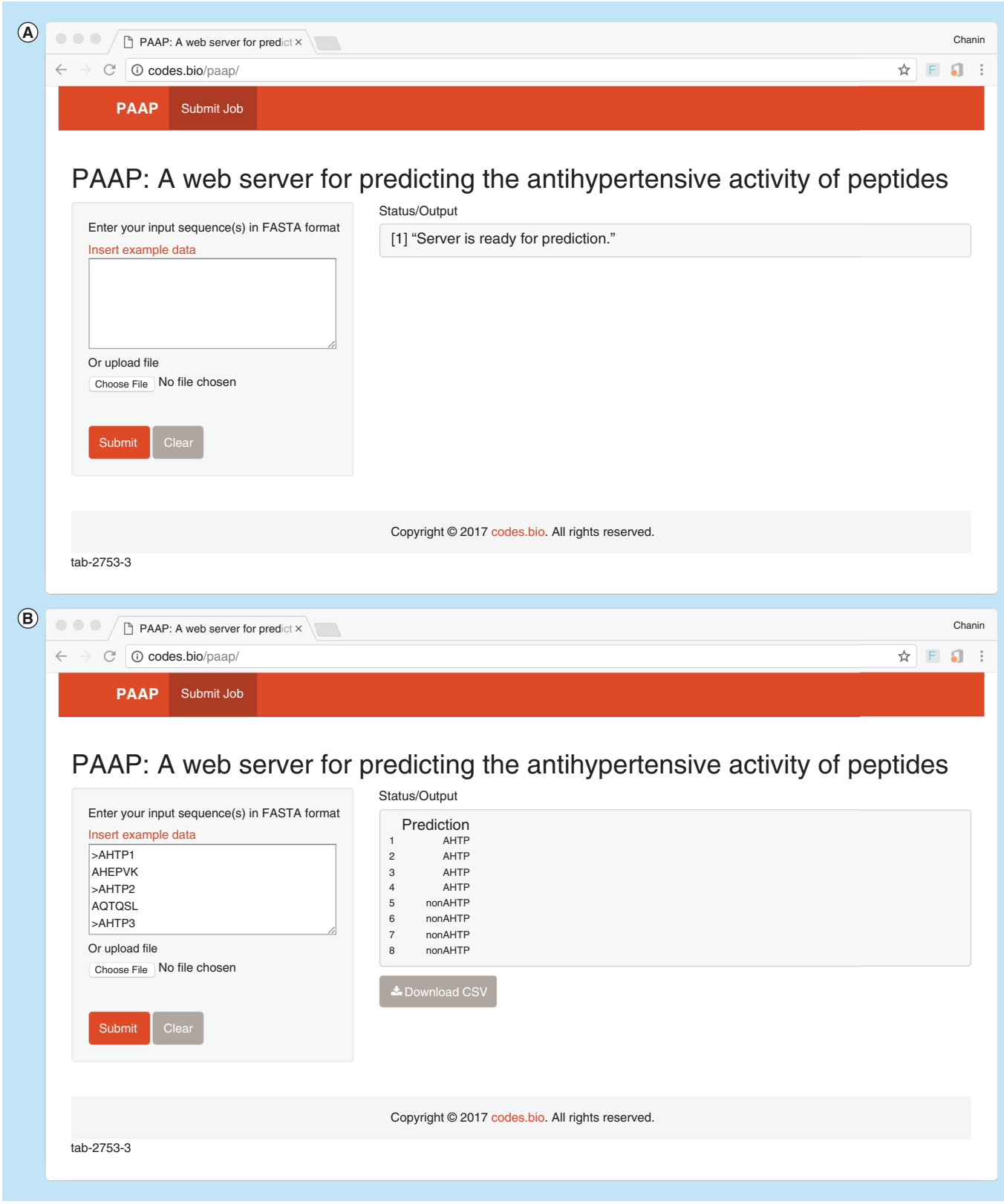

Figure 6. Screenshots of the predictor of antihypertensive activity of peptides web server. Shown are the screenshots before **(A)** and after **(B)** submission of input peptide sequence(s).
PAAP: Predictor of antihypertensive activity of peptide.

control of hypertension, their notable side effects and suboptimal control of this metabolic disease greatly warrants the discovery of new drugs. In this regard, bioactive peptides or AHTPs from natural sources have emerged as a therapeutic option. Thus, in order to help escalate the slow process of drug discovery, a reliable, accurate and effective computational approach for the identification of AHTPs is the need of the hour. Herein, we have provided a computational tool, known as PAAP for the rapid identification of ACE inhibition via discrimination of bioactive peptides into AHTPs and non-AHTPs using large-scale data analysis. Building upon past studies, the present study utilized RF algorithm cooperating with amino acid, dipeptide and PseAAC compositions for the development of a more effective and efficient prediction model. This study also sheds light on the biochemical and biophysical properties that can improve the activity of AHTPs. In addition, the prediction results obtained by rigorous cross-validations (LOO-CV, ten-fold CV and independent test), indicate that our proposed model PAAP is indeed promising and able to complementary role to the existing previous model. Most importantly, a free web-server is publicly available at https://codes.bio/paap/ for the prediction and discrimination of AHTPs.

Summary points

- Owing to labor-intensive and cost of experimental approaches, it is highly desirable to develop computational methods to timely identify antihypertensive activity of peptides.
- Correct identification of antihypertensive activity of peptides is important for supporting both basic research and drug development.
- The highest performance was achieved via the use of random forest as a function of amino acid and dipeptide compositions.
- Proline and hydrophobic amino acids at the C-terminal have been revealed to be crucial for improving angiotensin-converting enzyme-inhibitory activity.
- Almost all angiotensin-converting enzyme-inhibitory peptides with the highest activity were relatively short (2–6 amino acids).
- Predictor of antihypertensive activity of peptides web server was developed for lowering the barriers of access to the proposed predictive model described herein.

Financial & competing interests disclosure

This work is supported by the Office of Higher Education Commission and the Thailand Research Fund (No. MRG6180226); the New Researcher Grant (A31/2561) from Mahidol University; and the Center of Excellence on Medical Biotechnology (CEMB), S&T Postgraduate Education and Research Development Office (PERDO), Office of Higher Education Commission (OHEC), Thailand. Partial support from the annual budget grant (B.E.2557-2559) of Mahidol University is also acknowledged. The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as: •• of considerable interest

1. Varounis C, Katsi V, Nihoyannopoulos P, Lekakis J, Tousoulis D. Cardiovascular hypertensive crisis: recent evidence and review of the literature. *Front. Cardiovasc. Med.* 3, 51 (2016).
2. Ikeda N, Sapienza D, Guerrero R *et al.* Control of hypertension with medication: a comparative analysis of national surveys in 20 countries. *Bull. World Health Organ.* 92(1), C10–C19 (2014).
3. Loga-Zec S, Asceric M, Loga-Andrijic N, Kapetanovic B, Zerem E. The incidence of antihypertensive drug-induced side effects in patients with diabetes mellitus Type 2 and hypertension. *Med. Arch.* 68(6), 372–375 (2014).
4. Husserl FE, Messerli FH. Adverse effects of antihypertensive drugs. *Drugs* 22(3), 188–210 (1981).
5. Kovell LC, Ahmed HM, Misra S *et al.* US hypertension management guidelines: a review of the recent past and recommendations for the future. *J. Am. Heart Assoc.* 4(12), pii:e002315 (2015) (Epub ahead of print).
6. Bernstein KE, Giani JF, Shen XZ, Gonzalez-Villalobos RA. Renal angiotensin-converting enzyme and blood pressure control. *Curr. Opin. Nephrol. Hypertens.* 23(2), 106–112 (2014).
7. Wang X, Wang J, Lin Y *et al.* QSAR study on angiotensin-converting enzyme inhibitor oligopeptides based on a novel set of sequence information descriptors. *J. Mol. Model.* 17(7), 1599–1606 (2011).

•• **Creates benchmark datasets and develops the AHTPIN prediction model for predicting antihypertensive peptides.**

8. Kumar R, Chaudhary K, Singh Chauhan J *et al.* An *in silico* platform for predicting, screening and designing of antihypertensive peptides. *Sci. Rep.* 5, 12512 (2015).
9. Chou K-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* 273(1), 236–247 (2011).
10. Xiao N, Cao D-S, Zhu M-F, Xu Q-S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* 31(11), 1857–1859 (2015).
11. Win TS, Malik AA, Prachayasittikul V, S Wikberg JE, Nantasenamat C, Shoombuatong W. HemoPred: a web server for predicting the hemolytic activity of peptides. *Future Med. Chem.* 9(3), 275–291 (2017).
12. Simeon S, Shoombuatong W, Anuwongcharoen N *et al.* osFP: a web server for predicting the oligomeric states of fluorescent proteins. *J. Cheminform.* 8, 72 (2016).
13. Simeon S, Anuwongcharoen N, Shoombuatong W *et al.* Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking. *PeerJ* 4, e2322 (2016).
14. Anuwongcharoen N, Shoombuatong W, Tantimongcolwat T, Prachayasittikul V, Nantasenamat C. Exploring the chemical space of influenza neuraminidase inhibitors. *PeerJ* 4, e1958 (2016).
15. Shoombuatong W, Prachayasittikul V, Prachayasittikul V, Nantasenamat C. Prediction of aromatase inhibitory activity using the efficient linear method (ELM). *EXCLI J.* 14, 452–464 (2015).
16. Shoombuatong W, Prachayasittikul V, Anuwongcharoen N *et al.* Navigating the chemical space of dipeptidyl peptidase-4 inhibitors. *Drug Des. Devel. Ther.* 9, 4515–4549 (2015).
17. Shoombuatong W, Hongjaisee S, Barin F, Chaijaruwanich J, Samleerat T. HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees. *Comput. Biol. Med.* 42(9), 885–889 (2012).

•• **Proposes a powerful method used for not only classification and prediction problems but also for the estimation of feature importance.**

18. Breiman L. Random Forests. *Machine Learning* 45, 5–32 (2001).
19. Pratiwi R, Malik AA, Schaduangrat N *et al.* CryoProtect: a web server for classifying antifreeze proteins from nonantifreeze proteins. *J. Chem.* 2017, 9861752 (2017).
20. Wang L, Huang C, Yang JY. Predicting siRNA potency with random forests and support vector machines. *BMC. Genomics.* 11(Suppl. 3), S2 (2010).
21. Shoombuatong W, Nabu S, Simeon S *et al.* Extending proteochemometric modeling for unraveling the sorption behavior of compound–soil interaction. *Chemometr. Intell. Lab. Syst.* 151, 219–227 (2016).

•• **Develops a comprehensive database built by using experimentally validated peptides for analysis and presentation of antihypertensive peptides.**

22. Wu J, Aluko RE, Nakai S. Structural requirements of angiotensin I-converting enzyme inhibitory peptides: quantitative structure-activity relationship study of di- and tripeptides. *J. Agric. Food Chem.* 54(3), 732–738 (2006).
23. Kumar R, Chaudhary K, Sharma M *et al.* AHTPDB: a comprehensive platform for analysis and presentation of antihypertensive peptides. *Nucleic Acids Res.* 43, D956–D962 (2015).
24. Hernández-Ledesma B, del Mar Contreras M, Recio I. Antihypertensive peptides: production, bioavailability and incorporation into foods. *Adv. Colloid. Interface Sci.* 165(1), 23–35 (2011).
25. Yamamoto N, Ejiri M, Mizuno S. Biogenic peptides and their potential use. *Curr. Pharm. Des.* 9(16), 1345–1355 (2003).
26. Thévenet P, Shen Y, Maupetit J, Guyon F, Derreumaux P, Tufféry P. PEP-FOLD: an updated *de novo* structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic Acids Res.* 40, W288–W293 (2012).
27. DeLano WL. The PyMOL Molecular Graphics System. Schrödinger, LLC., NY, USA. http://pymol.org
28. Siebert KJ. Quantitative structure–activity relationship modeling of peptide and protein behavior as a function of amino acid composition. *J. Agric. Food Chem.* 49(2), 851–858 (2001).
29. Mishra B, Wang G. The importance of amino acid composition in natural amps: an evolutional, structural, and functional perspective. *Front. Immunol.* 3, 221 (2012).
30. Nie L, Siebert KJ. Modeling physicochemical properties and activity of aspartyl proteinases based on amino acid composition. *J. Agric. Food Chem.* 57(6), 2536–2543 (2009).
31. Yamamoto N, Maeno M, Takano T. Purification and characterization of an antihypertensive peptide from a yogurt-like product fermented by *Lactobacillus helveticus* CPN4. *J. Dairy Sci.* 82(7), 1388–1393 (1999).
32. Pan D, Luo Y, Tanokura M. Antihypertensive peptides from skimmed milk hydrolysate digested by cell-free extract of *Lactobacillus helveticus* JCM1004. *Food. Chem.* 91(1), 123–129 (2005).
33. Nakamura Y, Yamamoto N, Sakai K, Takano T. Antihypertensive effect of sour milk and peptides isolated from it that are inhibitors to angiotensin I-converting enzyme. *J. Dairy Sci.* 78(6), 1253–1257 (1995).

34. Sipola M, Finckenberg P, Korpela R, Vapaatalo H, Nurminen M-L. Effect of long-term intake of milk products on blood pressure in hypertensive rats. *J. Dairy Res.* 69(1), 103–111 (2002).
35. Lameu C, Hayashi MAF, Guerreiro JR *et al.* The central nervous system as target for antihypertensive actions of a proline-rich peptide from *Bothrops jararaca* venom. *Cytometry A* 77(3), 220–230 (2010).
36. Arcanjo DDR, Vasconcelos AG, Comerma-Steffensen SG *et al.* A novel vasoactive proline-rich oligopeptide from the skin secretion of the frog *Brachycephalus ephippium*. *PLoS ONE* 10(12), e0145071 (2015).
37. Jäkälä P, Vapaatalo H. Antihypertensive peptides from milk proteins. *Pharmaceuticals (Basel)* 3(1), 251–272 (2010).

•• **Describes the role of proline in increasing angiotensin-converting enzyme-inhibitory activity.**

38. Nakamura Y, Yamamoto N, Sakai K, Okubo A, Yamazaki S, Takano T. Purification and characterization of angiotensin I-converting enzyme inhibitors from sour milk. *J. Dairy Sci.* 78(4), 777–783 (1995).
39. Manzanares P, Salom JB, García-Tejedor A, Fernández-Musoles R, Ruiz-Giménez P, Gimeno-Alcañíz JV. Unraveling the mechanisms of action of lactoferrin-derived antihypertensive peptides: ACE inhibition and beyond. *Food Funct.* 6(8), 2440–2452 (2015).

•• **Demonstrates the importance of amino acid position at the C-terminal end to angiotensin-converting enzyme-inhibitory potency.**

40. Robert M-C, Razaname A, Mutter M, Juillerat MA. Identification of angiotensin-I-converting enzyme inhibitory peptides derived from sodium caseinate hydrolysates produced by *Lactobacillus helveticus* NCC 2765. *J. Agric. Food. Chem.* 52(23), 6923–6931 (2004).
41. Pripp AH. Initial proteolysis of milk proteins and its effect on formation of ACE-inhibitory peptides during gastrointestinal proteolysis: a bioinformatic, in silico, approach. *Eur. Food. Res. Technol.* 221(5), 712–716 (2005).
42. He R, Ma H, Zhao W *et al.* Modeling the QSAR of ACE-inhibitory peptides with ANN and its applied illustration. *Int. J. Pept.* 2012, 620609 (2012).
43. Mullally MM, Meisel H, FitzGerald RJ. Synthetic peptides corresponding to a-lactalbumin and b-lactoglobulin sequences with angiotensin-1-converting enzyme inhibitory activity. *Biol. Chem. Hoppe-Seyler.* 377(4), 259–260 (1996).
44. Pihlanto-Leppala A, Rokka T, Korhonen H. Angiotensin I converting enzyme inhibitory peptides derived from bovine milk proteins. *Int. Dairy J.* 8(4), 325–331 (1998).
45. Pihlanto-Leppälä A, Koskinen P, Piilola K, Tupasela T, Korhonen H. Angiotensin I-converting enzyme inhibitory properties of whey protein digests: concentration and characterization of active peptides. *J. Dairy Res.* 67(1), 53–64 (2000).
46. Pripp AH, Isaksson T, Stepaniak L, Sørhaug T. Quantitative structure–activity relationship modelling of ACE-inhibitory peptides derived from milk proteins. *Eur. Food Res. Technol.* 219(6), 579–583 (2004).
47. Gebauer S, Friebe S, Scherer G, Gubitz G, Krauss G-J. High performance liquid chromatography on calixarene-bonded silica gels. III. Seperations of cis/trans isomers of proline-containing peptides. *J. Chromatogr. Sci.* 36, 388–394 (1998).
48. Siltari A, Viitanen R, Kukkurainen S, Vapaatalo H, Valjakka J. Does the cis/trans configuration of peptide bonds in bioactive tripeptides play a role in ACE-1 enzyme inhibition? *Biologics* 8, 59–65 (2014).
49. Li G-H, Le G-W, Shi Y-H, Shrestha S. Angiotensin I– converting enzyme inhibitory peptidesderived from food proteins and their physiological andpharmacological effects. *Nutr. Res.* 24, 469–486 (2004).
50. Kohmura M, Nio N, Kubo K, Minoshima Y, Munekata E, Ariyoshi Y. Inhibition of angiotensin-converting enzyme by synthetic peptides of human β-casein. *Agric. Biol. Chem.* 53(8), 2107–2114 (1989).
51. Pripp AH, Ardo Y. Modelling relationship between angiotensin-(I)-converting enzyme inhibition and the bitter taste of peptides. *Food Chem.* 102(2007), 880–888 (2006).
52. Wu S, Qi W, Su R, Li T, Lu D, He Z. CoMFA and CoMSIA analysis of ACE-inhibitory, antimicrobial and bitter-tasting peptides. *Eur. J. Med. Chem.* 84, 100–106 (2014).
53. Pereira MA, Jacobs DR, Van Horn L, Slattery ML, Kartashov AI, Ludwig DS. Dairy consumption, obesity, and the insulin resistance syndrome in young adults: The CARDIA study. *JAMA* 287, 2081–2089 (2002).
54. Appel LJ, Moore TJ, Obarzanek E *et al.* A clinical trial of the effects of dietary patterns on blood pressure. *N. Engl. J. Med.* 336, 1117–1124 (1997).
55. van Beresteijn EC, van Schaik M, Schaafsma G. Milk: does it affect blood pressure? A controlled intervention study. *J. Intern. Med.* 228, 477–482 (1990).
56. Alonso A, Steffen LM, Folsom AR. Dairy intake and changes in blood pressure over 9 years: the ARIC study. *Eur. J. Clin. Nutr.* 63, 1272–1275 (2009).
57. Wang L. Dietary intake of dairy products, calcium, and vitamin D and the risk of hypertension in middle-aged and older women. *Hypertension* 51, 1073–1079 (2008).
58. Toledo E, Delgado-Rodríguez M, Estruch R *et al.* Low-fat dairy products and blood pressure: follow-up of 2290 older persons at high cardiovascular risk participating the PREDIMED study. *Br. J. Nutr.* 101, 59–67(2009).
59. He J, Whelton PK. Effect of dietary fiber and protein intake on blood pressure: a review of epidemiological evidence. *Clin. Exp. Hypertens.* 21, 785–796 (1999).

60. Meisel H, Bockelmann W. Bioactive peptides encrypted in milk proteins: proteolytic activation and thropho-functional properties. *Antonie Van Leeuwenhoek* 76, 207–215 (1999).

61. Sipola M, Finckenberg P, Santisteban J, Korpela R, Vapaatalo H, Nurminen ML. Long term intake of milk peptides attenuates development of hypertension in spontaneously hypertensive rats. *J. Physiol. Pharmacol.* 52, 745–754 (2001).

62. Hata Y, Yamamoto M, Ohni M, Nakajima K, Nakamura Y, Takano T. A placebo-controlled study of the effect of sour milk on blood pressure in hypertensive subjects. *Am. J. Clin. Nutr.* 64, 767–71 (1996).

63. Seppo L, Kerojoki O, Suomalainen T, Korpela R. The effect of a *Lactobacillus helveticus* LBK-16 H fermented milk on hypertension - a pilot study on humans. *Milchwissenschaft* 57, 124–127 (2002).

64. Xu J-Y, Qin L-Q, Wang P-Y, Li W, Chang C. Effect of milk tripeptides on blood pressure: a meta-analysis of randomized controlled trials. *Nutrition* 28, 933 (2008).

65. Pripp AH. Effect of peptides derived from food proteins on blood pressure: a meta-analysis of randomized controlled trials. *Food Nutr. Res.* 52, 1 (2008).

cells

MDPI

Article

PVPred-SCM: Improved Prediction and Analysis of Phage Virion Proteins Using a Scoring Card Method

Phasit Charoenkwan [1], Sakawrat Kanthawong [2], Nalini Schaduangrat [3], Janchai Yana[4] and Watshara Shoombuatong [3,*]

[1] Modern Management and Information Technology, College of Arts, Media and Technology, Chiang Mai University, Chiang Mai 50200, Thailand; phasit.c@cmu.ac.th
[2] Department of Microbiology, Faculty of Medicine, Khon Kaen University, Khon Kaen 40002, Thailand; sakawrat@kku.ac.th
[3] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand; nalini.sch@mahidol.edu
[4] Department of Chemistry, Faculty of Science and Technology, Chiang Mai Rajabhat University, Chiang Mai 50300, Thailand; janchai@g.cmru.ac.th
* Correspondence: watshara.sho@mahidol.ac.th; Tel.: +66-2-441-4371(ext. 2715)

Received: 26 December 2019; Accepted: 27 January 2020; Published: 3 February 2020

Abstract: Although, existing methods have been successful in predicting phage (or bacteriophage) virion proteins (PVPs) using various types of protein features and complex classifiers, such as support vector machine and naïve Bayes, these two methods do not allow interpretability. However, the characterization and analysis of PVPs might be of great significance to understanding the molecular mechanisms of bacteriophage genetics and the development of antibacterial drugs. Hence, we herein proposed a novel method (PVPred-SCM) based on the scoring card method (SCM) in conjunction with dipeptide composition to identify and characterize PVPs. In PVPred-SCM, the propensity scores of 400 dipeptides were calculated using the statistical discrimination approach. Rigorous independent validation test showed that PVPred-SCM utilizing only dipeptide composition yielded an accuracy of 77.56%, indicating that PVPred-SCM performed well relative to the state-of-the-art method utilizing a number of protein features. Furthermore, the propensity scores of dipeptides were used to provide insights into the biochemical and biophysical properties of PVPs. Upon comparison, it was found that PVPred-SCM was superior to the existing methods considering its simplicity, interpretability, and implementation. Finally, in an effort to facilitate high-throughput prediction of PVPs, we provided a user-friendly web-server for identifying the likelihood of whether or not these sequences are PVPs. It is anticipated that PVPred-SCM will become a useful tool or at least a complementary existing method for predicting and analyzing PVPs.

Keywords: phage virion protein; scoring card method; propensity score; interpretable model; physicochemical properties; machine learning

1. Introduction

The existence of viruses that can infect and multiply only in bacteria, known as bacteriophages (or phages or bacterial virus), can be found in several environments such as soil, freshwater, and marine. Infectious phage particles are composed of a nucleic acid molecule in the form of either DNA or RNA that is surrounded by a protein coat (capsid) [1]. These bacterial viruses are very species-specific with regards to their host, which is a single bacterial strain or species. Bacteriophages irreversibly attach themselves to the surface of a susceptible host, insert their genetic information and persist using two possible strategies: lytic or lysogenic life cycle [2]. Due to their characteristics, lack of toxicity for human cells, harmless to non-specific microbes (e.g., normal flora), and their potential

against antibiotic-resistant bacteria, bacteriophages are promising as alternatives to antibiotics [3]. Phage structural (virion) proteins (PVPs) including capsid proteins (coat proteins), tail proteins and phage particle enzymes, play important roles in the interaction between themselves and its host bacteria, which can be used to develop a new class of antimicrobial agents [4]. Previously, many experimental methods, such as mass spectrometry, sodium dodecyl sulfate polyacrylamide gel electrophoresis, and protein array [5–7], have been used for the identification of PVPs and non-PVPs. Although, the aforementioned methods are the gold standard and display high accuracy, they are too laborious and costly to be applied for large-scale analysis of PVPs from the sequence information. Hence, it is necessary to develop a computational model not only for discriminating PVPs from non-PVPs, but also for characterizing the biochemical and biophysical properties of PVPs.

Recently, many researchers have exploited various types of machine learning (ML) algorithms using sequence features to directly predict PVPs including Seguritan et al.'s method [8], Feng et al.'s method [9], PVPred [10], Zhang et al.'s method [11], PVP-SVM [12], PhagePred [13], Tan et al.'s method [14], Ru et al.'s method [15], and Pred-BVP-Unb [16], as summarized in Table 1. In 2012, Seguritan et al. [8] proposed the first predictor to identify viral structural proteins using an artificial neural network cooperating with a feature combination of amino acid composition (AAC) and protein isoelectric points. Shortly after, Feng et al. [9] established a benchmark dataset consisting of experimentally confirmed 99 PVPs and 208 non-PVPs, and developed a PVP predictor using the naïve Bayes (NB) method with AAC and dipeptide composition (DPC). Feng et al.'s method yielded an accuracy of 79.15% as assessed by the jackknife test. In the same research group [9], they also developed a new PVP predictor named PVPred [10] using support vector machine (SVM) with g-gap dipeptide composition (g-gap DPC). Moreover, this work constructed the first independent dataset consisting of 11 PVPs and 19 non-PVPs.

Until now, the state-of-the-art PVP predictors include PVP-SVM [12] and Pred-BVP-Unb [16] of which both provide high prediction accuracies of 79.80% and 83.06%, respectively, as evaluated by the independent validation test. PVP-SVM was developed by utilizing many types of sequence features, i.e., AAC, DPC, atomic composition, chain-transition-distribution and physicochemical properties, as input feature and SVM as the prediction engine. This study also gave a high number of PVPs and non-PVPs on the independent dataset (30 PVPs and 64 non-PVPs) by gathering protein sequences from the works of Ding et al. [10] and the Universal Protein Resource (UniProt). More recently, Arif et al. proposed a novel method called Pred-BVP-Unb [16] for discriminating PVPs from non-PVPs. Pred-BVP-Unb was developed by using the SVM model in conjunction with three types of protein features, i.e., composition and translation, split amino acid composition, and bi-profile position specific scoring matrix. This study also utilized feature selection (recursive feature elimination; RFE) and synthetic minority over-sampling (SMOTE) techniques for identifying an optimal feature set and overcoming the bias of an imbalanced dataset (i.e., 99 PVPs and 208 non-PVPs).

Although, Pred-BVP-Unb and other existing methods have their own merit and yielded encouraging performances with reasonably high prediction accuracies, the overall utility of the existing methods is limited in terms of few points as follows. Firstly, the underlying mechanisms of the investigated bioactivity for PVP-SVM [12] and Pred-BVP-Unb [16] affords limited interpretability for experimental scientists. Owing to the complex architecture of computational models and low interpretable features used in the study, it is not easy to identify and assess which features are beneficial for the biological activities of PVPs. As mentioned in a series of recent publications [17–33] and summarized in several comprehensive review papers [29,34–36], one of the main values of bioinformatics tools should be its ability to provide insight into mechanisms of action under study. Secondly, few existing methods were not assessed using an independent dataset, indicating that these methods might provide misleading results with overestimated accuracy. Finally, no web server was provided for many methods in this area including Seguritan et al.'s method [8], Feng et al.'s method [9], Zhang et al.'s method [11], PhagePred [13], Tan et al.'s method [14], Ru et al.'s method [15], and Pred-BVP-Unb [16]. Although, these methods have demonstrated good prediction performances,

they are not generalized or transferable to researchers with informatics background who can develop in-house prediction models.

Motivated by the above mentioned limitations, we developed a novel PVP predictor named PVPred-SCM, which is an efficient and interpretable method for predicting and analyzing PVPs. PVPred-SCM was developed by using an efficient and interpretable model named the scoring card method (SCM) cooperating with only DPC to estimate the propensity scores of 400 dipeptides for predicting and analyzing PVPs. To the best of our knowledge, this study is the first in which the SCM method has been applied for predicting and analyzing PVPs. In PVPred-SCM, the propensities of 400 dipeptides to be PVPs using the statistical discrimination approach between PVPs and non-PVPs on the benchmark dataset were calculated. To enhance the prediction performance of PVPred-SCM, the propensity scores of all dipeptides were optimized using our customized genetic algorithm. Experimental results over rigorous 10-fold cross-validation and independent validation test demonstrated that PVPred-SCM had promising performance, compared to existing SVM-based methods with various types of protein features. To identify unknown sequences, PVPred-SCM based on the weighted-sum function can easily be adopted using a single threshold for discriminating PVPs from non-PVPs. Furthermore, the propensity scores of amino acids were investigated to elucidate the biochemical and biophysical properties of PVPs by using the correlation coefficient (R) between the propensity scores and the physicochemical properties (PCPs) from AAindex [37]. Finally, PVPred-SCM was developed as a user-friendly and publicly accessible web server that allows robust predictions to be made without the need to develop in-house prediction models.

Table 1. Summary of some existing methods for predicting phage virion proteins.

Method	Classifier[a]	Sequence Feature [b]	Independent Test	Webserver
Seguritan et al.'s method [8]	ANN	AAC, PIP	-	-
Feng et al.'s method [9]	NB	AAC, DPC	-	-
PVPred [10]	SVM	g-gap DPC	✓	✓
PVP-SVM [12]	SVM	AAC, DPC, ATC, CTD, PCP	✓	✓
PhagePred [13]	Multinomial NB	g-gap DPC feature tree	-	✓[b]
Tan et al.'s method [14]	SVM	GDC	✓	-
Pred-BVP-Unb [16]	SVM	CT, SAAC, bi-PSSM	✓	-
PVPred-SCM (This study)	SCM	DPC	✓	✓

[a] ANN: artificial neural network, NB: Naïve Bayes, SCM: scoring card method, SVM: support vector machine. [b] AAC: amino acid composition, ATC: atomic composition, bi-PSSM: bi-profile position specific scoring matrix, CTD: chain-transition-distribution, CT: composition and translation, DPC: dipeptide composition, g-gap DPC: g-gap dipeptide composition, PCP: physicochemical properties, PIP: protein isoelectric points, SAAC: split amino acid composition. [b] The webserver was not functional during our manuscript preparation.

2. Results

In this study, PVPs and non-PVPs were predicted by the proposed method PVPred-SCM. Figure 1 summarizes the workflow of the computational approach of PVPred-SCM. It can be seen that four steps are involved in the development of this method as follows: (i) collecting both benchmark and independent datasets, (ii) calculating init-DPS using a statistical approach, (iii) optimizing init-DPS to obtain opti-DPS using the genetic algorithm (GA), (iv) PVP prediction, and v) PVP characterization using the propensity scores of dipeptides. Finally, to serve easy and rapid classification of query protein sequence, PVPred-SCM was utilized as a free prediction web server for discriminating PVPs and non-PVPs.

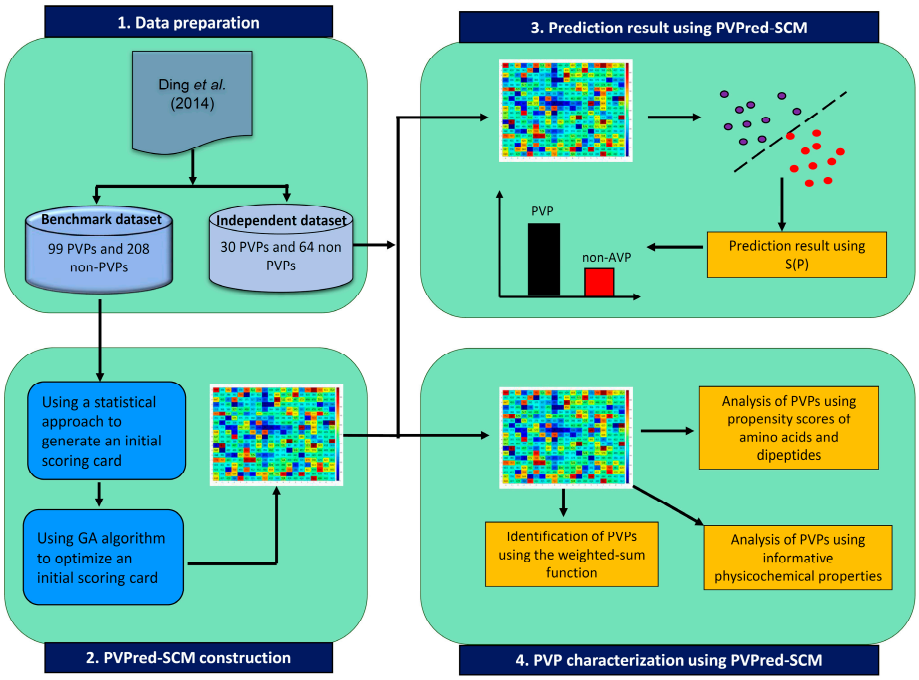

Figure 1. Schematic framework of PVPred-SCM for prediction and analysis of phage virion proteins (PVPs).

2.1. Prediction Performance

In this study, we performed both 10-fold cross-validation and independent validation test on the benchmark and independent datasets to demonstrate the efficient and effective prediction of the proposed method. To make a fair comparison with the existing PVP predictors [9,10,12–14], the same benchmark and independent datasets that have been used in previous studies [12] were used to develop our proposed model. Due to the non-deterministic characteristic of the GA algorithm [26,32], ten SCM models in conjunction with ten different optimized dipeptide propensity scores (opti-DPS) [21–27,38] were performed to generate ten different prediction results. Tables 2 and 3 list the performance comparisons of ten independent runs evaluated by 10-fold CV and independent validation test, respectively.

As seen in Table 2, the average prediction results of the ten individual experiments over 10-fold CV were 93.18% accuracy (ACC), 95.74% sensitivity (SN), 91.95% specificity (SP), 0.858 Matthews coefficient correlation (MCC), and 0.954 area under the receiver operating characteristic (ROC) curve (auROC). It could be noticed that the three-top ranked experiments exhibiting the highest prediction results (Ac, MCC, and auROC) were experiments #3 (95.11%, 0.894, and 0.966), #5 (94.44%, 0.882, and 0.960), and #10 (93.82%, 0.866, and 0.960), respectively. Meanwhile, Table 3 shows that the average prediction results of the ten individual experiments over the independent validation test were 74.47% ACC, 71.67% SN, 75.78% SP, 0.455 MCC, and 0.768 auROC. The three-top ranked experiments having highest prediction results (ACC, MCC, and auROC) were experiments #9 (77.66%, 0.523, and 0.781), #3 (76.60%, 0.482, and 0.781) and #5 (76.60%, 0.461, and 0.793), respectively. Although, experiment #9 was not in the three-top ranked experiments over 10-fold CV, it provides a promising result in terms of ACC, MCC, and auROC with 92.52%, 0.846, and 0.948, respectively, which was not significantly different from the result of experiment #3 (95.11%, 0.894, and 0.966). Moreover, due to the fact that the independent test was the most rigorous cross-validation method to demonstrate the robustness and reliability of the model in real-world applications [17–20,28,29,31,33,39–41], it could be noted that experiment #9 provided an important contribution to PVP prediction. For convenience, the best PVP predictor based on the SCM method in conjunction with the propensity scores of dipeptides from experiment #9 would be referred to as PVPred-SCM. Meanwhile, the opti-DPS from experiment #9

was also used for providing a good understanding of the biochemical and biophysical properties of PVPs, as depicted in Figure 2.

Table 2. Comparison of ten SCM models with ten different optimized dipeptide propensity scores (opti-DPS) over 10-fold cross-validation.

#Exp.	Fitness score	Threshold	ACC (%)	SN (%)	SP (%)	MCC	auROC
1	0.955	443.96	92.50	99.00	89.41	0.849	0.952
2	0.955	459.91	93.15	91.89	93.76	0.851	0.954
3	0.968	471.80	95.11	94.00	95.64	0.894	0.966
4	0.946	476.36	92.15	95.89	90.31	0.840	0.942
5	0.960	455.34	94.44	96.00	93.74	0.882	0.960
6	0.956	458.08	93.16	96.00	91.81	0.856	0.953
7	0.950	446.56	92.50	96.00	90.81	0.846	0.947
8	0.954	446.24	92.51	98.00	89.88	0.849	0.953
9	0.950	461.75	92.52	95.89	90.86	0.846	0.948
10	0.960	463.10	93.82	94.78	93.29	0.866	0.960
Mean	0.955	458.31	93.18	95.74	91.95	0.858	0.954
STD.	0.006	10.77	0.98	1.97	2.05	0.018	0.007

The threshold is an optimal score for discriminating PVPs from non-PVPs. Meanwhile, ACC, SN, SP, MCC, and auROC are accuracy, sensitivity, specificity, Matthews coefficient correlation, and area under the receiver operating characteristic (ROC) curve, respectively.

Table 3. Comparison of ten SCM models with ten different optimized dipeptide propensity scores (opti-DPS) over independent validation test.

#Exp.	Fitness score	Threshold	ACC (%)	SN (%)	SP (%)	MCC	auROC
1	0.955	443.96	74.47	80.00	71.88	0.486	0.782
2	0.955	459.91	75.53	73.33	76.56	0.476	0.743
3	0.968	471.80	76.60	70.00	79.69	0.482	0.781
4	0.946	476.36	76.60	63.33	82.81	0.461	0.775
5	0.960	455.34	76.60	63.33	82.81	0.461	0.793
6	0.956	458.08	71.28	73.33	70.31	0.410	0.749
7	0.950	446.56	72.34	76.67	70.31	0.440	0.749
8	0.954	446.24	70.21	73.33	68.75	0.395	0.742
9	0.950	461.75	77.66	76.67	78.13	0.523	0.781
10	0.960	463.10	73.40	66.67	76.56	0.417	0.787
Mean	0.955	458.31	74.47	71.67	75.78	0.455	0.768
STD.	0.006	10.77	2.56	5.72	5.22	0.040	0.020

The threshold is an optimal score for discriminating PVPs from non-PVPs. Meanwhile, ACC, SN, SP, MCC, and auROC are accuracy, sensitivity, specificity, Matthews coefficient correlation, and area under the ROC curve, respectively.

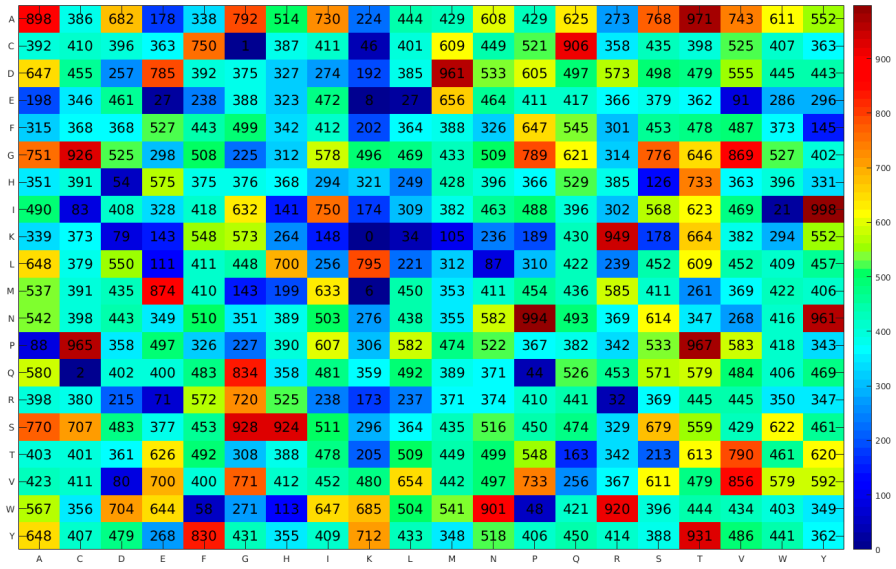

Figure 2. Heatmap of dipeptide propensity scores obtained from the PVPred-SCM method.

2.2. Contribution and Effectiveness of Dipeptide Propensity Scores

As mentioned in Tables 2 and 3, the opti-DPS obtained from experiment #9 outperformed the others in terms of Ac and MCC over the independent validation test. In this section, to investigate the effectiveness of the opti-DPS, we conducted a performance comparison between PVPred-SCM based on opti-DPS and initial dipeptide propensity (init-DPS) as assessed by the 10-fold CV and independent validation test, shown in Table 4. The init-DPS was calculated using the normalized dipeptide composition, while the opti-DPS was generated by optimizing the init-DPS with the GA algorithm as described in the scoring card method section.

From Table 4, PVPred-SCM based on the opti-DPS could enhance the values of ACC, SN, and MCC by 2%, 23%, and 11%, respectively, suggesting that opti-DPS provided more discriminative power than init-DPS. Moreover, the histogram was also used to demonstrate the discriminative power of opti-DPS and init-DPS for PVP prediction as depicted in Figure 3a,b, respectively, where blue and red lines, represent the score distribution of PVPs and non- PVPs, respectively. In practice, more overlap between the red and blue bars indicates that the feature was less capable in PVP prediction. Thus, it could be clearly seen in Figure 3 that opti-DPS was efficient and effective as the input feature for discriminating PVPs from non-PVPs.

Table 4. Comparison between PVPred-SCM based on optimized (opti-DPS) and initial (init-DPS) dipeptide propensity scores assessed by 10-fold cross-validation and independent validation test.

Method	10-fold CV		Independent Test			
	ACC (%)	MCC	ACC (%)	SN (%)	SP (%)	MCC
Init-DPS	85.99	0.705	75.53	53.33	85.94	0.414
opti-DPS	92.52	0.846	77.66	76.67	78.13	0.523

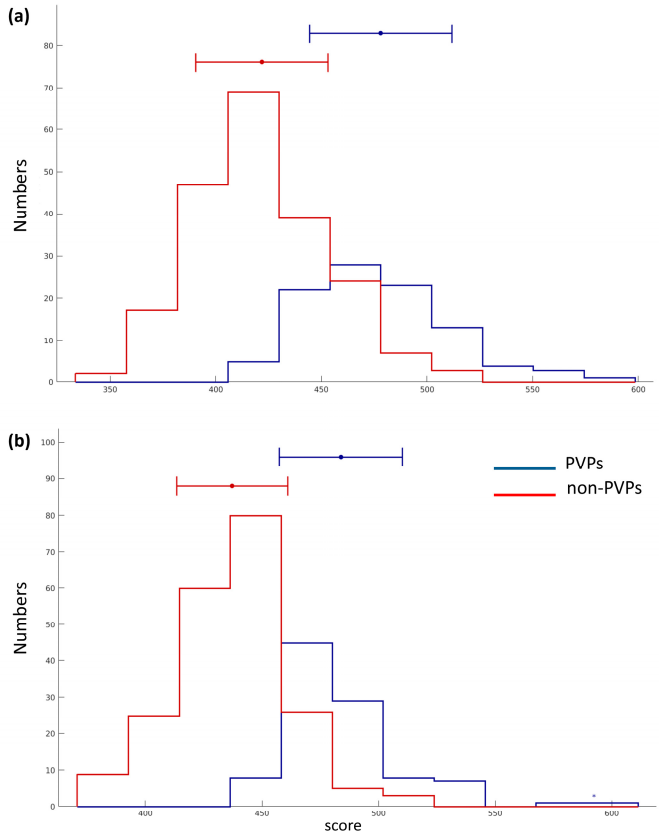

Figure 3. The histogram of scores of PVPs and non-PVPs derived from PVPred-SCM on the benchmark dataset by using initial (**a**) and optimized (**b**) dipeptide propensity scores, respectively.

2.3. Comparison with Existing Methods

As mentioned in the Introduction section, many research groups have made efforts to develop computation models for PVP prediction. To demonstrate the efficiency and strength of the proposed model, we made a comparison between PVPred-SCM and previously developed PVP predictors on the same benchmark (99 PVPs and 208 non-PVPs) and independent (30 PVPs and 64 non-PVPs) datasets, i.e., Feng et al.'s method [9], PVPred [10], PVP-SVM [12], PhagePred [13], and Tan et al.'s method [14], by performing both 10-fold CV and independent validation tests. As noticed from Table 5, there were two different experimental designs: (i) the prediction models were assessed using only 10-fold CV, i.e., Feng et al.'s method and PhagePred, and (ii) the prediction models were assessed on both 10-fold CV and independent tests, i.e., PVPred, PVP-SVM, and Tan et al.'s method**.**

Table 5 lists the performance comparisons between PVPred-SCM and the existing methods over the 10-fold CV and independent validation test. From Table 5, the highest prediction result over 10-fold CV (Ac, MCC) of (98.05%, 0.963) was achieved by PhagePred, while the second and third highest prediction results of (92.5%, 0.846) and (87.95%, 0.761) were achieved by PVPred-SCM and Tan et al.'s method, respectively. Based on the performance comparisons over an independent validation test, PVP-SVM outperforms other PVP predictors, while PVPred-SCM achieved very comparable or slightly worse than PVP-SVM in terms of Ac (79.80% vs. 77.66%) and MCC (0.531 vs. 0.523). On the other hand, the SN value obtained from PVPred-SCM was approximately 10.0% higher than that of PVP-SVM.

Table 5. Performance comparisons between PVPred-SCM and existing methods as assessed by the 10-fold cross-validation and independent validation test.

Method	10-fold CV		Independent Test			
	ACC (%)	MCC	ACC (%)	SN (%)	SP (%)	MCC
Feng et al.'s method[a]	79.15	-	-	-	-	-
PVPred[a]	85.02	-	71.30	60.00	76.50	0.357
PVP-SVM[a]	87.00	0.695	79.80	66.70	85.90	0.531
PhagePred[a]	98.05	0.963	-	-	-	-
Tan et al.'s method[a]	87.95	0.761	75.53	70.00	78.13	0.464
PVPred-SCM	92.52	0.846	77.66	76.67	78.13	0.523

[a] Results were reported from the work of Tan et al.'s method [14].

To validate the robustness of the proposed PVPred-SCM, we further tested and compared its performance using another independent dataset derived from Zhang et al.'s method [11] (containing 68 PVPs and 92 non-PVPs). Table S1 lists the results of this independent test. As noticed in Table A1, PVPred-SCM outperformed the two existing methods, i.e., Feng et al.'s method and PVPred. Meanwhile, Zhang et al.'s method obtained better prediction results than PVPred-SCM because the classifier employed was developed using a stack-based ensemble method in conjunction with various sequence features [11].

With regard to the performance comparison between PVPred-SCM and the existing methods as discussed above, apparently, the best performance of PVPred-SCM contributed to the best SN and the runner-up ACC and MCC. This result demonstrates that the proposed PVPred-SCM could be reliable for identifying truer PVPs (true positives) and fewer false positives. These prediction results revealed that the novelty and significance of PVPred-SCM was as follows: (i) amongst various types of ML algorithms and features employed in developing PVP predictors, the SCM method in conjunction with the propensity scores of dipeptides has never before been employed in developing a PVP predictor. Previously, many studies reported that this feature has been successfully exploited to predict and analyze many protein functions [21–27,32]; (ii) PVPred-SCM was constructed using only a single type of protein feature and the prediction results of unknown proteins were obtained by using the easy-to-use weighted-sum function as illustrated in the Scoring Card Method Section. It could be indicated that PVPred-SCM can be easily understood and manipulated by experimental scientists who want to use the prediction method without the interest to follow the mathematical and statistical details. (iii) Most of the existing PVP predictors [9,10,12–14] developed using NB and SVM methods were used for increasing prediction results. However, the underlying biochemical and biophysical implications remained hard to interpret. Therefore, the interpretability of the proposed method with a satisfied prediction result is a more useful and practical approach. We thus claimed that PVPred-SCM was very promising, as compared to PVP-SVM (77.66% vs 79.80%) considering its simplicity, interpretability, and implementation.

2.4. Identification of Phage Virion Proteins

One of the most important benefits of PVPred-SCM is to facilitate researchers by easily identifying their desired protein with very simple weighted-sum function (S(P)) [21,25]. In this study, PVPred-SCM was used to calculate the PVP scores on the benchmark dataset for finding a new PVP, which has not yet been experimentally characterized, where the PVP scores are calculated using the scoring function S(P) as illustrated in the Scoring Card Method Section. Table 6 lists the top ten potential proteins to be PVPs having the highest PVP scores along with their name, PDB ID, UniProt ID, and source. The mean, minimum, and maximum PVP scores on the benchmark dataset were 452.27, 370.75, and 605.69, respectively. As noticed in Table 6, all top ten proteins had PVP scores with greater than 461.75, where P was classified as a PVP when the PVP score was greater than 461.75; otherwise, P was classified as non-PVP. It could be noted that the selected ten proteins had high potential to be PVPs. Among the top ten proteins having the highest PVP scores, there were only four PVPs that had their structures elucidated via the fiber diffraction method, i.e., Capsid protein G8P, Capsid protein G8P, G VIII capsid protein Precursor, and Major coat protein as shown in Figure 4,

while the structures of the rest six PVPs have not yet been determined. As noticed in Table 6, most of the top ten potential phage virion proteins with the highest PVP scores played an important role in the same function such as capsid or coat proteins. Remarkably, among the mentioned top ten potential phage virion proteins, there were two PVPs appearing to be the major coat protein (gene VIII product; G8P or pVIII; p8) subunit of filamentous phage PH75 and phage Xf. The p8 subunit is a helix consisting of around 50 amino acid residues, which form a filament-like capsid that is involved with phage virion assembly [42]. Therefore, for further analysis, capsid proteins were chosen in the characterization of PVPs. In practice, bacteriophage capsids are present in a variety of shapes, ranging from icosahedral to highly elongated [43].

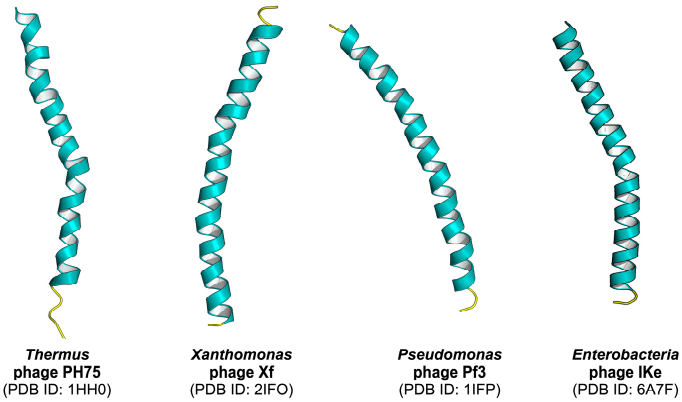

Figure 4. Structures of selected PVPs elucidated via the fiber diffraction method. Each structure is labeled by a common name followed by the Protein Data Bank identification number (PDB ID) in parenthesis on the subsequent line.

Table 6. Top ten potential phage virion proteins having the highest of their PVP scores.

Name (Uniprot)	PVP score	PDBID	UniProtID	Source
Capsid protein G8P	605.69	1HH0	P82889	*Thermus* phage PH75
Capsid protein G8P	581.12	2IFO	P03622	*Xanthomonas* phage Xf
HIS6-pVII fusion protein	541.13		ADR00487	VCSM13 HIS6-pVII modified interference-resistant helper phage
G VIII capsid protein Precursor	538.64	6A7F	NP_040575	Enterobacteria phage Ike
P34	538.52		YP_009639974	Enterobacteria phage PRD1
Major coat protein	534.58	1IFP	NP_040652	*Pseudomonas* phage Pf3
Structural protein P7	532.24		NP_049902	*Pseudoalteromonas* virus PM2
Transclycosylase	529.49		YP_009639979	Enterobacteria phage PRD1
MULTISPECIES: major coat protein	529.45		WP_015979773	Enterobacteriaceae
Hypothetical protein	519.23		WP_015975197	*Salmonella enterica*

2.5. Analysis of Phage Virion Proteins using Propensity Scores of Amino Acids and Dipeptides

The characterization and analysis of feature importance can provide a unique understanding of PVPs. In this study, the propensity score of amino acids and dipeptides were used to identify the important biochemical and biophysical properties of PVPs. Table 7 and Figure 2 list the propensity scores of amino acids and dipeptides, respectively, derived from the benchmark dataset consisting of 99 PVPs and 208 non-PVPs. The propensity scores of dipeptides were obtained from the opti-DPS from experiment #9, while the propensity scores of amino acids (PVP score) were calculated from such opti-DPS using a straightforward statistical approach as mentioned in our previous studies

[26,32]. Amino acids and dipeptides with the highest propensity scores are the most important elements in PVPs.

As seen in Table 7, Ala (529.50), Thr (511.43), Val (506.88), Gly (506.68), and Ser (504.63) were the five top-ranked amino acids present in PVPs, while Leu (395.85), Arg (383.15), His (378.45), Glu (358.93), and Lys (310.90) were the five top-ranked informative amino acids present in non-PVPs. Interestingly, the five top-ranked amino acids representing PVPs and non-PVPs were significantly different between the classes at the level of $p < 0.05$, except for Leu (0.160). These results revealed that the informative amino acids were beneficial for PVP prediction. Based on the propensity scores of dipeptides as shown in Figure 2, the twenty top-ranked dipeptides represented in PVPs included IY, NP, AT, PT, PC, DM, NY, KR, YT, SG, GC, SH, WR, CQ, WN, AA, ME, GV, VV, and QG with propensity scores of 998, 994, 971, 967, 965, 961, 961, 949, 931, 928, 926, 924, 920, 906, 901, 898, 874, 869, 856, and 834, respectively. Meanwhile, the twenty top-ranked dipeptides represented in non-PVPs included KK, CG, QC, MK, EK, IW, EE, EL, RR, KL, QP, CK, WP, HD, WF, RE, KD, VD, IC, and LN with propensity scores of 0, 1, 2, 6, 8, 21, 27, 27, 32, 34, 44, 46, 48, 54, 58, 71, 79, 80, 83, and 87, respectively.

Table 7. The propensity scores of twenty amino acids to be phage virion proteins (score) along with amino acid compositions (%) of PVPs and non- PVPs.

Amino Acid	PVP (%)	Non-PVP (%)	Difference	Score	*P*-value
A-Ala	9.98	8.09	1.89(1)	529.50(1)	<0.05
T-Thr	6.90	5.49	1.41(4)	511.43(2)	<0.05
V-Val	8.09	6.39	1.71(3)	506.88(3)	<0.05
G-Gly	8.20	6.42	1.78(2)	506.68(4)	<0.05
S-Ser	7.29	6.10	1.19(5)	504.63(5)	<0.05
Y-Tyr	3.37	3.50	−0.12(13)	479.13(6)	0.571
N-Asn	4.69	4.64	0.05(9)	471.50(7)	0.866
P-Pro	4.13	3.76	0.38(7)	462.15(8)	0.178
Q-Gln	4.18	3.72	0.46(6)	452.83(9)	0.106
I-Ile	6.23	6.06	0.17(8)	443.18(10)	0.629
W-Trp	1.38	1.50	−0.12(12)	442.33(11)	0.408
D-Asp	5.23	5.87	−0.64(16)	435.45(12)	<0.05
C-Cys	0.66	1.05	−0.39(14)	426.58(13)	<0.05
M-Met	2.64	2.73	−0.09(11)	426.15(14)	0.626
F-Phe	3.91	3.91	0.00(10)	423.45(15)	0.989
L-Leu	7.80	8.34	−0.54(15)	395.85(16)	0.160
R-Arg	4.28	5.48	−1.21(18)	383.15(17)	<0.05
H-His	1.04	1.80	−0.76(17)	378.45(18)	<0.05
E-Glu	4.85	7.36	−2.51(19)	358.93(19)	<0.05
K-Lys	5.14	7.81	−2.67(20)	310.90(20)	<0.05

Interestingly, our results from the propensity scores of amino acids were consistent with the computational results made by Ding et al. [10]. Their results were derived using the analysis of variance (ANOVA) and visualized via a heat map. This study reported that Ala, Gly, Pro, Ser, Thr, and Val were preferred in PVPs. Glu, Lys, Leu, and Arg were preferred in non-PVPs. Based on our analysis results, the ranks of propensity scores for Ala, Thr, Val, Gly, and Ser were 1, 2, 3, 4, and 5, while the ranks of propensity scores for Leu, Arg, Glu, and Lys, were 16, 17, 19, and 20, respectively as shown in Table 7. Furthermore, Lin et al. [44] analyzed the amino acid composition of filamentous bacterial virus xf (*Xanthomonas oryzae*) coat protein, which showed that His, Cys, and Phe were absent from the xf protein, indicating that these three amino acid were not preferred in PVPs. Apparently, the ranks of propensity scores for His, Cys, and Phe were 18, 13, and 15.

2.6. Analysis of PVPs using Informative Physicochemical Properties

In this section, the PVPred-SCM method was exploited to analyze the crucial physicochemical properties (PCPs) of PVPs. Previously, many studies have reported that the biochemical and

biophysical properties such as side-chain [45], alpha-helix propensity [46,47], and hydrophobicity [48] affect the biological activities of PVPs. To analyze the important characteristics of PVPs, Pearson correlation coefficient (R) between the propensity scores of amino acids and the PCPs in the AA index were used to identify informative PCPs that were useful in the analysis of PVPs. As seen in Table 8, the three selected PCPs were QIAN880126 (R = 0.502), WOLR790101 (R = 0.484), and side-chain (R = -0.516), respectively, according to the important biochemical and biophysical properties for PVPs as mentioned above. Additional details of the top ten informative PCPs having highest and lowest R values are listed in Tables S2 and S3. The analysis results cooperating with the propensity scores of amino acids and the three selected PCPs are discussed below.

KOEP990101, which can be described as the "Alpha-helix propensity" [49], had a high positive correlation (R = 0.502). Structure-based conformational preferences of amino acids were reported based on the knowledge of the template backbone on a physical potential energy function, which is composed of a Lennard-Jones potential, electrostatics based on the Coulomb potential, and interatomic contact areas [50]. The high positive correlation suggests that PVPs favor a high alpha-helix propensity. As noticed in Table 8, there were two amino acids, i.e., Gly and Thr, found in the five top-ranked amino acids having the highest propensity scores as well as alpha-helix propensity. The ranks of propensity scorers (PS, alpha-helix) for Gly and Thr were (4, 1) and (2, 3), respectively. Although, the rank of Ala was (1, 12), Pace et al. found that the estimated differences in free energy or $\Delta(\Delta G)$ values in kcal/mol of Ala has been set to zero because it is usually the amino acid with the most favorable helix propensity [46]. They also mentioned that the helix propensity was an important contribution to protein stability. In the backbone effect, a α-helix was stabilized by a strong H-bond and Van der Waals interactions. This study reported that if the α-helix was substituted with Ala, favorable interactions were observed and the helix was more stable than the coil formation. However, if the α-helix was composed of Gly, unfavorable interactions were observed, which resulted in the coil formation being more stable than the helix due to a large reduction in phi–psi space available to residues. Meanwhile, the H-atom of Gly was replaced by a CH_3 in Ala; this is almost entirely an effect of the CH_3 group on the conformational entropy of the backbone in the coil formation. On the other hand, the CH_3 group of Ala experiences no loss in conformational entropy when a randomly coiled polypeptide folds to an α-helix.

Table 8. The three important physicochemical properties (PCPs) derived from PVPred-SCM.

Amino Acid	PS	KOEP990101	Side-Chain [50]	WOLR790101
A-Ala	529.50(1)	−0.04(12)	15(19)	1.12(5)
T-Thr	511.43(2)	0.39(3)	45(15)	−0.02(10)
V-Val	506.88(3)	−0.06(13)	43(16)	1.13(4)
G-Gly	506.68(4)	1.24(1)	1(20)	1.20(1)
S-Ser	504.63(5)	0.15(7)	31(18)	−0.05(11)
Y-Tyr	479.13(6)	0.05(8)	107(2)	−0.23(13)
N-Asn	471.50(7)	0.25(5)	58(11)	−0.83(16)
P-Pro	462.15(8)	0.00(9)	42(17)	0.54(9)
Q-Gln	452.83(9)	−0.02(11)	72(9)	−0.78(14)
I-Ile	443.18(10)	−0.26(17)	57(12)	1.16(3)
W-Trp	442.33(11)	0.21(6)	130(1)	−0.19(12)
D-Asp	435.45(12)	0.27(4)	59(10)	−0.83(17)
C-Cys	426.58(13)	0.57(2)	47(14)	0.59(7)
M-Met	426.15(14)	−0.09(14)	75(6)	0.55(8)
F-Phe	423.45(15)	−0.01(10)	91(4)	0.67(6)
L-Leu	395.85(16)	−0.38(20)	57(13)	1.18(2)
R-Arg	383.15(17)	−0.30(18)	101(3)	−2.55(20)
H-His	378.45(18)	−0.11(15)	82(5)	−0.93(19)
E-Glu	358.93(19)	−0.33(19)	73(7)	−0.92(18)
K-Lys	310.90(20)	−0.18(16)	73(8)	−0.80(15)
Correlation R	1.000	0.502	−0.516	0.484

It could be noticed that the property of side-chain [47] of amino acid had a high negative correlation (R = −0.516), indicating that PVPs favor small amino acids. As seen in Table 8, five top-ranked important amino acids present in PVPs were Ala, Thr, Val, Gly, and Ser, while according to the property of side-chain [50], such five amino acids (PS, side-chain) were ranked at (1,19), (2,15), (3,16), (4,20), and (5,18), respectively. Our analysis result was well consistent with the work of a previous study [45]. In this study, Kuzmicheva et al. established a fusion phage protein containing the β-galactosidase-binding peptide at the N-terminal region by random mutations in amino acids 12–19 adjoining the guest peptide. Based on their experimental results, they suggested that small amino acids played a vital role in ensuring a high binding affinity for the domain C of the phage major coat protein. For the case under their discussion, it could be found that large amino acids tended to have a deleterious effect on the affinity of the guest peptides towards its target. In this study, they mentioned two reasons that form important chemical properties of small amino acids for the phage with the highest affinity: (i) low radius of octapeptide constituting domain C could constrain alpha helix regions and (ii) low steric hindrances leading to low conformation number (i.e., smaller side chain has lower steric hindrances).

In addition, WOLR790101, which can be described as the "Hydrophobicity index" [47], had a high positive correlation (R = 0.484). The hydrophobicity index is a measure of the relative hydrophobicity of each amino acid, which correlated to their solubility in water. Amino acids with hydrophobic side chains also show positive correlation with the hydrophobicity index. As could be seen in Table 8, there were three important amino acids found in the ten top-ranked highest propensity scorers and hydrophobicity index, i.e., Ala (1, 5), Val (3, 4), and Gly (4, 1). Previously, various experimental studies have mentioned that amino acid substitution could affect the stability of procapsids and virions of bacteriophages. For example, Roth et al. constructed minimized-M13 filamentous bacteriophage major coat proteins (P8) to study the minor contributors to the phage assembly process. The result showed that mimi-P8 containing only 50-residues, which comprised of 4 Gly, 37 Ala, and 9 larger residues that were required for phage coat incorporation [51]. In bacteriophage P22, the coat protein E-loop residues Glu52, Glu59, and Glu72 were hypothesized to play a role in stabilizing the capsid by making salt bridges with an adjacent subunit, the P-domain residues R102, R109, and K118, respectively within a capsomer. To verify this hypothesis, a mutant coat protein was constructed by Ala substitution at the E-loop site to remove the charge from Glu without increasing the electrostatic repulsion. The result found that the number of virions assembled within the wild-type coat protein decreased to 50% at 2 M urea, while phages assembled with mutant coat protein showed a 50% survival rate in a higher concentration of urea. It could be noted that the substitution of Glu with Ala increased the stability of procapsids and virions of bacteriophages. Although electrostatic repulsion could cause destabilization between the E-loop residue and spine helix, hydrophobic interaction could relieve this effect by balancing the repulsive electrostatic interactions [48]. As noticed in Table 8, the ranks of propensity scores (PS, hydrophobicity index) for Glu and Ala were (18, 19) and (1, 5), respectively, while the propensity scores of charged amino acid side chains of Lys, Glu, His, Arg, and Asp were ranked at 20, 19, 18, 17, and 12, respectively.

2.7. Web Server Implementation

In an effort to facilitate high-throughput prediction of PVPs, the best predictive model was deployed as a PVPred-SCM web server and is made freely available online at http://camt.pythonanywhere.com/PVPred-SCM where users can submit query protein sequences for determining the likelihood of these proteins being PVPs. Below, we provide step-by-step guidelines on how to use the PVPred-SCM web server in order to obtain the desired results. Firstly, the user opens the web server at http://camt.pythonanywhere.com/PVPred-SCM and the user will see the top screen of PVPpred on the user's computer screen, as shown in Figure 4. Secondly, the user enters the query sequence into the text box or uploads a FASTA file by clicking on the "Choose file" button. Thirdly, the user clicks on the "Submit" button in order to start the prediction process. Finally, after finishing the prediction process, the results are outputted as shown on the right-hand side of the web server. The user can see examples of FASTA-formatted sequences by clicking on the "example file"

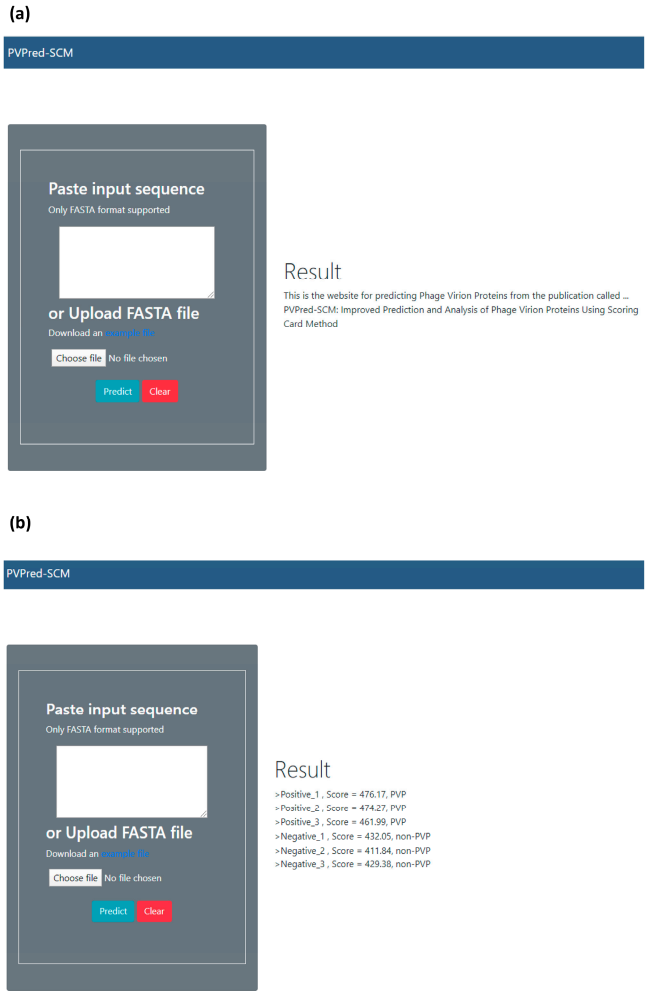

Figure 4. Screenshots of the PVPred-SCM web server before (**a**) and after (**b**) submission of a query protein. Prediction results are represented with the PVP scores derived from the scoring function (S(P)) and predicted classes.

button. In fact, user-friendly and publicly accessible web-servers that can display the findings manipulated by users according to their needs, might significantly enhance their impact, driving medicinal chemistry into an unprecedented revolution [52–56]. Keeping this point in our mind, we shall make efforts in our future work to provide a web-server with such functionality.

2.8. Reproducible Research

To guarantee and ensure the reproducibility of the proposed model, herein, all codes and datasets used in the construction of the predictive models are available on GitHub at https://github.com/Shoombuatong2527/PVPred-SCM.

3. Materials and Methods

In order to establish a robust and interpretable sequence-based tool for modeling the investigated PVPs, we followed the six prime keys as mentioned in a series of recent publications [17–33] and summarized in several comprehensive review papers [29,34–36]: (i) establishing a reliable dataset that contains experimentally validated sequences for training and validating the model; (ii) representing protein sequences with interpretable features; (iii) developing interpretable learning algorithms so as to allow the interpretation of important features responsible for the biological

activity; (iv) assessing the prediction model using standard cross-validation tests; (v) constructing a user-friendly web-server for obtaining the prediction without the need to understand complex mathematical and statistical details; and (vi) analyzing and characterizing the important features derived from the developed model to provide a better understanding of the biophysical and biochemical properties of proteins. Figure 2 shows the workflow of PVPred-SCM, which works in predicting and analyzing PVPs.

3.1. Dataset Preparation

One of the most crucial steps is to establish a reliable and stringent benchmark dataset to train and validate the proposed method. To make a fair comparison with existing methods, the same benchmark (S_{TR}) and independent (S_{TS}) datasets derived from the work of [12] were taken to develop and validate the proposed model, respectively. These two datasets were specifically used to develop PVP predictors for discriminating PVPs from non-PVPs. The main reason for selecting these two datasets was that the protein sequences were non-redundant, while protein sequences with more than 40% similarity were removed from the datasets, thereby avoiding misleading results with overestimated accuracy. In this study, the benchmark dataset consists of 99 PVPs and 208 non-PVPs, while the independent dataset consists of 30 PVPs and 64 non-PVPs. We noted that none of the PVPs in the independent dataset are identical to PVPs in the benchmark or training dataset. The benchmark (S_{TR}) and independent (S_{TS}) datasets used in this study can be summarized by the following formula:

$$\mathbf{S_{TR}} = \mathbf{S_{TR}^{+}} \cup \mathbf{S_{TR}^{-}} \quad (1)$$

$$\mathbf{S_{TR}} = \mathbf{S_{TR}^{+}} \cup \mathbf{S_{TR}^{-}} \quad (2)$$

where S^+ and S^- represent peptide sequences of PVPs and non-PVPs, respectively, while the symbol $\cup$ represents the union from the set theory.

3.2. Feature representation

Given a protein sequence (P), it can be represented as:

$$\mathbf{P} = \mathrm{p_1 p_2 p_3 \cdots p_{1N}} \quad (3)$$

where p_i and N denote the i^{th} residue in the protein $\mathbf{P}$ and the peptide length, respectively. Note that the residue p_i is in the set of natural amino acid, i.e., A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y. In the development of a sequence-based predictor for characterizing and analyzing the biophysical and biochemical property of peptides, one of the most crucial aspects is how to best represent the peptides in such a way as to afford a comprehensive and proper description of the feature that could well reflect their functions. Until now, various convenient tools have been developed for representing a protein sequence using a fixed-length feature vector, such as BioSeq-Analysis2.0 [57], iFeature [58], protr/ProtrWeb [59], etc. Many studies have reported that DPC is one of the most interpretable and effective features for predicting and analyzing various types of protein and peptide functions [12,17–26,28–33,39,40,60–66]. Basically, DPC are the proportion of dipeptide in a peptide sequence P that are expressed as a fixed length of 400. Thus, in terms of the DPC feature, a protein sequence P can be expressed by a vector with 400D (dimension), as formulated by:

$$\mathbf{P} = [\mathrm{dp_1}, \mathrm{dp_2}, \ldots, \mathrm{dp_{400}}]^{\mathbf{T}} \quad (4)$$

where T is the transposed operator, while $\mathrm{dp_1}$, $\mathrm{dp_2}$..., $\mathrm{dp_{400}}$ are occurrence frequencies of the 400 native dipeptides, respectively, in a protein sequence P.

3.3. Scoring Card Method

Amongst various types of machine learning algorithms that are available, SCM method has been successfully implemented to predict and characterize many protein functions [21–27] by using only the sequence information. SCM method is an efficient and simple approach that performs well not only for predicting proteins by using an efficient function, but also for gaining insight into the

characteristic of proteins based on the propensity scores of dipeptides and amino acids. The advantages of the SCM method are threefold. Firstly, the prediction procedure of SCM method is significantly simpler than the SVM model. For the prediction of unknown proteins, the SCM method exploits a weighted summation of opti-DPS and the occurrence frequencies of the 400 native dipeptides. Secondly, the propensity scores of amino acids and dipeptides play an important role in understanding the underlying biophysical and biochemical properties governing biological activities of proteins and peptides. Finally, the SCM model is a general-purpose method for predicting protein and peptide functions by using an efficient function. Therefore, in this study, the SCM method was utilized for developing the PVP predictor named PVPred-SCM. More details for the SCM method were described in our previous studies [26,27,32]. Herein, the basic concepts and associated parameter optimizations for the SCM method is briefly described as follows:

Step 1. Preparing a benchmark dataset (S_{TR}) for training and developing PVP prediction.

Step 2. Generating init-DPS consisting of 400 propensity scores of dipeptides and then normalizing such propensity scores into [0,1000], where dipeptides with the highest propensity scores are the most important residues in PVPs. It could be stated that the init-DPS is derived from the normalized dipeptide composition.

Step 3. Calculating propensity scores of amino acid AA_i by averaging 40 propensity scores of dipeptides that contain AA_i

Step 4. Optimizing init-DPS by using the GA method. In this study, the fitness function of the GA method was used to maximize both the prediction performance in terms of auROC and the Pearson's correlation coefficient (R) between init-DPS and opti-DPS. To obtain the best opti-DPS, the fitness function (Fit-Funt) is assessed by 10-fold CV defined as follows:

$$\text{Fit} - \text{Funt} = W_1 \times \text{auROC} + W_2 \times \text{R} \quad (5)$$

where the values of W_1 and W_2 are 0.9 and 0.1, respectively. The detail for the optimization of these two parameters was described in our previous studies [26,27,32].

Step 5. Predicting an unknown protein sequence ($P_{unknown}$) by using the scoring function (S(P)). The procedure of the prediction of an unknown sequence is briefly described as follows (i) choosing the best opti-DPS with the highest score derived from Fit-Funt and (ii) calculating the summation of 400 propensity scores (S_i) and frequencies (w_i) of dipeptides, predicting $P_{unknown}$ as PVP if S(.) is greater than the threshold value, otherwise $P_{unknown}$ is predicted as non-PVP.

$$S(P_{unknown}) = \sum_{i=1}^{400} w_i S_i \quad (6)$$

3.4. Characterization of Phage Virion Proteins

In this work, we performed the propensity score of amino acids and dipeptides using the SCM method to provide a better understanding of the biophysical and biochemical properties of PVPs. The calculation of propensity scores of amino acids and dipeptides was described in the above section. The values of propensity scores represent the influence of each amino acid and dipeptide on the biological, functional, and structural properties of proteins.

Furthermore, the informative physicochemical property (PCP) of amino acids was identified using PVPred-SCM. PCPs are one of the most intuitive features associated with biophysical and biochemical reactions. Previously, our studies utilized these features for predicting and analyzing various functions of proteins and peptides from primary sequences [17,19,20,26,28,29,32,33,67]. In fact, there are 544 PCPs of amino acids extracted from the amino acid index database (AAindex) [37], which is a collection of published literature as well as different biochemical and biophysical properties of amino acids. Each PCP consists of a set of 20 numerical values for amino acids. PCPs with not applicable (NA) as their amino acid indices were excluded, and a total of 531 PCPs were further used in this study.

The procedure of determining informative PCPs using PVPred-SCM consists of two main steps as follows: (i) calculating the R value between the propensity scores of amino acids and the 531 PCPs.

PCPs with highest absolute R values indicate that these PCPs are highly correlated with the biological, functional, and structural properties of PVPs, and (ii) PCPs with an absolute R value of greater than 0.5 are preferred as a candidate for PVP characterization [22,23,25].

3.5. Performance Evaluation

Cross-validation of the proposed method is an essential step in the development of computational models [29,35,36,68–74]. These cross-validation procedures were used in order to assess the success and error rates of the proposed method. In practice, there are three CV methods that are traditional approaches, i.e., the sub-sampling test or k-fold cross-validation (k-fold CV), jackknife test, and independent validation test or external test. Among these methods, the jackknife test is recognized as the least arbitrary and most objective one, as mention by Equations (28)–(32) in [75]. Meanwhile, the external test is considered as one of the most rigorous and objective methods for cross-validation in statistics. In the k-fold cross-validation procedure, the training set is randomly separated into k subsets. From the k subsets, a single subset is taken as the testing set to validate the prediction model trained and learned by the remaining k-1 subsets. This process is repeated k times, until each subset had been used as the testing set. During the jackknifing process, a single sample in the whole dataset having N samples is taken as the testing set and the remaining N-1 samples were used for training the model. This process was repeated N times, until each sample was used as the testing set.

In order to evaluate the prediction ability of the model, the following sets of four metrics were used as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \tag{7}$$

$$\mathrm{SN} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{8}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})} \tag{9}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{10}$$

where ACC, SN, SP, and MCC are called the accuracy, sensitivity, specificity, and Matthews coefficient correlation, respectively. TP, TN, FP, and FN represent the instances of true positive, true negative, false positive, and false negative, respectively. Moreover, the receiver operating characteristic (ROC) curves were exploited to evaluate the prediction performance of models using threshold-independent parameters. The area under the ROC curve (auROC) was used to measure the prediction performance, where the area under the curve (AUC) values of 0.5 and 1 were indicative of perfect and random models, respectively.

4. Conclusions

A phage virion protein (PVP) plays a crucial role in the research of bacterial infections, especially in bacterial drug resistance. Thus, a computational method that can both effectively predict and analyze PVPs may help researchers in reducing cost as well as the time spent in providing a better understanding of the biochemical and biophysical properties of PVPs. Unfortunately, there is no study so far that provides a systematic effort for the prediction and analysis of PVPs. Herein, we have developed an efficient and interpretable PVP predictor named PVPred-SCM in a systematic manner by taking advantage of a scoring card method (SCM) and the sequence features of propensity scores of amino acids and dipeptides. PVPred-SCM predicts an unknown sequence using the weighted-sum function that can easily be adopted using a single threshold for discriminating PVPs from non-PVPs. Unlike existing methods in pursuit of high prediction performances, PVPred-SCM aims to maximize the interpretability of both classifier and feature usage while maximizing predictive performance of a model. Upon comparing with the state-of-the-art method (PVP-SVM) on the same benchmark

dataset, it was found that the performance of PVPred-SCM was comparable to that of PVP-SVM (79.80% vs. 77.66%) as assessed by the rigorous independent validation test, indicating that the proposed model PVPred-SCM is very promising considering its simplicity, interpretability, and implementation. Additionally, the propensity score of amino acids and informative physicochemical properties derived from PVPred-SCM were used to provide insights into the biochemical and biophysical properties of PVPs. Finally, a web server named PVPred-SCM was established and made freely available online at http://camt.pythonanywhere.com/PVPred-SCM. Due to the high potential of our systematic approach employed in this study, PVPred-SCM could be extended for predicting and analyzing many other types of protein and peptide functions without any major modifications, such as predicting HIV-1 CRF01-AE co-receptor usage [18,62], prediction of the human leukocyte antigen gene [30,76], predicting antifreeze proteins [33], predicting the hemolytic activity of peptides [29], predicting anticancer activity of peptides [20], predicting antiviral activity of peptides [40], and predicting antihypertensive activity of peptides [28]. Furthermore, the proposed model could be integrated with other beneficial peptide features such as pseudo amino acid composition [50,77] or amphiphilic pseudo amino acid composition [78] as proposed by Chou [77,79–82] for further improving the PVP prediction.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Table S1. Performance comparisons between PVPred-SCM with existing methods on independent dataset derived from Zhang et al's method [11]. Table S2. The ten top-ranked informative physicohemical properties having the highest pearson correlation (R) with the propensity scores of amino acids. Table S3. The ten top-ranked informative physicohemical properties having the lowest pearson correlation (R) with the propensity scores of amino acids

Author Contributions: W.S. conceived, designed and analyzed the experiments. W.S. and P.C. designed and performed the experiments. J.Y. S.K and W.S analyzed the data. P.C contributed the code for constructing PVPred-SCM model and the web server. W.S., S.K., J.Y., N.S. and P.C. drafted the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was co-supported by the TRF Research Grant for New Scholar (No. MRG6180226) and College of Arts, Media and Technology, Chiang Mai University, and partially supported by Chiang Mai University and Mahidol University.

Acknowledgments: We thank the reviewers for their great comments in helping to improve this manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Clark, J.R.; March, J.B., Bacteriophages and biotechnology: Vaccines, gene therapy and antibacterials. *Trends Biotechnol* **2006,** 24, 212–218.
2. Samson, J.E.; Magadan, A.H.; Sabri, M.; Moineau, S., Revenge of the phages: Defeating bacterial defences. *Nat Rev Microbiol* **2013,** 11, 675–687.
3. Lyon, J., Phage Therapy's Role in Combating Antibiotic-Resistant Pathogens. *JAMA* **2017,** 318, 1746–1748.
4. Lekunberri, I.; Subirats, J.; Borrego, C.M.; Balcazar, J.L., Exploring the contribution of bacteriophages to antibiotic resistance. *Env. Pollut* **2017,** 220, (Pt B), 981-984.
5. Jara-Acevedo, R.; Díez, P.; González-González, M.; Dégano, R.M.; Ibarrola, N.; Góngora, R.; Orfao, A.; Fuentes, M., Screening phage-display antibody libraries using protein arrays. In *Phage Display*, Springer: 2018; pp 365–380.
6. Lavigne, R.; Ceyssens, P.-J.; Robben, J., Phage proteomics: Applications of mass spectrometry. In *Bacteriophages*, Springer: 2009; pp 239–251.
7. Yuan, Y.; Gao, M., Proteomic analysis of a novel bacillus jumbo phage revealing glycoside hydrolase as structural component. *Front. Microbiol.* **2016**, *7*, 745.

8. Seguritan, V.; Alves Jr, N.; Arnoult, M.; Raymond, A.; Lorimer, D.; Burgin Jr, A.B.; Salamon, P.; Segall, A.M., Artificial neural networks trained to detect viral and phage structural proteins. *Plos Comput. Biol.* **2012,** 8, e1002657.
9. Feng, P.-M.; Ding, H.; Chen, W.; Lin, H., Na\ve Bayes Classifier with Feature Selection to Identify Phage Virion Proteins. vol: 2013.
10. Ding, H.; Feng, P.-M.; Chen, W.; Lin, H., Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Mol. Biosyst.* **2014,** 10, 2229–2235.
11. Zhang, L.; Zhang, C.; Gao, R.; Yang, R., An ensemble method to distinguish bacteriophage virion from non-virion proteins based on protein sequence characteristics. *Int. J. Mol. Sci.* **2015,** 16, 21734–21758.
12. Manavalan, B.; Shin, T.H.; Lee, G., PVP-SVM: Sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.* **2018**, *9*, 476.
13. Pan, Y.; Gao, H.; Lin, H.; Liu, Z.; Tang, L.; Li, S., Identification of Bacteriophage Virion Proteins Using Multinomial Naive Bayes with g-Gap Feature Tree. *Int. J. Mol. Sci.* **2018,** 19, (6), 1779.
14. Tan, J.-X.; Dao, F.-Y.; Lv, H.; Feng, P.-M.; Ding, H., Identifying phage virion proteins by using two-step feature selection methods. *Molecules* **2018,** 23, 2000.
15. Ru, X.; Li, L.; Wang, C., Identification of phage viral proteins with hybrid sequence features. *Front. Microbiol.* **2019,** 10.
16. Arif, M.; Ali, F.; Ahmad, S.; Kabir, M.; Ali, Z.; Hayat, M., Pred-BVP-Unb: Fast prediction of bacteriophage Virion proteins using un-biased multi-perspective properties with recursive feature elimination. *Genomics* **2019**.
17. Shoombuatong, W.; Schaduangrat, N.; Pratiwi, R.; Nantasenamat, C., THPep: A machine learning-based approach for predicting tumor homing peptides. *Comput. Biol. Chem.* **2019**, *80*, 441–451.
18. Hongjaisee, S.; Nantasenamat, C.; Carraway, T.S.; Shoombuatong, W., HIVCoR: A sequence-based tool for predicting HIV-1 CRF01_AE coreceptor usage. *Comput. Biol. Chem.* **2019**, *80*, 419–432.
19. Laengsri, V.; Nantasenamat, C.; Schaduangrat, N.; Nuchnoi, P.; Prachayasittikul, V.; Shoombuatong, W., TargetAntiAngio: A Sequence-Based Tool for the Prediction and Analysis of Anti-Angiogenic Peptides. *Int. J. Mol. Sci.* **2019,** 20, 2950.
20. Schaduangrat, N.; Nantasenamat, C.; Prachayasittikul, V.; Shoombuatong, W., ACPred: A Computational Tool for the Prediction and Analysis of Anticancer Peptides. *Molecules* **2019,** 24, 1973.
21. Vasylenko, T.; Liou, Y.-F.; Chiou, P.-C.; Chu, H.-W.; Lai, Y.-S.; Chou, Y.-L.; Huang, H.-L.; Ho, S.-Y., SCMBYK: Prediction and characterization of bacterial tyrosine-kinases based on propensity scores of dipeptides. *Bmc Bioinform.* **2016,** 17, 514.
22. Vasylenko, T.; Liou, Y.-F.; Chen, H.-A.; Charoenkwan, P.; Huang, H.-L.; Ho, S.-Y. In *SCMPSP: Prediction and characterization of photosynthetic proteins based on a scoring card method*, BMC bioinformatics, 2015; BioMed Central: P S8.
23. Liou, Y.-F.; Vasylenko, T.; Yeh, C.-L.; Lin, W.-C.; Chiu, S.-H.; Charoenkwan, P.; Shu, L.-S.; Ho, S.-Y.; Huang, H.-L., SCMMTP: Identifying and characterizing membrane transport proteins using propensity scores of dipeptides. *Bmc Genom.* **2015,** 16, S6.
24. Huang, H.-L., Propensity scores for prediction and characterization of bioluminescent proteins from sequences. *Plos One* **2014,** 9, e97158.
25. Liou, Y.-F.; Charoenkwan, P.; Srinivasulu, Y.S.; Vasylenko, T.; Lai, S.-C.; Lee, H.-C.; Chen, Y.-H.; Huang, H.-L.; Ho, S.-Y., SCMHBP: Prediction and analysis of heme binding proteins using propensity scores of dipeptides. *Bmc Bioinform.* **2014,** 15, S4.

26. Charoenkwan, P.; Shoombuatong, W.; Lee, H.-C.; Chaijaruwanich, J.; Huang, H.-L.; Ho, S.-Y., SCMCRYS: Predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. *Plos One* **2013,** 8, e72368.
27. Huang, H.-L.; Charoenkwan, P.; Kao, T.-F.; Lee, H.-C.; Chang, F.-L.; Huang, W.-L.; Ho, S.-J.; Shu, L.-S.; Chen, W.-L.; Ho, S.-Y. In *Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition*, Bmc Bioinformatics, 2012; BioMed Central: P S3.
28. Win, T.S.; Schaduangrat, N.; Prachayasittikul, V.; Nantasenamat, C.; Shoombuatong, W., PAAP: A web server for predicting antihypertensive activity of peptides. *Future Med. Chem.* **2018,** 10, 1749–1767.
29. Win, T.S.; Malik, A.A.; Prachayasittikul, V.; S.; Wikberg, J.E.; Nantasenamat, C.; Shoombuatong, W., HemoPred: A web server for predicting the hemolytic activity of peptides. *Future Med. Chem.* **2017,** 9, 275–291.
30. Shoombuatong, W.; Prachayasittikul, V.; Anuwongcharoen, N.; Songtawee, N.; Monnor, T.; Prachayasittikul, S.; Prachayasittikul, V.; Nantasenamat, C., Navigating the chemical space of dipeptidyl peptidase-4 inhibitors. *Drug Des. Dev. Ther.* **2015**, *9*, 4515.
31. Shoombuatong, W.; Nabu, S.; Simeon, S.; Prachayasittikul, V.; Lapins, M.; Wikberg, J.E.; Nantasenamat, C., Extending proteochemometric modeling for unraveling the sorption behavior of compound–soil interaction. *Chemom. Intell. Lab. Syst.* **2016**, *151*, 219–227.
32. Shoombuatong, W.; Huang, H.-L.; Chaijaruwanich, J.; Charoenkwan, P.; Lee, H.-C.; Ho, S.-Y. In *Predicting protein crystallization using a simple scoring card method*, Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2013 IEEE Symposium on, 2013; IEEE: Pp 23-30.
33. Pratiwi, R.; Malik, A.A.; Schaduangrat, N.; Prachayasittikul, V.; Wikberg, J.E.; Nantasenamat, C.; Shoombuatong, W., CryoProtect: A Web Server for Classifying Antifreeze Proteins from Nonantifreeze Proteins. *J. Chem.* **2017,** 2017.
34. Shoombuatong, W.; Schaduangrat, N.; Nantasenamat, C., Unraveling the bioactivity of anticancer peptides as deduced from machine learning. *Excli J.* **2018**, *17*, 734.
35. Shoombuatong, W.; Schaduangrat, N.; Nantasenamat, C., Towards understanding aromatase inhibitory activity via QSAR modeling. *Excli J.* **2018**, *17*, 688.
36. Shoombuatong, W.; Prathipati, P.; Prachayasittikul, V.; Schaduangrat, N.; Malik, A.A.; Pratiwi, R.; Wanwimolruk, S.; Wikberg, J.E.; Gleeson, M.P.; Spjuth, O., Towards Predicting the Cytochrome P450 Modulation: From QSAR to Proteochemometric Modeling. *Curr. Drug Metab.* **2017,** 18, 540–555.
37. Kawashima, S.; Kanehisa, M., AAindex: Amino acid index database. *Nucleic Acids Res.* **2000,** 28, 374–374.
38. Shoombuatong, W.; Prachayasittikul, V.; Prachayasittikul, V.; Nantasenamat, C., Prediction of aromatase inhibitory activity using the efficient linear method (ELM). *Excli J.* **2015**, *14*, 452.
39. Simeon, S.; Shoombuatong, W.; Anuwongcharoen, N.; Preeyanon, L.; Prachayasittikul, V.; Wikberg, J.E.; Nantasenamat, C., osFP: A web server for predicting the oligomeric states of fluorescent proteins. *J. Cheminformatics* **2016,** 8, 72.
40. Schaduangrat, N.; Nantasenamat, C.; Prachayasittikul, V.; Shoombuatong, W., Meta-iAVP: A Sequence-Based Meta-Predictor for Improving the Prediction of Antiviral Peptides Using Effective Feature Representation. *Int. J. Mol. Sci.* **2019,** 20, 5743.
41. Suvannang, N.; Preeyanon, L.; Malik, A.A.; Schaduangrat, N.; Shoombuatong, W.; Worachartcheewan, A.; Tantimongcolwat, T.; Nantasenamat, C., Probing the origin of estrogen receptor alpha inhibition via large-scale QSAR study. *Rsc Adv.* **2018,** 8, 11344–11356.

42. Pederson, D.M.; Welsh, L.C.; Marvin, D.A.; Sampson, M.; Perham, R.N.; Yu, M.; Slater, M.R., The protein capsid of filamentous bacteriophage PH75 from Thermus thermophilus. *J Mol Biol* **2001,** 309, 401–421.
43. Ackermann, H.W., Bacteriophage taxonomy in 1987. *Microbiol Sci* **1987,** 4, 214–218.
44. Lin, J.Y.; Wu, C.C.; Kue, T.T., Amino acid analysis of the coat protein of the filamentous bacterial virus xf from Xanthomonas oryzae. *Virology* **1971,** 45, 38–41.
45. Kuzmicheva, G.; Jayanna, P.; Eroshkin, A.; Grishina, M.; Pereyaslavskaya, E.; Potemkin, V.; Petrenko, V., Mutations in fd phage major coat protein modulate affinity of the displayed peptide. *Protein Eng. Des. Sel.* **2009,** 22, 631–639.
46. Pace, C.N.; Scholtz, J.M., A helix propensity scale based on experimental studies of peptides and proteins. *Biophys J* **1998,** 75, 422–427.
47. Wolfenden, R.V.; Cullis, P.M.; Southgate, C.C., Water, protein folding, and the genetic code. *Science* **1979,** 206, , 575–577.
48. Asija, K.; Teschke, C.M., Of capsid structure and stability: The partnership between charged residues of E-loop and P-domain of the bacteriophage P22 coat protein. *Virology* **2019**, *534*, 45–53.
49. Koehl, P.; Levitt, M., Structure-based conformational preferences of amino acids. *Proc Natl Acad Sci U S A* **1999,** 96, 12524–12529.
50. Shen, H.-B.; Chou, K.-C., PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.* **2008,** 373, 386–388.
51. Roth, T.A.; Weiss, G.A.; Eigenbrot, C.; Sidhu, S.S., A minimized M13 coat protein defines the requirements for assembly into the bacteriophage particle. *J Mol Biol* **2002,** 322, 357–367.
52. Cheng, X.; Xiao, X.; Chou, K.-C., pLoc-mPlant: Predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC. *Mol. Biosyst.* **2017,** 13, 1722–1727.
53. Cheng, X.; Xiao, X.; Chou, K.-C., pLoc-mVirus: Predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC. *Gene* **2017**, *628*, 315–321.
54. Cheng, X.; Zhao, S.-G.; Lin, W.-Z.; Xiao, X.; Chou, K.-C., pLoc-mAnimal: Predict subcellular localization of animal proteins with both single and multiple sites. *Bioinformatics* **2017,** 33, 3524–3531.
55. Cheng, X.; Zhao, S.-G.; Xiao, X.; Chou, K.-C., iATC-mISF: A multi-label classifier for predicting the classes of anatomical therapeutic chemicals. *Bioinformatics* **2016,** 33, 341–346.
56. Cheng, X.; Zhao, S.-G.; Xiao, X.; Chou, K.-C., iATC-mHyb: A hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals. *Oncotarget* **2017,** 8, 58494.
57. Liu, B.; Gao, X.; Zhang, H., BioSeq-Analysis2. 0: An updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* **2019,** 47, e127¨Ce127.
58. Chen, Z.; Zhao, P.; Li, F.; Leier, A.; Marquez-Lago, T.T.; Wang, Y.; Webb, G.I.; Smith, A.I.; Daly, R.J.; Chou, K.-C., iFeature: A python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* **2018,** 34, 2499–2502.
59. Xiao, N.; Cao, D.-S.; Zhu, M.-F.; Xu, Q.-S., protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* **2015,** 31, (11), 1857-1859.
60. Simeon, S.; Li, H.; Win, T.S.; Malik, A.A.; Kandhro, A.H.; Piacham, T.; Shoombuatong, W.; Nuchnoi, P.; Wikberg, J.E.; Gleeson, M.P., PepBio: Predicting the bioactivity of host defense peptides. *Rsc Adv.* **2017,** 7, 35119–35134.

61. Shoombuatong, W.; Traisathit, P.; Prasitwattanaseree, S.; Tayapiwatana, C.; Cutler, R.; Chaijaruwanich, J., Prediction of the disulphide bonding state of cysteines in proteins using Conditional Random Fields. *Int. J. Data Min. Bioinform.* **2011,** 5, 449–464.
62. Shoombuatong, W.; Hongjaisee, S.; Barin, F.; Chaijaruwanich, J.; Samleerat, T., HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees. *Comput. Biol. Med.* **2012,** 42, 885–889.
63. Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G., mAHTPred: A sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* **2018**.
64. Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G., Meta-4mCpred: A Sequence-Based Meta-Predictor for Accurate DNA 4mC Site Prediction Using Effective Feature Representation. *Mol. Ther. -Nucleic Acids* **2019**, *16*, 733–744.
65. Manavalan, B.; Govindaraj, R.G.; Shin, T.H.; Kim, M.O.; Lee, G., iBCE-EL: A new ensemble learning framework for improved linear B-cell epitope prediction. *Front. Immunol.* **2018**, *9*, 1695.
66. Manavalan, B.; Shin, T.H.; Kim, M.O.; Lee, G., PIP-EL: A new ensemble learning method for improved proinflammatory peptide predictions. *Front. Immunol.* **2018**, *9*, 1783.
67. Charoenkwan, P.; Schaduangrat, N.; Nantasenamat, C.; Piacham, T.; Shoombuatong, W., iQSP: A Sequence-Based Tool for the Prediction and Analysis of Quorum Sensing Peptides via Chou's 5-Steps Rule and Informative Physicochemical Properties. *Int. J. Mol. Sci.* **2020,** 21, 75.
68. Worachartcheewan, A.; Prachayasittikul, V.; Anuwongcharoen, N.; Shoombuatong, W.; Prachayasittikul, V.; Nantasenamat, C., On the origins of hepatitis C virus NS5B polymerase inhibitory activity using machine learning approaches. *Curr. Top. Med. Chem.* **2015,** 15, 1814–1826.
69. Prachayasittikul, V.; Worachartcheewan, A.; Shoombuatong, W.; Songtawee, N.; Simeon, S.; Prachayasittikul, V.; Nantasenamat, C., Computer-aided drug design of bioactive natural products. *Curr. Top. Med. Chem.* **2015,** 15, 1780–1800.
70. Khatun, M.; Hasan, M.; Kurata, H., PreAIP: Computational prediction of anti-inflammatory peptides by integrating multiple complementary features. *Front. Genet.* **2019**, *10*, 129.
71. Hasan, M.M.; Manavalan, B.; Khatun, M.S.; Kurata, H., i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome. *Int. J. Biol. Macromol.* **2019**.
72. Hasan, M.; Khatun, M.; Kurata, H., Large-Scale Assessment of Bioinformatics Tools for Lysine Succinylation Sites. *Cells* **2019,** 8, 95.
73. Hasan, M.M.; Manavalan, B.; Khatun, M.S.; Kurata, H., Prediction of S-nitrosylation sites by integrating support vector machines and random forest. *Mol. Omics* **2019,** 15, 451–458.
74. Hasan, M.M.; Khatun, M.S.; Kurata, H., A comprehensive review of in silico analysis for protein S-sulfenylation sites. *Protein Pept. Lett.* **2018,** 25, 815–821.
75. Chou, K.-C., Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **2011,** 273, 236–247.
76. Shoombuatong, W.; Mekha, P.; Waiyamai, K.; Cheevadhanarak, S.; Chaijaruwanicha, J., Prediction of human leukocyte antigen gene using k-nearest neighbour classifier based on spectrum kernel. *ScienceAsia* **2013,** 39, 42–49.
77. Liu, B.; Liu, F.; Wang, X.; Chen, J.; Fang, L.; Chou, K.-C., Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* **2015,** 43, W65–W71.

78. Chou, K.-C., Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2004,** 21, 10–19.
79. Chou, K.-C., Impacts of bioinformatics to medicinal chemistry. *Med. Chem.* **2015,** 11, 218–234.
80. Chou, K.-C., An unprecedented revolution in medicinal chemistry driven by the progress of biological science. *Curr. Top. Med. Chem.* **2017,** 17, 2337–2358.
81. Chou, K.-C., Progresses in predicting post-translational modification. *Int. J. Pept. Res. Ther.* **2019**, 1-16.
82. Chou, K.-C., Advance in Predicting Subcellular Localization of Multi-label Proteins and its Implication for Developing Multi-target Drugs. *Curr. Med. Chem.* **2019**.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

International Journal of
Molecular Sciences

MDPI

Article

TargetAntiAngio: A Sequence-Based Tool for the Prediction and Analysis of Anti-Angiogenic Peptides

Vishuda Laengsri [1,2], Chanin Nantasenamat [3], Nalini Schaduangrat [3], Pornlada Nuchnoi [1,2], Virapong Prachayasittikul [4] and Watshara Shoombuatong [3,*]

1 Department of Clinical Microscopy, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand; l.vishuda@gmail.com (V.L.); pornlada.nuc@mahidol.ac.th (P.N.)
2 Center for Research and Innovation, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
3 Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand; chanin.nan@mahidol.edu (C.N.); nalini.schaduangrat@gmail.com (N.S.)
4 Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand; virapong.pra@mahidol.ac.th
* Correspondence: watshara.sho@mahidol.ac.th; Tel.: +66-2-441-4371 (ext. 2715)

Received: 24 May 2019; Accepted: 14 June 2019; Published: 17 June 2019

Abstract: Cancer remains one of the major causes of death worldwide. Angiogenesis is crucial for the pathogenesis of various human diseases, especially solid tumors. The discovery of anti-angiogenic peptides is a promising therapeutic route for cancer treatment. Thus, reliably identifying anti-angiogenic peptides is extremely important for understanding their biophysical and biochemical properties that serve as the basis for the discovery of new anti-cancer drugs. This study aims to develop an efficient and interpretable computational model called TargetAntiAngio for predicting and characterizing anti-angiogenic peptides. TargetAntiAngio was developed using the random forest classifier in conjunction with various classes of peptide features. It was observed via an independent validation test that TargetAntiAngio can identify anti-angiogenic peptides with an average accuracy of 77.50% on an objective benchmark dataset. Comparisons demonstrated that TargetAntiAngio is superior to other existing methods. In addition, results revealed the following important characteristics of anti-angiogenic peptides: (i) disulfide bond forming Cys residues play an important role for inhibiting blood vessel proliferation; (ii) Cys located at the C-terminal domain can decrease endothelial formatting activity and suppress tumor growth; and (iii) Cyclic disulfide-rich peptides contribute to the inhibition of angiogenesis and cell migration, selectivity and stability. Finally, for the convenience of experimental scientists, the TargetAntiAngio web server was established and made freely available online.

Keywords: anti-angiogenic peptide; therapeutic peptides; interpretable model; random forest; machine learning; classification

1. Introduction

Cancer constitutes a group of diseases involving the unregulated proliferation of abnormal cells. It is capable of both invading surrounding normal tissue and spreading throughout the body via the circulatory or lymphatic system in a process known as metastasis [1]. Cancer is the second leading cause of death globally accounting for an estimated 9.6 million cases of death in 2018 [2]. However, early identification and treatment are able to increase the chances of survival for patients. As such, a combination of targeted drugs accompanied with chemotherapy or radiation is an essential strategy for ensuring an optimal outcome for patients [3,4].

Angiogenesis is a process by which new blood vessels are formed and it is seen as one of the key processes for the proliferation and metastatic spread of cancer cells. It promotes the circulation of oxygenated blood, supplies nutrients, and removes waste products from the body [5]. Furthermore, angiogenesis is regulated by both activator and inhibitor molecules with stimulation occurring when tumor tissues require oxygen and nutrients. However, the upregulation of the activity of angiogenic factors alone is not enough for the angiogenesis of neoplasm, as anti-angiogenic factors also needs to be downregulated [6].

Until now, various proteins have been identified and characterized as pro-angiogenic molecules including vascular endothelial growth factor (VEGF), transforming growth factor (TGF)-α, TGF-β, basic fibroblast growth factor (bFGF), angiogenin, and platelet-derived endothelial growth factor (PDGF) [7]. VEGF plays a central role in angiogenesis and is greatly expressed in cancer cells. Although there are many naturally occurring proteins that can inhibit angiogenesis (e.g., endostatin, angiostatin, platelet factor 4, and thrombospondin) [8], many researchers are still attempting to develop new anti-angiogenic molecules for inhibiting VEGF [9,10]. The conceptual basis for the discovery of novel anti-angiogenic molecules as a therapeutic route by means of VEGF inhibition is summarized in Figure 1. In addition, monoclonal anti-VEGF antibody (Avastin or Bevacizumab) is the first anti-angiogenic drug that is known for its ability to inhibit tumor blood vessel growth as well as its ability to increase survival rate in cancer patients [11,12]. However, blocking VEGF alone is not sufficient for stalling cancer cell growth and progression. Moreover, Sorafenib and Sunitinib [13], which are small molecules that can inhibit the VEGF receptor tyrosine kinase, has been approved for renal cell carcinoma and colorectal cancer treatments [14]. The production of various pro-angiogenic molecules besides VEGF for promoting tumor angiogenesis is a challenging endeavor, thus constructing new anti-angiogenic peptides represents an interesting avenue for novel therapeutics [15]. Lee et al. [16] reported a novel collagen IV derived biomimetic peptide that can inhibit breast cancer growth and metastasis. Its use in combination treatment with HER2 and VEGF peptides mimicked the induction of potent anti-tumor responses both in vitro and in vivo [17]. Therefore, the efficacy of anti-angiogenic peptides is dependent upon the cancer type. It is worthy to note that anti-angiogenic drugs are more efficient in well vascularized cancers.

Apart from cancer, excessive vascular growth could promote blindness, rheumatoid arthritis and psoriasis [18]. Several studies have shown the effectiveness of anti-angiogenic peptides [15,19,20]. For example, luteolin has been demonstrated to be a potent anti-angiogenic agent for retinal neovascularization by suppressing the VEGF expression [21]. Additionally, anti-neuropilin-1 (anti-NP-1) was synthesized to block the function of NP-1, which is responsible for the induction of increased synoviocyte survival and angiogenesis. Thus, anti-NP-1 is useful in alleviating chronic arthritis [22]. Currently, peptide therapeutics are increasingly being used in medical practices against various diseases including cancers [15], microbial infections [23], and cardiovascular diseases [24]. Owing to several advantageous peptide properties such as high selectivity, efficacy, and being relatively safe, their use in the search for novel targeted drugs has gained much interest. However, several concerning factors should also be addressed, for instance, rapid degradation and excretion in humans, stability during storage, and low oral bioavailability [25]. In addition, most endogenous anti-angiogenic proteins are complex and too large, thereby causing them difficulty in penetrating target tissues. In this regard, therapeutic peptides were able to overcome these limitations with the development of anti-angiogenic peptides not exceeding 50 amino acids in length. Moreover, some of these peptides have been optimized and modified such as amino acid substitutions and conversion from L-to-D amino acids [26]. The effort of developing small peptide fragments that represent similar anti-angiogenic properties and could be applied to inhibit tumor angiogenesis in cancer patients represent valuable challenges. Besides, it is beneficial to construct specifically functional peptides as a result of an imbalance between activators and inhibitors that contributes to different pathological conditions.

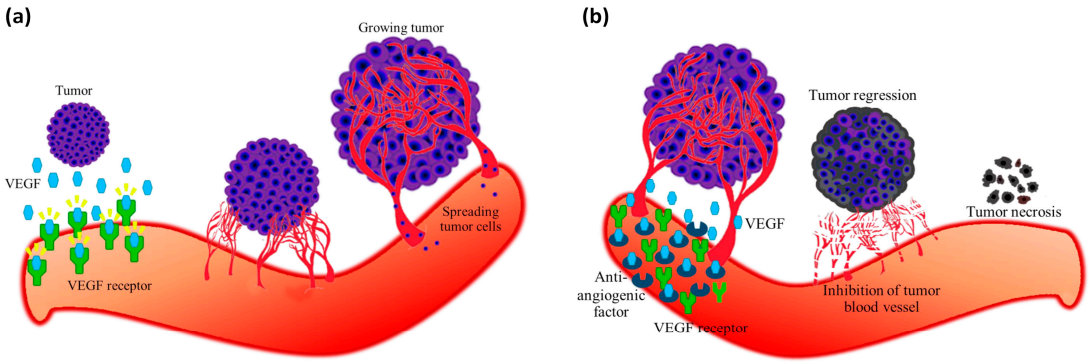

Figure 1. Angiogenesis is regulated by a local equilibrium between pro-angiogenic such as vascular endothelial growth factor (VEGF), platelet-derived endothelial growth factor (PDGF), fibroblast growth factor (FGF), and angiopoietins and anti-angiogenic molecules such as endostatin, PF4 and TSP-1. It is switched on when tumor cells require oxygen and nutrients. Tumor cells produce VEGF and then secretes them into surrounding tissues. When VEGF binds to its receptor on the outer surface of endothelial cells, it activates endothelial cells that subsequently drives the development of new blood vessels from pre-existing vasculatures. Blood vessels gradually grow and expand to tumor cells whereby tumor cells continuously proliferate and spread into the blood circulation. Cancer progression is induced by an overexpression of pro-angiogenic factors (**a**). Disruption of the vascular supply can be mediated by blocking pro-angiogenic factors or via the use of anti-angiogenic factors as therapeutic drug is anticipated to increase the survival rate of cancer patients. Anti-angiogenic factor binds to VEGF thereby leading to the inhibition of neovascularization and tumor growth thereby leading to a decrease of metastasis. Eventually, tumor cells which are devoid of fuels (e.g., oxygen and nutrients) gently regress and become tumor necrosis (**b**).

With the avalanche of post-genomic data, peptide sequences are abundantly available in databases. However, conventional experimental techniques for the identification and development of anti-angiogenic peptides have been very slow, expensive and laborious. Therefore, it is highly desirable to develop computational methods for predicting and characterizing anti-angiogenic peptides. However, until now, few efforts have been made to develop methods for accurately predicting anti-angiogenic peptides as summarized in Table 1. Ramaprasad et al. [27] first proposed a computational model named AntiAngioPred by using support vector machine in conjunction with amino acid composition and information on the first fifteen residues at the N-terminus region. This method gave accuracies of 75.00% and 74.96% as assessed by an independent validation test from 1 and 5 rounds of random splits, respectively. In 2018, Blanco et al. [28] utilized the three basic sequential features consisting of amino acid, dipeptide and tripeptide compositions for training and learning their prediction models. Their comparison results indicated that the best model based on generalized linear model yielded 86% accuracy in which such model utilized the top 200 informative features for model building. Recently, Zahiri et al. [29] developed the AntAngioCOOL R package and executed performance comparisons amongst various machine learning techniques and types of peptide features. Based on their results on performance comparisons, regression, and survival trees model employing descriptors consisting of pseudo amino acid composition, *k*-mer composition, *k*-mer composition (reduced alphabet), physico-chemical profile, and atomic profile yielded the highest accuracy of 77% over an independent validation test from 1 round of random split.

Table 1. Summary of existing methods for predicting anti-angiogenic peptides.

Method	Classifier [a]	Sequence Feature (No. of Feature Used) [b]	Independent Test	Web Server
AntiAngioPred [27]	SVM	AAC (20)	Yes	Yes
Blanco et al.'s method [28]	glmnet	AAC, DPC, TC (200)	No	No
AntAngioCOOL [29]	PART	PseAAC, *k*-mer composition, RAAC, PCP, AC (2,343)	No	No
TargetAntiAngio (this study)	RF	AAC, PseAAC, Am-PseAAC (48)	Yes	Yes

[a] glmnet: a generalized linear model, PART: recursive partitioning for classification, regression and survival trees, RF: random forest, SVM: support vector machine. [b] AAC: amino acid composition, AC: atomic profile, Am-PseAAC: amphiphilic pseudo amino acid composition, DPC: dipeptide composition, PCP: physicochemical properties, PseACC: pseudo amino acid composition, RACC: reduce amino acid composition, TC: tripeptide composition. The method is assessed by an independent validation test with *N* rounds of random splits.

Although all these methods yielded encouraging results and played an important role in simulating the development of anti-angiogenic peptides identification, there is still room for further improvement in the prediction performance of anti-angiogenic peptides. The following research gaps have been elucidated as follows: (i) Blanco et al.'s method [28] and AntAngioCOOL [29] were assessed by the independent validation test from only 1 round of random split, hence, their prediction results were not yet satisfactory; (ii) the studies of Blanco et al.'s method [28] and AntAngioCOOL [29] did not provide a web server, hence, their usage was quite limited; and (iii) AntiAngioPred [27] and Blanco et al.'s method were not straight-forward enough to provide the underlying mechanism of anti-angiogenic peptides due of the lack of interpretability of the model.

Motivated by these considerations, this work attempts to develop a new sequence-based predictor for predicting and analyzing anti-angiogenic peptides, called the TargetAntiAngio, which utilizes the random forest classifier in cooperation with various types of peptide features including amino acid composition, dipeptide composition, physicochemical properties, pseudo amino acid composition, and amphiphilic pseudo amino acid composition. Rigorous cross-validation tests indicated that TargetAntiAngio outperformed the existing methods. Furthermore, this study also identified sequence features that contributed to high prediction accuracy as well as provided better understanding on the biophysical and biochemical properties of anti-angiogenic peptides by means of feature importance analysis. Finally, based on the proposed method, a user-friendly web server, called the TargetAntiAngio, was established for the prediction of anti-angiogenic peptides.

2. Results and Discussion

In this study, both 5-fold CV and independent validation test was performed on the benchmark (S_{main}) and NT15 (S_{NT15}) datasets. As mentioned in the section on the Benchmark dataset, the training and testing sets were constructed with the random sampling process. Furthermore, data splitting was performed with ten independent iterations to avoid the possible bias of the random sampling procedure. The final prediction results of the 5-fold CV and independent validation test were obtained by averaging the ten independent experiments. After which, comparisons of the prediction performances between the proposed method and the existing methods were conducted. Moreover, the informative features of AAC, DPC, and PCP were investigated to provide important biophysical and biochemical properties of anti-angiogenic activities of peptides. Finally, TargetAntiAngio was established as a free web server. Figure 2 shows the workflow of TargetAntiAngio which works in discriminating peptides as anti-angiogenic or non-antiangiogenic peptides.

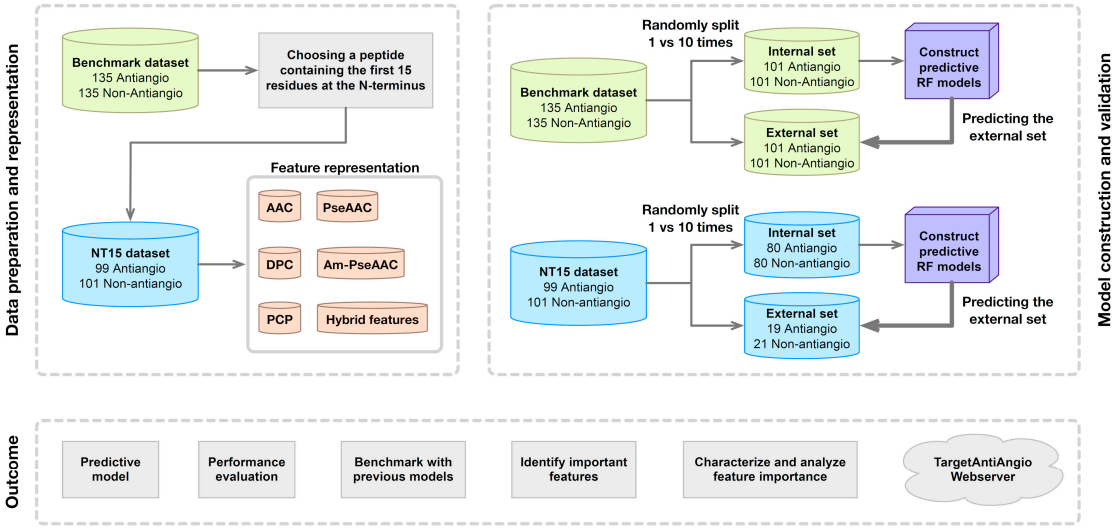

Figure 2. Schematic framework of TargetAntiAngio.

2.1. Prediction Performance

In order to predict and characterize anti-angiogenic peptides, it is very important to choose a useful classifier with informative features for the design of an accurate predictor as well as providing good understanding of anti-angiogenic activities of peptides. In this study, the five basic features (i.e., AAC, DPC, PCP, PseAAC, and Am-PseAAC) as well as their combinations (i.e., AAC+PseACC, AAC+Am-PseACC, PseACC+Am-PseACC, and AAC+PseACC+Am-PseACC) were selected as input features for training RF models followed identifying good combination of the five aforementioned features.

Performance comparisons of the various feature types was performed for models built via 5-fold CV and independent validation test on the S_{main} data set that was subjected to 1 random split and 10 rounds of random splits on the dataset as shown in Tables 2 and 3, respectively. As noticed in Table 2, the highest test accuracy and MCC of 72.22% and 0.45, respectively, was achieved using the PseAAC feature. Meanwhile, the Am-PseAAC and ACC performed well with the second and third highest test accuracies of 72.22% and 72.12%, respectively. In order to yield better prediction performance, we also utilized the combinations of the top 3 important features (i.e., ACC, PseAAC and Am-PseAAC) to train the prediction models. The combination of PseACC and Am-PseACC reached a test accuracy and MCC of 77.78% and 0.56, respectively, while the combination of AAC and PseACC provided the second highest test accuracy and MCC of 75.93% and 0.52, respectively. In the case of the prediction results from 10 rounds of random splits, from amongst the top 3 important features, Table 3 shows that AAC had the best performance with a test accuracy and MCC of 73.33 ± 1.01% and 0.47 ± 0.02, respectively. Meanwhile, the combined features of AAC+PseACC and AAC+PseAAC+Am-PseAAC yielded the first and second highest test accuracy and MCC of 74.81 ± 1.01%/0.50 ± 0.02 and 74.07 ± 1.31%/0.49 ± 0.02, respectively.

Table 2. Performance comparison of RF models built with various types of sequence features. Models were evaluated by means of five-fold cross-validation and independent validation test using benchmark and NT15 datasets subjected to one round of random split.

Feature	Dataset	5-Fold CV			Independent Test				
		Ac (%)	MCC	auROC	Ac (%)	Sn (%)	Sp (%)	MCC	auROC
ACC	Benckmark	71.03	0.42	0.81	72.12	67.86	76.92	0.45	0.77
	NT15	75.00	0.50	0.80	77.50	90.48	63.16	0.56	0.82
PseAAC	Benckmark	73.83	0.48	0.78	72.22	85.71	57.69	0.45	0.81
	NT15	73.75	0.48	0.80	72.50	85.71	57.90	0.46	0.83
Am-PseAAC	Benckmark	71.96	0.44	0.76	72.22	82.14	61.54	0.45	0.76
	NT15	72.50	0.45	0.80	75.00	76.19	73.68	0.50	0.80
DPC	Benckmark	68.22	0.37	0.75	70.37	82.14	57.69	0.41	0.72
	NT15	72.50	0.45	0.79	72.50	95.24	47.37	0.49	0.69
PCP	Benckmark	60.75	0.22	0.67	61.11	67.86	53.85	0.22	0.65
	NT15	67.50	0.36	0.72	67.50	76.19	57.90	0.35	0.74
AAC+PseAAC	Benckmark	72.43	0.45	0.79	75.93	85.71	65.39	0.52	0.80
	NT15	74.38	0.50	0.77	77.00	85.71	68.42	0.55	0.83
AAC+Am-PseAAC	Benckmark	70.09	0.41	0.76	74.07	89.29	57.69	0.50	0.83
	NT15	74.38	0.50	0.81	75.00	71.43	78.95	0.50	0.79
PseAAC+Am-PseAAC	Benckmark	72.90	0.46	0.77	77.78	82.14	73.08	0.56	0.83
	NT15	75.00	0.50	0.82	75.00	85.71	63.16	0.50	0.85
AAC+PseAAC+Am-PseAAC	Benckmark	71.03	0.42	0.78	74.07	75.00	73.08	0.48	0.82
	NT15	75.00	0.50	0.82	77.50	90.48	63.16	0.56	0.84

Parameters of PseAAC (weight^1 and lamda^1) and Am-PseAAC (weight^2 and lamda^2) were optimized by varying their values and assessed by a 5-fold CV procedure. Values of weight^1, weight^2, lamda^1, and lamda^2 as performed on the benchmark and NT15 datasets are (0.9, 0.9, 1, and 1) and (0.1, 0.2, 2, and 3), respectively.

Table 3. Performance comparison of RF models built with various types of sequence features. Models were evaluated by means of five-fold cross-validation and independent validation test using benchmark and NT15 datasets subjected to ten rounds of random splits.

Feature	Dataset	5-Fold CV			Independent Test				
		Ac (%)	MCC	auROC	Ac (%)	Sn (%)	Sp (%)	MCC	auROC
ACC	Benckmark	70.84 ± 1.54	0.42 ± 0.03	0.79 ± 0.01	73.33 ± 1.01	77.14 ± 8.60	69.23 ± 9.81	0.47 ± 0.02	0.79 ± 0.02
	NT15	74.12 ± 2.10	0.49 ± 0.04	0.80 ± 0.02	77.00 ± 2.09	84.76 ± 6.21	68.42 ± 8.32	0.55 ± 0.04	0.82 ± 0.02
PseAAC	Benckmark	71.78 ± 2.13	0.44 ± 0.04	0.77 ± 0.01	72.96 ± 1.66	80.00 ± 4.07	65.38 ± 6.08	0.46 ± 0.03	0.81 ± 0.02
	NT15	71.62 ± 1.44	0.43 ± 0.03	0.78 ± 0.02	73.50 ± 1.37	80.95 ± 8.25	65.26 ± 11.53	0.48 ± 0.03	0.81 ± 0.03
Am-PseAAC	Benckmark	70.47 ± 2.10	0.41 ± 0.04	0.75 ± 0.02	72.96 ± 2.11	75.71 ± 9.58	70.00 ± 8.34	0.46 ± 0.05	0.79 ± 0.04
	NT15	72.38 ± 2.31	0.45 ± 0.05	0.81 ± 0.01	73.50 ± 1.37	76.19 ± 12.14	70.53 ± 12.68	0.48 ± 0.02	0.79 ± 0.04
DPC	Benckmark	68.32 ± 0.84	0.37 ± 0.02	0.74 ± 0.01	69.63 ± 2.11	78.57 ± 9.45	60.00 ± 5.83	0.40 ± 0.05	0.74 ± 0.02
	NT15	71.50 ± 2.01	0.43 ± 0.04	0.78 ± 0.02	69.50 ± 3.26	78.09 ± 10.43	60.00 ± 7.98	0.40 ± 0.08	0.75 ± 0.07
PCP	Benckmark	60.19 ± 2.44	0.20 ± 0.05	0.65 ± 0.02	61.85 ± 2.11	62.14 ± 10.59	61.54 ± 9.81	0.24 ± 0.04	0.66 ± 0.03
	NT15	68.00 ± 1.03	0.36 ± 0.02	0.74 ± 0.02	67.50 ± 2.50	72.38 ± 7.82	62.11 ± 11.41	0.35 ± 0.05	0.72 ± 0.07
AAC+PseAAC	Benckmark	72.24 ± 0.53	0.45 ± 0.01	0.79 ± 0.01	74.81 ± 1.01	81.43 ± 4.66	67.69 ± 5.83	0.50 ± 0.02	0.81 ± 0.04
	NT15	73.50 ± 1.44	0.47 ± 0.03	0.79 ± 0.02	76.50 ± 1.37	84.76 ± 7.82	67.37 ± 10.79	0.54 ± 0.02	0.82 ± 0.05
AAC+Am-PseAAC	Benckmark	70.37 ± 1.13	0.41 ± 0.02	0.77 ± 0.02	73.33 ± 1.01	85.00 ± 4.66	60.77 ± 5.70	0.48 ± 0.02	0.78 ± 0.04
	NT15	73.00 ± 0.81	0.47 ± 0.02	0.80 ± 0.02	75.50 ± 1.12	83.81 ± 7.97	66.32 ± 9.56	0.52 ± 0.02	0.82 ± 0.06
PseAAC+Am-PseAAC	Benckmark	73.18 ± 1.57	0.47 ± 0.03	0.78 ± 0.01	73.33 ± 2.81	80.71 ± 1.96	65.38 ± 5.44	0.47 ± 0.05	0.78 ± 0.05
	NT15	73.88 ± 2.14	0.48 ± 0.04	0.80 ± 0.02	75.00 ± 1.77	80.95 ± 5.83	68.42 ± 6.45	0.50 ± 0.04	0.80 ± 0.05
AAC+PseAAC+Am-PseAAC	Benckmark	70.37 ± 1.22	0.41 ± 0.02	0.77 ± 0.02	74.07 ± 1.31	82.14 ± 5.05	65.38 ± 7.20	0.49 ± 0.02	0.81 ± 0.01
	NT15	74.62 ± 1.57	0.50 ± 0.03	0.81 ± 0.01	77.50 ± 1.77	84.76 ± 10.32	69.47 ± 8.65	0.56 ± 0.03	0.83 ± 0.03

Parameters of PseAAC (weight^1 and lamda^1) and Am-PseAAC (weight^2 and lamda^2) were optimized by varying their values and assessed by a 5-fold CV procedure. Values of weight^1, weight^2, lamda^1, and lamda^2 as performed on the benchmark and NT15 datasets are (0.9, 0.9, 1, and 1) and (0.1, 0.2, 2, and 3), respectively.

As mentioned in the section on the Benchmark dataset, it is not fair to compare our results with existing methods because AntiAngioPred was trained on the S_{NT15} dataset. Therefore, in this study, the S_{NT15} dataset was also utilized to develop the prediction models for comparative purposes. Performance comparisons of the RF models with various sequence features are summarized in Tables 2 and 3. The highest test accuracy and MCC of 77.50 ± 1.77% and 0.56 ± 0.03 was achieved by using the combined features consisting of AAC, PseACC, and Am-PseACC. Meanwhile, the AAC feature and the combined feature of AAC+PseACC performed well as it afforded the second and third highest test accuracy and MCC of 77.00 ± 2.09%/0.55 ± 0.04 and 75.50 ± 1.12%/0.52 ± 0.02, respectively. As seen in Tables 2 and 3, prediction results for the S_{NT15} dataset are quite consistent with that of the S_{main} dataset.

Furthermore, from Tables 2 and 3, the experimental results can be briefly summarized hereafter. Each of the three single features including AAC, PseACC, and Am-PseACC are benefical for predicting anti-angiogenic peptides with test accuracies of >73% and >77% when performed on S_{main} and S_{NT15} datasets, respectively. Furthermore, prediction results for the S_{NT15} dataset were better than that of the S_{main} dataset thereby indicating that the position of the first fifteen residues plays a vital role in discriminating anti-angiogenic from non-antiangiogenic peptides (Tables 2 and 3). This observation is in good consistency with the study of Ramaprasad et al. [27]. The best prediction performance for both S_{main} and S_{NT15} datasets as evaluated via independent validation test from 10 rounds of random splits were achieved by using the combined features of AAC, PseACC, and Am-PseACC. For convenience, we will refer to this RF method built with the combined feature of AAC, PseACC, and Am-PseACC as TargetAntiAngio.

2.2. Comparison with Other Methods

It is necessary to compare our proposed method TargetAntiAngio with that of the existing methods by performing both cross-validation and independent validation tests so as to ascertain its efficiency and strength for the prediction of anti-angiogenic peptides. Until now, there are only three sequence-based predictors that have been developed for identifying anti-angiogenic peptides consisting of AntiAngioPred [27], Blanco et al.'s method [28], and AntAngioCOOL, as summarized in Table 1. From amongst these three predictors, only AntiAngioPred provided prediction results that are rigorously assessed by both cross-validation and independent validation tests as assessed by more than 1 round of random split (i.e., evaluated on both the S_{main} and S_{NT15} datasets). In view of this point, herein, we only compared our method TargetAntiAngio with AntiAngioPred [27]. Table 4 lists the performance comparisons between TargetAntiAngio and AntiAngioPred over 5-fold cross-validation and independent validation tests using the S_{NT15} dataset.

Table 4. Performance comparisons between TargetAntiAngio and AntiAngioPred assessed by 5-fold cross-validation and independent validation tests on NT15 dataset.

Sampling Time	Method	Cross-Validation		Independent Test			
		Ac (%)	MCC	Ac (%)	Sn (%)	Sp (%)	MCC
1 round	AntiAngioPred [a]	80.90	0.62	75.00	-	-	0.51
	TargetAntiAngio	75.00	0.50	77.50	90.48	63.16	0.56
N rounds [b]	AntiAngioPred [a]	-	-	74.96	72.90	76.80	0.50
	TargetAntiAngio	74.62	0.50	77.50	84.76	69.47	0.56

[a] Results were reported from the work of AntiAngioPred. [b] N represents the number of 5 and 10 rounds of random splits for performing the prediction results of AntiAngioPred and TargetAntiAngio, respectively.

Based on the results from Table 4, it can be seen that TargetAntiAngio afforded a lower performance as compared to AntiAngioPred (75.00% vs 80.90% accuracy) as assessed by 5-fold cross-validation from one round of random split. On the other hand, prediction results for models evaluated by the independent validation test using dataset obtained from one round of random split indicated that TargetAntiAngio achieved better performances than that of AntiAngioPred as observed from the

values of accuracy (77.50% vs 75.00%) and MCC (0.56 vs. 0.51). Furthermore, TargetAntiAngio was also found to outperform AntiAngioPred with improvements of 3% and 6% for both accuracy and MCC, respectively, as evaluated by a robust independent validation test using datasets obtained from 10 rounds of random splits.

2.3. Biological Space

The analysis of feature importance can provide a better understanding of the mechanistic details governing the anti-angiogenic activity of peptides. As mentioned above, in this study, the informative features of AAC, DPC, and PCP were used to characterize the anti-angiogenic activity of peptides. In order to select informative features, this study utilized the RF model because of its built-in ability of feature importance estimation and its great prediction performance. The value of mean decrease of Gini index (MDGI) is adopted to rank and estimate the importance of each AAC and DPC features. Such information is derived from analysis of the S_{main} dataset that consists of 137 anti-angiogenic and 137 non-antiangiogenic peptides. Table 5 lists the percentage values of the 20 amino acids for both anti-angiogenic and non-anti-angiogenic along with their amino acid compositional difference between the two classes along with their MDGI values. Features with the highest MDGI value is considered as the most important as it significantly contributed to the prediction performance. As seen in Table 5, the 10 top-ranked informative amino acids with the highest MDGI values are Cys, Ser, Val, Ala, Leu, Arg, Glu, Lys, and Pro afforded MDGI values of 15.90, 14.43, 9.58, 9.21, 8.41, 8.31, 6.68, 6.59, 6.40, and 5.52, respectively. Meanwhile, from amongst the 10 informative amino acids, the analysis of AAC with the percentage of certain residues on anti-angiogenic peptides suggested that Cys, Ser, Arg, and Pro are dominant in anti-angiogenic peptides, while Val, Ala, Leu, Glu, and Ile are dominant in non-antiangiogenic peptides at the significance level of p-value ≤ 0.05.

Table 5. Amino acid compositions (%) of antiangiogenic (Angio) and non-antiangiogenic (non-Angio) peptides along with their mean decrease of Gini index (MDGI) values.

Amino acid	Anti-Angio (%)	Non-Anti-Angio (%)	Difference	p-Value	MDGI
A-Ala	0.053	0.086	−0.033 (20)	<0.05	9.21 (4)
C-Cys	0.047	0.014	0.034 (2)	<0.05	15.90 (1)
D-Asp	0.047	0.052	−0.005 (13)	0.568	5.02 (12)
E-Glu	0.046	0.065	−0.019 (17)	<0.05	6.68 (7)
F-Phe	0.030	0.043	−0.013 (15)	<0.05	4.89 (13)
G-Gly	0.081	0.073	0.008 (7)	0.420	5.18 (11)
H-His	0.030	0.024	0.007 (8)	0.373	4.58 (14)
I-Ile	0.046	0.064	−0.017 (16)	<0.05	5.26 (10)
K-Lys	0.056	0.056	0.001 (9)	0.933	6.59 (8)
L-Leu	0.067	0.095	−0.028 (19)	<0.05	8.41 (5)
M-Met	0.019	0.023	−0.004 (12)	0.366	3.45 (20)
N-Asn	0.037	0.040	−0.003 (11)	0.657	3.71 (18)
P-Pro	0.060	0.045	0.016 (4)	<0.05	6.40 (9)
Q-Gln	0.039	0.042	−0.002 (10)	0.757	4.47 (15)
R-Arg	0.088	0.055	0.032 (3)	<0.05	8.31 (6)
S-Ser	0.096	0.057	0.039 (1)	<0.05	14.43 (2)
T-Thr	0.062	0.054	0.008 (6)	0.232	3.77 (17)
V-Val	0.048	0.073	−0.025 (18)	<0.05	9.58 (3)
W-Trp	0.023	0.012	0.012 (5)	<0.05	3.95 (16)
Y-Tyr	0.023	0.029	−0.007 (14)	0.210	3.45 (19)

Furthermore, the sequence logo of the first and last fifteen residues at the N- and C-terminal regions of both anti-angiogenic and non-antiangiogenic peptides were created to visualize the positional information for each amino acid as shown in Figure 3. The overall stack height of each position indicates its sequence conservation while the size of the residue represents its propensity. Figure 3a,c shows that Pro, Ser, Trp, Cys, and Gly as well as Cys, Ser, Gly, Pro, and Arg are abundant at the first 15

residues from the N- and C-terminal regions, respectively, of anti-angiogenic peptides. However, only Leu and Ala are abundant at the last 15 residues from the C-terminal region of non-antiangiogenic peptides. Thus, information gathered from the sequence logo illustration shows crucial amino acid residues that could potentially be used for discriminating anti-angiogenic from non-antiangiogenic peptides. Moreover, Cys, Ser, and Arg are seen to be favored by anti-angiogenic peptides, especially at the C-terminal region. These analyses were in good consistency with the feature importance as estimated using MDGI values where Cys, Ser, and Arg are ranked 1, 2, and 6, respectively (Table 6).

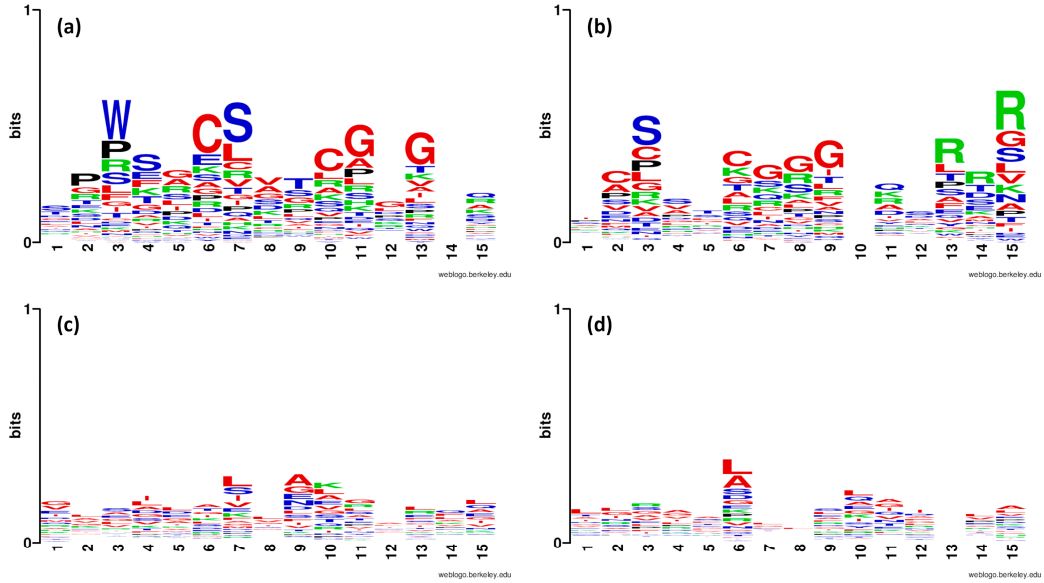

Figure 3. Sequence logo representations of antiangiogenic and non-antiangiogenic peptides. Shown are the sequence logo of the first and last 15 residues at N- and C-terminal regions from antiangiogenic peptides (**a**,**b**) and non-antiangiogenic peptides (**c**,**d**).

Table 6. Ten top-ranked physicochemical properties from the AAindex having the highest MDGI values.

Rank	AAindex	MDGI	Description
1	CHOP780216	0.73	Normalized frequency of the 2nd and 3rd residues in turn (Chou-Fasman, 1978b)
2	CHOP780215	0.61	Frequency of the 4th residue in turn (Chou-Fasman, 1978b)
3	MIYS990104	0.58	Optimized relative partition energies—method C (Miyazawa-Jernigan, 1999)
4	CHOP780214	0.54	Frequency of the 3rd residue in turn (Chou-Fasman, 1978b)
5	ENGD860101	0.54	Hydrophobicity index (Engelman et al., 1986)
6	OLSK800101	0.53	Average internal preferences (Olsen, 1980)
7	MIYS990105	0.53	Optimized relative partition energies—method D (Miyazawa-Jernigan, 1999)
8	LEVM780104	0.52	Normalized frequency of alpha-helix, unweighted (Levitt, 1978)
9	MIYS990101	0.52	Relative partition energies derived by the Bethe approximation (Miyazawa-Jernigan, 1999)
10	KIDA850101	0.50	Hydrophobicity-related index (Kidera et al., 1985)

MDGI: Mean decrease of Gini index.

The heatmap of feature importance for the DPC feature can be seen in Figure 4, from which, the 20 top-ranked informative dipeptides with the highest MDGI values are SP, TC, CG, CS, SC, TR, RT, PF, AS, HG, LI, PC, RP, AA, SL, AL, ST, IV, RR, and AD. From amongst the top 20 informative dipeptides, there are 6 dipeptides (SP, TC, CG, CS, SC, and TR) with MDGI values larger than 1.45. In addition, 4 out of the 6 top-ranked informative dipeptides (TC, CG, CS, and SC) consist of Cys, while 3 out of

the 20 top-ranked informative dipeptides (TR, RT, and RP) consist of Arg. As mentioned previously, Cys and Arg were the first and sixth important amino acids with the highest MDGI values of 15.90 and 8.31, respectively. These results reinforced the importance of Cys and Arg for the anti-angiogenic activity of peptides. Furthermore, detailed analysis of these two amino acids are discussed below.

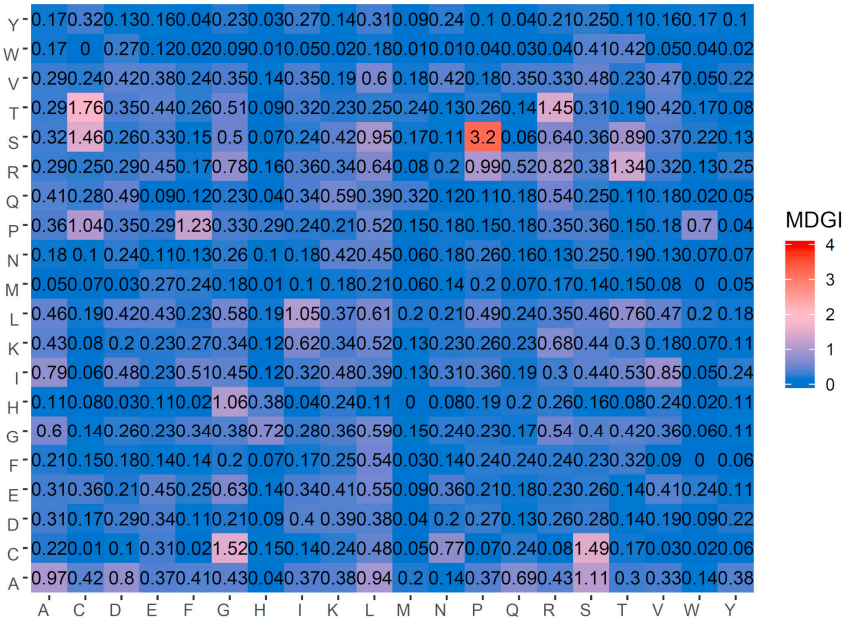

Figure 4. Heat map of the mean decrease of Gini index (MDGI) of dipeptide compositions. It should be noted that features with the largest value of MDGI are the most important.

Cys provided the largest MDGI value and results shown in Table 5 displayed that the percentage composition of Cys residues are found to be significantly different in a comparison between anti-angiogenic (0.047%) and non-antiangiogenic (0.014%) peptides producing significant p-value < 0.05. Many studies have reported that Cys is the preferred residue for anti-angiogenic activity [30–33]. Cys is classified as a polar, non-charged amino acid containing sulfur which, when oxidized, could form a disulfide bond. It stabilizes the tridimensional structure, which is essential for extracellular proteins that might be exposed to virulent conditions. Peptides containing multiple disulfide bridges are more resistant to thermal denaturation and is also crucial for maintaining their biological activity [34]. In 1997, a globular protein namely, endostatin was first discovered by Folkman and coworkers as an endogenous inhibitor of angiogenesis [35]. Mass spectrometry demonstrated that endostatin contains two disulfide bonds: Cys162-302 and Cys264-294 [31,32]. In addition, histological sections of tumors from saline-versus-endostatin-treated Lewis lung carcinomas were analyzed for apoptosis and angiogenesis. The results showed that the apoptotic index of tumor cells increased 7-folds (p-value < 0.001) while angiogenesis was completely suppressed in tumor cells (p-value < 0.001) for the endostatin treated mice [35]. Furthermore, Hiraki et al. [36] performed site-directed mutagenesis of chondromodulin-1 (ChM-1) as to assess the importance of Cys toward the function of ChM-1. The results disclosed that the ChM-1 mutant, which had all eight Cys residues replaced by Ser, lost the inhibitory effect of VEGF-A that subsequently stimulated the migration of human umbilical vein endothelial cells (HUVEC) due to the lack of disulfide bonds. Remarkably, Ser at positions 83 and 99 on the replaced ChM-1, revealed a decreased cell migration (150%) as compared to that of VEGF-A (350%). This result indicated that the disruption of one disulfide bond cannot neutralize its migratory effect. In addition, the Δ (Cys83 Cys99) rhChM-1 mutant lacking the 17 amino acid residues from Cys^{83}-Cys^{99} and but retained three disulfide bonds, still appeared to exhibit its inhibitory effect [37]. Similarly, Chlenski et al. [20] designed and synthesized two peptides consisting of FSEC (CELDENNTPMC) and FSEN (CQNHAKHGKVC) from FS-E (CQNHCKHGKVCELDENNTPMC)

by linking Cys 1 to Cys 3 and Cys 2 to Cys 4, owing to the need to construct simpler peptides with less complex structures. FS-E is classified in the group of secreted protein acidic and rich in cysteine (SPARC). In this study [20], the authors divided the experimental processes into three parts including: (i) endothelial cell migration assay (ii) inhibition of neuroblastoma tumor growth and (iii) inhibition of tumor induced angiogenesis. Firstly, in order to evaluate the capability of the two simple peptides to inhibit endothelial cell migration, HUVEC were treated with serial dilution of FSEC and FSEN by monitoring the percentage of stimulation compared with beta-fibroblast growth factor (bFGF) as a positive control. For the former, in vitro experiment was demonstrated by the inhibition of human umbilical vein endothelial cells (HUVECs) migration with an EC_{50} of 1 pM. Secondly, an in vivo experiment was demonstrated via a mice model in which mice with subcutaneous neuroblastoma xenografts were treated with the FSEC peptide for 2 weeks. The FSEC-treated mice were compared to the control group (PBS) and it was revealed that the inhibition of tumor growth was observed as deduced from the decreasing tumor weight ($p = 0.01$). Lastly, a paraffin section of xenografted mice was stained using green CD31 (PCAM-1) positive endothelial cells and red SMA-positive pericytes whereby the quantity of tumor blood vessels was calculated as the area occupied of staining. Results revealed that FSEC was significantly reduced in FSEC treated xenografts as compared to vesicle treated control (p-value < 0.001). This study also indicated that FSEC, which is a modified linear peptide containing disulfide bonds, has the ability to completely abrogate angiogenesis thereby leading to tumor growth inhibition. Their results is consistent with previous studies that SPARC can inhibit breast cancer progression [38], ovarian metastasis [33] with the overexpression of endogenous angiogenic inhibitors such as somatostatin, angiostatin, and endostatin, which also represents negative correlation with poor prognosis of cancer patients [39,40].

Furthermore, Yang X et al. [41] modified wild-type (WT) kringle5 (K5), which has been shown to contain anti-angiogenic activity with higher potential than angiostatin, by disruption of its disulfide bond distribution. K5mut1 was designed by deleting amino acid residues outside the kringle domain whereas Cys462-Cys451 is still located in the WT K5. Additionally, K5mut2 was constructed by removing Cys462, thereby leading to the loss of one disulfide bond. The effect of WT K5 and its deleted mutation on endothelial cell proliferation, cell apoptosis, and tumor growth were evaluated by the percentage of cell viability, flow cytometry and tumor weight, respectively. In vitro results showed that K5mut1 was able to decrease endothelial cell proliferation by 2-fold, enhancing endothelial cell apoptosis. Moreover, in vivo experiment was revealed that the weight of liver tumor in a mouse model was gradually decreased compared to mice treated with wild-type K5. Meanwhile, K5mut2 lacking one Cys, lost all its inhibitory effects. In summary, anti-angiogenic peptides containing Cys residues that formed disulfide bonds play an important role in (i) inhibiting blood vessel proliferation through the activation of angiostatin contributes to a lack of nutrients and blood supply to tumor cells [20,42], (ii) increasing anti-angiogenesis via reduction of specific receptors for pro-angiogenic molecules, (iii) inducing cell apoptosis [35,43], and (iv) balancing opposing signals in the tumor microenvironment [44].

Although our prediction model showed that Cys is the most important amino acid for the inhibition of blood vessel proliferation and tumor growth, other peptides which does not contain Cys have also demonstrated anti-angiogenic activity. Recent advances in biotechnology have led to the discovery of numerous biologically active peptides. The challenge is to also increase other physicochemical properties such as the bioavailability and as such pharmaceutical techniques such as liposome, hydrogel, nanoparticle, and targeted drug delivery system should be utilized for improvement of the potency of anti-angiogenic peptides. For example, tumstatin peptides binds to avB3 integrin on proliferating endothelial cells and also localizes to the target tumor. Moreover, when combined with bevacizumab (anti-VEGF antibody), an increase in its efficacy against tumor progression was observed [45]. Thus, the design of therapeutic peptides utilizes appropriate amino acids for bringing about the intended effect as to target specific mechanisms of interest. Representing the sixth largest MDGI value (Table 5), the percentage composition of Arg residues is found to be significantly different between anti-angiogenic

(0.088%) and non-antiangiogenic (0.055%) peptides at a significance level of p-value < 0.05. Bae et al. [46] identified hexapeptides from peptide libraries in order to investigate their effects on the binding of VEGF to their receptors. The authors found that the most important amino acids for inhibitory activity included Arg, Lys, and His. Meanwhile, three peptides RRKRRR, RKKRKR, and hexa-arginine (RRRRRR) were demonstrated to be the most effective inhibitor with IC_{50} values of 2, 3.4, and 3.8 μM, respectively. In addition, the interaction between hexapeptides and VEGF was investigated by monitoring the binding of labeled $VEGF_{165}$ to endothelial cells. Results showed that Arg-rich (AR) hexapeptides directly binds to $VEGF_{165}$ (K_D = 5, 2 and 22 μM). Furthermore, the proliferation assay also confirmed that AR hexapeptides inhibited HUVE cell proliferation by $VEGF_{165}$ in a concentration-dependent manner without cytotoxicity. Moreover, the essential role of hexapeptides containing basic charged amino acid resides was elucidated via blocking the metastasis of human colon carcinoma cells. Results disclosed that RRKRRR decreased the number of metastatic nodules by 16% as compared to that of the control whereas hexa-Lys (KKKKKK) showed minor inhibitory effects (80% of control). Conversely, the peptide with negative charge (EEFDDA) appeared to show no inhibitory activity at all. In addition, Xiong et al. [47] demonstrated that treatment of cells with 0.05 mmol/L of L-Arg for 7 days caused endothelial dysfunction as measured by the enhanced superoxide anion and decreased NO production. Thus, the chronic L-Arg supplementation is potent for accelerating endothelial cell senescence expression with the up-regulation of Arg-II. Moreover, Arg was utilized to create a synthetic RGD (Arg-Gly-Asp) integrin ligand sequence for improving the tumor cell targeting capability of therapeutic peptides [48]. Xu et al. [49] synthesized HM3 peptide (IRRADRAAVPGGGG) and added RGD (IRRADRAAVPGGGG-RGD) in their investigation on the inhibitory effect. The experimental result showed that it could significantly inhibit the migration of the HM3 peptide into endothelial cells. Besides, Matrigel and aortic ring tests conducted in a mice model also revealed that HM3 could potentially inhibit angiogenesis. Similarly, Buerkle et al. [50] explored the effect of cyclic RGD peptide as an α_v-integrin antagonist on angiogenesis, microcirculation, growth, and metastatic formation of solid tumors. Results indicated that the cyclic RDG peptide reduced blood vessel density as well as diminished tumor growth and metastasis. Additionally, Kando et al. [51] developed a liposomal drug targeted to membrane type-1 matrix metalloproteinase by modification with stearoyl Gly-Pro-Leu-Pro-Leu-Arg (GPLPLR). The authors observed that the modified liposome showed high binding ability to HUVEC and increased accumulation in tumor cell (> 4-fold). In summary, peptides containing Arg induced anti-angiogenic activity and contributed to the inhibition of tumor growth via (i) the binding of peptides to the main body of VEGF including the N- and C-terminal ends [46] and (ii) increasing the specificity to targeted tumors as Arg confers a small positive charge there allowing cell binding via electrostatic interactions with the negatively charged cell membranes thus, leading to arrested tumor growth [52].

2.4. Mechanistic Interpretation of Informative PCP

Physicochemical properties of amino acids play an essential role as effective features for identifying and characterizing the functions of protein or peptide from their primary sequences [53–57]. It is well known that PCP [58], such as molecular volume, exposure or accessible surface, polarity (hydrophobicity/hydrophilicity), charge/pK, hydrogen-bonding potential and so forth are correlated with the structure and function of the amino acid sequence [59]. Herein, we have obtained the 10 top-ranked informative PCPs corresponding to their MDGI values, as shown in Table 6. As seen in Table 6, CHOP780215, CHOP780214, and CHOP780213 represents the second, third, and fourth important PCPs with corresponding MDGI values of 0.61, 0.54, and 0.50, respectively. Meanwhile, another important PCP is CHOP780209 with a corresponding MDGI value of 0.34 was not found in the top 10 informative PCPs. The four important PCPs of anti-angiogenic peptides were analyzed and discussed below.

2.4.1. Peptides Having Cys Locating at the C-terminal Domain Can Decrease Endothelial Formation Activity and Suppress Tumor Growth

CHOP780209 with a corresponding MDGI value of 0.34 is described as the normalized frequency of the C-terminal β-sheet. The secondary structure prediction by the Chou-Fasman method demonstrated that the conformational preferences of Cys in adopting the β-strand structure is 1.19 [60]. It is well-known that the β-strand is a stretch of polypeptide chain containing approximately 3–10 residues in length. The interaction among more than two β-strands (around six β-strands) via hydrogen bonds could form the β-sheet structure [61]. Based on this PCP, it could be stated that the β-sheet structure and Cys residue are important for anti-angiogenic activities of peptides. Previous studies reported that endostatin [62,63], thrombospondin (TSP) [64,65], somatostatin [66], ChM-I [67], and TeM [65] containing Cys rich domain at the C-terminal region is likely to adopt the β-sheet structure. Figure 5 shows the three-dimensional structures of endostatin (a), somatostatin (b) and Platelet factor-4 (c). Moreover, an endogenous angiogenic inhibitor was revealed for controlling angiogenic balance. Hohenester et al. [63] reported that the crystal structure of endogenous angiogenic inhibitor namely endostatin, is a fragment derived from the C-terminal domain of collagen XVIII, containing two disulfide bridges located in the β-sheet [63]. After blocking angiogenesis, it is accompanied by high proliferation that is balanced by apoptosis in tumor cells [35]. In addition, Talaboletti et al. [64] showed that TSP has the ability to inhibit tumor cells migration. TSP-1 is an essential fragment of TSP at the C-terminal region that includes four Cys residues (two intra disulfide bonds). TSP-1 binds to CD36 receptors on endothelial cells that could allow for endothelial-cell apoptosis thereby leading to the inhibition of angiogenesis [65]. Furthermore, Ginj et al. [66] synthesized and evaluated the biological activity of somatostatin-based radiopeptides. The authors found that the peptides improved the binding affinity toward tumor cells and enhanced the internalization into cells expressing somatostatin receptor [44]. Hiraki et al. [36] revealed that mature human chondromodulin-I (ChM-I) consists of 120 amino acids and the C-terminal hydrophobic domain (Phe42 to Val120) in the β-sheet region indicates a functional domain for the inhibition of vascular endothelial cell growth in vitro. In order to present the structural requirements for ChM-I to exert its anti-angiogenic activity, Miura et al. [37] observed that the C-terminal domain containing eight Cys residues successfully inhibited cell migration with the decrease in percentage of migrated cells (200%) as compared to VEGF-A (>400%). Additionally, Oshima et al. [67] reported that the linear peptide, Tenomodulin (TeM), has the potential to be an anti-angiogenic peptide in vitro and in vivo. TeM is a well-known cartilage-derived angiogenesis inhibitor containing eight cysteine residues and a unique disulfide bridged hydrophobic domain at the C-terminal region, which adopts a β-sheet structure. The authors confirmed the functional role of Cys at the C-terminal region of TeM that are essential for the anti-angiogenic activity by monitoring the Matrigel tube formation and measuring the tube length. They further cleaved and constructed a secreted C-terminal domain of TeM (shTEM) containing the Cys-rich domain from human TeM. As in vitro result, their experiment showed that shTeM had more potential to restrain HUVEC cells at a low concentration of 50 μM when compared to the MOCK control (220 μM). In addition, they demonstrated in vivo experimental result that shTeM transduced melanoma cells formed tumors in C57BL/6 mice that were 46% smaller than enhanced green fluorescent protein (EGFP) transduction. In summary, Cys present at the C-terminal domain is one crucial factor influencing synthetic anti-angiogenic peptides in order to decrease endothelial formation activity and suppress tumor growth.

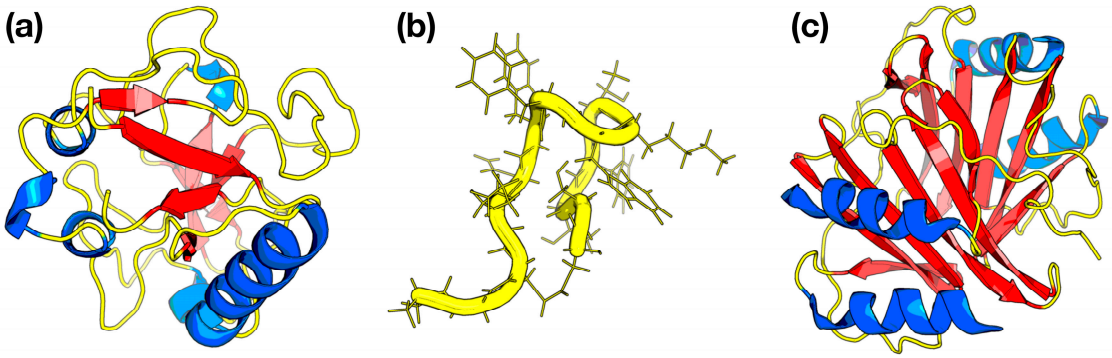

Figure 5. Three-dimensional structures of established anti-angiogenic inhibitors consisting of endostatin (PDB id 1KOE) (**a**), somatostatin (PDB id 2MI1) (**b**), and Platelet factor-4 (PDB id 1RHP) (**c**). α-helix, β-sheet, and loop are shown in blue, red and yellow colors, respectively.

2.4.2. Cyclic Disulfide-Rich Peptides Provide Greater Inhibition of Angiogenesis and Cell Migration, Selectivity, and Stability than Linear Peptides

The second, third, and fourth important PCPs (CHOP780215, CHOP780214, and CHOP780213, respectively) describes the turn of a protein structure. β-turn is thought to be involved in protein folding initiation while its conformational structure also contributes to protein stability and the free energy of proteins [68]. Moreover, the Chou-Fasman method of secondary structure prediction showed that the conformational preferences of Cys to be β-turn is 1.19 [60]. Many studies have reported that the increase in stability of protein structures as afforded by the β-turn is represented by disulfide bridges of the CXXC motif that is formed when the peptide brings together the first and fourth Cys to form a cyclic disulfide bond [69,70]. Although disulfide bonds located in linear peptides demonstrated anti-angiogenic activity, however, many researchers used the cyclic modified peptide conformations in order to promote their stability. Accordingly, cyclic peptides were able to exhibit enhanced cell permeability, increased bioavailability, and binding specificity [71]. Miura et al. [37] revealed the structural requirements of ChM-1 for enhancing its anti-angiogenic property by observing its prevention of the VEGF-A induced migration of HUVEC. The result showed that the synthetic ChM-1 cyclic peptide linked by disulfide bond between Cys^{83} and Cys^{99} promoted a migratory effect via a dose response curve (ID_{50} value of 2 μM) as compared to the ChM-1 linear peptide. In addition, the ChM-1 cyclic peptide was modified at the hydrophobic C-terminal tail and then examined for its effects on tumor angiogenesis. The authors revealed that the tailed ChM-1 cyclic peptide clearly decreased the tumor volume. Similarly, researchers [72,73] have stated that cyclic disulfide-rich peptides, including six inter-cysteine loops (namely *Momordica cochinchinensis* trypsin inhibitor-II (MCoTI-II)), have a high enzymatic and thermal stability. Thus, Chan et al. [19] designed a second-generation grafted cyclic peptide (MCoAA-02) combined with anti-angiogenic epitopes (somatostatin (SST)-01 and pigment epithelium-derived factor (PEDF)) [74]. The authors observed a high potency of 50% inhibition of HUVEC migration at 1 nM using the MCoAA-02 peptide. Based on its potency, MCoAA-02 was evaluated using the chorioallantoic membrane (CAM) assay to monitor its effects on blood vessel growth in vivo. Results revealed that MCoAA-02 had comparable capability to that of cyclic peptide-based drugs namely, octreotide at 10 μM. In addition, MCoAA-02 also exhibited high stability as compared to orally active sunitimib as observed through the percentage of peptide remaining over 24 h in a serum stability assay [19]. In conclusion, re-engineering linear peptides into cyclic peptides resulted in enhancing the inhibition of angiogenesis, cell migration, selectivity, and stability. Furthermore, the combination between cyclic peptides and potent anti-angiogenic agents provided a synergistic effect which poses an opportunity for the exploration of better therapeutic peptides [75–78].

2.5. TargetAntiAngio Web Server

To afford wide utilization of the prediction capability of the QSAR model, we had constructed a web server called the *TargetAntiAngio*. The web interface was established using the *Shiny* package under the R programming environment. The web server is freely accessible at http://codes.bio/targetantiangio/ (accessed on: 1 April 2019). Screenshots of the TargetAntiAngio web server is shown in Figure 6. A step-by-step guide on using the web server to get the desired results is given below:

- Step 1. Open the web server TargetAntiAngio at http://codes.bio/targetantiangio/ (accessed on: day 1 April 2019)
- Step 2. Either enter the query sequence into the Input box or upload the sequence file by clicking on the "Choose file" button (i.e., found below "the Enter your input sequence(s) in FASTA format heading").
- Step 3. Press on the "Submit" button to initiate the prediction process.
- Step 4. Prediction results are automatically displayed in the grey box found below the "Status/Output" heading. Typically, it takes a few seconds for the server to process the task. Users can also download the prediction results as a CSV file by pressing on the "Download CSV button".

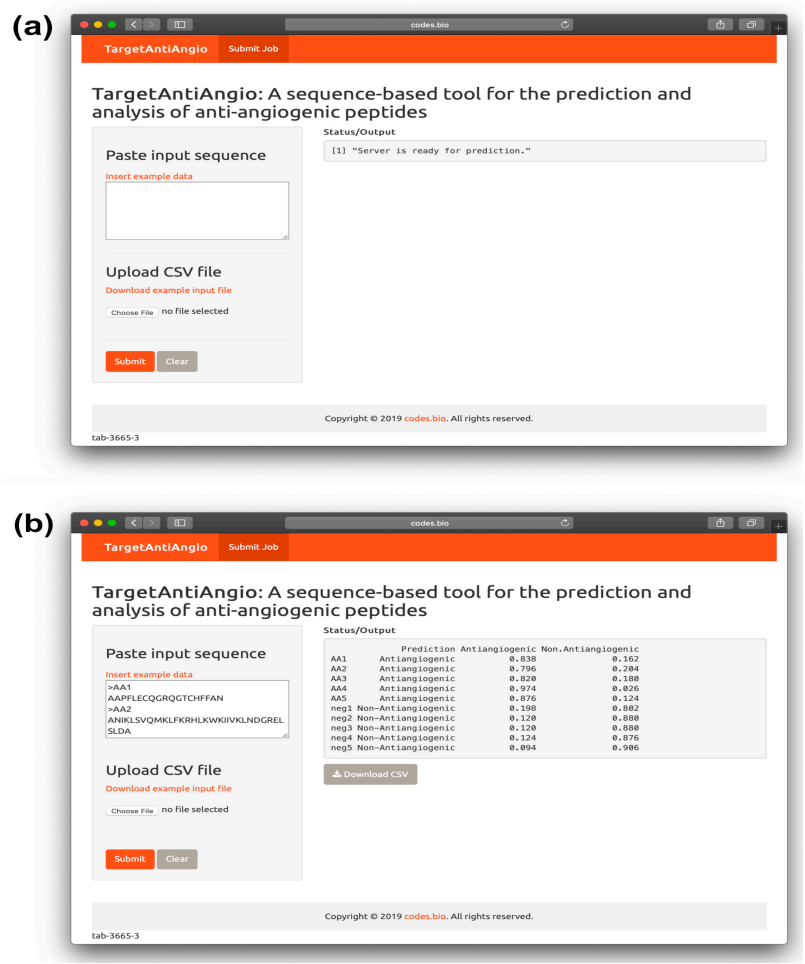

Figure 6. Screenshot of the TargetAntiAngio web server before (**a**) and after (**b**) submission of the input query sequence.

Additionally, users could also run a local copy of TargetAntiAngio on their own computer using a one-line code as follows in an R environment:

shiny::runGitHub('targetantiangio', 'chaninlab', subdir = " targetantiangio_shiny_server ")

However, prior to running the aforementioned code, it is recommended that users first install the prerequisite R packages. This can be performed by using the following code:

install.packages(c('shiny', 'shinyjs', 'shinythemes', 'protr', 'seqinr', 'caret', 'markdown')).

3. Materials and Methods

3.1. Benchmark Dataset

The first and most important consideration for developing a promising computational model is to construct a reliable benchmark dataset. In this study, the benchmark dataset was obtained from the work of Ramaprasad et al. [27], which has been used for developing recent prediction models of anti-angiogenic peptides [28,29]. Initially, the benchmark dataset had 257 peptide sequences in the anti-angiogenic class as derived from various articles and patents. To obtain a good quality benchmark dataset, the following steps were considered. Firstly, to avoid a dataset containing many redundant peptides, anti-angiogenic peptides having >90% of sequence similarity was filtered out using the CD-HIT program [79]. Secondly, anti-angiogenic peptides containing special characters, such as X and U, were removed. After such screening procedures, a set of 135 peptide sequences belonging to the anti-angiogenic class was obtained. Due to the lack of peptide sequences for the non-antiangiogenic class, 135 random peptides were used as non-antiangiogenic peptides. The benchmark dataset (S_{main}) used in this study can be summarized by the following formula:

$$S_{main} = S^{+}_{main} \cup S^{-}_{main} \quad (1)$$

where S^{+}_{main} and S^{-}_{main} represents peptide sequences of anti-angiogenic and non-antiangiogenic classes, respectively, from the S_{main} dataset while the symbol $\cup$ represents the union from the set theory. Based on the study of Ramaprasad et al. [27], AntiAngioPred was developed by using the NT15 dataset, which provided the highest prediction accuracy thus far. Therefore, to make a fair comparison with this method, the N-terminus dataset was considered. In this respect, the NT15 dataset containing the first 15 residues from the N-terminus region of peptide was used in our comparative investigation. After preparing the NT15 dataset, it consisted of 99 anti-angiogenic and 101 non-antiangiogenic peptides. The NT15 dataset (S_{NT15}) used in this study can be formulated as:

$$S_{NT15} = S^{+}_{NT15} \cup S^{-}_{NT15} \quad (2)$$

where S^{+}_{NT15} and S^{-}_{NT15} represents the peptide sequences of anti-angiogenic and non-antiangiogenic classes, respectively, from the S_{NT15} dataset. Moreover, the S_{main} and S_{NT15} datasets were randomly divided into training (for cross-validation test) and testing (for independent validation test) sets, where 80% and 20% of the two datasets were used as training and testing sets. A summary of the size distribution of the training and testing sets is provided in Table 7.

Table 7. Summary of two datasets for evaluating the predictors of anti-angiogenic peptides as obtained from Ramaprasad et al.

Dataset	S_{main}		S_{NT15}	
	Anti-angio	**Non-anti-angio**	**Anti-angio**	**Non-anti-angio**
Original data	137	137	99	101
Training set	101	101	80	80
Testing set	36	36	19	21

Anti-angio and non-anti-angio represent anti-angiogenic and non-antiangiogenic peptides, respectively.

3.2. Feature Representation

In order to develop a robust and interpretable sequence-based computational model, the critical process is to represent the peptides in such a way so as to afford a comprehensive and proper description of the feature that could well reflect their functions. Amongst the various types of sequence features that are available, easy and interpretable features are those pertaining to the amino acid composition (AAC), dipeptide composition (DPC), and physicochemical properties (PCP).

AAC and DPC are the proportions of each amino acid and dipeptide in a peptide sequence **P** that are expressed as fixed lengths of 20 and 400, respectively. Thus, in terms of AAC and DPC features, a peptide **P** can be expressed by vectors with 20D and 400D (dimension) spaces, respectively, as formulated by:

$$\mathbf{P} = [\mathrm{aa}_1, \mathrm{aa}_2, \ldots, \mathrm{aa}_{20}]^{\mathbf{T}} \tag{3}$$

$$\mathbf{P} = \left[\mathrm{dp}_1, \mathrm{dp}_2, \ldots, \mathrm{dp}_{400}\right]^{\mathbf{T}} \tag{4}$$

where T is the transposed operator, while aa_1, aa_2, … , aa_{20} and dp_1, dp_2, … , dp_{400} are occurrence frequencies of the 20 and 400 native amino acids and dipeptides, respectively, in a peptide sequence **P**.

PCP is one of the most intuitive features associated with biophysical and biochemical reactions. In fact, a total of 531 PCPs without NA values were derived from version 9.0 of the Amino acid index database (AAindex) [58], which is a collection of the published literature pertaining to different physicochemical and biophysical properties of amino acids and pairs of amino acids (http://www.genome.jp/aaindex/). Each physicochemical property consisted of a set of 20 numerical values for amino acids. The PCP feature has been extensively used for the prediction and analysis of various protein [53,55,80] and peptide [56,57] functions. To utilize PCP features for extracting a peptide sequence, peptide with the length of L amino acid residues is encoded into an L-dimensional vector of 531 PCPs (531D).

As mentioned in previous studies [81,82] and shown in Equations (3) and (4), AAC and DPC features only provide the compositional information of a peptide sequence, but all the sequence-order information may be completely lost. To remedy this limitation, the pseudo amino acid composition (PseAAC) and amphiphilic pseudo amino acid composition (Am-PseAAC) approaches were proposed by Chou [81,82]. According to Chou's PseAAC, the general form of PseAAC for a peptide **P** is formulated by:

$$\mathbf{P} = [\Psi_1, \Psi_2, \ldots, \Psi_u, \ldots, \Psi_\Omega]^{\mathbf{T}} \tag{5}$$

where the subscript Ω is an integer to reflect the feature's dimension. The value of Ω and the component of Ψ_u, where $u = 1, 2, \ldots, \Omega$ is dependent on the protein or peptide sequences. In this study, the parameters of PseAAC (i.e., the discrete correlation factor λ and weight of the sequence information ϖ) were estimated by using the optimization procedure as described hereafter. The dimension of PseAAC feature is $20 + \lambda \times \varpi$. Since the hydrophobic and hydrophilic properties of proteins play an important role in the folding and interaction of proteins, Am-PseAAC was introduced by Chou [82]. The dimension of Am-PseAAC feature is $20 + 2\lambda$. The first 20 components are the 20 basic AAC ($p_1, p_2, \ldots, p_{20}$) while the next 2λ ones denote the set of correlation factors that reveal the physicochemical properties such as hydrophobicity and hydrophilicity along a protein or peptide sequence as formulated by:

$$\mathbf{P} = [p_1, p_2, \ldots, p_{20}, p_{20+\lambda}, p_{20+\lambda+1}, \ldots p_{20+2\lambda}]^{\mathbf{T}} \tag{6}$$

In this study, the five aforementioned features of peptide sequences were generated by using the protr package in the R programming environment [83]. The parameters of PseAAC (weight^1 and lamda^1) and Am-PseAAC (weight^2 and lamda^2) were optimized by varying weight and lambda values from 0 to 1 and 1 to 10 with step sizes of 0.1 and 1, respectively, on the whole S_{main} and S_{NT15} datasets as assessed by a 5-fold CV procedure.

3.3. Random Forest

The learning classifier employed herein was constructed using the original RF algorithm [84,85]. This model is an ensemble model consisting of many classification and regression tree (CART) classifiers that improves prediction performances of CART classifiers by growing a number of weak CART classifiers. Prediction results from the classification task is obtained by using simple voting from amongst outputs all trees to arrive at one final prediction. In regression, a final prediction is the average of prediction results from many trees. In order to construct the RF model, each tree is built as follows: (i) a bootstrap sample, which is used as a training set for the current tree, is obtained from the whole training set consisting of N peptides. In the meanwhile, peptides which are not used for constructing the current tree are place in an out-of-bag (OOB) set, where the size of the OOB set is around N/3, (ii) the m selected features from the whole M features is derived from the best split by CART model, (iii) each tree is grown to the largest possible extent, (iv) if there is no pruning then the procedure is terminated. Herein, the RF classifier was established using the randomForest package in the R software [84]. To enhance the performance of the RF model, two parameters including ntree (i.e., the number of trees used for constructing the RF classifier) and mtry (i.e., the number of random candidate features) were determined using the caret R package [86] with the 5-fold CV approach. The search space of ntree and mtry are {100,200, . . . ,500} and {1,2, . . . ,10} with the steps of 100 and 1, respectively. Previously, RF model has been successfully used in the prediction of various functions and properties of peptides and proteins [55–57,87,88] as well as other biological or chemical entities [89–93].

3.4. Identification of Important Features

In this work, we performed the analysis and identification of feature importance for each type of sequence feature by using the RF method to provide a better understanding of the biophysical and biochemical properties of anti-angiogenic activities of peptides. In RF method, the OOB approach is used for evaluating the feature importance as follows: (i) two-thirds of the training data is utilized to construct the predictive classifier while the remaining is used for evaluating the performance of such classifier and (ii) the feature importance of each feature can be evaluated by measuring the decrease of the prediction performance. It should be noted that the performance evaluation of the model can be either accuracy or Gini index. In summary, the RF method provides two measures for ranking feature importance, i.e., the mean decrease of Gini index (MDGI) and the mean decrease of prediction accuracy. Since Calle and Urrea [94] demonstrated that the MDGI provided a more robust result as compared to the mean decrease of prediction accuracy, we utilized the MDGI value to rank the importance of interpretable features including AAC, DPC, and PCP. Until now, these three features have been used to characterize many peptides and proteins, such as predicting HIV-1 CRF01-AE co-receptor usage [95], predicting protein crystallization [53,96], predicting the oligomeric states of fluorescent proteins [88], predicting the bioactivity of host defense peptides [87], prediction of human leukocyte antigen gene [80,97], predicting antifreeze proteins [55], predicting the hemolytic activity of peptides [56], and predicting antihypertensive activity of peptides [57].

The Gini index can be defined as MDGI is an impurity measure that corresponds to the ability of each feature in discriminating the sample classes. The Gini index can be defined as

$$1 - \sum_{c=1}^{2} p^2(c|t) \tag{7}$$

where $p^2(c|t)$ denotes the estimated class probability for node t in a tree classifier and c in the class label (i.e., either anti-angiogenic or non-antiangiogenic peptides). In order to increase the reliability for identifying the feature importance, 10 RF models were constructed by varying the mtry parameter settings from 2 to 20 (mtry = 2, 3, 5, 7, 9, 11, 13, 15, 17, 20) and fixing the ntree parameter with 100 [55–57]. Finally, the average value of MDGI on 10 runs of feature importance estimations were used in this study. Features with the largest MDGI value is considered to be an important feature as it significantly

contributes to the prediction performance. Herein, the MDGI values of feature importance for each type of sequence feature was estimated using the randomForest package in the R software [84].

3.5. Performance Evaluation

In a statistical prediction, the following three testing methods are often used to evaluate the prediction performance in practical applications: sub-sampling test or k-fold cross-validation (k-fold CV), jackknife test, and independent validation test or external test. The sub-sampling and jackknife test are popular cross-validation methods to assess the predictive capability of the model. Meanwhile, the external test is considered as one of the most rigorous and reliable methods for the cross-validation purposes in statistics. In k-fold cross-validation procedure, the training set is randomly separated into k subsets. From the k subsets, a single subset is taken as the testing set to validate the prediction model that is trained and learned by the remaining k-1 subsets. This process is repeated k times, until each subset had been used as the testing set. During the jackknifing process, a single sample in the whole dataset having N samples is taken as the testing set and the remaining N-1 samples are used for training the model. This process is repeated N times, until each ssample had been used as the testing set.

In order to evaluate the prediction ability of the model, the following sets of four metrics are used as follows:

$$\mathrm{Ac} = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \tag{8}$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{9}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})} \tag{10}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{11}$$

where Ac, Sn, Sp, and MCC are called accuracy, sensitivity, specificity, and Matthews correlation coefficient, respectively. TP, TN, FP, and FN represent the instances of true positive, true negative, false positive, and false negative, respectively. Moreover, in order to evaluate the prediction performance of models using threshold-independent parameters, the receiver operating characteristic (ROC) curves were plotted by the pROC package in the R software [98]. The area under the ROC curve (auAUC) was used to measure the prediction performance, where AUC values of 0.5 and 1 are indicative of perfect and random models, respectively.

3.6. Reproducible Research

To ensure the reproducibility of the models proposed herein, all R codes and datasets used in the construction of the predictive models, graphical figures and the TargetAntiAngio web server are available on GitHub at https://github.com/Shoombuatong2527/targetantiangio (accessed on: 1 April 2019) and https://github.com/chaninlab/targetantiangio-webserver (accessed on: 1 April 2019).

4. Conclusions

Anti-angiogenesis plays a fundamental role in tumor growth, invasion, and metastatic dissemination. Several anti-angiogenic peptides have been developed in order to promote effective cancer treatment as well as enhanced survival rate. Therefore, computational methods that can predict and analyze anti-angiogenic peptides based on peptide sequences are highly desirable. In this study, we have developed a new computational model named TargetAntiAngio for predicting and analyzing anti-angiogenic peptides based on sequence information. TargetAntiAngio is developed using the random forest classifier in conjunction with a combination of amino acid composition, pseudo amino acid composition and amphiphilic pseudo amino acid composition. The prediction results

for both cross-validation and independent validation tests on the benchmark dataset demonstrated that TargetAntiAngio can pick out informative features as well as improve prediction performances. In addition, a thorough analysis of the peptide feature importance was conducted to unravel and rationalize the biophysical and biochemical properties of anti-angiogenic activities of peptides. Finally, to help potential users of TargetAntiAngio, a web-server based on the optimal model has been established and made freely available online at http://codes.bio/targetantiangio/ thereby allowing users easy access to their desired results.

Author Contributions: W.S. conceived, designed, performed, and analyzed the experiments. V.L., N.S., P.N., V.P., and C.N. analyzed the data. W.S. and V.L. drafted the manuscript. W.S. and C.N. contributed the code for constructing the web server. C.N. vetted the manuscript. All authors read and approved the manuscript.

Funding: This work is supported by the TRF Research Grant for New Scholar (No. MRG6180226) and the TRF Research Career Development Grant (No. RSA6280075) from the Thailand Research Fund, the Office of Higher Education Commission and Mahidol University; and the New Researcher Grant (A31/2561) from Mahidol University.

Acknowledgments: We thank the reviewers for their great comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674. [CrossRef] [PubMed]
2. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2018. *CA Cancer J. Clin.* **2018**, *68*, 7–30. [CrossRef] [PubMed]
3. Zhang, H.; Chen, J. Current status and future directions of cancer immunotherapy. *J. Cancer* **2018**, *9*, 1773–1781. [CrossRef] [PubMed]
4. Zugazagoitia, J.; Guedes, C.; Ponce, S.; Ferrer, I.; Molina-Pinelo, S.; Paz-Ares, L. Current challenges in cancer treatment. *Clin. Ther.* **2016**, *38*, 1551–1566. [CrossRef] [PubMed]
5. Stephenson, J.; Goddard, J.; Al-Taan, O.; Dennison, A.; Morgan, B. Tumour angiogenesis: A growth area—From John Hunter to Judah Folkman and beyond. *J. Cancer Res.* **2013**, *2013*. [CrossRef]
6. Kubota, Y. Tumor angiogenesis and anti-angiogenic therapy. *Keio J. Med.* **2012**, *61*, 47–56. [CrossRef] [PubMed]
7. Sund, M.; Zeisberg, M.; Kalluri, R. Endogenous stimulators and inhibitors of angiogenesis in gastrointestinal cancers: Basic science to clinical application. *Gastroenterology* **2005**, *129*, 2076–2091. [CrossRef] [PubMed]
8. Lenz, H.-J. Antiangiogenic agents in cancer therapy. *Oncology* **2005**, *19*, 17–25. [PubMed]
9. Senger, D.R.; Claffey, K.P.; Benes, J.E.; Perruzzi, C.A.; Sergiou, A.P.; Detmar, M. Angiogenesis promoted by vascular endothelial growth factor: Regulation through α1β1 and α2β1 integrins. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 13612–13617. [CrossRef]
10. Johnson, K.E.; Wilgus, T.A. Vascular endothelial growth factor and angiogenesis in the regulation of cutaneous wound repair. *Adv. Wound Care* **2014**, *3*, 647–661. [CrossRef]
11. Shih, T.; Lindley, C. Bevacizumab: An angiogenesis inhibitor for the treatment of solid malignancies. *Clin. Ther.* **2006**, *28*, 1779–1802. [CrossRef] [PubMed]
12. Su, Y.; Yang, W.-B.; Li, S.; Ye, Z.-J.; Shi, H.-Z.; Zhou, Q. Effect of angiogenesis inhibitor bevacizumab on survival in patients with cancer: A meta-analysis of the published literature. *PLoS ONE* **2012**, *7*, e35629. [CrossRef] [PubMed]
13. Kim, A.; Balis, F.M.; Widemann, B.C. Sorafenib and sunitinib. *Oncologist* **2009**, *14*, 800–805. [CrossRef] [PubMed]
14. Grandinetti, C.A.; Goldspiel, B.R. Sorafenib and sunitinib: Novel targeted therapies for renal cell cancer. *Pharmacother. J. Hum. Pharmacol. Drug Ther.* **2007**, *27*, 1125–1144. [CrossRef] [PubMed]
15. Rosca, E.V.; Koskimaki, J.E.; Rivera, C.G.; Pandey, N.B.; Tamiz, A.P.; Popel, A.S. Anti-angiogenic peptides for cancer therapeutics. *Curr. Pharm. Biotechnol.* **2011**, *12*, 1101–1116. [CrossRef] [PubMed]
16. Lee, E.; Lee, S.J.; Koskimaki, J.E.; Han, Z.; Pandey, N.B.; Popel, A.S. Inhibition of breast cancer growth and metastasis by a biomimetic peptide. *Sci. Rep.* **2014**, *4*, 7139. [CrossRef] [PubMed]

17. Foy, K.C.; Liu, Z.; Phillips, G.; Miller, M.; Kaumaya, P.T. Combination treatment with HER-2 and VEGF peptide mimics induces potent anti-tumor and anti-angiogenic responses in vitro and in vivo. *J. Biol. Chem.* **2011**, *286*, 13626–13637. [CrossRef]
18. Wong, W. Combining anti-inflammatory and anti-angiogenic therapy. *Sci. Signal.* **2013**, *6*, ec224. [CrossRef]
19. Chan, L.Y.; Craik, D.J.; Daly, N.L. Dual-targeting anti-angiogenic cyclic peptides as potential drug leads for cancer therapy. *Sci. Rep.* **2016**, *6*, 35347. [CrossRef]
20. Chlenski, A.; Guerrero, L.J.; Peddinti, R.; Spitz, J.A.; Leonhardt, P.T.; Yang, Q.; Tian, Y.; Salwen, H.R.; Cohn, S.L. Anti-angiogenic SPARC peptides inhibit progression of neuroblastoma tumors. *Mol. Cancer* **2010**, *9*, 138. [CrossRef]
21. Park, S.W.; Cho, C.S.; Jun, H.O.; Ryu, N.H.; Kim, J.H.; Yu, Y.S.; Kim, J.S.; Kim, J.H. Anti-angiogenic effect of luteolin on retinal neovascularization via blockade of reactive oxygen species production. *Investig. Ophthalmol. Vis. Sci.* **2012**, *53*, 7718–7726. [CrossRef] [PubMed]
22. Kong, J.S.; Yoo, S.A.; Kim, J.W.; Yang, S.P.; Chae, C.B.; Tarallo, V.; Falco, S.D.; Ryu, S.H.; Cho, C.S.; Kim, W.U. Anti–neuropilin-1 peptide inhibition of synoviocyte survival, angiogenesis, and experimental arthritis. *Arthritis Rheum. Off. J. Am. Coll. Rheumatol.* **2010**, *62*, 179–190. [CrossRef] [PubMed]
23. Mahlapuu, M.; Håkansson, J.; Ringstad, L.; Björn, C. Antimicrobial peptides: An emerging category of therapeutic agents. *Front. Cell. Infect. Microbiol.* **2016**, *6*, 194. [CrossRef] [PubMed]
24. Recio, C.; Maione, F.; Iqbal, A.J.; Mascolo, N.; De Feo, V. The potential therapeutic application of peptides and peptidomimetics in cardiovascular disease. *Front. Pharmacol.* **2017**, *7*, 526. [CrossRef] [PubMed]
25. Lau, J.L.; Dunn, M.K. Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic Med. Chem.* **2018**, *26*, 2700–2707. [CrossRef] [PubMed]
26. Sulochana, K.; Ge, R. Developing antiangiogenic peptide drugs for angiogenesis-related diseases. *Curr. Pharm. Des.* **2007**, *13*, 2074–2086. [CrossRef] [PubMed]
27. Ramaprasad, A.S.E.; Singh, S.; Venkatesan, S. AntiAngioPred: A server for prediction of anti-angiogenic peptides. *PLoS ONE* **2015**, *10*, e0136990.
28. Blanco, J.L.; Porto-Pazos, A.B.; Pazos, A.; Fernandez-Lozano, C. Prediction of high anti-angiogenic activity peptides in silico using a generalized linear model and feature selection. *Sci. Rep.* **2018**, *8*, 15688. [CrossRef] [PubMed]
29. Zahiri, J.; Khorsand, B.; Yousefi, A.A.; Kargar, M.; Zade, R.S.H.; Mahdevar, G. AntAngioCOOL: Computational detection of anti-angiogenic peptides. *J. Transl. Med.* **2019**, *17*, 71. [CrossRef] [PubMed]
30. Jia, H.; Lohr, M.; Jezequel, S.; Davis, D.; Shaikh, S.; Selwood, D.; Zachary, I. Cysteine-rich and basic domain HIV-1 Tat peptides inhibit angiogenesis and induce endothelial cell apoptosis. *Biochem. Biophys. Res. Commun.* **2001**, *283*, 469–479. [CrossRef] [PubMed]
31. Agarwal, A.; Munoz-Nájar, U.; Klueh, U.; Shih, S.-C.; Claffey, K.P. N-acetyl-cysteine promotes angiostatin production and vascular collapse in an orthotopic model of breast cancer. *Am. J. Pathol.* **2004**, *164*, 1683–1696. [CrossRef]
32. John, H.; Forssmann, W.G. Determination of the disulfide bond pattern of the endogenous and recombinant angiogenesis inhibitor endostatin by mass spectrometry. *Rapid Commun. Mass Spectrom. RCM* **2001**, *15*, 1222–1228. [CrossRef] [PubMed]
33. Naczki, C.; John, B.; Patel, C.; Lafferty, A.; Ghoneum, A.; Afify, H.; White, M.; Davis, A.; Jin, G.; Kridel, S.; et al. SPARC inhibits metabolic plasticity in ovarian cancer. *Cancers* **2018**, *10*, 385. [CrossRef] [PubMed]
34. Muskal, S.M.; Holbrook, S.R.; Kim, S.H. Prediction of the disulfide-bonding state of cysteine in proteins. *Protein Eng.* **1990**, *3*, 667–672. [CrossRef] [PubMed]
35. O'Reilly, M.S.; Boehm, T.; Shing, Y.; Fukai, N.; Vasios, G.; Lane, W.S.; Flynn, E.; Birkhead, J.R.; Olsen, B.R.; Folkman, J. Endostatin: An endogenous inhibitor of angiogenesis and tumor growth. *Cell* **1997**, *88*, 277–285. [CrossRef]
36. Hiraki, Y.; Mitsui, K.; Endo, N.; Takahashi, K.; Hayami, T.; Inoue, H.; Shukunami, C.; Tokunaga, K.; Kono, T.; Yamada, M.; et al. Molecular cloning of human chondromodulin-I, a cartilage-derived growth modulating factor, and its expression in Chinese hamster ovary cells. *Eur. J. Biochem.* **1999**, *260*, 869–878. [CrossRef] [PubMed]

37. Miura, S.; Kondo, J.; Kawakami, T.; Shukunami, C.; Aimoto, S.; Tanaka, H.; Hiraki, Y. Synthetic disulfide-bridged cyclic peptides mimic the anti-angiogenic actions of chondromodulin-I. *Cancer Sci.* **2012**, *103*, 1311–1318. [CrossRef] [PubMed]
38. Ma, J.; Gao, S.; Xie, X.; Sun, E.; Zhang, M.; Zhou, Q.; Lu, C. SPARC inhibits breast cancer bone metastasis and may be a clinical therapeutic target. *Oncol. Lett.* **2017**, *14*, 5876–5882. [CrossRef]
39. Huang, Y.; Zhang, J.; Zhao, Y.-Y.; Jiang, W.; Xue, C.; Xu, F.; Zhao, H.-Y.; Zhang, Y.; Zhao, L.-P.; Hu, Z.-H.; et al. SPARC expression and prognostic value in non-small cell lung cancer. *Chin. J. Cancer* **2012**, *31*, 541–548. [CrossRef]
40. Zhu, A.; Yuan, P.; Du, F.; Hong, R.; Ding, X.; Shi, X.; Fan, Y.; Wang, J.; Luo, Y.; Ma, F.; et al. SPARC overexpression in primary tumors correlates with disease recurrence and overall survival in patients with triple negative breast cancer. *Oncotarget* **2016**, *7*, 76628–76634. [CrossRef]
41. Yang, X.; Cai, W.; Xu, Z.; Chen, J.; Li, C.; Liu, S.; Yang, Z.; Pan, Q.; Li, M.; Ma, J.; et al. High efficacy and minimal peptide required for the anti-angiogenic and anti-hepatocarcinoma activities of plasminogen K5. *J. Cell. Mol. Med.* **2010**, *14*, 2519–2530. [CrossRef] [PubMed]
42. Scappaticci, F.A.; Smith, R.; Pathak, A.; Schloss, D.; Lum, B.; Cao, Y.; Johnson, F.; Engleman, E.G.; Nolan, G.P. Combination angiostatin and endostatin gene transfer induces synergistic antiangiogenic activity in vitro and antitumor efficacy in leukemia and solid tumors in mice. *Mol. Ther. J. Am. Soc. Gene Ther.* **2001**, *3*, 186–196. [CrossRef] [PubMed]
43. Nor, J.E.; Mitra, R.S.; Sutorik, M.M.; Mooney, D.J.; Castle, V.P.; Polverini, P.J. Thrombospondin-1 induces endothelial cell apoptosis and inhibits angiogenesis by activating the caspase death pathway. *J. Vasc. Res.* **2000**, *37*, 209–218. [CrossRef] [PubMed]
44. Florio, T.; Morini, M.; Villa, V.; Arena, S.; Corsaro, A.; Thellung, S.; Culler, M.D.; Pfeffer, U.; Noonan, D.M.; Schettini, G.; et al. Somatostatin inhibits tumor angiogenesis and growth via somatostatin receptor-3-mediated regulation of endothelial nitric oxide synthase and mitogen-activated protein kinase activities. *Endocrinology* **2003**, *144*, 1574–1584. [CrossRef] [PubMed]
45. Eikesdal, H.P.; Sugimoto, H.; Birrane, G.; Maeshima, Y.; Cooke, V.G.; Kieran, M.; Kalluri, R. Identification of amino acids essential for the antiangiogenic activity of tumstatin and its use in combination antitumor activity. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 15040–15045. [CrossRef] [PubMed]
46. Bae, D.G.; Gho, Y.S.; Yoon, W.H.; Chae, C.B. Arginine-rich anti-vascular endothelial growth factor peptides inhibit tumor growth and metastasis by blocking angiogenesis. *J. Biol. Chem.* **2000**, *275*, 13588–13596. [CrossRef]
47. Xiong, Y.; Fru, M.F.; Yu, Y.; Montani, J.P.; Ming, X.F.; Yang, Z. Long term exposure to L-arginine accelerates endothelial cell senescence through arginase-II and S6K1 signaling. *Aging* **2014**, *6*, 369–379. [CrossRef]
48. Ruoslahti, E. RGD and other recognition sequences for integrins. *Annu. Rev. Cell Dev. Biol.* **1996**, *12*, 697–715. [CrossRef]
49. Xu, H.; Pan, L.; Ren, Y.; Yang, Y.; Huang, X.; Liu, Z. RGD-modified angiogenesis inhibitor HM-3 dose: Dual function during cancer treatment. *Bioconjugate Chem.* **2011**, *22*, 1386–1393. [CrossRef]
50. Buerkle, M.A.; Pahernik, S.A.; Sutter, A.; Jonczyk, A.; Messmer, K.; Dellian, M. Inhibition of the alpha-nu integrins with a cyclic RGD peptide impairs angiogenesis, growth and metastasis of solid tumours in vivo. *Br. J. Cancer* **2002**, *86*, 788–795. [CrossRef]
51. Kondo, M.; Asai, T.; Katanasaka, Y.; Sadzuka, Y.; Tsukada, H.; Ogino, K.; Taki, T.; Baba, K.; Oku, N. Anti-neovascular therapy by liposomal drug targeted to membrane type-1 matrix metalloproteinase. *Int. J. Cancer* **2004**, *108*, 301–306. [CrossRef] [PubMed]
52. Li, Y.; Wang, J.; Gao, Y.; Zhu, J.; Wientjes, M.G.; Au, J.L. Relationships between liposome properties, cell membrane binding, intracellular processing, and intracellular bioavailability. *AAPS J.* **2011**, *13*, 585–597. [CrossRef] [PubMed]
53. Charoenkwan, P.; Shoombuatong, W.; Lee, H.-C.; Chaijaruwanich, J.; Huang, H.-L.; Ho, S.-Y. SCMCRYS: Predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. *PLoS ONE* **2013**, *8*, e72368. [CrossRef] [PubMed]
54. Huang, H.-L. Propensity scores for prediction and characterization of bioluminescent proteins from sequences. *PLoS ONE* **2014**, *9*, e97158. [CrossRef] [PubMed]

55. Pratiwi, R.; Malik, A.A.; Schaduangrat, N.; Prachayasittikul, V.; Wikberg, J.E.; Nantasenamat, C.; Shoombuatong, W. CryoProtect: A web server for classifying antifreeze proteins from nonantifreeze proteins. *J. Chem.* **2017**, *2017*. [CrossRef]
56. Win, T.S.; Malik, A.A.; Prachayasittikul, V.; Wikberg, J.E.S.; Nantasenamat, C.; Shoombuatong, W. HemoPred: A web server for predicting the hemolytic activity of peptides. *Future Med. Chem.* **2017**, *9*, 275–291. [CrossRef]
57. Win, T.S.; Schaduangrat, N.; Prachayasittikul, V.; Nantasenamat, C.; Shoombuatong, W. PAAP: A web server for predicting antihypertensive activity of peptides. *Future Med. Chem.* **2018**, *10*, 1749–1767. [CrossRef] [PubMed]
58. Kawashima, S.; Kanehisa, M. AAindex: Amino acid index database. *Nucleic Acids Res.* **2000**, *28*, 374. [CrossRef] [PubMed]
59. Tsai, C.S. *Biomacromolecules: Introduction to Structure, Function and Informatics*; John Wiley & Sons: Hoboken, NJ, USA, 2007.
60. Argos, P.; Hanei, M.; Garavito, R.M. The Chou-Fasman secondary structure prediction method with an extended data base. *FEBS Lett.* **1978**, *93*, 19–24. [CrossRef]
61. Nowick, J.S. Exploring beta-sheet structure and interactions with chemical model systems. *Acc. Chem. Res.* **2008**, *41*, 1319–1330. [CrossRef]
62. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinform.* **2008**, *9*, 40. [CrossRef] [PubMed]
63. Hohenester, E.; Sasaki, T.; Olsen, B.R.; Timpl, R. Crystal structure of the angiogenesis inhibitor endostatin at 1.5 A resolution. *EMBO J.* **1998**, *17*, 1656–1664. [CrossRef] [PubMed]
64. Carlson, C.B.; Lawler, J.; Mosher, D.F. Structures of thrombospondins. *Cell. Mol. Life Sci. CMLS* **2008**, *65*, 672–686. [CrossRef] [PubMed]
65. Taraboletti, G.; Roberts, D.D.; Liotta, L.A. Thrombospondin-induced tumor cell migration: Haptotaxis and chemotaxis are mediated by different molecular domains. *J. Cell Biol.* **1987**, *105*, 2409–2415. [CrossRef]
66. Ginj, M.; Schmitt, J.S.; Chen, J.; Waser, B.; Reubi, J.C.; de Jong, M.; Schulz, S.; Maecke, H.R. Design, synthesis, and biological evaluation of somatostatin-based radiopeptides. *Chem. Biol.* **2006**, *13*, 1081–1090. [CrossRef] [PubMed]
67. Oshima, Y.; Sato, K.; Tashiro, F.; Miyazaki, J.; Nishida, K.; Hiraki, Y.; Tano, Y.; Shukunami, C. Anti-angiogenic action of the C-terminal domain of tenomodulin that shares homology with chondromodulin-I. *J. Cell Sci.* **2004**, *117 Pt 13*, 2731–2744. [CrossRef]
68. Marcelino, A.M.C.; Gierasch, L.M. Roles of beta-turns in protein folding: From peptide models to protein engineering. *Biopolymers* **2008**, *89*, 380–391. [CrossRef] [PubMed]
69. Karagiannis, E.D.; Popel, A.S. A systematic methodology for proteome-wide identification of peptides inhibiting the proliferation and migration of endothelial cells. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 13775–13780. [CrossRef] [PubMed]
70. Hsu, H.-J.; Chang, H.-J.; Peng, H.-P.; Huang, S.-S.; Lin, M.-Y.; Yang, A.-S. Assessing computational amino acid β-turn propensities with a phage-displayed combinatorial library and directed evolution. *Structure* **2006**, *14*, 1499–1510. [CrossRef]
71. Millward, S.W.; Fiacco, S.; Austin, R.J.; Roberts, R.W. Design of cyclic peptides that bind protein surfaces with antibody-like affinity. *ACS Chem. Biol.* **2007**, *2*, 625–634. [CrossRef]
72. Tien, P.G.; Kayama, F.; Konishi, F.; Tamemoto, H.; Kasono, K.; Hung, N.T.; Kuroki, M.; Ishikawa, S.E.; Van, C.N.; Kawakami, M. Inhibition of tumor growth and angiogenesis by water extract of Gac fruit (*Momordica cochinchinensis* Spreng). *Int. J. Oncol.* **2005**, *26*, 881–889. [CrossRef] [PubMed]
73. Hernandez, J.F.; Gagnon, J.; Chiche, L.; Nguyen, T.M.; Andrieu, J.P.; Heitz, A.; Trinh Hong, T.; Pham, T.T.; Le Nguyen, D. Squash trypsin inhibitors from *Momordica cochinchinensis* exhibit an atypical macrocyclic structure. *Biochemistry* **2000**, *39*, 5722–5730. [CrossRef] [PubMed]
74. Torras, A.S.; Carvalho, A.; Abasolo, I.; Zapata, M.; Distefano, L.; Schwartz, S.; Garcia-Arumi, J. In vitro studies on the antiangiogenic effects of Pigment Epithelium Derived Factor and Somatostatin. *Investig. Ophthalmol. Vis. Sci.* **2013**, *54*, 4660.
75. Chan, L.Y.; Craik, D.J.; Daly, N.L. Cyclic thrombospondin-1 mimetics: Grafting of a thrombospondin sequence into circular disulfide-rich frameworks to inhibit endothelial cell migration. *Biosci. Rep.* **2015**, *35*, e00270. [CrossRef] [PubMed]

76. Maeshima, Y.; Manfredi, M.; Reimer, C.; Holthaus, K.A.; Hopfer, H.; Chandamuri, B.R.; Kharbanda, S.; Kalluri, R. Identification of the anti-angiogenic site within vascular basement membrane-derived tumstatin. *J. Biol. Chem.* **2001**, *276*, 15240–15248. [CrossRef] [PubMed]
77. Northfield, S.E.; Wang, C.K.; Schroeder, C.I.; Durek, T.; Kan, M.-W.; Swedberg, J.E.; Craik, D.J. Disulfide-rich macrocyclic peptides as templates in drug design. *Eur. J. Med. Chem.* **2014**, *77*, 248–257. [CrossRef] [PubMed]
78. Cemazar, M.; Kwon, S.; Mahatmanto, T.; Ravipati, A.S.; Craik, D.J. Discovery and applications of disulfide-rich cyclic peptides. *Curr. Top. Med. Chem.* **2012**, *12*, 1534–1545. [CrossRef]
79. Huang, Y.; Niu, B.; Gao, Y.; Fu, L.; Li, W. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics* **2010**, *26*, 680–682. [CrossRef]
80. Shoombuatong, W.; Mekha, P.; Chaijaruwanich, J. Sequence based human leukocyte antigen gene prediction using informative physicochemical properties. *Int. J. Data Min. Bioinform.* **2015**, *13*, 211–224. [CrossRef]
81. Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **2011**, *273*, 236–247. [CrossRef]
82. Chou, K.-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2004**, *21*, 10–19. [CrossRef] [PubMed]
83. Xiao, N.; Cao, D.-S.; Zhu, M.-F.; Xu, Q.-S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* **2015**, *31*, 1857–1859. [CrossRef] [PubMed]
84. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
85. Breiman, L. *Classification and Regression Trees*; Routledge: Abingdon, UK, 2017.
86. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
87. Simeon, S.; Li, H.; Win, T.S.; Malik, A.A.; Kandhro, A.H.; Piacham, T.; Shoombuatong, W.; Nuchnoi, P.; Wikberg, J.E.; Gleeson, M.P. PepBio: Predicting the bioactivity of host defense peptides. *RSC Adv.* **2017**, *7*, 35119–35134. [CrossRef]
88. Simeon, S.; Shoombuatong, W.; Anuwongcharoen, N.; Preeyanon, L.; Prachayasittikul, V.; Wikberg, J.E.; Nantasenamat, C. osFP: A web server for predicting the oligomeric states of fluorescent proteins. *J. Cheminformatics* **2016**, *8*, 72. [CrossRef] [PubMed]
89. Phanus-umporn, C.; Shoombuatong, W.; Prachayasittikul, V.; Anuwongcharoen, N.; Nantasenamat, C. Correction: Privileged substructures for anti-sickling activity via cheminformatic analysis. *RSC Adv.* **2018**, *8*, 8233. [CrossRef]
90. Prachayasittikul, V.; Worachartcheewan, A.; Shoombuatong, W.; Prachayasittikul, V.; Nantasenamat, C. Classification of P-glycoprotein-interacting compounds using machine learning methods. *Excli J.* **2015**, *14*, 958.
91. Simeon, S.; Anuwongcharoen, N.; Shoombuatong, W.; Malik, A.A.; Prachayasittikul, V.; Wikberg, J.E.; Nantasenamat, C. Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking. *PeerJ* **2016**, *4*, e2322. [CrossRef]
92. Suvannang, N.; Preeyanon, L.; Malik, A.A.; Schaduangrat, N.; Shoombuatong, W.; Worachartcheewan, A.; Tantimongcolwat, T.; Nantasenamat, C. Probing the origin of estrogen receptor alpha inhibition via large-scale QSAR study. *RSC Adv.* **2018**, *8*, 11344–11356. [CrossRef]
93. Worachartcheewan, A.; Prachayasittikul, V.; Anuwongcharoen, N.; Shoombuatong, W.; Prachayasittikul, V.; Nantasenamat, C. On the origins of hepatitis C virus NS5B polymerase inhibitory activity using machine learning approaches. *Curr. Top. Med. Chem.* **2015**, *15*, 1814–1826. [CrossRef] [PubMed]
94. Calle, M.L.; Urrea, V. Letter to the editor: Stability of random forest importance measures. *Brief. Bioinform.* **2010**, *12*, 86–89. [CrossRef]
95. Shoombuatong, W.; Hongjaisee, S.; Barin, F.; Chaijaruwanich, J.; Samleerat, T. HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees. *Comput. Biol. Med.* **2012**, *42*, 885–889. [CrossRef] [PubMed]
96. Shoombuatong, W.; Huang, H.-L.; Chaijaruwanich, J.; Charoenkwan, P.; Lee, H.-C.; Ho, S.-Y. Predicting protein crystallization using a simple scoring card method. In Proceedings of the 2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Singapore, 16–19 April 2013; pp. 23–30.

97. Shoombuatong, W.; Mekha, P.; Waiyamai, K.; Cheevadhanarak, S.; Chaijaruwanicha, J. Prediction of human leukocyte antigen gene using k-nearest neighbour classifier based on spectrum kernel. *ScienceAsia* **2013**, *39*, 42–49. [CrossRef]
98. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.-C.; Müller, M. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **2011**, *12*, 77. [CrossRef] [PubMed]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

BMC Medical Informatics and Decision Making

TECHNICAL ADVANCE Open Access

ThalPred: a web-based prediction tool for discriminating thalassemia trait and iron deficiency anemia

V. Laengsri[1,2], W. Shoombuatong[3], W. Adirojananon[2], C. Nantasenamat[3], V. Prachayasittikul[4] and P. Nuchnoi[1,2*]

Abstract

Background: The hypochromic microcytic anemia (HMA) commonly found in Thailand are iron deficiency anemia (IDA) and thalassemia trait (TT). Accurate discrimination between IDA and TT is an important issue and better methods are urgently needed. Although considerable RBC formulas and indices with various optimal cut-off values have been developed, distinguishing between IDA and TT is still a challenging problem due to the diversity of various anemic populations. To address this problem, it is desirable to develop an improved and automated prediction model for discriminating IDA from TT.

Methods: We retrospectively collected laboratory data of HMA found in Thai adults. Five machine learnings, including *k*-nearest neighbor (*k*-NN), decision tree, random forest (RF), artificial neural network (ANN) and support vector machine (SVM), were applied to construct a discriminant model. Performance was assessed and compared with thirteen existing discriminant formulas and indices.

Results: The data of 186 patients (146 patients with TT and 40 with IDA) were enrolled. The interpretable rules derived from the RF model were proposed to demonstrate the combination of RBC indices for discriminating IDA from TT. A web-based tool 'ThalPred' was implemented using an SVM model based on seven RBC parameters. ThalPred achieved prediction results with an external accuracy, MCC and AUC of 95.59, 0.87 and 0.98, respectively.

Conclusion: ThalPred and an interpretable rule were provided for distinguishing IDA from TT. For the convenience of health care team experimental scientists, a web-based tool has been established at http://codes.bio/thalpred/ by which users can easily get their desired screening test result without the need to go through the underlying mathematical and computational details.

Keywords: Thalassemia trait, Iron deficiency anemia, Machine learning, Support vector machine, Random forest, Discrimination, Decision making

Background

Anemia is the condition of decreased number of red blood cells (RBCs) or of the concentration of hemoglobin (Hb). Anemia is a health problem affecting both developing and developed countries. The global prevalence of anemia in 2010 was 32.9% and it was especially common in Central Africa, the Middle East, the Mediterranean and Southeast Asia [1]. It can occur from one of three causes, acute blood loss, increased hemolysis or ineffective hematopoiesis.

In Thailand, both iron deficiency anemia (IDA) and thalassemia trait (TT) are highly prevalent. Iron is an important element in our body, being a component of many enzymes and playing a role in hemoglobin synthesis. Therefore, a lack of iron can lead to IDA. The prevalence of IDA among Thai people is estimated to be 1.5–8% [2]. Thalassemia is an inherited hematological disorder that is caused by abnormal production of alpha (α)- or beta (β)-globin chains. The prevalence of TT is approximately 20–25% in the Thai population [3]. The levels of serum ferritin, serum iron, total iron binding capacity and

* Correspondence: pornlada.nuc@mahidol.ac.th
[1]Center for Research and Innovation, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand
[2]Department of Clinical Microscopy, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand
Full list of author information is available at the end of the article

© The Author(s). 2019, corrected publication 2019. **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

percentage of transferrin saturation are the most commonly used assays to confirm IDA [4]. Meanwhile, Hb testing via high performance liquid chromatography or capillary electrophoresis to detect abnormal Hb levels or DNA analysis are the assays commonly used for TT diagnosis [5–7]. In practice, these latter techniques are not available in routine laboratories as they require special machines, and are time- and cost-consuming. Basically, when patients were assumed to be TT or IDA, clinician often prescribed the cassette of hematological tests covering both TT and IDA diagnosis. These laboratory workups consume personnel work load and governmental budget. This leads to financial crisis for national health care of low-middle income countries. In order to guide clinician for rational lab use, we therefore developed web-based tool for assisting clinician to prescribe rational and cost-effective laboratory testing for TT and IDA diagnosis. Multiple formulas and indices have been proposed for such discrimination including the Bessman index (BI) [8], Ehsani formula (EF) [9], England & Fraser index (E&F) [10], Green & King index (G&K) [11], Mentzler index (MI) [12], Red Cell Distribution Width index (RDWI) [13], Ricerca index (RI) [14], Shine and Lal index (S&L) [15], Siridah index (SI) [16], Srivastava formula (SF) [17], Sirachainan formula (SiF) [18], Kandhro 1 formula (KF1) [19] and Kandhro 2 formula (KF2) [19].

The aforementioned discriminant indices and formulas yielded quite encouraging prediction results. However, the prediction results across different populations, especially in sensitivity and specificity, are still unsatisfactory. Some of studies are contentious due to difference of gender, age or ethnicity [20–22]. In order to improve prediction results, the computation of optimal cut-off values for specific populations in different countries is needed [23, 24]. The potential of machine learning techniques has been demonstrated for near-term translational impact. For instance, in the case of the biomedicine, major applications of machine learning are medical/radiological diagnosis and drug discovery. Thus, it may be possible to develop discriminant models based on machine learning techniques for providing effective large-scale analyses of laboratory data. To the best of our knowledge, only one discriminant model has been proposed to differentiate IDA and TT in the Thai population [18]. In this study, we exerted an effort to develop a powerful model to discriminate IDA from TT using a support vector machine (SVM) which we named *ThalPred*.

The previous works had demonstrated the meaningful of data mining and the increasing of computational power in various aspect of biomedical application [25–28]. The present study aims to establish a reliable and interpretable computational model. Therefore, the important procedures as the following are considered: (i) collect clear and reliable laboratory dataset for training and validating; (ii) demonstrate characteristic or descriptor of dataset for intrinsic properties prediction; (iii) identify feature of importance for improving interpretability; (iv) develop a simple and interpretable model; (v) perform rigorous cross validation for analyzing internal and external predictive power; (vi) develop affordable and user friendly based web tool for implementation in the healthcare community.

Material and methods

Data collection

This was a retrospective study of encoded and unlinked clinical laboratory data obtained from the Center of Medical Laboratory Services, the Faculty of Medical Technology, Mahidol University from the period July 2014 to September 2016. The abbreviations of hematological laboratory testing were listed in Table 1. We collected 237 sets of hematological data from Thai subjects, age 18 to 60 years, that showed hypochromic microcytic anemia (HMA) and were diagnosed as reflecting IDA or TT by two independent medical technologists. Other causes of HMA, such as hypothyroidism, anemia of chronic disorders, hepatitis B / C / or D infection, or *Helicobacter pylori* infection, were excluded from the study in order to specify IDA and TT only. Hemoglobin electrophoresis and serum ferritin results were analyzed to discriminate TT and IDA, respectively. The level of ferritin was determined by electrochemiluminescence method (Cobas® 2014© Roche, Switzerland). Variant hemoglobin testing was performed using low pressure liquid chromatography (Bio-Rad Laboratories, USA). The final dataset consisted of 186 subjects (146 TT and 40 IDA cases) which were used for internal and external analyses. The study was conducted under the approval of the Mahidol University Central Institutional Review Board (MU-CIRB; CODE No. MU-CIRB 2016/084.0311). We received a participant consent waiver from MU-CIRB. All information of subjects was de-identified prior data analysis.

Table 1 List of laboratory testing abbreviation used in this study

Full Name	Abbreviation
Hemoglobin	Hb
Hematocrit	Hct
Hypochromic microcytic anemia	HMA
Iron deficiency anemia	IDA
Mean corpuscular volume	MCV
Mean corpuscular hemoglobin	MCH
Mean corpuscular hemoglobin concentration	MCHC
Red blood cell	RBC
Red blood cell distribution width	RDW
Thalassemia trait	TT

Statistical analyses

The statistical analysis of this study was computed using SPSS software, version 20 (SPSS Inc., New York, Armonk, USA). As an exploratory statistical analysis, univariate statistical analysis using mean and standard deviation (SD) was performed to investigate the different patterns and trends of individual hematological parameters between the data of groups TT and IDA. The Kolmogorov-Smirnov test was used for normality testing. Since the data of this study was not normally distributed. The comparison between groups TT and IDA was performed by using the Mann-Whitney U test. A p-value < 0.05 was considered significant. Furthermore, in order to perform multivariate statistical analyses amongst RBC indices, principal component analysis (PCA) was performed. PCA has probably been the most popular technique to perform multivariate statistical analysis for the tasks of data exploration and pattern recognition. The advantages of PCA are to: (i) extract the most important information and represent it with only a few dimensions, called principal components (PCs); (ii) compress the dimensions of the dataset by keeping only the important information; and (iii) analyze the characteristics and structure of the data set. In this study, PCA was performed by using the *FactoMineR* package [29] in R program to represent the distributions of IDA and TT cases, and identify the RBC indices important for distinguishing IDA and TT cases.

Models construction

The following five popular computational models, namely k-nearest neighbor (k-NN), decision tree (DT), random forest (RF), artificial neural network (ANN) and support vector machine learning (SVM), were applied to develop discriminant models for effectively distinguishing TT from IDA. The fundamental and associated parameter optimization for the five classifiers are briefly described (Fig. 1).

k-NN is one of the most popular and lazy machine learning algorithms for a variety of problem domains [25, 30, 31]. This algorithm is conceptually based on a distance function, such as the Euclidean distance, to measure the similarity between a pair of training and unknown datasets. For obtaining the best k-NN model, the suitable number of neighbors ($k \in \{5, 7, 9,..., 43\}$ was determined by using the *caret* package [28, 32] over 5-fold cross-validation (5-fold CV) scheme.

DT model can be used in the classification and regression tasks by constructing a model in the form of a tree structure. Herein, we constructed the DT model by implementing the J48 algorithm from the RWeka package in R program (version 3.3.2) [33] using default parameters. Briefly, the J48 algorithm is a re-implementation of the C4.5 algorithm [34] based on Javascript. The feature with the highest information gain is select to build a model. Finally, because of its built-in feature selector, the DT

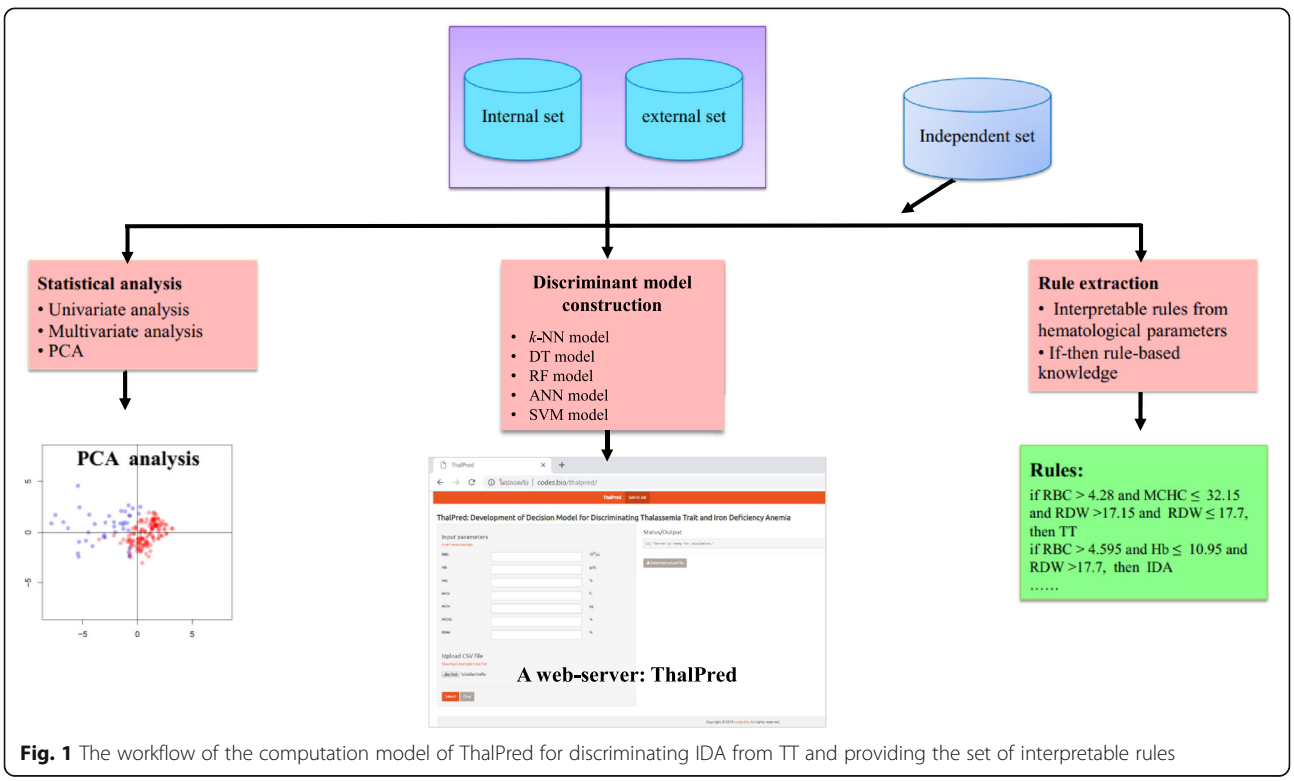

Fig. 1 The workflow of the computation model of ThalPred for discriminating IDA from TT and providing the set of interpretable rules

model will provide the feature usage score for ranking the feature importance.

RF is an ensemble classification and regression tree (CART) classifier [34, 35]. The RF model is a classifier derived from gathering many weak CART trees for improving the prediction performance. To construct the model, RF takes advantage of two well-known machine learning algorithms, i.e. bagging and random feature selection. To estimate informative features, RF model utilizes two measures, i.e. mean decrease in accuracy and Gini index [35]. In this study, the RF classifier was implemented using the *randomForest R* package [34]. To obtain an optimal RF model, two parameters, namely, *ntree* (i.e., the number of trees used for constructing the RF classifier) and *mtry* (i.e., the number of random candidate features), were tuned by using a grid search procedure based on 5-fold cross-validation (5-fold CV), where *ntree* ∈ {100, 200, 300, 400, 500} was determined, while *mtry* was estimated using the tuneRF function in the *randomForest* R package [34].

ANN is computing systems originally inspired by the way biological nervous systems process information [36]. Previously, many researchers reported that ANN accomplished well in many domains, such as protein sequence analysis, image recognition, speech recognition, and natural language processing [27, 37, 38]. In practice, there are two important types of ANN, i.e. the perceptron and the sigmoid neuron, while stochastic gradient descent is known as the standard estimating parameter algorithm for ANN. For achieving the best ANN model, two parameters, namely, *size* (i.e., the number of hidden nodes) and *decay* (i.e., the number of weight decay), were subjected to optimization. Particularly, *size* ∈ {1, 2, 3,..., 10} and *decay* ∈ {0.1, 0.2, 0.3, 0.4, 0.5} were determined by using the *caret* package [28, 32] over a 5-fold CV scheme.

SVM is a statistical learning approach based on the principle of structure-risk minimization and a kernel method (as proposed by Vapnik [39]) which are used to construct a maximum-margin-separating hyperplane for distinguishing the two classes of interest. The radial basis function kernel was used to transform the original feature space into a higher dimensional space in which the SVM classifier can linearly separate the inherent classes of the dependent variable via a maximum separating hyperplane [40]. Optimization of the SVM parameters consisting of the cost ∈ {2^{8}, 2^{7}, 2^{6},..., 2^{7}, 2^{8}} and γ ∈ {2^{8}, 2^{7}, 2^{6},..., 2^{7}, 2^{8}} were determined via a grid search spanning the search space evaluated by 5-fold CV scheme using the *e1071*R package [41].

Cross-validation for identification of discriminant capability

The validation of an empirical predictive model is essential. In order to train and evaluate the discriminant models, the data set of this study was randomly partitioned into internal and external sets with 80 and 20%, respectively, of the data set from both TT and IDA. The internal set was evaluated using the 5-fold CV scheme [42] to confirm the reliability and robustness of the proposed discriminative model. The external set was used to assess the generalizability of the model when extrapolating to unknown samples. To avoid the possibility of bias arising from a single data split upon model training, data splitting was performed for 100 independent iterations. The final prediction performances of the 5-fold CV and external validation tests of the proposed discriminative model were reported by using the mean and standard deviation values of analyzed parameters (Fig. 1).

Rule extraction

In addition to model accuracy and ability to discriminate IDA from TT, there is a possibility that a simple and meaningful rule may be extracted from an RF model. From the *nroot* to a leaf node is a rule for a tree. The main purpose of extraction rules is to ease utilization and make the model more interpretable compared to black-box approaches [25] such as SVM. In this study, the extraction rules were obtained by using the R package *inTrees* (interpretable trees) [34] and only 100 decision trees were established for each encoding to train the RF model. *InTrees* is a powerful package for extracting, measuring, pruning, selecting and summarizing from an RF model (Fig. 1).

Performance assessment

Performance assessment is important step for developing reliable and useful predictor. In order to discriminate TT from IDA, pattern recognition is applied for classification concept. Five standard statistical parameters, namely, accuracy (Ac), sensitivity (Sn), specificity (Sp), Matthew's correlation coefficient (MCC), and Youden's index (YI), were addressed to evaluate the predictive performance of the proposed methods. These five parameters were computed as follows [43–46]:

$$\mathbf{Ac} = \frac{TP + TN}{(TP + TN + FP + FN)} \times 100$$

$$\mathbf{Sn} = \frac{TP}{(TP + FN)} \times 100$$

$$\mathbf{Sp} = \frac{TN}{(TN + FP)} \times 100$$

$$\mathbf{MCC} = \frac{(TP)(TN) - (FP)(Fn)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$\mathbf{YI} = [(\mathrm{Sn} + \mathrm{Sp}) - 100]/100$$

where TP, FP, TN and FN are true positive, false positive, true negative and false negative, respectively. Furthermore,

the receiver operating characteristic (ROC) curve, which plots the true positive rate (Sn) against the false positive rate (1-Sp) at different thresholds, was also used for model evaluation. A predictor with perfect classification has a ROC curve passing through the top-left corner, i.e. Sn = 100% and Sp = 100%. To plot the ROC curve and calculate the value of area under the curve (AUC), the ROCR package [47] was used.

Results

Differences in hematological data of TT and IDA using univariate and multivariate analyses

HMA data was classified into two groups: 146 TT cases and 40 IDA cases. The characteristics and hematological data of both groups of IDA and TT are summarized in Table 2. Age of two groups was not significantly different (*p*-value 0.435). All RBC parameters were significantly different between the two groups ($p < 0.05$). All averages of parameters in TT group were higher than IDA group except RDW. Furthermore, PCA was applied as multivariate analysis was presented as scores (Fig. 2a) and loadings plots (Fig. 2b). The percentage of variance explained by the first two PCs (83. 33%) were considered sufficient in describing the behavior of the data. The graphs of Fig. 2a and Fig. 2b also revealed that RDW contributes greatly to IDA, while the remaining RBC parameters contribute to TT. These results were well reflected by the mean values of IDA (15.88 ± 1.13) and TT (20. 48 ± 3.23). Interestingly, the outputs of two analyses were consistent. Therefore, these suggested that all RBC indices may be important to discriminate between patients with IDA and those with TT.

Evaluation of existing discriminant formulas and indices

To objectively evaluate the performance of our proposed discriminant model and fairly compare it with existing discriminant formulas and indices including BI [8], EF [9], E&F [10], G&K [11], MI [12], RDWI [13], RI [14], S&L [15], SI [16], SF [17], SiF [18], KF1 [19] and KF2 [19], the same dataset was used for evaluation of each. Their cut-off, Ac, Sn, Sp, MCC, YI, and AUC were shown in Table 3.

As noticed in Fig. 3, most existing formulas and indices achieved an AUC value that was greater than 0.5 [except for S&L (0. 31) and SiF (0.02)]. The highest MCC and AUC of 0.74 and 0.98, respectively, were achieved by the KF2 formula. Meanwhile, RI and G&K formulas performed well with the second and third highest MCC/AUC of 0.70/0.98 and 0.66/0.93, respectively. Meanwhile, the formulas and indices with the lowest values of MCC/AUC were SiF, S&L and EF (-05.6/0.02, 0.01/0. 31 and 0.17/0.70, respectively). These prediction performances indicated that only a few formulas and indices, e.g. G&K, RI and KF2, performed well on our dataset.

Evaluation of the proposed discriminant model

To develop a machine learning-based discriminating model, prediction performance depends on the parameter(s) used. In this study, five popular classifiers were considered, e.g. *k*-NN, DT, RF, ANN and SVM. To make fair comparison with existing formulas and indices, the same dataset was used. Furthermore, a 5-fold CV and external validation tests were carried out. As described above, the internal and external datasets were constructed with a random sampling procedure. To objectively evaluate the impact of the random sampling procedure, we repeated this construction 100 times. Table 4 and Fig. 4 list performance comparisons of various models using different methods over the 5-fold CV and external validation schemes.

As seen in Table 4, the DT model yielded the highest prediction results over the 5-fold CV with a mean Ac, Sp, Sn, MCC, YI, and AUC of 98.03, 96.15, 98.54%, 0.94, 0.93, and 1.00, respectively, while the SVM and RF models performed well with the second and third best prediction results. On the other hand, the SVM model

Table 2 The age and red blood cell parameters of study subjects with thalassemia trait or iron deficiency anemia

Parameters	TT (*N* = 146)	IDA (*N* = 40)	*p*-value
Age (yrs.)	37.79 ± 7.86 (18.00–50.00)	39.15 ± 9.61 (23.00–58.00)	0.435
RBC (10^6/μL)	5.32 ± 0.48 (4.35–6.83)	4.03 ± 0.96 (1.69–5.77)	< 0.001*
Hb (g/dL)	11.99 ± 1.11 (8.10–14.60)	7.97 ± 2.26 (2.50–10.90)	< 0.001*
Hct (%)	36.46 ± 4.10 (34.10–42.40)	26.19 ± 6.63 (9.90–34.30)	< 0.001*
MCV (fL)	69.49 ± 6.14 (52.30–79.70)	65.12 ± 9.32 (48.70–81.00)	0.010*
MCH (pg)	22.71 ± 2.25 (17.70–26.60)	19.65 ± 3.33 (12.50–25.70)	< 0.001*
MCHC (%)	32.67 ± 1.14 (30.10–35.70)	30.13 ± 1.93 (24.90–35.70)	< 0.001*
RDW (%)	15.88 ± 1.13 (13.50–22.00)	20.48 ± 3.23 (14.90–26.70)	< 0.001*

The data are shown as mean ± standard deviation

Hb Hemoglobin; *Hct* Hematocrit; *IDA* Iron deficiency anemia; *MCH* mean corpuscular hemoglobin; *MCHC* mean corpuscular hemoglobin concentration; *MCV* mean corpuscular volume; *RBC* red blood cell count; *RDW* red blood cell distribution width; *TT* Thalassemia trait

* Mann-Whitney U test *p*-value < 0.05

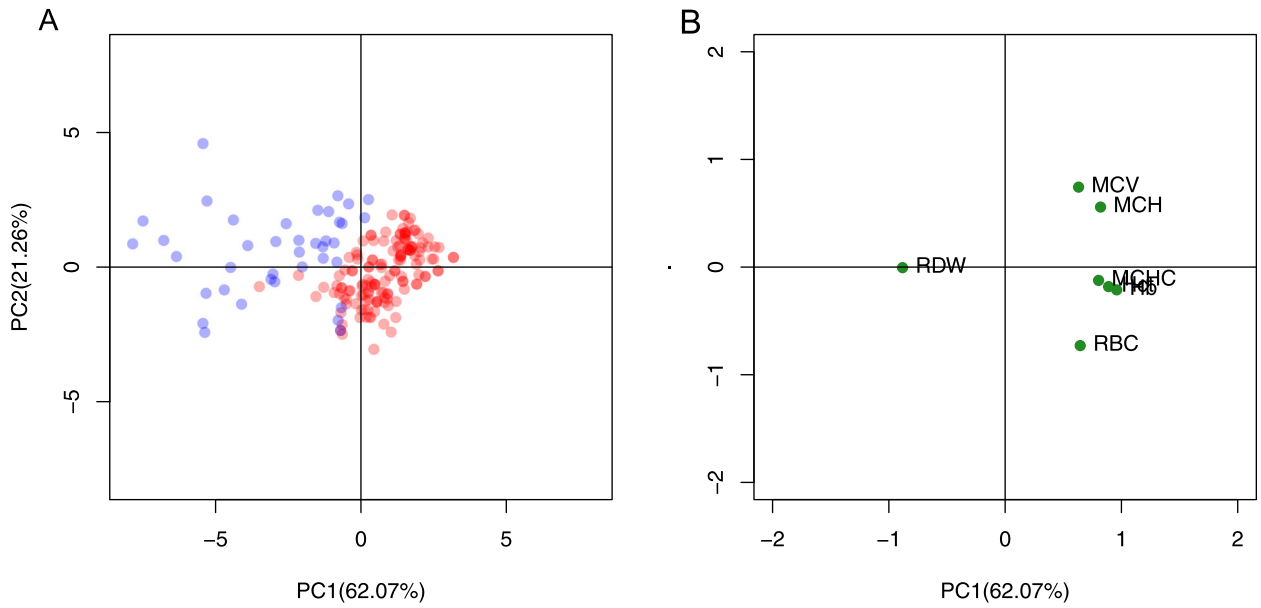

Fig. 2 Multivariate analysis using principal component analysis (PCA) of our laboratory data consisting of 146 TT cases (red circles) and 40 IDA cases (blue circles) derived from PCA scores (**a**) and loadings (**b**) plots

was more effective than the DT model over the external validation test with a mean Ac, Sp, Sn, MCC, YI, and AUC of 95.59, 92. 49%, 96.74%, 0.87, 0.87, and 0.98, respectively. Interestingly, the SVM model still achieved levels of discrimination between IDA and TT on the independent dataset with means of Ac, Sp, Sn, MCC, YI, and AUC being 96.08, 92.59, 100.00%, 0.92, 0.92, and 1.00, respectively. As mentioned in the section *Data collection*, the dataset used in this study is imbalance among IDA and TT samples. Thus, we utilized a resampling approach to randomly generate a balanced internal set consisting of 32 IDA and 32 TT samples, the remaining samples consisting of 8 IDA and 114 TT sample were used as an external set. To objectively evaluate the impact of random sampling, we repeated this process with 100 independent iterations to generate the balanced internal and new external sets for constructing SVM models. The average prediction results (Ac and MCC)

Table 3 Performance comparisons of existing discriminant formulas and indices proposed for differentiation of iron deficiency anemia from thalassemia trait

Indices/ formulas	Cut-off	Ac (%)	Sn (%)	Sp (%)	MCC	YI	AUC
BI = RDW	15	36.02	19.18	97.50	0.19	0.17	0.71
EF = MCV − 10 × RBC	15	52.15	45.89	75.00	0.17	0.21	0.70
E&F = MCV - RBC − 5Hb - 6.4	0	67.42	60.95	92.50	0.44	0.54	0.91
G&K = MCV^2×RDW/100Hb	72	84.95	82.88	92.50	0.66	0.75	0.93
MI = MCV/RBC	13	54.30	47.95	77.50	0.21	0.25	0.75
RDWI = MCV ×RDW/RBC	220	67.20	60.96	90.00	0.42	0.51	0.92
RI = RDW/RBC	3.3	87.63	86.30	92.50	0.70	0.79	0.98
S&L = MCV^2×MCH/100	1530	74.19	92.47	7.50	−0.01	0.00	0.31
SI = MCV − RBC - 3Hb	27	53.76	45.21	85.00	0.26	0.302	0.80
SF = MCH/RBC	3.8	37.63	26.71	77.50	0.04	0.04	0.68
SiF = 1.5Hb − 0.05MCV	14	24.73	31.51	0.00	−0.56	−0.69	0.02
KF1 = RBC/Hct + 0.5RDW	8.2	70.97	64.38	95.00	0.49	0.59	0.93
KF2 = 5RDW/RBC	16.8	89.79	89.04	92.50	0.74	0.82	0.98

Ac Accuracy; *AUC* Area under receiver operating curve; *Hb* Hemoglobin; *Hct* Hematocrit; *IDA* Iron deficiency anemia; *MCC* Matthew's correlation coefficient; *MCH* mean corpuscular hemoglobin; *MCV* mean corpuscular volume; *RBC* red blood cell count; *RDW* red blood cell distribution width; *Sn* Sensitivity; *Sp* Specificity; *TT* Thalassemia trait; *YI* Youden's index

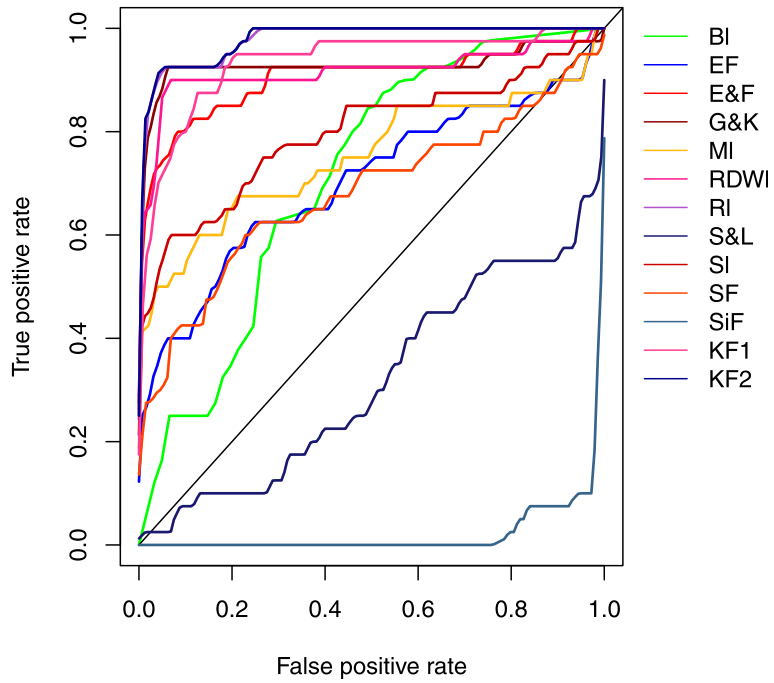

Fig. 3 Performance comparison of existing discriminant formulas and indices using ROC curves

Table 4 Performance comparisons between DT, RF and SVM models in differentiation of iron deficiency anemia from thalassemia trait

Classifier	Validation	Ac (%)	Sp (%)	Sn (%)	MCC	YI	AUC
k-NN	5-fold CV	92.36 ± 1.67	90.48 ± 3.62	92.80 ± 1.76	0.76 ± 0.06	0.83 ± 0.04	0.81 ± 0.07
	External	92.54 ± 4.26	90.09 ± 10.79	93.35 ± 4.22	0.77 ± 0.14	0.83 ± 0.13	0.80 ± 0.10
	Independent	90.20	83.33	100.00	0.82	0.83	0.85
DT	5-fold CV	98.03 ± 0.91	96.15 ± 3.40	98.54 ± 0.99	0.94 ± 0.026	0.93 ± 0.06	1.00 ± 0.00
	External	93.83 ± 4.10	86.86 ± 11.16	96.22 ± 3.77	0.82 ± 0.12	0.82 ± 0.04	0.92 ± 0.06
	Independent	92.16	86.21	100.00	0.85	0.85	1.00
RF	5-fold CV	94.17 ± 1.26	88.15 ± 3.49	95.75 ± 0.95	0.83 ± 0.04	0.83 ± 0.04	0.97 ± 0.01
	External	94.62 ± 3.29	90.07 ± 8.99	96.13 ± 3.19	0.84 ± 0.10	0.84 ± 0.07	0.98 ± 0.02
	Independent	92.16	86.21	100.00	0.85	0.85	1.00
ANN	5-fold CV	94.11 ± 1.31	86.75 ± 3.58	96.14 ± 1.06	0.83 ± 0.04	0.83 ± 0.04	0.97 ± 0.02
	External	93.78 ± 3.71	86.81 ± 10.84	96.22 ± 3.30	0.82 ± 0.11	0.83 ± 0.11	0.98 ± 0.02
	Independent	94.12	89.29	100.00	0.89	0.89	1.00
SVM	5-fold CV	95.05 ± 1.06	89.81 ± 2.66	96.45 ± 0.96	0.85 ± 0.03	0.85 ± 0.04	0.97 ± 0.01
	External	95.59 ± 2.76	92.49 ± 8.47	96.74 ± 2.59	0.87 ± 0.08	0.87 ± 0.10	0.98 ± 0.03
	Independent	96.08	92.59	100.00	0.92	0.92	1.00

The data are shown as mean ± standard deviation (100 times)
Ac Accuracy; *ANN* Artificial neural network; *AUC* Area under receiver operating curve; *DT* Decision tree; *k-NN* k-nearest neighbor; *MCC* Matthew's correlation coefficient; *RF* Random forest; *Sn* Sensitivity; *Sp* Specificity; *SVM* Support vector machine; *YI* Youden's index; *5-fold CV* 5-fold cross validation
Parameters of *k*-NN (*k*), RF (ntree, mtry), ANN (*size*, *decay*) and SVM (cost, γ) were optimized by a 5-fold CV procedure. Values of *k,* ntree, mtry, *size*, *decay,* cost and γ are 5, 200, 2, 4, 0.5, 8 and 0.5

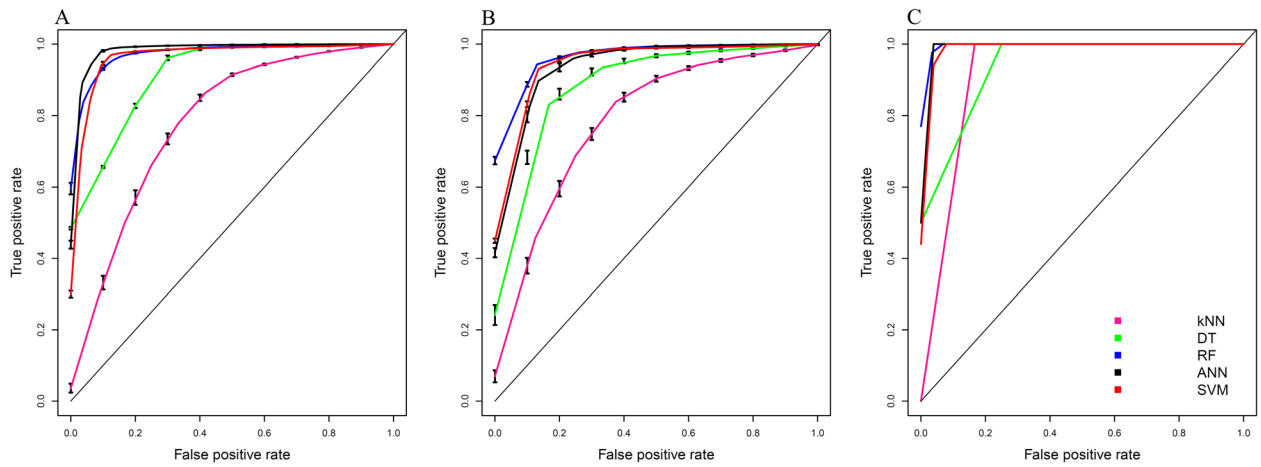

Fig. 4 Performance comparisons among k-NN, DT, RF, ANN and SVM models using ROC curves over 5-fold CV (**a**) external test (**b**) and independent test (**c**), where k-NN, DT, RF, ANN and SVM models are represented by pink, green, red, black and blue, respectively

performed on the balanced internal and new external sets were (92.75%, 0.88) and (94. 30%, 0.77), respectively. These results indicated that SVM model can tackle the imbalanced dataset problem and provide desirable prediction results [48].

By observing the performance comparisons in Table 4 and Fig. 4, we conclude that DT model shows the highest performance level when evaluated by 5-fold CV, while SVM model outperform that other conventional models over the external validation test and independent dataset. This observation is consistent with the previous works [28, 49]. Furthermore, many studies have mentioned that the overfitting is the major problem of DT model [50, 51]. For convenience, this best predictor for discriminating between IDA and TT (based on SVM model) will be referred to as 'ThalPred'.

For convenience, this best predictor for discriminating between IDA and TT (based on SVM model) will be referred to as 'ThalPred'.

Extracted important rules obtained from the RF model

The interpretable rules were established by using the RF model to demonstrate the combination of RBC indices for discriminating IDA from TT. Table 5 presents the eight interpretable rules as conditions in a simple linguistic manner, where *Frequency (%)* is the percentage of a data satisfying a condition, *Error* is the error percentage of a rule and *Prediction* is the outcome of a rule.

The first rule, which covers 63.98% of the whole data, was constructed with single RBC indices, e.g. Hb. This rule has one criterion: if the value of Hb is larger than 10.95 g/dL, then the prediction is TT. The second rule, which covers 15.05% of whole data, was constructed with two RBC indices, e.g. RBC and Hb. This rule has two criteria: if (i) the value of RBC is equal to or less than 4.5×10^6/μL and (ii) the value of Hb is equal to or less than 10.45 g/dL, then the prediction is IDA. The third rule, which covers 11.19% of the whole data, was depicted with single RBC indices, e.g. RDW. This rule has one criterion: if the value of RDW is equal or less than 17.15%, then the prediction is TT. The fourth rule, which covers 4.84% of the whole data, is constructed with three RBC indices, e.g. RBC, Hb and RDW. This rule has three criteria: if (i) the value of RBC is greater than 4.59×10^6/μL and (ii) the value of Hb is equal to or

Table 5 The extracted interpretable rules derived from RF model in differentiation of iron deficiency anemia from thalassemia trait

Length	Frequency (%)	Error (%)	Condition	Prediction
1	63.98	0.00	Hb > 10.95	TT
2	15.05	0.00	RBC ≤ 4.5 and Hb ≤ 10.45	IDA
1	11.29	0.00	RDW ≤ 17.15	TT
3	4.84	0.00	RBC > 4.59 and Hb ≤ 10.95 and RDW > 17.7	IDA
4	2.15	0.00	RBC > 4.28 and MCHC ≤32.15 and RDW > 17.15 and RDW ≤ 17.7	TT
3	1.08	0.00	Hb ≤ 11.45 and MCHC > 31.35 and RDW > 17.4	IDA
1	1.61	33.00	Else	TT

Hb Hemoglobin; *IDA* Iron deficiency anemia; *MCHC* mean corpuscular hemoglobin concentration; *RBC* red blood cell count; *RDW* red blood cell distribution width; *TT* Thalassemia trait

less than 10.95 g/dL and (iii) the value of RDW is greater than 17.7%, then the prediction is IDA. The fifth rule, which covers 2.15% of whole data, was constructed with three RBC indices, e.g. RBC, MCHC and RDW. This rule has four criteria: if (i) the value of RBC is greater than 4.28 × 10^6/μL and (ii) the value of MCHC is equal to or less than 32.15 g/dL and (iii) the value of RDW is greater than 17.15% and (iv) the value of RDW is greater than 17.7%, then the prediction is TT. The sixth rule, which covers 1.08% of whole data, was constructed with three RBC indices, e.g. Hb, MCHC and RDW. This rule has three criteria: if (i) the value of Hb is equal to or less than 11.45 g/dL, (ii) the value of MCHC is greater than 31. 35 g/dL, and (iii) the value of RDW is greater than 17.4%, then the prediction is IDA. The seventh rule covers 1. 61% of the whole data. If a query has RBC indices which do not satisfy any of the seven interpretable rules, then it is classified as TT with an error of 33.00%.

Web-based tool implementation

For the convenience of health care team, based on the best model (e.g. SVM model) proposed in the present work, a publicly accessible web-based tool for ThalPred has been established. A screenshot of the ThalPred is shown in Fig. 5. Furthermore, to maximize user's convenience, a step-by-step walkthrough of the procedures for using the ThalPred web based tool is provided below.

Step 1. Open the web-based tool at http://codes.bio/thalpred/.

Step 2. There are two options for prediction:

1) Key the single data in the *input parameters panel.*

2) Upload the query patient's RBC indices (RBC, Hb, Hct, MCV, MCH, MCHC and RDW) into the input box by clicking on the *Choose file* button. The input data should be in CSV format. For example, of patient's RBC indices in CSV format, click the *insert example data* above the input box. Finally, press on the *Submit* button to initiate the prediction process.

Step 3. Prediction results are automatically displayed in a grey box found below the Status/Output heading. Users can also download the prediction results as a CSV file by pressing on the *Download results* bottom.

Discussion

Anemia is crucial problem occurring with population in Thailand and affecting to health and economic system. HMA, such as TT and IDA, are found commonly in clinical laboratory. Discrimination between TT and IDA group is essential for correct genetic counseling and effective treatment. RBC indices, consisting of RBC, Hb, Hct, MCV, MCH, MCHC and RDW, are generated from complete blood count (CBC) analysis using automated instruments. Therefore, using parameters from the CBC result for differentiation is of much interest. Generally, the RBC count, Hb, MCV, MCH and RDW are used to formulate a new index. As a result of the anemic condition, subjects with TT are found to have an increased RBC count [22, 52, 53]. While blood films from individuals with TT or IDA have shown different levels of

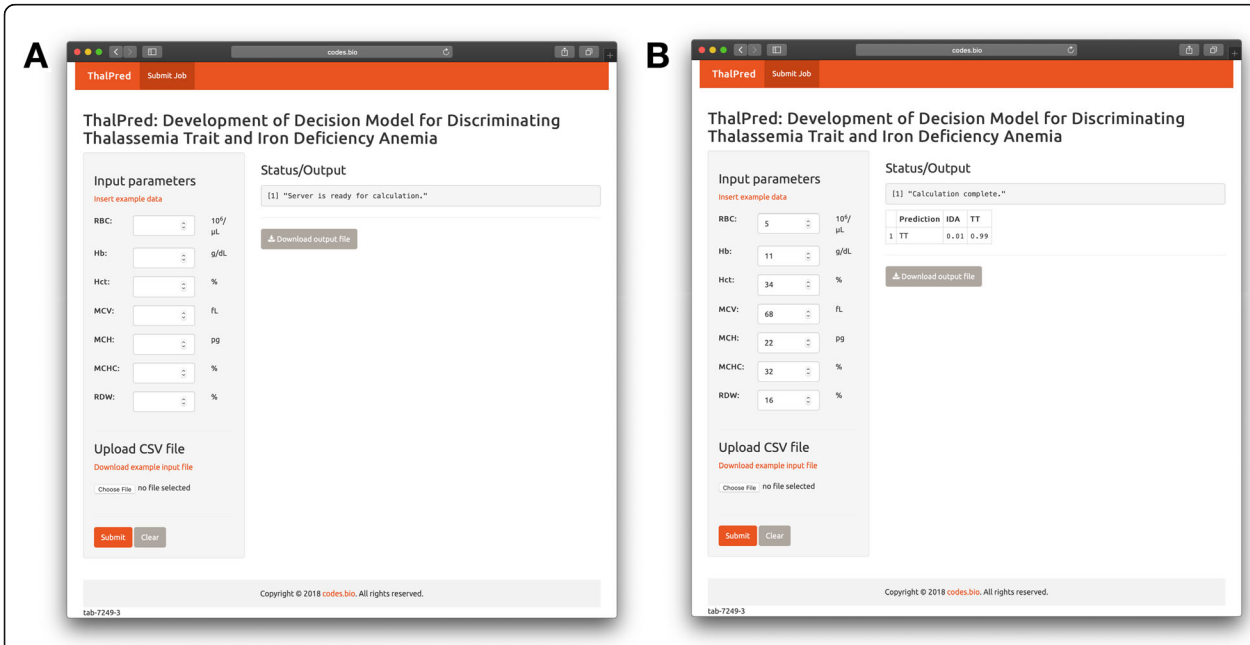

Fig. 5 Screenshots of the ThalPred web-based tool before (**a**) and after (**b**, **c**) submission of laboratory data, which is available at http://codes.bio/thalpred/

Table 6 The prediction results derived from DT, RF and SVM models in differentiation of iron deficiency anemia (IDA) from thalassemia trait (TT)

No.	RBC	Hb	Hct	MCV	MCH	MCHC	RDW	Diagnosis	Prediction				
									k-NN	DT	RF	ANN	SVM
1	5.35	10.6	33.0	61.7	19.8	32.1	13.7	TT	TT	TT	TT	TT	TT
2	5.40	10.9	34.1	63.2	20.2	32.0	14.5	TT	TT	TT	TT	TT	TT
3	5.40	10.3	33.0	61.1	19.1	31.2	14.2	TT	TT	TT	TT	TT	TT
4	6.01	12.3	38.3	63.7	20.5	32.1	13.2	TT	TT	TT	TT	TT	TT
5	5.55	11.0	33.4	60.2	19.8	32.9	13.2	TT	TT	TT	TT	TT	TT
6	6.04	12.9	39.6	65.6	21.4	32.6	13.5	TT	TT	TT	TT	TT	TT
7	5.90	13.1	40.0	67.8	22.2	32.8	12.7	TT	TT	TT	TT	TT	TT
8	5.95	12.6	39.0	65.6	21.2	32.3	13.4	TT	TT	TT	TT	TT	TT
9	6.11	12.7	39.5	64.7	20.8	32.2	13.7	TT	TT	TT	TT	TT	TT
10	5.45	11.8	36.6	67.1	21.7	32.3	14.0	TT	TT	TT	TT	TT	TT
11	5.40	11.0	34.0	63.0	20.4	32.4	12.8	TT	TT	TT	TT	TT	TT
12	5.40	11.0	34.0	63.0	20.4	32.4	14.0	TT	TT	TT	TT	TT	TT
13	6.20	12.4	37.2	60.0	20.0	33.3	12.6	TT	TT	TT	TT	TT	TT
14	5.40	10.7	33.0	61.1	19.8	32.4	13.6	TT	TT	TT	TT	TT	TT
15	6.11	12.3	38.0	62.2	20.1	32.4	12.6	TT	TT	TT	TT	TT	TT
16	3.40	7.7	24.1	70.9	22.7	32.0	20.1	IDA	IDA	IDA	IDA	IDA	IDA
17	4.66	11.3	34.4	73.8	24.3	32.9	21.0	IDA	TT	TT	TT	IDA	TT
18	4.54	10.6	32.8	72.3	23.4	32.3	21.0	IDA	IDA	IDA	IDA	IDA	IDA
19	3.50	7.9	25.2	72.0	22.6	31.4	21.2	IDA	IDA	IDA	IDA	IDA	IDA
20	4.15	9.9	29.0	69.9	23.9	34.1	20.2	IDA	TT	IDA	IDA	IDA	IDA
21	3.90	8.9	28.0	71.8	22.8	31.8	20.2	IDA	IDA	IDA	IDA	IDA	IDA
22	4.17	9.9	29.0	69.5	23.7	34.1	21.1	IDA	IDA	IDA	IDA	IDA	IDA
23	3.85	8.5	27.5	71.4	22.1	30.9	21.3	IDA	IDA	IDA	IDA	IDA	IDA
24	4.24	9.8	30.5	71.9	23.1	32.1	20.2	IDA	IDA	IDA	IDA	IDA	IDA
25	4.83	11.1	34.0	70.4	23.0	32.7	19.0	IDA	TT	TT	TT	TT	TT
26	4.64	10.6	33.0	71.1	22.8	32.1	20.5	IDA	IDA	IDA	IDA	IDA	IDA
27	4.01	9.5	29.0	72.3	23.7	32.8	20.4	IDA	IDA	IDA	IDA	IDA	IDA
28	4.80	11.0	34.0	70.8	22.9	32.4	21.1	IDA	TT	TT	TT	IDA	IDA
29	3.65	8.4	26.0	71.2	23.0	32.3	20.4	IDA	IDA	IDA	IDA	IDA	IDA
30	4.00	9.2	28.0	70.0	23.0	32.9	20.1	IDA	IDA	IDA	IDA	IDA	IDA
31	4.45	10.2	31.4	70.6	22.9	32.5	19.8	IDA	IDA	IDA	IDA	IDA	IDA
32	4.44	10.3	32.0	72.1	23.2	32.2	21.3	IDA	IDA	IDA	IDA	IDA	IDA
33	4.56	10.4	32.0	70.2	22.8	32.5	19.4	IDA	IDA	IDA	IDA	IDA	IDA
34	4.84	11.0	34.0	70.3	22.7	32.4	20.7	IDA	TT	TT	TT	IDA	IDA
36	3.83	8.7	27.0	70.5	22.72	32.22	20.0	IDA	IDA	IDA	IDA	IDA	IDA
37	3.6	8.7	25.0	69.44	24.17	34.8	19.3	IDA	IDA	IDA	IDA	IDA	IDA
38	3.52	8.0	25.0	71.02	22.73	32	21.1	IDA	IDA	IDA	IDA	IDA	IDA
39	3.98	9.0	27.8	69.85	22.61	32.37	20.7	IDA	IDA	IDA	IDA	IDA	IDA
40	4.02	9.0	28.5	70.9	22.39	31.58	21.1	IDA	IDA	IDA	IDA	IDA	IDA
41	4.43	10.1	31.4	70.88	22.8	32.17	19.8	IDA	IDA	IDA	IDA	IDA	IDA
42	4.24	9.6	30.0	70.75	22.64	32.0	21.0	TT	TT	TT	TT	TT	TT
43	5.94	12.6	39.0	65.7	21.2	32.3	13.0	TT	IDA	TT	TT	IDA	TT

Table 6 The prediction results derived from DT, RF and SVM models in differentiation of iron deficiency anemia (IDA) from thalassemia trait (TT) *(Continued)*

No.	RBC	Hb	Hct	MCV	MCH	MCHC	RDW	Diagnosis	Prediction				
									k-NN	DT	RF	ANN	SVM
44	5.80	11.7	36.3	62.6	20.2	32.2	13.1	TT	IDA	TT	TT	IDA	TT
45	5.45	11.4	35.0	64.2	20.9	32.6	13.0	TT	IDA	TT	TT	IDA	TT
46	5.40	11.0	33.7	62.4	20.4	32.6	12.5	TT	IDA	TT	TT	IDA	TT
47	6.11	12.6	38.5	63.0	20.6	32.7	14.3	TT	IDA	TT	TT	IDA	TT
48	5.80	12.2	36.4	62.8	21.0	33.5	14.0	TT	IDA	TT	TT	IDA	TT
49	5.44	10.8	34.0	62.5	19.9	31.8	12.7	TT	IDA	TT	TT	IDA	TT
50	5.40	10.8	33.0	61.1	20.0	32.7	13.6	TT	TT	TT	TT	TT	TT
51	6.05	12.3	38.0	62.8	20.3	32.4	14.0	TT	TT	TT	TT	TT	TT

ANN Artificial neural network; *DT* Decision tree; *Hb* Hemoglobin; *Hct* Hematocrit; *IDA* Iron deficiency anemia; *k-NN k*-nearest neighbor; *MCH* mean corpuscular hemoglobin; *MCHC* mean corpuscular hemoglobin concentration; *MCV* mean corpuscular volume; *RBC* red blood cell count; *RDW* red blood cell distribution width; *RF* Random forest; *SVM* Support vector machine; *TT* Thalassemia trait. Parameters of *k*-NN (*k*), RF (ntree, mtry), ANN (*size*, *decay*) and SVM (cost, γ) were optimized by a 5-fold CV procedure. Values of *k,* ntree, mtry, *size*, *decay,* cost and γ are 5, 200, 2, 4, 0.5, 8 and 0.5

anisocytosis (a condition of variation in sizes of RBC). Most of IDA patients appeared to have higher RDW than TT patients [22, 54]. The results of both univariate (Table 2) and multivariate (Fig. 2) analyses are agreeable with previous studies. Hence, one parameter is not enough to discriminate between these two conditions and all RBC parameters have potential role on the differentiation.

Previously, several discriminant formulas and indices have been proposed by different researchers for discrimination between IDA from TT [8–19]. However, these formulas and indices are not appropriate for every population. In Thailand, many researchers attempted to utilize these existing formulas and indices with different interest groups, e.g. school children [55], adults [56], anemic vegetarians [57], etc. In this study, the group of dataset composed of adults between 18 to 50 years old. Our finding showed that G&K, RI and KF2 performed well on our dataset. Interestingly, the top- three formulas and indices demonstrated the highest performance harboring of RDW and 2 of 3 consisting of RBC parameter.

Prior studies utilizing data from Thai people also found that most formulas containing RBC and RDW yielded a high performance. Pornprasert et al. suggested that SI and SF are the most reliable formulas [55], whereas RI was found to be most efficient in studies by Plengsuree et al. [56] and Sirachainan et al [18]. Conversely, S&L (which does not contain RBC and RDW) was proposed as a suitable index for differentiation in young Asians. It achieved 100% Sn and high Sp in Indian and Taiwanese [22, 54]. Our findings demonstrated consistent results to the previous publications that RBC and RDW play key roles in accurately distinguishing these two diseases. However, 89.8% Ac is not sufficient for medical decision making [58, 59]. Therefore, we attempted to increase the efficiency and accuracy of discrimination between TT and IDA in Thai adults by constructing a new model from hematological indices via machine learning approach.

As seen in Table 3 and Fig. 3**,** the prediction performances of the existing formulas and indices are still not satisfactory and there is room for improvement. Computational models based on machine learning approaches may further enhance prediction performance as well as provide effective large-scale analysis of available clinical data. The final prediction performances of the 5-fold CV and external validation tests were obtained by averaging the 100 corresponding performances of the internal and external sets, respectively. Furthermore, an independent dataset (Table 6) was used to assess the true predictive power of our proposed discriminating model.

As noticed in Table 6, the four or five prediction models give the prediction results on samples 17 and 25 as TT case, but the correct result is IDA case. These results might be due to the distribution of samples 17 and 25 is close to TT case, as illustrated in Fig. 6**.** Another reason supported our results is that thalassemia and IDA are most common etiologies of hypochromic microcytic anemia in Thailand therefore, both of them are close differential diagnosis and can coexist together. Prior studies showed that there is frequent occurrence of IDA in TT [60, 61]. Interestingly, this finding inspires us to perform future research for construction new efficient model to distinguish three groups including TT, IDA and coexisting of these two conditions.

As shown in Table 4, SVM and RF models perform well with the highest and second highest prediction results, respectively, for correctly discriminating IDA from TT. However, SVM models do not lend themselves as easily to interpretability. SVM is classified as black box model that could work perfectly with unknown data [62] because it overcame the problem of overfitting from DT

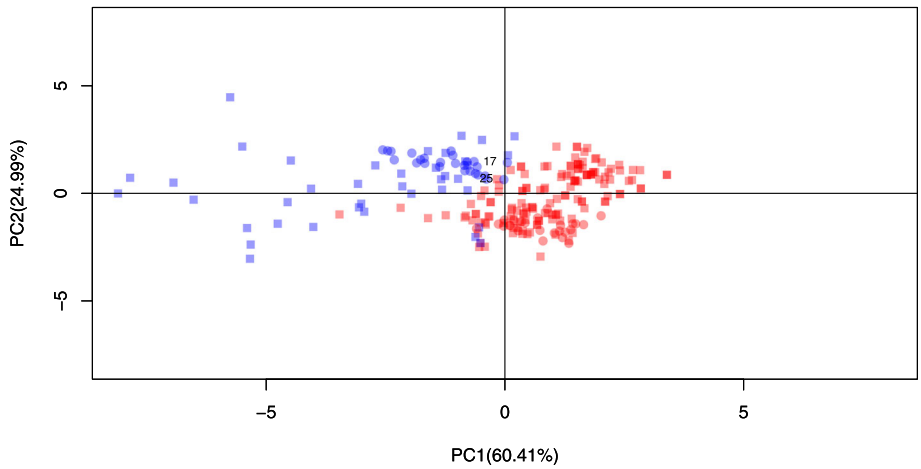

Fig. 6 Multivariate analysis using principal component analysis (PCA) of internal (square) and external (circles) sets, where red and blue represent TT and IDA cases, respectively

[63]. According to DT model showed good performance on the training dataset but poor classification to other dataset in case the data is various detail and complex. RF could solve this limitation of DT [64]. Therefore, the seven interpretable rules (Table 5) extracting from RF were provided to represent criteria of each parameters for discrimination.

By observing the prediction results listed in Tables 3 and 4, we can clearly find that ThalPred (based on SVM model) has good discriminating power and outperforms all existing formulas and indices with the highest values of MCC (0.92), AUC (1.00) and YI (0.92). Improvements of MCC/YI with 18%/10 and 22%/12% for MCC and YI, respectively, were observed when compared with the best (KF2) and second-best (RI) existing formulas. Hence, ThalPred had better generalization capability for discrimination between IDA from TT than the existing formulas and indices. Further, ThalPred developed in the web server using SVM model [65] is stable and reliable to assist health care team for discrimination. This study is based on the small size [66, 67] of IDA and performed unbalancing dataset [68, 69]. So, the predictor may not be robust enough to apply on a very diverse dataset. As soon as more patients' data require retraining as the new independent dataset to make the predictor more robust. This suggested that partition of the patients' age for training dataset to make more sensitive and specific should be considered for the future development of effective models. The mobile application also be one choice for more convenient and available for all users.

Conclusion

Discrimination between patients with IDA and TT is still a challenging problem due to the diversity of populations with anemia. Computational models can accelerate the process of screening HMA patients and save a lot of expenses and time. In this study, we have extracted an interpretable rule and established a web-based tool for discriminating IDA from TT. The prediction results for both cross-validation and independent validation tests on our laboratory data demonstrate the superiority of ThalPred over existing indices and formulas. Furthermore, a user-friendly web-based tool for ThalPred was established at http://codes.bio/thalpred/, by which users can easily obtain the prediction result without the need to follow the mathematical and computational details. We believe that the proposed ThalPred will supplement the existing indices and formulas as well as facilitate the health care provider.

Abbreviations

5-fold CV: Five-fold cross validation; Ac: Accuracy; ANN: Artificial neural network; AUC: Area under receiver operating characteristics curve; CART: classification and regression tree; CE: Capillary electrophoresis; DT: Decision tree; Hb: Hemoglobin; Hct: Hematocrit; HMA: Hypochromic microcytic anemia; IDA: Iron deficiency anemia; *k*-NN: *k*-nearest neighbor; LPLC: Low pressure liquid chromatography; MCC: Matthew correlation coefficient; MCH: Mean corpuscular hemoglobin; MCHC: Mean corpuscular hemoglobin concentration; MCV: Mean corpuscular volume; PCA: Principal component analysis; PCs: Principal components; RBC: Red blood cell; RDW: Red blood cell distribution width; RF: Random forest; SD: Standard deviation; Sn: Sensitivity; Sp: Specificity; SVM: Support vector machine; TT: Thalassemia trait; YI: Youden's Index

Acknowledgements

We thank the editor and three anonymous reviewers for valuable comments and suggestion on our study.

Authors' contributions

All authors have read and approved of the final version of the manuscript. VL: data/machine learning analysis, result interpretation and manuscript preparation. WS: computational model, algorithm analysis, manuscript preparation and website construction. WA: correlation and interpretation of laboratory data and data processing. CN: project discussion and website construction. VP: result analysis and discussion. PN: project initiation, study

design, correlation and interpretation of laboratory data, and manuscript preparation.

Funding
This study was supported by the Faculty of Medical Technology, Mahidol University, Office of the Higher Education Commission (OHEC) and the Thailand Research Fund to PN (MRG5460082) and to WS (MRG6180226) as well as the Ministry of Europe and Foreign Affairs (MEAE), the Ministry of Higher Education, Research and Innovation (MESRI) and the Office of the Higher Education Commission (OHEC) via the Franco - Thai Cooperation Program in Higher Education and Research (Frnco-Thai Mobility Progra/PHC SIAM) to PN. The funding body has no role in the design of the study, data collection, analysis and interpretation of data and in writing the manuscript.

Availability of data and materials
The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Ethics approval and consent to participate
The study was conducted under the approval of Mahidol University Central Institutional Review Board (MU-CIRB: 2016/084.0311) We received a participant consent waiver from MU-CIRB. All information of subjects was de-identified prior data analysis.

Consent for publication
Not Applicable

Competing interests
The authors declare that they have no competing interest and the research was conducted in the absence of any commercial or financial relationship that could be influenced as potential conflict of interest.

Author details
[1]Center for Research and Innovation, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand. [2]Department of Clinical Microscopy, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand. [3]Center of Data Mining and Medical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand. [4]Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand.

Received: 11 October 2017 Accepted: 14 October 2019
Published online: 07 November 2019

References

1. Kassebaum NJ, Jasrasaria R, Naghavi M, Wulf SK, Johns N, Lozano R, et al. A systematic analysis of global anemia burden from 1990 to 2010. Blood. 2014;123:615–24. https://doi.org/10.1182/blood-2013-06-508325.
2. Winichagoon P. Prevention and control of anemia: Thailand experiences. J Nutr. 2002;132(4 Suppl):862S–6S.
3. Thong XY, Lee LY, Chia DA, Wong YC, Biswas A. Management and outcomes of fetal hydrops in a tertiary care centre in singapore. Ann Acad Med Singap. 2017;46:4–10.
4. Johnson-Wimbley TD, Graham DY. Diagnosis and management of iron deficiency anemia in the 21st century. Therap Adv Gastroenterol. 2011;4:177–84. https://doi.org/10.1177/1756283X11398736.
5. Nathalang O, Arnutti P, Nillakupt K. Thalassemia screening among Royal Thai Army medical cadets. Mil Med. 2004;169:274–6. https://doi.org/10.7205/MILMED.169.4.274.
6. Nathalang O, Nillakupt K, Arnutti P, Boonsiri T, Panichkul S, Areekul W. Screening for thalassemia and hemoglobinopathy in a rural area of Thailand: a preliminary study. J Med Assoc Thai. 2005;88(Suppl 3):S35–42.
7. Ou Z, Li Q, Liu W, Sun X. Elevated hemoglobin A2 as a marker for β-thalassemia trait in pregnant women. Tohoku J Exp Med. 2011;223:223–6.
8. Bessman JD, Feinstein DI. Quantitative anisocytosis as a discriminant between iron deficiency and thalassemia minor. Blood. 1979;53:288–93.
9. Ehsani MA, Shahgholi E, Rahiminejad MS, Seighali F, Rashidi A. A new index for discrimination between iron deficiency anemia and beta-thalassemia minor: results in 284 patients. Pak J Biol Sci. 2009;12:473–5.
10. England JM, Fraser PM. Differentiation of iron deficiency from thalassaemia trait by routine blood-count. Lancet. 1973;1:449–52.
11. Green R, King R. A new red cell discriminant incorporating volume dispersion for differentiating iron deficiency anemia from thalassemia minor. Blood Cells. 1989;15:481–91 discussion 492.
12. Mentzer WC. Differentiation of iron deficiency from thalassaemia trait. Lancet. 1973;1:882.
13. Jayabose S, Giamelli J, LevondogluTugal O, Sandoval C, Ozkaynak F, Visintainer P. Differentiating iron deficiency anemia from thalassemia minor by using an RDW-based index. J Pediatr Hematol Oncol. 1999;21(4):314.
14. Ricerca BM, Storti S, d'Onofrio G, Mancini S, Vittori M, Campisi S, et al. Differentiation of iron deficiency from thalassaemia trait: a new approach. Haematologica. 1987;72:409–13.
15. Shine I, Lal S. A strategy to detect beta-thalassaemia minor. Lancet. 1977;1:692–4.
16. Sirdah M, Tarazi I, Al Najjar E, Al HR. Evaluation of the diagnostic reliability of different RBC indices and formulas in the differentiation of the beta-thalassaemia minor from iron deficiency in Palestinian population. Int J Lab Hematol. 2008;30:324–30. https://doi.org/10.1111/j.1751-553X.2007.00966.x.
17. Srivastava PC, Bevington JM. Iron deficiency and-or thalassaemia trait. Lancet. 1973;1:832.
18. Sirachainan N, Iamsirirak P, Charoenkwan P, Kadegasem P, Wongwerawattanakoon P, Sasanakul W, et al. New mathematical formula for differentiating thalassemia trait and iron deficiency anemia in thalassemia prevalent area: a study in healthy school-age children. Southeast Asian J Trop Med Public Health. 2014;45:174–82.
19. Hafeez Kandhro A, Shoombuatong W, Prachayasittikul V, Nuchnoi P. New Bioinformatics-Based Discrimination Formulas for Differentiation of Thalassemia Traits From Iron Deficiency Anemia. Lab Med. 2017;48:230–7. https://doi.org/10.1093/labmed/lmx029.
20. van Alphen B, BHJ W, Frens MA. Age- and sex-related differences in contrast sensitivity in C57BL/6 mice. Invest Ophthalmol Vis Sci. 2009;50:2451–8. https://doi.org/10.1167/iovs.08-2594.
21. Ambayya A, Su AT, Osman NH, Nik-Samsudin NR, Khalid K, Chang KM, et al. Haematological reference intervals in a multiethnic population. PLoS ONE. 2014;9:e91968. https://doi.org/10.1371/journal.pone.0091968.
22. Huang T-C, Wu Y-Y, Chen Y-G, Lai S-W, Wu S-C, Ye R-H, et al. Discrimination index of microcytic anemia in young soldiers: a single institutional analysis. PLoS ONE. 2015;10:e0114061. https://doi.org/10.1371/journal.pone.0114061.
23. Miri-Moghaddam E, Sargolzaie N. Cut off determination of discrimination indices in differential diagnosis between iron deficiency anemia and β-thalassemia minor. Int J Hematol Oncol Stem Cell Res. 2014;8:27–32.
24. Bordbar E, Taghipour M, Zucconi BE. Reliability of Different RBC Indices and Formulas in Discriminating between β-Thalassemia Minor and other Microcytic Hypochromic Cases. Mediterr J Hematol Infect Dis. 2015;7:e2015022. https://doi.org/10.4084/MJHID.2015.022.
25. Shoombuatong W, Prathipati P, Prachayasittikul V, Schaduangrat N, Malik AA, Pratiwi R, et al. Towards Predicting the Cytochrome P450 Modulation: From QSAR to proteochemometric modeling. Curr Drug Metab. 2017. https://doi.org/10.2174/1389200218666170320121932.
26. Shoombuatong W, Prachayasittikul V, Anuwongcharoen N, Songtawee N, Monnor T, Prachayasittikul S, et al. Navigating the chemical space of dipeptidyl peptidase-4 inhibitors. Drug Des Devel Ther. 2015;9:4515–49. https://doi.org/10.2147/DDDT.S86529.
27. Shoombuatong W, Prachayasittikul V, Prachayasittikul V, Nantasenamat C. Prediction of aromatase inhibitory activity using the efficient linear method (ELM). EXCLI J. 2015;14:452–64. https://doi.org/10.17179/excli2015-140.
28. Win TS, Malik AA, Prachayasittikul V, JE SW, Nantasenamat C, Shoombuatong W. HemoPred: a web server for predicting the hemolytic activity of peptides. Future Med Chem. 2017;9:275–91. https://doi.org/10.4155/fmc-2016-0188.
29. Le S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. J Stat Software. 2008;25:1–8.
30. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna: the R Foundation for Statistical Computing; 2011. ISBN: 3–900051–07-0. http://www.R-project.org/
31. Quinlan JR. C4.5: Programs for Machine Learning. Amsterdam: Elsevier; 2014.
32. Cutler A. randomForest: Random Forests for Classification and Regression; 2006.
33. Holmes G, Donkin A, Witten IH. WEKA: a machine learning workbench. In: Proceedings of ANZIIS '94 - Australian New Zealand intelligent information systems conference. IEEE; 1994. p. 357–61. https://doi.org/10.1109/ANZIIS.1994.396988.

34. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and Regression Trees* Taylor & Francis Group: CRC Press; 1984.
35. Calle ML, Urrea V. Letter to the editor: stability of random forest importance measures. Brief Bioinform. 2011;12(1):86–9.
36. Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J Pharm Biomed Anal. 2000;22:717–27. https://doi.org/10.1016/S0731-7085(99)00272-1.
37. Liu X. Deep Recurrent Neural Network for Protein Function Prediction from Sequence. BioRxiv. 2017. https://doi.org/10.1101/103994.
38. Liang M, Hu X. Recurrent convolutional neural network for object recognition. In: 2015 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE; 2015. p. 3367–75. https://doi.org/10.1109/CVPR.2015.7298958.
39. Cortes C, Vapnik V. Support-vector networks. Mach.Learn. 1995;20(3):273–97.
40. Erfani SM, Rajasegarar S, Karunasekera S, Leckie C. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recognit. 2016;58:121–34. https://doi.org/10.1016/j.patcog.2016.03.028.
41. David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2017). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6–8. https://CRAN.R project.org/package=e1071
42. Motsinger AA, Ritchie MD. The effect of reduction in cross-validation intervals on the performance of multifactor dimensionality reduction. Genet Epidemiol. 2006;30:546–55. https://doi.org/10.1002/gepi.20166.
43. Baratloo A, Hosseini M, Negida A, El Ashal G. Part 1: Simple definition and calculation of accuracy, sensitivity and specificity. Emerg (Tehran). Spring. 2015;3(2):48–49.
44. Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol. 2008;56(1):45–50.
45. Gorodkin J. Comparing two K-category assignments by a K-category correlation coefficient. Comput Biol Chem. 2004;28(5-6):367–74.
46. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861–74.
47. Sing T, et al. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21(20):3940–1.
48. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS ONE. 2017;12:e0177678. https://doi.org/10.1371/journal.pone.0177678.
49. Agrawal P, Kumar S, Singh A, Raghava GPS, Singh IK. NeuroPIpred: a tool to predict, design and scan insect neuropeptides. Sci Rep. 2019;9:5129. https://doi.org/10.1038/s41598-019-41538-x.
50. Bramer M. Avoiding overfitting of decision trees. In: Principles of data mining. London: Springer London; 2013. p. 121–36. https://doi.org/10.1007/978-1-4471-4884-5_9.
51. Gupta D. Applied analytics through case studies using Sas and R: implementing predictive models and machine learning techniques. Berkeley: Apress; 2018. https://doi.org/10.1007/978-1-4842-3525-6.
52. Ntaios G, Chatzinikolaou A, Saouli Z, Girtovitis F, Tsapanidou M, Kaiafa G, et al. Discrimination indices as screening tests for beta-thalassemic trait. Ann Hematol. 2007;86:487–91. https://doi.org/10.1007/s00277-007-0302-x.
53. Okan V, Cigiloglu A, Cifci S, Yilmaz M, Pehlivan M. Red cell indices and functions differentiating patients with the beta-thalassaemia trait from those with iron deficiency anaemia. J Int Med Res. 2009;37:25–30. https://doi.org/10.1177/147323000903700103.
54. Piplani S, Madaan M, Mannan R, Manjari M, Singh T, Lalit M. Evaluation of various discrimination indices in differentiating iron deficiency anemia and beta thalassemia trait: A practical low cost solution. Annal Pathol Lab Med. 2016;3(6):A551–59.
55. Pornprasert S, Panya A, Punyamung M, Yanola J, Kongpan C. Red cell indices and formulas used in differentiation of β-thalassemia trait from iron deficiency in Thai school children. Hemoglobin. 2014;38:258–61. https://doi.org/10.3109/03630269.2014.930044.
56. Plengsuree S, Punyamung M, Yanola J, Nanta S, Jaiping K, Maneewong K, et al. Red Cell Indices and Formulas Used in Differentiation of β-Thalassemia Trait from Iron Deficiency in Thai Adults. Hemoglobin. 2015;39:235–9. https://doi.org/10.3109/03630269.2015.1048352.
57. Wongprachum K, Sanchaisuriya K, Sanchaisuriya P, Siridamrongvattana S, Manpeun S, Schlep FP. Proxy indicators for identifying iron deficiency among anemic vegetarians in an area prevalent for thalassemia and hemoglobinopathies. Acta Haematol. 2012;127:250–5. https://doi.org/10.1159/000337032.
58. Lurie JD, Sox HC. Principles of medical decision making. Spine. 1999;24:493–8.
59. Krieg AF, Abendroth TW, Bongiovanni MB. When is a diagnostic test result positive? Decision tree models based on net utility and threshold. Arch Pathol Lab Med. 1986;110:787–91.
60. Verma S, Gupta R, Kudesia M, Mathur A, Krishan G, Singh S. Coexisting iron deficiency anemia and Beta thalassemia trait: effect of iron therapy on red cell parameters and hemoglobin subtypes. ISRN Hematol. 2014;2014:293216. https://doi.org/10.1155/2014/293216.
61. Mohammad N, Chowdhury MJ, Alam MR, Kabir AL, Rahaman MFU, Chakrabarty B. Co-existence of iron deficiency in beta thalassaemia trait. J Med. 2017;19:44. https://doi.org/10.3329/jom.v19i1.34842.
62. Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24:1565–7. https://doi.org/10.1038/nbt1206-1565.
63. Hawkins DM. The problem of overfitting. J Chem Inf Comput Sci. 2004;44:1–12. https://doi.org/10.1021/ci0342472.
64. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci. 2003;43:1947–58. https://doi.org/10.1021/ci034160g.
65. Li YH, Xu JY, Tao L, Li XF, Li S, Zeng X, et al. SVM-Prot 2016: A Web-Server for Machine Learning Prediction of Protein Functional Families from Sequence Irrespective of Similarity. PLoS ONE. 2016;11:e0155290. https://doi.org/10.1371/journal.pone.0155290.
66. Dougherty ER, Dalton LA. Scientific knowledge is possible with small-sample classification. EURASIP J Bioinform Syst Biol. 2013;2013:10. https://doi.org/10.1186/1687-4153-2013-10.
67. Popovici V, Chen W, Gallas BG, Hatzis C, Shi W, Samuelson FW, et al. Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. Breast Cancer Res. 2010;12:R5. https://doi.org/10.1186/bcr2468.
68. Khan SH, Hayat M, Bennamoun M, Sohel FA, Togneri R. Cost-Sensitive Learning of Deep Feature Representations From Imbalanced Data. IEEE Trans Neural Netw Learn Syst. 2017. https://doi.org/10.1109/TNNLS.2017.2732482.
69. Guo H, Zhi W, Liu H, Xu M. Imbalanced learning based on logistic discrimination. Comput Intell Neurosci. 2016;2016:5423204. https://doi.org/10.1155/2016/5423204.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

Computational Biology and Chemistry 80 (2019) 441–451

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac

THPep: A machine learning-based approach for predicting tumor homing peptides

Watshara Shoombuatong[a,*], Nalini Schaduangrat[a], Reny Pratiwi[a,b], Chanin Nantasenamat[a,*]

[a] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[b] Department of Medical Laboratory Technology, Faculty of Health Science, Setia Budi University, Surakarta 57127, Indonesia

ARTICLE INFO

Keywords:
Tumor homing peptide
Therapeutic peptide
Classification
Machine learning
Random forest

ABSTRACT

In the present era, a major drawback of current anti-cancer drugs is the lack of satisfactory specificity towards tumor cells. Despite the presence of several therapies against cancer, tumor homing peptides are gaining importance as therapeutic agents. In this regard, the huge number of therapeutic peptides generated in recent years, demands the need to develop an effective and interpretable computational model for rapidly, effectively and automatically predicting tumor homing peptides. Therefore, a sequence-based approach referred herein as THPep has been developed to predict and analyze tumor homing peptides by using an interpretable random forest classifier in concomitant with amino acid composition, dipeptide composition and pseudo amino acid composition. An overall accuracy and Matthews correlation coefficient of 90.13% and 0.76, respectively, were achieved from the independent test set on an objective benchmark dataset. Upon comparison, it was found that THPep was superior to the existing method and holds high potential as a useful tool for predicting tumor homing peptides. For the convenience of experimental scientists, a web server for this proposed method is provided publicly at http://codes.bio/thpep/.

1. Introduction

The burden of cancer has become a major cause of concern worldwide owing to its associated high morbidity and mortality. According to the World Health Organization (WHO) and the National Cancer Institute (NCI), more than 10 million new cases are reported each year and 1 in 6 of the global deaths reported is due to cancer (N. I. o. Health, 2017; W. H. Organization, 2017). Despite the progression in cancer drug development, the lack of specificity of chemotherapeutic drugs towards tumor cells and its toxicity to normal cells still represent key challenges in this field (Mäe et al., 2009; Myrberg et al., 2008). In an attempt to discover a more potent and specific therapy for tumor, the research community has focused their efforts in exploiting peptides as a tool to improve the targeted-drug delivery system. Over the last decade, three distinct classes of peptide-mediated drug delivery systems have been developed including, homing peptides (HPs), peptide linked to cell-penetrating peptides (CPPs), and cell-penetrating homing peptides (CPHPs). HPs functions by delivering cargo (i.e. drugs) on to the cell surface, while CPPs facilitates the cargo internalization, and CPHPs enables cargo internalization without support form an external agent (Svensen et al., 2012). Herein, we have focused on a tumor homing peptide (THP), which is a short sequence of peptides with the specific ability to recognize and home to (i.e. reach out) to tumor cells or tumor vasculature (Laakkonen and Vuorinen, 2010).

Since the discovery of in vivo phage display technology and in vitro screening of synthetic peptide libraries, a large number of THPs have been identified (Svensen et al., 2012). The first generation of THPs contain common motifs like RGD (Arg-Gly-Asp) and NGR (Asn-Gly-Arg) (Laakkonen and Vuorinen, 2010). The RGD peptide binds specifically to α-integrin receptors, while the NGR-containing peptides bind to aminopeptidase N receptors, which are expressed in tumor vasculature (Pasqualini et al., 1997, 2000). Because of their short length, usually between 3–15 amino acids, these peptides are more favorable to travel across the body than other anti-tumor drugs as show in Fig. 1. In addition, there is a vast literature stating that THPs exhibited well-defined specificity towards receptors on target organs which may be part of the tumor vasculature or lymphatic system. Furthermore, these peptides are relatively non-immunogenic and generate low production costs as compared to other therapeutic vehicles (Svensen et al., 2012; Laakkonen and Vuorinen, 2010; Gautam et al., 2014).

Although experimental approach is an objective method to identify THPs, it is time-consuming. Particularly, with a huge number of therapeutic peptides generated in recent years, it is highly desirable to develop a computational model to discriminate THPs from non-THPs.

* Corresponding authors.
E-mail addresses: watshara.sho@mahidol.ac.th (W. Shoombuatong), chanin.nan@mahidol.edu (C. Nantasenamat).

https://doi.org/10.1016/j.compbiolchem.2019.05.008
Received 6 December 2018; Received in revised form 18 April 2019; Accepted 17 May 2019
Available online 24 May 2019
1476-9271/ © 2019 Elsevier Ltd. All rights reserved.

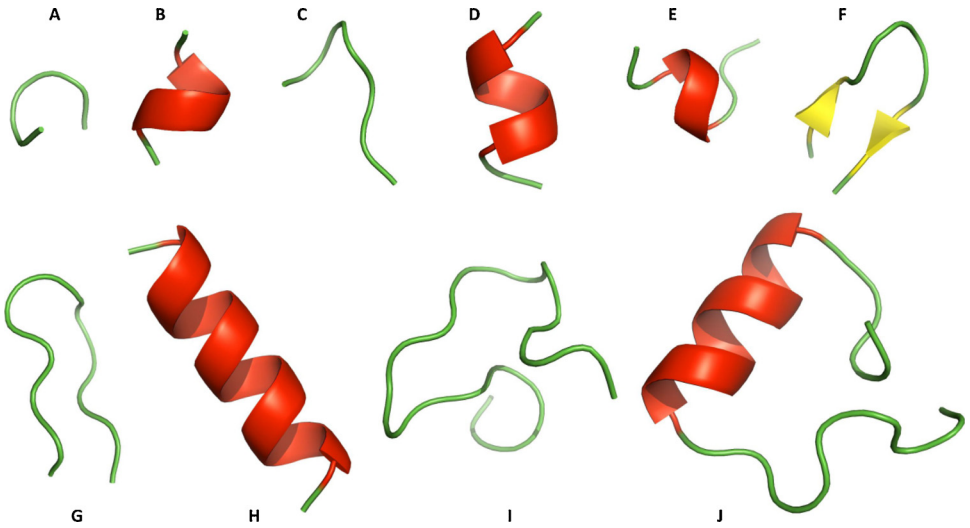

Fig. 1. Three-dimensional structure of THPs. Panels A–F are THPs with 5–10 amino acid residues while panels G–J are THPs larger than 10 amino acid residues. Peptide structures were constructed by PEP-FOLD server (http://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-FOLD3/) and visualized in PyMOL version 1.7.6.3 (https://pymol.org/).

In 2012, a database of tumor homing peptide (THP) namely TumorHoPe developed by Kapoor et al., stored comprehensive information of THPs to aid in designing peptide-based drugs and drug-delivery systems (Kapoor et al., 2012). To the best of our knowledge, there is only one computational model developed in this regard. (Sharma et al. (2013)) were the first to propose a sequence-based predictor called TumorHPD using support vector machine (SVM) as a prediction model and amino acid composition, dipeptide composition and binary profile patterns as feature input for discriminating THPs from non-THPs in peptide sequences. TumorHPD was able to yield a high accuracy of 86.56%. The aforementioned method has its own merit and played a major role in stimulating the development of this area although, bearing few limitations. Firstly, the peptide sequences in their datasets shared high-sequence similarity. It has been demonstrated that their datasets would lack statistical representations. Secondly, TumorHPD was trained and tested with such a dataset might yield overestimated prediction results. Lastly, the SVM model is well-recognized as a black-box approach that does not allow itself to interpretability, but the characterization and analysis of THPs are important for both basic research and drug development.

Given the aforementioned limitations of the existing method, it is desirable to develop an effective and interpretable a sequence-based model for discriminating THPs from non-THPs. Firstly, we constructed an objective and strict benchmark dataset by excluding peptides having a sequence identity of > 90%. Secondly, a prediction method named THPep was constructed by using an interpretable random forests (RF) method cooperated with amino acid composition, dipeptide composition and pseudo amino acid composition. This study utilizes the RF method because of its built-in ability of feature importance estimation. Thirdly, we estimated informative amino acids and dipeptides to provide insights of the biological implications of THPs. The prediction results over both 5-fold cross-validation and an independent dataset demonstrated that the proposed method THPep was superior to the existing method. Finally, a free web server was built to provide an efficient and useful tool for THP prediction.

2. Materials and methods

As elaborated by a series of publications from our group (Simeon et al., 2016, 2017; Pratiwi et al., 2017; Win et al., 2017; Shoombuatong et al., 2018a, b), to develop a useful and interpretable sequence-based predictor, the following five procedures should be considered: (i) establish or select a reliable benchmark dataset for training and validating the predictor and also calculating the dataset modelability as to evaluate the feasibility of obtaining a predictor with significant predictive power; (ii) represent protein or peptides sequences that can truly reflect their intrinsic properties to be predicted; (iii) introduce or develop an easily interpretable predictor that can provide a better understanding of the biological system; (iv) evaluate the predictive power of the predictor by performing rigorous cross-validation methods; (v) develop a user-friendly web-server that can easily reveal the desired results without needing to go through the mathematical and statistical details. Fig. 2 presents the workflow of the systems used to predict and analyze THPs, which involves dataset compilation, feature extraction, data splitting, model construction and evaluation and model post-analysis.

2.1. Benchmark datasets

A high quality benchmark dataset that was obtained from only experimentally verified peptides can guarantee the reliability and accuracy of a predictor. In this study, the benchmark datasets, i.e. Main (S_{Main}) and Small (S_{Small}) datasets, were derived from Sharma et al. (Sharma et al., 2013) to examine a predictor for its effectiveness in practical THP prediction as well as to fairly compare it with the existing method. The benchmark datasets S_{Main} and S_{Small} can be formulated as follows:

$$S_{Main} = S^{+}_{Main} \bigcup S^{-}_{Main} \quad (1)$$

$$S_{Small} = S^{+}_{Small} \bigcup S^{-}_{Small} \quad (2)$$

where S^{+}, S^{-} and the symbol $\bigcup$ represent the positive or THPs subset, the negative or Non-THPs subset and the union in the set theory, respectively. The subsets S^{+}_{Main} and S^{-}_{Main} contain 651 THPs and 651 Non-THPs, respectively, while the subsets S^{+}_{Small} and S^{-}_{Small} contain 469 THPs and 469 Non-THPs, respectively.

However, the peptides in Main (S_{Main}) and Small (S_{Small}) datasets share high-sequence similarity leading to a lack of statistical representation. As demonstrated in (Chou (2011)), a predictor trained and tested with such bias benchmark dataset, may yield misleading prediction results with overestimated accuracy. To remedy this problem in the proposed predictor, we constructed a new dataset, called Main90 dataset (S_{Main90}), by excluding peptides from the S_{Main} dataset with a sequence identity of 90% or above using CD-HIT (Huang et al., 2010). After such a procedure, the remaining dataset contained 176 THPs and 443 Non-THPs. The S_{Main90} dataset that can guarantee the reliability of the prediction model can be formulated as

$$S_{Main90} = S^{+}_{Main90} \bigcup S^{-}_{Main90} \quad (3)$$

where the subsets S^{+}_{Main90} and S^{-}_{Main90} contain 176 THPs and 443 Non-THPs, respectively. But, in the case of the S_{Small} dataset, we did not

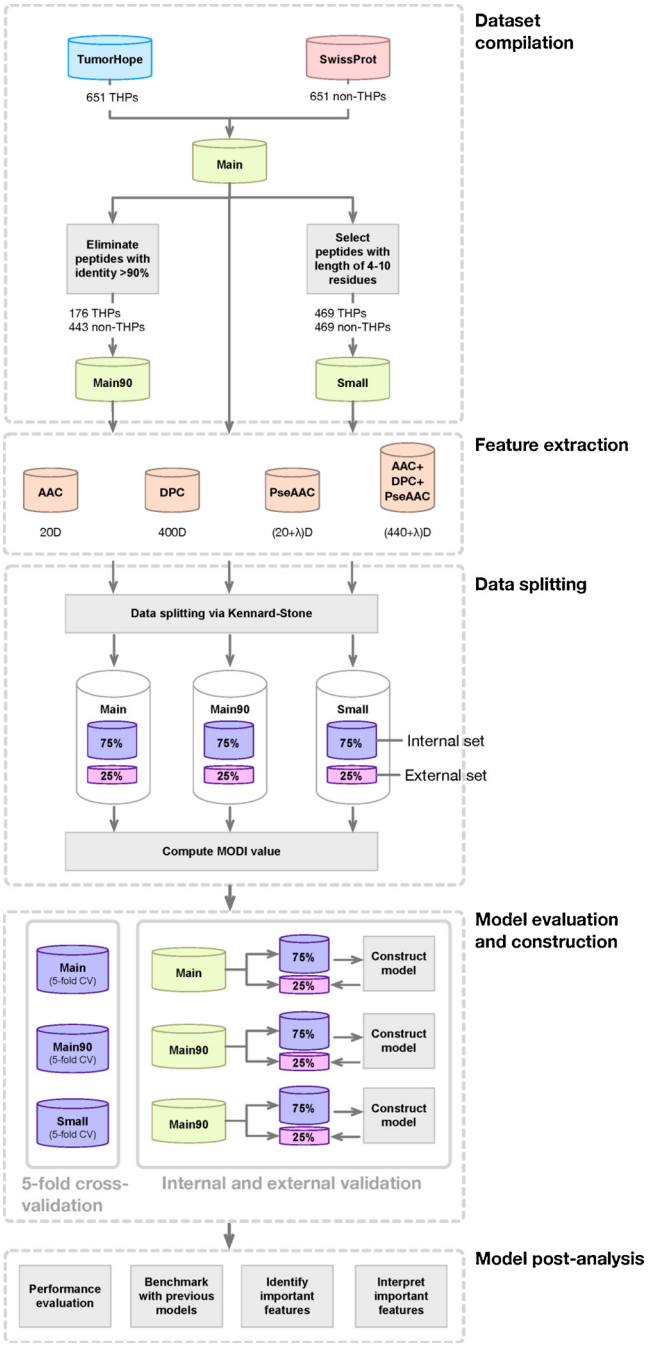

Fig. 2. Schematic illustration of the workflow for prediction and analysis of THPs.

Table 1
Summary of three benchmark datasets for evaluating the predictors of THPs. These benchmarked datasets were obtained from Sharma et al. (2013) (Sharma et al., 2013).

Dataset	Peptide identity	Whole dataset		Internal set		External set	
		THP	Non-THP	THP	Non-THP	THP	Non-THP
Main	100	651	651	490	490	161	161
Small	100	469	469	350	350	119	119
Main90	90	176	443	132	332	44	111

remove the homologous peptides because the short length of the peptides inside this dataset which may lead to a loss of information for establishing an efficient and effective predictor. Table 1 lists the distribution of THPs and non-THPs on all three benchmark datasets.

2.2. Feature representation

After obtaining the benchmark datasets, the next step is to represent each peptide into numerical vectors that can considerably depict the perspective of its biological and chemical properties. For peptide or protein sequences, the most classic and interpretable features are amino acid composition (AAC) and dipeptide composition (DPC). Until now, AAC and DPC have been applied in the investigation of many peptides and proteins, such as predicting HIV-1 CRF01-AE co-receptor usage (Shoombuatong et al., 2012), predicting protein crystallization (Shoombuatong et al., 2013; Charoenkwan et al., 2013), predicting the oligomeric states of fluorescent proteins (Simeon et al., 2016), predicting the bioactivity of host defense peptides (Simeon et al., 2017), predicting antifreeze proteins (Pratiwi et al., 2017), and predicting the hemolytic activity of peptides (Win et al., 2017).

According to the classical definition, the AAC, DPC and TPC features are expressed as a fixed length of 20, 400 and 8000 with each representing the frequency of one of 20, 400 and 8000 native amino acids, dipeptides and tripeptides, respectively, in the peptide sequence **P**. Thus, in the terms of AAC, DPC and TC features, a peptide **P** can be expressed by vectors with 20D and 400D (dimension) spaces, respectively as formulated by:

$$\mathbf{P} = [aa_1, aa_2, ..., aa_{20}]^T \quad (4)$$

$$\mathbf{P} = [dp_1, dp_2, ..., dp_{400}]^T \quad (5)$$

where T is the transpose operator, while aa_1, aa_2, …, aa_{20} and dp_1, dp_2, …, dp_{400} are occurrence frequencies of the 20 and 400 native amino acids and dipeptides, respectively, in a peptide sequence **P**.

As shown in Eq. 4 and 5, AAC and DPC features only provide the 20D AA and 400D DP compositions to represent a peptide **P**, but such three vectors defined by the concept of AAC and DPC may completely lose the sequence-order information. To deal with such dilemma, the pseudo amino acid composition (PseAAC) approach was proposed. According to Chou's PseAAC (Shoombuatong et al., 2018a), the general form of PseAAC for a peptide **P** is formulated by:

$$\mathbf{P} = [\Psi_1, \Psi_2, ..., \Psi_\Omega]^T \quad (6)$$

where the subscript Ω is an integer and its value as well as the components Ψ_1, Ψ_2, … will depend on the method used to extract the desired information from the amino acid sequence of a peptide **P**. Below, we describe the method used to generate the PseAAC feature from the three benchmark datasets to define the peptide sequences via Eq. 7.

2.3. Dataset modelability

One of the important factors for estimating the prediction performance of a QSAR model is the modelability of the dataset. In 2004, Golbraikh et al. introduced (Golbraikh et al., 2014; Fourches et al., 2016) the “Modelability Index” (MODI) as a simple and quick method to estimate the feasibility in developing robust and predictive QSAR model. The key idea behind this metric is that a pair of peptides having similar sequence patterns or characteristics are deemed to belong to the same biological class. In this study, the MODI score for each type of peptide features were calculated using an in-house developed R code. A threshold value of 0.65 separates the dataset into modelable or non-modelable classes. The steps involved in the calculation of the MODI score are described as follows:

Step 1: For a given dataset, the normalized Euclidean distance D'_{ij} between a pair of peptides M_i and M_j is computed as follows:

$$D_{ij} = \|\mathrm{M}_i - \mathrm{M}_j\| = \sqrt{\sum_{k=1}^{m} (M_{ik} - M_{jk})^2} \tag{7}$$

$$\bar{D}_i = \frac{\sum_{j=1}^{n} D_{ij}}{n-1} \tag{8}$$

$$D'_{ij} = \frac{D_{ij} - \min(\bar{D}_i)}{\max(\bar{D}_i) - \min(\bar{D}_i)} \tag{9}$$

where D_{ij}, $\bar{D}_i$ and n represents the distance scores between two peptides, the mean Euclidean distance and the number of peptides, respectively.

Step 2: For each peptide in a dataset, the MODI can be computed by identifying its first nearest neighbor (i.e. a peptide with the smallest Euclidean distance) belonging to the same or different class as follows:

$$MODI = \frac{1}{N_C} \sum_{l=1}^{N_C} \frac{N_i^{same}}{N_i^{total}} \tag{10}$$

where N_C is the number of classes (i.e. $N_C = 2$ denotes THP and non-THP), N_i^{same} is the number of peptides of the i^{th} class that have their first nearest neighbors belonging to the same i^{th} class and N_i^{total} is the number of peptides belonging to the i^{th} class. A dataset is considered to be modelable if the MODI index is greater than the threshold value of 0.65.

2.4. Random forest

(Breiman (2001)) proposed the random forest (RF) algorithm, which is an ensemble of tree-structured classifiers obtained by growing many weak classification and regression trees (CART) (Breiman et al., 1984) for enhancing the prediction performance. RF model takes advantage of two powerful machine-learning techniques: bagging (Breiman, 1996) and random feature selection. Growing many classification trees and using randomness in building RF model, yields a better prediction performance and enhances robustness of the model in a dataset. Each classification tree can be grown according to the following rules: (1) in bagging, each classification tree is trained on a bootstrap sample, which is a subset of the whole training N peptide. Peptides not included in the bootstrap sample are placed in the out-of-bag (OOB) set of M peptides, where $M \approx N/3$, which is then used to estimate the effectiveness of the RF model; (2) for each node of a classification tree, the RF model randomly selects *mtry* features and uses them to determine the best possible split with the Gini index as the splitting criteria; (3) each classification tree is grown to the largest extent possible i.e. no pruning is conducted. Herein, the RF classifier was established using the *randomForest* R package (Breiman, 2006). To enhance the performance of the RF model, two parameters namely *ntree* (i.e. the number of trees used for constructing the RF classifier) and *mtry* (i.e. the number of random candidate features) were subjected to optimization using the *caret* R package (Kuhn et al., 2017). The search for optimal values regarding the two parameters were carried out according to the following:

$$1 \leq mtry \leq 5\,,\ step\ \ \Delta = 1 \tag{11}$$

2.5. Feature importance

The RF method is useful for evaluating feature importance with the use of the OOB set. In order to provide a better understanding of the tumor homing activity of peptides, herein, the RF models were utilized for the identification of informative features for each type of peptide sequential feature. This is due to its efficient and effective built-in feature importance estimator. Generally, in the system of RF model construction, two measures based on the mean decrease of the Gini index (MDGI) and prediction accuracy are available for ranking feature importance. Considering the recommendations of (Calle and Urrea (2011)) that MDGI is more robust than the mean decrease of the prediction accuracy, we thus utilize the value of MDGI to rank the importance of each type of peptide sequential features. In the system of ranking feature importance, we constructed 10 RF models by fixing the *mtree* parameter to 100 and varying the *mtry* parameter setting from 2 to 20 ($mtry \in \{2,3,\ 5,7,\ 9,11,13,15,17,19,20\}$). The construction of multiple RF models for estimating important features can reveal potential bias of using a single value of *mtry* and thereby increases the reliability of the estimated informative features. Features with the largest MDGI value is considered to be the most important feature as it contributes the most to the prediction performance of the model.

2.6. Performance evaluation

In a statistical prediction, three cross-validation methods, namely sub-sampling test (2-, 5- or 10-fold cross-validation), jackknife test and independent (or external) testing dataset are often used to evaluate the prediction performance. Among the three aforementioned cross-validation methods, the jackknife test can yield a unique result for a given benchmark dataset, however, it is time-consuming. Thus, in this study, the 5-fold cross-validation (5-fold CV) and external testing dataset were used to evaluate the prediction performance of our models. The 5-fold CV approach was performed to benchmark and used to compared with the existing method (Sharma et al., 2013), while the external testing dataset was used to ensure that no overlap existed between the training and testing sets. Thus, it could be concluded that the predictor having high accuracy over the external testing dataset determines its high predictive ability on an unknown dataset. The procedure of external testing dataset is briefly described as follows. Firstly, the internal and external sets with the ratio of 3:1 were constructed using the Kennard-Stone algorithm via the R package *prospectr* (Stevens et al., 2015). Subsequently, the internal set was used to estimate the parameter of the model via a 5-fold CV and further, a prediction model with the optimum parameter was constructed to determine the class label on the external set.

In order to evaluate the prediction ability of the model, the following sets of four metrics are used as follows:

$$Ac = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{12}$$

$$Sn = \frac{TP}{(TP + FN)} \tag{13}$$

$$Sp = \frac{TN}{(TN + FP)} \tag{14}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{15}$$

where *Ac*, *Sn*, *Sp* and MCC are called accuracy, sensitivity, specificity and Matthews coefficient correlation, respectively. *TP*, *TN*, *FP*, and *FN* represent the instances of true positive, true negative, false positive and false negative, respectively. Moreover, in order to evaluate the prediction performance of models using threshold-independent parameters, the receiver operating characteristic (ROC) curves were plotted by the *pROC* package in the R software (Robin et al., 2011). The area under the ROC curve (auAUC) was used to measure the prediction performance, where AUC values of 0.5 and 1 are indicative of perfect and random models, respectively.

2.7. Physicochemical properties

When generating the PseAAC feature of a peptide, the following parameter set was considered: (1) the PseAAC mode; (2) the physicochemical properties (PCPs) used; (3) the two parameters (The lambda parameter λ and the weighting factor w, where the parameter λ represents the global or long-range sequence order effect, while the parameter w is the weight factor of a long-range effect usually within

the range of 0–1.

The parameter optimization procedure of PseAAC feature is described as follows. Firstly, the PseAAC model was set by using the default. Secondly, we selected the informative PCPs set from 531 PCPs (Kawashima et al., 2007) using the values of MDGI. During the determination of optimal subset of PCPs procedure, the RF model was constructed using the 5,10,15,20,… top-ranked informative PCPs and examined with 5-fold CV on the three benchmark datasets S_{Main}, S_{Small} and S_{Main90}. Finally, after obtaining the optimal PCPs subset, adjustments were made to the two parameters (λ and w) in order to achieve the best prediction accuracy. The search for the optimal values of the two parameters were carried out according to the following:

$$0.1 \le w \le 5\,,\ step\ \ \Delta = 0.1 \tag{16}$$

According to Eq. 16, total of 10 × 10 = 100 individual combinations should be investigated for determining the optimal parameter combination. Herein, the 5-fold CV approach was used to assess the performance for each parameter combination in the process of parameter optimization.

2.8. Reproducible research

To ensure the reproducibility of the models proposed herein, all R codes and the three benchmark datasets used in the construction of the predictive models, graphical figures and the THPep web server are available on GitHub at https://github.com/shoombuatong2527/thpep.

3. Results

3.1. Parameter determination

As mentioned above, the PseAAC feature depends on the PCPs and the two parameters (λ and w). At this step, the RF model was constructed using the 5, 10, 15, …, top-ranked informative PCPs and assessed by the 5-fold CV approach. The feature set with the highest prediction performance (Ac and MCC) were considered. Table S1 shows the performance comparisons among different subsets of informative PCPs on the three benchmark datasets. As noticed in Table S1, the 15, 30 and 25 top-ranked informative PCPs for S_{Main}, S_{Small} and S_{Main90} datasets showed the highest value of Ac/MCC of 84.56%/0.69, 81.77%/0.64 and 88.51%/0.71, respectively. Additional details of the informative PCPs on the three benchmark datasets are summarized in Table S2. After obtaining the optimal subset of informative PCPs, the two parameters of λ and w were estimated using the grid search within the range as mentioned in Eq. 16. The parameter combinations having the lowest prediction error was used to construct the THP predictor. In the present study, the optimal combination parameters λ and w for S_{Main}, S_{Small} and S_{Main90} datasets were 2/0.1, 3/0.6 and 1/0.5, respectively, in the following predictor development.

3.2. Prediction performance

In this study, we made an effort to develop a THP predictor that utilizes the RF model as the prediction model and AAC, PseAAC, DPC and all combination features, i.e. the combination of AAC + PseAAC, AAC + DPC, DPC + PseAAC and AAC + PseAAC + DPC, as the input features. In order to determine the features that are beneficial to the prediction of THPs, RF models cooperating with the different features were performed on the three-benchmark datasets via 5-fold CV. Moreover, in order to avoid the influence of redundancy, the proposed predictor was trained and tested with the S_{Main90} dataset having a sequence identity of 90% or less. Furthermore, in order to describe the predictable capability of our method on unknown samples, data splitting was performed for a 100 independent iterations on the three benchmarked datasets. The number of peptides was then used to create the models which were further evaluated using two cross-validation methods as shown in Table 1.

Table 2
Summary of MODI index as derived from various types of peptide features on both internal and external sets.

Feature	Internal set			External set		
	Main	Small	Main90	Main	Small	Main90
AAC	0.71	0.71	0.76	0.72	0.73	0.70
PseAAC	0.75	0.72	0.74	0.79	0.70	0.73
DPC	0.66	0.69	0.68	0.62	0.64	0.74
AAC + PseAAC	0.75	0.72	0.74	0.79	0.70	0.74
AAC + DPC	0.70	0.71	0.75	0.68	0.68	0.77
PseAAC + DPC	0.75	0.72	0.74	0.79	0.69	0.74
AAC + PseAAC + DPC	0.75	0.72	0.74	0.79	0.70	0.74

Before construction of the predictive model, the modelability of the dataset was assessed by computing the MODI index. This metric helps to estimate the feasibility of obtaining robust and reliable predictive model by using the three types of peptide features and their possible combinations for both internal and external sets. For binary classification problem, if the value of MODI index is greater than 0.65, the dataset is considered to be reliable for classification modelling. As seen in Table 2, most features from internal and external sets met these criteria with the exception of models built using DPC features on an external set.

Table 3–4 and Fig. 3 list the results of performance comparisons between different peptide features as evaluated by 5-fold CV and external testing dataset, respectively. Moreover, Figs. 3 shows the ROC curves of the two cross-validation methods. As noticed in Table 3, the combination feature of AAC + PseAAC affords the highest prediction performance in terms of Ac (86.10% and 90.80%) and MCC (0.72 and 0.77) on S_{Main} and S_{Main90} datasets, respectively. Meanwhile, the AAC feature performed well with the second highest Ac of 85.71% and 89.66% on S_{Main} and S_{Main90} datasets, respectively. In the case of the S_{Small} dataset, the AAC feature reached a maximum Ac of 83.37%, while the combination features of AAC + DPC and AAC + PseAAC performed well with the second and third highest Ac of 83.26% and 82.94%, respectively.

To infer the true predictive power of the proposed THP predictor, performance comparisons as evaluated by external testing dataset were also performed as shown in Table 4 and Fig. 3. Interestingly, the results of the performance comparisons were well correlated with the results derived from the 5-fold CV method (Table 3). As seen in Table 4, the combination feature of AAC + PseAAC and AAC + DPC reached a maximum Ac of 85.28%/83.74%, 90.13%/89.47% and 77.78%/80.77% on the S_{Main}, S_{Main90} and S_{Small} datasets, respectively. Furthermore, the combination feature of AAC + PseAAC outperformed the three single peptide features with an improvement of greater than 4–13% on MCC under the external testing dataset.

The results of performance comparisons can be briefly summarized as follows. Each of the three single peptide features can be effectively used to discriminate THPs from non-THPs. The AAC feature was observed as the most beneficial to be develop as a predictor than the DPC and PseAAC features. In the case of the combination features, the AAC + PseAAC is the effective and efficient peptide feature that performed well on the three benchmark datasets over the two standard cross-validation methods. All of the prediction results were seen to be well related with the MODI values. For convenience of the subsequence description, we will refer to this method as THPep.

3.3. Comparison with the existing prediction method

It is necessary to compare the proposed method with the existing method. To the best of our knowledge, there is only one other study which focused on the prediction of tumor homing peptides (Sharma

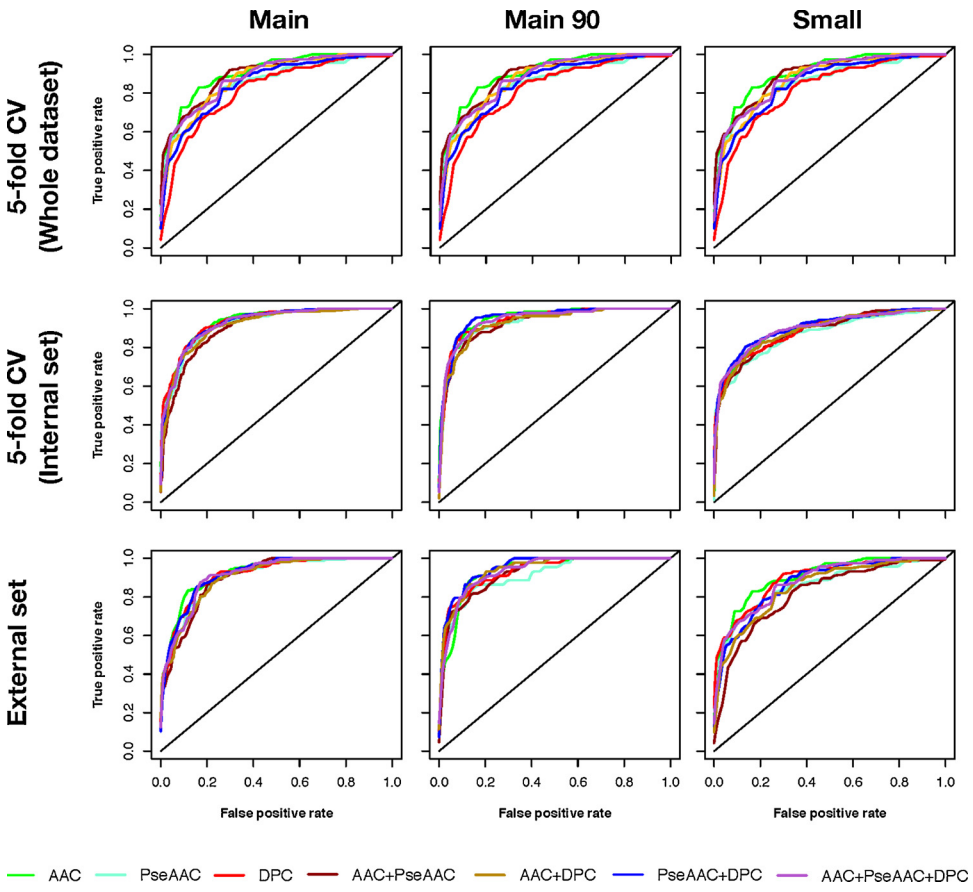

Fig. 3. ROC curve of RF models as assessed by 5-fold cross-validation. Models were evaluated via 5-fold CV on the whole dataset (fist row), 5-fold CV on the internal set (second row) and external test (third row). Performance of the three benchmarked datasets consisting of Main, Main 90 and Small are shown in the first, second and third columns, respectively.

Table 3
Performance comparison of RF models with various classes of protein sequences assesses by 5-fold cross-validation.

Dataset	Feature	Ac (%)	Sn (%)	Sp (%)	MCC	auAUC
Main	AAC	85.71	87.20	84.34	0.71	0.93
	PseAAC	84.56	84.72	84.40	0.69	0.92
	DPC	82.33	80.82	84.01	0.65	0.91
	AAC + PseAAC	86.10	87.07	85.18	0.72	0.93
	AAC + DPC	84.41	85.67	83.23	0.69	0.92
	PseAAC + DPC	84.72	84.98	84.45	0.69	0.92
	AAC + PseAAC + DPC	84.33	85.09	83.61	0.69	0.93
Small	AAC	83.37	81.24	85.81	0.67	0.91
	PseAAC	81.77	78.88	85.31	0.64	0.89
	DPC	80.06	80.00	80.13	0.60	0.87
	AAC + PseAAC	82.94	80.00	86.52	0.66	0.91
	AAC + DPC	83.26	80.95	85.94	0.67	0.91
	PseAAC + DPC	82.52	78.72	87.47	0.66	0.89
	AAC + PseAAC + DPC	82.62	79.54	86.43	0.66	0.90
Main90	AAC	89.66	90.22	87.92	0.74	0.95
	PseAAC	88.51	90.24	83.54	0.71	0.93
	DPC	86.86	87.79	83.80	0.67	0.92
	AAC + PseAAC	90.80	91.80	87.97	0.77	0.95
	AAC + DPC	88.51	88.05	90.15	0.71	0.95
	PseAAC + DPC	87.19	86.15	91.53	0.68	0.94
	AAC + PseAAC + DPC	88.34	88.03	89.47	0.71	0.95

et al., 2013). Thus, the prediction results between THPep and the existing method (TumorHPD (Sharma et al., 2013)) was compared. In 2013, TumorHPD was developed using a computational model based on the SVM method cooperating with AAC feature on the NTCT10 dataset. For facilitating comparison, the same benchmark datasets (S_{Main} and S_{Small}) and cross-validation method (5-fold CV) were employed. The results of the performance comparison between THPep and TumorHPD are listed in Table 5.

The reported results of TumorHPD as shown in Table 5 come directly from (Sharma et al., 2013). Since we are the first to perform a prediction model using a non-redundant dataset S_{Main90}, Table 5 does not show the prediction result of TumorHPD on this dataset. As noticed in Table 5, THPep outperformed TumorHPD with improvements of > 6-8%, > 2% and 2–3% for Sn, MCC and auAUC, respectively, as evaluated by the 5-fold CV approach. Although the Ac and Sp obtained from THPep were little lower than TumorHPD as evaluated by 5-fold CV approach, the MCC and auAUC are the most reliable statistic measures and are less influenced by an imbalanced dataset (Ganguly and Sadaoui, 2017). Considering that the external testing dataset is the most rigorous cross-validation method and a high value of MCC and auAUC are preferable, it could be stated that THPep is not only more effective than TumorHPD, but also more stable.

To further verify the power of THPep, some conventional classifiers such as *k*-nearest neighbor (*k*-NN), decision tree (DT), artificial neural network (ANN) and support vector machine (SVM) were implemented for performance comparisons. Such classifiers were tested on the three benchmark datasets and implemented using R programming (Gentleman, 2008). Table 6 lists the performance comparisons of THPep with the conventional classifiers based on the combination features of AAC and PseAAC. By observing the results listed in Table 6, we can clearly demonstrate that THPep outperforms all the four classifiers by achieving the highest Ac and MCC of 90.80 and 0.77, respectively as evaluated by 5-fold CV. Moreover, these results also indicated that the combination feature of AAC and PseAAC (83.79% Ac) are more effective than the proposed features in the work of (Sharma et al. (2013)) (82.52% Ac). The comparison results above leave no doubt that, THPep, as proposed in this study is quite promising and holds potential to become a useful tool for discriminating THPs from non-THPs. Furthermore, the proposed method, THPep, can play a complementary role to the existing method in the same area.

Table 4
Performance comparison of RF models with various types of protein sequences assessed by the 5-fold cross-validation and external sets.

Dataset	Feature	5-fold CV test					External testing dataset				
		Ac (%)	Sn (%)	Sp (%)	MCC	auAUC	Ac (%)	Sn (%)	Sp (%)	MCC	auAUC
Main	AAC	84.53	85.47	83.63	0.69	0.92	83.13	82.21	84.05	0.66	0.91
	PseAAC	83.50	82.63	84.42	0.67	0.92	83.44	82.82	84.05	0.67	0.91
	DPC	80.84	77.51	85.08	0.62	0.91	78.83	82.82	74.85	0.58	0.89
	AAC + PseAAC	84.63	86.11	83.27	0.69	0.93	85.28	82.82	87.73	0.71	0.92
	AAC + DPC	85.25	85.68	84.82	0.70	0.92	83.74	78.53	88.96	0.68	0.91
	PseAAC + DPC	83.81	83.13	84.52	0.68	0.92	82.52	82.21	82.82	0.65	0.90
	AAC + PseAAC + DPC	84.84	85.86	83.86	0.70	0.92	84.05	79.75	88.34	0.68	0.92
Small	AAC	81.68	80.55	82.89	0.63	0.91	78.63	83.76	73.50	0.58	0.87
	PseAAC	78.27	75.84	81.19	0.57	0.87	74.36	76.07	72.65	0.49	0.82
	DPC	79.69	78.17	81.38	0.59	0.87	72.22	75.21	69.23	0.45	0.82
	AAC + PseAAC	80.68	78.88	82.73	0.61	0.89	77.78	85.47	70.09	0.56	0.85
	AAC + DPC	80.40	78.16	83.02	0.61	0.90	80.77	83.76	77.78	0.62	0.87
	PseAAC + DPC	79.69	75.30	85.91	0.60	0.88	77.35	82.05	72.65	0.55	0.84
	AAC + PseAAC + DPC	80.11	77.60	83.12	0.60	0.89	78.21	81.20	75.21	0.57	0.85
Main90	AAC	87.53	88.51	84.40	0.69	0.94	88.82	93.52	77.27	0.72	0.95
	PseAAC	87.96	89.94	82.35	0.70	0.93	84.87	88.89	75.00	0.63	0.91
	DPC	85.12	85.79	82.65	0.62	0.93	88.16	90.74	81.82	0.72	0.93
	AAC + PseAAC	89.72	90.88	86.32	0.74	0.94	90.13	94.44	79.55	0.76	0.93
	AAC + DPC	85.34	85.44	84.95	0.63	0.94	89.47	97.22	70.45	0.74	0.94
	PseAAC + DPC	84.46	83.96	86.75	0.60	0.93	88.16	96.30	68.18	0.70	0.94
	AAC + PseAAC + DPC	85.78	84.57	91.36	0.64	0.95	89.47	97.22	70.45	0.74	0.95

Table 5
Performance comparisons between THPep and TumorHPD over 5-fold cross-validation.

Dataset	Method	Ac (%)	Sn (%)	Sp (%)	MCC	AUC
Main	TumorHPD	86.56	80.63	89.71	0.70	0.91
	THPep	86.10	87.07	85.18	0.72	0.93
Small	TumorHPD	81.88	73.13	90.92	0.65	0.88
	THPep	83.37	81.24	85.81	0.67	0.91
Main90	TumorHPD	–	–	–	–	–
	THPep	90.80	91.8	87.97	0.77	0.95

Table 6
Performance comparisons between THPep and some conventional methods over 5-fold cross-validation.

Dataset	Method	Ac (%)	Sn (%)	Sp (%)	MCC
Main	*k*-NN	79.95	75.26	86.79	0.61
	DT	80.57	80.24	80.90	0.61
	ANN	81.87	81.30	82.47	0.64
	SVM	83.79	83.33	84.27	0.68
	THPep	86.10	87.07	85.18	0.72
Small	*k*-NN	79.00	82.54	76.15	0.58
	DT	76.97	77.56	76.41	0.54
	ANN	77.08	79.53	75.00	0.54
	SVM	79.10	81.52	77.03	0.58
	THPep	83.37	81.24	85.81	0.67
Main90	*k*-NN	81.94	67.93	88.00	0.57
	DT	84.40	73.68	88.58	0.62
	ANN	85.88	77.11	89.16	0.65
	SVM	86.54	85.07	86.95	0.66
	THPep	90.80	91.8	87.97	0.77

3.4. Mechanistic interpretation of feature importance

To the best of our knowledge, this study represents the first systematic effort for the selection and characterization of relevant features for THPs prediction. The characterization and analysis of feature importance can provide a unique understanding of THPs. The ability of THPs to recognize and bind with tumor cell receptors and tumor vasculature is considered their most important attribute. In order to function as a homing peptide, the presence of certain amino acid residues are required. In that regard, the existence of certain key residues in the sequence of THPs and their associated functions that determine activity will be discussed herein. Moreover, in this study, simple and easily interpretable features consisting of AAC and DPC were used. The efficient built-in feature importance estimator of the RF method was utilized to determine informative features for AAC and DPC. The value of MDGI was used to rank feature importance. The most important features are represented with the largest value of MDGI. The feature importance for AAC and DPC is shown in Fig. 4. The top three informative amino acids were shown to be Cysteine (Cys or C), Tryptophan (Trp or W) and Arginine (Arg or R) having MDGI values of 139.48, 46.56 and 45.40, respectively, while RC, GR, CR, and CG were the top four informative dipeptides having MDGI values of 9.22, 8.74, 8.41 and 7.86, respectively. Therefore, it can be inferred that peptides containing such amino acids and dipeptides have a higher chance of possessing THP activity. Interestingly, our results are consistent with the observations made by Sharma et al., where C, R, and W, were among the most abundant residues in THPs and are considered as the preferred residues present at the first position of a peptide sequence (Sharma et al., 2013).

3.5. THPep web server

To enable application of the proposed method, a user-friendly and freely accessible web server called THPep (i.e. based on the S_{Main} dataset) was created using the Shiny package under the R programming environment. As mentioned above, the best prediction performance was afforded by the RF model that makes use of the combination of AAC + PseAAC features. However, in practice, the length of this combination feature is a 420 + λ-dimensional vector, thereby resulting in a time-consuming procedure for both the construction of the prediction model and the prediction of the desired peptide sequence. Thus, selected AAC feature with 20-dimensional vector, which were found to be comparable with the aforementioned feature, was used for the construction of the THPep web server. Screenshots of the THPep web server is shown in Fig. 5. Furthermore, a step-by-step guide on how to use the THPep web server is given as follows:

Step 1. Browse to the THPep web server at http://codes.bio/THPep/ and you will see the top page of THPep on your computer screen as shown in Fig. 5(A).

Step 2. Either type or copy/paste the query peptide sequence into the Input box or upload the file of peptide sequences by clicking on the

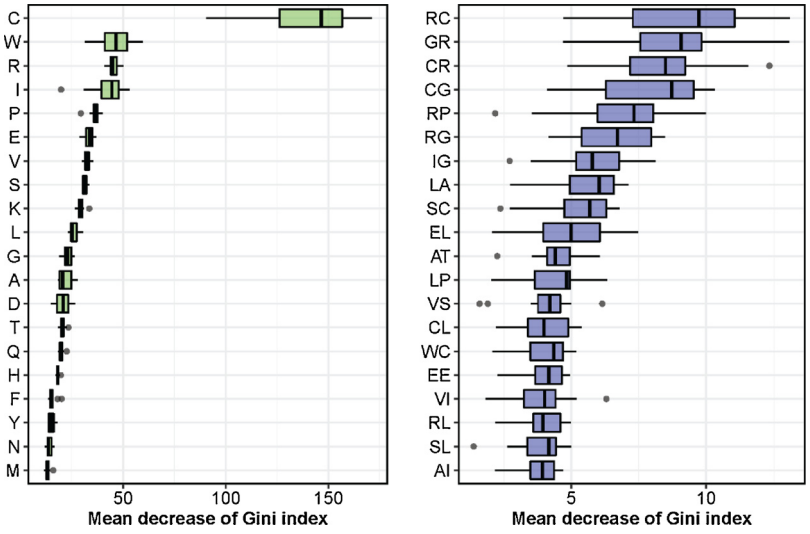

Fig. 4. Feature importance as deduced from the mean decrease of Gini index. Amino acid and dipeptide compositions are shown in the left and right panels, respectively.

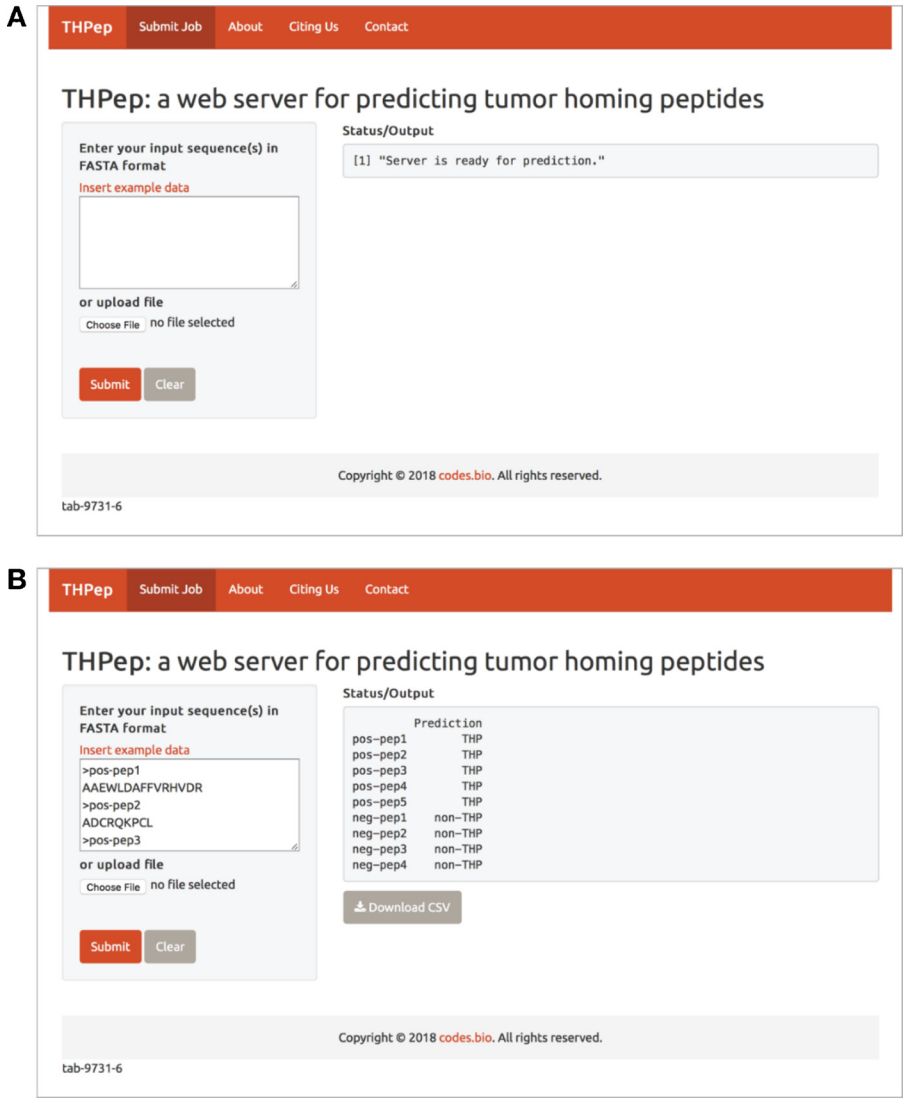

Fig. 5. Screenshot of the THPep web server. Before (A) and after (B) the initiation of the prediction process is shown.

Choose file button. The input peptide sequence should be in the FASTA format. Finally, press on the *Submit* button to initiate the prediction process.

Step 3. Click on the *Submit* button to initiate the modeling process and retrieve the prediction results, as shown in Fig. 5(B). Users can also download the prediction results as a CSV file by pressing on the *Download CSV* button.

Additionally, to maximize convenience, users could also build the THPep web server on their own computer by using the following code in an R environment:

shiny::runGitHub('THPep', 'chaninlab', subdir = "THPep_shiny _server")

Before running the aforementioned code, users must install prerequisite R packages by using the following code:

install.packages(c('shiny', 'shinyjs', 'shinythemes', 'protr', 'seqinr', 'randomForest', '
markdown'))

4. Discussion

The capability of current cancer therapies are limited by their ability in discriminating between tumor and healthy tissues. To solve such limitations, THPs have been developed for navigating therapeutic agents towards specific targets i.e. tumour tissues. Since the identification and development of new THPs using conventional experimental methods are labor intensive and time-consuming, computational models are the need of the hour for helping experimental scientists by offering analysis results. In the past decades, most model developers utilize complex computational models such as SVM, to predict many proteins and peptides, however those prediction models lacked mechanism prediction and interpretation. Rather than only focusing on the improvement of prediction results, model developers should also be concerned with the ability of their models which can help users in understanding and rationalizing their data of interest. In 2012, Huang et al. introduced a scoring card method (SCM) to predict protein solubility (Huang et al., 2012). For the prediction system, the SCM model used only a single threshold value to discriminate between soluble and insoluble proteins. In this work, they mentioned that an interpretable model with a comparable accuracy was more applicable. Most recently, Shoombuatong et al. provoked the importance and privilege of using interpretable models for researchers and scientists (Shoombuatong et al., 2017). Therefore, the goal of this study is to develop a systematic effort for predicting and analyzing THPs.

Table 3–4 and Fig. 3 compare the quality of prediction performances of RF model using seven feature sets as assessed by 5-fold CV and external testing dataset: AAC, PseAAC, DPC, AAC + PseAAC, AAC + DPC, PseAAC + DP and AAC + PseAAC + DPC. The performance comparisons on S_{Main}, S_{Small} and S_{Main90} datasets showed that the combination set of AAC + PseAAC resulted in superior predictions. Our prediction results was also shown to be consistent with previous related studies (Pratiwi et al., 2017; Win et al., 2017; Ganguly and Sadaoui, 2017). For instance, (Liaw et al. (2013)) developed an antibody amyloidogenesis model by using RF cooperating with AAC, DPC and PCP. This study suggested that the combined usage of the three feature types might improve prediction performances due to the fact that the usage of multiple features can capture the pattern of dataset more than by using a single feature. From the prediction results on each of the three basic features, the simple AAC feature performed better than PseAAC and DPC and also provided comparable results with the optimal combination sets mentioned above. Previously, the AAC feature has been widely used in the prediction and analysis of various protein and peptide functions, such as predicting the hemolytic activity of peptides (Win et al., 2017) or antifreeze proteins (Pratiwi et al., 2017).

As aforementioned, the top three informative amino acids were shown to be Cysteine (Cys or C), Tryptophan (Trp or W) and Arginine (Arg or R). Owing to its unique ability to form disulfide bonds through sulfhydryl side chains, Cys plays an important role in the function of THPs due to their abundance in the sequences. It has been reported that the cyclic peptide with two disulfide bonds, such as CDCRGDCFC, selectively accumulates in α_v integrins receptors in tumor blood vessels (Pasqualini et al., 1997). Moreover, another cyclic homing peptide i.e. cCPGPEGAGC (PEGA) containing one disulfide bridge is known to concentrate in the breast tumor vasculature and as a result, increases the efficacy of its conjugated-drug (Myrberg et al., 2008). Furthermore, a major limitation to the therapeutic effects of THPs is their short half-life in circulation, caused by renal clearance. Hence, coupling the peptide with a carrier molecule provides an effective strategy for prolonging its half-life by limiting renal clearance. This effect was demonstated by Pang et al. whereby the authors added a free Cys residue in the cyclic iRGD tumor-targeting peptide (CRGDK/RGPD/EC) that provided an extension to the half-life of the peptide and improved its accumulation in tumors. The added Cys conjugated with serum albumin through a disulfide bond and thus, afforded a higher molecular weight to the peptide that was above the renal clearance limit (Pang et al., 2014).

Furthermore, Colombo et al. investigated the importance of Cys residues using both, cyclic (CNGRC) and linear (GNGRG) peptides containing the NGR (Asn-Gly-Arg) motif using molecular dynamic simulations. The primary features under investigation were the stabilization of peptide conformations and their ability to deliver anti-tumor compounds (i.e. TNF-α) to tumor vessels. However, the authors found that the linear peptide containing Gly residues (GNGRG) had the propensity to form β-turn thus, lowering tumor targeting efficiency (Colombo et al., 2002) (Huang et al., 2012). The propensity of the linear peptide to form turns may favour the NGR motif through intramolecular stabilizing interactions (i.e. hydrogen bonding) thus, allowing them to be recognized by receptors and mediate anti-tumor compounds to tumor cells, albeit with lower efficiency. This in turn highlights the critical importance of Cys disulfide bridge formation in the enhanced stabilization and increased tumor targeting efficieny. Cyclic CNGRC peptides are known to bind with aminopeptidase N receptors expressed on the surface of cancer cells. Cys residues on the peptide formed a disulfide bridge which allowed the exposure of the N-terminus to bind with the active site of receptors, while the C-terminus carries the cargo (drug or chemical tag) (Graziadio et al., 2016).

Trp represents an amino acid having a highly hydrophobic indole ring side chain. Trp is important for the activity of THPs mainly due to the fact that their receptors reside in the membrane of tumor cells. Although considered a rare aromatic residue, Trp is highly enriched in membrane proteins (Von Heijne, 2007). A molecular dynamic study revealed the important role of Trp as a response modulator of hydrophobic mismatch to stabilize protein membrane anchoring in the phospholipid membrane (Jesus and Allen, 2013). It is likely that THPs containing Trp will be able to attach in the membrane protein of tumor cells through similar mechanisms as membrane protein anchoring. In that regard, an engineered M13 filamentous phage used for screening proteins present in the cell cytoplasm known as iPhage, was recently developed. The CPP sequence (RQIKIWFQNRRMKWKK) was attached to the pVIII capsid of the phage to allow internalization of the iPhage. Intriguingly, total functional loss of the iPhage was observed with Ala residue replacements, suggesting that the two Trp residues of the peptide sequence are essential for the intracellular translocation (Rangel et al., 2013).

In addition, the presence of Arg in a peptide sequence with the unique ability of conducting tumor homing, has been well demonstrated in both the NGR motif discussed above as well as in the RGD (Arg-Gly-Asp) motif (Pasqualini et al., 1997). These two motifs represent the first-generation of tumor homing peptides that appear to be independent of tumor type, indicating that their receptors are unregulated during angiogenesis. The RGD peptides have a high affinity towards integrins $\alpha_v\beta_3$ and $\alpha_v\beta_5$ which are specifically over-expressed during angiogenesis and nearly absent in the normal tissue (Varner and

Cheresh, 1996). One of the problems associated with drug permeation into tumor tissue is that the blood flow through tumor vessels may be abnormal (Hambley and Hait, 2009). Furthermore, it is well known that Arg is primarily involved in the improvement of blood circulation in the body through its metabolism of nitric oxide (Scibior and Czeczot, 2004). Thus, it can be inferred that peptides containing Arg as part of its tumor homing motif, may benefit from this essential amino acid that has also been demonstrated to protect against the early stages of cancer (Geiger et al., 2016). In addition, novel RGD peptides have been adapted to include a tissue penetrating motif that binds to integrins and then upon proteolytic cleavage, enables binding to neuropilin-1 (NRP-1), thus triggering tissue penetration (Sugahara et al., 2009, 2010). Besides, various forms of RGD (e.g. RGD-4C, iRGD and cRGD) have been used to target arthritis and various cancers by coupling with apoptotic peptides and anti-cancer drugs (Sugahara et al., 2010; Elayadi et al., 2007; Gerlag et al., 2001). Furthermore, the first cell penetrating homing peptide to enter clinical trials, Cilengitide (Merck) is a salt of cRGD that acts as an anti-angiogenic agent and $\alpha_v\beta_3$ and $\alpha_v\beta_5$ integrin antagonist (Eskens et al., 2003; Fink et al., 2010; Gilbert et al., 2011).

Furthermore, (Lee et al. (2013)) identified peptides using bioinformatics that has recently progressed to the pre-clinical phase whereby the basement membrane of endothelial cells were explored via protein-protein interactions. Similarly, the peptide-peptide combination of AARP (Ala-Ala-Arg-Pro), coupled with an assembly of the metalloproteinase inhibitory peptides (MMP-2 and 9; CTTHWGFTLC) were conjugated to two different anti-angiogenic agents. Thus, the recombinant AARP-CTT comprises one single compound which was designed to block angiogenesis and inhibit MMP2 and -9 at the same time. This conjugate has been used to target various human solid tumors in mouse models and shown to inhibit angiogenesis, tumor growth, and metastasis, efficiently (Wang et al., 2014). Additionally, a novel, nine amino acid long peptide (CGLSGLGVA) known as CooP, for use against malignant brain tumour as well as in drug delivery, has been ascertained. CooP binds to the over-expressed cancer cell only, mammary-derived growth inhibitor and has further demonstrated exceptional homing peptide capabilities with its targeting of tumor-associated blood vessels in the brain (Hyvonen et al., 2014).

Due to their high specificity, low antigenicity, flexibility, and simple production, peptides represent a promising new avenue of possibilities for the development of anti-cancer therapeutic agents for various tumors. In the long term, such an approach can become personalized as the current use of THPs for drug delivery, such that novel mechanisms can be targeted which cannot otherwise be reached with traditional small molecules. However, while the identification of specific THPs is relatively fast, the identification of their target receptors represents a bottleneck of the system. As such, no universal approach exists for drug design although promising methods are reliant on the binding affinity between peptide and receptor. Since THPs are administered with the aim that they bind to their targets via circulation, efficient, reliable and generalize attachment needs to be developed.

5. Conclusions

Much evidence has been acquired to indicate the favourable nature of tumor homing peptides as promising candidates for cancer therapy. Thus, accurate identification and efficient characterization based on a computational model of tumor homing peptides are essential for understating their role in drug development as well as reducing time and costs. In this study, an interpretable and powerful sequence-based model termed THPep has been developed by using an interpretable random forest classifier and multiple peptide features. Rigorous cross-validation using both 5-fold cross-validation and independent testing dataset have indicated that the proposed THPep model is extremely powerful and promising. It is anticipated that THPep might become an important tool for both basic research and drug development.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors thank TS Win and V Prachayasittikul for helpful comments and suggestion on this manuscript. This work is supported by the TRF Research Grant for New Scholar (No. MRG6180226) and the TRF Research Career Development Grant (No. RSA6280075) from the Thailand Research Fund, the Office of Higher Education Commission and Mahidol University; the New Researcher Grant (A31/2561) from Mahidol University; and the Ph.D. scholarship from the Ministry of Research, Technology and Higher Education of the Republic of Indonesia.

References

N. I. o. Health, 2017. Cancer Statistics of National Cancer Institute.
W. H. Organization, 2017. World Health Organization Cancer Fact Sheet.
Mäe, M., Myrberg, H., El-Andaloussi, S., Langel, Ü., 2009. Design of a tumor homing cell-penetrating peptide for drug delivery. Int. J. Pept. Res. Ther. 15, 11–15.
Myrberg, H., Zhang, L., Mäe, M., Langel, Ü., 2008. Design of a tumor-homing cell-penetrating peptide. Bioconjug. Chem. 19, 70–75.
Svensen, N., Walton, J.G.A., Bradley, M., 2012. Peptides for cell-selective drug delivery. Trends Pharmacol. Sci. 33, 186–192.
Laakkonen, P., Vuorinen, K., 2010. Homing peptides as targeted delivery vehicles. Integr. Biol. 2, 326–337.
Pasqualini, R., Koivunen, E., Rouslahti, E., 1997. Av integrins as receptors for tumor targeting by circulating ligands. Nat. Biotechnol. 15, 542–546.
Pasqualini, R., et al., 2000. Aminopeptidase N is a receptor for tumor-homing peptides and a target for inhibiting angiogenesis. Cancer Res. 60, 722–727.
Gautam, A., et al., 2014. Tumor homing peptides as molecular probes for cancer therapeutics, diagnostics and theranostics. Curr. Med. Chem. 21, 2367–2391.
Kapoor, P., Singh, H., Gautam, A., Chaudhary, K., Kumar, R., Raghava, G.P.S., 2012. TumorHoPe: a database of tumor homing peptides. PLoS One 7, 1–6.
Sharma, A., et al., 2013. Computational approach for designing tumor homing peptides. Sci. Rep. 3, 1607.
Simeon, S., et al., 2016. osFP: a web server for predicting the oligomeric states of fluorescent proteins. J. Cheminform. 8, 72.
Simeon, S., et al., 2017. PepBio: predicting the bioactivity of host defense peptides. RSC Adv. 7, 35119–35134.
Pratiwi, R., et al., 2017. CryoProtect: a web server for classifying antifreeze proteins from nonantifreeze proteins. J. Chem. 2017.
Win, T.S., Malik, A.A., Prachayasittikul, V., Wikberg, J.E.S., Nantasenamat, C., Shoombuatong, W., 2017. HemoPred: a web server for predicting the hemolytic activity of peptides. Future Med. Chem. 9, 275–291.
Shoombuatong, W., Schaduangrat, N., Nantasenamat, C., 2018a. Towards understanding aromatase inhibitory activity via QSAR modeling. EXCLI J. 17, 688.
Shoombuatong, W., Schaduangrat, N., Nantasenamat, C., 2018b. Unraveling the bioactivity of anticancer peptides as deduced from machine learning. EXCLI J. 17, 734.
Chou, K.-C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236–247.
Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W., 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26, 680–682.
Shoombuatong, W., Hongjaisee, S., Barin, F., Chaijaruwanich, J., Samleerat, T., 2012. HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees. Comput. Biol. Med. 42, 885–889.
Shoombuatong, W., Huang, H.-L., Chaijaruwanich, J., Charoenkwan, P., Lee, H.-C., Ho, S.-Y., 2013. Predicting protein crystallization using a simple scoring card method. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) 23–30.
Charoenkwan, P., Shoombuatong, W., Lee, H.-C., Chaijaruwanich, J., Huang, H.-L., Ho, S.-Y., 2013. SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. PLoS One 8, e72368.
Golbraikh, A., Muratov, E., Fourches, D., Tropsha, A., 2014. Data set modelability by {QSAR}. J. Chem. Inf. Model. 54, 1–4.
Fourches, D., Muratov, E., Tropsha, A., 2016. Trust, but verify II: a practical guide to chemogenomics data curation. J. Chem. Inf. Model. 56, 1243–1252.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.
Breiman, L., 2006. random{F}orest: {B}reiman and {C}utler's random forests for classification and regression.
Kuhn, M., et al., 2017. caret: classification and regression training.
Calle, M.L., Urrea, V., 2011. Letter to the editor: stability of Random Forest importance measures. Brief. Bioinformatics 12, 86–89.

Stevens, A., Ramirez-Lopez, L., Stevens, M.A., Rcpp, L., 2015. prospectr. Miscellaneous Functions for Processing and Sample Selection of vis-NIR Diffuse Reflectance Data. pp. 32.
Robin, X., et al., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12, 77.
Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., Kanehisa, M., 2007. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 36, D202–D205.
Ganguly, S., Sadaoui, S., 2017. Classification of imbalanced auction fraud data. Canadian Conference on Artificial Intelligence 84–89.
Gentleman, R., 2008. R Programming for Bioinformatics. Chapman and Hall/CRC.
Huang, H.-L., et al., 2012. Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition. BMC Bioinformatics 13, S3.
Shoombuatong, W., et al., 2017. Towards the revival of interpretable QSAR models. In: Roy, K. (Ed.), Advances in QSAR Modeling, pp. 3–55.
Liaw, C., Tung, C.-W., Ho, S.-Y., 2013. Prediction and analysis of antibody amyloidogenesis from sequences. PLoS One 8, e53235.
Pang, H.-b., et al., 2014. A free cysteine prolongs the half-life of a homing peptide and improves its tumor-penetrating activity. J. Control. Release 175, 48–53.
Colombo, G., et al., 2002. Structure-activity relationships of linear and cyclic peptides containing the NGR tumor-homing motif. J. Biol. Chem. 277, 47891–47897.
Graziadio, A., et al., 2016. NGR tumor-homing peptides: structural requirements for effective APN (CD13) targeting. Bioconjug. Chem. 27, 1332–1340.
Von Heijne, G., 2007. The membrane protein universe: What's out there and why bother? J. Intern. Med. 261, 543–557.
Jesus, A.J.D., Allen, T.W., 2013. The role of tryptophan side chains in membrane protein anchoring and hydrophobic mismatch. Biochimica et Biophysica Acta (BBA) - Biomembranes 1828, 864–876.
Rangel, R., et al., 2013. Targeting mammalian organelles with internalizing phage (iPhage) libraries. Nat. Protoc. 8, 1916–1939.
Varner, J.A., Cheresh, D.A., 1996. Integrins and cancer. Curr. Opin. Cell Biol. 8, 724–730.
Hambley, T.W., Hait, W.N., 2009. Is Anticancer Drug Development Heading in the Right Direction? Cancer Res. 69, 1259–1262.
Scibior, D., Czeczot, H., 2004. Arginine-metabolism and functions in the human organism. Postepy Hig. Med. Dosw. 58, 321–332.
Geiger, R., et al., 2016. L-arginine modulates t cell metabolism and enhances survival and anti-tumor activity. Cell 167, 829–842 e13.
Sugahara, K.N., et al., 2009. Tissue-penetrating delivery of compounds and nanoparticles into tumors. Cancer Cell 16, 510–520.
Sugahara, K.N., et al., 2010. Coadministration of a tumor-penetrating peptide enhances the efficacy of Cancer drugs. Science 328, 1031–1035.
Elayadi, A.N., et al., 2007. A peptide selected by biopanning identifies the integrin αvβ6 as a prognostic biomarker for nonsmall cell lung Cancer. Cancer Res. 67, 5889–5895.
Gerlag, D.M., et al., 2001. Suppression of murine collagen-induced arthritis by targeted apoptosis of synovial neovasculature. Arthritis Res. 3, 357.
Eskens, F.A.L.M., et al., 2003. Phase I and pharmacokinetic study of continuous twice weekly intravenous administration of Cilengitide (EMD 121974), a novel inhibitor of the integrins alphavbeta3 and alphavbeta5 in patients with advanced solid tumours. Eur. J. Cancer 39, 917–926.
Fink, K., et al., 2010. Long-term effects of cilengitide, a novel integrin inhibitor, in recurrent glioblastoma: A randomized phase {IIa} study. J. Clin. Oncol. 28, 2010.
Gilbert, M.R., et al., 2011. Cilengitide in patients with recurrent glioblastoma: the results of NABTC 03-02, a phase II trial with measures of treatment delivery. J. Neurooncol. 106, 147–153.
Lee, E., Koskimaki, J.E., Pandey, N.B., Popel, A.S., 2013. Inhibition of Lymphangiogenesis and angiogenesis in breast tumor xenografts and lymph nodes by a peptide derived from transmembrane protein 45A. Neoplasia 15, 112–IN6.
Wang, H., Yang, Z., Gu, J., 2014. Therapeutic targeting of angiogenesis with a recombinant CTT peptide-endostatin mimic-kringle 5 protein. Mol. Cancer Ther. 13, 2674–2687.
Hyvonen, M., et al., 2014. Novel Target for Peptide-Based Imaging and Treatment of Brain Tumors. Mol. Cancer Ther. 13, 996–1007.