รายงานวิจัยฉบับสมบูรณ์

โครงการ: การตรวจหาอันตรกิริยาแบบอิพิสเตซิสบริสุทธิ์ระหว่างหลายตำแหน่งพันธุกรรมโดยการทดสอบการเรียงสับเปลี่ยนเชิงสุ่มที่รวมผลการวิเคราะห์อันตรกิริยาระหว่างตำแหน่งพันธุกรรมสองตำแหน่ง

Detecting Purely Epistatic Multi-locus Interactions by an Omnibus Permutation Test on Ensembles of Two-Locus Analyses

โดย รองศาสตราจารย์ ดร.ณชล ไชยรัตนะ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

เมษายน 2554

สัญญาเลขที่ RMU5380004

รายงานวิจัยฉบับสมบูรณ์

โครงการ: การตรวจหาอันตรกิริยาแบบอิพิสเตซิสบริสุทธิ์ระหว่าง หลายตำแหน่งพันธุกรรมโดยการทดสอบการเรียงสับเปลี่ยนเชิงสุ่มที่รวมผล การวิเคราะห์อันตรกิริยาระหว่างตำแหน่งพันธุกรรมสองตำแหน่ง

Detecting Purely Epistatic Multi-locus Interactions by an Omnibus Permutation Test on Ensembles of Two-Locus Analyses

รองศาสตราจารย์ ดร.ณชล ไชยรัตนะ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

สนับสนุนโดยสำนักงานคณะกรรมการการอุดมศึกษา
และสำนักงานกองทุนสนับสนุนการวิจัย

(ความเห็นในรายงานนี้เป็นของผู้วิจัย สกอ. และ สกว.ไม่จำเป็นต้องเห็นด้วยเสมอไป)

บทคัดย่อ

รหัสโครงการ: RMU5380004

ชื่อโครงการ: การตรวจหาอันตรกิริยาแบบอิพิสเตซิสบริสุทธิ์ระหว่างหลายตำแหน่งพันธุกรรมโดยการทดสอบการเรียงสับเปลี่ยนเชิงสุ่มที่รวมผลการวิเคราะห์อันตรกิริยาระหว่างตำแหน่งพันธุกรรมสองตำแหน่ง

ชื่อนักวิจัย: รองศาสตราจารย์ ดร.ณชล ไชยรัตนะ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

E-mail Address: nchl@kmutnb.ac.th

ระยะเวลาโครงการ: 2 ปี

ระยะเวลาสัญญา: 1 ปี

รายงานฉบับนี้สนใจโรคพันธุกรรมสองโรคที่สามารถประยุกต์ใช้การรู้จำแบบในการวิเคราะห์ โรคแรกที่สนใจคือโรคเบาหวานชนิดที่สอง 2LOmb ซึ่งทำการคัดเลือกลักษณะประจำโดยการทดสอบการเรียงสับเปลี่ยนเชิงสุ่มที่รวมผลการวิเคราะห์อันตรกิริยาระหว่างตำแหน่งพันธุกรรมสองตำแหน่งเป็นขั้นตอนวิธีที่พัฒนาขึ้นสำหรับปัญหาแรก ขั้นตอนวิธีประกอบด้วยสี่ขั้นได้แก่ การวิเคราะห์แบบสองตำแหน่งพันธุกรรม การทดสอบการเรียงสับเปลี่ยนเชิงสุ่ม การคำนวณค่าความน่าจะเป็นวงกว้าง และการค้นหาแบบก้าวหน้าเพื่อระบุการรวมการวิเคราะห์แบบสองตำแหน่งที่ดีที่สุด สมรรถนะของ 2LOmb ได้รับการเปรียบเทียบกับสมรรถนะของวิธีการหาความสัมพันธ์ทางพันธุกรรมแบบเซต การคัดเลือกลักษณะประจำแบบอิงสหสัมพันธ์ และเทคนิค ReliefF แบบปรับปรุง ผลการจำลองจากปัญหาอันตรกิริยาหลายตำแหน่งแสดงให้เห็นว่า 2LOmb ให้ค่าผิดพลาดแบบบวกเท็จต่ำ นอกจากนั้น 2LOmb สามารถระบุภาวะพหุสัณฐานนิวคลีโอไทด์เดี่ยวที่เป็นสาเหตุโรคได้ทั้งหมด ซึ่งชี้ให้เห็นว่า 2LOmb มีกำลังการตรวจจับสูง 2LOmb ได้รับการนำไปประยุกต์ใช้กับข้อมูลโรคเบาหวานชนิดที่สองที่ประกอบด้วย 7,065 ภาวะพหุสัณฐานนิวคลีโอไทด์เดี่ยวจาก 370 ยีน 2LOmb ระบุ 11 ภาวะพหุสัณฐานนิวคลีโอไทด์เดี่ยวจากสี่ยีนที่สัมพันธ์กับโรคเบาหวานชนิดที่สอง โรคที่สองที่สนใจคือโรคธาลัสซีเมีย โดยปัญหาที่สนใจคือการจำแนกคุณสมบัติของเลือดโดยใช้ต้นไม้ตัดสินใจ C4.5 ตัวจำแนกเบย์อย่างง่าย และมัลติเลเยอร์เพอร์เซ็บตรอนสำหรับการคัดกรองโรคธาลัสซีเมีย เป้าหมายคือการจำแนกประเภทความผิดปกติธาลัสซีเมีย 18 กลุ่มซึ่งมีความชุกสูงในประเทศไทยและคนปกติหนึ่งกลุ่ม

โดยอาศัยข้อมูลจากการตรวจความสมบูรณ์ของเม็ดเลือดและการตรวจชนิดฮีโมโกลบิน ผลการทดลองแสดงให้เห็นว่าได้สมรรถนะการจำแนกสูงสุดซึ่งมีค่าความแม่นเฉลี่ย 93.23% (ค่าเบี่ยงเบนมาตรฐาน 1.67%) และค่าความแม่นเฉลี่ย 92.60% (ค่าเบี่ยงเบนมาตรฐาน 1.75%) จากตัวจำแนกเบย์อย่างง่ายและมัลติเลเยอร์เพอร์เซ็บตรอนตามลำดับเมื่อทำให้ลักษณะประจำมีค่าแบบภินทนะก่อน และผลจากการคัดเลือกลักษณะประจำแบบห่อแสดงให้เห็นว่าค่าความเข้มข้นฮีโมโกลบินซึ่งเป็นลักษณะประจำไม่มีผลต่อสมรรถนะการจำแนก นอกจากนี้สมรรถนะการจำแนกนี้ยังสูงกว่าสมรรถนะการจำแนกที่ได้เมื่อใช้ข้อมูลจากการตรวจชนิดฮีโมโกลบินเป็นข้อมูลเข้าสำหรับตัวจำแนกเพียงอย่างเดียว

คำหลัก: การคัดเลือกลักษณะประจำ การจำแนก การรู้จำแบบ ธาลัสซีเมีย เบาหวานชนิดที่สอง

Abstract

Project Code:	RMU5380004
Project Title:	Detecting Purely Epistatic Multi-locus Interactions by an Omnibus Permutation Test on Ensembles of Two-Locus Analyses
Investigator:	Associate Professor Dr. Nachol Chaiyaratana King Mongkut’s University of Technology North Bangkok
E-mail Address:	nchl@kmutnb.ac.th
Project Period:	2 Years
Contract Period:	1 Year

This report interests in two genetic disease problems that can be tackled using pattern recognition. The first problem covers a genetic association study of type 2 diabetes mellitus (T2D). 2LOmb which performs attribute selection via an omnibus permutation test on ensembles of two-locus analyses is proposed. The algorithm consists of four main steps: two-locus analysis, a permutation test, global *p*-value determination and a progressive search for the best ensemble. 2LOmb is benchmarked against a set association approach, a correlation-based feature selection technique and a tuned ReliefF technique. The simulation results from multi-locus interaction problems indicate that 2LOmb has a low false-positive error. Moreover, 2LOmb has the best performance in terms of an ability to identify all causative single nucleotide polymorphisms (SNPs), which signifies a high detection power. 2LOmb is subsequently applied to a T2D data set, which contains 7,065 SNPs from 370 genes. The 2LOmb search reveals that 11 SNPs in four genes are associated with T2D. The second problem involves the classification of blood characteristics by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening.

The aim is to classify 18 classes of thalassaemia abnormality, which have a high prevalence in Thailand, and one control class by inspecting data characterised by a complete blood count (CBC) and haemoglobin typing. The stratified 10-fold cross-validation results indicate that the best classification performance with average accuracy of 93.23% (standard deviation = 1.67%) and 92.60% (standard deviation = 1.75%) is achieved when the naïve Bayes classifier and the multilayer perceptron are respectively applied to samples which have been pre-processed by attribute discretisation. The results from wrapper attribute selection also suggest that the haemoglobin concentration attribute is redundant. Moreover, the achieved classification performance is significantly higher than that obtained using only haemoglobin typing attributes as classifier inputs.

Keywords: Attribute Selection, Classification, Pattern Recognition, Thalassaemia, Type 2 Diabetes Mellitus

Acknowledgements

I acknowledge the financial support from the Office of the Higher Education Commission and the Thailand Research Fund.

Nachol Chaiyaratana

April 2011

Executive Summary

This report interests in two genetic disease problems that can be tackled using pattern recognition. The first problem covers a genetic association study of type 2 diabetes mellitus (T2D). 2LOmb which performs attribute selection via an omnibus permutation test on ensembles of two-locus analyses is proposed. The algorithm consists of four main steps: two-locus analysis, a permutation test, global *p*-value determination and a progressive search for the best ensemble. 2LOmb is benchmarked against three attribute selection techniques namely a set association approach, a correlation-based feature selection technique and a tuned ReliefF technique. The simulation results from multi-locus interaction problems indicate that 2LOmb has a low false-positive error. Moreover, 2LOmb has the best performance in terms of an ability to identify all causative single nucleotide polymorphisms (SNPs), which signifies a high detection power. 2LOmb is subsequently applied to a genome-wide T2D data set. After primarily screening for SNPs that locate within or near candidate genes and exhibit no marginal single-locus effects, the T2D data set is reduced to 7,065 SNPs from 370 genes. The 2LOmb search reveals that 11 SNPs in four genes are associated with T2D. The findings provide an alternative explanation for the aetiology of T2D.

The second problem involves the classification of blood characteristics by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening. The aim is to classify 18 classes of thalassaemia abnormality, which have a high prevalence in Thailand, and one control class by inspecting data characterised by a complete blood count (CBC) and haemoglobin typing. Two indices namely a haemoglobin concentration (HB) and a mean corpuscular volume (MCV) are the chosen CBC attributes. On the other hand, known types of haemoglobin from six ranges of retention time identified via high performance liquid chromatography (HPLC) are the chosen haemoglobin typing attributes. The stratified 10-fold cross-validation results indicate that the best classification performance with average accuracy of 93.23% (standard deviation = 1.67%) and 92.60% (standard deviation = 1.75%) is achieved when the naïve Bayes classifier and the multilayer perceptron are respectively applied to samples which have been pre-processed by attribute discretisation. The results from wrapper attribute selection also suggest that the HB attribute is redundant. Moreover, the achieved classification performance is significantly higher than that obtained using only haemoglobin typing attributes as classifier inputs. Subsequently, the naïve Bayes classifier and the multilayer perceptron are applied to an additional data set in a clinical trial which respectively results in accuracy of 99.39% and 99.71%. These results suggest that a combination of CBC and haemoglobin typing analysis with a naïve Bayes classifier or a multilayer perceptron is highly suitable for automatic thalassaemia screening.

Contents

1. Introduction

Mathematical models play a crucial role in scientific studies. This is because they serve many purposes including function approximation, interpolation and extrapolation. Pattern recognition interests in the mapping between inputs and discrete-valued outputs, which are commonly referred to as classes. In other words, pattern recognition concentrates on creating a mathematical model that captures an input-output relationship which leads to the correct identification of output classes after observing input features or attributes. Since the procedure of capturing the input-output relationship involves many steps, research in pattern recognition can be categorised according to these steps. Examples of these steps include attribute discretisation, attribute selection and classification model construction. Attribute discretisation involves a transformation of continuous-valued attributes into discrete-valued attributes. It is required when a number of classifiers including decision trees (Quinlan, 1993) are employed. Attribute selection interests in the identification of optimal attribute subset that leads to the maximum classification accuracy. Classification model construction concentrates on maximising the capability of identifying the correct class of an unknown pattern based on the available pattern data.

In this report, all three areas of pattern recognition are explored. Specifically, they are applied to two genetic disease problems: genetic association of type 2 diabetes mellitus and thalassaemia classification. Genetic association studies involve the identification of single nucleotide polymorphisms (SNPs) which are associated with the disease of interest. Since there are over 3,000,000 SNPs in the human genome, the problem can be formulated as an attribute selection problem. It will be demonstrated that a procedure involving an omnibus permutation test on ensembles of two-locus analyses and a decision table (Ritchie et al., 2001) provides a sufficient means for extracting informative SNPs from large-scaled data. In contrast, the thalassaemia classification problem covers a smaller number of attributes. Moreover, the attributes are continuous-valued attributes. It will be shown that the problem can be solved using a procedure which includes an information-theoretic attribute discretisation technique (Fayyad and Irani, 1993), wrapper attribute selection (Kohavi and John, 1997) and a naïve Bayes classifier (Mitchell, 1997) as well as a multilayer perceptron (Rumelhart and McClelland, 1986).

The organisation of this report is as follows. In Section 2, the genetic association of type 2 diabetes mellitus is explained. All necessary techniques for the study are also explained. Next, the thalassaemia classification problem and how it can be solved are discussed in Section 3. Finally, the conclusions are given in Section 4.

2. Genetic Association of Type 2 Diabetes Mellitus

2.1. Background

Complex diseases cannot generally be explained by Mendelian inheritance (Risch and Merikangas, 1996) because they are influenced by gene-gene and gene-environment interactions. Many common diseases such as asthma, cancer, diabetes, hypertension and obesity are widely accepted and acknowledged to be results of complex interactions between multiple genetic factors (Musani et al., 2007). Attempts to identify factors that could be the causes of complex diseases have led to many genome-wide association studies (The Wellcome Trust Case Control Consortium, 2007; The GAIN Collaborative Research Group, 2007). Raw results from these attempts produce a large amount of single nucleotide polymorphism (SNP) data from every individual participating in the trials.

For genetic epidemiologists, data sets from genome-wide association studies present many challenges, particularly the correct identification of SNPs that associate with the disease of interest from all available SNPs (Heidema et al., 2006). This challenge can be treated as a pattern recognition problem which aims to identify an attribute or SNP set that can lead to the correct classification of recruited samples. Heidema et al. (2006) and Motsinger et al. (2007) have reviewed and identified many machine learning techniques that are suitable to the task. Among many strategies and techniques, the protocol that appears to be most promising for genome-wide association studies involves two main steps: SNP set reduction and classification model construction (Moore et al., 2006).

The success of the two-step pattern recognition approach relies heavily on the attribute selection step (Hall and Holmes, 2003). In case-control studies, epistatic effects play a vital role in establishing the difficulty level of SNP screening problems

(Cordell, 2002). Epistasis in the simplest form can be represented by disease models that require genotype inputs from two interacting SNPs (Neuman and Rice, 1992; Schork et al., 1993). Many attempts have been made to produce consistent definitions and categorisation of different types of epistasis models (Musani et al., 2007, Cordell, 2002; Li and Reich, 2000; Marchini et al., 2005; Hallgrímsdóttir and Yuster, 2008). According to Musani et al. (2007), a pure epistasis model (Culverhouse et al., 2002) is difficult because each SNP exhibits no marginal single-locus effect in the model. As a result, it is impossible to detect the pure epistasis by univariate statistical tests.

Many genetic association studies reveal that various complex diseases are results of putative multi-locus interactions (Hoh et al., 2001; Heidema et al., 2007). With the constraints on a computational capability, exhaustive multi-locus analysis in large-scale or genome-wide association studies would be infeasible (Gayán et al., 2008). On the other hand, single-locus analysis would be unsuitable for the detection of pure epistasis. One possible approach that provides a trade-off between a computational limitation and an epistasis detection capability is to capture a multi-locus interaction by combining multiple results from two-locus analysis. To achieve this, it is necessary to prove that once a multi-locus interaction model is broken down into a combination of two-locus models, all or some of these models remain detectable through two-locus analysis. Although it is hinted in an early work on two-locus analysis (Gayán et al., 2008) that the proposed approach is plausible, explicit experimentation and testing has never been conducted.

In this report, the feasibility of employing an ensemble of two-locus analyses for the multi-locus interaction determination is demonstrated. Specifically, the significance of the two-locus analysis ensemble is assessed by an omnibus permutation test. The primary function of the proposed method is to detect possible association and assess its significance through the exploration of different ensembles of two-locus analyses. Hence, the proposed method is equally interested in both ensemble selection and testing for significant association.

The proposed method is benchmarked against a simple exhaustive two-locus analysis technique, the set association approach (Hoh et al., 2001), the correlation-based feature selection technique (Hall and Holmes, 2003) and the tuned ReliefF technique (Moore and White, 2007). These filter-based attribute selection techniques are suitable for the benchmark trial since they are capable of detecting association. The case-control classification models constructed from screened SNPs via a multifactor dimensionality reduction method (Ritchie et al., 2001) are also provided.

2.2. Methods

2.2.1. Omnibus Permutation Test on Ensembles of Two-Locus Analyses

The proposed algorithm performs an omnibus permutation test on ensembles of two-locus analyses and is referred to as a 2LOmb technique. The algorithm consists of four steps and can be described as follows.

2.2.1.1. Two-Locus Analysis

Consider a case-control genetic association study with n_m SNPs, for each pair of SNPs, a 2 × 9 contingency table with rows for disease status and columns for genotype configurations is created. A χ^2 test statistic and the corresponding *p*-value can subsequently be computed. With the total of n_m SNPs, there are $\binom{n_m}{2} = \frac{n_m!}{(n_m - 2)!2!}$ possible SNP pairs. As a result, the *p*-value from each two-locus analysis must be adjusted by a Bonferroni correction. The Bonferroni-corrected *p*-

value from each analysis is the lower value between $\binom{n_m}{2}\times$ the uncorrected *p*-value and one.

2.2.1.2. Permutation Test

The *p*-value p_0^e for the null hypothesis H_0^e that ensemble *e*—an ensemble of two-locus analyses of interest—is not associated with the disease can be evaluated by a permutation test. To achieve this, a scalar statistic is first computed from a function that combines the Bonferroni-corrected χ^2's *p*-values of individual two-locus tests. A suitable combining function must (a) be non-increasing in each *p*-value, (b) attain its maximum value when any *p*-value equals to zero and (c) have a finite critical value that is less than its maximum for any significant level greater than zero. In this study, a Fisher's combining function $(-2\sum_i \log(p_i))$ is selected. The *p*-value for the ensemble of two-locus analyses is assessed via a permutation simulation. In each permutation replicate, samples are constructed such that the case/control status of each sample is randomly permuted while the total numbers of case and control samples remain unchanged. A χ^2 contingency table with new entries and a Bonferroni-corrected *p*-value for the two-locus analysis within each permutation replicate are then obtained. This, in turn, leads to a new Fisher's test statistic. Let T_i^e denote the value of Fisher's test statistic obtained for the *i*th permutation replicate, p_0^e is the fraction of permutation replicates with a test statistic greater than or equal to the test statistic obtained from the original case-control data (T_0^e). In other words,

$$p_0^e = \left|\left\{i : 1 \le i \le t, T_i^e \ge T_0^e\right\}\right| / t\,, \tag{2-1}$$

where *t* is the number of permutation replicates which is set to 10,000 in this study and $|.|$ denotes the size of a set.

2.2.1.3. Global *p*-value Determination

There are many candidate ensembles of two-locus analyses that can be explored. Let $H_0 = \bigcap_{1 \le e \le E} H_0^e$ be the global null hypothesis that none of *E* explored ensembles of two-locus analyses is associated with the disease, the test of the global null hypothesis leads to the global *p*-value and provides the genetic association explanation. In the second step, the *p*-value p_0^e for a fixed hypothesis H_0^e is a raw or unadjusted *p*-value. To account for the correlation among multiple hypotheses that have been tested during the exploration through many candidate ensembles, the testing result of the global null hypothesis depends on $p_0^{\min} = \min_e p_0^e$. In other words, the global null hypothesis is rejected if the minimum of the raw *p*-values is sufficiently small. The distribution of $p_0^{\min}$ can again be determined by a permutation simulation. However, a nested simulation is unnecessary since the same set of permutation replicates for the p_0^e determination can be reused in the estimation of the empirical distribution of $p_0^{\min}$. The unadjusted *p*-value for the permutation replicate *i* of each hypothesis *e* is thus given by

$$p_i^e = \left|\left\{ j : 0 \le j \le t, j \ne i, T_j^e \ge T_i^e \right\}\right| \Big/ t \,. \tag{2-2}$$

Let $p_i^{\min} = \min_e p_i^e$ be the minimum of unadjusted *p*-values over all explored ensembles of two-locus analyses in the *i*th permutation replicate, the *p*-value for the global null hypothesis H_0 is defined by

$$p_{\text{global}} = \left|\left\{ i : 1 \le i \le t, p_i^{\min} \le p_0^{\min} \right\}\right| \Big/ t \,. \tag{2-3}$$

2.2.1.4. Search for the Best Ensemble of Two-Locus Analyses

A simple progressive search is used to identify the best ensemble of two-locus analyses. The search begins by locating the best two-SNP unit with the smallest Bonferroni-corrected χ^2's *p*-value from the first step. A permutation test is then performed for this two-locus analysis, yielding both raw and global *p*-values since only one hypothesis has been explored. Next, the search attempts to combine the existing best two-SNP unit with the two-SNP unit that possesses the next smallest Bonferroni-corrected χ^2's *p*-value from the first step and does not have a higher permutation *p*-value than the first two-SNP unit. If this new ensemble yields either a higher raw *p*-value or the same raw *p*-value but a higher global *p*-value from a permutation test, the search is terminated and the association is explained by the previously identified two-locus analysis. Otherwise, the best ensemble of two-locus analyses is updated and the process of appending more two-SNP units to the ensemble continues. The progressive search terminates when deterioration in the raw or global *p*-value is detected, or all possible two-locus analyses have been included in the ensemble. It is recalled from the third step that for the best ensemble containing $E-1<\binom{n_m}{2}$ two-locus analyses, its global *p*-value is obtained from the evaluation of E hypotheses.

2.2.2. Set Association Approach

A set association approach (SAA) is an association detection technique based on an omnibus permutation test on sets of candidate SNPs (Hoh et al., 2001). The test captures information about genotyping errors, deviation from Hardy-Weinberg equilibrium (HWE) and allelic association. In the first step, the genotype distribution for each SNP in the control samples is checked for HWE. Then, the number of SNPs that is to be excluded from the study (n_d) is set to the number of SNPs in the control samples that deviate from HWE. Two test statistics are subsequently calculated for each SNP: an allelic association statistic and a statistic for the deviation from HWE of each SNP in the case samples. The allelic association statistic is a χ^2 statistic which is calculated from the contingency table of alleles or genotypes with disease status. On the other hand, a χ^2 statistic for the deviation from HWE of each SNP in the case samples indicates the level of association. A large deviation from the equilibrium usually signifies strong association between a SNP and the disease. However, an excessively large deviation may be the result of genotyping errors. n_d SNPs with largest test statistics for the deviation from HWE are hence excluded from the consideration.

The test statistics for the allelic association and deviation from HWE are multiplied together to form a single *s* statistic for each remaining SNP. SNPs are then ranked according to their *s* statistics. A preset number of SNPs with highest ranks are considered for association. The first candidate SNP set contains only the SNP with the highest rank (the highest *s* statistic). The *p*-value for this first set is determined from a permutation simulation where the case and control labels are randomly permuted while the numbers of case and control samples remain unchanged. In each permutation replicate, a new genotype contingency table is constructed and a new *s* statistic is subsequently obtained. The *p*-value is given by the fraction of permutation replicates with an *s* statistic greater than or equal to the *s* statistic from the original data. The second candidate SNP set consists of the first two SNPs in the rank list. The test statistic for this SNP set is the sum of *s* statistics from both SNPs. The *p*-value for the second candidate SNP set is also obtained through the permutation simulation. By progressively adding the remaining SNP with the highest rank to the previously considered candidate set and performing the permutation simulation, *p*-values for all candidate SNP sets are estimated. The sizes of candidate SNP sets have the range of one to the preset number. Among all candidate sets, the SNP set that best describes genetic association has the lowest *p*-value.

Since multiple hypotheses are postulated during the construction of candidate SNP sets, the global *p*-value for the selected candidate set must be evaluated. This is achieved through a permutation simulation in which the current raw *p*-value for the chosen candidate set is now used as the test statistic. The existing permutation replicates, created for the early estimation of the raw *p*-value, can be reused and a nested permutation simulation is hence avoided. In this study, the maximum allowable size of the candidate SNP set is the total number of available SNPs while the number of permutation replicates for *p*-value evaluation is set to 10,000. The allelic association statistic employed in the study is the χ^2 statistic that is obtained through the contingency table of genotypes with disease status.

2.2.3. Correlation-Based Feature Selection Technique

A correlation-based feature selection (CFS) technique (Hall and Holmes, 2003) is an attribute (SNP) subset evaluation heuristic that considers both the usefulness of individual features (SNPs) in the (case-control) classification task and the level of inter-correlation among features. Each attribute subset is assigned a score given by

$$Merit_F = \frac{n_c \bar{r}_{cf}}{\sqrt{n_c + n_c(n_c - 1)\bar{r}_{ff}}} \tag{2-4}$$

where $Merit_F$ is the heuristic merit of an n_c-attribute subset *F*, $\bar{r}_{cf}$ is the average feature-class correlation and $\bar{r}_{ff}$ is the average feature-feature inter-correlation. An attribute subset receives a high merit score if it contains features that are highly correlated with the class and at the same time have low inter-correlation among one another. An application of a best-first search for the best subset identification is carried out to avoid searching through all possible attribute subsets.

2.2.4. Tuned ReliefF

A tuned ReliefF (TuRF) algorithm is a ranking algorithm for identifying genetic markers which are important in case-control classification (Moore and White, 2007). TuRF is built on a ReliefF engine (Robnik-Šikonja and Kononenko, 2003). ReliefF randomly picks a sample from the (case-control) data and identifies its n_k nearest neighbours from the same class and another n_k nearest neighbours from the opposite class. The attribute values—the genotypes in this application—of the neighbour samples are compared to that of the randomly picked sample and are subsequently used to update the relevance score for each attribute (genetic marker). This process is repeated for a specified number of samples, which is limited by the total sample size. The rationale of ReliefF is that an attribute which is important for the classification should have different values for samples from different classes and have the same value for samples from the same class. The relevance score of an attribute have a range from -1 (not relevant) to +1 (highly relevant). TuRF exploits the capability of ReliefF by repeatedly executing ReliefF and removing a portion of worst attributes at the end of each execution. This leads to the reevaluation of remaining attributes and, hence, reduces the effects of attribute noise on the attribute screening. In this study, the number of repetitions for random sample picking in the ReliefF part is equal to the total number of case-control samples while the neighbourhood size (n_k) for the relevance score calculation is set to ten. Furthermore, the worst 1% of SNPs is removed at the end of each ReliefF iteration (TuRF 1%).

2.2.5. Multifactor Dimensionality Reduction

A multifactor dimensionality reduction (MDR) method is a wrapper-based technique that is capable of identifying the best genetic marker combination among possible markers for the separation between case and control samples (Ritchie et al., 2001). Similar to other wrapper-based methods, an n_f-fold cross-validation technique provides a means to determine the classification accuracy of the candidate marker model. Basically, the combined case and control samples are randomly divided into n_f folds where $n_f - 1$ folds of samples are used to construct a decision table while the remaining fold of samples is used to identify the classification capability of the constructed decision table. The decision table construction and testing procedure is repeated n_f times. Hence, the samples in each fold are always used both to construct and to test the decision table. The number of cells in a decision table is given by G^{n_c} where n_c is the number of candidate markers selected from possible markers and *G* is the number of possible genotypes according to the marker. For a SNP, which is a bi-allelic marker, *G* is equal to three. During the decision table construction, each cell in the table is filled with case and control samples that have their genotype corresponds to the cell label. The ratio between numbers of case and control samples provides the decision for each cell whether the corresponding genotype is a protective or disease-predisposing genotype. An example of decision table construction is illustrated in Figure 2.1.

The classification accuracy of the decision table is subsequently evaluated by counting the numbers of case and control samples in the testing fold that their disease status can correctly be identified using the constructed decision rules. The process of decision table construction and evaluation must be cycled through all or some of possible $2^{n_m} - 1$ combinations where n_m is the total number of available markers in the study. The best genetic marker combination is determined from two criteria: classification accuracy and cross-validation consistency. Each time that a testing fold is used for the classification accuracy determination, the accuracy of the interesting marker combination model is compared with that from other models that also contain the same number of markers. The model that consistently ranks the first in comparison to other choices with the same number of markers has high cross-validation consistency (CVC). Classification accuracy is the main criterion for decision making while cross-validation consistency is only used as an auxiliary measure. Cross-validation consistency generally confirms that the high rank model can consistently be identified regardless of how the samples are divided for cross-validation. In a situation where two or more models with different number of markers are equally good in terms of classification accuracy and cross-validation consistency, the most parsimonious model—the combination with the least number of markers—is chosen as the best model.

After the best model has been selected, a permutation test is used to assess the probability of obtaining classification accuracy that is at least as large as or larger than that observed in the original data from randomised data. This represents the probability that the null hypothesis of no association is true. Each permutation replicate is constructed by randomly assigning the case/control status to each sample with the numbers of case and control samples remaining fixed. MDR analysis is subsequently carried out to obtain the classification accuracy of each permutation replicate. The empirical *p*-value is denoted by the fraction of permutation replicates with the classification accuracy greater than or equal to the classification accuracy obtained from the original data.

2.2.6. Implementation

2LOmb is implemented in a C programming language. The program can be compiled by Microsoft Visual Studio and GNU C compilers. The program has been successfully tested for the execution under Windows and Linux operating systems. 2LOmb can tackle problems in quadratic time. 2LOmb in its present form occupies one processor during the program execution. A parallel version of 2LOmb for genome-wide data is under development. All results included in the study are collected from the execution of computer programs in a Beowulf cluster. The computational platform consists of 12 nodes. Each node is equipped with dual Xeon 2.8 GHz processors and 4GB of main memory. The Rocks Cluster Distribution is installed on all nodes.

2.3. Results and Discussions

2.3.1 Testing with Simulated Data

2LOmb is benchmarked against SAA (Hoh et al., 2001), CFS (Hall and Holmes, 2003) and TuRF (Moore and White, 2007) in a simulation trial. The simulation covers data with causative SNPs, which signify pure epistasis. An efficient algorithm should produce a result with a high number of correctly-identified causative SNPs. This signifies the detection capability. An efficient algorithm should also produce a result with a low number of erroneous SNPs, which are irrelevant to the correct association explanation. This provides an indication for the false-positive error. These two measures on the number of SNPs in the results are used as the performance indicators.

Each simulated data set contains 1,000, 2,000 or 4,000 SNPs in which pure epistasis is governed by two, three or four causative SNPs. The allele frequencies of all causative SNPs are 0.5 while the minor allele frequencies of the remaining SNPs are between 0.05 and 0.5. The data set consists of balanced case-control samples of sizes 400, 800 or 1,600. All SNPs in control samples are in HWE. The genotype distribution of causative interacting SNPs follows the pure epistasis model by Culverhouse et al. (2002), leading to the heritability of 0.01. Every SNP in each data set exhibits no marginal single-locus effect (Bonferroni-corrected χ^2's p-value > 0.05). Seventy-five independent data sets for each simulation setting are generated via a genomeSIM package (Dudek et al., 2006). A paired *t*-test is suitable to assess the significance of results since the same simulated data sets are used during the algorithm benchmarking.

The results from the two-, three- and four-locus interaction problems are shown in Figure 2.2. Clearly, 2LOmb significantly outperforms other techniques in terms of the low number of erroneous SNPs, the high number of correctly-identified causative SNPs or both in every interaction problem (a paired *t*-test on 675 benchmark results yields a p-value < 0.05). The statistical power analysis also reveals that the benchmark trial with 75 independent data sets for each simulation setting is sufficient for an accurate evaluation of the overall algorithm performance (power > 0.95 for a Type I error rate of 0.05).

2LOmb also has an advantage in terms of computational time over the other three techniques. The computational time for all four techniques to finish screening the SNPs is provided in Table 2.1 to demonstrate this strength of 2LOmb. It can be clearly seen that the maximum time required by 2LOmb to screen SNPs in the largest data set is 419 seconds or approximately seven minutes. This is much less than the computational time required by the other techniques for the same data set.

2.3.2. Testing with Real Data

2LOmb has been applied to study a type 2 diabetes mellitus (T2D) data set, collected and investigated by the Wellcome Trust Case Control Consortium (WTCCC) (The Wellcome Trust Case Control Consortium, 2007). The data set consists of 1,999 case samples from affected individuals in the UK and 3,004 control samples, which are the results of a merging between 1,500 samples from the UK blood services and 1,504 samples from the 1958 British birth cohort. The original genome-wide data set contains 500,568 SNPs that are obtained through the Affymetrix GeneChip 500K Mapping Array Set. The SNP set is primarily reduced by screening for SNPs within and near 372 candidate genes collected by the Human Genome Epidemiology Network (HuGENet) (Yu et al., 2008). These candidate genes cover genes from both positive and negative genetic association reports, in which studies are conducted in various ethnic groups and populations. The SNP set is further reduced by removing SNPs that exhibit strong evidence of genetic association via single-locus analysis. The final SNP set contains 7,065 SNPs from 370 candidate genes. All SNPs in the reduced data set exhibit no marginal single-locus effects (Bonferroni-corrected χ^2's p-value > 0.05).

The 2LOmb search in the reduced T2D data set takes 3,456 seconds (57.6 minutes) of computational time on the Beowulf cluster. The possible genetic association is detected from 11 intronic SNPs in four genes (global p-value < 0.0001). Details of these SNPs, the two-SNP units that exhibit marginal two-locus effects and the identified genes are given in Table 2.2. A two-SNP unit is located in *LMX1A*. A two-SNP unit is also detected in *PARK2*. In addition, there is one SNP in common among SNPs in both *GYS2* two-SNP units. Similarly, there is one common SNP among three two-SNP units located in *PGM1*. Nonetheless, a two-SNP unit in which each SNP is located in a different gene is absent, indicating that there is no evidence of gene-gene interactions which can be observed from the 2LOmb result. Linkage disequilibrium (LD) analysis is subsequently performed using a JLIN package (Carter et al., 2006) and the resulting LD patterns are illustrated in Figure 2.3. It is noted that there is strong LD among SNPs within each gene due to high values of D' and r^2. The detection of these two-SNP units is thus believed to be the consequence of haplotype effects. An early T2D association study also reveals similar haplotype effects in FUSION data (Epstein and Satten, 2003). Next, an interaction dendrogram (Jakulin and Bratko, 2003; Jakulin et al., 2003) constructed from the 11 SNPs is given in Figure 2.4. A strong synergistic effect between the two SNPs in *PARK2* is clearly observed. In contrast, the interactions between *PGM1*, *LMX1A*, *PARK2* and *GYS2* are clearly absent.

Since many early genetic association studies of T2D and metabolic syndrome employ MDR analysis (Cho et al., 2004; Hsieh et al., 2006; Qi et al., 2007; Fiorito et al., 2007), additional MDR analysis would be useful for the comparison. The screened T2D case-control data set which contains 11 SNPs identified by 2LOmb is further subjected to MDR analysis. The classification accuracy of the best MDR model is summarised in Table 2.3. The model covers six SNPs in three genes: *PGM1, PARK2* and *GYS2*. These SNPs are also present in three two-SNP units identified by 2LOmb. It is noted that the classification accuracy in this real data set is much less than that from the simulated data sets. Nevertheless, the attainment of low classification accuracy does not necessarily suggest that there is no genetic association. Early works involving genetic association studies of T2D and metabolic syndrome in various populations via MDR analysis produce similar values of classification accuracy as summarised in Table 2.4. The classification accuracy by MDR from most studies is in the range of 0.5-0.6. The only genetic association study of T2D that the classification accuracy is distinctively high is conducted in a Korean population (Cho et al., 2004). Differences in genetic background, candidate genes and selected SNPs are the main causes of variation in the genetic association results. Although MDR does not select five SNPs from the 2LOmb output, these SNPs should not be regarded as erroneous SNPs because there is strong linkage disequilibrium among SNPs in each gene. Moreover, genotype and haplotype relative risk analysis clearly indicates that each gene, identified by 2LOmb, plays a role in the T2D association explanation. Overall, the analysis with the methods above only confirms the positive association for *PGM1*, *LMX1A*, *PARK2* and *GYS2* while gene-gene interactions are clearly absent. This signifies that, for the current study, there is no interaction between each pair of the identified genes that can be described by purely epistatic two-locus interaction models. In addition, there are no interactions between these four genes that can be described by purely epistatic multi-locus interaction models with marginal two-locus effects.

The four genes selected by 2LOmb regulate many pathways that involve in the disease development (Kanehisa and Goto, 2000; Kanehisa et al., 2006; Kanehisa et al., 2008). The genetic association studies involving these genes have been previously conducted in different populations. For instance, *LMX1A* has been chosen as a positional and biological candidate gene for a case-control study of T2D in Pima Indians (Thameem et al., 2002). This gene is chosen as a candidate because a linkage of T2D to chromosome 1q21-q23 has been previously reported (Hanson et al., 1998). In addition, *LMX1A* is one of LIM homeobox genes that are expressed in pancreas and has been shown to activate insulin gene transcription. Although SNPs have been carefully selected from the entire gene, no association between these SNPs in *LMX1A* and T2D has been found in this ethnic group.

PARK2 is another candidate gene that is also selected for case-control studies, based on evidence from genome-wide linkage analysis (Leak et al., 2008). A linkage of T2D in an African American population to chromosome 6q24-q27 has been previously identified (Sale et al., 2004). Although *PARK2* mainly involves in the development of Parkinson's disease, single-locus analysis reveals strong evidence of association between SNPs, which are in the vicinity of SNPs identified by 2LOmb, and T2D in African Americans.

In contrast to *LMX1A* and *PARK2*, which are candidate genes in typical T2D case-control studies, *GYS2* is considered in a study to identify genes responsible for troglitazone-associated hepatotoxicity in Japaneses with T2D (Watanabe et al., 2003). In other words, both case and control samples in the study are drawn from troglitazone-treated T2D patients, in which case patients exhibit an abnormal increase in alanine transaminase (ALT) and aspartate transaminase (AST) levels. *GYS2* regulates starch and sucrose metabolism and an insulin signalling pathway. The selected SNPs in *GYS2* are not found to associate with troglitazone-induced hepatotoxicity.

Similar to the study of *GYS2*, the association study involving *PGM1* is not carried out as a typical T2D case-control study. In fact, an attempt to identify association between *PGM1* polymorphisms and obesity has been conducted among T2D affected individuals in Italy (Gloria-Bottini et al., 2007). *PGM1* regulates glycolysis and gluconeogenesis, starch and sucrose metabolism, galactose metabolism, a pentose phosphate pathway, and streptomycin biosynthesis. Isozyme polymorphisms (Spencer et al., 1964; March et al., 1993), which are defined through structural differences in PGM1 protein, are used instead of SNPs in the study where positive association is identified.

In summary, positive association has been reported from previous studies involving *PARK2* in African Americans and *PGM1* in Italians. In contrast, negative association has been reported from previous studies about *LMX1A* in Pima Indians and *GYS2* in Japaneses. Both *GYS2* and *PGM1* regulate starch and sucrose metabolism while *LMX1A* and *PARK2* govern insulin gene transcription and Parkinson's disease development, respectively. The above discussion strengthens the importance of conducting large-scale association studies due to two main reasons. Firstly, a gene that does not contribute to the aetiology of a complex disease in one population may be important for association explanation in another population. Secondly, the absence of interacting candidate genes from a study may lead to negative association due to a lack of necessary genetic information. A two-locus interaction can occur between SNPs from genes that regulate one specific pathway (Hsieh et al., 2006) or between SNPs from genes that regulate different pathways (Qi et al., 2007). Furthermore, a multi-locus interaction may involve both SNPs from genes that regulate the same pathway and SNPs from genes that govern different pathways. Hence, candidate genes should be selected by considering all pathways that directly and indirectly contribute to the disease development.

This study produces evidence of association between 11 intronic SNPs in *PGM1*, *LMX1A*, *PARK2* and *GYS2*, and T2D in a UK population. Although there are other independent genome-wide T2D data sets, the association detection within these data using a similar methodology to the presented method has never been attempted because the methodology employed in the majority of genome-wide association studies is based on single-locus analysis (The Wellcome Trust Case Control Consortium, 2007; Zeggini et al., 2008). It is recalled that each SNP explored in the reduced T2D data set exhibits no marginal single-locus effect. Hence, the most logical approach to confirm the possibility of replicating association results from the current study is to perform the same detection method on these independent data sets. This is certainly important to gain further understanding of the genetic role in T2D susceptibility.

2.4. Conclusions

In this report, a method for detecting epistatic multi-locus interactions in case-control data is presented. The study focuses on pure epistasis (Musani et al., 2007), which cannot be detected via single-locus analysis (Culverhouse et al., 2002). To overcome this difficulty, the proposed method performs an omnibus permutation test on ensembles of two-locus analyses and is thus referred to as 2LOmb. The detection performance of 2LOmb is evaluated using both simulated and real data. From the simulation, 2LOmb produces a low false-positive error and a high detection power. Furthermore, 2LOmb outperforms a set association approach (SAA) (Hoh et al., 2001), a correlation-based feature selection (CFS) technique (Hall and Holmes, 2003) and a tuned ReliefF (TuRF) technique (Moore and White, 2007) in various interaction scenarios. These scenarios are set up by varying the number of causative SNPs, the number of SNPs in data and the sample size. 2LOmb is subsequently applied to a real case-control type 2 diabetes mellitus (T2D) data set, which is collected from a UK population by the Wellcome Trust Case Control Consortium (WTCCC) (The

Wellcome Trust Case Control Consortium, 2007). The original genome-wide data sets are first reduced by selecting only SNPs that locate within or near candidate genes reported by the Human Genome Epidemiology Network (HuGENet) (Yu et al., 2008). In addition, the selected SNPs must exhibit no marginal single-locus effects. The final T2D data set contains 7,065 SNPs from 370 genes. 2LOmb identifies 11 SNPs in four genes that are associated with T2D. This evidence of genetic association leads to an alternative explanation for the aetiology of T2D in the UK population. It also implies that SNPs from genome-wide data which are usually discarded after single-locus analysis confirms the null hypothesis of no association can still be useful for genetic association studies of complex diseases.

Table 2.1 Computational time required by all four techniques to detect interactions in simulated data sets. Only one computing processor in a Beowulf cluster is occupied during the analysis of one data set. The displayed time is evaluated from the processing of 75 independent data sets for each simulation setting. The computational time from the benchmark trial is the maximum time needed by each method to detect interactions in one data set.

Number of causative SNPs	Sample size	Computational time required by each technique (sec)							
		2LOmb		SAA		CFS		TuRF	
		1,000 SNPs	4,000 SNPs	1,000 SNPs	4,000 SNPs	1,000 SNPs	4,000 SNPs	1,000 SNPs	4,000 SNPs
2	400	15	135	2,838	3,560	2,182	4,721	2,137	80,312
	800	21	224	2,222	5,065	6,884	10,135	3,739	161,032
	1,600	36	400	2,997	10,105	20,054	31,034	7,134	322,084
3	400	22	140	2,788	3,539	1,302	5,546	2,319	78,892
	800	30	229	3,239	5,059	6,888	10,328	3,521	170,936
	1,600	50	406	5,758	10,393	19,395	30,865	5,827	322,870
4	400	32	150	3,070	5,075	3,050	6,427	2,071	73,654
	800	46	236	3,306	6,914	8,038	12,985	3,823	157,369
	1,600	70	419	5,508	11,368	22,407	36,693	6,780	340,824

Table 2.2 2LOmb identifies 11 intronic SNPs, which are located in four genes, from the reduced T2D data. Association between these SNPs and the disease is possible (global *p*-value < 0.0001). Seven two-SNP units are present in the ensemble where each unit contains a pair of SNPs from the same gene.

Gene	Chromosome and location	Two-SNP unit in the ensemble
PGM1	1p31	(rs2269241, rs3790857)
		(rs2269239, rs3790857)
		(rs3790857, rs2269238)
LMX1A	1q22-q23	(rs2348250, rs6702087)
PARK2	6q25.2-q27	(rs1893551, rs6924502)
GYS2	12p12.2	(rs6487236, rs1871142)
		(rs1871142, rs10770836)

Table 2.3 Classification accuracy of the best MDR model constructed from the 2LOmb output. The model contains six SNPs from *PGM1*, *PARK2* and *GYS2*. A permutation test with 1,000 randomised replicates of case-control data for this model reveals that the empirical *p*-value for the null hypothesis of no association is *p* < 0.001.

Description	Value
SNP and gene	rs2269241 (*PGM1*), rs3790857 (*PGM1*), rs1893551 (*PARK2*), rs6924502 (*PARK2*), rs1871142 (*GYS2*), rs10770836 (*GYS2*)
Classification accuracy (%)	54.02
CVC	9/10

Table 2.4 Summary of classification accuracy by MDR from early genetic association studies of T2D in a Korean population, a Han Chinese population from Taiwan, a female population from the US, and that from an early genetic association study of metabolic syndrome in an Italian population from the Centre East Coast Italy. A permutation test with 1,000 randomised replicates is performed to obtain the empirical *p*-value for the null hypothesis of no association in the studies conducted in the US and Italian populations. In contrast, a permutation test with 100 randomised replicates is performed to obtain the empirical *p*-value in the study conducted in the Korean population.

Reference	Population	Gene	Classification acc. (%)	CVC	Permutation *p*-value
Cho et al. (2004)	Korean	*PPARG*, *UCP2*	79.57	9/10	0.01
Hsieh et al. (2006)	Han Chinese	*RXRG*, *EGFR*	62.70	11/12	N/A
Qi et al. (2007)	US	*KCNJ11*, *HNF4A*	54.20	10/10	0.010
Fiorito et al. (2007)	Italian	*PPARG*, *DIO2*	61.70	10/10	0.005

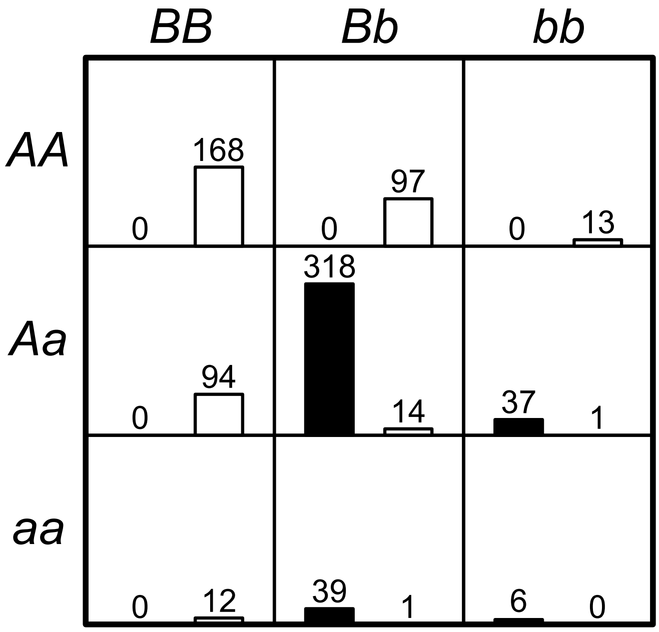


Figure 2.1 An MDR decision table that is constructed using a balanced case-control data set with the sample size of 800. The genotype of each sample is determined from two SNPs. The table consists of nine cells where each cell represents a unique genotype. The left (black) bar in each cell represents the number of case samples while the right (white) bar represents the number of control samples. The cells with genotypes *AABB*, *AABb*, *AAbb*, *AaBB* and *aaBB* are labelled as protective genotypes while the cells with genotypes *AaBb*, *Aabb*, *aaBb* and *aabb* are labelled as disease-predisposing genotypes.

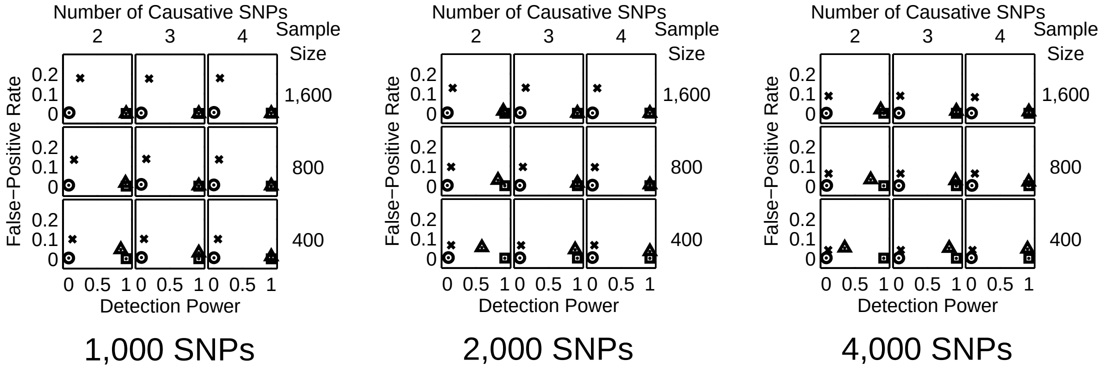


Figure 2.2 Performance of SAA, CFS, TuRF and 2LOmb in all simulation scenarios. Detection is declared for SAA and 2LOmb if the *p*-values used as detection indicators in their results are less than 0.05. The results from SAA, CFS, TuRF and 2LOmb are displayed using circle, cross, triangle and square markers, respectively.

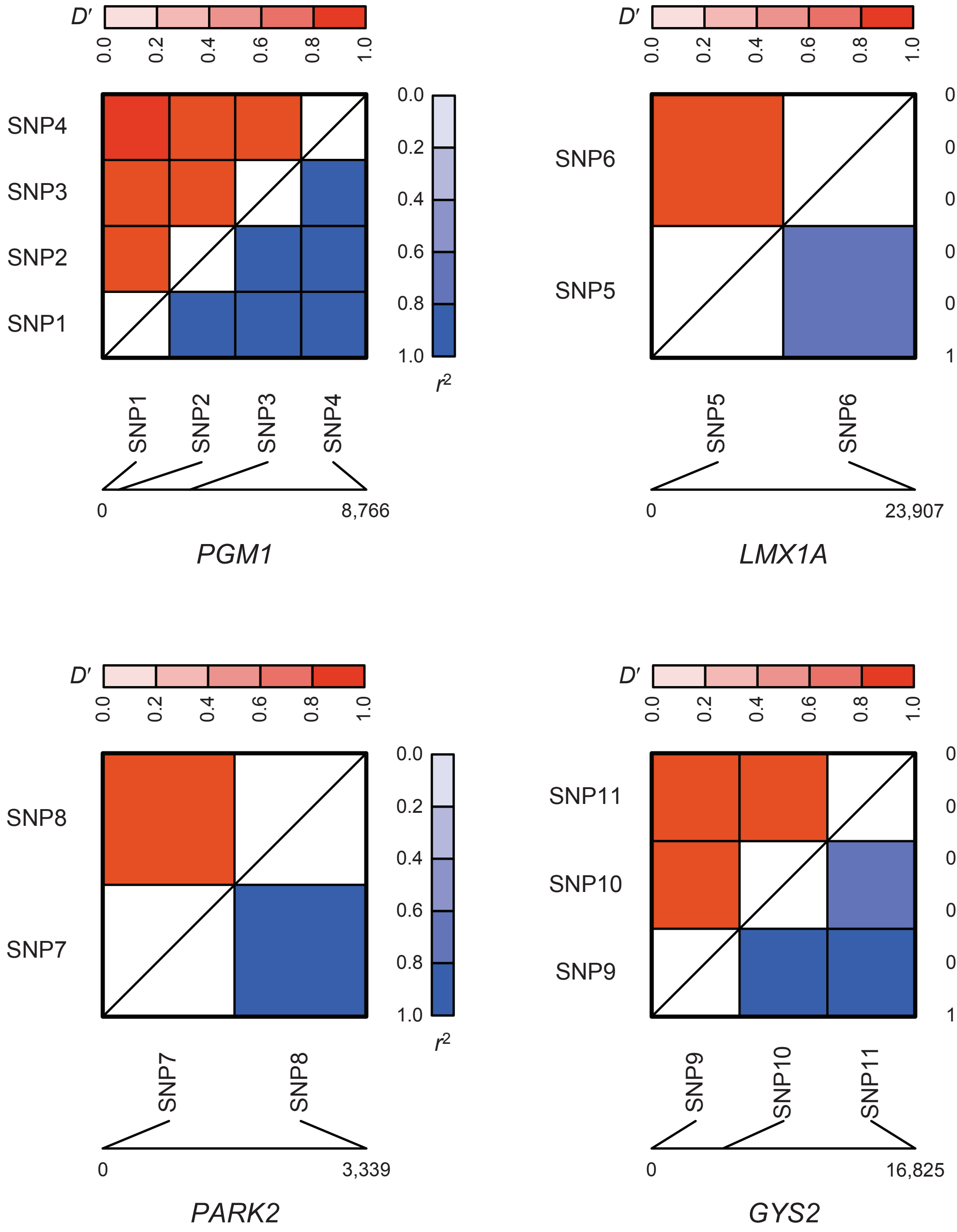


Figure 2.3 Linkage disequilibrium (LD) patterns of SNPs in *PGM1*, *LMX1A*, *PARK2* and *GYS2*. LD is explained via *D*′ displayed in the upper triangle and r^2 displayed in the lower triangle. Dark colours indicate high values while pale colours indicate low values. Distances between SNPs are given in terms of the number of base pairs. SNP1 = rs2269241, SNP2 = rs2269239, SNP3 = rs3790857, SNP4 = rs2269238, SNP5 = rs2348250, SNP6 = rs6702087, SNP7 = rs1893551, SNP8 = rs6924502, SNP9 = rs6487236, SNP10 = rs1871142 and SNP11 = rs10770836.

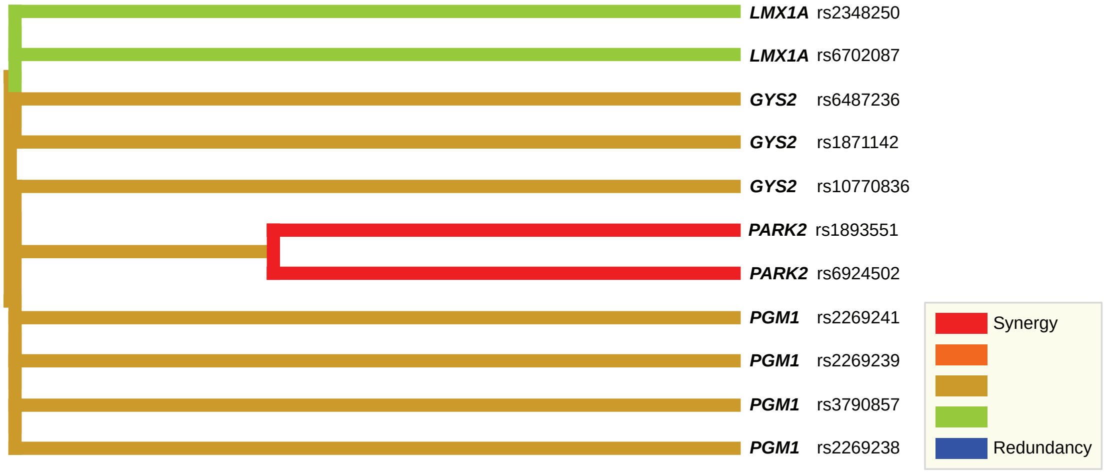


Figure 2.4 Interaction dendrogram produced from 11 SNPs that are chosen by 2LOmb. The colours in the dendrogram comprise a spectrum of colours representing a transition from synergy to redundancy. Synergy denotes the situation in which the entropy-based interaction between two SNPs provides more information than the entropy-based correlation between the pair. Redundancy refers to the situation in which the entropy-based interaction between two SNPs provides less information than the entropy-based correlation between the pair (Moore et al., 2006).

3. Thalassaemia Classification

3.1. Background

Thalassaemia is a genetic disease that has a high prevalence in Thailand (Fucharoen and Winichagoon, 1997). It is a result of an abnormality in genes that govern the formation of a protein called globin, which is a major component of haemoglobin (Hb). Since each red blood cell contains approximately 300 million molecules of haemoglobin, a modification of globin structure affects the cell structure and functionality. This subsequently leads to the reduction in the life span of a red blood cell (Weatherall and Clegg, 2001). The globin protein contains two components: α-globin and β-globin. The α-globin and β-globin synthesis is regulated by genes on chromosomes 16 and 11, respectively. Since the transmission mode of abnormal genes is autosomal recessive, a person must have two copies of a recessive gene on the same chromosome to develop the disease. In general, blood characteristics are analysed during the course of disease diagnosis. A complete blood count (CBC) and haemoglobin typing are the primary screening tests for a laboratory diagnosis of thalassaemia. Nonetheless, there is still a limitation in the data analysis due to a large number of candidate blood characteristics. Moreover, there are many types of thalassaemia and thalassaemia trait. (Persons with thalassaemia trait do not have the disease but inherit genes the cause the disease.) As a result, manual diagnosis is needed to be carried out by trained professionals (Jimenez et al., 1995; Demir et al., 2002; Ntaios et al., 2007; Sripichai et al., 2008).

Early attempts to develop an automatic diagnostic tool involve CBC data analysis using image processing (Lund and Barnes, 1972), statistical (Engle et al., 1976) and clustering techniques (Barosi et al., 1985). Later, the research interest has shifted to the use of expert systems where both rule-based (Quaglini et al., 1986;

Quaglini et al., 1988; Lanzola et al., 1990) and hybrid neural network/rule-based systems (Birndorf et al., 1996) have been successfully implemented for clinical trials. Nonetheless, these tools broadly differentiate between a wide range of blood disorders including various types of anaemia. As a result, many subsequent research works focus on the development of a diagnostic tool that only differentiates between thalassaemic patients, persons with thalassaemia traits and normal subjects. These works cover the implementation of a multilayer perceptron (Amendolia et al., 2002; Amendolia et al., 2003; Wongseree et al., 2007), a *k*-nearest neighbour technique (Amendolia et al., 2003), a support vector machine (Amendolia et al., 2003) and genetic programming (Wongseree et al., 2007) as thalassaemic diagnostic tools. Among these machine learning techniques, the multilayer perceptron emerges as the most suitable tool for the thalassaemia classification problem in Thailand (Wongseree et al., 2007) which covers higher varieties of haemoglobinopathies than other countries (Fucharoen and Winichagoon, 1997). The multilayer perceptron is capable of handling a problem with 13 classes of thalassaemia abnormality and two classes of normal subjects with and without iron deficiency. However, the classification accuracy during the clinical trial that covers 300 samples is only 81.6% (Wongseree et al., 2007).

A significant improvement in thalassaemia classification accuracy has been achieved through the application of machine learning techniques in conjunction with haemoglobin typing inputs. The techniques that have been applied to the problem include a C4.5 decision tree, a random forest and a multilayer perceptron. The C4.5 decision tree is proven to be the most suitable technique where the classification accuracy during a clinical trial involving 1,000 samples from 13 classes of thalassaemia abnormality and a normal subject class is 93.1% (Piroonratana et al., 2009). The improvement in classification accuracy can be achieved because CBC and haemoglobin typing data represents different aspects of blood characteristics. CBC information is useful for the diagnosis of various types of anaemia (Quaglini et al., 1986) while haemoglobin typing information can confirm the haemoglobinopathies (Sirichotiyakul et al., 2009). Nonetheless, haemoglobin typing alone is insufficient for the classification between certain types of thalassaemia abnormality. For instance, haemoglobin typing characteristics cannot be used to differentiate between a person with α-thalassaemia 1 trait, a person with α-thalassaemia 2 trait and a normal subject (Old, 2003).

Since both CBC and haemoglobin typing data are usually available for the laboratory diagnosis of thalassaemia, an attempt to develop an automated diagnostic tool that takes both forms of data should be carried out. Consequently, this should lead to an improvement of classification accuracy. In this report, the possibility of using CBC and haemoglobin typing data in thalassaemia classification by machine learning is investigated. The choices of machine learning technique include a multilayer perceptron, a C4.5 decision tree and a naïve Bayes classifier. The first two techniques are chosen because they are proven to be suitable in the early investigations involving CBC (Wongseree et al., 2007) and haemoglobin typing inputs (Piroonratana et al., 2009), respectively. In contrast, a naïve Bayes classifier is selected due to its classification efficacy and implementation simplicity (Hall and Holmes, 2003). As a result, it is common to provide classification accuracy achieved by a naïve Bayes classifier as a comparison baseline (Hall and Holmes, 2003; Polat and Günes, 2006).

With the availability of CBC and haemoglobin typing data and selected choices of machine learning technique, an investigation can be conducted as follows.

Firstly, the data is pre-processed via input attribute discretisation. Informative attributes identified in the previous works (Wongseree et al., 2007; Piroonratana et al., 2009) are discretised since it has been reported that a proper discretisation of continuous-valued attributes can significantly improve the classification accuracy of both multilayer perceptron and C4.5 decision tree (Piroonratana et al., 2009). The attributes are thus discretised via an information-theoretic technique proposed by Fayyad and Irani (1993). As a result, the data with discrete-valued attributes is available for classifier benchmarking. Since the information contained within the CBC attributes and necessary for the classification may overlap with that contained within the haemoglobin typing attributes, redundant attributes that can be removed without affecting the classification accuracy are also identified during classifier benchmarking. Correlation analysis via symmetrical uncertainty measurement (Press et al., 1988) is subsequently performed on the attributes necessary for the classification. After the classifier performance evaluation is completed, the best classifier together with the pruned attribute set is chosen for a clinical trial involving a separate data set. This independent data set contains more samples than those in the early clinical trials by Wongseree et al. (2007) and Piroonratana et al. (2009). Finally, classification analysis of the clinical trial results is carried out to determine the feasibility of the chosen classifier and pruned attribute set. Every step in the procedure described above is illustrated in Figure 3.1 and is implemented using a WEKA package (Witten and Frank, 2005).

3.2. Materials and Methods

3.2.1. CBC and Haemoglobin Typing Data Sets

The data sets for thalassaemia classification consist of various blood characteristics obtained through CBC and haemoglobin typing. In this study, the data sets consist of eight input attributes. Details of these attributes are given in Table 3.1. The first two attributes are obtained through CBC: a haemoglobin concentration (HB) and a mean corpuscular volume (MCV). These two attributes are chosen since they are proven to be highly informative in the early studies (Amendolia et al., 2003; Wongseree et al., 2007). The last six attributes are obtained through haemoglobin typing. Multiple haemoglobin typing attributes, characterised through the use of high performance liquid chromatography (HPLC) (Clarke and Higgins, 2000), are necessary for classification since a blood specimen generally contains more than one type of haemoglobin. As a result, many types of thalassaemia abnormality can be identified through the difference in proportion of haemoglobin contents (Old, 2003; Ou and Rognerud, 2001; Colah et al., 2007). The haemoglobin typing attributes are extracted from elution chromatograms (Joutovsky et al., 2004). Typical elution chromatograms of two different specimens are illustrated in Figure 3.2. The first chromatogram shows that the specimen is mostly made up from Hb A_0. If the MCV value obtained via CBC is greater than 75 fL, this specimen is most likely to be taken from a normal person. Otherwise, this specimen must be obtained from a person with α-thalassaemia 1 trait. The second chromatogram indicates that the specimen consists of Hb A_0 and Hb E. This means that the specimen is taken from a person with Hb E trait. It is noticeable that some types of thalassaemia abnormality can be diagnosed via haemoglobin typing analysis alone while the diagnosis of other types requires both CBC and haemoglobin typing information. Figure 3.2 also illustrates that different types of haemoglobin are detectable in the form of elution peaks at different retention time. Hence, a chromatogram can be divided into multiple sections where each section occupies a non-overlapping range of retention time. Consequently, each chromatogram section represents a unique attribute for the classification where the percentage of haemoglobin in the elution profile corresponds to the attribute value. The last six attributes and the associated types of haemoglobin are also summarised in Table 3.1. It is noticed that two attributes representing the elution profiles which occupy the retention time between 161 and 199 seconds and between 231 and 249 seconds are not needed. These two attributes correspond to unknown types of haemoglobin and are proven to be uninformative for the classification task (Piroonratana et al., 2009).

Two confirmed diagnosis data sets are acquired for this study. The first data set is created for the evaluation of classifier performance while the second set is used in a clinical trial. The data set for the classifier evaluation consists of 1,402 samples which represent the majority of blood specimens from adults that need to be screened for thalassaemia. On the other hand, the data set for the clinical trial contains 8,054 samples and is at least eight times larger than those from the previous studies (Wongseree et al., 2007; Piroonratana et al., 2009). The clinical trial data set represents a typical distribution of specimens which are submitted for screening during a fixed time period. Both data sets are collected from Siriraj Hospital, Bangkok, Thailand during 1 January 2007 and 31 December 2008. The data acquisition has been conducted in accordance with the Faculty of Medicine Siriraj Hospital Ethics Committee’s guideline and in accordance with the Helsinki Declaration. In addition, informed consent has been obtained from all individuals. The description of both data sets is summarised in Table 3.2. The samples are made up from seven groups of thalassaemic patients, ten groups of persons with thalassaemia trait, one group of persons with abnormal haemoglobin and one group of normal subjects. Some types of thalassaemia abnormality in the data set for classifier benchmarking are not presented in the specimens collected for the clinical trial. In other words, the classifier benchmarking data set covers more types of thalassaemia abnormality than those presented in the data set for clinical trial. This is carried out to increase the possibility of using the chosen classifier in other data sets without or with minimal additional classifier training.

3.2.2. Attribute Discretisation

The thalassaemia data sets are pre-processed by means of attribute discretisation. The discretisation technique selected for this study is developed by Fayyad and Irani (1993). The technique, which has been successfully applied in an early investigation into thalassaemia classification (Piroonratana et al., 2009), is an information-theoretic technique that employs entropy-based splitting and minimum description length stopping criteria. A chosen cut point within the range of each attribute value is ensured to lie at a boundary between two classes. A new cut point is introduced recursively to each sample subset and is accepted if a significant information gain—the difference between the information values before and after the split—is achieved. The class entropy of a sample set *S*, which consists of samples from $|C|$ classes, is defined as

$$Ent(S) = -\sum_{c \in C} p(c) \log_2 p(c) \tag{3-1}$$

where *p* denotes the probability and *c* is a class. A cut point *T*, which is performed on an attribute *A* of the sample set *S*, creates a partition that has the class information entropy

$$E(A,T,S)=\frac{|S_1|}{|S|}Ent(S_1)+\frac{|S_2|}{|S|}Ent(S_2) \qquad (3\text{-}2)$$

where S_1 and S_2 are the sample subsets of *S* and $S_1+S_2=S$. The cut point *T* is accepted according to the minimum description length stopping criterion if and only if

$$Gain(A,T,S)>\frac{\log_2(|S|-1)}{|S|}+\frac{\log_2(3^{|C|}-2)-[|C|Ent(S)-|C_1|Ent(S_1)-|C_2|Ent(S_2)]}{|S|}$$

(3-3)

where $|C_1|$ and $|C_2|$ are the numbers of classes in the subsets S_1 and S_2, respectively and *Gain*(*A*, *T*, *S*) is the information gain of the cut point, which is defined as

$$Gain(A,T,S)=Ent(S)-\frac{|S_1|}{|S|}Ent(S_1)-\frac{|S_2|}{|S|}Ent(S_2)\,. \qquad (3\text{-}4)$$

3.2.3. Symmetrical Uncertainty

Symmetrical uncertainty is an information-theoretic measure discussed by Press et al. (1988) that defines a correlation between two discrete-valued variables. Consider a sample set in which each sample is described by *m* discrete-valued attributes $A_1,...,A_m$. The entropy of an attribute A_i before and after observing an attribute A_j is respectively given by

$$H(A_i)=-\sum_{a_i\in A_i}p(a_i)\log_2 p(a_i) \qquad (3\text{-}5)$$

and

$$H(A_i|A_j)=-\sum_{a_j\in A_j}p(a_j)\sum_{a_i\in A_i}p(a_i|a_j)\log_2 p(a_i|a_j) \qquad (3\text{-}6)$$

where a_i is a value of the attribute A_i and a_j is a value of the attribute A_j. The degree of correlation between attributes A_i and A_j can be estimated via symmetrical uncertainty (SU) which is defined by

$$\begin{aligned} SU &= 2\times\left[\frac{H(A_i)-H(A_i|A_j)}{H(A_i)+H(A_j)}\right] \\ &= 2\times\left[\frac{H(A_j)-H(A_j|A_i)}{H(A_i)+H(A_j)}\right] \\ &= 2\times\left[\frac{H(A_i)+H(A_j)-H(A_i,A_j)}{H(A_i)+H(A_j)}\right] \end{aligned} \quad . \qquad (3\text{-}7)$$

The value range of symmetrical uncertainty is [0, 1]. An *SU* value close to zero indicates a weak correlation while an *SU* value close to one indicates a strong correlation (Pierrakos and Paliouras, 2010).

3.2.4. C4.5 Decision Tree

A C4.5 decision tree is one of the most widely used inductive inference tools (Quinlan, 1993). The tree is generally constructed in a top-down manner. The construction begins at the root node where each attribute is evaluated using a statistical test to determine how well it can classify the training samples. The best attribute is chosen as the test at the root node of the tree. A descendant of the root node is then created for either each possible value of this attribute if it is a discrete-valued attribute or each possible discretised interval of this attribute if it is a continuous-valued attribute. Next, the training samples are sorted to the appropriate descendant node. The process is repeated using the training samples associated with each descendant node to select the best attribute for testing at that point in the tree. This forms a greedy search for a decision tree, in which the algorithm never backtracks to reconsider earlier node choices. Although it is possible to add a new node to the tree until all samples that are assigned to one node belong to the same class, the tree is not allowed to grow to its maximum depth. A node is introduced to the tree only when there are a sufficient number of samples left from sorting. After the complete tree is constructed, a tree pruning is usually carried out to avoid data over-fitting.

A statistical test used in C4.5 for assigning an attribute to each node in the tree also employs an entropy-based measure. The assigned attribute is the one with the highest information gain ratio among attributes available at that tree construction point. The information gain ratio *GainRatio*(*A*, *S*) of an attribute *A* relative to the sample set *S* is defined as

$$GainRatio(A,S) = \frac{Gain(A,S)}{SplitInformation(A,S)} \tag{3-8}$$

where $$Gain(A,S) = Ent(S) - \sum_{a \in A} \frac{|S_a|}{|S|} Ent(S_a) \tag{3-9}$$

and $$SplitInformation(A,S) = -\sum_{a \in A} \frac{|S_a|}{|S|} \log_2 \frac{|S_a|}{|S|}. \tag{3-10}$$

S_a is the subset of *S* for which the attribute *A* has the value *a*. Obviously, the information gain ratio can be calculated straightaway for discrete-valued attributes. In contrast, continuous-valued attributes are needed to be discretised prior to the information gain ratio calculation.

3.2.5. Naïve Bayes Classifier

A naïve Bayes classifier is a classification system in which the class prediction is based on the application of Bayes theorem (Mitchell, 1997). Consider a set of training samples where each sample is made up from *m* discrete-valued attributes and a class from a finite set *C*. The naïve Bayes classifier can probabilistically predict the class of an unknown sample using the available training sample set to calculate the most probable output. The most probable class c_{NB} of an unknown sample with the conjunction a_1, a_2, …, a_m is given by

$$c_{NB} = \arg\max_{c \in C} p(c | a_1, a_2, ..., a_m) . \tag{3-11}$$

With the use of Bayes theorem, this expression can be rewritten as

$$\begin{aligned} c_{NB} &= \arg\max_{c \in C} \frac{p(a_1, a_2, ..., a_m | c) p(c)}{p(a_1, a_2, ..., a_m)} \quad . \\ &= \arg\max_{c \in C} p(a_1, a_2, ..., a_m | c) p(c) \end{aligned} \tag{3-12}$$

The naïve Bayes classifier functions by assuming that the attributes are conditionally independent given the class. In other words, given the class of the sample the probability of observing the conjunction a_1, a_2, …, a_m is the product of the probability of observing each attribute:

$$p(a_1, a_2, ..., a_m | c) = \prod_i p(a_i | c) . \tag{3-13}$$

Substituting equation (3-13) into equation (3-12), the most probable class as predicted by the naïve Bayes classifier is

$$c_{NB} = \arg\max_{c \in C} p(c) \prod_i p(a_i | c) . \tag{3-14}$$

A Laplace estimate (Cestnik, 1990) is used to calculate $p(a_i | c)$, that is

$$p(a_i | c) = \frac{\left| S_{a_i|c} \right| + 1}{|S_c| + |A_i|} \tag{3-15}$$

where $\left| S_{a_i|c} \right|$ is the number of samples from the class c in which the ith attribute (A_i) has the value a_i, $|S_c|$ is the number of samples from the class c and $|A_i|$ is the number of possible values for the attribute A_i.

3.2.6. Multilayer Perceptron

A neural network is an interconnected group of artificial neurons that uses a computational model for information processing. The neural network selected for this study is a multilayer perceptron (Rumelhart and McClelland, 1986). The model of a neuron shown in Figure 3.3(a) indicates that *q* input signals are received by the neuron. These inputs are weighted and summed together. The threshold, which is treated as an extra connection weight, is then applied to the weighted-sum result. Thus, the linear combiner output (*z*) or input to the activation function is given by

$$z = \sum_i w_i u_i \tag{3-16}$$

where u_i is the *i*th input to the neuron and w_i is the connection weight for the for the input u_i. In addition, $u_0 = -1$ and w_0 is the threshold. The neuron output ($h(z)$) is the output from the activation function and is denoted by

$$h(z) = \frac{1}{1 + \exp(-z)} . \tag{3-17}$$

As a result, the output signal from each neuron is limited by a logistic sigmoid function. The neuron model described above is used throughout the multilayer feed-forward network illustrated in Figure 3.3(b). The multilayer perceptron is implemented in this manner because the early studies indicate that a multilayer perceptron of this form works well with both CBC (Wongseree et al., 2007) and haemoglobin typing inputs (Piroonratana et al., 2009).

Since the multilayer perceptron is used as a classifier, the number of network inputs is equal to the number of attributes while the number of network outputs is equal to the number of classes. Each output neuron represents one of possible classes with the highest-valued output taken as the network prediction. This is often referred to as a 1-of-*n* output encoding technique (Mitchell, 1997). In this study, the network has a single hidden layer, which is a layer of neurons that receive attributes as inputs and send signals to output neurons. The number of neurons in the hidden layer is a design parameter and can generally be set using a heuristic rule. The number of hidden nodes is set to $\lfloor$(number of attributes + number of classes)/2$\rfloor$ in this study (Witten and Frank, 2005).

The multilayer perceptron is trained by a back-propagation algorithm. The algorithm employs a sample-by-sample updating rule for adjusting connection weights. A training sample is presented to the network during the iteration. The signal is fed in a forward manner through the network until the network output is obtained. The error between actual and target network outputs is then calculated and used to adjust the connection weights. The adjustment procedure, which is based on a gradient descent method, is first applied to connection weights in the output layer. Next, connection weights in the hidden layer are adjusted. The iteration is completed when all connection weights have been adjusted.

3.3. Results and Discussions

3.3.1. Attribute Discretisation

The original CBC and haemoglobin typing data contains eight attributes as given in Table 3.1. The attributes in classifier evaluation samples are discretised using the information-theoretic technique (Fayyad and Irani, 1993) described earlier. The discrete intervals of each attribute are illustrated in Table 3.3. The discrete intervals for attributes 3, 4, 6 and 8, which are haemoglobin typing attributes, are similar to those reported in Piroonratana et al. (2009) while more discrete intervals for attributes 5 and 7 are introduced. This implies that the discrete intervals for some attributes from the previous study by Piroonratana et al. (2009) are sufficient for the classification task in the current study. However, new discrete intervals for the other attributes are also required to accommodate more types of thalassaemia. Furthermore, a cut point at MCV = 74.95 fL is introduced to attribute 2. This conforms to the expert rule explained earlier for the differentiation between normal subjects and persons with α-thalassaemia 1 trait.

3.3.2. Classifier Performance Evaluation

Three candidate classifiers—a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron—are benchmarked using the classifier evaluation data in stratified 10-fold cross-validation experiments (Kohavi, 1995). The training of a multilayer perceptron is terminated prior to the occurrence of data over-fitting. The suitable number of training epochs is chosen using a training and validation approach (Mitchell, 1997). Basically, a validation data set is employed to detect the training epoch where the training error continues to decrease while the validation error begins to increase. This is the point where data over-fitting occurs. In this study, 10% of data samples are used as the validation samples. In contrast to a multilayer perceptron, C4.5 has a built-in mechanism for data over-fitting avoidance. This is achieved via a rule post-pruning strategy (Quinlan, 1993).

The classification performance of C4.5, a naïve Bayes classifier and a multilayer perceptron on the classifier evaluation data obtained from stratified 10-fold cross-validation is summarised in Table 3.4. The results from each classifier are tallied from 30 runs where new cross-validation folds are generated for each run. The results obtained by excluding CBC attributes from classifier inputs are also given for comparison purposes. The exclusion of haemoglobin typing attributes from classifier inputs lead to unacceptably low classification accuracy; these results are not shown. It is noticeable that both naïve Bayes classifier and multilayer perceptron have the highest classification accuracy. Moreover, the use of six haemoglobin typing attributes in conjunction with the MCV attribute as classifier inputs is proven to be sufficient while the HB attribute appears to be a redundant attribute. The symmetrical uncertainty analysis of seven attributes necessary for the classification in Table 3.5 reveals that only the Hb A_0, Hb E and Hb A_2 attributes are moderately correlated among one another while the remaining attributes are uncorrelated.

Among three classifiers, it can be clearly seen that both naïve Bayes classifier and multilayer perceptron have higher classification performance than C4.5 when all attributes are used and when only HB is not used as an input (*t*-test's *p*-value < 0.0001). Moreover, the performance difference between naïve Bayes classifier and multilayer perceptron in both cases is statistically insignificant ($p > 0.05$). However, when both HB and MCV are excluded from classifier inputs the classification accuracy of C4.5 is significantly better than that of the multilayer perceptron ($p < 0.01$). In the early work by Wongseree et al. (2007), which involves the application of a multilayer perceptron and a genetic programming based decision tree to the classification problem with CBC attributes, the multilayer perceptron is proven to be the best classifier. In contrast, C4.5 is proven to be the best classifier in comparison to a multilayer perceptron and a random forest in the classification problem with haemoglobin typing attributes (Piroonratana et al., 2009). Based on the evidence from the early and present studies, it is possible to deduce that C4.5 is more suitable to the problem when only haemoglobin typing attributes are considered as inputs while a multilayer perceptron is more suitable to the problem that involves CBC attributes.

As mentioned earlier, it is hypothesised that using both CBC and haemoglobin typing attributes as inputs to the classifiers should lead to an improvement in the classification accuracy. The classification performance of each classifier clearly supports this hypothesis where the difference between the classification accuracy obtained using all attributes and that using only haemoglobin typing attributes is statistically significant ($p < 0.0001$). Nonetheless, MCV is the only necessary CBC attribute for the classification. This is deduced from the results which clearly indicate that the difference between the classification accuracy obtained using all attributes and that with the exclusion of HB attribute is not statistically significant ($p > 0.05$). In addition, the removal of MCV from the classifier inputs also leads to a significant degradation of classification performance ($p < 0.0001$). Further analysis of the classification accuracy reveals that the increase in classification error mainly stems from an inability to differentiate between normal, α-thalassaemia 1 trait and α-thalassaemia 2 trait samples. This conforms to the early explanation regarding the limitation of using haemoglobin typing attributes as the sole inputs for the classification task involving these three classes (Old, 2003; Piroonratana et al., 2009). Although the early studies by Amendolia et al. (2003) and Wongseree et al. (2007) suggest that both HB and MCV are informative CBC attributes, it is most likely that the use of haemoglobin typing attributes is sufficient to satisfy the need for using the HB attribute.

It is observed that the redundancy of the HB attribute can only be identified via the inspection of the variation in classification accuracy after changing the attribute combination, which is a simple form of wrapper attribute selection (Kohavi and John, 1997). An attempt to perform attribute selection by other techniques including a correlation-based feature selection technique (Hall and Holmes, 2003) and a ReliefF technique (Robnik-Šikonja and Kononenko, 2003) fails to identify this redundancy. These techniques are considered because the attributes selected by the techniques are not transformed, which makes the clinical interpretation of results a straightforward task. Moreover, the correlation-based feature selection technique is proven to be capable of identifying informative haemoglobin typing attributes in the early study by Piroonratana et al. (2009).

Detailed classification performance of all three classifiers obtained without the HB input is subsequently analysed. The results from an example run of the stratified 10-fold cross-validation given in Tables 3.6-3.8 indicate that the classification accuracy of each classifier for the classes with a small number of samples is low. Nonetheless, most of the misclassified samples from these classes are identified as samples from classes which are closely related to the true classes. For instance, mixed Hb E and Hb Constant Spring trait samples as well as mixed Hb E and abnormal haemoglobin samples are classified as Hb E trait, mixed α-thalassaemia 1 and Hb E trait and homozygous Hb E samples. Similarly, samples from persons with Hb H-Constant Spring disease are misclassified as homozygous Hb Constant Spring samples and vice versa. The results from other runs and from stratified 10-fold cross-validation with other settings for attribute combination also have a similar trend and hence are not shown.

Based on the above discussion of classification results, it can be concluded that the necessary attributes for this study are six haemoglobin typing attributes and the MCV attribute while the suitable classifiers for use with these attributes are a naïve Bayes classifier and a multilayer perceptron. Moreover, a multilayer perceptron is equally suitable to a naïve Bayes classifier in terms of the actual implementation for clinical trials. This is because the storage of trained connection weights for a multilayer perceptron requires a similar amount of space to that of probability values for a naïve Bayes classifier. In lieu of these reasons, a naïve Bayes classifier and a multilayer perceptron are the chosen classifiers for a clinical trial.

3.3.3. Clinical Trial

In the previous sub-section, both naïve Bayes classifier and multilayer perceptron with six haemoglobin attributes and the MCV attribute are proven to be the best approach for thalassaemia classification. The naïve Bayes classifier and the multilayer perceptron are subsequently used in a clinical trial involving 8,054 samples. The distribution of classes within the clinical trial data set has been given in Table 3.2. A naïve Bayes classifier and a multilayer perceptron, which are trained with classifier evaluation samples, are applied to the clinical trial data set where the classification accuracy of 99.39% and 99.71% are respectively achieved. Detailed classification performance of both classifiers is given in Tables 3.9-3.10. It is noticeable that nearly all misclassified samples from both classifiers stem from mixed α-thalassaemia 1 and Hb E trait samples. They are classified as Hb E trait samples, which belong to the class that are closely related to the true class. This indicates that both naïve Bayes classifier and multilayer perceptron are highly suitable to the present classification problem.

3.4. Conclusions

In this report, a thalassaemia classification problem in Thailand is investigated. The aim is to identify whether the human subject is a person with abnormal haemoglobin, a person with thalassaemia trait, a thalassaemic patient or a normal person using complete blood count (CBC) and haemoglobin typing data. The data sets contain eight attributes and 19 classes. The first two attributes are CBC attributes: a haemoglobin concentration (HB) and a mean corpuscular volume (MCV). On the other hand, the last six attributes reflect the percentages of haemoglobin at a specific range of retention time. In other words, these attributes represent different types of haemoglobin. The study is divided into two main parts: classifier and attribute subset selection and a clinical trial. Candidate classifiers for the task include a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron. The experiment involving stratified 10-fold cross-validation reveals that both naïve Bayes classifier and multilayer perceptron are the most suitable classifier for the data that has been pre-processed by attribute discretisation. The experiment results also suggest that using both CBC and haemoglobin typing attributes as classifier inputs can significantly improve the classification accuracy over that achieved using only haemoglobin typing attributes. Furthermore, the experiment results indicate that HB is a redundant attribute for this study. The naïve Bayes classifier and the multilayer perceptron are subsequently applied to an additional data set in a clinical trial. The analysis of classification errors from the trial indicates that most of the misclassified samples are identified as samples from classes which are closely related to the true classes. This helps emphasise the suitability of a naïve Bayes classifier and a multilayer perceptron as an automated thalassaemic classification tool.

Table 3.1 Input features or attributes for thalassaemia classification. The first two attributes are CBC attributes while the last six attributes represent different types of haemoglobin.

Attribute	Attribute name	Description (measurement unit)
1	HB	Haemoglobin concentration (gram/decilitre, g/dL)
2	MCV	Mean corpuscular volume (femtolitre, fL)
3	Hb Bart's	Percentage of haemoglobin at retention time 0-68 sec (%)
4	Hb A_{1C}/Hb F	Percentage of haemoglobin at retention time 69-160 sec (%)
5	Hb A_0	Percentage of haemoglobin at retention time 200-230 sec (%)
6	Hb E	Percentage of haemoglobin at retention time 250-280 sec (%)
7	Hb A_2	Percentage of haemoglobin at retention time 281-289 sec (%)
8	Hb D/Hb S/ Hb Constant Spring/Hb C	Percentage of haemoglobin at retention time 290-320 sec (%)

Table 3.2 Two data sets for thalassaemia classification. The first set contains 1,402 samples and is used for classifier benchmarking. The second set consists of 8,054 samples which are collected for a clinical trial.

Class	Description	Category	Number of samples Classifier Bench-marking	 Clinical trial
1	Normal subject	Normal	78	436
2	α-Thalassaemia 2 trait	Trait	150	978
3	Hb Constant Spring trait	Trait	23	10
4	Hb E trait	Trait	548	4,292
5	Hb E trait + Hb Constrant Spring trait	Trait	6	1
6	Hb E trait + abnormal haemoglobin	Trait	5	0
7	α-Thalassaemia 1 trait	Trait	180	1,008
8	α-Thalassaemia 1 trait + Hb E trait	Trait	116	431
9	Homozygous Hb E	Trait	105	571
10	β-Thalassaemia trait	Trait	74	267
11	HPFH	Trait	9	1
12	Abnormal haemoglobin	N/A	11	1
13	Hb H disease	Disease	39	27
14	Hb H-Constant Spring disease	Disease	5	0
15	Homozygous Hb Constant Spring	Disease	8	0
16	EA Bart's disease	Disease	11	1
17	β^0-Thalassaemia/Hb E	Disease	15	22
18	β^+-Thalassaemia/Hb E	Disease	9	3
19	Homozygous β-thalassaemia	Disease	10	5
	Total		1,402	8,054

Table 3.3 Discrete intervals of the attributes.

Attribute	Attribute name	Interval	Range
1	HB	1	HB ≤ 8.65
		2	8.65 < HB ≤ 10.65
		3	10.65 < HB ≤ 14.55
		4	HB > 14.55
2	MCV	1	MCV ≤ 56.45
		2	56.45 < MCV ≤ 61.35
		3	61.35 < MCV ≤ 64.05
		4	64.05 < MCV ≤ 69.95
		5	69.95 < MCV ≤ 72.15
		6	72.15 < MCV ≤ 74.95
		7	74.95 < MCV ≤ 78.75
		8	78.75 < MCV ≤ 79.95
		9	79.95 < MCV ≤ 84.05
		10	84.05 < MCV ≤ 87.45
		11	MCV > 87.45

Table 3.3 Discrete intervals of the attributes (cont.).

Attribute	Attribute name	Interval	Range
3	Hb Bart's	1	% of Hb ≤ 1.70
		2	% of Hb > 1.70
4	Hb A_{1C}/Hb F	1	% of Hb ≤ 0.58
		2	0.58 < % of Hb ≤ 3.35
		3	3.35 < % of Hb ≤ 10.85
		4	10.85 < % of Hb ≤ 74.35
		5	74.35 < % of Hb ≤ 83.60
		6	% of Hb > 83.60
5	Hb A_0	1	% of Hb ≤ 1.05
		2	1.05 < % of Hb ≤ 49.10
		3	49.10 < % of Hb ≤ 52.75
		4	52.75 < % of Hb ≤ 62.85
		5	62.85 < % of Hb ≤ 65.15
		6	65.15 < % of Hb ≤ 67.65
		7	67.65 < % of Hb ≤ 72.35
		8	72.35 < % of Hb ≤ 77.05
		9	77.05 < % of Hb ≤ 83.25
		10	83.25 < % of Hb ≤ 83.35
		11	% of Hb > 83.35
6	Hb E	1	% of Hb ≤ 7.70
		2	7.70 < % of Hb ≤ 18.65
		3	18.65 < % of Hb ≤ 24.95
		4	24.95 < % of Hb ≤ 40.75
		5	40.75 < % of Hb ≤ 59.80
		6	59.80 < % of Hb ≤ 75.80
		7	% of Hb > 75.80
7	Hb A_2	1	% of Hb ≤ 0.05
		2	0.05 < % of Hb ≤ 1.05
		3	1.05 < % of Hb ≤ 2.15
		4	2.15 < % of Hb ≤ 2.55
		5	2.55 < % of Hb ≤ 3.55
		6	3.55 < % of Hb ≤ 11.40
		7	11.40 < % of Hb ≤ 20.10
		8	% of Hb > 20.10
8	Hb D/Hb S/ Hb Constant Spring/Hb C	1	% of Hb ≤ 0.05
		2	% of Hb > 0.05

Table 3.4 Summarised classification performance of the C4.5 decision tree, naïve Bayes classifier and multilayer perceptron. The results are averaged over 30 runs of stratified 10-fold cross-validation involving 1,402 samples. The numbers in the brackets are standard deviations.

	Classification accuracy (%)		
Attribute set	C4.5 decision tree	Naïve Bayes classifier	Multilayer perceptron
Complete	89.88 (1.74)	92.77 (1.74)	92.34 (1.69)
Without HB	89.30 (1.82)	93.23 (1.67)	92.60 (1.75)
Without MCV	79.80 (2.58)	80.49 (2.34)	77.04 (2.65)
Without both HB and MCV	77.98 (2.23)	79.27 (2.13)	76.31 (2.44)

Table 3.5 A symmetrical uncertainty (SU) analysis of six haemoglobin typing attributes and the MCV attribute. An *SU* value close to zero denotes a weak correlation while an *SU* value close to one denotes a strong correlation. The description of each attribute has been given in Table 3.1.

Attribute	2	3	4	5	6	7	8
2	1.0000	0.0202	0.0196	0.0992	0.1360	0.0430	0.0031
3		1.0000	0.0043	0.0298	0.0543	0.0713	0.0606
4			1.0000	0.0921	0.0870	0.0473	0.0053
5				1.0000	0.4960	0.3130	0.0157
6					1.0000	0.4950	0.0231
7						1.0000	0.0202
8							1.0000

Table 3.6 Detailed classification performance of the C4.5 decision tree from one run of stratified 10-fold cross-validation without the HB attribute. The classification accuracy is 89.90%, which sufficiently represents the average accuracy of 89.30% (standard deviation = 1.82%) obtained from 30 runs. The description of each class has been given in Table 3.2.

Actual	Identified class (number of samples)																			Class
class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	acc. (%)
1	78																			100.00
2		146	1				3													97.33
3	1	4	13				5													56.52
4	1	3		538			1	4	1											98.18
5				2	1			3												16.67
6				3				1									1			0.00
7		3		2			169	1		2			3							93.89
8					1			115												99.14
9									103								2			98.10
10	8	7					12			44	3									59.46
11	3	1					1			1	1	2								11.11
12	1	2					4	1		1	1							1		0.00
13	1	2					12						24							61.54
14			2				2						1							0.00
15	1		5										2							0.00
16						1	1	1								8				72.73
17				3					1								11			73.33
18								2										7		77.78
19	4	1								4									1	10.00

Table 3.7 Detailed classification performance of the naïve Bayes classifier from one run of stratified 10-fold cross-validation without the HB attribute. The classification accuracy is 93.37%, which sufficiently represents the average accuracy of 93.23% (standard deviation = 1.67%) obtained from 30 runs. The description of each class has been given in Table 3.2.

Actual	Identified class (number of samples)																			Class
class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	acc. (%)
1	74									2	2									94.87
2		145					3			1			1							96.67
3		2	21																	91.30
4	1	1		540				3	1			2								98.54
5				2				4												0.00
6				2				1	2											0.00
7		3		1			169	1		4			2							93.89
8				6				110												94.83
9									104								1			99.05
10	3	2					1			63	5									85.14
11	3						1			1	4									44.44
12	1	1		2			2	1			2						1	1		0.00
13			1				5						33							84.62
14													1	2	2					40.00
15			3											2	3					37.50
16								1								10				90.91
17									1								14			93.33
18								2										7		77.78
19																			10	100.00

Table 3.8 Detailed classification performance of the multilayer perceptron from one run of stratified 10-fold cross-validation without the HB attribute. The classification accuracy is 92.65%, which sufficiently represents the average accuracy of 92.60% (standard deviation = 1.75%) obtained from 30 runs. The description of each class has been given in Table 3.2.

Actual	Identified class (number of samples)																			Class
class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	acc. (%)
1	73			1						3	1									93.59
2		142	1				4			1		1	1							94.67
3		1	20												2					86.96
4	1			538		1		4	1	1		2								98.18
5			1	1	2			2												33.33
6				2				1	1			1								0.00
7		3		2			166	2		2	1	1	3							92.22
8				1	2			111								1		1		95.69
9									103								2			98.10
10	1	2					3			63	5									85.14
11										4	3	2								33.33
12	1	1		1			2	1		1	1	1					1	1		9.09
13		1					3						34		1					87.18
14		1											2		2					0.00
15			4												4					50.00
16								1								10				90.91
17									2			1					12			80.00
18								1										8		88.89
19										1									9	90.00

Table 3.9 Detailed classification performance of the naïve Bayes classifier from the clinical trial involving 8,054 samples. The classification accuracy is 99.39%. The description of each class has been given in Table 3.2.

Actual	Identified class (number of samples)																			Class
class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	acc. (%)
1	436																			100.00
2		977					1													99.90
3			10																	100.00
4				4291					1											99.98
5				1																0.00
6																				
7							1008													100.00
8				37				394												91.42
9									570								1			99.82
10		7								260										97.38
11											1									100.00
12																	1			0.00
13													27							100.00
14																				
15																				
16																1				100.00
17																	22			100.00
18																		3		100.00
19																			5	100.00

Table 3.10 Detailed classification performance of the multilayer perceptron from the clinical trial involving 8,054 samples. The classification accuracy is 99.71%. The description of each class has been given in Table 3.2.

Actual	Identified class (number of samples)																			Class
class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	acc. (%)
1	436																			100.00
2		976					1			1										99.80
3			10																	100.00
4				4289		2			1											99.93
5		1																		0.00
6																				
7	1						1007													99.90
8				13				418												96.98
9									571											100.00
10										267										100.00
11											1									100.00
12												1								100.00
13													27							100.00
14																				
15																				
16																1				100.00
17									3								19			86.36
18																		3		100.00
19																			5	100.00

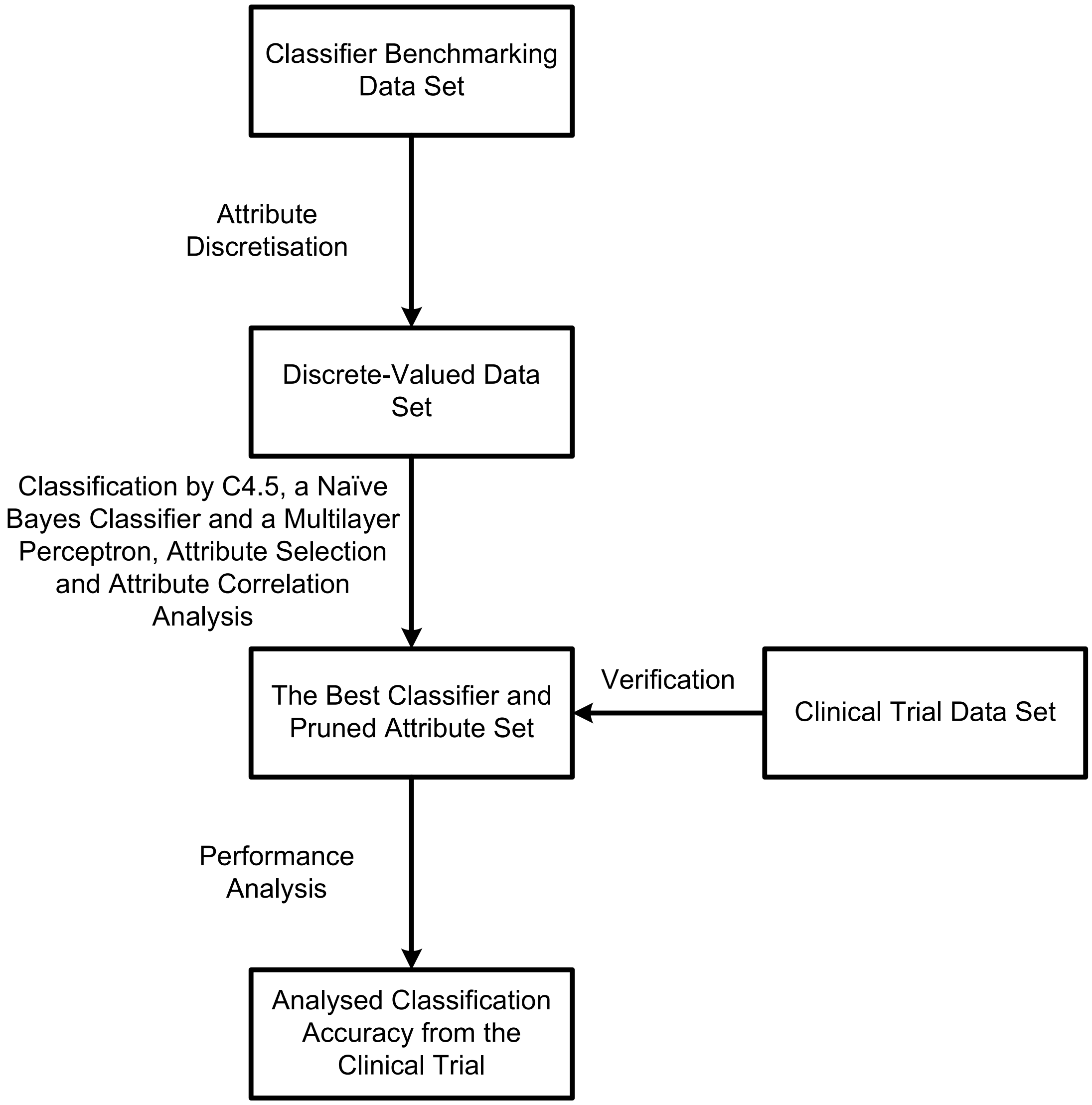
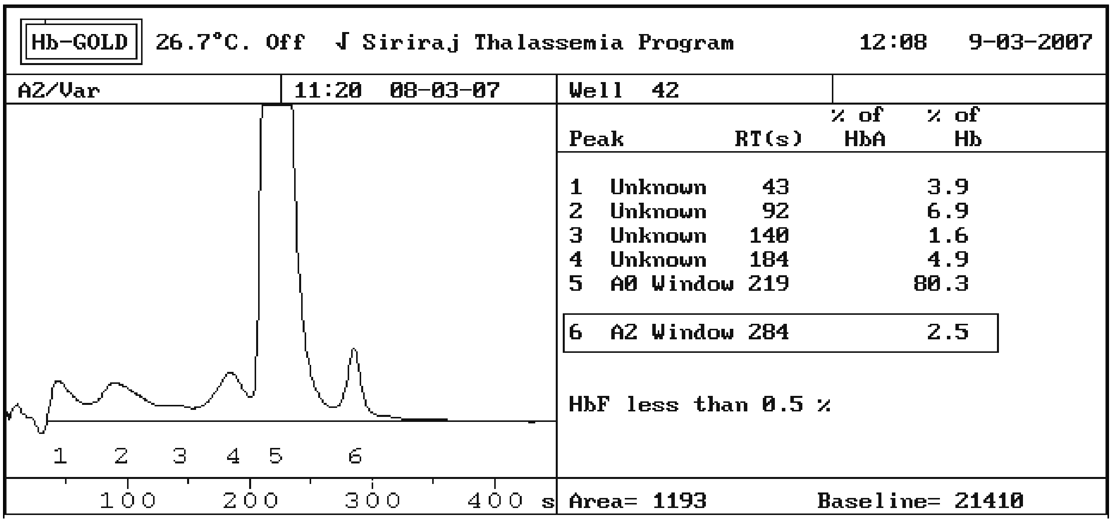
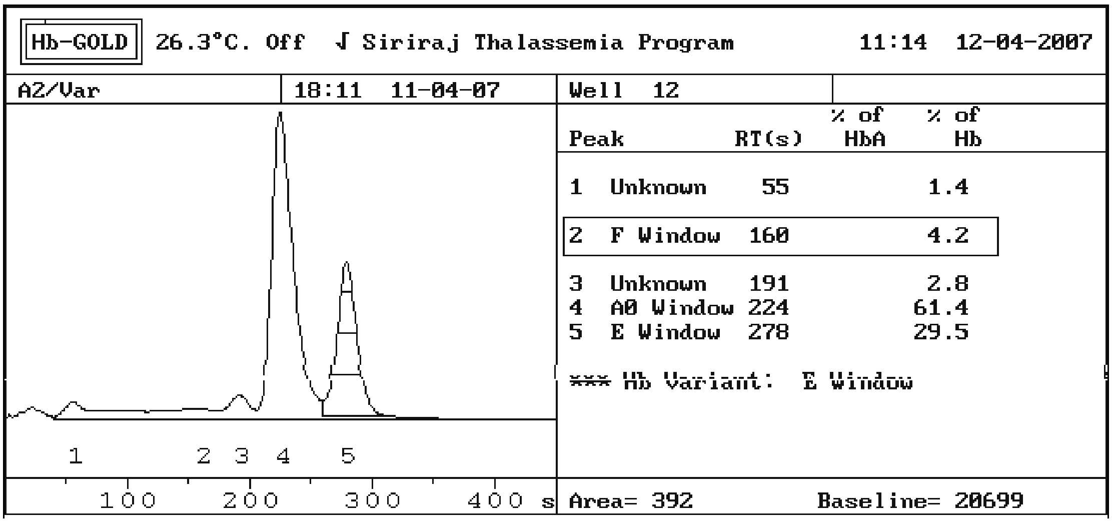

Classifier Benchmarking Data Set
Attribute Discretisation
Discrete-Valued Data Set
Classification by C4.5, a Naïve Bayes Classifier and a Multilayer Perceptron, Attribute Selection and Attribute Correlation Analysis
The Best Classifier and Pruned Attribute Set
Verification
Clinical Trial Data Set
Performance Analysis
Analysed Classification Accuracy from the Clinical Trial

Figure 3.1 Schematic diagram for the methodology employed in the study.



```
Hb-GOLD  26.7°C. Off  √ Siriraj Thalassemia Program        12:08   9-03-2007
A2/Var              11:20  08-03-07    Well  42
                                                      % of    % of
                                       Peak        RT(s)   HbA      Hb

                                       1  Unknown    43             3.9
                                       2  Unknown    92             6.9
                                       3  Unknown   140             1.6
                                       4  Unknown   184             4.9
                                       5  A0 Window 219            80.3

                                       6  A2 Window 284             2.5

                                       HbF less than 0.5 %

                                       Area= 1193          Baseline= 21410
```

(a)



```
Hb-GOLD  26.3°C. Off  √ Siriraj Thalassemia Program        11:14  12-04-2007
A2/Var              18:11  11-04-07    Well  12
                                                      % of    % of
                                       Peak        RT(s)   HbA      Hb

                                       1  Unknown    55             1.4

                                       2  F Window  160             4.2

                                       3  Unknown   191             2.8
                                       4  A0 Window 224            61.4
                                       5  E Window  278            29.5

                                       *** Hb Variant:  E Window

                                       Area= 392           Baseline= 20699
```

(b)

Figure 3.2 Elution chromatograms of (a) either a normal specimen or a specimen from a person with α-thalassaemia 1 trait depending on the MCV value and (b) a specimen from a person with Hb E trait that are obtained from an Hb Gold HPLC system. RT(s) represents the retention time in seconds for each fraction of elute. % of Hb represents the percentage of haemoglobin in the elution peak.

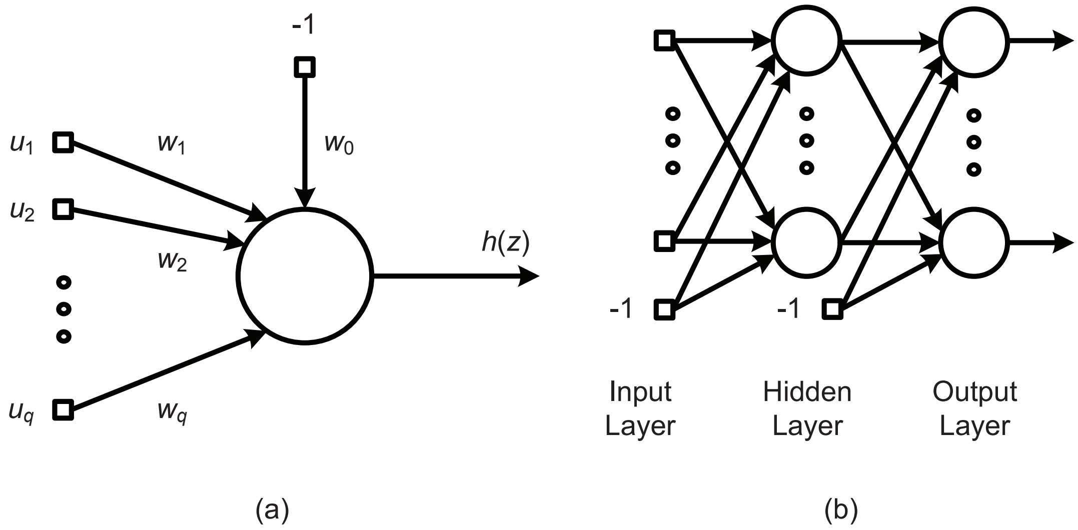


Figure 3.3 Schematic diagram of a multilayer perceptron: (a) computational model of a neuron and (b) feed-forward network with one hidden layer.

4. Conclusions

In this report, two genetic disease problems have been tackled using pattern recognition. The first problem involves the identification of single nucleotide polymorphisms (SNPs) that are useful for genetic association study of type 2 diabetes mellitus (T2D). The problem is difficult because the number of discrete-valued attributes is large while the sample size is small. As a result, the problem can be treated as an attribute selection problem. An omnibus permutation test on ensembles of two-locus analyses has been developed for the task. The technique is proven to be a vital part of the SNP extraction procedure. The findings provide an alternative explanation for the aetiology of T2D.

The second problem of interest is a thalassaemia classification problem. In contrast to the first problem, this problem involves continuous-valued attributes. The procedure for solving the problem hence begins with information-theoretic attribute discretisation (Fayyad and Irani, 1993). Then informative attributes are identified via wrapper attribute selection (Kohavi and John, 1997). Finally, the classification models are constructed using a naïve Bayes classifier (Mitchell, 1997) and a multilayer perceptron (Rumelhart and McClelland, 1986). The result indicates that significantly high classification accuracy can be achieved. Overall, the results from both problems suggest that pattern recognition techniques are highly efficient and are proven to be useful for problems in genetics.

Output from the Project

The research results have been published in three international journal articles; details of these articles follow.

1. Wongseree, W., Assawamakin, A., Piroonratana, T., Sinsomros, S., Limwongse, C. and Chaiyaratana, N. (2009). Detecting purely epistatic multi-locus interactions by an omnibus permutation test on ensembles of two-locus analyses. *BMC Bioinformatics*, *10*, 294. (2009 Journal Impact Factor = 3.428)
2. Piroonratana, T., Wongseree, W., Assawamakin, A., Paulkhaolarn, N., Kanjanakorn, C., Sirikong, M., Thongnoppakhun, W., Limwongse, C. and Chaiyaratana, N. (2009). Classification of haemoglobin typing chromatograms by neural networks and decision trees for thalassaemia screening. *Chemometrics and Intelligent Laboratory Systems*, *99*, 101-110. (2009 Journal Impact Factor = 2.111)
3. Setsirichok, D., Piroonratana, T., Wongseree, W., Usavanarong, T., Paulkhaolarn, N., Kanjanakorn, C., Sirikong, M., Limwongse, C. and Chaiyaratana, N. (2011). Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening. *Biomedical Signal Processing and Control*, in press. (2009 Journal Impact Factor = 0.620)

Appendix

Publication of the Research Results

A.1. BMC Bioinformatics

Wongseree, W., Assawamakin, A., Piroonratana, T., Sinsomros, S., Limwongse, C. and Chaiyaratana, N. (2009). Detecting purely epistatic multi-locus interactions by an omnibus permutation test on ensembles of two-locus analyses. *BMC Bioinformatics*, *10*, 294. (2009 Journal Impact Factor = 3.428)

BMC Bioinformatics is an open access journal. All articles published in the journal are publicly available in both html and pdf formats. The above article can be accessed at www.biomedcentral.com/1471-2105/10/294. Updated information regarding a published article can be found in the comment section of the article at the publisher repository. The latest comment about the above article, which was published on 7 April 2011, is also included in the appendix.

BMC Bioinformatics IMPACT FACTOR 3.43

Log on/register
Feedback | Support | My details
home | journals A-Z | subject areas | advanced search | authors | reviewers | libraries | about | my BioMed Central

Methodology article

Open Access

Detecting purely epistatic multi-locus interactions by an omnibus permutation test on ensembles of two-locus analyses

Waranyu Wongseree[1], Anunchai Assawamakin[2], Theera Piroonratana[1], Saravudh Sinsomros[1], Chanin Limwongse[2] and Nachol Chaiyaratana[1,2]

1 Department of Electrical Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok, 1518 Piboolsongkram Road, Bangsue, Bangkok 10800, Thailand
2 Division of Molecular Genetics, Department of Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University, 2 Prannok Road, Bangkoknoi, Bangkok 10700, Thailand

author email corresponding author email

BMC Bioinformatics 2009, 10:294 doi:10.1186/1471-2105-10-294

The electronic version of this article is the complete one and can be found online at: http://www.biomedcentral.com/1471-2105/10/294

Received: 5 March 2009
Accepted: 17 September 2009
Published: 17 September 2009

BMC Bioinformatics
Volume 10

Viewing options:
Abstract
Full text
PDF (1.6MB)
Additional files

Associated material:
Readers' comments (1)
PubMed record

Related literature:
Articles citing this article
on Google Scholar
on ISI Web of Science
on PubMed Central
Other articles by authors
on Google Scholar
on PubMed
Related articles/pages
on Google
on Google Scholar
on PubMed

Tools:
Download citation(s)
Download XML
Email to a friend
Order reprints
Post a comment

Post to:
Citeulike
Connotea
Del.icio.us
Facebook
Mendeley
Twitter

Abstract

Background

Purely epistatic multi-locus interactions cannot generally be detected via single-locus analysis in case-control studies of complex diseases. Recently, many two-locus and multi-locus analysis techniques have been shown to be promising for the epistasis detection. However, exhaustive multi-locus analysis requires prohibitively large computational efforts when problems involve large-scale or genome-wide data. Furthermore, there is no explicit proof that a combination of multiple two-locus analyses can lead to the correct identification of multi-locus interactions.

Results

The proposed 2LOmb algorithm performs an omnibus permutation test on ensembles of two-locus analyses. The algorithm consists of four main steps: two-locus analysis, a permutation test, global *p*-value determination and a progressive search for the best ensemble. 2LOmb is benchmarked against an exhaustive two-locus analysis technique, a set association approach, a correlation-based feature selection (CFS) technique and a tuned ReliefF (TuRF) technique. The simulation results indicate that 2LOmb produces a low false-positive error. Moreover, 2LOmb has the best performance in terms of an ability to identify all causative single nucleotide polymorphisms (SNPs) and a low number of output SNPs in purely epistatic two-, three- and four-locus interaction problems. The interaction models constructed from the 2LOmb outputs via a multifactor dimensionality reduction (MDR) method are also included for the confirmation of epistasis detection. 2LOmb is subsequently applied to a type 2 diabetes mellitus (T2D) data set, which is obtained as a part of the UK genome-wide genetic epidemiology study by the Wellcome Trust Case Control Consortium (WTCCC). After primarily screening for SNPs that locate within or near 372 candidate genes and exhibit no marginal single-locus effects, the T2D data set is reduced to 7,065 SNPs from 370 genes. The 2LOmb search in the reduced T2D data reveals that four intronic SNPs in *PGM1* (phosphoglucomutase 1), two intronic SNPs in *LMX1A* (LIM homeobox transcription factor 1, alpha), two intronic SNPs in *PARK2* (Parkinson disease (autosomal recessive, juvenile) 2, parkin) and three intronic SNPs in *GYS2* (glycogen synthase 2 (liver)) are associated with the disease. The 2LOmb result suggests that there is no interaction between each pair of the identified genes that can be described by purely epistatic two-locus interaction models. Moreover, there are no interactions between these four genes that can be described by purely epistatic multi-locus interaction models with marginal two-locus effects. The findings provide an alternative explanation for the aetiology of T2D in a UK population.

Conclusion

An omnibus permutation test on ensembles of two-locus analyses can detect purely epistatic multi-locus interactions with marginal two-locus effects. The study also reveals that SNPs from large-scale or genome-wide case-control data which are discarded after single-locus analysis detects no association can still be useful for genetic epidemiology studies.

Background

Complex diseases cannot generally be explained by Mendelian inheritance [1] because they are influenced by gene-gene and gene-environment interactions. Many common diseases such as asthma, cancer, diabetes, hypertension and obesity are widely accepted and acknowledged to be results of complex interactions between multiple genetic factors [2]. Attempts to identify factors that could be the causes of complex diseases have led to many genome-wide association studies [3,4]. Raw results from these attempts produce a large amount of single nucleotide polymorphism (SNP) data from every individual participating in the trials.

For genetic epidemiologists, data sets from genome-wide association studies present many challenges, particularly the correct identification of SNPs that associate with the disease of interest from all available SNPs [5]. This challenge can be treated as a pattern recognition problem which aims to identify an attribute or SNP set that can lead to the correct classification of recruited samples. Heidema et al. [5] and Motsinger et al. [6] have reviewed and identified many machine learning techniques that are suitable to the task. Among many strategies and techniques, the protocol that appears to be most promising for genome-wide association studies involves two main steps: SNP set reduction and classification model construction [7]. From a machine learning viewpoint, attribute selection techniques can be divided into three main categories: filter, wrapper and embedded approaches [8]. In a filter approach, a measure or an index is used to determine the correlation between attributes and classes, e.g. affected and unaffected status in a case-control study. Attributes that are deemed to be important for the classification according to the measure are then selected. The filter approach includes χ^2 and odds ratio tests [9,10], omnibus permutation tests [11-13], a correlation-based feature selection technique [14], a ReliefF technique [15] and a tuned ReliefF technique [16]. In a wrapper approach, the significance of an attribute subset is evaluated from the classification performance by a classifier. The capability of the wrapper approach to identify significant attributes thus depends on the chosen classifier and the search algorithm for the identification of the best attribute subset. Combinatorial [17] and restricted partitioning methods [18], a multifactor dimensionality reduction method [19-25] and a polymorphism interaction analysis technique [26] are examples of the wrapper approach. An embedded approach concentrates on informative attributes during the construction of a classification model. Examples of the embedded approach include a genetic algorithm with Boolean algebra [27], genetic programming based decision trees [28,29], random forests [30-32] and evolutionary neural networks [33,34]. Based on this categorisation, classification models are not direct outputs from filter-based techniques. On the other hand, classification models are readily prepared as outputs from the wrapper and embedded approaches. In other words, the last two approaches can also be regarded as classification model construction techniques.

The success of the two-step pattern recognition approach relies heavily on the attribute selection step [14]. In case-control studies, epistatic effects play a vital role in establishing the difficulty level of SNP screening problems [35,36]. Epistasis in the simplest form can be represented by disease models that require genotype inputs from two interacting SNPs [37,38]. Many attempts have been made to produce consistent definitions and categorisation of different types of epistasis models [2,35,39-41]. According to Musani et al. [2], a pure epistasis model [42] is difficult because each SNP exhibits no marginal single-locus effect in the model. As a result, it is impossible to detect the pure epistasis by univariate statistical tests. Examples of complex diseases that case-control studies have uncovered putatively pure epistasis include type 2 diabetes mellitus (T2D) [43-46] and metabolic syndrome [47]. Due to the difficulty of screening for each SNP independently, it is suggested that attention should be focused on the analysis of differences between two-locus genotype distribution within case and control groups [40] and multi-locus Bayesian statistical analysis [48,49].

A number of SNP screening and association detection techniques have adopted the two-locus genotype monitoring strategy as their core engines [40,50-52]. The search for interactions can be carried out via either exhaustive analysis [52] or the analysis that can be divided into two stages, incorporating single-locus analysis for the pre-screening purpose [40,50,51]. In the two-stage mode, at least one SNP that involves in the construction of two-locus genotype unit must be a strong candidate for the association explanation, usually verified through univariate statistical tests. Each mode of the two-locus analysis possesses different strengths and weaknesses. The exhaustive analysis has a full capability of detecting pure epistasis but requires larger computational efforts [52]. In contrast, the two-stage analysis is more practical for large-scale data but with some risk of missing possible pure epistasis [50]. More practical usage of both two-locus analysis modes in real case-control studies is required before the feasibility issue can be fully addressed.

Many genetic association studies reveal that various complex diseases are results of putative multi-locus interactions [11,46,53]. With the constraints on a computational capability, exhaustive multi-locus analysis in large-scale or genome-wide association studies would be infeasible [52]. On the other hand, single-locus analysis would be unsuitable for the detection of pure epistasis. One possible approach that provides a trade-off between a computational limitation and an epistasis detection capability is to capture a multi-locus interaction by combining multiple results from two-locus analysis. To achieve this, it is necessary to prove that once a multi-locus interaction model is broken down into a combination of two-locus models, all or some of these models remain detectable through two-locus analysis. Although it is hinted in an early work on two-locus analysis [52] that the proposed approach is plausible, explicit experimentation and testing has never been conducted.

In this article, the feasibility of employing an ensemble of two-locus analyses for the multi-locus interaction determination is demonstrated. Specifically, the significance of the two-locus analysis ensemble is assessed by an omnibus permutation test [54]. The proposed method is inspired by a set association approach [11], in which a limited number of sets that contain different numbers of SNPs are explored for possible association. These SNP sets are crucial in the global *p*-value calculation of the selected set via a permutation test and thus the decision to accept or reject the null hypothesis of no association. In other words, SNP set exploration and selection is required to assess the significance of the identified association. This means that the set association approach is equally interested in both SNP set selection and testing for significant association. The primary function of the proposed method is to detect possible association and assess its significance through the exploration of different ensembles of two-locus analyses. Hence, the proposed method is also equally interested in both ensemble selection and testing for significant association.

The proposed method is benchmarked against a simple exhaustive two-locus analysis technique, the set association approach [11], the correlation-based feature selection technique [14] and the tuned ReliefF technique [16]. These filter-based attribute selection techniques are suitable for the benchmark trial since they are capable of detecting association. The case-control classification models constructed from screened SNPs via a multifactor dimensionality reduction method [19] are also provided.

Results and discussion

Algorithm

The proposed algorithm performs an omnibus permutation test on ensembles of two-locus analyses and is referred to as a 2LOmb technique. The algorithm consists of four steps as illustrated in Figure 1 and can be described as follows.

Figure 1. Outline of 2LOmb. In this example, the algorithm takes a balanced case-control data set that consists of 400 samples and 1,000 SNPs. Each genotype is represented by an integer: 0 denotes a homozygous wild-type genotype, 1 denotes a heterozygous genotype and 2 denotes a homozygous variant or homozygous mutant genotype. A χ^2 contingency table is then constructed for each pair of SNPs in two-locus analysis. This results in the total of $\binom{1,000}{2}$ = 499,500 two-locus analyses. Thus, the Bonferroni-corrected χ^2's *p*-value for each two-locus analysis is the lower value between 499,500 × its uncorrected *p*-value and one. In one ensemble, Bonferroni-corrected χ^2's *p*-values from multiple two-locus analyses are combined together via a Fisher's combining function, which in turn provides a Fisher's test statistic result. The raw *p*-value for the ensemble is obtained through a permutation test, which is composed of 10,000 randomised permutation replicates. Since multiple ensembles may be tried during the identification of the best association explanation, a global *p*-value is calculated to account for multiple hypothesis testing. The global *p*-value is estimated through the same permutation test that gives the raw *p*-value for each ensemble. The progressive search for the best association explanation is carried out by incrementally adding a two-SNP unit to the current best ensemble. The condition for search termination is based on both the raw *p*-value for the explored ensemble and the global *p*-value. In this example, the search is terminated after the fourth ensemble is explored due to an increase in the raw *p*-value. Subsequently, the best SNP set for association explanation contains SNP1, SNP2 and SNP3 where the global *p*-value that accounts for testing of four hypotheses is $p < 0.0001$.

Two-locus analysis

Consider a case-control genetic association study with n_m SNPs, for each pair of SNPs, a 2 × 9 contingency table with rows for disease status and columns for genotype configurations is created. A χ^2 test statistic and the corresponding *p*-value can subsequently be computed. With the total of n_m SNPs, there are $\binom{n_m}{2}$ = $n_m!/((n_m - 2)!2!)$ possible SNP pairs. As a result, the *p*-value from each two-locus analysis must be adjusted by a Bonferroni correction. The Bonferroni-corrected *p*-value from each analysis is the lower value between $\binom{n_m}{2}$ × the uncorrected *p*-value and one.

Permutation test

The *p*-value p_0^e for the null hypothesis H_0^e that ensemble *e*--an ensemble of two-locus analyses of interest--is not associated with the disease can be evaluated by a permutation test. To achieve this, a scalar statistic is first computed from a function that combines the Bonferroni-corrected χ^2's *p*-values of individual two-locus tests. A suitable combining function must (a) be non-increasing in each *p*-value, (b) attain its maximum value when any *p*-value equals to zero and (c) have a finite critical value that is less than its maximum for any significant level greater than zero [54]. In this study, a Fisher's combining function ($-2\sum_i \log(p_i)$) is selected [55]. The *p*-value for the ensemble of two-locus analyses is assessed via a permutation simulation. In each permutation replicate, samples are constructed such that the case/control status of each sample is randomly permuted while the total numbers of case and control samples remain unchanged. A χ^2 contingency table with new entries and a Bonferroni-corrected *p*-value for the two-locus analysis within each permutation replicate are then obtained. This, in turn, leads to a new Fisher's test statistic. Let T_i^e denote the value of Fisher's test statistic obtained for the *i*th permutation replicate, p_0^e is the fraction of permutation replicates with a test statistic greater than or equal to the test statistic obtained from the original case-control data (T_0^e). In other words,

$$p_0^e = |\{i : 1 \le i \le t, T_i^e \ge T_0^e\}| \;/\; t,$$

where *t* is the number of permutation replicates which is set to 10,000 in this study and |·| denotes the size of a set.

Global *p*-value determination

There are many candidate ensembles of two-locus analyses that can be explored. Let $H_0 = \bigcap_{1 \le e \le E} H_0^e$ be the global null hypothesis that none of *E* explored ensembles of two-locus analyses is associated with the disease, the test of the global null hypothesis leads to the global *p*-value and provides the genetic association explanation. In step 2, the *p*-value p_0^e for a fixed hypothesis H_0^e is a raw or unadjusted *p*-value. To account for the correlation among multiple hypotheses that have been tested during the exploration through many candidate ensembles, the testing result of the global null hypothesis depends on $p_0^{\min} = \min_e p_0^e$. In other words, the global null hypothesis is rejected if the minimum of the raw *p*-values is sufficiently small. The distribution of $p_0^{\min}$ can again be determined by a permutation simulation. However, a nested simulation is unnecessary since the same set of permutation replicates for the p_0^e determination can be reused in the estimation of the empirical distribution of $p_0^{\min}$ [56]. This strategy has been successfully implemented in a number of genetic association detection techniques, including a set association approach [11] and a haplotype interaction approach embedded in FAMHAP [57,58]. The unadjusted *p*-value for the permutation replicate *i* of each hypothesis *e* is thus given by

$$p_i^e = |\{j : 0 \le j \le t, j \ne i, T_j^e \ge T_i^e\}| \;/\; t.$$

Let $p_i^{\min} = \min_e p_i^e$ be the minimum of unadjusted *p*-values over all explored ensembles of two-locus analyses in the *i*th permutation replicate, the *p*-value for the global null hypothesis H_0 is defined by

$$p_{\text{global}} = |\{i : 1 \le i \le t, p_i^{\min} \le p_0^{\min}\}| \;/\; t.$$

Search for the best ensemble of two-locus analyses

A simple progressive search is used to identify the best ensemble of two-locus analyses. The search begins by locating the best two-SNP unit with the smallest Bonferroni-corrected χ^2's *p*-value from step 1. A permutation test is then performed for this two-locus analysis, yielding both raw and global *p*-values since only one hypothesis has been explored. Next, the search attempts to combine the existing best two-SNP unit with the two-SNP unit that possesses the next smallest Bonferroni-corrected χ^2's *p*-value from step 1 and does not have a higher permutation *p*-value than the first two-SNP unit. If this new ensemble yields either a higher raw *p*-value or the same raw *p*-value but a higher global *p*-value from a permutation test, the search is terminated and the association is explained by the previously identified two-locus analysis. Otherwise, the best ensemble of two-locus analyses is updated and the process of appending more two-SNP units to the ensemble continues. The progressive search terminates when deterioration in the raw or global *p*-value is detected, or all possible two-locus analyses have been included in the ensemble. It is recalled from step 3 that for the best ensemble containing $E - 1 < \binom{n_m}{2}$ two-locus analyses, its global *p*-value is obtained from the evaluation of *E* hypotheses.

Validity of the algorithm

A permutation replicate in 2LOmb is constructed by randomly assigning the case or control status to each sample while maintaining the original proportion of case and control samples. Once the construction of a permutation replicate is finished, the assigned case and control labels remain fixed to the samples. The pattern of case and control labels in each permutation replicate is thus constant and unique. Therefore, the Bonferroni-corrected χ^2's *p*-values from any two-SNP units within a permutation replicate are calculated from the same case-control data set. Hence, the combining of these Bonferroni-corrected χ^2's *p*-values via a Fisher's combining function is attainable. The calculation of Fisher's test statistics from all permutation replicates and the original data set leads to the raw or unadjusted *p*-value p_0^e for the null hypothesis H_0^e of the ensemble *e* as given in equation 1. Since the same set of permutation replicates is always used during the evaluation of each ensemble, the raw *p*-values for the null hypotheses from all ensembles can be directly compared against one another. Furthermore, the global *p*-value calculation is based on this set of permutation replicates. This is possible because the unadjusted *p*-value for the permutation replicate *i* of ensemble *e* or p_i^e can be calculated in a similar manner to the raw *p*-value p_0^e as defined in equation 2. The unadjusted *p*-values for the same permutation replicate but different ensembles can also be directly compared and the subsequent calculation of $p_i^{\min} = \min_e p_i^e$ is attainable. With $p_i^{\min}$ and $p_0^{\min} = \min_e p_0^e$, the *p*-value for the global null hypothesis $H_0 = \bigcap_{1 \le e \le E} H_0^e$ that incorporates all *E* explored hypotheses can be determined by equation 3. In summary, only one set of permutation replicates is required for the calculation of both the raw *p*-value for the null hypothesis of every ensemble and the global *p*-value. The *p*-values can be compared in each step of 2LOmb. Consequently, the selection of the best ensemble for association explanation can be carried out via a *p*-value comparison.

Testing with simulated data

2LOmb is benchmarked against a simple exhaustive two-locus analysis technique, a set association approach (SAA) [11], a correlation-based feature selection (CFS) technique [14] and a tuned ReliefF (TuRF) technique [16] in a simulation trial. The exhaustive two-locus analysis is simply the two-locus analysis procedure from the first step of the 2LOmb algorithm. An interaction is declared if at least one two-SNP unit with a Bonferroni-corrected χ^2's *p*-value below 0.05 is detected. The exhaustive two-locus analysis reports all SNPs that meet this detection condition. The simulation covers two main data categories: null data of no significant genetic association and data with causative SNPs which signify pure epistasis. The algorithm performance on the null data provides an indication for the false-positive error. On the other hand, the algorithm performance on the data with causative SNPs indicates the detection capability. An efficient algorithm should produce an output with a low number of SNPs and a high number of correctly-identified causative SNPs when epistasis is present. Similarly, it should also report that there are no causative SNPs in the null data. These two measures on the number of SNPs in the results are used as the performance indicators.

Each simulated data set contains 1,000, 2,000 or 4,000 SNPs in which either there are no causative SNPs or there is pure epistasis, governed by two, three or four causative SNPs. The allele frequencies of all causative SNPs are 0.5 while the minor allele frequencies of the remaining SNPs are between 0.05 and 0.5. The data set consists of balanced case-control samples of sizes 400, 800 or 1,600. All SNPs in control samples are in Hardy-Weinberg equilibrium (HWE) [59]. The genotype distribution of causative interacting SNPs follows the pure epistasis model by Culverhouse et al. [42], leading to three interesting values of heritability: 0.01, 0.025 and 0.05. Every SNP in each data set exhibits no marginal single-locus effect (Bonferroni-corrected χ^2's *p*-value > 0.05). Twenty-five independent data sets for each simulation setting are generated via a genomeSIM package [60]. A paired *t*-test is suitable to assess the significance of results since the same simulated data sets are used during the algorithm benchmarking.

The results from the null data problem are summarised in Figure 2 while the results from the two-, three-and four-locus interaction problems are shown in Figures 3-4, 5-6 and 7-8, respectively. Clearly, 2LOmb significantly outperforms other techniques in terms of the low number of output SNPs, the high number of correctly-identified causative SNPs or both in every interaction problem (a paired *t*-test on 675 benchmark results yields a *p*-value < 0.05). On the other hand, both 2LOmb and SAA have the lowest false-positive error when compared to other techniques in the null data problem (a paired *t*-test on 225 benchmark results yields a *p*-value < 0.05). The statistical power analysis also reveals that the benchmark trial with 25 independent data sets for each simulation setting is sufficient for an accurate evaluation of the overall algorithm performance (power > 0.95 for a Type I error rate of 0.05). These results can be further interpreted as follows.

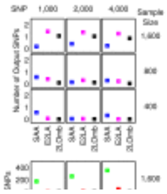
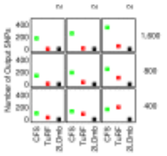


Figure 2. Performance of the exhaustive two-locus analysis, SAA, CFS, TuRF and 2LOmb in the null data problem. The results are averaged over 25 independent simulations. False detection is declared for the exhaustive two-locus analysis, SAA and 2LOmb if the *p*-values used as detection indicators in their results are less than 0.05. The results from the exhaustive two-locus analysis (E2LA), SAA, CFS, TuRF and 2LOmb are displayed using magenta, blue, green, red and black markers, respectively. In each chart, the horizontal axis represents the detection algorithm while the vertical axis represents the number of output SNPs



reported by the algorithm. The top nine charts are displayed using a finer scale than the bottom nine charts.

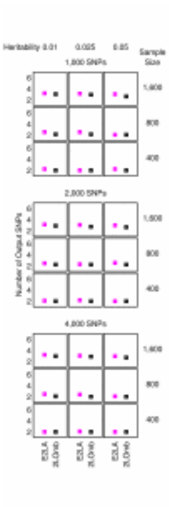


Figure 3. Performance of the exhaustive two-locus analysis and 2LOmb in the two-locus interaction problem. The results are averaged over 25 independent simulations. Detection is declared for the exhaustive two-locus analysis and 2LOmb if the *p*-values used as detection indicators in their results are less than 0.05. The results from the exhaustive two-locus analysis (E2LA) and 2LOmb are displayed using magenta and black markers, respectively. In each chart, the horizontal axis represents the detection algorithm while the vertical axis represents the number of output SNPs reported by the algorithm. All causative SNPs are present in outputs from both the exhaustive two-locus analysis and 2LOmb in all simulations.

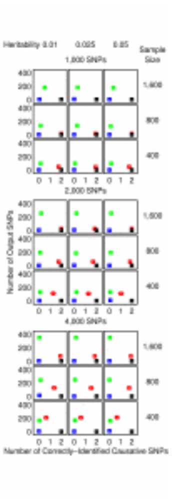


Figure 4. Performance of SAA, CFS, TuRF and 2LOmb in the two-locus interaction problem. The results are averaged over 25 independent simulations. Detection is declared for SAA and 2LOmb if the *p*-values used as detection indicators in their results are less than 0.05. The results from SAA, CFS, TuRF and 2LOmb are displayed using blue, green, red and black markers, respectively. In each chart, the horizontal axis represents the number of correctly-identified causative SNPs while the vertical axis represents the number of output SNPs reported by the algorithm. The charts on which the red markers are invisible denote the situations in which the performance of TuRF and 2LOmb is similar. The charts in this figure are displayed using a coarser scale than the charts in Figure 3.

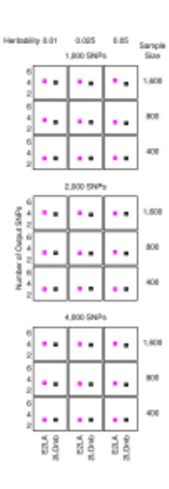


Figure 5. Performance of the exhaustive two-locus analysis and 2LOmb in the three-locus interaction problem. The explanation for how the results are obtained and displayed is the same as that given in Figure 3.

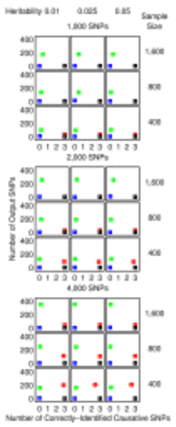


Figure 6. Performance of SAA, CFS, TuRF and 2LOmb in the three-locus interaction problem. The explanation for how the results are obtained and displayed is the same as that given in Figure 4.

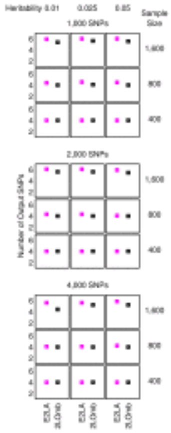


Figure 7. Performance of the exhaustive two-locus analysis and 2LOmb in the four-locus interaction problem. The explanation for how the results are obtained and displayed is the same as that given in Figure 3.

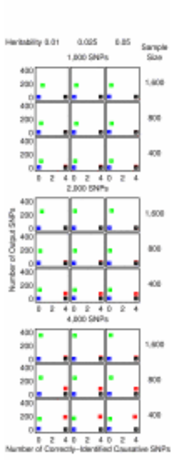


Figure 8. Performance of SAA, CFS, TuRF and 2LOmb in the four-locus interaction problem. The explanation for how the results are obtained and displayed is the same as that given in Figure 4.

The performance of many existing attribute selection techniques for pattern recognition depends on the level of attribute interactions. A number of techniques, including CFS, appear to function well under a moderate level of interactions. However, the performance of CFS appears to be significantly reduced when the interaction level becomes too high [14,61] because CFS favours an attribute that is strongly correlated with the classification outcome--disease status in this study--while at the same time is not correlated with other attributes. Since the main driving force behind epistasis is the interaction between SNPs, which are themselves attributes, CFS would not intuitively select all causative SNPs. Consequently, the SNP set produced by CFS appears to contain only uncorrelated SNPs. Obviously, a SNP that is a part of the interaction model would occasionally be picked up by CFS but CFS never successfully identifies all causative SNPs in any interaction problems. In addition, CFS reports more erroneous SNPs than other techniques in the null data problem and all three interaction problems due to many SNPs being uncorrelated.

The benchmarking of attribute selection techniques by Hall and Holmes [14] also reveals that ReliefF [15] is better than CFS in problems with a high level of interactions. Since ReliefF is essentially the core engine of TuRF, the results from this study are in agreement with the early benchmark trial. This finding strengthens the observation that the interaction level of SNPs in pure epistasis models is too high for CFS to handle. Similar to its predecessor, the performance of TuRF still depends on both the number of attributes and sample size. TuRF performs well in the majority of simulation scenarios with 1,000-2,000 SNPs and 800-1,600 samples. These scenarios are relatively easy since the number of SNPs is small while the sample size is large. However, the size of output SNP set, reported by TuRF from the null data problem and all three interaction problems, increases significantly when the difficulty level rises by either reducing the sample size or increasing the number of SNPs. This implies that when the problem contains a large number of candidate SNPs, the only way to ensure that TuRF reports a proper SNP set is to use a relatively large sample size, making it impractical in real genetic association studies due to many factors including disease prevalence, population size and genotyping cost.

The global *p*-values in most of the SAA results from the null data problem and all three interaction problems exceed 0.05, showing that SAA reports a low false-positive result in the null data problem. Nonetheless, SAA remains unsuitable for detecting pure epistasis because of its high false-negative error. This poor performance can be traced back to the manner in which SAA exploits an omnibus permutation test. As stated earlier, single-locus analysis does not detect any association between a SNP and the disease in this study. Hoh et al. [11] have demonstrated that genetic association can be more significantly observed when the single-locus test statistics are combined together. Nonetheless, there is an additional requirement that each causative SNP must exhibit a marginal single-locus effect. In the current study, the association signal from each causative SNP is lower than the required threshold, leading to similar test statistics and global *p*-values for both combinations of multiple SNPs which include causative SNPs and those which exclude causative SNPs.

Both 2LOmb and exhaustive two-locus analysis technique are capable of identifying all causative SNPs. However, the size of output SNP set from 2LOmb is significantly smaller than that from the exhaustive two-locus analysis. Appended SNPs to the causative SNPs in the output from 2LOmb and those from the exhaustive two-locus analysis are erroneous SNPs. These erroneous SNPs are parts of false two-SNP units with Bonferroni-corrected χ^2's *p*-values less than 0.05. A similar trend of results regarding the size of output SNP set is also observed in the benchmark trial involving the application of 2LOmb and exhaustive two-locus analysis to the null data. This signifies that the permutation test and the progressive search embedded in 2LOmb can help reducing the number of erroneous SNPs in the output.

As mentioned earlier, 2LOmb produces the best results among five techniques in the benchmark trial. 2LOmb has a low false-positive error in the null data problem and is capable of detecting all causative SNPs in every simulated data set in all three interaction problems. This performance is further strengthened by highly significant global *p*-values in 2LOmb results from all three interaction problems (p < 0.0001) and the presence of a SNP in common among some or all pairs of two-SNP units in the three- and four-locus interaction problems. Nonetheless, some of the 2LOmb outputs contain a few erroneous SNPs which are irrelevant to the correct association explanation. Since all three interaction problems involving different numbers of causative SNPs are investigated by varying the total number of SNPs, the sample size and the level of heritability, these parameters may influence the number of erroneous SNPs in the 2LOmb results. Similarly, the total number of SNPs and the sample size may affect the number of erroneous SNPs in the 2LOmb results from the null data problem. ANOVA reveals that the only source of variation that significantly affects the number of erroneous SNPs in the null data, two-locus interaction and three-locus interaction problems is the sample size (p < 0.000001). In addition, the sample size must be greater than 800 for an increase in the number of erroneous SNPs to be significant. In contrast, ANOVA reveals that two sources of variation that affect the number of erroneous SNPs in the four-locus interaction problem are the sample size (p < 0.000001) and the total number of SNPs (p < 0.00005). Similar to the null data, two-locus interaction and three-locus interaction problems, the sample size in the four-locus interaction problem must be greater than 800 to create a significant increase in the number of erroneous SNPs. On the other hand, the number of erroneous SNPs appears to decrease when the total number of SNPs increases. These two sources of variation also interact with each other (p < 0.005). However, the interaction is most evident only when the sample size is large, i.e. when the sample size is 1,600.

ANOVA shows that the number of erroneous SNPs in the 2LOmb results is influenced by the sample size and the total number of SNPs but not by the heritability. It is observed that the number of erroneous SNPs increases when the sample size is large. This counterintuitive phenomenon can be explained as follows. As 2LOmb combines *p*-values that are determined from χ^2 tests, the number of entries for the contingency table construction is large when the sample size is large. This subsequently leads to a significantly large χ^2 statistic and hence an extremely small *p*-value if the SNPs under consideration are causative SNPs. At the same time, the possibility that a reasonably large χ^2 statistic and a small *p*-value can be obtained by chance from a two-SNP unit which is irrelevant to the correct association explanation also inevitably increases. With the increase in the possibility of erroneous SNP inclusion, the size of output SNP set gets bigger when the sample size is large. Another observation that appears to be counterintuitive is the reduction in the number of erroneous

SNPs when the total number of SNPs increases. This phenomenon is the result of the Bonferroni correction usage. When the total number of SNPs is doubled, the Bonferroni correction factor in 2LOmb is quadrupled. A higher correction factor leads to a more stringent criterion for SNP selection. This subsequently leads to the reduction in the number of erroneous SNPs when the total number of SNPs is large.

In contrast to the first two parameters, different levels of heritability appear to have no effect on the 2LOmb results because all simulated data sets have balanced case-control samples and the embedded interaction models have the same architecture. For instance, a two-locus interaction model leads to zero penetrances for genotypes *AABB*, *AABb*, *AaBB*, *Aabb*, *aaBb* and *aabb*. Hence, the penetrances for these six genotypes are always equal to zero regardless of the heritability. On the other hand, genotypes *AAbb*, *AaBb* and *aaBB* have non-zero penetrances (see Methods for details). Therefore, different heritability levels certainly lead to different penetrances for genotypes *AAbb*, *AaBb* and *aaBB*. However, the ratios between the penetrances of these three genotypes are fixed and independent of the heritability. This model description can be generalised to cover the other multi-locus interaction models. In addition, the maximum penetrance in any two-locus or multi-locus interaction models always stays below 0.1 even though the heritability is at the highest level (see Methods for details). This means that case samples are always over-sampled from affected individuals to achieve a balanced case-control data set. Since all explored heritability levels lead to the same case over-sampling pattern, the simulated data sets of which the only primary difference being the heritability levels are indistinguishable from one another. This leads to the result similarities in interaction problems with the same number of SNPs in the data set, sample size and number of causative SNPs but different levels of heritability as shown in Figures 3, 4, 5, 6, 7 and 8. The result trend is also independent of the number of simulated data sets used in the benchmark trial.

In a permutation test, the ability to differentiate between two *p*-values is influenced by the number of permutation replicates. With *t* permutation replicates, the test declares an actual *p*-value that is less than 1/*t* to be zero. During the progressive search for the best ensemble, the inclusion of a new two-SNP unit is accepted if this inclusion does not worsen the current result. If the number of permutation replicates is too low, the search may include erroneous two-SNP units that are irrelevant to the correct association explanation. The analysis is confirmed as the number of output SNPs from 2LOmb is equal to the number of causative SNPs in most of simulation results. This phenomenon suggests that the number of permutation replicates employed in this study (t = 10,000) is high enough to screen off most of the erroneous two-SNP units. In other words, the inclusion of these erroneous two-SNP units leads to an increase in the *p*-value by at least 1/*t*. Nonetheless, the fact that 2LOmb results are not entirely free from erroneous SNPs suggests that there are erroneous two-SNP units with extremely small *p*-values. It is advisable to perform a genotype relative risk calculation for the elimination of erroneous SNPs. If the presence of an erroneous two-SNP unit is suspected, its result on two-locus genotype relative risk would not be as significant as that from the other two-SNP units in the ensemble. Alternatively, an additional means for further SNP screening by other techniques such as MDR is also recommended. The chance of erroneous SNP discovery would be further minimised by employing two consecutive attribute selection techniques. The same concept has been adopted for the implementation of MDR software, in which many additional filters including a χ^2 test, an odds ratio test, ReliefF and TuRF are available for SNP screening prior to the MDR analysis.

The two-, three- and four-locus interaction data sets which have been screened for causative SNPs by 2LOmb are subsequently subjected to MDR analysis. MDR has successfully identified all erroneous SNPs and the correct interaction models have been constructed from all data sets. The prediction accuracy from the MDR analysis is illustrated in Figure 9. It is noted that the prediction accuracy from all data sets is quite high due to the manner in which the pure epistasis model is defined [42]. Using the penetrance table for a two-locus interaction model with the heritability = 0.01 (see Methods for details), the two-locus genotype distribution of causative SNPs in a balanced case-control sample set from simulated data with 800 samples can be estimated and shown in Figure 10.

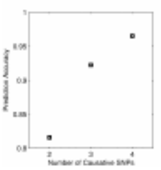


Figure 9. Prediction accuracy from the MDR analysis. A 10-fold cross-validation strategy is applied during the accuracy evaluation. The best MDR model is located by exploring all possible SNP combinations. All erroneous SNPs, which are left over after the screening by 2LOmb, have been successfully identified. All MDR models contain the correct number of causative SNPs. In addition, the MDR cross-validation consistency is 10/10.

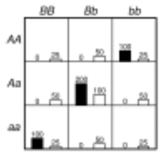


Figure 10. Genotype distribution of two causative SNPs in a balanced case-control data set with the sample size of 800. The left (black) bar in each cell represents the number of case samples while the right (white) bar represents the number of control samples. The cells with genotypes *AABB*, *AABb*, *AaBB*, *Aabb*, *aaBb* and *aabb* are labelled as protective genotypes while the cells with genotypes *AAbb*, *AaBb* and *aaBB* are labelled as disease-predisposing genotypes.

Six genotypes in Figure 10 namely *AABB*, *AABb*, *AaBB*, *Aabb*, *aaBb* and *aabb* are protective genotypes. In other words, a sample with one of these six genotypes is a control sample because the penetrances for these genotypes are zero. It is also noted that the control samples with all nine genotypes precisely follow the distribution as jointly described by independent single-locus genotype distribution from loci A and B. In contrast, three remaining genotypes in Figure 10 namely *AAbb*, *AaBb* and *aaBB* are labelled as disease-predisposing genotypes because the majority of samples with these three genotypes are case samples. Samples with these genotypes may be either case or control samples because the penetrances for these genotypes are between zero and one. In fact, the probabilities that persons with these genotypes to have the disease are quite low since the penetrances for these genotypes are small. However, case samples must be over-sampled from affected individuals to ensure a balanced case-control data set because the disease prevalence for this two-locus interaction model is only 0.004975. In addition, each case sample must contain one of these three genotypes because the penetrances for the other genotypes are zero. As a result, the case samples with these genotypes do not follow the same two-locus genotype distribution as in the control samples. With six genotypes being exclusively specific to control samples and the majority of three remaining genotypes being found in case samples, the MDR prediction accuracy for the two-locus interaction model is high. This explanation can also be generalised to cover the MDR results from the other multi-locus interaction data sets.

Another advantage of using 2LOmb for SNP screening prior to the MDR analysis is the reduction in computational time for interaction detection. The computational time for 2LOmb to finish screening the SNPs is provided to demonstrate this strength of 2LOmb. Moreover, the computational time required to identify causative SNPs by the MDR analysis and that by the combined approach which involves SNP screening by 2LOmb and follows by the MDR analysis is given. The previously-described simulated data sets with causative SNPs are used to produce the computational time results from the SNP screening by 2LOmb and the combined approach. All possible interaction models that can be constructed from the 2LOmb outputs are explored by MDR in the combined approach. On the other hand, the data sets for the direct MDR analysis are prepared by restricting the number of SNPs in each data set to 100. Only SNPs that are irrelevant to the correct association explanation are removed from the original simulated data sets. Furthermore, MDR only explores the interaction models that do not cover more than four SNPs in the data for this latter simulation setting. The summary of computational time required for the SNP screening by 2LOmb and that for both direct MDR and combined approaches to correctly identify all causative SNPs is given in Table 1. The maximum time required by 2LOmb to screen SNPs in the largest data set is 419 seconds or approximately seven minutes. Moreover, the combined 2LOmb and MDR approach discovers the correct causative SNPs much faster than MDR. This time reduction is achieved even though the problems have been simplified for the direct MDR analysis. A direct application of MDR to the original simulated data sets is certainly impractical.

Table 1. Computational time required by 2LOmb, a combined 2LOmb and MDR approach, and direct MDR analysis to detect interactions in simulated data sets with different sizes and different numbers of causative SNPs.

The simulated multi-locus interaction problems in this article are based on the pure epistasis model by Culverhouse et al. [42]. It is possible to capture a number of multi-locus interactions with marginal two-locus effects via a combination of two-locus analyses. However, there are many multi-locus interaction scenarios without marginal two-locus effects. In such cases, 2LOmb and the exhaustive two-locus analysis technique are unable to detect interactions. Among the explored techniques, TuRF and MDR have a better chance of detection. Nonetheless, TuRF functions well only when the total number of SNPs in data is small and the sample size is large enough while the total number of SNPs in data affects the practicality of direct MDR analysis.

Every attribute selection technique has a limitation in terms of the maximum numbers of samples and attributes that it can handle. Single-locus analysis techniques always have a higher limit than multi-locus analysis techniques. Because attribute subset evaluation is usually integrated into multi-locus analysis techniques, consequently the number of possible attribute subsets that can be explored is extremely large when the candidate attribute set is large. Together with a potentially large sample size, a higher computational requirement for multi-locus analysis techniques is inevitable. As a result, the direct application of multi-locus analysis techniques to a much larger data set than those presented in this article, which is usually considered in genome-wide association studies, would be impractical. However, it is reasonable to expect that both marginal single-locus and epistatic effects are present in any genome-wide data sets. A multi-stage strategy that incorporates multiple techniques, designed for different detection modes, would be more suitable to handle large data. For instance, the marginal single-locus effects should be the first priority and, as such, be detected by single-locus analysis. Then, a special case of pure epistasis [2] or semi-purely epistatic events, in which a SNP displaying a marginal single-locus effect interacts with a SNP that exhibits no marginal single-locus effect, should be considered. Many two-locus analysis techniques have been proven to be well suited to this type of epistasis [40,50,51]. Finally, the detection of pure epistasis is carried out in the last stage. With the reduction of SNPs from the first stage, the chance that some multi-locus analysis techniques are applicable to the remaining SNPs increases. In addition to the multi-stage approach, a prior knowledge regarding the previously reported association can be exploited to select candidate genes based upon ontology and pathways. This practice is due to the necessity for the derivation of plausible interpretation. The screening for SNPs within or near candidate genes before the association detection also increases the chance that multi-locus analysis techniques can be applied to the remaining data.

Testing with real data

2LOmb has been applied to study a type 2 diabetes mellitus (T2D) data set, collected and investigated by the Wellcome Trust Case Control Consortium (WTCCC) [3]. The data set consists of 1,999 case samples from affected individuals in the UK and 3,004 control samples, which are the results of a merging between 1,500 samples from the UK blood services and 1,504 samples from the 1958 British birth cohort. The original genome-wide data set contains 500,568 SNPs that are obtained through the Affymetrix GeneChip 500 K Mapping Array Set. The SNP set is primarily reduced by screening for SNPs within and near 372 candidate genes collected by the Human Genome Epidemiology Network (HuGENet) [62]. These candidate genes cover genes from both positive and negative genetic association reports, in which studies are conducted in various ethnic groups and populations. The SNP set is further reduced by removing SNPs that exhibit strong evidence of genetic association via single-locus analysis. The final SNP set contains 7,065 SNPs from 370 candidate genes. All SNPs in the reduced data set exhibit no marginal single-locus effects (Bonferroni-corrected χ^2's *p*-value > 0.05). Detailed description of the final SNP set is given in the supplement (see Additional file 1).

Additional file 1. List of SNPs for the association study of T2D. This Excel spreadsheet file contains the information about 7,065 SNPs which are explored during the genetic association study of T2D. Bonferroni-corrected and uncorrected χ^2's *p*-values from single-locus analyses are also provided.

Format: XLS Size: 847KB Download file

This file can be viewed with: Microsoft Excel Viewer

OPEN DATA

The 2LOmb search in the reduced T2D data set takes 3,456 seconds (57.6 minutes) of computational time on the Beowulf cluster. The possible genetic association is detected from 11 intronic SNPs in four genes (global *p*-value < 0.0001). Details of these SNPs, the two-SNP units that exhibit marginal two-locus effects and the identified genes are given in Table 2. A two-SNP unit is located in *LMX1A*. A two-SNP unit is also detected in *PARK2*. In addition, there is one SNP in common among SNPs in both *GYS2* two-SNP units. Similarly, there is one common SNP among three two-SNP units located in *PGM1*. Nonetheless, a two-SNP unit in which each SNP is located in a different gene is absent, indicating that there is no evidence of gene-gene interactions which can be observed from the 2LOmb result. Linkage disequilibrium (LD) analysis is subsequently performed using a JLIN package [63] and the resulting LD patterns are illustrated in Figure 11. It is noted that there is strong LD among SNPs within each gene due to high values of *D'* [64] and r^2 [65]. The genotype and haplotype relative risks are then calculated and the results are presented in Tables 3, 4, 5, 6, 7, 8, 9 and 10. Haplotype inference is carried out using an expectation-maximisation method [66]. The analysis reveals that a more prominent indication of a relative risk is observed when two-SNP units are considered. It is also noted that the genotype relative risk is directly influenced by the haplotype relative risk once a genotype is phased into all possible haplotype pairs. The detection of these two-SNP units is thus believed to be the consequence of haplotype effects. An early T2D association study also reveals similar haplotype effects in FUSION data [67]. Next, an interaction dendrogram [68,69] constructed from the 11 SNPs by MDR software is given in Figure 12. A strong synergistic effect between the two SNPs in *PARK2* is clearly observed. In contrast, the interactions between *PGM1*, *LMX1A*, *PARK2* and *GYS2* are clearly absent.

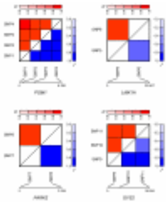


Figure 11. Linkage disequilibrium (LD) patterns of SNPs in *PGM1*, *LMX1A*, *PARK2* and *GYS2*. LD is explained via *D'* displayed in the upper triangle and r^2 displayed in the lower triangle. Dark colours indicate high values while pale colours indicate low values. Distances between SNPs are given in terms of the number of base pairs. SNP1 = rs2269241, SNP2 = rs2269239, SNP3 = rs3790857, SNP4 = rs2269238, SNP5 = rs2348250, SNP6 = rs6702087, SNP7 = rs1893551, SNP8 = rs6924502, SNP9 = rs6487236, SNP10 = rs1871142 and SNP11 = rs10770836.

Figure 12. Interaction dendrogram produced from 11 SNPs that are chosen by 2LOmb. The colours in the dendrogram comprise a spectrum of colours representing a transition from synergy to redundancy. Synergy denotes the situation in which the entropy-based interaction between two SNPs provides more information than the entropy-based correlation between the pair. Redundancy refers to the situation in which the entropy-based interaction between two SNPs provides less information than the entropy-based correlation between the pair [7].

Table 2. 2LOmb identifies 11 intronic SNPs, which are located in four genes, from the reduced T2D data.

Table 3. Genotype relative risk evaluated from genotype distribution of SNPs in *PGM1*.

Table 4. Haplotype relative risk evaluated from genotype distribution of SNPs in *PGM1*.

Table 5. Genotype relative risk evaluated from genotype distribution of SNPs in *LMX1A*.

Table 6. Haplotype relative risk evaluated from genotype distribution of SNPs in *LMX1A*.

Table 7. Genotype relative risk evaluated from genotype distribution of SNPs in *PARK2*.

Table 8. Haplotype relative risk evaluated from genotype distribution of SNPs in *PARK2*.

Table 9. Genotype relative risk evaluated from genotype distribution of SNPs in *GYS2*.

Table 10. Haplotype relative risk evaluated from genotype distribution of SNPs in *GYS2*.

Since many early genetic association studies of T2D and metabolic syndrome employ MDR analysis [43-45,47], additional MDR analysis would be useful for the comparison. The screened T2D case-control data set which contains 11 SNPs identified by 2LOmb is further subjected to MDR analysis. The prediction accuracy of the best MDR model is summarised in Table 11. The model covers six SNPs in three genes: *PGM1*, *PARK2* and *GYS2*. These SNPs are also present in three two-SNP units identified by 2LOmb. It is noted that the prediction accuracy in this real data set is much less than that from the simulated data sets. Nevertheless, the attainment of low prediction accuracy does not necessarily suggest that there is no genetic association. Early works involving genetic association studies of T2D and metabolic syndrome in various populations via MDR analysis produce similar values of prediction accuracy as summarised in Table 12. The prediction accuracy by MDR from most studies is in the range of 0.5-0.6. The only genetic association study of T2D that the prediction accuracy is distinctively high is conducted in a Korean population [43]. Differences in genetic background, candidate genes and selected SNPs are the main causes of variation in the genetic association results. Although MDR does not select five SNPs from the 2LOmb output, these SNPs should not be regarded as erroneous SNPs because there is strong linkage disequilibrium among SNPs in each gene. Moreover, early genotype and haplotype relative risk analysis clearly indicates that each gene, identified by 2LOmb, plays a role in the T2D association explanation. Overall, the analysis with the methods above only confirms the positive association for *PGM1*, *LMX1A*, *PARK2* and *GYS2* while gene-gene interactions are clearly absent. This signifies that, for the current study, there is no interaction between each pair of the identified genes that can be described by purely epistatic two-locus interaction models. In addition, there are no interactions between these four genes that can be described by purely epistatic multi-locus interaction models with marginal two-locus effects.

Table 11. Prediction accuracy of the best MDR model constructed from the 2LOmb output.

Table 12. Summary of prediction accuracy by MDR from early genetic association studies of T2D in a Korean population, a Han Chinese population from Taiwan, a female population from the US, and that from an early genetic association study of metabolic syndrome in an Italian population from the Centre East Coast Italy.

The four genes selected by 2LOmb regulate many pathways that involve in the disease development [70-72]. The genetic association studies involving these genes have been previously conducted in different populations. For instance, *LMX1A* has been chosen as a positional and biological candidate gene for a case-control study of T2D in Pima Indians [73]. This gene is chosen as a candidate because a linkage of T2D to chromosome 1q21-q23 has been previously reported [74]. In addition, *LMX1A* is one of LIM homeobox genes that are expressed in pancreas and has been shown to activate insulin gene transcription. Although SNPs have been carefully selected from the entire gene, no association between these SNPs in *LMX1A* and T2D has been found in this ethnic group.

PARK2 is another candidate gene that is also selected for case-control studies, based on evidence from genome-wide linkage analysis [75]. A linkage of T2D in an African American population to chromosome 6q24-q27 has been previously identified [76]. Although *PARK2* mainly involves in the development of Parkinson's disease, single-locus analysis reveals strong evidence of association between SNPs, which are in the vicinity of SNPs identified by 2LOmb, and T2D in African Americans.

In contrast to *LMX1A* and *PARK2*, which are candidate genes in typical T2D case-control studies, *GYS2* is considered in a study to identify genes responsible for troglitazone-associated hepatotoxicity in Japaneses with T2D [77]. In other words, both case and control samples in the study are drawn from troglitazone-treated T2D patients, in which case patients exhibit an abnormal increase in alanine transaminase (ALT) and aspartate transaminase (AST) levels. *GYS2* regulates starch and sucrose metabolism and an insulin signalling pathway. The selected SNPs in *GYS2* are not found to associate with troglitazone-induced hepatotoxicity.

Similar to the study of *GYS2*, the association study involving *PGM1* is not carried out as a typical T2D case-control study. In fact, an attempt to identify association between *PGM1* polymorphisms and obesity has been conducted among T2D affected individuals in Italy [78]. *PGM1* regulates glycolysis and gluconeogenesis, starch and sucrose metabolism, galactose metabolism, a pentose phosphate pathway, and streptomycin biosynthesis. Isozyme polymorphisms [79,80], which are defined through structural differences in PGM1 protein, are used instead of SNPs in the study where positive association is identified.

In summary, positive association has been reported from previous studies involving *PARK2* in African Americans and *PGM1* in Italians. In contrast, negative association has been reported from previous studies about *LMX1A* in Pima Indians and *GYS2* in Japaneses. Both *GYS2* and *PGM1* regulate starch and sucrose metabolism while *LMX1A* and *PARK2* govern insulin gene transcription and Parkinson's disease development, respectively. The above discussion strengthens the importance of conducting large-scale association studies due to two main reasons. Firstly, a gene that does not contribute to the aetiology of a complex disease in one population may be important for association explanation in another population. Secondly, the absence of interacting candidate genes from a study may lead to negative association due to a lack of necessary genetic information. A two-locus interaction can occur between SNPs from genes that regulate one specific pathway [44] or between SNPs from genes that regulate different pathways [45]. Furthermore, a multi-locus interaction may involve both SNPs from genes that regulate the same pathway and SNPs from genes that govern different pathways. Hence, candidate genes should be selected by considering all pathways that directly and indirectly contribute to the disease development.

This study produces evidence of association between 11 intronic SNPs in *PGM1*, *LMX1A*, *PARK2* and *GYS2*, and T2D in a UK population. Although there are other independent genome-wide T2D data sets, the association detection within these data using a similar methodology to the presented method has never been attempted because the methodology employed in the majority of genome-wide association studies is based on single-locus analysis [3,81]. It is recalled that each SNP explored in the reduced T2D data set exhibits no marginal single-locus effect. Hence, the most logical approach to confirm the possibility of replicating association results from the current study is to perform the same detection method on these independent data sets. This is certainly important to gain further understanding of the genetic role in T2D susceptibility.

Implementation

2LOmb is implemented in a C programming language. All functions within the program are written by the first author except the χ^2 distribution function, which is taken from the Numerical Recipes in C [82]. The program can be compiled by Microsoft Visual Studio and GNU C compilers. The program has been successfully tested for the execution under Windows and Linux operating systems. The time required by 2LOmb to complete a problem containing n attributes is $T(n) = \binom{n}{2} = n!/((n-2)!2!) = n(n-1)/2$. 2LOmb thus has the order of n^2 time complexity ($T(n) \in O(n^2)$). Consequently, 2LOmb can tackle problems in quadratic time. 2LOmb in its present form occupies one processor during the program execution. A parallel version of 2LOmb for genome-wide data is under development. All results included in the study are collected from the execution of computer programs in a Beowulf cluster. The computational platform consists of 12 nodes. Each node is equipped with dual Xeon 2.8 GHz processors and 4GB of main memory. The Rocks Cluster Distribution is installed on all nodes.

Conclusion

In this article, a method for detecting epistatic multi-locus interactions in case-control data is presented. The study focuses on pure epistasis [2], which cannot be detected via single-locus analysis [42]. To overcome this difficulty, the proposed method performs an omnibus permutation test [54] on ensembles of two-locus analyses and is thus referred to as 2LOmb. The detection performance of 2LOmb is evaluated using both simulated and real data. From the simulation, 2LOmb produces a low false-positive error when the tests on null data of no association are performed. Furthermore, 2LOmb can identify all causative SNPs and outperforms a simple exhaustive two-locus analysis technique, a set association approach (SAA) [11], a correlation-based feature selection (CFS) technique [14] and a tuned ReliefF (TuRF) technique [16] in various interaction scenarios with marginal two-locus effects. These scenarios are set up by varying the number of causative SNPs, the number of SNPs in data, the sample size and the heritability. ANOVA reveals that the number of SNPs in data and the sample size influence the number of erroneous SNPs appended to the correctly-identified causative SNPs in the 2LOmb output. In contrast, the results from 2LOmb appear to be insensitive to the variation in heritability. After subjecting the data sets containing only SNPs that are screened by 2LOmb to multifactor dimensionality reduction (MDR) analysis [19], all erroneous SNPs are successfully removed. In addition, an insight into the MDR models is provided. 2LOmb is subsequently applied to a real case-control type 2 diabetes mellitus (T2D) data set, which is collected from a UK population by the Wellcome Trust Case Control Consortium (WTCCC) [3]. The original genome-wide data set is first reduced by selecting only SNPs that locate within or near 372 candidate genes reported by the Human Genome Epidemiology Network (HuGENet) [62]. In addition, the selected SNPs must exhibit no marginal single-locus effects. The final data set, which consists of 1,999 case samples and 3,004 control samples, contains 7,065 SNPs from 370 candidate genes. 2LOmb identifies 11 intronic SNPs that are associated with the disease. These SNPs are located in *PGM1*, *LMX1A*, *PARK2* and *GYS2*. The 2LOmb result suggests that there is no interaction between each pair of the identified genes that can be described by purely epistatic two-locus interaction models. Moreover, there are no interactions between these four genes that can be described by purely epistatic multi-locus interaction models with marginal two-locus effects. This evidence of genetic association for these four genes leads to an alternative explanation for the aetiology of T2D in the UK population. It also implies that SNPs from genome-wide data which are usually discarded after single-locus analysis confirms the null hypothesis of no association can still be useful for genetic association studies of complex diseases.

Methods

Pure epistasis model

The pure epistasis model of interest is proposed by Culverhouse et al. [42]. The model describes a restriction or constraint for penetrance of each genotype constituting the interaction model. Consider a two-locus model that captures an interaction between loci A and B, let *A* and *a* be the major (common) and minor (rare) alleles at locus A. Similarly, let *B* and *b* be the major and minor alleles at locus B. At each locus, the genotype is represented by characters 0, 1 or 2 where 0 denotes a homozygous wild-type genotype (*AA* and *BB*), 1 denotes a heterozygous genotype (*Aa* and *Bb*) and 2 denotes a homozygous variant or homozygous mutant genotype (*aa* and *bb*). $f_{ij} \in [0, 1]$ is defined as the disease penetrance of the two-locus genotype *ij* that consists of genotype *i* at locus A and genotype *j* at locus B. The marginal penetrances M_{Ai} for genotype *i* at locus A and M_{Bj} for genotype *j* at locus B are given by

$$M_{Ai} = p_B^2 f_{i0} + 2p_B(1 - p_B)f_{i1} + (1 - p_B)^2 f_{i2}, i \in \{0, 1, 2\},$$

and

$$M_{Bj} = p_A^2 f_{0j} + 2p_A(1 - p_A)f_{1j} + (1 - p_A)^2 f_{2j}, j \in \{0, 1, 2\},$$

where p_A and p_B are the major allele frequencies. Equations 4 and 5 are usually represented by a penetrance table as illustrated in Table 13. The two-locus interaction model is a pure epistasis model if

Table 13. Penetrances for a two-locus interaction model.

$$M_{Ai} = M_{Bj} = K, \forall i, j \in \{0, 1, 2\},$$

where *K* is the disease prevalence. Obviously, many combinations of penetrance f_{ij} satisfy the condition given in equation 6. Culverhouse et al. [42] suggest that a pure epistasis model with the maximum heritability is particularly useful in association studies. The heritability (h^2) of the two-locus interaction model is defined by

$$h^2 = V_I / V_T,$$

where $V_T = K(1 - K)$ is the total variance of the dichotomous phenotypes in the population and V_I is the epistatic variation attributable to the genotypes. V_I is defined by

$$\begin{aligned} V_I = {} & p_A^2 p_B^2 (f_{00} - K)^2 + 2p_A^2 p_B(1 - p_B)(f_{01} - K)^2 + p_A^2(1 - p_B)^2(f_{02} - K)^2 \\ & + 2p_A(1 - p_A)p_B^2(f_{10} - K)^2 + 4p_A(1 - p_A)p_B(1 - p_B)(f_{11} - K)^2 + 2p_A(1 - p_A)(1 - p_B)^2(f_{12} - K)^2 \\ & + (1 - p_A)^2 p_B^2(f_{20} - K)^2 + 2(1 - p_A)^2 p_B(1 - p_B)(f_{21} - K)^2 + (1 - p_A)^2(1 - p_B)^2(f_{22} - K)^2. \end{aligned}$$

The search for feasible penetrance f_{ij} that also maximises the heritability or other variance-based objectives can be treated as a constraint optimisation problem. Many algorithms including a double description method [42] and a genetic algorithm [83] have been proven to be suitable for the task.

Culverhouse et al. [42] have identified the maximum heritability of purely epistatic two-locus and multi-locus interaction models for various values of disease prevalence. For instance, the maximum heritability of a two-locus interaction model for $p_A = p_B = 0.5$ with the penetrances in Table 14 is

Table 14. Two-locus penetrances that lead to the maximum heritability $h^2_{\max}(K) = 2K/(1 - K)$ for $K \in (0, 1/4]$.

$$h^2_{\max}(K) = 2K / (1 - K), K \in (0, 1/4].$$

When a two-locus interaction model is expanded into a multi-locus interaction model, the marginal penetrance equality constraint must be extended to cover all loci. Furthermore, the expression for V_I must also be expanded to cover additional genotypes while the expression for V_T remains unchanged. With the necessary model expansion, the maximum heritability of a three-locus interaction model for $p_A = p_B = p_C = 0.5$ with the penetrances in Table 15 is given by

Table 15. Three-locus penetrances that lead to the maximum heritability $h^2_{\max}(K) = 9K/(1 - K)$ for $K \in (0, 1/16]$.

$$h^2_{\max}(K) = 9K/(1-K), K \in (0, 1/16].$$

Similarly, the maximum heritability of a four-locus interaction model for $p_A = p_B = p_C = p_D = 0.5$ with the penetrances in Table 16 is

Table 16. Four-locus penetrances that lead to the maximum heritability $h^2_{\max}(K) = 35K/(1 - K)$ for $K \in (0, 1/64]$.

$$h^2_{\max}(K) = 35K/(1-K), K \in (0, 1/64].$$

Additional details about the maximum heritability and the corresponding two-locus and multi-locus penetrance tables for other values of disease prevalence can be found in Culverhouse et al. [42]. In this article, the simulated data sets are generated to achieve the maximum heritability of 0.01, 0.025 and 0.05. The values of disease prevalence that lead to the target heritability for two-, three- and four-locus interaction models are given in Table 17.

Table 17. Disease prevalence that gives the target maximum heritability of 0.01, 0.025 and 0.05 for two-, three- and four-locus interaction models.

genomeSIM

genomeSIM is a simulation package for generating case-control samples in large-scale and genome-wide association studies [60]. The package is capable of producing many realistic scenarios, which can be observed in a population and genetic samples, including linkage disequilibrium, phenocopy and genotyping errors. The case/control status of each sample is determined from the penetrance-based genetic models or interaction models. As a result, the package can accommodate many epistasis models including the one proposed by Culverhouse et al. [42]. A data set can be produced via two modes: a population-based simulation and a probability-based simulation. In the population-based simulation, an initial population is generated according to the predefined allele frequency of each SNP. Then further generations are created by crossing the genotype strings within successive generations until the specified number of generations is reached. The resulting data set contains a population-dependent case and control samples that follow a forward-time simulation strategy. In contrast, genotype strings are incrementally generated without any string crossing for only one generation in the probability-based simulation. The creation of new strings is terminated only when the desired numbers of case and control samples are obtained. In this study, the probability-based simulation is used to produce all case and control samples where the simulation parameter setting is given in the supplement (see Additional file 2). genomeSIM is available upon request to Scott M. Dudek at the Vanderbilt University dudek@chgr.mc.vanderbilt.edu.

Additional file 2. genomeSIM parameters. This text file contains an example of parameter setting in the genomeSIM simulation package.

Format: TXT Size: 2KB Download file

OPEN DATA

Set association approach

A set association approach (SAA) is an association detection technique based on an omnibus permutation test on sets of candidate SNPs [11]. The test captures information about genotyping errors, deviation from Hardy-Weinberg equilibrium (HWE) and allelic association. In the first step, the genotype distribution for each SNP in the control samples is checked for HWE. Then, the number of SNPs that is to be excluded from the study (n_d) is set to the number of SNPs in the control samples that deviate from HWE. Two test statistics are subsequently calculated for each SNP: an allelic association statistic and a statistic for the deviation from HWE of each SNP in the case samples. The allelic association statistic is a χ^2 statistic which is calculated from the contingency table of alleles or genotypes with disease status. On the other hand, a χ^2 statistic for the deviation from HWE of each SNP in the case samples indicates the level of association. A large deviation from the equilibrium usually signifies strong association between a SNP and the disease. However, an excessively large deviation may be the result of genotyping errors. n_d SNPs with largest test statistics for the deviation from HWE are hence excluded from the consideration.

The test statistics for the allelic association and deviation from HWE are multiplied together to form a single *S* statistic for each remaining SNP. SNPs are then ranked according to their *S* statistics. A preset number of SNPs with highest ranks are considered for association. The first candidate SNP set contains only the SNP with the highest rank (the highest *S* statistic). The *p*-value for this first set is determined from a permutation simulation where the case and control labels are randomly permuted while the numbers of case and control samples remain unchanged. In each permutation replicate, a new genotype contingency table is constructed and a new *S* statistic is subsequently obtained. The *p*-value is given by the fraction of permutation replicates with an *S* statistic greater than or equal to the *S* statistic from the original data. The second candidate SNP set consists of the first two SNPs in the rank list. The test statistic for this SNP set is the sum of *S* statistics from both SNPs. The *p*-value for the second candidate SNP set is also obtained through the permutation simulation. By progressively adding the remaining SNP with the highest rank to the previously considered candidate set and performing the permutation simulation, *p*-values for all candidate SNP sets are estimated. The sizes of candidate SNP sets have the range of one to the preset number. Among all candidate sets, the SNP set that best describes genetic association has the lowest *p*-value.

Since multiple hypotheses are postulated during the construction of candidate SNP sets, the global *p*-value for the selected candidate set must be evaluated. This is achieved through a permutation simulation in which the current raw *p*-value for the chosen candidate set is now used as the test statistic. The existing permutation replicates, created for the early estimation of the raw *p*-value, can be reused and a nested permutation simulation is hence avoided. In this study, the maximum allowable size of the candidate SNP set is the total number of available SNPs while the number of permutation replicates for *p*-value evaluation is set to 10,000. The allelic association statistic employed in the study is the χ^2 statistic that is obtained through the contingency table of genotypes with disease status. A PASCAL program for the set association approach can be obtained from the website for *S* statistic in gene mapping [84].

Correlation-based feature selection technique

A correlation-based feature selection (CFS) technique [14] is an attribute (SNP) subset evaluation heuristic that considers both the usefulness of individual features (SNPs) in the (case-control) classification task and the level of inter-correlation among features. Each attribute subset is assigned a score given by

$$Merit_F = \frac{n_C \bar{r}_{cf}}{\sqrt{n_C + n_C(n_C - 1)\bar{r}_{ff}}},$$

where $Merit_F$ is the heuristic merit of an n_c-attribute subset *F*, $\bar{r}_{cf}$ is the average feature-class correlation and $\bar{r}_{ff}$ is the average feature-feature inter-correlation. An attribute subset receives a high merit score if it contains features that are highly correlated with the class and at the same time have low inter-correlation among one another. An application of a best-first search for the best subset identification is carried out to avoid searching through all possible attribute subsets. CFS has been integrated into a Weka package [85,86].

Tuned ReliefF

A tuned ReliefF (TuRF) algorithm is a ranking algorithm for identifying genetic markers which are important in case-control classification [16]. TuRF is built on a ReliefF engine [15]. ReliefF randomly picks a sample from the (case-control) data and identifies its n_k nearest neighbours from the same class and another n_k nearest neighbours from the opposite class. The attribute values--the genotypes in this application--of the neighbour samples are compared to that of the randomly picked sample and are subsequently used to update the relevance score for each attribute (genetic marker). This process is repeated for a specified number of samples, which is limited by the total sample size. The rationale of ReliefF is that an attribute which is important for the classification should have different values for samples from different classes and have the same value for samples from the same class. The relevance score of an attribute have a range from -1 (not relevant) to +1 (highly relevant). TuRF exploits the capability of ReliefF by repeatedly executing ReliefF and removing a portion of worst attributes at the end of each execution. This leads to the reevaluation of remaining attributes and, hence, reduces the effects of attribute noise on the attribute screening. In this study, the number of repetitions for random sample picking in the ReliefF part is equal to the total number of case-control samples while the neighbourhood size (n_k) for the relevance score calculation is set to ten. Furthermore, the worst 1% of SNPs is removed at the end of each ReliefF iteration (TuRF 1%). TuRF has been integrated into the current distribution of multifactor dimensionality reduction (MDR) software.

Multifactor dimensionality reduction

A multifactor dimensionality reduction (MDR) method is a wrapper-based technique that is capable of identifying the best genetic marker combination among possible markers for the separation between case and control samples [19]. Similar to other wrapper-based methods, an n_f -fold cross-validation technique provides a means to determine the prediction accuracy of the candidate marker model. Basically, the combined case and control samples are randomly divided into n_f folds where n_f - 1 folds of samples are used to construct a decision table while the remaining fold of samples is used to identify the prediction capability of the constructed decision table. The decision table construction and testing procedure is repeated n_f times. Hence, the samples in each fold are always used both to construct and to test the decision table. The number of cells in a decision table is given by G^{n_c} where n_c is the number of candidate markers selected from possible markers and G is the number of possible genotypes according to the marker. For a SNP, which is a bi-allelic marker, G is equal to three. During the decision table construction, each cell in the table is filled with case and control samples that have their genotype corresponds to the cell label. The ratio between numbers of case and control samples provides the decision for each cell whether the corresponding genotype is a protective or disease-predisposing genotype. An example of decision table construction is illustrated in Figure 13.

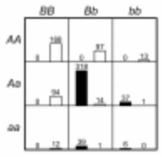


Figure 13. An MDR decision table that is constructed using a balanced case-control data set with the sample size of 800. The genotype of each sample is determined from two SNPs. The table consists of nine cells where each cell represents a unique genotype. The left (black) bar in each cell represents the number of case samples while the right (white) bar represents the number of control samples. The cells with genotypes *AABB*, *AABb*, *AAbb*, *AaBB* and *aaBB* are labelled as protective genotypes while the cells with genotypes *AaBb*, *Aabb*, *aaBb* and *aabb* are labelled as disease-predisposing genotypes.

The prediction accuracy of the decision table is subsequently evaluated by counting the numbers of case and control samples in the testing fold that their disease status can correctly be identified using the constructed decision rules. The process of decision table construction and evaluation must be cycled through all or some of possible 2^{n_m} - 1 combinations where n_m is the total number of available markers in the study. The best genetic marker combination is determined from two criteria: prediction accuracy and cross-validation consistency. Each time that a testing fold is used for the prediction accuracy determination, the accuracy of the interesting marker combination model is compared with that from other models that also contain the same number of markers. The model that consistently ranks the first in comparison to other choices with the same number of markers has high cross-validation consistency. Prediction accuracy is the main criterion for decision making while cross-validation consistency is only used as an auxiliary measure. Cross-validation consistency generally confirms that the high rank model can consistently be identified regardless of how the samples are divided for cross-validation. In a situation where two or more models with different number of markers are equally good in terms of prediction accuracy and cross-validation consistency, the most parsimonious model--the combination with the least number of markers--is chosen as the best model.

After the best model has been selected, a permutation test is used to assess the probability of obtaining prediction accuracy that is at least as large as or larger than that observed in the original data from randomised data. This represents the probability that the null hypothesis of no association is true. Each permutation replicate is constructed by randomly assigning the case/control status to each sample with the numbers of case and control samples remaining fixed. MDR analysis is subsequently carried out to obtain the prediction accuracy of each permutation replicate. The empirical p-value is denoted by the fraction of permutation replicates with the prediction accuracy greater than or equal to the prediction accuracy obtained from the original data. MDR software, which incorporates many additional features including interaction visualisation via dendrograms and genetic marker screening via a χ^2 test, an odds ratio test, ReliefF and TuRF, is available from its homepage [87].

JLIN

JLIN or a Java LINkage disequilibrium plotter is a computer program for visualisation of linkage disequilibrium analysis [63]. The program is capable of displaying many statistical measures including D' [64] and r^2 [65]. The program is publicly available from the Centre for Genetic Epidemiology and Biostatistics, University of Western Australia [88].

Interaction dendrogram

An interaction dendrogram is a graphical tool for the visualisation of relationships among attributes (SNPs) [68,69]. The interaction dendrogram is constructed via hierarchical clustering analysis and is embedded into MDR software [87]. The dendrogram illustrates the entropy-based interaction between attributes by displaying interacting or related attributes closely together as adjacent leaves in a tree. At the same time, independent attributes are placed far apart from one another. In addition, the conclusion regarding whether the interaction between attributes is synergistic or redundancy is present can be deduced.

Availability and requirements

The 2LOmb program for Windows platforms and examples of simulated data are available at http://code.google.com/p/nachol/w/list webcite.

List of abbreviations

2LOmb: omnibus permutation test on ensembles of two-locus analyses; ALT: alanine transaminase; ANOVA: analysis of variance; AST: aspartate transaminase; CFS: correlation-based feature selection; CI: confidence interval; CVC: cross-validation consistency; *DIO2*: deiodinase, iodothyronine, type II; E2LA: exhaustive two-locus analysis; *EGFR*: epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian); FAMHAP: software for single-marker analysis and joint analysis of unphased genotype data from tightly linked markers (haplotype analysis); FUSION: Finland-United States Investigation of NIDDM Genetics; genomeSIM: simulation package for generating case-control samples in large-scale and genome-wide association studies; *GYS2*: glycogen synthase 2 (liver); *HNF4A*: hepatocyte nuclear factor 4, alpha; HuGENet: Human Genome Epidemiology Network; HWE: Hardy-Weinberg equilibrium; JLIN: Java LINkage disequilibrium plotter; *KCNJ11*: potassium inwardly-rectifying channel, subfamily J, member 11; LD: linkage disequilibrium; LIM domains: protein structural domains that are named after their initial discovery in the proteins Lin11, Isl-1 and Mec-3; *LMX1A*: LIM homeobox transcription factor 1, alpha; MDR: multifactor dimensionality reduction; NIDDM: noninsulin-dependent diabetes mellitus; *PARK2*: Parkinson disease (autosomal recessive, juvenile) 2, parkin; *PGM1*: phosphoglucomutase 1; *PPARG*: peroxisome proliferator-activated receptor gamma; *RXRG*: retinoid X receptor, gamma; SAA: set association approach; SNP: single nucleotide polymorphism; T2D: type 2 diabetes mellitus; TuRF: tuned ReliefF; *UCP2*: uncoupling protein 2 (mitochondrial, proton carrier); Weka: Waikato environment for knowledge analysis; WTCCC: Wellcome Trust Case Control Consortium.

Authors' contributions

WW conducted the literature survey, formulated the research question, implemented the proposed algorithm, designed the experiment, and collected and interpreted the computational results. AA conducted the literature survey, formulated the research question, designed the experiment and secured the access to the genomeSIM package. TP performed the statistical analysis and interpreted the statistical results. SS monitored and oversaw the execution of computer programs on the Beowulf cluster. CL provided additional comments about the genetic association study of T2D. NC conducted the literature survey, formulated the research question, designed the proposed algorithm, designed the experiment, secured the access to the T2D data from WTCCC, selected the candidate genes for the T2D association study, discussed all results, drew the conclusions and wrote the manuscript. All authors read and approved the final manuscript.

Authors' information

WW is a Ph.D. student at the Department of Electrical Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok. He also received his B.Eng. and M.Eng. degrees in electrical engineering from King Mongkut's University of Technology North Bangkok. His current research interests include machine learning, evolutionary computation and bioinformatics.

AA is a Ph.D. student at the Department of Immunology, Faculty of Medicine Siriraj Hospital, Mahidol University. He also received his B.Sc. degree in pharmacy from Mahidol University. His current research interests include human genetics, genetic epidemiology, population genetics and bioinformatics.

TP is a Ph.D. student at the Department of Electrical Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok. He also received his B.Eng. and M.Eng. degrees in production engineering from King Mongkut's University of Technology North Bangkok. His current research interests include evolutionary multi-objective optimisation and machine learning.

SS is a part-time research assistant at the Department of Electrical Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok. He received his B.Eng. and M.Eng. degrees in electrical engineering from Thammasat University and King Mongkut's University of Technology North Bangkok, respectively. His current research interests include machine learning and genetic epidemiology.

CL is the Head of Division of Molecular Genetics at the Department of Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University. He also received his M.D. degree from Mahidol University. His current research interests include human genetics and genetic diseases.

NC is an associate professor of electrical engineering at King Mongkut's University of Technology North Bangkok and an adjunct professor of genetic epidemiology at Mahidol University. He received his B.Eng. and Ph.D. degrees from the Department of Automatic Control and Systems Engineering, University of Sheffield. His current research interests include evolutionary computation, machine learning and genetic epidemiology.

Acknowledgements

The authors are extremely grateful to four anonymous reviewers and Dr. Danielle Talbot for their valuable comments and suggestions, which have contributed a lot towards improving the content and presentation of this article. The authors are also extremely grateful to Dr. Kuntinee Maneeratana, Dr. Pensiri Tongpadungrod, Dr. Graeme James Sheppard and Mr.

Edward L. Karas for their efforts on proofreading the manuscript. The authors wish to thank Dr. Marong Phadoongsidhi and Dr. Vara Varavithya and for their assistance on processing the T2D data and providing an access to the Beowulf cluster, respectively. The authors also wish to thank Dr. Sissades Tongsima for his assistance on testing of multiple hypotheses, providing many suggestions for the manuscript revision and allowing the usage of computer facilities at the Genome Institute, National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA) during the algorithm development. The authors acknowledge Scott M. Dudek at the Vanderbilt University for providing an access to the genomeSIM package. Furthermore, the authors acknowledge Dr. Audrey Duncanson and the Wellcome Trust Case Control Consortium for granting an access to the genome-wide case-control data. WW, AA and TP were supported by the Thailand Research Fund (TRF) through the Royal Golden Jubilee Ph.D. Programme (Grant No. PHD/1.E.KN.49/A.1, PHD/4.I.MU.45/C.1 and PHD/1.E.KN.50/A.1, respectively). CL was supported by the Mahidol Research Grant. NC was supported by the Thailand Research Fund and the National Science and Technology Development Agency through the NSTDA Chair Professor Grant.

References

1. Risch N, Merikangas K: The future of genetic studies of complex human diseases.
   Science 1996, 273:1516-1517. PubMed Abstract | Publisher Full Text

2. Musani SK, Shriner D, Liu N, Feng R, Coffey CS, Yi N, Tiwari HK, Allison DB: Detection of gene × gene interactions in genome-wide association studies of human population data.
   Hum Hered 2007, 63:67-84. PubMed Abstract | Publisher Full Text

3. The Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.
   Nature 2007, 447:661-678. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

4. The GAIN Collaborative Research Group: New models of collaboration in genome-wide association studies: the Genetic Association Information Network.
   Nat Genet 2007, 39:1045-1051. PubMed Abstract | Publisher Full Text

5. Heidema AG, Boer JMA, Nagelkerke N, Mariman ECM, van der A DL, Feskens EJM: The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases.
   BMC Genet 2006, 7:23. PubMed Abstract | BioMed Central Full Text | PubMed Central Full Text

6. Motsinger AA, Ritchie MD, Reif DM: Novel methods for detecting epistasis in pharmacogenomics studies.
   Pharmacogenomics 2007, 8:1229-1241. PubMed Abstract | Publisher Full Text

7. Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC: A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility.
   J Theor Biol 2006, 241:252-261. PubMed Abstract | Publisher Full Text

8. Saeys Y, Inza I, Larrañaga P: A review of feature selection techniques in bioinformatics.
   Bioinformatics 2007, 23:2507-2517. PubMed Abstract | Publisher Full Text

9. Lewis CM: Genetic association studies: design, analysis and interpretation.
   Brief Bioinform 2002, 3:146-153. PubMed Abstract | Publisher Full Text

10. Montana G: Statistical methods in genetics.
    Brief Bioinform 2006, 7:297-308. PubMed Abstract | Publisher Full Text

11. Hoh J, Wille A, Ott J: Trimming, weighting, and grouping SNPs in human case-control association studies.
    Genome Res 2001, 11:2115-2119. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

12. Potter DM: Omnibus permutation tests of the association of an ensemble of genetic markers with disease in case-control studies.
    Genet Epidemiol 2006, 30:438-446. PubMed Abstract | Publisher Full Text

13. Chapman J, Clayton D: Detecting association using epistatic information.
    Genet Epidemiol 2007, 31:894-909. PubMed Abstract | Publisher Full Text

14. Hall MA, Holmes G: Benchmarking attribute selection techniques for discrete class data mining.
    IEEE Trans Knowl Data Eng 2003, 15:1437-1447. Publisher Full Text

15. Robnik-Šikonja M, Kononenko I: Theoretical and empirical analysis of ReliefF and RReliefF. *Mach Learn* 2003, 53:23-69. Publisher Full Text

16. Moore JH, White BC: Tuning ReliefF for genome-wide genetic analysis. In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Edited by Marchiori E, Moore JH, Rajapakse JC. Berlin, Heidelberg: Springer; 2007:166-175.
[Goos G, Hartmanis J, van Leeuwen J (Founding and Former Series Editors): Lecture Notes in Computer Science, vol 4447].

17. Nelson MR, Kardia SLR, Ferrell RE, Sing CF: A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* 2001, 11:458-470. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

18. Culverhouse R, Klein T, Shannon W: Detecting epistatic interactions contributing to quantitative traits. *Genet Epidemiol* 2004, 27:141-152. PubMed Abstract | Publisher Full Text

19. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001, 69:138-147. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

20. Hahn LW, Ritchie MD, Moore JH: Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 2003, 19:376-382. PubMed Abstract | Publisher Full Text

21. Bush WS, Dudek SM, Ritchie MD: Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene-gene interactions. *Bioinformatics* 2006, 22:2173-2174. PubMed Abstract | Publisher Full Text

22. Chung Y, Lee SY, Elston RC, Park T: Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. *Bioinformatics* 2007, 23:71-76. PubMed Abstract | Publisher Full Text

23. Bush WS, Edwards TL, Dudek SM, McKinney BA, Ritchie MD: Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. *BMC Bioinformatics* 2008, 9:238. PubMed Abstract | BioMed Central Full Text | PubMed Central Full Text

24. Lou XY, Chen GB, Yan L, Ma JZ, Mangold JE, Zhu J, Elston RC, Li MD: A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies. *Am J Hum Genet* 2008, 83:457-467. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

25. Edwards TL, Lewis K, Velez DR, Dudek SM, Ritchie MD: Exploring the performance of multifactor dimensionality reduction in large scale SNP studies and in the presence of genetic heterogeneity among epistatic disease models. *Hum Hered* 2009, 67:183-192. PubMed Abstract | Publisher Full Text

26. Mechanic LE, Luke BT, Goodman JE, Chanock SJ, Harris CC: Polymorphism Interaction Analysis (PIA): a method for investigating complex gene-gene interactions. *BMC Bioinformatics* 2008, 9:146. PubMed Abstract | BioMed Central Full Text | PubMed Central Full Text

27. Liang KH, Hwang Y, Shao WC, Chen EY: An algorithm for model construction and its applications to pharmacogenomic studies. *J Hum Genet* 2006, 51:751-759. PubMed Abstract | Publisher Full Text

28. Estrada-Gil JK, Fernández-López JC, Hernández-Lemus E, Silva-Zolezzi I, Hidalgo-Miranda A, Jiménez-Sánchez G, Vallejo-Clemente EE: GPDTI: a Genetic Programming Decision Tree Induction method to find epistatic effects in common complex diseases. *Bioinformatics* 2007, 23:i167-i174. PubMed Abstract | Publisher Full Text

29. Nunkesser R, Bernholt T, Schwender H, Ickstadt K, Wegener I: Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics* 2007, 23:3280-3288. PubMed Abstract | Publisher Full Text

30. Lunetta KL, Hayward LB, Segal J, van Eerdewegh P: Screening large-scale association study data: exploiting interactions using random forests.

BMC Genet 2004, 5:32. PubMed Abstract | BioMed Central Full Text | PubMed Central Full Text

31. Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, van Eerdewegh P: Identifying SNPs predictive of phenotype using random forests.
Genet Epidemiol 2005, 28:171-182. PubMed Abstract | Publisher Full Text

32. Chen X, Liu CT, Zhang M, Zhang H: A forest-based approach to identifying gene and gene-gene interactions.
Proc Natl Acad Sci USA 2007, 104:19199-19203. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

33. Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH: Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases.
BMC Bioinformatics 2003, 4:28. PubMed Abstract | BioMed Central Full Text | PubMed Central Full Text

34. Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD: Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology.
Genet Epidemiol 2008, 32:325-340. PubMed Abstract | Publisher Full Text

35. Cordell HJ: Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans.
Hum Mol Genet 2002, 11:2463-2468. PubMed Abstract | Publisher Full Text

36. Wilson SR: Epistasis. In *Nature Encyclopedia of the Human Genome*. *Volume 2*. Edited by Cooper DN. London: Nature Publishing Group; 2004:317-320.

37. Neuman RJ, Rice JP: Two-locus models of disease.
Genet Epidemiol 1992, 9:347-365. PubMed Abstract | Publisher Full Text

38. Schork NJ, Boehnke M, Terwilliger JD, Ott J: Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits.
Am J Hum Genet 1993, 53:1127-1136. PubMed Abstract | PubMed Central Full Text

39. Li W, Reich J: A complete enumeration and classification of two-locus disease models.
Hum Hered 2000, 50:334-349. PubMed Abstract | Publisher Full Text

40. Marchini J, Donnelly P, Cardon LR: Genome-wide strategies for detecting multiple loci that influence complex diseases.
Nat Genet 2005, 37:413-417. PubMed Abstract | Publisher Full Text

41. Hallgrímsdóttir IB, Yuster DS: A complete classification of epistatic two-locus models.
BMC Genet 2008, 9:17. PubMed Abstract | BioMed Central Full Text | PubMed Central Full Text

42. Culverhouse R, Suarez BK, Lin J, Reich T: A perspective on epistasis: limits of models displaying no main effect.
Am J Hum Genet 2002, 70:461-471. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

43. Cho YM, Ritchie MD, Moore JH, Park JY, Lee KU, Shin HD, Lee HK, Park KS: Multifactor-dimensionality reduction shows a two-locus interaction associated with type 2 diabetes mellitus.
Diabetologia 2004, 47:549-554. PubMed Abstract | Publisher Full Text

44. Hsieh CH, Liang KH, Hung YJ, Huang LC, Pei D, Liao YT, Kuo SW, Bey MSJ, Chen JL, Chen EY: Analysis of epistasis for diabetic nephropathy among type 2 diabetic patients.
Hum Mol Genet 2006, 15:2701-2708. PubMed Abstract | Publisher Full Text

45. Qi L, van Dam RM, Asselbergs FW, Hu FB: Gene-gene interactions between *HNF4A* and *KCNJ11* in predicting type 2 diabetes in women.
Diabet Med 2007, 24:1187-1191. PubMed Abstract | Publisher Full Text

46. Zhang Z, Zhang S, Wong MY, Wareham NJ, Sha Q: An ensemble learning approach jointly modeling main and interaction effects in genetic association studies.
Genet Epidemiol 2008, 32:285-300. PubMed Abstract | Publisher Full Text

47. Fiorito M, Torrente I, De Cosmo S, Guida V, Colosimo A, Prudente S, Flex E, Menghini R, Miccoli R, Penno G,

Pellegrini F, Tassi V, Federici M, Trischitta V, Dallapiccola B: Interaction of *DIO2* T92A and *PPARγ2* P12A polymorphisms in the modulation of metabolic syndrome.
Obesity 2007, 15:2889-2895. PubMed Abstract | Publisher Full Text

48. Albrechtsen A, Castella S, Andersen G, Hansen T, Pedersen O, Nielsen R: A Bayesian multilocus association method: allowing for higher-order interaction in association studies.
Genetics 2007, 176:1197-1208. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

49. Zhang Y, Liu JS: Bayesian inference of epistatic interactions in case-control studies.
Nat Genet 2007, 39:1167-1173. PubMed Abstract | Publisher Full Text

50. Evans DM, Marchini J, Morris AP, Cardon LR: Two-stage two-locus models in genome-wide association.
PLoS Genet 2006, 2:e157. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

51. Ionita I, Man M: Optimal two-stage strategy for detecting interacting genes in complex diseases.
BMC Genet 2006, 7:39. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

52. Gayán J, González-Pérez A, Bermudo F, Sáez ME, Royo JL, Quintas A, Galan JJ, Morón FJ, Ramirez-Lorca R, Real LM, Ruiz A: A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis.
BMC Genomics 2008, 9:360. PubMed Abstract | BioMed Central Full Text | PubMed Central Full Text

53. Heidema AG, Feskens EJM, Doevendans PAFM, Ruven HJT, van Houwelingen HC, Mariman ECM, Boer JMA: Analysis of multiple SNPs in genetic association studies: comparison of three multi-locus methods to prioritize and select SNPs.
Genet Epidemiol 2007, 31:910-921. PubMed Abstract | Publisher Full Text

54. Pesarin F: *Multivariate Permutation Tests with Applications to Biostatistics*. Chichester: Wiley; 2001.

55. Fisher RA: *Statistical Methods for Research Workers*. 4th edition. London: Oliver and Boyd; 1932.

56. Westfall PH, Young SS: *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*. New York: John Wiley and Sons; 1993.

57. Becker T, Schumacher J, Cichon S, Baur MP, Knapp M: Haplotype interaction analysis of unlinked regions.
Genet Epidemiol 2005, 29:313-322. PubMed Abstract | Publisher Full Text

58. Herold C, Becker T: Genetic association analysis with FAMHAP: a major program update.
Bioinformatics 2009, 25:134-136. PubMed Abstract | Publisher Full Text

59. Hardy GH: Mendelian proportions in a mixed population.
Science 1908, 28:49-50. PubMed Abstract | Publisher Full Text

60. Dudek SM, Motsinger AA, Velez DR, Williams SM, Ritchie MD: Data simulation software for whole-genome association and other studies in human genetics. In *Proceedings of the Pacific Symposium on Biocomputing 2006: 3-7 January 2006; Maui*. Edited by Altman RB, Dunker AK, Hunter L, Murray T, Klein TE. Singapore: World Scientific; 2006:499-510. PubMed Abstract | Publisher Full Text

61. Guyon I, Elisseeff A: An introduction to variable and feature selection.
J Mach Learn Res 2003, 3:1157-1182. Publisher Full Text

62. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ: A navigator for human genome epidemiology.
Nat Genet 2008, 40:124-125. PubMed Abstract | Publisher Full Text

63. Carter KW, McCaskie PA, Palmer LJ: JLIN: a java based linkage disequilibrium plotter.
BMC Bioinformatics 2006, 7:60. PubMed Abstract | BioMed Central Full Text | PubMed Central Full Text

64. Lewontin RC: The interaction of selection and linkage. I. general considerations; heterotic models.
Genetics 1964, 49:49-67. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

65. Hill WG, Robertson A: Linkage disequilibrium in finite populations.
Theor Appl Genet 1968, 38:226-231. Publisher Full Text

66. Excoffier L, Slatkin M: Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population.

Mol Biol Evol 1995, 12:921-927. PubMed Abstract | Publisher Full Text

67. Epstein MP, Satten GA: Inference on haplotype effects in case-control studies using unphased genotype data.
Am J Hum Genet 2003, 73:1316-1329. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

68. Jakulin A, Bratko I, Smrke D, Demšar J, Zupan B: Attribute interactions in medical data analysis. In *Artificial Intelligence in Medicine*. Edited by Dojat M, Keravnou E, Barahona P. Berlin, Heidelberg: Springer; 2003:229-238.
PubMed Abstract | Publisher Full Text
[Carbonell JG, Siekmann J (Series Editors): Lecture Notes in Artificial Intelligence, vol 2780].

69. Jakulin A, Bratko I: Analyzing attribute dependencies. In *Knowledge Discovery in Databases: PKDD 2003*. Edited by Lavrač N, Gamberger D, Todorovski L, Blockeel H. Berlin, Heidelberg: Springer; 2003:229-240.
[Carbonell JG, Siekmann J (Series Editors): Lecture Notes in Artificial Intelligence, vol 2838].

70. Kanehisa M, Goto S: KEGG: Kyoto Encyclopedia of Genes and Genomes.
Nucleic Acids Res 2000, 28:27-30. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

71. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita K, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: From genomics to chemical genomics: new developments in KEGG.
Nucleic Acids Res 2006, 34:D354-D357. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

72. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y: KEGG for linking genomes to life and the environment.
Nucleic Acids Res 2008, 36:D480-D484. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

73. Thameem F, Wolford JK, Wang J, German MS, Bogardus C, Prochazka M: Cloning, expression and genomic structure of human *LMX1A*, and variant screening in Pima Indians.
Gene 2002, 290:217-225. PubMed Abstract | Publisher Full Text

74. Hanson RL, Ehm MG, Pettitt DJ, Prochazka M, Thompson DB, Timberlake D, Foroud T, Kobes S, Baier L, Burns DK, Almasy L, Blangero J, Garvey WT, Bennett PH, Knowler WC: An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians.
Am J Hum Genet 1998, 63:1130-1138. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

75. Leak TS, Mychaleckyj JC, Smith SG, Keene KL, Gordon CJ, Hicks PJ, Freedman BI, Bowden DW, Sale MM: Evaluation of a SNP map of 6q24-27 confirms diabetic nephropathy loci and identifies novel associations in type 2 diabetes patients with nephropathy from an African-American population.
Hum Genet 2008, 124:63-71. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

76. Sale MM, Freedman BI, Langefeld CD, Williams AH, Hicks PJ, Colicigno CJ, Beck SR, Brown WM, Rich SS, Bowden DW: A genome-wide scan for type 2 diabetes in African-American families reveals evidence for a locus on chromosome 6q.
Diabetes 2004, 53:830-837. PubMed Abstract | Publisher Full Text

77. Watanabe I, Tomita A, Shimizu M, Sugawara M, Yasumo H, Koishi R, Takahashi T, Miyoshi K, Nakamura K, Izumi T, Matsushita Y, Furukawa H, Haruyama H, Koga T: A study to survey susceptible genetic factors responsible for troglitazone-associated hepatotoxicity in Japanese patients with type 2 diabetes mellitus.
Clin Pharmacol Ther 2003, 73:435-455. PubMed Abstract | Publisher Full Text

78. Gloria-Bottini F, Magrini A, Antonacci E, La Torre M, Di Renzo L, De Lorenzo A, Bergamaschi A, Bottini E: Phosphoglucomutase genetic polymorphism and body mass.
Am J Med Sci 2007, 334:421-425. PubMed Abstract | Publisher Full Text

79. Spencer N, Hopkinson DA, Harris H: Phosphoglucomutase polymorphism in man.
Nature 1964, 204:742-745. PubMed Abstract | Publisher Full Text

80. March RE, Putt W, Hollyoake M, Ives JH, Lovegrove JU, Hopkinson DA, Edwards YH, Whitehouse DB: The classical human phosphoglucomutase (PGM1) isozyme polymorphism is generated by intragenic recombination.
Proc Natl Acad Sci USA 1993, 90:10730-10733. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

81. Zeggini E, Scott LJ, Saxena R, Voight BF: Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes.
Nat Genet 2008, 40:638-645. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

82. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: *Numerical Recipes in C: The Art of Scientific Computing*. 2nd edition. Cambridge: Cambridge University Press; 1992.

83. Moore JH, Hahn LW, Ritchie MD, Thornton TA, White BC: Routine discovery of complex genetic models using genetic algorithms.
Appl Soft Comput 2004, 4:79-86. Publisher Full Text

84. *S* Statistic in Gene Mapping [http://www.genemapping.cn/sumstat.html] webcite

85. Witten IH, Frank E: *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd edition. San Francisco: Morgan Kaufmann; 2005.

86. Weka 3: Data Mining Software in Java [http://www.cs.waikato.ac.nz/ml/weka/] webcite

87. Multifactor Dimensionality Reduction [http://www.multifactordimensionalityreduction.org/] webcite

88. JLIN: A Java Based Linkage Disequilibrium Plotter [http://www.genepi.org.au/jlin.html] webcite

Have something to say? Post a comment on this article!

BMC Bioinformatics

IMPACT FACTOR 3.43

Log on/register

Feedback | Support | My details

home | journals A-Z | subject areas | advanced search | authors | reviewers | libraries | about | my BioMed Central

Comments(1)

Detecting purely epistatic multi-locus interactions by an omnibus permutation test on ensembles of two-locus analyses

Waranyu Wongseree ✉, Anunchai Assawamakin ✉, Theera Piroonratana ✉, Saravudh Sinsomros ✉, Chanin Limwongse ✉ and Nachol Chaiyaratana ✉

BMC Bioinformatics 2009, 10:294 doi:10.1186/1471-2105-10-294

Additional Authors' Acknowledgements

Nachol Chaiyaratana (07 April 2011) King Mongkut's University of Technology North Bangkok ✉

This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of investigators who contributed to the generation of data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. NC was supported by the Thailand Research Fund and the Office of the Higher Education Commission.

Competing interests

None declared.

top

BMC Bioinformatics
Volume 10

Viewing options:
Abstract
Full text
PDF (1.6MB)
Additional files

Associated material:
Readers' comments (1)
PubMed record

Related literature:
Articles citing this article
on Google Scholar
on ISI Web of Science
on PubMed Central
Other articles by authors
on Google Scholar
on PubMed
Related articles/pages
on Google
on Google Scholar
on PubMed

Tools:
Download citation(s)
Download XML
Email to a friend
Order reprints
Post a comment

Post to:
Citeulike
Connotea
Del.icio.us
Facebook
Mendeley
Twitter

Have something to say? Post a comment on this article!

Terms and Conditions Privacy statement Information for advertisers Jobs at BMC Contact us

BMC Bioinformatics

BioMed Central

Methodology article

Detecting purely epistatic multi-locus interactions by an omnibus permutation test on ensembles of two-locus analyses

Waranyu Wongseree[1], Anunchai Assawamakin[2], Theera Piroonratana[1], Saravudh Sinsomros[1], Chanin Limwongse[2] and Nachol Chaiyaratana*[1,2]

Address: [1]Department of Electrical Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok, 1518 Piboolsongkram Road, Bangsue, Bangkok 10800, Thailand and [2]Division of Molecular Genetics, Department of Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University, 2 Prannok Road, Bangkoknoi, Bangkok 10700, Thailand

Email: Waranyu Wongseree - waranyu.wongseree@gmail.com; Anunchai Assawamakin - anunchai_ice@yahoo.com; Theera Piroonratana - theepi@gmail.com; Saravudh Sinsomros - 2saravudh@gmail.com; Chanin Limwongse - siclw@mahidol.ac.th; Nachol Chaiyaratana* - n.chaiyaratana@gmail.com

* Corresponding author

Published: 17 September 2009

BMC Bioinformatics 2009, **10**:294 doi:10.1186/1471-2105-10-294

Received: 5 March 2009

Accepted: 17 September 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/294

Abstract

Background: Purely epistatic multi-locus interactions cannot generally be detected via single-locus analysis in case-control studies of complex diseases. Recently, many two-locus and multi-locus analysis techniques have been shown to be promising for the epistasis detection. However, exhaustive multi-locus analysis requires prohibitively large computational efforts when problems involve large-scale or genome-wide data. Furthermore, there is no explicit proof that a combination of multiple two-locus analyses can lead to the correct identification of multi-locus interactions.

Results: The proposed 2LOmb algorithm performs an omnibus permutation test on ensembles of two-locus analyses. The algorithm consists of four main steps: two-locus analysis, a permutation test, global *p*-value determination and a progressive search for the best ensemble. 2LOmb is benchmarked against an exhaustive two-locus analysis technique, a set association approach, a correlation-based feature selection (CFS) technique and a tuned ReliefF (TuRF) technique. The simulation results indicate that 2LOmb produces a low false-positive error. Moreover, 2LOmb has the best performance in terms of an ability to identify all causative single nucleotide polymorphisms (SNPs) and a low number of output SNPs in purely epistatic two-, three- and four-locus interaction problems. The interaction models constructed from the 2LOmb outputs via a multifactor dimensionality reduction (MDR) method are also included for the confirmation of epistasis detection. 2LOmb is subsequently applied to a type 2 diabetes mellitus (T2D) data set, which is obtained as a part of the UK genome-wide genetic epidemiology study by the Wellcome Trust Case Control Consortium (WTCCC). After primarily screening for SNPs that locate within or near 372 candidate genes and exhibit no marginal single-locus effects, the T2D data set is reduced to 7,065 SNPs from 370 genes. The 2LOmb search in the reduced T2D data reveals that four intronic SNPs in *PGM1* (phosphoglucomutase 1), two intronic SNPs in *LMX1A* (LIM homeobox transcription factor 1, alpha), two intronic SNPs in *PARK2* (Parkinson disease (autosomal recessive, juvenile) 2, parkin) and three intronic SNPs in *GYS2* (glycogen synthase 2 (liver)) are associated with the disease. The 2LOmb result suggests that there is no interaction between each pair of the identified genes that can be described by purely epistatic two-locus interaction models. Moreover, there are no interactions between these four genes that can be described by purely epistatic multi-locus interaction models with marginal two-locus effects. The findings provide an alternative explanation for the aetiology of T2D in a UK population.

Conclusion: An omnibus permutation test on ensembles of two-locus analyses can detect purely epistatic multi-locus interactions with marginal two-locus effects. The study also reveals that SNPs from large-scale or genome-wide case-control data which are discarded after single-locus analysis detects no association can still be useful for genetic epidemiology studies.

Background

Complex diseases cannot generally be explained by Mendelian inheritance [1] because they are influenced by gene-gene and gene-environment interactions. Many common diseases such as asthma, cancer, diabetes, hypertension and obesity are widely accepted and acknowledged to be results of complex interactions between multiple genetic factors [2]. Attempts to identify factors that could be the causes of complex diseases have led to many genome-wide association studies [3,4]. Raw results from these attempts produce a large amount of single nucleotide polymorphism (SNP) data from every individual participating in the trials.

For genetic epidemiologists, data sets from genome-wide association studies present many challenges, particularly the correct identification of SNPs that associate with the disease of interest from all available SNPs [5]. This challenge can be treated as a pattern recognition problem which aims to identify an attribute or SNP set that can lead to the correct classification of recruited samples. Heidema et al. [5] and Motsinger et al. [6] have reviewed and identified many machine learning techniques that are suitable to the task. Among many strategies and techniques, the protocol that appears to be most promising for genome-wide association studies involves two main steps: SNP set reduction and classification model construction [7]. From a machine learning viewpoint, attribute selection techniques can be divided into three main categories: filter, wrapper and embedded approaches [8]. In a filter approach, a measure or an index is used to determine the correlation between attributes and classes, e.g. affected and unaffected status in a case-control study. Attributes that are deemed to be important for the classification according to the measure are then selected. The filter approach includes χ^2 and odds ratio tests [9,10], omnibus permutation tests [11-13], a correlation-based feature selection technique [14], a ReliefF technique [15] and a tuned ReliefF technique [16]. In a wrapper approach, the significance of an attribute subset is evaluated from the classification performance by a classifier. The capability of the wrapper approach to identify significant attributes thus depends on the chosen classifier and the search algorithm for the identification of the best attribute subset. Combinatorial [17] and restricted partitioning methods [18], a multifactor dimensionality reduction method [19-25] and a polymorphism interaction analysis technique [26] are examples of the wrapper approach. An embedded approach concentrates on informative attributes during the construction of a classification model. Examples of the embedded approach include a genetic algorithm with Boolean algebra [27], genetic programming based decision trees [28,29], random forests [30-32] and evolutionary neural networks [33,34]. Based on this categorisation, classification models are not direct outputs from filter-based techniques. On the other hand, classification models are readily prepared as outputs from the wrapper and embedded approaches. In other words, the last two approaches can also be regarded as classification model construction techniques.

The success of the two-step pattern recognition approach relies heavily on the attribute selection step [14]. In case-control studies, epistatic effects play a vital role in establishing the difficulty level of SNP screening problems [35,36]. Epistasis in the simplest form can be represented by disease models that require genotype inputs from two interacting SNPs [37,38]. Many attempts have been made to produce consistent definitions and categorisation of different types of epistasis models [2,35,39-41]. According to Musani et al. [2], a pure epistasis model [42] is difficult because each SNP exhibits no marginal single-locus effect in the model. As a result, it is impossible to detect the pure epistasis by univariate statistical tests. Examples of complex diseases that case-control studies have uncovered putatively pure epistasis include type 2 diabetes mellitus (T2D) [43-46] and metabolic syndrome [47]. Due to the difficulty of screening for each SNP independently, it is suggested that attention should be focused on the analysis of differences between two-locus genotype distribution within case and control groups [40] and multi-locus Bayesian statistical analysis [48,49].

A number of SNP screening and association detection techniques have adopted the two-locus genotype monitoring strategy as their core engines [40,50-52]. The search for interactions can be carried out via either exhaustive analysis [52] or the analysis that can be divided into two stages, incorporating single-locus analysis for the prescreening purpose [40,50,51]. In the two-stage mode, at least one SNP that involves in the construction of two-locus genotype unit must be a strong candidate for the association explanation, usually verified through univariate statistical tests. Each mode of the two-locus analysis possesses different strengths and weaknesses. The exhaustive analysis has a full capability of detecting pure epistasis but requires larger computational efforts [52]. In contrast, the two-stage analysis is more practical for large-scale data but with some risk of missing possible pure epistasis [50]. More practical usage of both two-locus analysis modes in real case-control studies is required before the feasibility issue can be fully addressed.

Many genetic association studies reveal that various complex diseases are results of putative multi-locus interactions [11,46,53]. With the constraints on a computational capability, exhaustive multi-locus analysis in large-scale or genome-wide association studies would be infeasible [52]. On the other hand, single-locus analysis would be unsuitable for the detection of pure epistasis. One possible approach that provides a trade-off between a computational limitation and an epistasis detection capability is to capture a multi-locus interaction by combining multiple results from two-locus analysis. To achieve this, it is necessary to prove that once a multi-locus interaction model is broken down into a combination of two-locus models, all or some of these models remain detectable through two-locus analysis. Although it is hinted in an early work on two-locus analysis [52] that the proposed approach is plausible, explicit experimentation and testing has never been conducted.

In this article, the feasibility of employing an ensemble of two-locus analyses for the multi-locus interaction determination is demonstrated. Specifically, the significance of the two-locus analysis ensemble is assessed by an omnibus permutation test [54]. The proposed method is inspired by a set association approach [11], in which a limited number of sets that contain different numbers of SNPs are explored for possible association. These SNP sets are crucial in the global *p*-value calculation of the selected set via a permutation test and thus the decision to accept or reject the null hypothesis of no association. In other words, SNP set exploration and selection is required to assess the significance of the identified association. This means that the set association approach is equally interested in both SNP set selection and testing for significant association. The primary function of the proposed method is to detect possible association and assess its significance through the exploration of different ensembles of two-locus analyses. Hence, the proposed method is also equally interested in both ensemble selection and testing for significant association.

The proposed method is benchmarked against a simple exhaustive two-locus analysis technique, the set association approach [11], the correlation-based feature selection technique [14] and the tuned ReliefF technique [16]. These filter-based attribute selection techniques are suitable for the benchmark trial since they are capable of detecting association. The case-control classification models constructed from screened SNPs via a multifactor dimensionality reduction method [19] are also provided.

Results and discussion

Algorithm

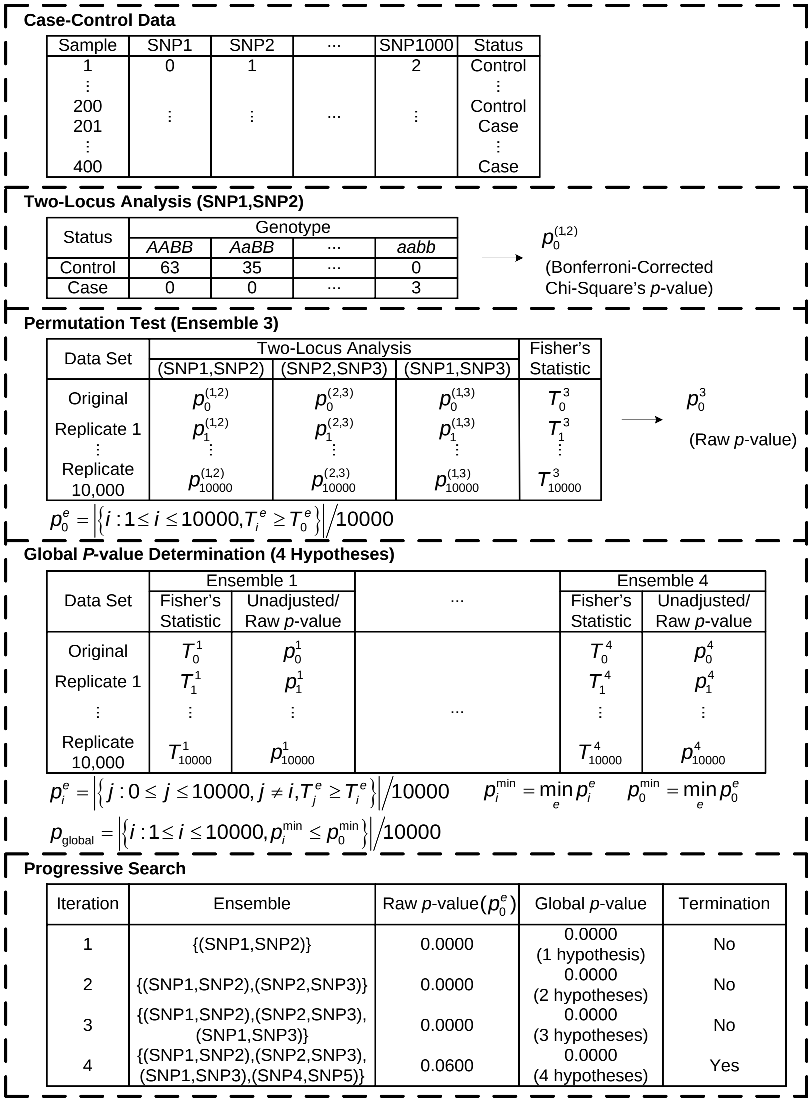
The proposed algorithm performs an omnibus permutation test on ensembles of two-locus analyses and is referred to as a 2LOmb technique. The algorithm consists of four steps as illustrated in Figure 1 and can be described as follows.

Two-locus analysis

Consider a case-control genetic association study with n_m SNPs, for each pair of SNPs, a 2 × 9 contingency table with rows for disease status and columns for genotype configurations is created. A χ^2 test statistic and the corresponding *p*-value can subsequently be computed. With the total of n_m SNPs, there are $\binom{n_m}{2} = n_m!/((n_m - 2)!2!)$ possible SNP pairs. As a result, the *p*-value from each two-locus analysis must be adjusted by a Bonferroni correction. The Bonferroni-corrected *p*-value from each analysis is the lower value between $\binom{n_m}{2}$ × the uncorrected *p*-value and one.

Permutation test

The *p*-value p_0^e for the null hypothesis H_0^e that ensemble *e*--an ensemble of two-locus analyses of interest--is not associated with the disease can be evaluated by a permutation test. To achieve this, a scalar statistic is first computed from a function that combines the Bonferroni-corrected χ^2's *p*-values of individual two-locus tests. A suitable combining function must (a) be non-increasing in each *p*-value, (b) attain its maximum value when any *p*-value equals to zero and (c) have a finite critical value that is less than its maximum for any significant level greater than zero [54]. In this study, a Fisher's combining function ($-2\sum_i \log(p_i)$) is selected [55]. The *p*-value for the ensemble of two-locus analyses is assessed via a permutation simulation. In each permutation replicate, samples are constructed such that the case/control status of each sample is randomly permuted while the total numbers of



Case-Control Data

Sample	SNP1	SNP2	···	SNP1000	Status
1	0	1		2	Control
⋮					⋮
200	⋮	⋮	···	⋮	Control
201					Case
⋮					⋮
400					Case

Two-Locus Analysis (SNP1,SNP2)

Status	Genotype			
	AABB	*AaBB*	···	*aabb*
Control	63	35	···	0
Case	0	0	···	3

→ $p_0^{(1,2)}$ (Bonferroni-Corrected Chi-Square's *p*-value)

Permutation Test (Ensemble 3)

Data Set	Two-Locus Analysis			Fisher's Statistic
	(SNP1,SNP2)	(SNP2,SNP3)	(SNP1,SNP3)	
Original	$p_0^{(1,2)}$	$p_0^{(2,3)}$	$p_0^{(1,3)}$	T_0^3
Replicate 1	$p_1^{(1,2)}$	$p_1^{(2,3)}$	$p_1^{(1,3)}$	T_1^3
⋮	⋮	⋮	⋮	⋮
Replicate 10,000	$p_{10000}^{(1,2)}$	$p_{10000}^{(2,3)}$	$p_{10000}^{(1,3)}$	T_{10000}^3

→ p_0^3 (Raw *p*-value)

$$p_0^e = \left|\left\{i : 1 \le i \le 10000, T_i^e \ge T_0^e\right\}\right| / 10000$$

Global *P*-value Determination (4 Hypotheses)

Data Set	Ensemble 1		···	Ensemble 4	
	Fisher's Statistic	Unadjusted/ Raw *p*-value		Fisher's Statistic	Unadjusted/ Raw *p*-value
Original	T_0^1	p_0^1		T_0^4	p_0^4
Replicate 1	T_1^1	p_1^1		T_1^4	p_1^4
⋮	⋮	⋮	···	⋮	⋮
Replicate 10,000	T_{10000}^1	p_{10000}^1		T_{10000}^4	p_{10000}^4

$$p_i^e = \left|\left\{j : 0 \le j \le 10000, j \ne i, T_j^e \ge T_i^e\right\}\right| / 10000 \qquad p_i^{\min} = \min_e p_i^e \qquad p_0^{\min} = \min_e p_0^e$$

$$p_{\text{global}} = \left|\left\{i : 1 \le i \le 10000, p_i^{\min} \le p_0^{\min}\right\}\right| / 10000$$

Progressive Search

Iteration	Ensemble	Raw *p*-value(p_0^e)	Global *p*-value	Termination
1	{(SNP1,SNP2)}	0.0000	0.0000 (1 hypothesis)	No
2	{(SNP1,SNP2),(SNP2,SNP3)}	0.0000	0.0000 (2 hypotheses)	No
3	{(SNP1,SNP2),(SNP2,SNP3), (SNP1,SNP3)}	0.0000	0.0000 (3 hypotheses)	No
4	{(SNP1,SNP2),(SNP2,SNP3), (SNP1,SNP3),(SNP4,SNP5)}	0.0600	0.0000 (4 hypotheses)	Yes

Figure 1

Outline of 2LOmb. In this example, the algorithm takes a balanced case-control data set that consists of 400 samples and 1,000 SNPs. Each genotype is represented by an integer: 0 denotes a homozygous wild-type genotype, 1 denotes a heterozygous genotype and 2 denotes a homozygous variant or homozygous mutant genotype. A χ^2 contingency table is then constructed for each pair of SNPs in two-locus analysis. This results in the total of $\binom{1,000}{2}$ = 499,500 two-locus analyses. Thus, the Bonferroni-corrected χ^2's *p*-value for each two-locus analysis is the lower value between 499,500 × its uncorrected *p*-value and one. In one ensemble, Bonferroni-corrected χ^2's *p*-values from multiple two-locus analyses are combined together via a Fisher's combining function, which in turn provides a Fisher's test statistic result. The raw *p*-value for the ensemble is obtained through a permutation test, which is composed of 10,000 randomised permutation replicates. Since multiple ensembles may be tried during the identification of the best association explanation, a global *p*-value is calculated to account for multiple hypothesis testing. The global *p*-value is estimated through the same permutation test that gives the raw *p*-value for each ensemble. The progressive search for the best association explanation is carried out by incrementally adding a two-SNP unit to the current best ensemble. The condition for search termination is based on both the raw *p*-value for the explored ensemble and the global *p*-value. In this example, the search is terminated after the fourth ensemble is explored due to an increase in the raw *p*-value. Subsequently, the best SNP set for association explanation contains SNP1, SNP2 and SNP3 where the global *p*-value that accounts for testing of four hypotheses is $p < 0.0001$.

case and control samples remain unchanged. A χ^2 contingency table with new entries and a Bonferroni-corrected p-value for the two-locus analysis within each permutation replicate are then obtained. This, in turn, leads to a new Fisher's test statistic. Let T_i^e denote the value of Fisher's test statistic obtained for the ith permutation replicate, p_0^e is the fraction of permutation replicates with a test statistic greater than or equal to the test statistic obtained from the original case-control data (T_0^e). In other words,

$$p_0^e = |\{i : 1 \le i \le t, T_i^e \ge T_0^e\}| / t, \qquad (1)$$

where t is the number of permutation replicates which is set to 10,000 in this study and $|\cdot|$ denotes the size of a set.

Global p*-value determination*

There are many candidate ensembles of two-locus analyses that can be explored. Let $H_0 = \bigcap_{1 \le e \le E} H_0^e$ be the global null hypothesis that none of E explored ensembles of two-locus analyses is associated with the disease, the test of the global null hypothesis leads to the global p-value and provides the genetic association explanation. In step 2, the p-value p_0^e for a fixed hypothesis H_0^e is a raw or unadjusted p-value. To account for the correlation among multiple hypotheses that have been tested during the exploration through many candidate ensembles, the testing result of the global null hypothesis depends on $p_0^{\min} = \min_e p_0^e$. In other words, the global null hypothesis is rejected if the minimum of the raw p-values is sufficiently small. The distribution of $p_0^{\min}$ can again be determined by a permutation simulation. However, a nested simulation is unnecessary since the same set of permutation replicates for the p_0^e determination can be reused in the estimation of the empirical distribution of $p_0^{\min}$ [56]. This strategy has been successfully implemented in a number of genetic association detection techniques, including a set association approach [11] and a haplotype interaction approach embedded in FAMHAP [57,58]. The unadjusted p-value for the permutation replicate i of each hypothesis e is thus given by

$$p_i^e = |\{j : 0 \le j \le t, j \ne i, T_j^e \ge T_i^e\}| / t. \qquad (2)$$

Let $p_i^{\min} = \min_e p_i^e$ be the minimum of unadjusted p-values over all explored ensembles of two-locus analyses in the ith permutation replicate, the p-value for the global null hypothesis H_0 is defined by

$$p_{\text{global}} = |\{i : 1 \le i \le t, p_i^{\min} \le p_0^{\min}\}| / t. \qquad (3)$$

Search for the best ensemble of two-locus analyses

A simple progressive search is used to identify the best ensemble of two-locus analyses. The search begins by locating the best two-SNP unit with the smallest Bonferroni-corrected χ^2's p-value from step 1. A permutation test is then performed for this two-locus analysis, yielding both raw and global p-values since only one hypothesis has been explored. Next, the search attempts to combine the existing best two-SNP unit with the two-SNP unit that possesses the next smallest Bonferroni-corrected χ^2's p-value from step 1 and does not have a higher permutation p-value than the first two-SNP unit. If this new ensemble yields either a higher raw p-value or the same raw p-value but a higher global p-value from a permutation test, the search is terminated and the association is explained by the previously identified two-locus analysis. Otherwise, the best ensemble of two-locus analyses is updated and the process of appending more two-SNP units to the ensemble continues. The progressive search terminates when deterioration in the raw or global p-value is detected, or all possible two-locus analyses have been included in the ensemble. It is recalled from step 3 that for the best ensemble containing $E - 1 < \binom{n_m}{2}$ two-locus analyses, its global p-value is obtained from the evaluation of E hypotheses.

Validity of the algorithm

A permutation replicate in 2LOmb is constructed by randomly assigning the case or control status to each sample while maintaining the original proportion of case and control samples. Once the construction of a permutation replicate is finished, the assigned case and control labels remain fixed to the samples. The pattern of case and control labels in each permutation replicate is thus constant and unique. Therefore, the Bonferroni-corrected χ^2's p-values from any two-SNP units within a permutation replicate are calculated from the same case-control data set. Hence, the combining of these Bonferroni-corrected χ^2's p-values via a Fisher's combining function is attainable. The calculation of Fisher's test statistics from all permutation replicates and the original data set leads to the raw or unadjusted p-value p_0^e for the null hypothesis H_0^e of the ensemble e as given in equation 1. Since the same set of permutation replicates is always used during the evaluation of each ensemble, the raw p-values for the null hypotheses from all ensembles can be directly compared against one another. Furthermore, the global p-value calculation is based on this set of permutation replicates. This is possible because the unadjusted p-value for the permutation replicate i of ensemble e or p_i^e can be calculated in a similar manner to the raw p-value p_0^e as defined in equation 2. The unadjusted p-values for the same permutation replicate but different ensembles can also be directly compared and the subsequent calculation of $p_i^{\min} = \min_e p_i^e$ is attainable. With $p_i^{\min}$ and $p_0^{\min} = \min_e p_0^e$, the p-value for the global null hypothesis $H_0 = \bigcap_{1 \le e \le E} H_0^e$ that incorporates all E explored hypotheses can be determined by equation 3. In summary, only one set of permutation replicates is required for the calculation of both the raw p-value for the null hypothesis of every ensemble and the global p-value. The p-values can be compared in each step of 2LOmb. Consequently, the selection of the best ensemble for association explanation can be carried out via a p-value comparison.

Testing with simulated data

2LOmb is benchmarked against a simple exhaustive two-locus analysis technique, a set association approach (SAA) [11], a correlation-based feature selection (CFS) technique [14] and a tuned ReliefF (TuRF) technique [16] in a simulation trial. The exhaustive two-locus analysis is simply the two-locus analysis procedure from the first step of the 2LOmb algorithm. An interaction is declared if at least one two-SNP unit with a Bonferroni-corrected χ^2's p-value below 0.05 is detected. The exhaustive two-locus analysis reports all SNPs that meet this detection condition. The simulation covers two main data categories: null data of no significant genetic association and data with causative SNPs which signify pure epistasis. The algorithm performance on the null data provides an indication for the false-positive error. On the other hand, the algorithm performance on the data with causative SNPs indicates the detection capability. An efficient algorithm should produce an output with a low number of SNPs and a high number of correctly-identified causative SNPs when epistasis is present. Similarly, it should also report that there are no causative SNPs in the null data. These two measures on the number of SNPs in the results are used as the performance indicators.

Each simulated data set contains 1,000, 2,000 or 4,000 SNPs in which either there are no causative SNPs or there is pure epistasis, governed by two, three or four causative SNPs. The allele frequencies of all causative SNPs are 0.5 while the minor allele frequencies of the remaining SNPs are between 0.05 and 0.5. The data set consists of balanced case-control samples of sizes 400, 800 or 1,600. All SNPs in control samples are in Hardy-Weinberg equilibrium (HWE) [59]. The genotype distribution of causative interacting SNPs follows the pure epistasis model by Culverhouse et al. [42], leading to three interesting values of heritability: 0.01, 0.025 and 0.05. Every SNP in each data set exhibits no marginal single-locus effect (Bonferroni-corrected χ^2's p-value > 0.05). Twenty-five independent data sets for each simulation setting are generated via a genomeSIM package [60]. A paired t-test is suitable to assess the significance of results since the same simulated data sets are used during the algorithm benchmarking.

The results from the null data problem are summarised in Figure 2 while the results from the two-, three-and four-locus interaction problems are shown in Figures 3-4, 5-6 and 7-8, respectively. Clearly, 2LOmb significantly outperforms other techniques in terms of the low number of output SNPs, the high number of correctly-identified causative SNPs or both in every interaction problem (a paired t-test on 675 benchmark results yields a p-value < 0.05). On the other hand, both 2LOmb and SAA have the lowest false-positive error when compared to other techniques in the null data problem (a paired t-test on 225 benchmark results yields a p-value < 0.05). The statistical power analysis also reveals that the benchmark trial with 25 independent data sets for each simulation setting is sufficient for an accurate evaluation of the overall algorithm performance (power > 0.95 for a Type I error rate of 0.05). These results can be further interpreted as follows.

The performance of many existing attribute selection techniques for pattern recognition depends on the level of attribute interactions. A number of techniques, including CFS, appear to function well under a moderate level of interactions. However, the performance of CFS appears to be significantly reduced when the interaction level becomes too high [14,61] because CFS favours an attribute that is strongly correlated with the classification outcome--disease status in this study--while at the same time is not correlated with other attributes. Since the main driving force behind epistasis is the interaction between SNPs, which are themselves attributes, CFS would not intuitively select all causative SNPs. Consequently, the SNP set produced by CFS appears to contain only uncorrelated SNPs. Obviously, a SNP that is a part of the interaction model would occasionally be picked up by CFS but CFS never successfully identifies all causative SNPs in any interaction problems. In addition, CFS reports more erroneous SNPs than other techniques in the null data problem and all three interaction problems due to many SNPs being uncorrelated.

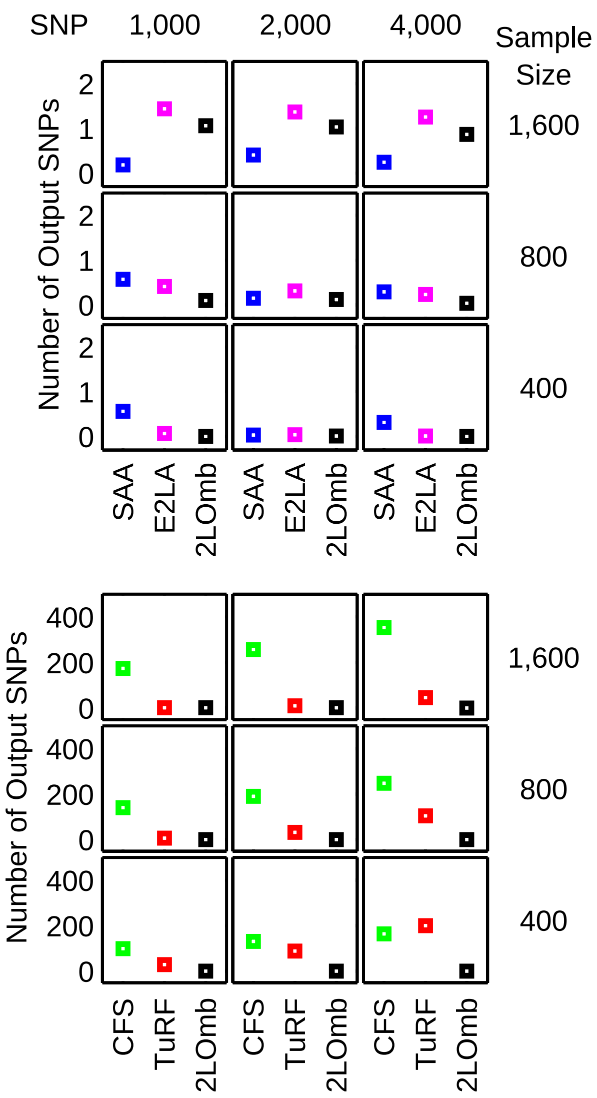


Figure 2

Performance of the exhaustive two-locus analysis, SAA, CFS, TuRF and 2LOmb in the null data problem. The results are averaged over 25 independent simulations. False detection is declared for the exhaustive two-locus analysis, SAA and 2LOmb if the *p*-values used as detection indicators in their results are less than 0.05. The results from the exhaustive two-locus analysis (E2LA), SAA, CFS, TuRF and 2LOmb are displayed using magenta, blue, green, red and black markers, respectively. In each chart, the horizontal axis represents the detection algorithm while the vertical axis represents the number of output SNPs reported by the algorithm. The top nine charts are displayed using a finer scale than the bottom nine charts.

The benchmarking of attribute selection techniques by Hall and Holmes [14] also reveals that ReliefF [15] is better than CFS in problems with a high level of interactions. Since ReliefF is essentially the core engine of TuRF, the results from this study are in agreement with the early benchmark trial. This finding strengthens the observation that the interaction level of SNPs in pure epistasis models is too high for CFS to handle. Similar to its predecessor, the performance of TuRF still depends on both the number of attributes and sample size. TuRF performs well in the majority of simulation scenarios with 1,000-2,000 SNPs and 800-1,600 samples. These scenarios are relatively easy since the number of SNPs is small while the sample size is large. However, the size of output SNP set, reported by TuRF from the null data problem and all three interaction problems, increases significantly when the difficulty level rises by either reducing the sample size or increasing the number of SNPs. This implies that when the problem contains a large number of candidate SNPs, the only way to ensure that TuRF reports a proper SNP set is to use a relatively large sample size, making it impractical in real genetic association studies due to many factors including disease prevalence, population size and genotyping cost.

The global *p*-values in most of the SAA results from the null data problem and all three interaction problems exceed 0.05, showing that SAA reports a low false-positive result in the null data problem. Nonetheless, SAA remains unsuitable for detecting pure epistasis because of its high false-negative error. This poor performance can be traced back to the manner in which SAA exploits an omnibus permutation test. As stated earlier, single-locus analysis does not detect any association between a SNP and the disease in this study. Hoh et al. [11] have demonstrated that genetic association can be more significantly observed when the single-locus test statistics are combined together. Nonetheless, there is an additional requirement that each causative SNP must exhibit a marginal single-locus effect. In the current study, the association signal from each causative SNP is lower than the required threshold, leading to similar test statistics and global *p*-values for both combinations of multiple SNPs which include causative SNPs and those which exclude causative SNPs.

Both 2LOmb and exhaustive two-locus analysis technique are capable of identifying all causative SNPs. However, the size of output SNP set from 2LOmb is significantly smaller than that from the exhaustive two-locus analysis. Appended SNPs to the causative SNPs in the output from 2LOmb and those from the exhaustive two-locus analysis are erroneous SNPs. These erroneous SNPs are parts of false two-SNP units with Bonferroni-corrected χ^2's *p*-val-

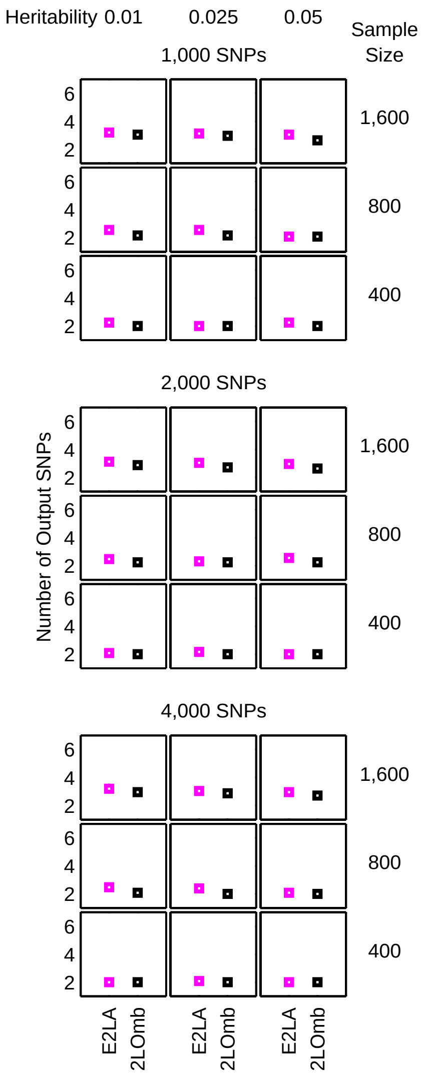


Figure 3
Performance of the exhaustive two-locus analysis and 2LOmb in the two-locus interaction problem. The results are averaged over 25 independent simulations. Detection is declared for the exhaustive two-locus analysis and 2LOmb if the *p*-values used as detection indicators in their results are less than 0.05. The results from the exhaustive two-locus analysis (E2LA) and 2LOmb are displayed using magenta and black markers, respectively. In each chart, the horizontal axis represents the detection algorithm while the vertical axis represents the number of output SNPs reported by the algorithm. All causative SNPs are present in outputs from both the exhaustive two-locus analysis and 2LOmb in all simulations.

ues less than 0.05. A similar trend of results regarding the size of output SNP set is also observed in the benchmark trial involving the application of 2LOmb and exhaustive two-locus analysis to the null data. This signifies that the permutation test and the progressive search embedded in 2LOmb can help reducing the number of erroneous SNPs in the output.

As mentioned earlier, 2LOmb produces the best results among five techniques in the benchmark trial. 2LOmb has a low false-positive error in the null data problem and is capable of detecting all causative SNPs in every simulated data set in all three interaction problems. This performance is further strengthened by highly significant global *p*-values in 2LOmb results from all three interaction problems ($p < 0.0001$) and the presence of a SNP in common among some or all pairs of two-SNP units in the three- and four-locus interaction problems. Nonetheless, some of the 2LOmb outputs contain a few erroneous SNPs which are irrelevant to the correct association explanation. Since all three interaction problems involving different numbers of causative SNPs are investigated by varying the total number of SNPs, the sample size and the level of heritability, these parameters may influence the number of erroneous SNPs in the 2LOmb results. Similarly, the total number of SNPs and the sample size may affect the number of erroneous SNPs in the 2LOmb results from the null data problem. ANOVA reveals that the only source of variation that significantly affects the number of erroneous SNPs in the null data, two-locus interaction and three-locus interaction problems is the sample size ($p < 0.000001$). In addition, the sample size must be greater than 800 for an increase in the number of erroneous SNPs to be significant. In contrast, ANOVA reveals that two sources of variation that affect the number of erroneous SNPs in the four-locus interaction problem are the sample size ($p < 0.000001$) and the total number of SNPs ($p < 0.00005$). Similar to the null data, two-locus interaction and three-locus interaction problems, the sample size in the four-locus interaction problem must be greater than 800 to create a significant increase in the number of erroneous SNPs. On the other hand, the number of erroneous SNPs appears to decrease when the total number of SNPs increases. These two sources of variation also interact with each other ($p < 0.005$). However, the interaction is most evident only when the sample size is large, i.e. when the sample size is 1,600.

ANOVA shows that the number of erroneous SNPs in the 2LOmb results is influenced by the sample size and the total number of SNPs but not by the heritability. It is observed that the number of erroneous SNPs increases when the sample size is large. This counterintuitive phenomenon can be explained as follows. As 2LOmb com-

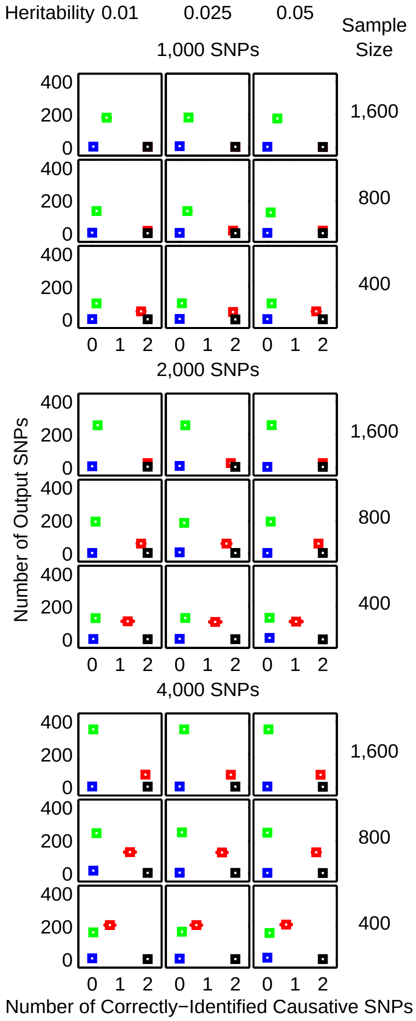


Figure 4

Performance of SAA, CFS, TuRF and 2LOmb in the two-locus interaction problem. The results are averaged over 25 independent simulations. Detection is declared for SAA and 2LOmb if the *p*-values used as detection indicators in their results are less than 0.05. The results from SAA, CFS, TuRF and 2LOmb are displayed using blue, green, red and black markers, respectively. In each chart, the horizontal axis represents the number of correctly-identified causative SNPs while the vertical axis represents the number of output SNPs reported by the algorithm. The charts on which the red markers are invisible denote the situations in which the performance of TuRF and 2LOmb is similar. The charts in this figure are displayed using a coarser scale than the charts in Figure 3.

bines *p*-values that are determined from χ^2 tests, the number of entries for the contingency table construction is large when the sample size is large. This subsequently leads to a significantly large χ^2 statistic and hence an extremely small *p*-value if the SNPs under consideration are causative SNPs. At the same time, the possibility that a reasonably large χ^2 statistic and a small *p*-value can be obtained by chance from a two-SNP unit which is irrelevant to the correct association explanation also inevitably increases. With the increase in the possibility of erroneous SNP inclusion, the size of output SNP set gets bigger when the sample size is large. Another observation that appears to be counterintuitive is the reduction in the number of erroneous SNPs when the total number of SNPs increases. This phenomenon is the result of the Bonferroni correction usage. When the total number of SNPs is doubled, the Bonferroni correction factor in 2LOmb is quadrupled. A higher correction factor leads to a more stringent criterion for SNP selection. This subsequently leads to the reduction in the number of erroneous SNPs when the total number of SNPs is large.

In contrast to the first two parameters, different levels of heritability appear to have no effect on the 2LOmb results because all simulated data sets have balanced case-control samples and the embedded interaction models have the same architecture. For instance, a two-locus interaction model leads to zero penetrances for genotypes *AABB*, *AABb*, *AaBB*, *Aabb*, *aaBb* and *aabb*. Hence, the penetrances for these six genotypes are always equal to zero regardless of the heritability. On the other hand, genotypes *AAbb*, *AaBb* and *aaBB* have non-zero penetrances (see Methods for details). Therefore, different heritability levels certainly lead to different penetrances for genotypes *AAbb*, *AaBb* and *aaBB*. However, the ratios between the penetrances of these three genotypes are fixed and independent of the heritability. This model description can be generalised to cover the other multi-locus interaction models. In addition, the maximum penetrance in any two-locus or multi-locus interaction models always stays below 0.1 even though the heritability is at the highest level (see Methods for details). This means that case samples are always over-sampled from affected individuals to achieve a balanced case-control data set. Since all explored heritability levels lead to the same case over-sampling pattern, the simulated data sets of which the only primary difference being the heritability levels are indistinguishable from one another. This leads to the result similarities in interaction problems with the same number of SNPs in the data set, sample size and number of causative SNPs but different levels of heritability as shown in Figures 3, 4, 5, 6, 7 and 8. The result trend is also independent of the number of simulated data sets used in the benchmark trial.

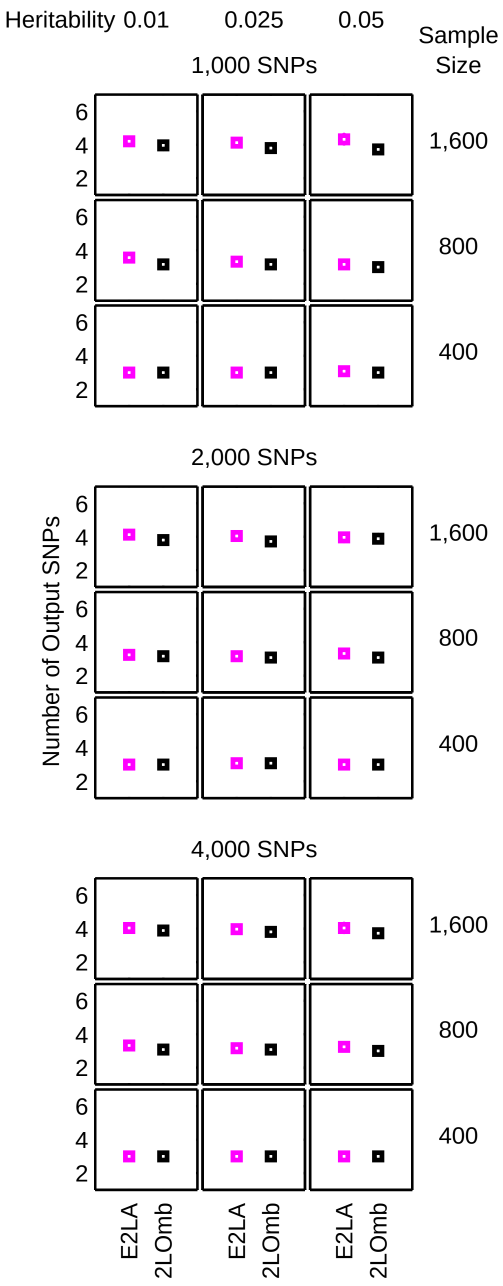


Figure 5

Performance of the exhaustive two-locus analysis and 2LOmb in the three-locus interaction problem. The explanation for how the results are obtained and displayed is the same as that given in Figure 3.

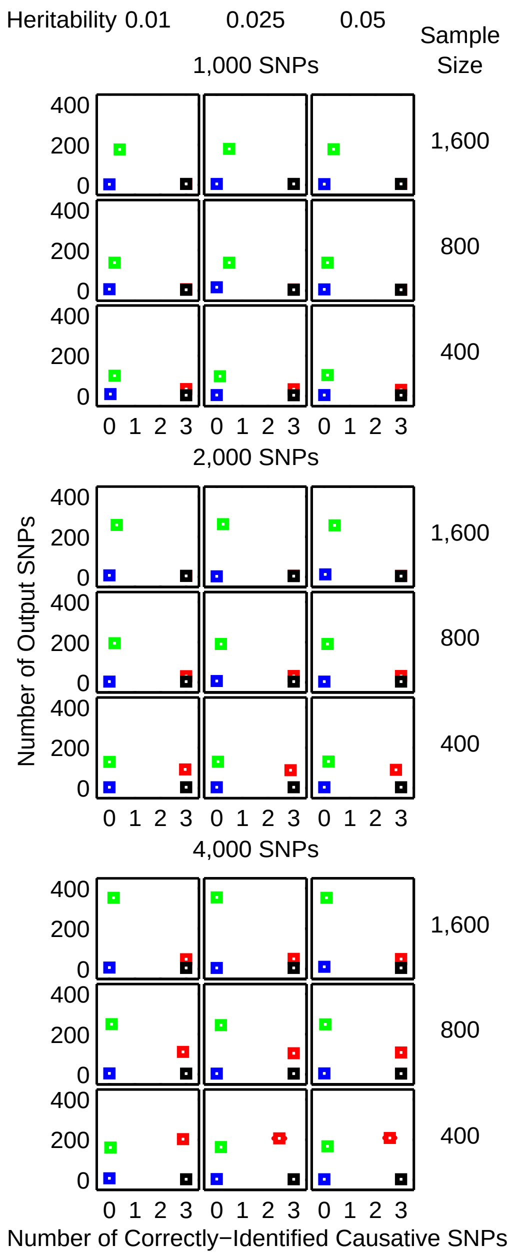


Figure 6

Performance of SAA, CFS, TuRF and 2LOmb in the three-locus interaction problem. The explanation for how the results are obtained and displayed is the same as that given in Figure 4.

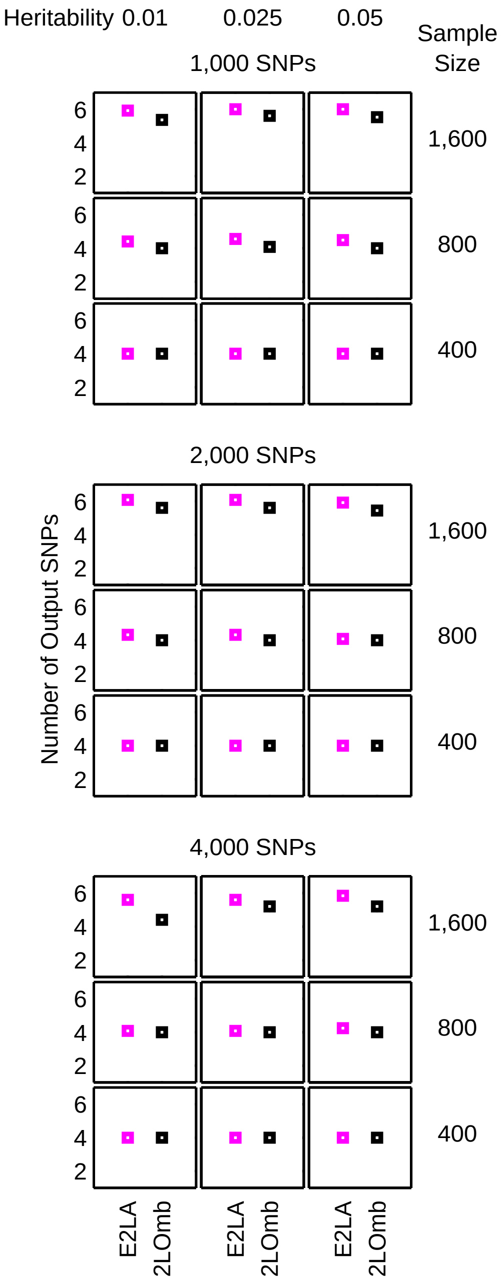


Figure 7

Performance of the exhaustive two-locus analysis and 2LOmb in the four-locus interaction problem. The explanation for how the results are obtained and displayed is the same as that given in Figure 3.

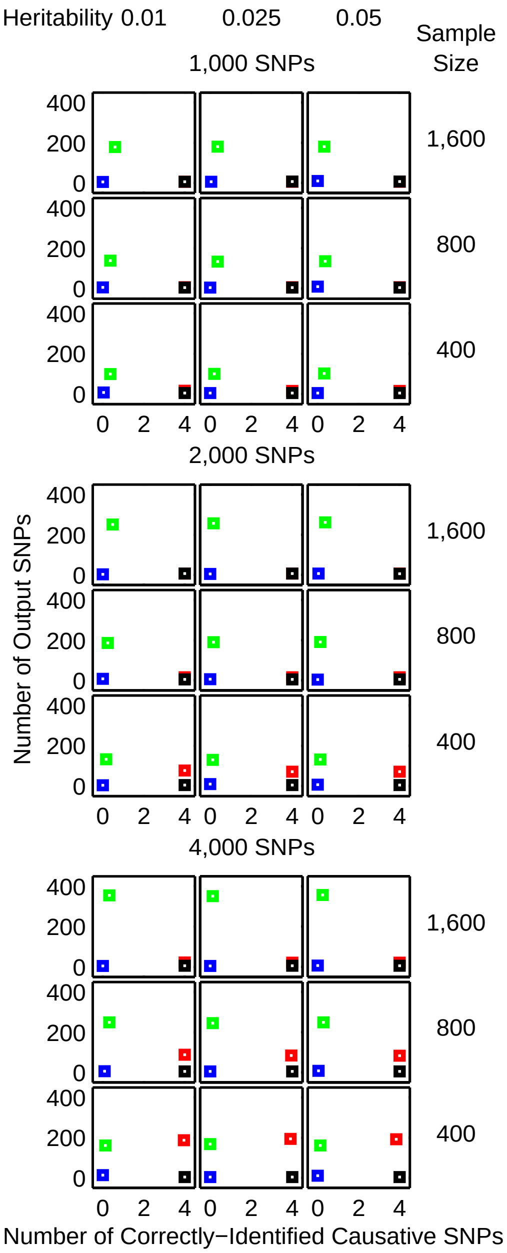


Figure 8

Performance of SAA, CFS, TuRF and 2LOmb in the four-locus interaction problem. The explanation for how the results are obtained and displayed is the same as that given in Figure 4.

In a permutation test, the ability to differentiate between two *p*-values is influenced by the number of permutation replicates. With *t* permutation replicates, the test declares an actual *p*-value that is less than $1/t$ to be zero. During the progressive search for the best ensemble, the inclusion of a new two-SNP unit is accepted if this inclusion does not worsen the current result. If the number of permutation replicates is too low, the search may include erroneous two-SNP units that are irrelevant to the correct association explanation. The analysis is confirmed as the number of output SNPs from 2LOmb is equal to the number of causative SNPs in most of simulation results. This phenomenon suggests that the number of permutation replicates employed in this study (t = 10,000) is high enough to screen off most of the erroneous two-SNP units. In other words, the inclusion of these erroneous two-SNP units leads to an increase in the *p*-value by at least $1/t$. Nonetheless, the fact that 2LOmb results are not entirely free from erroneous SNPs suggests that there are erroneous two-SNP units with extremely small *p*-values. It is advisable to perform a genotype relative risk calculation for the elimination of erroneous SNPs. If the presence of an erroneous two-SNP unit is suspected, its result on two-locus genotype relative risk would not be as significant as that from the other two-SNP units in the ensemble. Alternatively, an additional means for further SNP screening by other techniques such as MDR is also recommended. The chance of erroneous SNP discovery would be further minimised by employing two consecutive attribute selection techniques. The same concept has been adopted for the implementation of MDR software, in which many additional filters including a χ^2 test, an odds ratio test, ReliefF and TuRF are available for SNP screening prior to the MDR analysis.

The two-, three- and four-locus interaction data sets which have been screened for causative SNPs by 2LOmb are subsequently subjected to MDR analysis. MDR has successfully identified all erroneous SNPs and the correct interaction models have been constructed from all data sets. The prediction accuracy from the MDR analysis is illustrated in Figure 9. It is noted that the prediction accuracy from all data sets is quite high due to the manner in which the pure epistasis model is defined [42]. Using the penetrance table for a two-locus interaction model with the heritability = 0.01 (see Methods for details), the two-locus genotype distribution of causative SNPs in a balanced case-control sample set from simulated data with 800 samples can be estimated and shown in Figure 10.

Six genotypes in Figure 10 namely *AABB*, *AABb*, *AaBB*, *Aabb*, *aaBb* and *aabb* are protective genotypes. In other words, a sample with one of these six genotypes is a control sample because the penetrances for these genotypes are zero. It is also noted that the control samples with all nine genotypes precisely follow the distribution as jointly described by independent single-locus genotype distribution from loci A and B. In contrast, three remaining genotypes in Figure 10 namely *AAbb*, *AaBb* and *aaBB* are labelled as disease-predisposing genotypes because the majority of samples with these three genotypes are case samples. Samples with these genotypes may be either case or control samples because the penetrances for these genotypes are between zero and one. In fact, the probabilities that persons with these genotypes to have the disease are quite low since the penetrances for these genotypes are small. However, case samples must be over-sampled from affected individuals to ensure a balanced case-control data set because the disease prevalence for this two-locus interaction model is only 0.004975. In addition, each case sample must contain one of these three genotypes because the penetrances for the other genotypes are zero. As a result, the case samples with these genotypes do not follow the same two-locus genotype distribution as in the control samples. With six genotypes being exclusively specific to control samples and the majority of three remaining genotypes being found in case samples, the MDR prediction accuracy for the two-locus interaction model is high. This explanation can also be generalised to cover the MDR results from the other multi-locus interaction data sets.

Another advantage of using 2LOmb for SNP screening prior to the MDR analysis is the reduction in computational time for interaction detection. The computational time for 2LOmb to finish screening the SNPs is provided to demonstrate this strength of 2LOmb. Moreover, the computational time required to identify causative SNPs by the MDR analysis and that by the combined approach which involves SNP screening by 2LOmb and follows by the MDR analysis is given. The previously-described simulated data sets with causative SNPs are used to produce the computational time results from the SNP screening by 2LOmb and the combined approach. All possible interaction models that can be constructed from the 2LOmb outputs are explored by MDR in the combined approach. On the other hand, the data sets for the direct MDR analysis are prepared by restricting the number of SNPs in each data set to 100. Only SNPs that are irrelevant to the correct association explanation are removed from the original simulated data sets. Furthermore, MDR only explores the interaction models that do not cover more than four SNPs in the data for this latter simulation setting. The summary of computational time required for the SNP screening by 2LOmb and that for both direct MDR and combined approaches to correctly identify all causative SNPs is given in Table 1. The maximum time required by 2LOmb to screen SNPs in the largest data set is 419 seconds or approximately seven minutes. Moreover, the combined 2LOmb and MDR approach discovers the correct causative SNPs much faster than MDR. This time reduction is achieved even though the problems have been simplified for the direct MDR analysis. A direct application of MDR to the original simulated data sets is certainly impractical.

The simulated multi-locus interaction problems in this article are based on the pure epistasis model by Culverhouse et al. [42]. It is possible to capture a number of multi-locus interactions with marginal two-locus effects via a combination of two-locus analyses. However, there are many multi-locus interaction scenarios without marginal two-locus effects. In such cases, 2LOmb and the exhaustive two-locus analysis technique are unable to detect interactions. Among the explored techniques, TuRF and MDR have a better chance of detection. Nonetheless, TuRF functions well only when the total number of SNPs in data is small and the sample size is large enough while the total number of SNPs in data affects the practicality of direct MDR analysis.

Every attribute selection technique has a limitation in terms of the maximum numbers of samples and attributes that it can handle. Single-locus analysis techniques always have a higher limit than multi-locus analysis techniques.

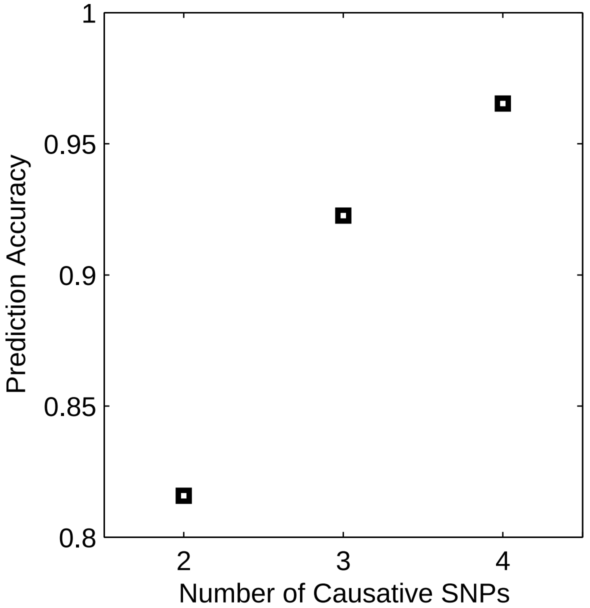


Figure 9
Prediction accuracy from the MDR analysis. A 10-fold cross-validation strategy is applied during the accuracy evaluation. The best MDR model is located by exploring all possible SNP combinations. All erroneous SNPs, which are left over after the screening by 2LOmb, have been successfully identified. All MDR models contain the correct number of causative SNPs. In addition, the MDR cross-validation consistency is 10/10.

Because attribute subset evaluation is usually integrated into multi-locus analysis techniques, consequently the number of possible attribute subsets that can be explored is extremely large when the candidate attribute set is large. Together with a potentially large sample size, a higher computational requirement for multi-locus analysis techniques is inevitable. As a result, the direct application of multi-locus analysis techniques to a much larger data set than those presented in this article, which is usually considered in genome-wide association studies, would be impractical. However, it is reasonable to expect that both marginal single-locus and epistatic effects are present in any genome-wide data sets. A multi-stage strategy that incorporates multiple techniques, designed for different detection modes, would be more suitable to handle large data. For instance, the marginal single-locus effects should be the first priority and, as such, be detected by single-locus analysis. Then, a special case of pure epistasis [2] or semi-purely epistatic events, in which a SNP displaying a marginal single-locus effect interacts with a SNP that exhibits no marginal single-locus effect, should be considered. Many two-locus analysis techniques have been proven to be well suited to this type of epistasis [40,50,51]. Finally, the detection of pure epistasis is carried out in the last stage. With the reduction of SNPs from the first stage, the chance that some multi-locus analysis techniques are applicable to the remaining SNPs increases. In addition to the multi-stage approach, a prior knowledge regarding the previously reported association can be exploited to select candidate genes based upon ontology and pathways. This practice is due to the necessity for the derivation of plausible interpretation. The screening for SNPs within or near candidate genes before the association detection also increases the chance that multi-locus analysis techniques can be applied to the remaining data.

Testing with real data

2LOmb has been applied to study a type 2 diabetes mellitus (T2D) data set, collected and investigated by the Wellcome Trust Case Control Consortium (WTCCC) [3]. The data set consists of 1,999 case samples from affected individuals in the UK and 3,004 control samples, which are the results of a merging between 1,500 samples from the UK blood services and 1,504 samples from the 1958 British birth cohort. The original genome-wide data set contains 500,568 SNPs that are obtained through the Affymetrix GeneChip 500 K Mapping Array Set. The SNP set is primarily reduced by screening for SNPs within and near 372 candidate genes collected by the Human Genome Epidemiology Network (HuGENet) [62]. These candidate genes cover genes from both positive and negative genetic association reports, in which studies are conducted in various ethnic groups and populations. The SNP set is further reduced by removing SNPs that exhibit

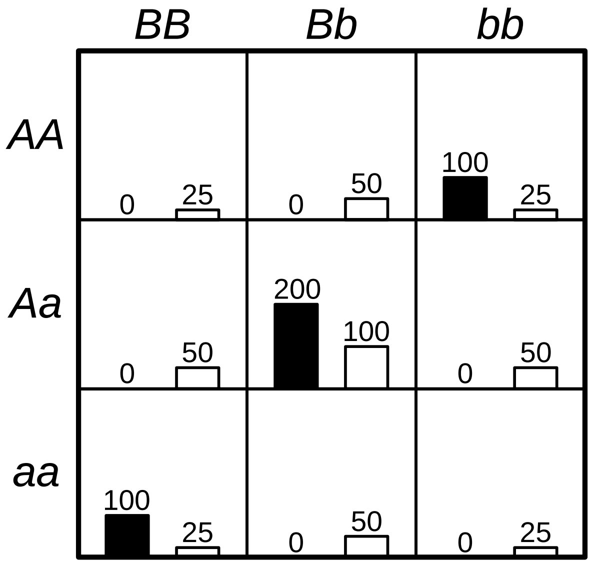


Figure 10
Genotype distribution of two causative SNPs in a balanced case-control data set with the sample size of 800. The left (black) bar in each cell represents the number of case samples while the right (white) bar represents the number of control samples. The cells with genotypes *AABB*, *AABb*, *AaBB*, *Aabb*, *aaBb* and *aabb* are labelled as protective genotypes while the cells with genotypes *AAbb*, *AaBb* and *aaBB* are labelled as disease-predisposing genotypes.

strong evidence of genetic association via single-locus analysis. The final SNP set contains 7,065 SNPs from 370 candidate genes. All SNPs in the reduced data set exhibit no marginal single-locus effects (Bonferroni-corrected χ^2's *p*-value > 0.05). Detailed description of the final SNP set is given in the supplement (see Additional file 1).

The 2LOmb search in the reduced T2D data set takes 3,456 seconds (57.6 minutes) of computational time on the Beowulf cluster. The possible genetic association is detected from 11 intronic SNPs in four genes (global *p*-value < 0.0001). Details of these SNPs, the two-SNP units that exhibit marginal two-locus effects and the identified genes are given in Table 2. A two-SNP unit is located in *LMX1A*. A two-SNP unit is also detected in *PARK2*. In addition, there is one SNP in common among SNPs in both *GYS2* two-SNP units. Similarly, there is one common SNP among three two-SNP units located in *PGM1*. Nonetheless, a two-SNP unit in which each SNP is located in a different gene is absent, indicating that there is no evidence of gene-gene interactions which can be observed from the 2LOmb result. Linkage disequilibrium (LD) analysis is subsequently performed using a JLIN package [63] and the resulting LD patterns are illustrated in Figure 11. It is noted that there is strong LD among SNPs within each gene due to high values of *D*' [64] and r^2 [65]. The genotype and haplotype relative risks are then calculated and the results are presented in Tables 3, 4, 5, 6, 7, 8, 9 and 10. Haplotype inference is carried out using an expectation-maximisation method [66]. The analysis reveals that a more prominent indication of a relative risk is observed when two-SNP units are considered. It is also noted that the genotype relative risk is directly influenced by the haplotype relative risk once a genotype is phased into all possible haplotype pairs. The detection of these two-SNP units is thus believed to be the consequence of haplotype effects. An early T2D association study also reveals similar haplotype effects in FUSION data [67]. Next, an interaction dendrogram [68,69] constructed from the 11 SNPs by MDR software is given in Figure 12. A strong synergistic effect between the two SNPs in *PARK2* is clearly observed. In contrast, the interactions between *PGM1*, *LMX1A*, *PARK2* and *GYS2* are clearly absent.

Since many early genetic association studies of T2D and metabolic syndrome employ MDR analysis [43-45,47], additional MDR analysis would be useful for the comparison. The screened T2D case-control data set which contains 11 SNPs identified by 2LOmb is further subjected to MDR analysis. The prediction accuracy of the best MDR model is summarised in Table 11. The model covers six SNPs in three genes: PGM1, PARK2 and GYS2. These SNPs are also present in three two-SNP units identified by 2LOmb. It is noted that the prediction accuracy in this real data set is much less than that from the simulated data sets. Nevertheless, the attainment of low prediction accuracy does not necessarily suggest that there is no genetic association. Early works involving genetic association studies of T2D and metabolic syndrome in various populations via MDR analysis produce similar values of prediction accuracy as summarised in Table 12. The prediction accuracy by MDR from most studies is in the range of 0.5-0.6. The only genetic association study of T2D that the prediction accuracy is distinctively high is conducted in a Korean population [43]. Differences in genetic background, candidate genes and selected SNPs are the main causes of variation in the genetic association results. Although MDR does not select five SNPs from the 2LOmb output, these SNPs should not be regarded as erroneous SNPs because there is strong linkage disequilibrium among SNPs in each gene. Moreover, early genotype and haplotype relative risk analysis clearly indicates that each gene, identified by 2LOmb, plays a role in the T2D association explanation. Overall, the analysis with the methods above only confirms the positive association for PGM1, LMX1A, PARK2 and GYS2 while gene-gene interactions are clearly absent. This signifies that, for the current study, there is no interaction between each pair of the identified genes that can be described by purely epistatic two-locus interaction models. In addition, there are no interactions

Table 1: Computational time required by 2LOmb, a combined 2LOmb and MDR approach, and direct MDR analysis to detect interactions in simulated data sets with different sizes and different numbers of causative SNPs.

		Computational Time Required by Each Approach (sec)						
		2LOmb			2LOmb+MDR			MDR
Number of Causative SNPs	Sample Size	1,000 SNPs	2,000 SNPs	4,000 SNPs	1,000 SNPs	2,000 SNPs	4,000 SNPs	100 SNPs
2	400	15	37	135	17	39	137	7,656
	800	21	59	224	23	61	226	15,990
	1,600	36	106	400	38	108	402	31,222
3	400	22	43	140	24	45	142	7,721
	800	30	65	229	32	67	231	16,206
	1,600	50	115	406	52	117	408	31,232
4	400	32	55	150	34	57	152	7,841
	800	46	80	236	48	82	238	16,285
	1,600	70	133	419	72	135	421	31,637

Only one computing processor in a Beowulf cluster is occupied during the analysis of one data set. The test problems for the direct MDR analysis have been simplified by reducing the number of SNPs in each data set to achieve attainable computational time. The displayed time is collected from the processing of multiple independent data sets for each simulation setting. The computational time from the benchmark trial involving 2LOmb, and the combined 2LOmb and MDR approach is the maximum time needed by each method to detect interactions in one data set. In contrast, the computational time from the direct MDR analysis is the minimum time for the completion of interaction detection in one data set. The computational time required by 2LOmb for the null data problem is similar to that for the two-locus interaction problem.

between these four genes that can be described by purely epistatic multi-locus interaction models with marginal two-locus effects.

The four genes selected by 2LOmb regulate many pathways that involve in the disease development [70-72]. The genetic association studies involving these genes have been previously conducted in different populations. For instance, *LMX1A* has been chosen as a positional and biological candidate gene for a case-control study of T2D in Pima Indians [73]. This gene is chosen as a candidate because a linkage of T2D to chromosome 1q21-q23 has been previously reported [74]. In addition, *LMX1A* is one of LIM homeobox genes that are expressed in pancreas and has been shown to activate insulin gene transcription. Although SNPs have been carefully selected from the entire gene, no association between these SNPs in *LMX1A* and T2D has been found in this ethnic group.

PARK2 is another candidate gene that is also selected for case-control studies, based on evidence from genome-wide linkage analysis [75]. A linkage of T2D in an African American population to chromosome 6q24-q27 has been previously identified [76]. Although *PARK2* mainly involves in the development of Parkinson's disease, single-locus analysis reveals strong evidence of association between SNPs, which are in the vicinity of SNPs identified by 2LOmb, and T2D in African Americans.

In contrast to *LMX1A* and *PARK2*, which are candidate genes in typical T2D case-control studies, *GYS2* is considered in a study to identify genes responsible for troglitazone-associated hepatotoxicity in Japaneses with T2D [77]. In other words, both case and control samples in the study are drawn from troglitazone-treated T2D patients, in which case patients exhibit an abnormal increase in alanine transaminase (ALT) and aspartate transaminase (AST) levels. *GYS2* regulates starch and sucrose metabolism and an insulin signalling pathway. The selected SNPs in *GYS2* are not found to associate with troglitazone-induced hepatotoxicity.

Similar to the study of *GYS2*, the association study involving *PGM1* is not carried out as a typical T2D case-control study. In fact, an attempt to identify association between *PGM1* polymorphisms and obesity has been conducted among T2D affected individuals in Italy [78]. *PGM1* regulates glycolysis and gluconeogenesis, starch and sucrose metabolism, galactose metabolism, a pentose phosphate pathway, and streptomycin biosynthesis. Isozyme polymorphisms [79,80], which are defined through structural differences in PGM1 protein, are used instead of SNPs in the study where positive association is identified.

In summary, positive association has been reported from previous studies involving *PARK2* in African Americans and *PGM1* in Italians. In contrast, negative association has been reported from previous studies about *LMX1A* in Pima Indians and *GYS2* in Japaneses. Both *GYS2* and *PGM1* regulate starch and sucrose metabolism while *LMX1A* and *PARK2* govern insulin gene transcription and Parkinson's disease development, respectively. The above discussion strengthens the importance of conducting large-scale association studies due to two main reasons. Firstly, a gene that does not contribute to the aetiology of a complex disease in one population may be important for association explanation in another population. Secondly, the absence of interacting candidate genes from a study may lead to negative association due to a lack of necessary genetic information. A two-locus interaction can occur between SNPs from genes that regulate one specific pathway [44] or between SNPs from genes that regulate different pathways [45]. Furthermore, a multi-locus interaction may involve both SNPs from genes that regulate the same pathway and SNPs from genes that govern different pathways. Hence, candidate genes should be selected by considering all pathways that directly and indirectly contribute to the disease development.

This study produces evidence of association between 11 intronic SNPs in *PGM1*, *LMX1A*, *PARK2* and *GYS2*, and T2D in a UK population. Although there are other independent genome-wide T2D data sets, the association detection within these data using a similar methodology to the presented method has never been attempted because the methodology employed in the majority of genome-wide association studies is based on single-locus analysis [3,81]. It is recalled that each SNP explored in the reduced T2D data set exhibits no marginal single-locus effect. Hence, the most logical approach to confirm the possibility of replicating association results from the current study is to perform the same detection method on these independent data sets. This is certainly important to gain further understanding of the genetic role in T2D susceptibility.

Implementation

2LOmb is implemented in a C programming language. All functions within the program are written by the first author except the χ^2 distribution function, which is taken from the Numerical Recipes in C [82]. The program can be compiled by Microsoft Visual Studio and GNU C compilers. The program has been successfully tested for the execution under Windows and Linux operating systems. The time required by 2LOmb to complete a problem containing n attributes is $T(n) = \binom{n}{2} = n!/((n-2)!2!) = n(n-1)/2$. 2LOmb thus has the order of n^2 time complexity ($T(n) \in O(n^2)$). Consequently, 2LOmb can tackle problems in quadratic time. 2LOmb in its present form occupies one processor during the program execution. A parallel version of 2LOmb for genome-wide data is under development. All results included in the study are collected from the execution of computer programs in a Beowulf cluster. The computational platform consists of 12 nodes. Each node is equipped with dual Xeon 2.8 GHz processors and 4GB of main memory. The Rocks Cluster Distribution is installed on all nodes.

Conclusion

In this article, a method for detecting epistatic multi-locus interactions in case-control data is presented. The study focuses on pure epistasis [2], which cannot be detected via single-locus analysis [42]. To overcome this difficulty, the proposed method performs an omnibus permutation test [54] on ensembles of two-locus analyses and is thus referred to as 2LOmb. The detection performance of 2LOmb is evaluated using both simulated and real data. From the simulation, 2LOmb produces a low false-positive error when the tests on null data of no association are performed. Furthermore, 2LOmb can identify all causative SNPs and outperforms a simple exhaustive two-locus analysis technique, a set association approach (SAA) [11], a correlation-based feature selection (CFS) technique [14] and a tuned ReliefF (TuRF) technique [16] in various interaction scenarios with marginal two-locus effects. These scenarios are set up by varying the number of causative SNPs, the number of SNPs in data, the sample size and the heritability. ANOVA reveals that the number of SNPs in data and the sample size influence the number of erroneous SNPs appended to the correctly-identified causative SNPs in the 2LOmb output. In contrast, the results from 2LOmb appear to be insensitive to the variation in heritability. After subjecting the data sets containing only SNPs that are screened by 2LOmb to multifactor dimensionality reduction (MDR) analysis [19], all erroneous SNPs are successfully removed. In addition, an insight into the MDR models is provided. 2LOmb is subsequently applied to a real case-control type 2 diabetes mellitus (T2D) data set, which is collected from a UK population by the Wellcome Trust Case Control Consortium (WTCCC) [3]. The original genome-wide data set is first reduced by selecting only SNPs that locate within or near 372 candidate genes reported by the Human Genome Epidemiology Network (HuGENet) [62]. In addition, the selected SNPs must exhibit no marginal single-locus effects. The final data set, which consists of 1,999 case samples and 3,004 control samples, contains 7,065 SNPs from 370 candidate genes. 2LOmb identifies 11 intronic SNPs that are associated with the disease. These SNPs are located in *PGM1*, *LMX1A*, *PARK2* and *GYS2*. The 2LOmb result suggests that there is no interaction between each pair of the identified genes that can be described by purely epistatic two-locus interaction models. Moreover, there are no interactions between these four genes that can be described by purely epistatic multi-locus interaction models with marginal two-locus effects. This evidence of genetic association for these four genes leads to an alternative explanation for the aetiology of T2D in the UK population. It also implies that SNPs from genome-wide data which are usually discarded after single-locus analysis confirms the null hypothesis of no association can still be useful for genetic association studies of complex diseases.

Methods

Pure epistasis model

The pure epistasis model of interest is proposed by Culverhouse et al. [42]. The model describes a restriction or constraint for penetrance of each genotype constituting the

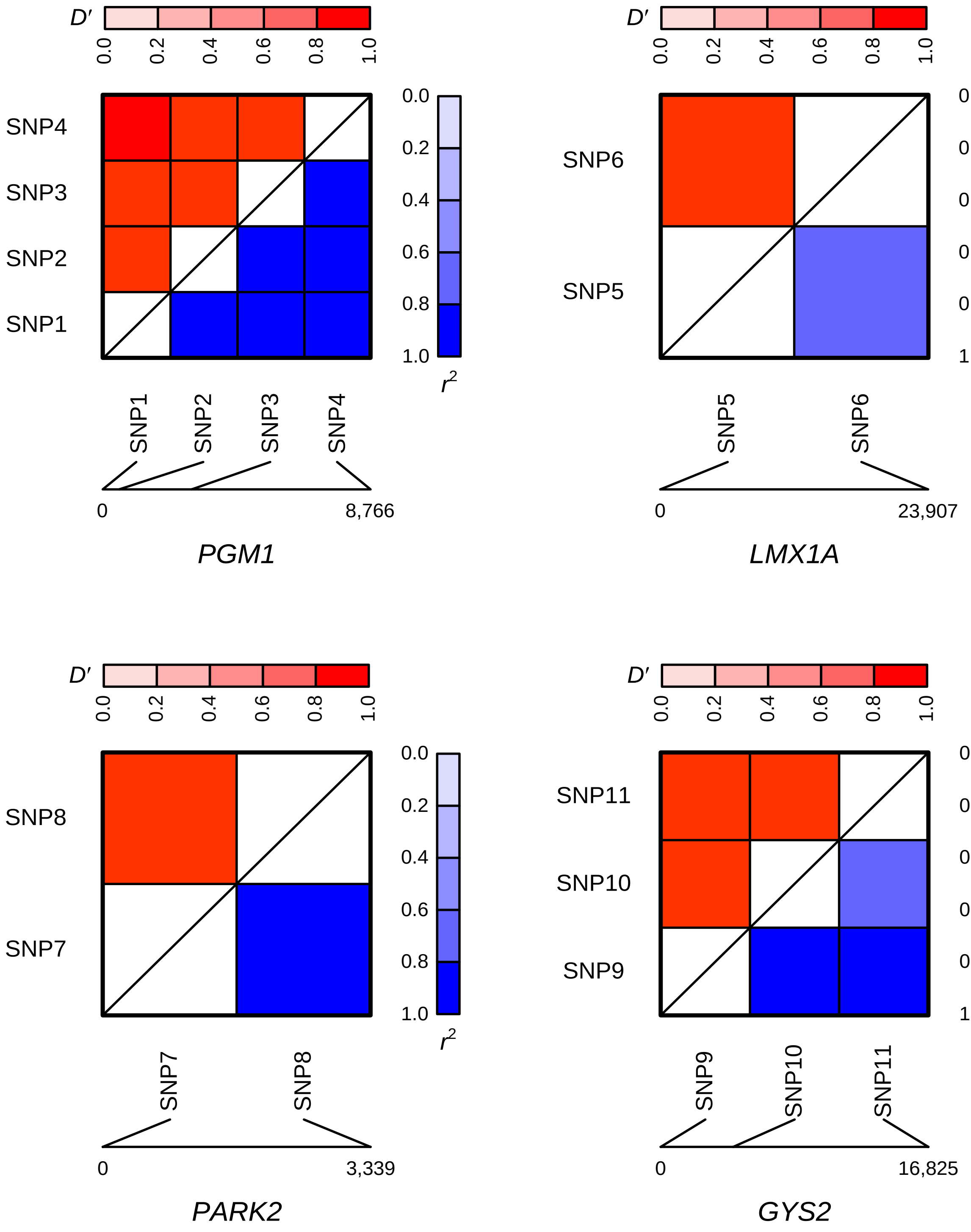


Figure 11

Linkage disequilibrium (LD) patterns of SNPs in *PGM1*, *LMX1A*, *PARK2* and *GYS2*. LD is explained via *D*' displayed in the upper triangle and r^2 displayed in the lower triangle. Dark colours indicate high values while pale colours indicate low values. Distances between SNPs are given in terms of the number of base pairs. SNP1 = rs2269241, SNP2 = rs2269239, SNP3 = rs3790857, SNP4 = rs2269238, SNP5 = rs2348250, SNP6 = rs6702087, SNP7 = rs1893551, SNP8 = rs6924502, SNP9 = rs6487236, SNP10 = rs1871142 and SNP11 = rs10770836.

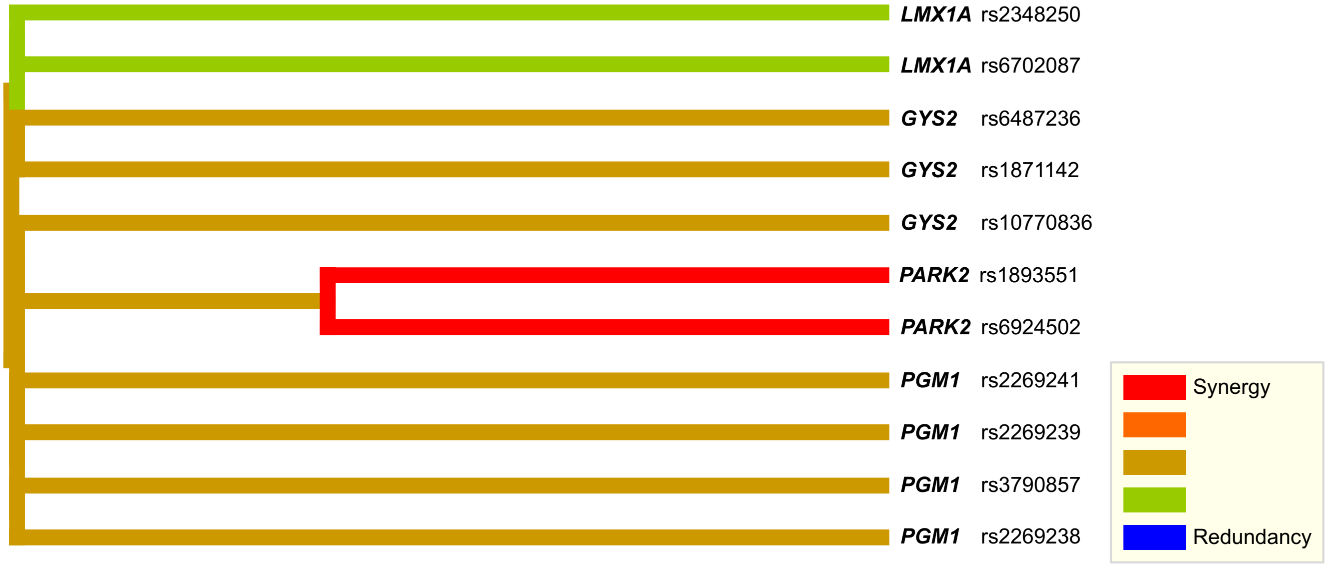


Figure 12

Interaction dendrogram produced from 11 SNPs that are chosen by 2LOmb. The colours in the dendrogram comprise a spectrum of colours representing a transition from synergy to redundancy. Synergy denotes the situation in which the entropy-based interaction between two SNPs provides more information than the entropy-based correlation between the pair. Redundancy refers to the situation in which the entropy-based interaction between two SNPs provides less information than the entropy-based correlation between the pair [7].

interaction model. Consider a two-locus model that captures an interaction between loci A and B, let *A* and *a* be the major (common) and minor (rare) alleles at locus A. Similarly, let *B* and *b* be the major and minor alleles at locus B. At each locus, the genotype is represented by characters 0, 1 or 2 where 0 denotes a homozygous wild-type genotype (*AA* and *BB*), 1 denotes a heterozygous genotype (*Aa* and *Bb*) and 2 denotes a homozygous variant or homozygous mutant genotype (*aa* and *bb*). $f_{ij} \in [0, 1]$ is defined as the disease penetrance of the two-locus genotype *ij* that consists of genotype *i* at locus A and genotype *j* at locus B. The marginal penetrances M_{Ai} for genotype *i* at locus A and M_{Bj} for genotype *j* at locus B are given by

$$M_{Ai} = p_B^2 f_{i0} + 2p_B(1 - p_B)f_{i1} + (1 - p_B)^2 f_{i2}, i \in \{0, 1, 2\}, \tag{4}$$

and

$$M_{Bj} = p_A^2 f_{0j} + 2p_A(1 - p_A)f_{1j} + (1 - p_A)^2 f_{2j}, j \in \{0, 1, 2\}, \tag{5}$$

where p_A and p_B are the major allele frequencies. Equations 4 and 5 are usually represented by a penetrance table as illustrated in Table 13. The two-locus interaction model is a pure epistasis model if

Table 2: 2LOmb identifies 11 intronic SNPs, which are located in four genes, from the reduced T2D data.

Gene	Chromosome and Location	Two-SNP Unit in the Ensemble
PGM1 (phosphoglucomutase 1)	1p31	(rs2269241, rs3790857) (rs2269239, rs3790857) (rs3790857, rs2269238)
LMX1A (LIM homeobox transcription factor 1, alpha)	1q22-q23	(rs2348250, rs6702087)
PARK2 (Parkinson disease (autosomal recessive, juvenile) 2, parkin)	6q25.2-q27	(rs1893551, rs6924502)
GYS2 (glycogen synthase 2 (liver))	12p12.2	(rs6487236, rs1871142) (rs1871142, rs10770836)

Association between these SNPs and the disease is possible (global *p*-value < 0.0001). Seven two-SNP units are present in the ensemble where each unit contains a pair of SNPs from the same gene.

Table 3: Genotype relative risk evaluated from genotype distribution of SNPs in *PGM1*.

		Frequency			
SNP	**Genotype**	**Case**	**Control**	**Relative Risk**	**95% CI**
SNP1	0	0.6513	0.6528	0.9977	(0.9573-1.0399)
	1	0.3082	0.3076	1.0018	(0.9204-1.0905)
	2	0.0405	0.0396	1.0229	(0.7757-1.3488)
SNP2	0	0.6493	0.6521	0.9957	(0.9552-1.0379)
	1	0.3087	0.3079	1.0024	(0.9209-1.0910)
	2	0.0420	0.0399	1.0519	(0.8006-1.3822)
SNP3	0	0.6153	0.6568	0.9368	(0.8972-0.9782)
	1	0.3472	0.3056	1.1361	(1.0479-1.2316)
	2	0.0375	0.0376	0.9974	(0.7490-1.3281)
SNP4	0	0.6638	0.6668	0.9956	(0.9564-1.0364)
	1	0.2991	0.2969	1.0074	(0.9237-1.0988)
	2	0.0370	0.0363	1.0202	(0.7636-1.3631)
(SNP1, SNP3)	0 0	0.6043	0.6448	0.9372	(0.8966-0.9796)
	0 1	**0.0470**	**0.0080**	**5.8858**	**(3.7730-9.1816)**
	1 0	0.0110	0.0120	0.9183	(0.5420-1.5561)
	1 1	0.2961	0.2943	1.0064	(0.9222-1.0983)
	1 2	0.0010	0.0013	0.7514	(0.1378-4.0984)
	2 1	0.0040	0.0033	1.2022	(0.4753-3.0408)
	2 2	0.0365	0.0363	1.0064	(0.7523-1.3463)
(SNP2, SNP3)	00	0.6038	0.6448	0.9364	(0.8958-0.9789)
	01	**0.0455**	**0.0073**	**6.2159**	**(3.9154-9.8681)**
	10	0.0115	0.0117	0.9875	(0.5853-1.6661)
	11	0.2966	0.2949	1.0058	(0.9218-1.0975)
	12	0.0005	0.0013	0.3757	(0.0420-3.3588)
	21	0.0050	0.0033	1.5028	(0.6266-3.6038)
	22	0.0370	0.0363	1.0202	(0.7636-1.3631)
(SNP3, SNP4)	00	0.6138	0.6551	0.9369	(0.8971-0.9785)
	01	0.0015	0.0017	0.9017	(0.2157-3.7686)
	10	**0.0500**	**0.0117**	**4.2936**	**(2.9340-6.2831)**
	11	0.2971	0.2936	1.0121	(0.9274-1.1044)
	21	0.0005	0.0017	0.3006	(0.0351-2.5706)
	22	0.0370	0.0360	1.0297	(0.7702-1.3765)

The relative risk is calculated from the ratio between the probability of a genotype of interest occurring in the case group and that of the same genotype occurring in the control group. Only the relative risks based on genotypic information from each SNP and SNP pairs identified by 2LOmb are considered. Characters 0, 1 and 2 represent different genotypes at each locus where 0 denotes a homozygous wild-type genotype, 1 denotes a heterozygous genotype and 2 denotes a homozygous variant or homozygous mutant genotype. The relative risk displayed in boldface is statistically significant. The major/minor alleles for rs2269241, rs2269239, rs3790857 and rs2269238 are T/C, G/C, C/T and G/T, respectively. The allelic information is extracted from the original T2D data. SNP1 = rs2269241, SNP2 = rs2269239, SNP3 = rs3790857 and SNP4 = rs2269238.

$$M_{Ai} = M_{Bj} = K, \forall i, j \in \{0,1,2\}, \qquad (6)$$

where K is the disease prevalence. Obviously, many combinations of penetrance f_{ij} satisfy the condition given in equation 6. Culverhouse et al. [42] suggest that a pure epistasis model with the maximum heritability is particularly useful in association studies. The heritability (h^2) of the two-locus interaction model is defined by

$$h^2 = V_I / V_T, \qquad (7)$$

where $V_T = K(1 - K)$ is the total variance of the dichotomous phenotypes in the population and V_I is the epistatic variation attributable to the genotypes. V_I is defined by

$$\begin{aligned} V_I = {} & p_A^2 p_B^2 (f_{00} - K)^2 + 2 p_A^2 p_B (1 - p_B)(f_{01} - K)^2 + p_A^2 (1 - p_B)^2 (f_{02} - K)^2 \\ & + 2 p_A (1 - p_A) p_B^2 (f_{10} - K)^2 + 4 p_A (1 - p_A) p_B (1 - p_B)(f_{11} - K)^2 + 2 p_A (1 - p_A)(1 - p_B)^2 (f_{12} - K)^2 \\ & + (1 - p_A)^2 p_B^2 (f_{20} - K)^2 + 2(1 - p_A)^2 p_B (1 - p_B)(f_{21} - K)^2 + (1 - p_A)^2 (1 - p_B)^2 (f_{22} - K)^2. \end{aligned} \qquad (8)$$

The search for feasible penetrance f_{ij} that also maximises the heritability or other variance-based objectives can be treated as a constraint optimisation problem. Many algorithms including a double description method [42] and a genetic algorithm [83] have been proven to be suitable for the task.

Culverhouse et al. [42] have identified the maximum heritability of purely epistatic two-locus and multi-locus interaction models for various values of disease prevalence. For instance, the maximum heritability of a two-

Table 4: Haplotype relative risk evaluated from genotype distribution of SNPs in *PGM1*.

		Frequency			
SNP	**Allele and Haplotype**	**Case**	**Control**	**Relative Risk**	**95% CI**
SNP1	0	0.8054	0.8066	0.9985	(0.9791-1.0183)
	1	0.1946	0.1934	1.0061	(0.9274-1.0916)
SNP2	0	0.8037	0.8061	0.9970	(0.9775-1.0168)
	1	0.1963	0.1939	1.0126	(0.9336-1.0982)
SNP3	0	0.7889	0.8096	0.9744	(0.9550-0.9943)
	1	0.2111	0.1904	1.1087	(1.0240-1.2003)
SNP4	0	0.8134	0.8152	0.9977	(0.9789-1.0170)
	1	0.1866	0.1848	1.0100	(0.9288-1.0982)
(SNP1, SNP3)	0 0	0.7812	0.8019	0.9742	(0.9543-0.9945)
	0 1	**0.0242**	**0.0047**	**5.1533**	**(3.3946-7.8231)**
	1 0	0.0077	0.0077	1.0000	(0.6349-1.5752)
	1 1	0.1869	0.1857	1.0064	(0.9257-1.0941)
(SNP2, SNP3)	00	0.7804	0.8017	0.9734	(0.9535-0.9938)
	01	**0.0232**	**0.0044**	**5.3216**	**(3.4557-8.1949)**
	10	0.0085	0.0079	1.0758	(0.6930-1.6701)
	11	0.1879	0.1861	1.0099	(0.9291-1.0977)
(SNP3, SNP4)	00	0.7881	0.8086	0.9747	(0.9552-0.9946)
	01	0.0008	0.0010	0.7661	(0.1943-3.0205)
	10	**0.0253**	**0.0067**	**3.7936**	**(2.6367-5.4582)**
	11	0.1858	0.1837	1.0113	(0.9298-1.0999)

The relative risk is calculated from the ratio between the probability of a haplotype of interest occurring in the case group and that of the same haplotype occurring in the control group. Haplotype inference is carried out using an expectation-maximisation method. Only the relative risks based on genotypic information from each SNP and SNP pairs identified by 2LOmb are considered. Characters 0 and 1 represent different alleles at each locus where 0 and 1 denote major and minor alleles, respectively. The relative risk displayed in boldface is statistically significant. The allelic information for each SNP is given in Table 3. SNP1 = rs2269241, SNP2 = rs2269239, SNP3 = rs3790857 and SNP4 = rs2269238.

locus interaction model for $p_A = p_B = 0.5$ with the penetrances in Table 14 is

$$h^2_{\max}(K) = 2K/(1-K), K \in (0,1/4]. \qquad (9)$$

When a two-locus interaction model is expanded into a multi-locus interaction model, the marginal penetrance equality constraint must be extended to cover all loci. Furthermore, the expression for V_I must also be expanded to cover additional genotypes while the expression for V_T remains unchanged. With the necessary model expansion, the maximum heritability of a three-locus interaction model for $p_A = p_B = p_C = 0.5$ with the penetrances in Table 15 is given by

$$h^2_{\max}(K) = 9K/(1-K), K \in (0,1/16]. \qquad (10)$$

Table 5: Genotype relative risk evaluated from genotype distribution of SNPs in *LMX1A*.

		Frequency			
SNP	**Genotype**	**Case**	**Control**	**Relative Risk**	**95% CI**
SNP5	0	0.8429	0.8642	0.9754	(0.9526-0.9987)
	1	0.1531	0.1315	1.1642	(1.0140-1.3366)
	2	0.0040	0.0043	0.9248	(0.3840-2.2271)
SNP6	0	0.8799	0.8492	1.0362	(1.0135-1.0594)
	1	0.1161	0.1465	0.7924	(0.6829-0.9193)
	2	0.0040	0.0043	0.9248	(0.3840-2.2271)
(SNP5, SNP6)	00	0.8329	0.8299	1.0036	(0.9784-1.0295)
	01	**0.0100**	**0.0343**	**0.2918**	**(0.1814-0.4695)**
	10	**0.0470**	**0.0193**	**2.4355**	**(1.7644-3.3618)**
	11	0.1061	0.1119	0.9482	(0.8061-1.1153)
	22	0.0040	0.0040	1.0018	(0.4103-2.4465)

The explanation for how the relative risks are obtained and displayed is the same as that given in Table 3. The major/minor alleles for rs2348250 and rs6702087 are G/A and G/C, respectively. The allelic information is extracted from the original T2D data. SNP5 = rs2348250 and SNP6 = rs6702087.

Table 6: Haplotype relative risk evaluated from genotype distribution of SNPs in *LMX1A*.

		Frequency			
SNP	**Allele and Haplotype**	**Case**	**Control**	**Relative Risk**	**95% CI**
SNP5	0	0.9195	0.9299	0.9887	(0.9774-1.0002)
	1	0.0805	0.0701	1.1494	(0.9997-1.3214)
SNP6	0	0.9380	0.9224	1.0168	(1.0059-1.0279)
	1	0.0620	0.0776	0.7997	(0.6892-0.9280)
(SNP5, SNP6)	00	0.9143	0.9124	1.0021	(0.9898-1.0145)
	01	**0.0051**	**0.0175**	**0.2931**	**(0.1829-0.4697)**
	10	**0.0236**	**0.0100**	**2.3640**	**(1.7150-3.2585)**
	11	0.0569	0.0601	0.9472	(0.8064-1.1127)

The explanation for how the relative risks are obtained and displayed is the same as that given in Table 4. The allelic information for each SNP is given in Table 5. SNP5 = rs2348250 and SNP6 = rs6702087.

Similarly, the maximum heritability of a four-locus interaction model for $p_A = p_B = p_C = p_D = 0.5$ with the penetrances in Table 16 is

$$h^2_{\max}(K) = 35K/(1-K), K \in (0,1/64]. \qquad (11)$$

Additional details about the maximum heritability and the corresponding two-locus and multi-locus penetrance tables for other values of disease prevalence can be found in Culverhouse et al. [42]. In this article, the simulated data sets are generated to achieve the maximum heritability of 0.01, 0.025 and 0.05. The values of disease prevalence that lead to the target heritability for two-, three- and four-locus interaction models are given in Table 17.

genomeSIM

genomeSIM is a simulation package for generating case-control samples in large-scale and genome-wide association studies [60]. The package is capable of producing many realistic scenarios, which can be observed in a population and genetic samples, including linkage disequilibrium, phenocopy and genotyping errors. The case/control status of each sample is determined from the penetrance-based genetic models or interaction models. As a result, the package can accommodate many epistasis models including the one proposed by Culverhouse et al. [42]. A data set can be produced via two modes: a population-based simulation and a probability-based simulation. In the population-based simulation, an initial population is generated according to the predefined allele frequency of each SNP. Then further generations are created by crossing the genotype strings within successive generations until the specified number of generations is reached. The resulting data set contains a population-dependent case and control samples that follow a forward-time simulation strategy. In contrast, genotype strings are incrementally generated without any string crossing for only one generation in the probability-based simulation. The creation of new strings is terminated only when the desired numbers

Table 7: Genotype relative risk evaluated from genotype distribution of SNPs in *PARK2*.

		Frequency			
SNP	**Genotype**	**Case**	**Control**	**Relative Risk**	**95% CI**
SNP7	0	0.4802	0.5110	0.9398	(0.8873-0.9954)
	1	0.4322	0.4041	1.0695	(1.0008-1.1429)
	2	0.0875	0.0849	1.0313	(0.8581-1.2395)
SNP8	0	0.4892	0.4923	0.9937	(0.9380-1.0527)
	1	0.4087	0.4121	0.9917	(0.9267-1.0613)
	2	0.1021	0.0955	1.0682	(0.9009-1.2665)
(SNP7, SNP8)	00	0.4492	0.4844	0.9275	(0.8726-0.9858)
	01	0.0285	0.0260	1.0982	(0.7841-1.5380)
	02	0.0025	0.0007	3.7569	(0.7296-19.3450)
	10	**0.0400**	**0.0080**	**5.0092**	**(3.1856-7.8767)**
	11	0.3792	0.3848	0.9854	(0.9169-1.0590)
	12	0.0130	0.0113	1.1492	(0.6918-1.9089)
	21	0.0010	0.0013	0.7514	(0.1378-4.0984)
	22	0.0865	0.0836	1.0358	(0.8606-1.2465)

The explanation for how the relative risks are obtained and displayed is the same as that given in Table 3. The major/minor alleles for rs1893551 and rs6924502 are G/A and T/C, respectively. The allelic information is extracted from the original T2D data. SNP7 = rs1893551 and SNP8 = rs6924502.

Table 8: Haplotype relative risk evaluated from genotype distribution of SNPs in *PARK2*.

		Frequency			
SNP	**Allele and Haplotype**	**Case**	**Control**	**Relative Risk**	**95% CI**
SNP7	0	0.6963	0.7130	0.9766	(0.9515-1.0023)
	1	0.3037	0.2870	1.0582	(0.9950-1.1254)
SNP8	0	0.6936	0.6984	0.9931	(0.9672-1.0198)
	1	0.3064	0.3016	1.0159	(0.9563-1.0793)
(SNP7, SNP8)	00	0.6726	0.6937	0.9696	(0.9434-0.9966)
	01	0.0238	0.0194	1.2248	(0.9369-1.6011)
	10	**0.0210**	**0.0048**	**4.4215**	**(2.8972-6.7480)**
	11	0.2826	0.2822	1.0016	(0.9397-1.0675)

The explanation for how the relative risks are obtained and displayed is the same as that given in Table 4. The allelic information for each SNP is given in Table 7. SNP7 = rs1893551 and SNP8 = rs6924502.

of case and control samples are obtained. In this study, the probability-based simulation is used to produce all case and control samples where the simulation parameter setting is given in the supplement (see Additional file 2). genomeSIM is available upon request to Scott M. Dudek at the Vanderbilt University dudek@chgr.mc.vanderbilt.edu.

Set association approach

A set association approach (SAA) is an association detection technique based on an omnibus permutation test on sets of candidate SNPs [11]. The test captures information about genotyping errors, deviation from Hardy-Weinberg equilibrium (HWE) and allelic association. In the first step, the genotype distribution for each SNP in the control samples is checked for HWE. Then, the number of SNPs that is to be excluded from the study (n_d) is set to the number of SNPs in the control samples that deviate from HWE. Two test statistics are subsequently calculated for each SNP: an allelic association statistic and a statistic for the deviation from HWE of each SNP in the case samples. The allelic association statistic is a χ^2 statistic which is cal-

Table 9: Genotype relative risk evaluated from genotype distribution of SNPs in *GYS2*.

		Frequency			
SNP	**Genotype**	**Case**	**Control**	**Relative Risk**	**95% CI**
SNP9	0	0.6473	0.6575	0.9846	(0.9447-1.0262)
	1	0.3167	0.3046	1.0396	(0.9558-1.1308)
	2	0.0360	0.0379	0.9491	(0.7105-1.2679)
SNP10	0	0.5933	0.6441	0.9211	(0.8805-0.9634)
	1	0.3712	0.3169	1.1713	(1.0839-1.2657)
	2	0.0355	0.0389	0.9119	(0.6828-1.2180)
SNP11	0	0.6058	0.6142	0.9864	(0.9427-1.0321)
	1	0.3507	0.3352	1.0461	(0.9675-1.1310)
	2	0.0435	0.0506	0.8601	(0.6650-1.1126)
(SNP9, SNP10)	00	0.5863	0.6335	0.9255	(0.8841-0.9689)
	01	**0.0610**	**0.0240**	**2.5463**	**(1.9135-3.3885)**
	10	0.0055	0.0107	0.5166	(0.2610-1.0224)
	11	0.3092	0.2919	1.0590	(0.9717-1.1540)
	12	0.0020	0.0020	1.0018	(0.2831-3.5456)
	21	0.0010	0.0010	1.0018	(0.1676-5.9902)
	22	0.0335	0.0370	0.9071	(0.6734-1.2218)
(SNP10, SNP11)	00	0.5463	0.5922	0.9224	(0.8776-0.9695)
	01	0.0455	0.0506	0.8997	(0.6982-1.1593)
	02	0.0015	0.0013	1.1271	(0.2525-5.0304)
	10	**0.0595**	**0.0220**	**2.7095**	**(2.0164-3.6408)**
	11	0.3032	0.2823	1.0739	(0.9839-1.1722)
	12	0.0085	0.0126	0.6723	(0.3805-1.1877)
	21	0.0020	0.0023	0.8587	(0.2517-2.9296)
	22	0.0335	0.0366	0.9153	(0.6791-1.2336)

The explanation for how the relative risks are obtained and displayed is the same as that given in Table 3. The major/minor alleles for rs6487236, rs1871142 and rs10770836 are A/G, G/A and G/A, respectively. The allelic information is extracted from the original T2D data. SNP9 = rs6487236, SNP10 = rs1871142 and SNP11 = rs10770836.

Table 10: Haplotype relative risk evaluated from genotype distribution of SNPs in *GYS2*.

SNP	Allele and Haplotype	Frequency		Relative Risk	95% CI
		Case	Control		
SNP9	0	0.8057	0.8098	0.9949	(0.9757-1.0146)
	1	0.1943	0.1902	1.0216	(0.9412-1.1087)
SNP10	0	0.7789	0.8026	0.9705	(0.9505-0.9908)
	1	0.2211	0.1974	1.1201	(1.0367-1.2102)
SNP11	0	0.7811	0.7818	0.9992	(0.9782-1.0205)
	1	0.2189	0.2182	1.0030	(0.9299-1.0818)
(SNP9, SNP10)	00	0.7740	0.7967	0.9715	(0.9512-0.9922)
	01	**0.0317**	**0.0131**	**2.4258**	**(1.8357-3.2056)**
	10	0.0049	0.0059	0.8330	(0.4807-1.4433)
	11	0.1894	0.1843	1.0276	(0.9455-1.1169)
(SNP10, SNP11)	00	0.7494	0.7692	0.9742	(0.9524-0.9965)
	01	0.0295	0.0334	0.8843	(0.7069-1.1061)
	10	**0.0318**	**0.0126**	**2.5275**	**(1.9064-3.3511)**
	11	0.1894	0.1848	1.0244	(0.9426-1.1134)

The explanation for how the relative risks are obtained and displayed is the same as that given in Table 4. The allelic information for each SNP is given in Table 9. SNP9 = rs6487236, SNP10 = rs1871142 and SNP11 = rs10770836.

culated from the contingency table of alleles or genotypes with disease status. On the other hand, a χ^2 statistic for the deviation from HWE of each SNP in the case samples indicates the level of association. A large deviation from the equilibrium usually signifies strong association between a SNP and the disease. However, an excessively large deviation may be the result of genotyping errors. n_d SNPs with largest test statistics for the deviation from HWE are hence excluded from the consideration.

The test statistics for the allelic association and deviation from HWE are multiplied together to form a single *S* statistic for each remaining SNP. SNPs are then ranked according to their *S* statistics. A preset number of SNPs with highest ranks are considered for association. The first candidate SNP set contains only the SNP with the highest rank (the highest *S* statistic). The *p*-value for this first set is determined from a permutation simulation where the case and control labels are randomly permuted while the numbers of case and control samples remain unchanged. In each permutation replicate, a new genotype contingency table is constructed and a new *S* statistic is subsequently obtained. The *p*-value is given by the fraction of permutation replicates with an *S* statistic greater than or equal to the *S* statistic from the original data. The second candidate SNP set consists of the first two SNPs in the rank list. The test statistic for this SNP set is the sum of *S* statistics from both SNPs. The *p*-value for the second candidate SNP set is also obtained through the permutation simulation. By progressively adding the remaining SNP with the highest rank to the previously considered candidate set and performing the permutation simulation, *p*-values for all candidate SNP sets are estimated. The sizes of candidate SNP sets have the range of one to the preset number. Among all candidate sets, the SNP set that best describes genetic association has the lowest *p*-value.

Since multiple hypotheses are postulated during the construction of candidate SNP sets, the global *p*-value for the selected candidate set must be evaluated. This is achieved through a permutation simulation in which the current raw *p*-value for the chosen candidate set is now used as the test statistic. The existing permutation replicates, created for the early estimation of the raw *p*-value, can be reused and a nested permutation simulation is hence avoided. In this study, the maximum allowable size of the candidate

Table 11: Prediction accuracy of the best MDR model constructed from the 2LOmb output.

Description	Value
SNP and Gene	rs2269241 (*PGM1*), rs3790857 (*PGM1*), rs1893551 (*PARK2*), rs6924502 (*PARK2*), rs1871142 (*GYS2*), rs10770836 (*GYS2*)
Classification (Training) Accuracy	0.5709
Prediction Accuracy	0.5402
Cross-Validation Consistency (CVC)	9/10

The model contains six SNPs from *PGM1*, *PARK2* and *GYS2*. A permutation test with 1,000 randomised replicates of case-control data for this model reveals that the empirical *p*-value for the null hypothesis of no association is $p < 0.001$.

Table 12: Summary of prediction accuracy by MDR from early genetic association studies of T2D in a Korean population, a Han Chinese population from Taiwan, a female population from the US, and that from an early genetic association study of metabolic syndrome in an Italian population from the Centre East Coast Italy.

Reference	Population	Gene	Prediction Accuracy	CVC	Permutation *p*-value
Cho et al. [43]	Korean	*PPARG, UCP2*	0.7957	9/10	0.01
Hsieh et al. [44]	Han Chinese	*RXRG, EGFR*	0.6270	11/12	N/A
Qi et al. [45]	US	*KCNJ11, HNF4A*	0.5420	10/10	0.010
Fiorito et al. [47]	Italian	*PPARG, DIO2*	0.6170	10/10	0.005

A permutation test with 1,000 randomised replicates is performed to obtain the empirical *p*-value for the null hypothesis of no association in the studies conducted in the US and Italian populations. In contrast, a permutation test with 100 randomised replicates is performed to obtain the empirical *p*-value in the study conducted in the Korean population.

SNP set is the total number of available SNPs while the number of permutation replicates for *p*-value evaluation is set to 10,000. The allelic association statistic employed in the study is the χ^2 statistic that is obtained through the contingency table of genotypes with disease status. A PASCAL program for the set association approach can be obtained from the website for *S* statistic in gene mapping [84].

Correlation-based feature selection technique

A correlation-based feature selection (CFS) technique [14] is an attribute (SNP) subset evaluation heuristic that considers both the usefulness of individual features (SNPs) in the (case-control) classification task and the level of intercorrelation among features. Each attribute subset is assigned a score given by

$$Merit_F = \frac{n_c \overline{r}_{cf}}{\sqrt{n_c + n_c(n_c - 1)\overline{r}_{ff}}}, \qquad (12)$$

where $Merit_F$ is the heuristic merit of an n_c-attribute subset F, $\overline{r}_{cf}$ is the average feature-class correlation and $\overline{r}_{ff}$ is the average feature-feature inter-correlation. An attribute subset receives a high merit score if it contains features that are highly correlated with the class and at the same time have low inter-correlation among one another. An application of a best-first search for the best subset identification is carried out to avoid searching through all possible attribute subsets. CFS has been integrated into a Weka package [85,86].

Tuned ReliefF

A tuned ReliefF (TuRF) algorithm is a ranking algorithm for identifying genetic markers which are important in case-control classification [16]. TuRF is built on a ReliefF engine [15]. ReliefF randomly picks a sample from the (case-control) data and identifies its n_k nearest neighbours from the same class and another n_k nearest neighbours from the opposite class. The attribute values--the genotypes in this application--of the neighbour samples are compared to that of the randomly picked sample and are subsequently used to update the relevance score for each attribute (genetic marker). This process is repeated for a specified number of samples, which is limited by the total sample size. The rationale of ReliefF is that an attribute which is important for the classification should have different values for samples from different classes and have the same value for samples from the same class. The relevance score of an attribute have a range from -1 (not relevant) to +1 (highly relevant). TuRF exploits the capability of ReliefF by repeatedly executing ReliefF and removing a portion of worst attributes at the end of each execution. This leads to the reevaluation of remaining attributes and, hence, reduces the effects of attribute noise on the attribute screening. In this study, the number of repetitions for random sample picking in the ReliefF part is equal to the total number of case-control samples while the neighbourhood size (n_k) for the relevance score calculation is set to ten. Furthermore, the worst 1% of SNPs is removed at the end of each ReliefF iteration (TuRF 1%). TuRF has been integrated into the current distribution of multifactor dimensionality reduction (MDR) software.

Table 13: Penetrances for a two-locus interaction model.

	Penetrance of Genotype			
Genotype	*BB*	*Bb*	*bb*	Marginal Penetrance
AA	f_{00}	f_{01}	f_{02}	M_{A0}
Aa	f_{10}	f_{11}	f_{12}	M_{A1}
aa	f_{20}	f_{21}	f_{22}	M_{A2}
Marginal Penetrance	M_{B0}	M_{B1}	M_{B2}	K

f_{ij} is the disease penetrance of genotype *ij*. M_{Ai} and M_{Bj} are the marginal penetrances for genotype *i* at locus A and genotype *j* at locus B, respectively. $M_{Ai} = M_{Bj} = K$, $\forall i, j \in \{0, 1, 2\}$ in a pure epistasis model.

Table 14: Two-locus penetrances that lead to the maximum heritability $h^2_{\max}$ (*K*) = 2*K*/(1 - *K*) for *K* ∈ (0, 1/4].

Genotype	Penetrance of Genotype *BB*	*Bb*	*bb*
AA	0	0	4*K*
Aa	0	2*K*	0
aa	4*K*	0	0

All allele frequencies are equal ($p_A = p_B = 0.5$).

Multifactor dimensionality reduction

A multifactor dimensionality reduction (MDR) method is a wrapper-based technique that is capable of identifying the best genetic marker combination among possible markers for the separation between case and control samples [19]. Similar to other wrapper-based methods, an n_f-fold cross-validation technique provides a means to determine the prediction accuracy of the candidate marker model. Basically, the combined case and control samples are randomly divided into n_f folds where n_f - 1 folds of samples are used to construct a decision table while the remaining fold of samples is used to identify the prediction capability of the constructed decision table. The decision table construction and testing procedure is repeated n_f times. Hence, the samples in each fold are always used both to construct and to test the decision table. The number of cells in a decision table is given by G^{n_c} where n_c is the number of candidate markers selected from possible markers and *G* is the number of possible genotypes according to the marker. For a SNP, which is a bi-allelic marker, *G* is equal to three. During the decision table construction, each cell in the table is filled with case and control samples that have their genotype corresponds to the cell label. The ratio between numbers of case and control samples provides the decision for each cell whether the corresponding genotype is a protective or disease-predisposing genotype. An example of decision table construction is illustrated in Figure 13.

Table 15: Three-locus penetrances that lead to the maximum heritability $h^2_{\max}$ (*K*) = 9*K*/(1 - *K*) for *K* ∈ (0, 1/16].

Genotype	Penetrance of Genotype								
	CC			***Cc***			***cc***		
	BB	***Bb***	***bb***	***BB***	***Bb***	***bb***	***BB***	***Bb***	***bb***
AA	0	0	16*K*	0	0	0	0	0	0
Aa	0	0	0	0	4*K*	0	0	0	0
aa	0	0	0	0	0	0	16*K*	0	0

All allele frequencies are equal ($p_A = p_B = p_C = 0.5$).

Table 16: Four-locus penetrances that lead to the maximum heritability $h^2_{\max}$ (*K*) = 35*K*/(1 - *K*) for *K* ∈ (0, 1/64].

Genotype		Penetrance of Genotype								
		CC			***Cc***			***cc***		
		DD	***Dd***	***dd***	***DD***	***Dd***	***dd***	***DD***	***Dd***	***Dd***
	BB	0	0	0	0	0	0	0	0	0
AA	*Bb*	0	0	0	0	0	0	0	0	0
	bb	0	0	64*K*	0	0	0	0	0	0
	BB	0	0	0	0	0	0	0	0	0
Aa	*Bb*	0	0	0	0	8*K*	0	0	0	0
	bb	0	0	0	0	0	0	0	0	0
	BB	0	0	0	0	0	0	64*K*	0	0
aa	*Bb*	0	0	0	0	0	0	0	0	0
	bb	0	0	0	0	0	0	0	0	0

All allele frequencies are equal ($p_A = p_B = p_C = p_D = 0.5$).

The prediction accuracy of the decision table is subsequently evaluated by counting the numbers of case and control samples in the testing fold that their disease status can correctly be identified using the constructed decision rules. The process of decision table construction and evaluation must be cycled through all or some of possible 2^{n_m} - 1 combinations where n_m is the total number of available markers in the study. The best genetic marker combination is determined from two criteria: prediction accuracy and cross-validation consistency. Each time that a testing fold is used for the prediction accuracy determination, the accuracy of the interesting marker combination model is compared with that from other models that also contain the same number of markers. The model that consistently ranks the first in comparison to other choices with the same number of markers has high cross-validation consistency. Prediction accuracy is the main criterion for decision making while cross-validation consistency is only used as an auxiliary measure. Cross-validation consistency generally confirms that the high rank model can consistently be identified regardless of how the samples are divided for cross-validation. In a situation where two or more models with different number of markers are equally good in terms of prediction accuracy and cross-validation consistency, the most parsimonious model--the combination with the least number of markers--is chosen as the best model.

After the best model has been selected, a permutation test is used to assess the probability of obtaining prediction accuracy that is at least as large as or larger than that observed in the original data from randomised data. This

Table 17: Disease prevalence that gives the target maximum heritability of 0.01, 0.025 and 0.05 for two-, three- and four-locus interaction models.

Model	Prevalence (*K*)		
	$h^2_{\max}$ (*K*) = 0.01	$h^2_{\max}$ (*K*) = 0.025	$h^2_{\max}$ (*K*) = 0.05
Two-locus	0.004975	0.012346	0.024390
Three-locus	0.001110	0.002770	0.005525
Four-locus	0.000286	0.000714	0.001427

represents the probability that the null hypothesis of no association is true. Each permutation replicate is constructed by randomly assigning the case/control status to each sample with the numbers of case and control samples remaining fixed. MDR analysis is subsequently carried out to obtain the prediction accuracy of each permutation replicate. The empirical *p*-value is denoted by the fraction of permutation replicates with the prediction accuracy greater than or equal to the prediction accuracy obtained from the original data. MDR software, which incorporates many additional features including interaction visualisation via dendrograms and genetic marker screening via a χ^2 test, an odds ratio test, ReliefF and TuRF, is available from its homepage [87].

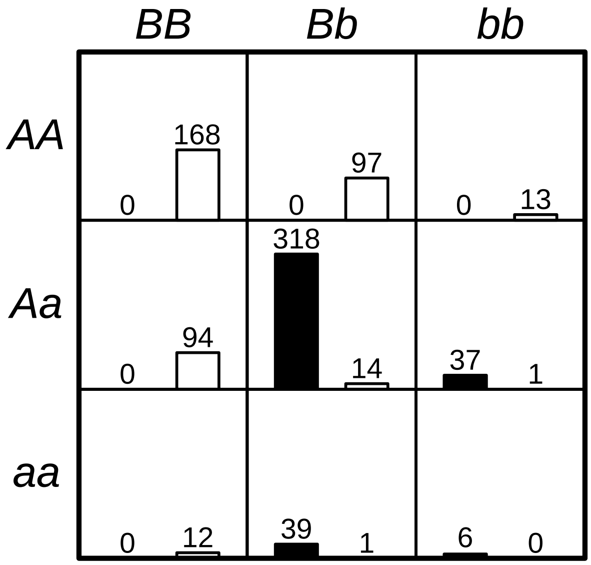


Figure 13

An MDR decision table that is constructed using a balanced case-control data set with the sample size of 800. The genotype of each sample is determined from two SNPs. The table consists of nine cells where each cell represents a unique genotype. The left (black) bar in each cell represents the number of case samples while the right (white) bar represents the number of control samples. The cells with genotypes *AABB*, *AABb*, *AAbb*, *AaBB* and *aaBB* are labelled as protective genotypes while the cells with genotypes *AaBb*, *Aabb*, *aaBb* and *aabb* are labelled as disease-predisposing genotypes.

JLIN

JLIN or a Java LINkage disequilibrium plotter is a computer program for visualisation of linkage disequilibrium analysis [63]. The program is capable of displaying many statistical measures including *D'* [64] and r^2 [65]. The program is publicly available from the Centre for Genetic Epidemiology and Biostatistics, University of Western Australia [88].

Interaction dendrogram

An interaction dendrogram is a graphical tool for the visualisation of relationships among attributes (SNPs) [68,69]. The interaction dendrogram is constructed via hierarchical clustering analysis and is embedded into MDR software [87]. The dendrogram illustrates the entropy-based interaction between attributes by displaying interacting or related attributes closely together as adjacent leaves in a tree. At the same time, independent attributes are placed far apart from one another. In addition, the conclusion regarding whether the interaction between attributes is synergistic or redundancy is present can be deduced.

Availability and requirements

The 2LOmb program for Windows platforms and examples of simulated data are available at http://code.google.com/p/nachol/w/list.

List of abbreviations

2LOmb: omnibus permutation test on ensembles of two-locus analyses; ALT: alanine transaminase; ANOVA: analysis of variance; AST: aspartate transaminase; CFS: correlation-based feature selection; CI: confidence interval; CVC: cross-validation consistency; *DIO2*: deiodinase, iodothyronine, type II; E2LA: exhaustive two-locus analysis; *EGFR*: epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian); FAMHAP: software for single-marker analysis and joint analysis of unphased genotype data from tightly linked markers (haplotype analysis); FUSION: Finland-United States Investigation of NIDDM Genetics; genomeSIM: simulation package for generating case-control samples in large-scale and genome-wide association studies; *GYS2*: glycogen synthase 2 (liver); *HNF4A*: hepatocyte nuclear factor 4, alpha; HuGENet: Human Genome Epidemiology Network; HWE: Hardy-Weinberg equilibrium; JLIN: Java LINkage disequilibrium plotter; *KCNJ11*: potassium inwardly-rectifying channel, subfamily J, member 11; LD: linkage disequilibrium; LIM domains: protein structural domains that are named after their initial discovery in the proteins Lin11, Isl-1 and Mec-3; *LMX1A*: LIM homeobox transcription factor 1, alpha; MDR: multifactor dimensionality reduction; NIDDM: noninsulin-dependent diabetes mellitus; *PARK2*: Parkinson disease (autosomal recessive, juvenile) 2, parkin; *PGM1*: phosphoglucomutase 1; *PPARG*: peroxisome proliferator-activated receptor gamma; *RXRG*: retinoid X receptor, gamma; SAA: set association approach; SNP: single nucleotide polymorphism; T2D: type 2 diabetes mellitus; TuRF: tuned ReliefF; *UCP2*: uncoupling protein 2 (mitochondrial, proton carrier); Weka: Waikato environment for knowledge analysis; WTCCC: Wellcome Trust Case Control Consortium.

Authors' contributions

WW conducted the literature survey, formulated the research question, implemented the proposed algorithm, designed the experiment, and collected and interpreted the computational results. AA conducted the literature survey, formulated the research question, designed the experiment and secured the access to the genomeSIM package. TP performed the statistical analysis and interpreted the statistical results. SS monitored and oversaw the execution of computer programs on the Beowulf cluster. CL provided additional comments about the genetic association study of T2D. NC conducted the literature survey, formulated the research question, designed the proposed algorithm, designed the experiment, secured the access to the T2D data from WTCCC, selected the candidate genes for the T2D association study, discussed all results, drew the conclusions and wrote the manuscript. All authors read and approved the final manuscript.

Authors' information

WW is a Ph.D. student at the Department of Electrical Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok. He also received his B.Eng. and M.Eng. degrees in electrical engineering from King Mongkut's University of Technology North Bangkok. His current research interests include machine learning, evolutionary computation and bioinformatics.

AA is a Ph.D. student at the Department of Immunology, Faculty of Medicine Siriraj Hospital, Mahidol University. He also received his B.Sc. degree in pharmacy from Mahidol University. His current research interests include human genetics, genetic epidemiology, population genetics and bioinformatics.

TP is a Ph.D. student at the Department of Electrical Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok. He also received his B.Eng. and M.Eng. degrees in production engineering from King Mongkut's University of Technology North Bangkok. His current research interests include evolutionary multi-objective optimisation and machine learning.

SS is a part-time research assistant at the Department of Electrical Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok. He received his B.Eng. and M.Eng. degrees in electrical engineering from Thammasat University and King Mongkut's University of Technology North Bangkok, respectively. His current research interests include machine learning and genetic epidemiology.

CL is the Head of Division of Molecular Genetics at the Department of Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University. He also received his M.D. degree from Mahidol University. His current research interests include human genetics and genetic diseases.

NC is an associate professor of electrical engineering at King Mongkut's University of Technology North Bangkok and an adjunct professor of genetic epidemiology at Mahidol University. He received his B.Eng. and Ph.D. degrees from the Department of Automatic Control and Systems Engineering, University of Sheffield. His current research interests include evolutionary computation, machine learning and genetic epidemiology.

Additional material

Additional file 1

***List of SNPs for the association study of T2D**. This Excel spreadsheet file contains the information about 7,065 SNPs which are explored during the genetic association study of T2D. Bonferroni-corrected and uncorrected χ^2's p-values from single-locus analyses are also provided.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-294-S1.xls]

Additional file 2

***genomeSIM parameters**. This text file contains an example of parameter setting in the genomeSIM simulation package.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-294-S2.txt]

Acknowledgements

The authors are extremely grateful to four anonymous reviewers and Dr. Danielle Talbot for their valuable comments and suggestions, which have contributed a lot towards improving the content and presentation of this article. The authors are also extremely grateful to Dr. Kuntinee Maneeratana, Dr. Pensiri Tongpadungrod, Dr. Graeme James Sheppard and Mr. Edward L. Karas for their efforts on proofreading the manuscript. The authors wish to thank Dr. Marong Phadoongsidhi and Dr. Vara Varavithya and for their assistance on processing the T2D data and providing an access to the Beowulf cluster, respectively. The authors also wish to thank Dr. Sissades Tongsima for his assistance on testing of multiple hypotheses, providing many suggestions for the manuscript revision and allowing the usage of computer facilities at the Genome Institute, National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA) during the algorithm development. The authors acknowledge Scott M. Dudek at the Vanderbilt University for providing an access to the genomeSIM package. Furthermore, the authors acknowledge Dr. Audrey Duncanson and the Wellcome Trust Case Control Consortium for granting an access to the genome-wide case-control data. WW, AA and TP were supported by the Thailand Research Fund (TRF) through the Royal Golden Jubilee Ph.D. Programme (Grant No. PHD/1.E.KN.49/A.1, PHD/4.I.MU.45/C.1 and PHD/1.E.KN.50/A.1, respectively). CL was supported by the Mahidol Research Grant. NC was supported by the Thailand Research Fund and the National Science and Technology Development Agency through the NSTDA Chair Professor Grant.

References

1. Risch N, Merikangas K: **The future of genetic studies of complex human diseases.** *Science* 1996, **273:**1516-1517.
2. Musani SK, Shriner D, Liu N, Feng R, Coffey CS, Yi N, Tiwari HK, Allison DB: **Detection of gene × gene interactions in genome-wide association studies of human population data.** *Hum Hered* 2007, **63:**67-84.
3. The Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447:**661-678.
4. The GAIN Collaborative Research Group: **New models of collaboration in genome-wide association studies: the Genetic Association Information Network.** *Nat Genet* 2007, **39:**1045-1051.
5. Heidema AG, Boer JMA, Nagelkerke N, Mariman ECM, van der A DL, Feskens EJM: **The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases.** *BMC Genet* 2006, **7:**23.
6. Motsinger AA, Ritchie MD, Reif DM: **Novel methods for detecting epistasis in pharmacogenomics studies.** *Pharmacogenomics* 2007, **8:**1229-1241.
7. Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC: **A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility.** *J Theor Biol* 2006, **241:**252-261.
8. Saeys Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics.** *Bioinformatics* 2007, **23:**2507-2517.
9. Lewis CM: **Genetic association studies: design, analysis and interpretation.** *Brief Bioinform* 2002, **3:**146-153.
10. Montana G: **Statistical methods in genetics.** *Brief Bioinform* 2006, **7:**297-308.
11. Hoh J, Wille A, Ott J: **Trimming, weighting, and grouping SNPs in human case-control association studies.** *Genome Res* 2001, **11:**2115-2119.
12. Potter DM: **Omnibus permutation tests of the association of an ensemble of genetic markers with disease in case-control studies.** *Genet Epidemiol* 2006, **30:**438-446.
13. Chapman J, Clayton D: **Detecting association using epistatic information.** *Genet Epidemiol* 2007, **31:**894-909.
14. Hall MA, Holmes G: **Benchmarking attribute selection techniques for discrete class data mining.** *IEEE Trans Knowl Data Eng* 2003, **15:**1437-1447.
15. Robnik-Šikonja M, Kononenko I: **Theoretical and empirical analysis of ReliefF and RReliefF.** *Mach Learn* 2003, **53:**23-69.
16. Moore JH, White BC: **Tuning ReliefF for genome-wide genetic analysis.** In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics* Edited by: Marchiori E, Moore JH, Rajapakse JC. Berlin, Heidelberg: Springer; 2007:166-175. [Goos G, Hartmanis J, van Leeuwen J (Founding and Former Series Editors): Lecture Notes in Computer Science, vol 4447].
17. Nelson MR, Kardia SLR, Ferrell RE, Sing CF: **A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation.** *Genome Res* 2001, **11:**458-470.
18. Culverhouse R, Klein T, Shannon W: **Detecting epistatic interactions contributing to quantitative traits.** *Genet Epidemiol* 2004, **27:**141-152.
19. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: **Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer.** *Am J Hum Genet* 2001, **69:**138-147.
20. Hahn LW, Ritchie MD, Moore JH: **Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions.** *Bioinformatics* 2003, **19:**376-382.
21. Bush WS, Dudek SM, Ritchie MD: **Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene-gene interactions.** *Bioinformatics* 2006, **22:**2173-2174.
22. Chung Y, Lee SY, Elston RC, Park T: **Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions.** *Bioinformatics* 2007, **23:**71-76.
23. Bush WS, Edwards TL, Dudek SM, McKinney BA, Ritchie MD: **Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction.** *BMC Bioinformatics* 2008, **9:**238.
24. Lou XY, Chen GB, Yan L, Ma JZ, Mangold JE, Zhu J, Elston RC, Li MD: **A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies.** *Am J Hum Genet* 2008, **83:**457-467.
25. Edwards TL, Lewis K, Velez DR, Dudek SM, Ritchie MD: **Exploring the performance of multifactor dimensionality reduction in large scale SNP studies and in the presence of genetic heterogeneity among epistatic disease models.** *Hum Hered* 2009, **67:**183-192.
26. Mechanic LE, Luke BT, Goodman JE, Chanock SJ, Harris CC: **Polymorphism Interaction Analysis (PIA): a method for investigating complex gene-gene interactions.** *BMC Bioinformatics* 2008, **9:**146.
27. Liang KH, Hwang Y, Shao WC, Chen EY: **An algorithm for model construction and its applications to pharmacogenomic studies.** *J Hum Genet* 2006, **51:**751-759.
28. Estrada-Gil JK, Fernández-López JC, Hernández-Lemus E, Silva-Zolezzi I, Hidalgo-Miranda A, Jiménez-Sánchez G, Vallejo-Clemente EE: **GPDTI: a Genetic Programming Decision Tree Induction method to find epistatic effects in common complex diseases.** *Bioinformatics* 2007, **23:**i167-i174.
29. Nunkesser R, Bernholt T, Schwender H, Ickstadt K, Wegener I: **Detecting high-order interactions of single nucleotide polymorphisms using genetic programming.** *Bioinformatics* 2007, **23:**3280-3288.
30. Lunetta KL, Hayward LB, Segal J, van Eerdewegh P: **Screening large-scale association study data: exploiting interactions using random forests.** *BMC Genet* 2004, **5:**32.
31. Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, van Eerdewegh P: **Identifying SNPs predictive of phenotype using random forests.** *Genet Epidemiol* 2005, **28:**171-182.
32. Chen X, Liu CT, Zhang M, Zhang H: **A forest-based approach to identifying gene and gene-gene interactions.** *Proc Natl Acad Sci USA* 2007, **104:**19199-19203.
33. Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH: **Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases.** *BMC Bioinformatics* 2003, **4:**28.
34. Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD: **Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology.** *Genet Epidemiol* 2008, **32:**325-340.
35. Cordell HJ: **Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans.** *Hum Mol Genet* 2002, **11:**2463-2468.
36. Wilson SR: **Epistasis.** In *Nature Encyclopedia of the Human Genome Volume 2*. Edited by: Cooper DN. London: Nature Publishing Group; 2004:317-320.

37. Neuman RJ, Rice JP: **Two-locus models of disease.** *Genet Epidemiol* 1992, **9:**347-365.
38. Schork NJ, Boehnke M, Terwilliger JD, Ott J: **Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits.** *Am J Hum Genet* 1993, **53:**1127-1136.
39. Li W, Reich J: **A complete enumeration and classification of two-locus disease models.** *Hum Hered* 2000, **50:**334-349.
40. Marchini J, Donnelly P, Cardon LR: **Genome-wide strategies for detecting multiple loci that influence complex diseases.** *Nat Genet* 2005, **37:**413-417.
41. Hallgrímsdóttir IB, Yuster DS: **A complete classification of epistatic two-locus models.** *BMC Genet* 2008, **9:**17.
42. Culverhouse R, Suarez BK, Lin J, Reich T: **A perspective on epistasis: limits of models displaying no main effect.** *Am J Hum Genet* 2002, **70:**461-471.
43. Cho YM, Ritchie MD, Moore JH, Park JY, Lee KU, Shin HD, Lee HK, Park KS: **Multifactor-dimensionality reduction shows a two-locus interaction associated with type 2 diabetes mellitus.** *Diabetologia* 2004, **47:**549-554.
44. Hsieh CH, Liang KH, Hung YJ, Huang LC, Pei D, Liao YT, Kuo SW, Bey MSJ, Chen JL, Chen EY: **Analysis of epistasis for diabetic nephropathy among type 2 diabetic patients.** *Hum Mol Genet* 2006, **15:**2701-2708.
45. Qi L, van Dam RM, Asselbergs FW, Hu FB: **Gene-gene interactions between *HNF4A* and *KCNJ11* in predicting type 2 diabetes in women.** *Diabet Med* 2007, **24:**1187-1191.
46. Zhang Z, Zhang S, Wong MY, Wareham NJ, Sha Q: **An ensemble learning approach jointly modeling main and interaction effects in genetic association studies.** *Genet Epidemiol* 2008, **32:**285-300.
47. Fiorito M, Torrente I, De Cosmo S, Guida V, Colosimo A, Prudente S, Flex E, Menghini R, Miccoli R, Penno G, Pellegrini F, Tassi V, Federici M, Trischitta V, Dallapiccola B: **Interaction of *DIO2* T92A and *PPARγ2* P12A polymorphisms in the modulation of metabolic syndrome.** *Obesity* 2007, **15:**2889-2895.
48. Albrechtsen A, Castella S, Andersen G, Hansen T, Pedersen O, Nielsen R: **A Bayesian multilocus association method: allowing for higher-order interaction in association studies.** *Genetics* 2007, **176:**1197-1208.
49. Zhang Y, Liu JS: **Bayesian inference of epistatic interactions in case-control studies.** *Nat Genet* 2007, **39:**1167-1173.
50. Evans DM, Marchini J, Morris AP, Cardon LR: **Two-stage two-locus models in genome-wide association.** *PLoS Genet* 2006, **2:**e157.
51. Ionita I, Man M: **Optimal two-stage strategy for detecting interacting genes in complex diseases.** *BMC Genet* 2006, **7:**39.
52. Gayán J, González-Pérez A, Bermudo F, Sáez ME, Royo JL, Quintas A, Galan JJ, Morón FJ, Ramirez-Lorca R, Real LM, Ruiz A: **A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis.** *BMC Genomics* 2008, **9:**360.
53. Heidema AG, Feskens EJM, Doevendans PAFM, Ruven HJT, van Houwelingen HC, Mariman ECM, Boer JMA: **Analysis of multiple SNPs in genetic association studies: comparison of three multi-locus methods to prioritize and select SNPs.** *Genet Epidemiol* 2007, **31:**910-921.
54. Pesarin F: *Multivariate Permutation Tests with Applications to Biostatistics* Chichester: Wiley; 2001.
55. Fisher RA: *Statistical Methods for Research Workers* 4th edition. London: Oliver and Boyd; 1932.
56. Westfall PH, Young SS: *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment* New York: John Wiley and Sons; 1993.
57. Becker T, Schumacher J, Cichon S, Baur MP, Knapp M: **Haplotype interaction analysis of unlinked regions.** *Genet Epidemiol* 2005, **29:**313-322.
58. Herold C, Becker T: **Genetic association analysis with FAMHAP: a major program update.** *Bioinformatics* 2009, **25:**134-136.
59. Hardy GH: **Mendelian proportions in a mixed population.** *Science* 1908, **28:**49-50.
60. Dudek SM, Motsinger AA, Velez DR, Williams SM, Ritchie MD: **Data simulation software for whole-genome association and other studies in human genetics.** In *Proceedings of the Pacific Symposium on Biocomputing 2006: 3-7 January 2006; Maui* Edited by: Altman RB, Dunker AK, Hunter L, Murray T, Klein TE. Singapore: World Scientific; 2006:499-510.
61. Guyon I, Elisseeff A: **An introduction to variable and feature selection.** *J Mach Learn Res* 2003, **3:**1157-1182.
62. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ: **A navigator for human genome epidemiology.** *Nat Genet* 2008, **40:**124-125.
63. Carter KW, McCaskie PA, Palmer LJ: **JLIN: a java based linkage disequilibrium plotter.** *BMC Bioinformatics* 2006, **7:**60.
64. Lewontin RC: **The interaction of selection and linkage. I. general considerations; heterotic models.** *Genetics* 1964, **49:**49-67.
65. Hill WG, Robertson A: **Linkage disequilibrium in finite populations.** *Theor Appl Genet* 1968, **38:**226-231.
66. Excoffier L, Slatkin M: **Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population.** *Mol Biol Evol* 1995, **12:**921-927.
67. Epstein MP, Satten GA: **Inference on haplotype effects in case-control studies using unphased genotype data.** *Am J Hum Genet* 2003, **73:**1316-1329.
68. Jakulin A, Bratko I, Smrke D, Demšar J, Zupan B: **Attribute interactions in medical data analysis.** In *Artificial Intelligence in Medicine* Edited by: Dojat M, Keravnou E, Barahona P. Berlin, Heidelberg: Springer; 2003:229-238. [Carbonell JG, Siekmann J (Series Editors): Lecture Notes in Artificial Intelligence, vol 2780].
69. Jakulin A, Bratko I: **Analyzing attribute dependencies.** In *Knowledge Discovery in Databases: PKDD 2003* Edited by: Lavrač N, Gamberger D, Todorovski L, Blockeel H. Berlin, Heidelberg: Springer; 2003:229-240. [Carbonell JG, Siekmann J (Series Editors): Lecture Notes in Artificial Intelligence, vol 2838].
70. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 2000, **28:**27-30.
71. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita K, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic Acids Res* 2006, **34:**D354-D357.
72. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y: **KEGG for linking genomes to life and the environment.** *Nucleic Acids Res* 2008, **36:**D480-D484.
73. Thameem F, Wolford JK, Wang J, German MS, Bogardus C, Prochazka M: **Cloning, expression and genomic structure of human *LMX1A*, and variant screening in Pima Indians.** *Gene* 2002, **290:**217-225.
74. Hanson RL, Ehm MG, Pettitt DJ, Prochazka M, Thompson DB, Timberlake D, Foroud T, Kobes S, Baier L, Burns DK, Almasy L, Blangero J, Garvey WT, Bennett PH, Knowler WC: **An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians.** *Am J Hum Genet* 1998, **63:**1130-1138.
75. Leak TS, Mychaleckyj JC, Smith SG, Keene KL, Gordon CJ, Hicks PJ, Freedman BI, Bowden DW, Sale MM: **Evaluation of a SNP map of 6q24-27 confirms diabetic nephropathy loci and identifies novel associations in type 2 diabetes patients with nephropathy from an African-American population.** *Hum Genet* 2008, **124:**63-71.
76. Sale MM, Freedman BI, Langefeld CD, Williams AH, Hicks PJ, Colicigno CJ, Beck SR, Brown WM, Rich SS, Bowden DW: **A genome-wide scan for type 2 diabetes in African-American families reveals evidence for a locus on chromosome 6q.** *Diabetes* 2004, **53:**830-837.
77. Watanabe I, Tomita A, Shimizu M, Sugawara M, Yasumo H, Koishi R, Takahashi T, Miyoshi K, Nakamura K, Izumi T, Matsushita Y, Furukawa H, Haruyama H, Koga T: **A study to survey susceptible genetic factors responsible for troglitazone-associated hepatotoxicity in Japanese patients with type 2 diabetes mellitus.** *Clin Pharmacol Ther* 2003, **73:**435-455.
78. Gloria-Bottini F, Magrini A, Antonacci E, La Torre M, Di Renzo L, De Lorenzo A, Bergamaschi A, Bottini E: **Phosphoglucomutase genetic polymorphism and body mass.** *Am J Med Sci* 2007, **334:**421-425.
79. Spencer N, Hopkinson DA, Harris H: **Phosphoglucomutase polymorphism in man.** *Nature* 1964, **204:**742-745.
80. March RE, Putt W, Hollyoake M, Ives JH, Lovegrove JU, Hopkinson DA, Edwards YH, Whitehouse DB: **The classical human phosphoglucomutase (PGM1) isozyme polymorphism is generated by intragenic recombination.** *Proc Natl Acad Sci USA* 1993, **90:**10730-10733.
81. Zeggini E, Scott LJ, Saxena R, Voight BF: **Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes.** *Nat Genet* 2008, **40:**638-645.

82. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: *Numerical Recipes in C: The Art of Scientific Computing* 2nd edition. Cambridge: Cambridge University Press; 1992.
83. Moore JH, Hahn LW, Ritchie MD, Thornton TA, White BC: **Routine discovery of complex genetic models using genetic algorithms.** *Appl Soft Comput* 2004, **4:**79-86.
84. ***S* Statistic in Gene Mapping** [http://www.genemapping.cn/sumstat.html]
85. Witten IH, Frank E: *Data Mining: Practical Machine Learning Tools and Techniques* 2nd edition. San Francisco: Morgan Kaufmann; 2005.
86. **Weka 3: Data Mining Software in Java** [http://www.cs.waikato.ac.nz/ml/weka/]
87. **Multifactor Dimensionality Reduction** [http://www.multifactordimensionalityreduction.org/]
88. **JLIN: A Java Based Linkage Disequilibrium Plotter** [http://www.genepi.org.au/jlin.html]

A.2. Chemometrics and Intelligent Laboratory Systems

Piroonratana, T., Wongseree, W., Assawamakin, A., Paulkhaolarn, N., Kanjanakorn, C., Sirikong, M., Thongnoppakhun, W., Limwongse, C. and Chaiyaratana, N. (2009). Classification of haemoglobin typing chromatograms by neural networks and decision trees for thalassaemia screening. *Chemometrics and Intelligent Laboratory Systems*, *99*, 101-110. (2009 Journal Impact Factor = 2.111)

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Classification of haemoglobin typing chromatograms by neural networks and decision trees for thalassaemia screening

Theera Piroonratana [a], Waranyu Wongseree [a], Anunchai Assawamakin [b], Nuttawut Paulkhaolarn [c], Chompunut Kanjanakorn [c], Monchan Sirikong [c], Wanna Thongnoppakhun [b], Chanin Limwongse [b], Nachol Chaiyaratana [a,b,*]

[a] *Department of Electrical Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand*
[b] *Division of Molecular Genetics, Department of Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand*
[c] *Siriraj Thalassemia Program Project, Department of Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand*

ARTICLE INFO

Article history:
Received 26 March 2008
Received in revised form 24 July 2009
Accepted 27 July 2009
Available online 3 August 2009

Keywords:
C4.5
Haemoglobin typing
Neural network
Random forests
Thalassaemia

ABSTRACT

This article presents an application of a neural network and decision trees in thalassaemia screening. The aim is to classify thirteen classes of thalassaemia abnormality and one control class by inspecting the distribution of multiple types of haemoglobin in blood specimens, which are identified via high performance liquid chromatography (HPLC). C4.5 and random forests are the chosen architecture for decision tree implementation. For comparison, multilayer perceptrons are explored in classification via a neural network. The stratified 10-fold cross-validation results indicate that the best classification performance with overall accuracy of 97.2% (sensitivity = 97.2% and specificity = 99.8%) is achieved when C4.5 is used in conjunction with samples which have been pre-processed with input attribute discretisation and redundant attribute removal. Subsequently, C4.5 is applied to an additional sample set in a clinical trial which results in overall accuracy of 93.1% (sensitivity = 93.1% and specificity = 99.5%). These results suggest that a combination of C4.5 with haemoglobin typing analysis via HPLC may give rise to a guideline for further investigation of thalassaemia classification.

1. Introduction

Thalassaemia is a genetic disease that causes a reduction in the life span of a red blood cell [1]. The disease is a result of an abnormality in the genes that regulate the formation of a protein called globin, which is a major component of haemoglobin (Hb). Each red blood cell contains approximately 300 million molecules of haemoglobin. Hence, a change in the structure of globin affects the structure and functionality of a red blood cell. A globin molecule contains two parts: α-globin and β-globin. The α-globin contains 141 amino acids, which are regulated by genes on chromosome 16. The β-globin consists of 146 amino acids, which are governed by genes on chromosome 11. Since the regulatory genes reside on two autosomes, the transmission mode of abnormal genes is autosomal recessive. Hence, a person must have two copies of a recessive gene on the same chromosome in order to have the disease. In order to make the diagnosis, the blood characteristics must be analysed. A complete blood count (CBC) and haemoglobin typing are the primary screening tests for a laboratory diagnosis of thalassaemia. However, there is still a limitation in the analysis of data due to a large number of possible candidate characteristics. In addition, there are various types of thalassaemia and thalassaemia trait. (Persons with thalassaemia trait do not have the disease but inherit genes that cause the disease.) As a result, a manual diagnostic process can only be carried out by specialists [2–4].

Early attempts to formulate an automated diagnostic tool concentrate on analysing CBC data with image analysis [5], statistical [6] and clustering techniques [7]. Later, the implementation protocol has shifted to the expert systems, in which both rule-based [8–10] and hybrid neural network/rule-based systems [11] have been successfully tested in clinical trials. Nonetheless, these tools broadly differentiate between a wide range of blood-related diseases including various types of anaemia. In order to narrow the diagnostic target down to the differentiation between thalassaemic patients, persons with thalassaemia trait and normal subjects, an alternative automated diagnostic tool is required. Recently, a successful implementation of a multilayer perceptron [12–14], a *k*-nearest neighbour technique [13], a support vector machine [13] and a genetic programming based decision tree [14] as a thalassaemic diagnostic tool has been reported. Among these tools, the multilayer perceptron [15] emerges as the most suitable tool for thalassaemia classification problems in Thailand which cover higher varieties of haemoglobinopathies than other

* Corresponding author. Department of Electrical Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok, 1518 Piboolsongkram Road, Bangsue, Bangkok 10800, Thailand. Tel.: +66 2 9132500x8410; fax: +66 2 5856149.
E-mail addresses: nchl@kmutnb.ac.th, n.chaiyaratana@gmail.com (N. Chaiyaratana).

0169-7439/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2009.07.014

countries [16]. In the early investigation, it has been demonstrated that a multilayer perceptron can handle a problem with 13 classes of thalassaemic abnormality and two classes of normal subjects with and without iron deficiency [14].

Although the use of a neural network with CBC inputs has been proven to be successful, further investigation into automated thalassaemia classification is still required. Specifically, haemoglobin typing data should also be considered as possible inputs [17]. This is because CBC and haemoglobin typing data represent different aspects of blood characteristics. CBC information is useful for the diagnosis of various types of anaemia while haemoglobin typing data alone can confirm the haemoglobinopathies. In this article, the possibility of using haemoglobin typing data in automatic thalassaemia classification is investigated. The choices of classifier for the task include a multilayer perceptron, a C4.5 decision tree [18] and random forests [19]. A multilayer perceptron is selected for this study because it has been proven to be a suitable classifier in early investigations [12–14]. In contrast, both C4.5 and random forests have never been used in thalassaemia classification problems. Nonetheless, both classifiers have been successfully applied to many chemometric applications. For instance, C4.5 has been used for ion chromatography detection [20,21] while random forests have been implemented for prediction of drug's chromatographic retention time [22], near-infrared spectrum analysis of red grape homogenates [23], and classification of prostate cancer [24] and agro-industrial products [25].

With the availability of haemoglobin typing data and selected choices of classifier, an investigation can be conducted as follows. Firstly, the data are pre-processed via input attribute (feature) discretisation and reduction. It has been reported that a proper discretisation of continuous-valued attributes can improve the classification performance of decision tree classifiers [26]. In addition, an early investigation indicates that measured blood characteristics usually contain non-informative attributes which can be eliminated without affecting the classification outcome [14]. The attributes are thus discretised via an information-theoretic technique [26] while redundant attributes are eliminated using a correlation-based feature selection technique [27]. As a result, two reduced-attribute data sets are available for classifier benchmarking: continuous- and discrete-valued data sets. Since the attribute discretisation is also required prior to the continuous-valued attribute reduction in correlation analysis [27], eliminated attributes from the original continuous-valued data coincide with those from the derived discrete-valued data. The use of both continuous- and discrete-valued data in the benchmarking of neural networks is necessary since there is not enough evidence which suggests that one attribute representation is better than the other. After the classifier performance evaluation for both cases of attribute representation is completed, the best classifier together with the suitable attribute format is picked for a clinical trial involving a separate data set. Finally, classification analysis of clinical trial results is carried out to determine the feasibility of the selected classifier. Every step in the procedure described above is illustrated in Fig. 1 and implemented using a WEKA package [28].

The organisation of this article is as follows. In Section 2, materials and methods are explained. These include the description of haemoglobin typing data, an information-theoretic attribute discretisation technique, a correlation-based feature selection technique, a multilayer perceptron, C4.5, and random forests. Results from attribute discretisation and redundancy elimination, classifier benchmarking and a clinical trial are then discussed in Section 3. Finally, the conclusions and further works are given in Section 4.

2. Materials and methods

2.1. Haemoglobin typing data sets

A blood specimen generally contains more than one type of haemoglobin. With the use of high performance liquid chromatography (HPLC) [29], the haemoglobin contents in each blood specimen can be characterised. Various types of thalassaemia are identifiable through the difference in proportion of haemoglobin contents [30–32]. Haemoglobin typing results are usually obtained in the form of elution chromatograms [33]. Typical elution chromatograms of a normal specimen and a specimen from a person with Hb E trait are illustrated in Fig. 2. The normal specimen is mostly made up from Hb A_0 while the specimen from a person with Hb E trait consists of Hb F, Hb A_0 and Hb E. Since different types of haemoglobin are detectable in the form of elution peaks at different retention time, a chromatogram can be divided into multiple sections where each section occupies a non-overlapping range of retention time. Each chromatogram section would represent a unique input feature or attribute for thalassaemia classification in which the percentage of haemoglobin in the elution profile corresponds to the attribute value. In this investigation, a chromatogram is divided into eight sections. The attribute set and the associated types of haemoglobin are summarised in Table 1. In Table 1, eight attributes are defined according to the possible range of retention time in a haemoglobin chromatogram. Some attributes are related to known types of haemoglobin while the others are corresponded to unknown haemoglobin.

Two confirmed diagnosis data sets are acquired for this investigation. The first data set is created for the evaluation of classifier performance while the second set is used in a clinical trial. The data set for classifier evaluation consists of 150 samples which represent the majority of blood specimens from adults that need to be screened for thalassaemia. On the other hand, the data set for clinical trial contains 1000 samples and represents a typical distribution of specimens which are submitted for screening during a fixed time period. This data set is collected from Siriraj Hospital, Bangkok, Thailand during 1 August 2007 and 31 October 2007. The data acquisition has been conducted in accordance with the Faculty of Medicine Siriraj Hospital Ethics Committee's guidelines and in accordance with the Helsinki Declaration. In addition, informed consent has been obtained from all individuals. The description of these two sample sets is summarised in Table 2. From Table 2, the samples are made up from seven groups of thalassaemic patients, five groups of persons with thalassaemia trait, one group of persons with abnormal haemoglobin and one group of normal subjects. It is noticed that some types of thalassaemia in the data set for classifier benchmarking are not presented in the specimens collected for the clinical trial. Further, samples from persons with α-thalassaemia 1 and α-thalassaemia 2 traits are not included in this study. This is because haemoglobin typing characteristics cannot be used to differentiate between these two groups and the normal subject group. Generally, CBC and genotyping confirmation is needed for the diagnosis of these two types of thalassaemia trait [31].

2.2. Attribute discretisation

When attributes in the classification problem of interest are continuous-valued attributes, it is possible to transform these attributes into discrete-valued attributes. This transformation can be viewed as a form of data pre-processing procedure. The attribute discretisation technique that is selected for the current application is proposed by Fayyad and Irani [26]. The technique is an information-theoretic technique that employs entropy-based splitting and minimum description length stopping criteria. A chosen cut point within the range of each attribute values is guaranteed to lie between two class boundaries. A candidate cut point is introduced recursively to each sample subset and is acceptable if a significant information gain—the difference between the information value with and without the split— is achieved. For a sample set S, which contains samples from m classes denoted by $C_1,\ldots,C_m$, the class entropy of S is defined as

$$Ent(S) = -\sum_{i=1}^{m} p(C_i,S)\log_2(p(C_i,S)) \quad (1)$$

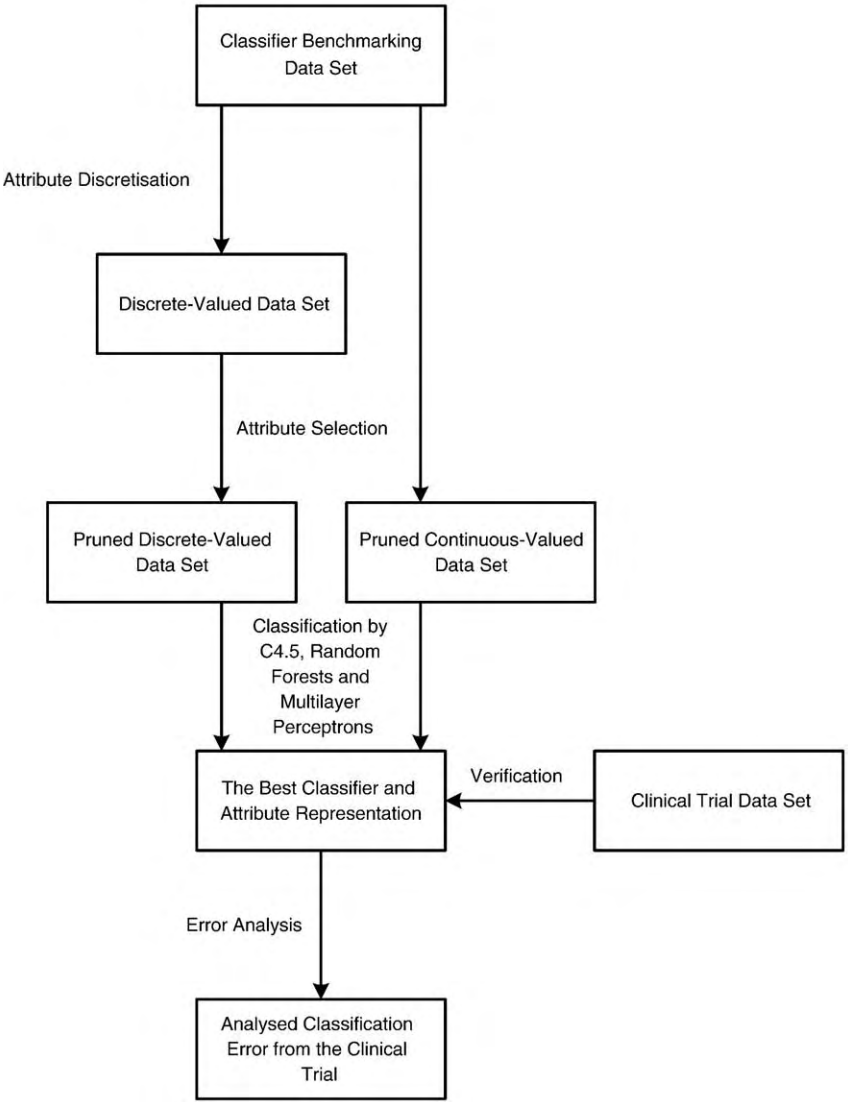

Classifier Benchmarking Data Set
Attribute Discretisation
Discrete-Valued Data Set
Attribute Selection
Pruned Discrete-Valued Data Set
Pruned Continuous-Valued Data Set
Classification by C4.5, Random Forests and Multilayer Perceptrons
The Best Classifier and Attribute Representation
Verification
Clinical Trial Data Set
Error Analysis
Analysed Classification Error from the Clinical Trial

Fig. 1. Schematic diagram for the methodology employed in the investigation.

where $p(C_i,S)$ is the proportion of samples in S that belong to class C_i. A cut point T, which is introduced to an attribute A in the sample set S, will create a partition that has a class information entropy

$$E(A,T,S) = \frac{|S_1|}{|S|}Ent(S_1) + \frac{|S_2|}{|S|}Ent(S_2) \tag{2}$$

where S_1 and S_2 are sample subsets of S and $S_1 + S_2 = S$. According to the minimum description length stopping criterion, a cut point T is accepted if and only if

$$Gain(A,T,S) > \frac{\log_2(|S|-1)}{|S|} + \frac{\log_2(3^m-2)-[mEnt(S)-m_1Ent(S_1)-m_2Ent(S_2)]}{|S|} \tag{3}$$

where m_1 and m_2 are the numbers of classes in the subsets S_1 and S_2, respectively and $Gain(A, T, S)$ is the information gain of the cut point, which is defined by

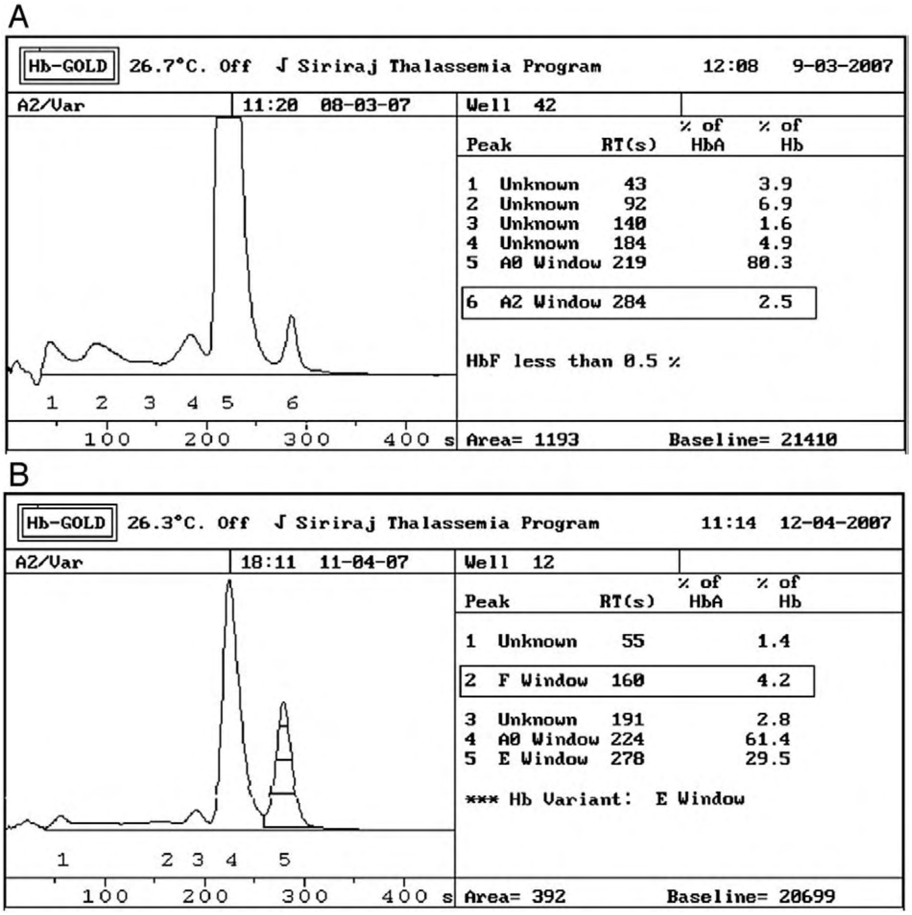
$$Gain(A,T,S) = Ent(S) - \frac{|S_1|}{|S|}Ent(S_1) - \frac{|S_2|}{|S|}Ent(S_2). \tag{4}$$

2.3. Attribute selection

Attribute selection is the process of identifying and removing irrelevant and redundant information from input features. This procedure has been proven to help improving the classification robustness in thalassaemia classification [13,14]. The chosen attribute selection method for the current investigation is a correlation-based feature selection technique [27]. The technique describes an attribute subset evaluation heuristic that considers both the usefulness of individual features in the classification task and the level of inter-correlation among features. Each attribute subset will be assigned a score given by

$$Merit_F = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k+1)\bar{r}_{ff}}} \tag{5}$$

where $Merit_F$ is the heuristic merit of a k-attribute subset F, $\bar{r}_{cf}$ is the average feature–class correlation and $\bar{r}_{ff}$ is the average feature–feature inter-correlation. An attribute subset will receive a high merit score if it contains features that are highly correlated with the class and at the same time have low inter-correlation among each other. Hall and Holmes [27] suggest that continuous-valued attributes should be



```
Hb-GOLD  26.7°C. Off  √ Siriraj Thalassemia Program        12:08   9-03-2007
A2/Var              11:20  08-03-07   Well  42
                                                     % of    % of
                                      Peak         RT(s)   HbA      Hb

                                      1  Unknown     43              3.9
                                      2  Unknown     92              6.9
                                      3  Unknown    140              1.6
                                      4  Unknown    184              4.9
                                      5  A0 Window 219              80.3

                                      6  A2 Window 284               2.5

                                      HbF less than 0.5 %

                                      Area= 1193         Baseline= 21410
```

```
Hb-GOLD  26.3°C. Off  √ Siriraj Thalassemia Program        11:14  12-04-2007
A2/Var              18:11  11-04-07   Well  12
                                                     % of    % of
                                      Peak         RT(s)   HbA      Hb

                                      1  Unknown     55              1.4

                                      2  F Window   160              4.2

                                      3  Unknown    191              2.8
                                      4  A0 Window 224              61.4
                                      5  E Window   278             29.5

                                      *** Hb Variant:  E Window

                                      Area= 392          Baseline= 20699
```

Fig. 2. Elution chromatograms of (A) a normal specimen and (B) a specimen from a person with Hb E trait that are obtained from an Hb Gold HPLC system. RT(s) represents the retention time in seconds for each fraction of elute. % of Hb represents the percentage of haemoglobin in the elution peak.

discretised using the information-theoretic technique [26] prior to the merit score calculation.

In order to avoid searching through all possible attribute subsets, Hall and Holmes [27] recommend an application of a simple hill-climbing technique for the best subset identification. Basically, the hill-climbing search begins with an empty set and evaluates each individual input feature to locate the best single attribute. The search then tries combining each remaining feature with the best attribute to identify the best attribute pair. Next, the search attempts the combination between each remaining feature and the best attribute pair to find the best three-attribute subset. The search continues until an attribute addition to the previously identified best subset does not show an improvement in the merit score.

Table 1
Input features or attributes for thalassaemia classification.

Attribute	Type of haemoglobin	Retention time (s)
1	Hb Bart's	0–68
2	Hb A_{1C}, Hb F	69–160
3	Unknown	161–199
4	Hb A_0	200–230
5	Unknown	231–249
6	Hb E	250–280
7	Hb A_2	281–289
8	Hb D, Hb S, Hb Constant Spring, Hb C	290–320

Each attribute represents different type of haemoglobin and occupies a unique range of retention time.

2.4. C4.5 decision tree

C4.5 decision tree is one of the most widely used and practiced tools for inductive inference [18]. A decision tree is generally constructed in a top-down manner. The tree construction begins at the root node where each input feature or attribute is evaluated using a statistical test to determine how well it alone classifies the training samples. The best attribute is selected and used as the test at the root node of the tree. A descendant of the root node is then created for

Table 2
Two data sets for thalassaemia classification.

Class	Description	Category	Number of samples	
			Classifier benchmarking	Clinical trial
1	Normal subject	Normal	15	281
2	Hb Constant Spring trait	Trait	9	35
3	Hb E trait	Trait	15	500
4	α-Thalassaemia 1 trait + Hb E trait	Trait	15	43
5	Homozygous Hb E	Trait	15	64
6	β-Thalassaemia trait	Trait	15	44
7	Abnormal haemoglobin	N/A	6	6
8	Hb H disease	Disease	12	4
9	Hb H disease with Constant Spring	Disease	12	0
10	EA Bart's disease	Disease	9	3
11	HPFH disease	Disease	10	2
12	Homozygous β-thalassaemia	Disease	5	0
13	β^0-Thalassaemia/Hb E	Disease	6	14
14	β^+-Thalassaemia/Hb E	Disease	6	4
	Total		150	1000

The first set contains 150 samples and is used for classification benchmarking. The second set consists of 1000 samples which are collected for a clinical trial.

either each possible value of this attribute if the attribute value is discrete or each possible discretised interval of this attribute if the attribute value is continuous. Next, the training samples are sorted to the appropriate descendant node. The entire process is subsequently repeated using the training samples associated with each descendant node to select the best attribute to test at that point in the tree. This forms a greedy search for an acceptable decision tree, in which the algorithm never backtracks to reconsider earlier choices. Although a new node can always be added to the tree until all samples which are assigned to one node belong to the same class, C4.5 does not allow the tree to grow to its maximum depth. As a result, a node is only introduced to the tree only when there are a sufficient number of samples left from sorting. After the complete tree has been constructed, a tree pruning is usually carried out to avoid data over-fitting.

The statistical test for assigning an attribute to each node in C4.5 also employs an entropy-based measure. The chosen attribute is the one with the highest information gain ratio among available attributes at the tree construction step considered. The information gain ratio $GainRatio(A, S)$ of an attribute A relative to a sample set S is defined as

$$GainRatio(A,S) = \frac{Gain(A,S)}{SplitInformation(A,S)} \tag{6}$$

$$\text{where} \quad Gain(A,S) = Ent(S) - \sum_{v \in V} \frac{|S_v|}{|S|} Ent(S_v) \tag{7}$$

$$\text{and} \quad SplitInformation(A,S) = -\sum_{v \in V} \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|}. \tag{8}$$

V is the set of all possible values of attribute A and S_v is the subset of S for which attribute A has value v. This information gain ratio can be calculated straightaway for discrete-valued attributes. For continuous-valued attributes, it is necessary to partition the attribute value into a discrete set of intervals prior to the calculation of information gain ratio. Quinlan [18] suggests that an appropriate threshold can be used to partition a continuous-valued attribute into two intervals. Let $\{v_1,v_2,\ldots,v_{|S|}\}$ be the sorted attribute values of attribute A. Candidate values that need to be considered are only the midpoints between v_i and v_{i+1}. The chosen midpoint is the value that leads to the partition that gives the highest information gain ratio according to Eq. (6). After the best midpoint is selected, C4.5 will pick the largest attribute value in A that does not exceed this midpoint as the threshold. Quinlan [18] explains that this strategy ensures that all threshold values appearing in the tree actually occur in the data.

2.5. Random forests

Random forests refer to a collection or ensemble of decision trees [19]. The technique takes a majority vote result from all trees as the class decision. Hence, the tree structures should be significantly diversified in order for the majority-vote concept to be applicable. This can be achieved by replacing the greedy search strategy for attribute selection as implemented in C4.5 with a stochastic attribute selection procedure. Breiman [19] suggests that an attribute for each node in a tree can be randomly selected from a small group of input features. Further, empirical studies indicate that a feature group size of one is sufficient. As a result, an attribute is randomly selected from available attributes for each node in this investigation. Another main difference between C4.5 and random forests is that each tree in random forests is allowed to grow to its maximum size. This would not lead to data over-fitting since the overall class decision would rely on outcomes from multiple trees within the forest [19].

Similar to C4.5, random forests can handle problems with discrete-valued attributes straightaway. Again, continuous-valued attributes must be split into discrete intervals during tree construction. Similar to C4.5 which uses an information gain ratio to determine the best split location on the attribute value range, Breiman et al. [34] introduces a *Gini* index for the same task. The *Gini* index is an "impurity" measure that directly relates to the proportion of classes in a sample set. The index reaches the value of zero when only one class is present in the collection and attains the maximum value when class sizes in the collection are equal. Using the same notation for Eq. (1), the *Gini* index of a sample set S is defined by

$$Gini(S) = \sum_{i \neq j} p(C_i,S)p(C_j,S) = 1 - \sum_i p(C_i,S)^2. \tag{9}$$

The best split location on attribute A is the one that most decreases the *Gini* index. This is achieved when

$$\Delta Gini(A,S) = Gini(S) - \frac{|S_1|}{|S|} Gini(S_1) - \frac{|S_2|}{|S|} Gini(S_2) \tag{10}$$

is minimal. S_1 and S_2 denotes the sample subsets of S after the split and $S_1 + S_2 = S$.

2.6. Neural network

A neural network is an interconnected group of artificial neurons that uses a mathematical or computational model for information processing based on a connectionist approach. The neural network that is selected for this implementation is a multilayer perceptron [15]. The model of a neuron is illustrated in Fig. 3(A). From Fig. 3(A), q input signals are received by the neuron. These inputs are weighted and linearly summed together. The threshold, which can be treated as an extra connection weight, is then applied to the weighted-sum result. Thus, the linear combiner output (z) or input to the activation function is given by

$$z = \sum_{i=0}^{q} w_i u_i \tag{11}$$

where u_i is the ith input to the neuron and w_i is the connection weight on input u_i. In addition, $u_0 = -1$ and w_0 is the threshold. The neuron output $(h(z))$ is the output from the activation function and is given by

$$h(z) = \frac{1}{1 + \exp(-z)}. \tag{12}$$

The output signal from each neuron is thus limited by a logistic sigmoid function. This described neuron model is used throughout the multilayer feed-forward network depicted in Fig. 3(B).

In this implementation, the number of network inputs is equal to the number of attributes or input features while the number of network outputs is equal to the number of output classes. Each output

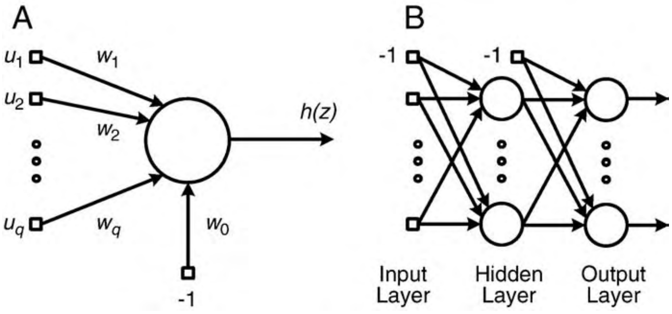


Fig. 3. Schematic diagram of a multilayer perceptron: (A) computational model of a neuron and (B) feed-forward network structure that contains one hidden layer.

neuron represents one of possible output classes with the highest-valued output taken as the network prediction. This is often referred to as a 1-of-*n* output encoding technique [35]. The network has a single hidden layer—a layer of neurons that receive attributes as inputs and send signals to output neurons—in which the number of neurons in the hidden layer can vary. As a result, the observation displays the effect of network non-linearity on the classification performance.

The multilayer perceptron will be trained by a back-propagation algorithm. The algorithm uses a sample-by-sample updating rule for adjusting connection weights in the network. In one algorithm iteration, a training sample is presented to the network. The signal is then fed in a forward manner through the network until the network output is obtained. The error between the actual and aimed network outputs is calculated and used to adjust the connection weights. Basically, the adjustment procedure, derived from a gradient descent method, is used to reduce the error magnitude. The procedure is firstly applied to the connection weights in the output layer, followed by the connection weights in the hidden layer. The iteration is completed after all connection weights in the network have been adjusted.

3. Results and discussions

3.1. Attribute discretisation and selection

The original haemoglobin typing data contains eight attributes as shown in Table 1. The attributes in classifier evaluation samples have been discretised using the information-theoretic technique [26] described in Section 2.2. The discrete interval of each attribute is illustrated in Table 3. After applying the correlation-based feature selection technique [27] described in Section 2.3 to the discretised attribute set, it is found that attributes 3 and 5 are redundant and can be omitted from the classification task. This decision is based on the merit scores shown in Table 4. These two attributes are related to unknown haemoglobin. This implies that each thalassaemia abnormality can be described by a unique combination of known types of haemoglobin.

3.2. Classifier performance evaluation

Three candidate classifiers—a multilayer perceptron, C4.5 and random forests—are benchmarked using the reduced-attribute evaluation data in stratified 10-fold cross-validation experiments [36]. All three classifiers can take both continuous- and discrete-valued attributes. The best network size for multilayer perceptron is identified by varying the number of hidden nodes. The numbers of hidden nodes used in the trial are 10, 15, 20 and 25. The training process of each multilayer perceptron is terminated prior to the occurrence of data over-fitting. The appropriate number of training epochs is identified via a training and validation approach [35]. A validation data set is generally employed to detect the point where the training error continues to decrease while the validation error starts to increase. This is the point where data over-fitting usually occurs. In this investigation, 10% of samples in the data set are used as the validation set.

In contrast to parametric techniques such as neural networks, random forests avoid the occurrence of data over-fitting by constructing multiple trees for the ensemble. In this implementation, the ensemble is made up from ten trees. Although this setting is significantly less than the number of trees recommended by Breiman [19], which is between 100 and 200, empirical studies indicate that the difference in the classification accuracy of random forests with 10 and 100 trees in this application is negligible. Similar to random forests, C4.5 also has a built-in mechanism for data over-fitting avoidance. However, since C4.5 produces a single decision tree, data over-fitting is avoided by employing a rule post-pruning strategy [18].

Table 3
Discrete intervals of the attributes.

Attribute	Type of haemoglobin	Interval	% of haemoglobin
1	Hb Bart's	1	$Hb \le 1.85$
		2	$1.85 < Hb \le 4.70$
		3	$4.70 < Hb \le 7.95$
		4	$7.95 < Hb \le 10.40$
		5	$10.40 < Hb \le 23.00$
		6	$Hb > 23.00$
2	Hb A_{1C}, Hb F	1	$Hb \le 0.35$
		2	$0.35 < Hb \le 3.50$
		3	$3.50 < Hb \le 15.25$
		4	$15.25 < Hb \le 24.35$
		5	$24.35 < Hb \le 67.75$
		6	$Hb > 67.75$
3	Unknown	1	$Hb \ge 0.00$
4	Hb A_0	1	$Hb \le 13.05$
		2	$13.05 < Hb \le 50.70$
		3	$50.70 < Hb \le 62.35$
		4	$62.35 < Hb \le 65.95$
		5	$65.95 < Hb \le 73.80$
		6	$Hb > 73.80$
5	Unknown	1	$Hb \le 0.50$
		2	$Hb > 0.50$
6	Hb E	1	$Hb \le 6.60$
		2	$6.60 < Hb \le 20.00$
		3	$20.00 < Hb \le 25.55$
		4	$25.55 < Hb \le 37.90$
		5	$37.90 < Hb \le 81.00$
		6	$Hb > 81.00$
7	Hb A_2	1	$Hb \le 0.40$
		2	$0.40 < Hb \le 2.15$
		3	$2.15 < Hb \le 3.65$
		4	$Hb > 3.65$
8	Hb D, Hb S, Hb Constant Spring, Hb C	1	$Hb \le 0.05$
		2	$0.05 < Hb \le 12.20$
		3	$Hb > 12.20$

The classification performance of the multilayer perceptron, C4.5 and random forests on reduced-attribute data obtained from stratified 10-fold cross-validation is summarised in Table 5. The results from each classifier are obtained from 30 runs where new cross-validation folds are generated for each run. The results suggest that for this task, discrete-valued attributes appear to be better than continuous-valued attributes at representing the problem inputs. In addition, decision trees also have higher classification accuracy than multilayer perceptrons in which C4.5 possesses the highest performance. These results can be interpreted as follows.

Firstly, consider the results from multilayer perceptrons with different number of hidden nodes. In the case of continuous-valued attribute, an increase in the number of hidden nodes leads to a significant increase in the classification accuracy. Hence, the most suitable number of hidden nodes is 25. The statistical significance is

Table 4
Merit scores for subsets of discrete-valued attributes.

Number of attributes	Attribute subset	Merit score
1	$\{Attr_6\}$	0.6905
2	$\{Attr_2, Attr_6\}$	0.7667
3	$\{Attr_2, Attr_4, Attr_6\}$	0.8178
4	$\{Attr_2, Attr_4, Attr_6, Attr_7\}$	0.8436
5	$\{Attr_1, Attr_2, Attr_4, Attr_6, Attr_7\}$	0.8521
6	$\{Attr_1, Attr_2, Attr_4, Attr_6, Attr_7, Attr_8\}$	0.8558
7	$\{Attr_1, Attr_2, Attr_4, Attr_5, Attr_6, Attr_7, Attr_8\}$	0.8313
8	$\{Attr_1, Attr_2, Attr_3, Attr_4, Attr_5, Attr_6, Attr_7, Attr_8\}$	0.5994

The displayed subsets are the subsets identified by a hill-climbing search. The merit scores indicate that attributes 3 and 5 are redundant and can be eliminated. Since the total number of attributes is relatively small, an exhaustive search for the best attribute subset is also carried out. The exhaustive search confirms that the displayed six-attribute subset is the optimal subset.

Table 5
Classification performance of the multilayer perceptron, C4.5 and random forests.

Index	Attribute	C4.5	Random forests	Multilayer perceptron			
				10 nodes	15 nodes	20 nodes	25 nodes
Accuracy (%)	Continuous	93.13 (0.82)	94.40 (1.01)	90.13 (1.17)	92.02 (1.34)	91.89 (1.08)	93.07 (1.16)
	Discrete	97.24 (0.89)	96.00 (1.15)	92.56 (0.94)	93.24 (0.69)	93.38 (0.78)	93.44 (0.76)
Sensitivity (%)	Continuous	93.13 (0.82)	94.40 (1.01)	90.13 (1.17)	92.02 (1.34)	91.89 (1.08)	93.07 (1.16)
	Discrete	97.24 (0.89)	96.00 (1.15)	92.56 (0.94)	93.24 (0.69)	93.38 (0.78)	93.44 (0.76)
Specificity (%)	Continuous	99.40 (0.10)	99.50 (0.11)	99.14 (0.13)	99.32 (0.13)	99.31 (0.13)	99.40 (0.11)
	Discrete	99.78 (0.07)	99.68 (0.11)	99.37 (0.08)	99.45 (0.10)	99.46 (0.09)	99.42 (0.08)

The results are averaged over 30 runs of stratified 10-fold cross-validation. The numbers in the brackets are standard deviations. The discretisation of attribute values leads to a statistically significant improvement in classification accuracy of all three classifiers ($p<0.05$).

determined via a *t*-test at a 95% confidence level where the results from different neural network settings are compared in a pair-wise manner. In contrast, the appropriate number of hidden nodes for the network that takes discrete-valued attributes is 15. This is because an increase in the number of hidden nodes does not produce a statistically significant change in classification accuracy after the number of hidden nodes is greater than 15. Nonetheless, further exploration of appropriate number of hidden nodes for the network with continuous-valued attributes appears to be unnecessary since a change in the classification accuracy is more driven by a transition from continuous-valued attributes to discrete-valued attributes.

The classification accuracy of decision trees is now compared. Random forests outperform C4.5 when continuous-valued attributes are used. However, C4.5 has higher classification accuracy than random forests in the case of discrete-valued attribute. Further inspection on the effect of attribute discretisation reveals that changing attribute representation can significantly improve classification performance of both neural networks and decision trees. In addition, this improvement is more prominent in C4.5 than random forests. This makes C4.5 with discrete-valued attributes the best decision tree classifier in this application.

Finally, the classification accuracy of multilayer perceptrons and C4.5 is compared. With the use of discrete-valued attributes, C4.5 outperforms the multilayer perceptron with 15 hidden nodes by 4%. This helps to confirm that the best classifier for the task at hand is C4.5. In addition to the superiority in accuracy, a single decision tree produced by C4.5 can give further insight into the relationship between attributes and output classes. The C4.5 decision tree, which is constructed using all classifier benchmarking samples, is illustrated in Fig. 4. In Fig. 4, the tree has the maximum depth of four in which attribute 6, which represents Hb E, is located at the root node. This attribute alone can be used to identify four classes including an EA Bart's disease, a person with both α-thalassaemia 1 trait and Hb E trait, a person with Hb E trait and a person with homozygous Hb E. A direct relationship between the Hb E attribute and three output classes among the mentioned classes is obvious. Together with the use of other attributes, a similar explanation can be deduced from the tree. It is noticed that the decision tree can be further simplified by merging a number of adjacent leave nodes that lead to the same class prediction into one node. These extra leave nodes are the result of the attribute discretisation prior to the tree construction.

3.3. Clinical trial

In the previous sub-section, C4.5 with discrete-valued attributes is proven to be the best approach for thalassaemia classification. In addition, the decision tree generated by C4.5 can also be represented by a set of decision rules; this is convenient for diagnostic interpretation. C4.5 has subsequently been used in a clinical trial involving 1000 samples. The distribution of classes within the sample set has been given in Table 2. The C4.5 decision tree illustrated in Fig. 4 is directly applied to the clinical trial data set where classification accuracy of 93.1% (sensitivity = 93.1% and specificity = 99.5%) has been achieved. In addition, the clinical trial data set with attribute discretisation and reduction is randomly split five times into 75% for training and 25% for testing of a C4.5 decision tree. The classification accuracy of 95.0% (sensitivity = 95.0% and specificity = 99.4%) has subsequently been achieved. A higher accuracy for this latter case is to be expected since the clinical trial data are used both to train and to test the C4.5 decision tree. The classification errors can be divided into three main categories: a misclassification within the same super-group, a false prediction of high severity and a false prediction of low severity. The categorisation of errors in this manner is crucial since the problem covers multiple types of abnormality. These error categories can be explained as follows.

The misclassification within the same super-group occurs when either a person with thalassaemia trait is identified as being a person with another type of thalassaemia trait, or a thalassaemic patient is misdiagnosed as being another type of patient. In this study, the persons with thalassaemia trait, thalassaemic patients and normal subjects are clustered into three super-groups based upon the severity of the exhibited thalassaemic characteristics. The details of all three super-groups are given in Table 6. From Table 6, the super-groups are made up from the minor trait/normal super-group, the major trait super-group and the disease super-group. It is noticed that normal subjects are gathered into the minor trait/normal super-group that also contains persons with Hb Constrant Spring and Hb E traits. They are grouped together since the blood characteristics of the sample within this super-group are very similar. Using the same argument, homozygous Hb E samples are placed in the same super-group as mixed α-thalassaemia 1 and Hb E trait, β-thalassaemia trait and abnormal haemoglobin samples.

The last two types of classification error are the false predictions of low and high severity. In this study, the false prediction of high severity refers to the situation when a sample is misidentified as belonging to a super-group with a higher severity of thalassaemic characteristics. On the other hand, the false prediction of low severity refers to the case when a sample is misclassified as being a member of a super-group with a lower severity of thalassaemic characteristics.

The numbers of misclassified samples, which are extracted from confusion matrices are given in Tables 7 and 8. It can be clearly seen that most of classification errors stem from samples in the minor trait/normal and major trait super-groups. The summary of classification errors from Tables 7 and 8 can be expressed as percentages as given in Table 9. It is noticeable that the misclassification within the disease super-group and the false predictions of severity levels for samples from thalassaemic patients are low. This agrees with the previous observation regarding the primary cause of errors.

4. Conclusions and further works

In this article, a thalassaemia classification problem is investigated. The objective is to identify automatically whether the human subject is a person with abnormal haemoglobin, a person with thalassaemia trait, a thalassaemic patient, or a normal person using haemoglobin typing data from HPLC. The derived data sets contain eight input features or attributes and 14 distinct classes. Each attribute reflects the percentage of haemoglobin at a specific chromatographic

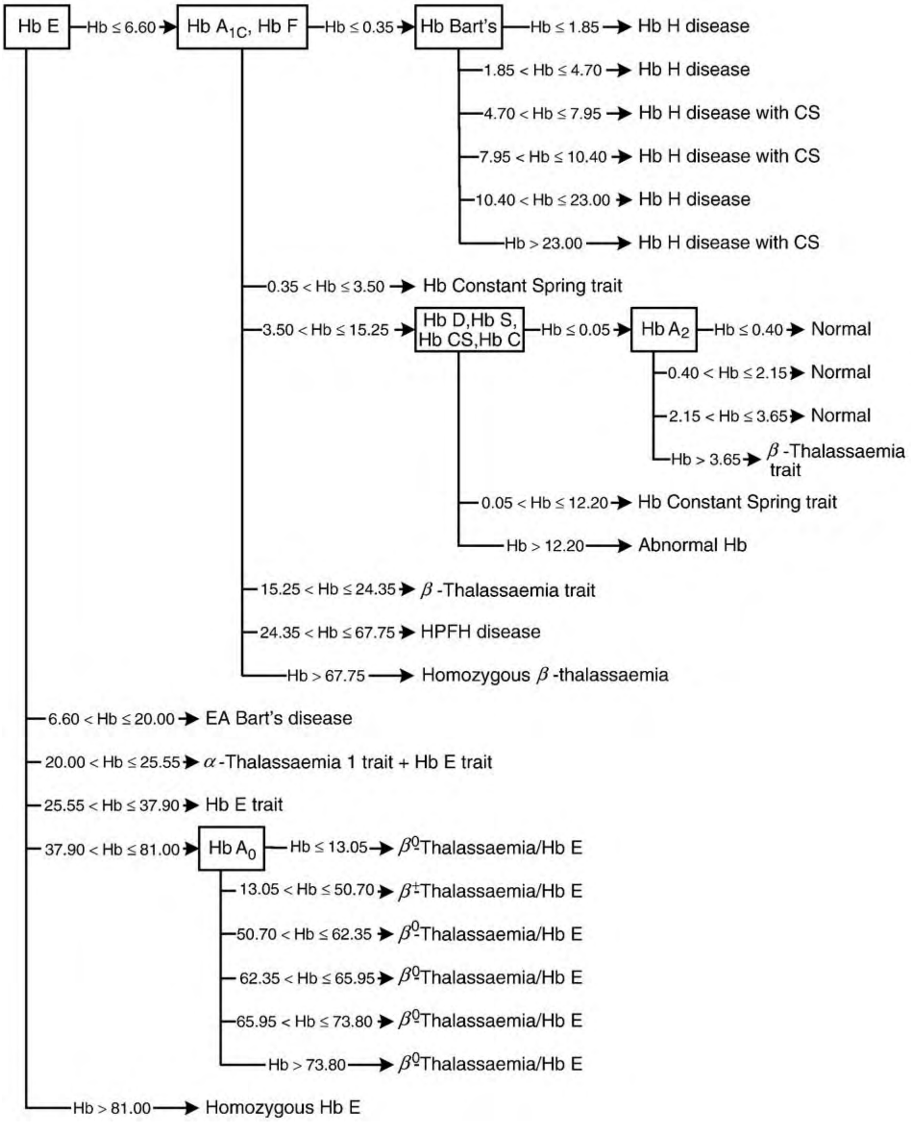

Hb E
Hb ≤ 6.60
Hb A1C, Hb F
Hb ≤ 0.35
Hb Bart's
Hb ≤ 1.85 Hb H disease
1.85 < Hb ≤ 4.70 Hb H disease
4.70 < Hb ≤ 7.95 Hb H disease with CS
7.95 < Hb ≤ 10.40 Hb H disease with CS
10.40 < Hb ≤ 23.00 Hb H disease
Hb > 23.00 Hb H disease with CS
0.35 < Hb ≤ 3.50 Hb Constant Spring trait
3.50 < Hb ≤ 15.25
Hb D,Hb S, Hb CS,Hb C
Hb ≤ 0.05
Hb A2
Hb ≤ 0.40 Normal
0.40 < Hb ≤ 2.15 Normal
2.15 < Hb ≤ 3.65 Normal
Hb > 3.65 β-Thalassaemia trait
0.05 < Hb ≤ 12.20 Hb Constant Spring trait
Hb > 12.20 Abnormal Hb
15.25 < Hb ≤ 24.35 β-Thalassaemia trait
24.35 < Hb ≤ 67.75 HPFH disease
Hb > 67.75 Homozygous β-thalassaemia
6.60 < Hb ≤ 20.00 EA Bart's disease
20.00 < Hb ≤ 25.55 α-Thalassaemia 1 trait + Hb E trait
25.55 < Hb ≤ 37.90 Hb E trait
37.90 < Hb ≤ 81.00
Hb A0
Hb ≤ 13.05 β0-Thalassaemia/Hb E
13.05 < Hb ≤ 50.70 β±Thalassaemia/Hb E
50.70 < Hb ≤ 62.35 β0-Thalassaemia/Hb E
62.35 < Hb ≤ 65.95 β0-Thalassaemia/Hb E
65.95 < Hb ≤ 73.80 β0-Thalassaemia/Hb E
Hb > 73.80 β0-Thalassaemia/Hb E
Hb > 81.00 Homozygous Hb E

Fig. 4. C4.5 decision tree which is constructed using discretised attributes. A set of screening rules can be extracted from the decision tree. For example, if the percentage of Hb E in a blood specimen is less than or equal to 6.60 while the combined percentage of Hb A_{1C} and Hb F from the same specimen is between 24.35 and 67.75, then it is most likely that the specimen is taken from an HPFH patient.

retention time. In other words, the attribute set covers multiple types of haemoglobin. The investigation is divided into two main parts: a classifier selection and a clinical trial. Candidate classifiers for the task include a multilayer perceptron, a C4.5 decision tree and random

Table 6
Clustering of persons with abnormal haemoglobin, persons with thalassaemia trait, thalassaemic patients and normal subjects into three super-groups.

Minor trait/normal	Major trait	Disease
Normal subject	α-Thalassaemia 1 trait + Hb E trait	Hb H disease
Hb Constant Spring trait	Homozygous Hb E	Hb H disease with Constant Spring
Hb E trait	β-Thalassaemia trait	EA Bart's disease
	Abnormal haemoglobin	HPFH disease
		Homozygous β-thalassaemia
		β^0-Thalassaemia/Hb E
		β^+-Thalassaemia/Hb E

Persons with abnormal haemoglobin are treated as persons with major thalassaemia trait due to their similarity in the blood characteristics.

Table 7
Detailed classification errors from applying the C4.5 decision tree in Fig. 4 to 1000 samples in the clinical trial.

Actual class		Identified class (number of misclassified samples)													
		Minor/normal			Major trait				Disease						
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Minor	1		19				2		6	2					
	2								1	1					
	3				9										
Major	4		1							1	3				
	5													5	
	6	6	4					1		1					
	7		1	1			2				1				1
Disease	8														
	9														
	10														
	11						1								
	12														
	13														
	14														

The description of each class has been given in Table 2. Almost all classification errors can be traced back to samples in the minor trait/normal and major trait super-groups.

Table 8
Detailed classification errors from the C4.5 decision tree which is trained and tested with 1000 samples in the clinical trial.

Actual class		Identified class (number of misclassified samples)													
		Minor/normal			Major trait				Disease						
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Minor	1		4				4		2						
	2														
	3				10										
Major	4	2									3				
	5													3	
	6	5	3												
	7			1			4				1				
Disease	8	4													
	9														
	10				2										
	11	2													
	12														
	13														
	14														

The description of each class has been given in Table 2. The numbers of misclassified samples are displayed in a manner that they can be directly compared against the numbers in Table 7. Similar to the results in Table 7, almost all classification errors can be traced back to samples in the minor trait/normal and major trait super-groups.

forests. The study involving stratified 10-fold cross-validation reveals that C4.5 is the most suitable classifier for the data that have been preprocessed by attribute discretisation and reduction. Subsequently, C4.5 is applied in the clinical trial and further analysis of the classification error indicates that the misclassification among disease classes and the false predictions of severity levels for samples from thalassaemic patients are low. This helps emphasise the suitability of C4.5 as an automated thalassaemic classification tool.

In order for the proposed automated classification procedure for thalassaemia screening to be applicable in clinical settings, a larger clinical trial may still be required. However, since C4.5 is the chosen classifier, an additional trial can be easily carried out. This is because a set of decision rules can be extracted from the tree illustrated in Fig. 4. The decision rules can subsequently be implemented in many user-friendly computer programs including spreadsheets and databases. This is an additional advantage of using C4.5 over a multilayer perceptron and random forests.

With the availability of knowledge regarding the relationship between haemoglobin typing inputs and thalassaemic class outputs and that between CBC inputs and similar outputs [14], the most obvious further study is to employ both types of inputs in the classification task. Since CBC and haemoglobin typing are always carried out for a laboratory diagnosis of thalassaemia, it is not difficult to acquire both types of data from the same blood specimen. In addition to thalassaemia classification, another possible further work is to apply the procedure for extracting informative features from chromatograms as described in this article to other pattern recognition problems. Examples of classification problems that involves the inspection of chromatograms include a diagnosis of liver and bile diseases [37], determination of herb's origins [38], differentiation of tea varieties [39] and ink identification for forensic purposes [40]. These examples illustrate that a wide range of chemometric applications can benefit from the feature extraction procedure explained in this investigation.

Table 9
Summary of classification errors from the clinical trial.

Type of classification error	Classification error (%)	
	Table 7	Table 8
Misclassification within the same super-group	2.2	0.8
Misclassification within the minor trait/normal super-group	1.9	0.4
Misclassification within the major trait super-group	0.3	0.4
Misclassification within the disease super-group	0.0	0.0
False prediction of high severity	3.3	2.3
Minor trait/normal identified as major trait	1.1	1.4
Major trait identified as disease	1.2	0.7
Minor trait/normal identified as disease	1.0	0.2
False prediction of low severity	1.4	1.9
Disease identified as major trait	0.1	0.2
Major trait identified as minor trait/normal	1.3	1.1
Disease identified as minor trait/normal	0.0	0.6
Total	6.9	5.0

The errors are described in terms of misclassification within the same super-group, false prediction of high severity and false prediction of low severity.

Acknowledgements

The authors are extremely grateful to two anonymous reviewers and Prof. Yi-Zeng Liang for their valuable comments and suggestions, which have contributed a lot towards improving the content and presentation of this article. The authors acknowledge Dr. Sissades Tongsima for his assistance on capturing chromatogram images from the Hb Gold HPLC system. Theera Piroonratana, Waranyu Wongseree and Anunchai Assawamakin were supported by the Thailand Research Fund (TRF) through the Royal Golden Jubilee Ph.D. Programme (Grant No. PHD/1.E.KN.50/A.1, PHD/1.E.KN.49/A.1 and PHD/4.I.MU.45/C.1, respectively). Chanin Limwongse was supported by the Mahidol Research Grant. Nachol Chaiyaratana was supported by the Thailand Research Fund and the Commission on Higher Education (CHE).

References

[1] D.J. Weatherall, J.B. Clegg, The Thalassemia Syndromes, 4th Edition, Blackwell Science, Malden, MA, 2001.
[2] C.V. Jimenez, J. Minchinela, J. Ros, Clinical and Laboratory Haematology 17 (1995) 151–155.
[3] A. Demir, N. Yarali, T. Fisgin, F. Duru, A. Kara, Pediatrics International 44 (2002) 612–616.
[4] G. Ntaios, A. Chatzinikolaou, Z. Saouli, F. Girtovitis, M. Tsapanidou, G. Kaiafa, Z. Kontoninas, A. Nikolaidou, C. Savopoulos, I. Pidonia, S. Alexiou-Daniel, Annals of Hematology 86 (2007) 487–491.
[5] P.R. Lund, R.D. Barnes, Lancet 300 (1972) 463–464.
[6] R.L. Engle, B.J. Flehinger, S. Allen, R. Friedman, M. Lipkin, B.J. Davis, L.L. Leveridge, Bulletin of the New York Academy of Medicine 52 (1976) 584–600.
[7] G. Barosi, M. Cazzola, C. Berzuini, S. Quaglini, M. Stefanelli, British Journal of Haematology 61 (1985) 357–370.
[8] S. Quaglini, M. Stefanelli, G. Barosi, A. Berzuini, Computers and Biomedical Research 19 (1986) 13–27.
[9] S. Quaglini, M. Stefanelli, G. Barosi, A. Berzuini, Computers and Biomedical Research 21 (1988) 307–323.
[10] G. Lanzola, M. Stefanelli, G. Barosi, L. Magnani, Computers and Biomedical Research 23 (1990) 560–582.
[11] N.I. Birndorf, J.O. Pentecost, J.R. Coakley, K.A. Spackman, Computers and Biomedical Research 29 (1996) 16–26.
[12] S.R. Amendolia, A. Brunetti, P. Carta, G. Cossu, M.L. Ganadu, B. Golosio, G.M. Mura, M.G. Pirastru, Medical Decision Making 22 (2002) 18–26.
[13] S.R. Amendolia, G. Cossu, M.L. Ganadu, B. Golosio, G.L. Masala, G.M. Mura, Chemometrics and Intelligent Laboratory Systems 69 (2003) 13–20.
[14] W. Wongseree, N. Chaiyaratana, K. Vichittumaros, P. Winichagoon, S. Fucharoen, Information Sciences 177 (2007) 771–786.
[15] D.E. Rumelhart, J.L. McClelland (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1, MIT Press, Cambridge, MA, 1986.
[16] S. Fucharoen, P. Winichagoon, Hemoglobin 21 (1997) 299–319.
[17] F. Kutlar, Hemoglobin 31 (2007) 243–250.
[18] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.
[19] L. Breiman, Machine Learning 45 (2001) 5–32.
[20] M. Mulholland, D.B. Hibbert, P.R. Haddad, P. Parslov, Chemometrics and Intelligent Laboratory Systems 30 (1995) 117–128.
[21] M. Mulholland, D.B. Hibbert, P.R. Haddad, C. Sammut, Chemometrics and Intelligent Laboratory Systems 27 (1995) 95–104.
[22] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y. Everingham, Chemometrics and Intelligent Laboratory Systems 76 (2005) 185–196.
[23] D. Donald, D. Coomans, Y. Everingham, D. Cozzolino, M. Gishen, T. Hancock, Chemometrics and Intelligent Laboratory Systems 82 (2006) 122–129.
[24] D. Donald, T. Hancock, D. Coomans, Y. Everingham, Chemometrics and Intelligent Laboratory Systems 82 (2006) 2–7.
[25] P.M. Granitto, C. Furlanello, F. Biasioli, F. Gasperi, Chemometrics and Intelligent Laboratory Systems 83 (2006) 83–90.

[26] U.M. Fayyad, K.B. Irani, in: R. Bajcsy (Ed.), Proceedings of the 13th International Joint Conference on Artificial Intelligence, 28 August–3 September 1993, Chambéry, France, Morgan Kaufmann, San Mateo, CA, 1993, pp. 1022–1027.

[27] M.A. Hall, G. Holmes, IEEE Transactions on Knowledge and Data Engineering 15 (2003) 1437–1447.

[28] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, San Francisco, CA, 2005.

[29] G.M. Clarke, T.N. Higgins, Clinical Chemistry 46 (2000) 1284–1290.

[30] C.-N. Ou, C.L. Rognerud, Clinica Chimica Acta 313 (2001) 187–194.

[31] J.M. Old, Blood Reviews 17 (2003) 43–53.

[32] R.B. Colah, R. Surve, P. Sawant, E. D'Souza, K. Italia, S. Phanasgaonkar, A.H. Nadkarni, A.C. Gorakshakar, Indian Journal of Pediatrics 74 (2007) 657–662.

[33] A. Joutovsky, J. Hadzi-Nesic, M.A. Nardi, Clinical Chemistry 50 (2004) 1736–1747.

[34] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth International Group, Belmont, CA, 1984.

[35] T.M. Mitchell, Machine Learning, McGraw-Hill, Singapore, 1997.

[36] R. Kohavi, in: C.S. Mellish (Ed.), Proceedings of the 14th International Joint Conference on Artificial Intelligence, 20–25 August 1995, Montréal, Québec, Canada, Morgan Kaufmann, San Mateo, CA, 1995, pp. 1137–1143.

[37] R.H. Zhao, B.F. Yue, J.Y. Ni, H.F. Zhou, Y.K. Zhang, Chemometrics and Intelligent Laboratory Systems 45 (1999) 163–170.

[38] C.-C. Chuang, W.-C. Wen, S.-J. Sheu, Journal of Separation Science 30 (2007) 1827–1832.

[39] A. Alcazar, O. Ballesteros, J.M. Jurado, F. Pablos, M.J. Martin, J.L. Vilches, A. Navalon, Journal of Agricultural and Food Chemistry 55 (2007) 5960–5965.

[40] A. Kher, M. Mulholland, E. Green, B. Reedy, Vibrational Spectroscopy 40 (2006) 270–277.

A.3. Biomedical Signal Processing and Control

Setsirichok, D., Piroonratana, T., Wongseree, W., Usavanarong, T., Paulkhaolarn, N., Kanjanakorn, C., Sirikong, M., Limwongse, C. and Chaiyaratana, N. (2011). Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening. *Biomedical Signal Processing and Control*, in press. (2009 Journal Impact Factor = 0.620)

Biomedical Signal Processing and Control xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Note

Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening

Damrongrit Setsirichok[a], Theera Piroonratana[a], Waranyu Wongseree[a], Touchpong Usavanarong[a], Nuttawut Paulkhaolarn[b], Chompunut Kanjanakorn[b], Monchan Sirikong[b], Chanin Limwongse[c], Nachol Chaiyaratana[a,c,*]

[a] Department of Electrical Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand
[b] Siriraj Thalassemia Program Project, Department of Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand
[c] Division of Molecular Genetics, Department of Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand

ARTICLE INFO

Article history:
Received 4 September 2010
Received in revised form 21 January 2011
Accepted 2 March 2011
Available online xxx

Keywords:
C4.5
Complete blood count
Haemoglobin typing
Multilayer perceptron
Naïve Bayes classifier
Thalassaemia

ABSTRACT

This article presents the classification of blood characteristics by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening. The aim is to classify eighteen classes of thalassaemia abnormality, which have a high prevalence in Thailand, and one control class by inspecting data characterised by a complete blood count (CBC) and haemoglobin typing. Two indices namely a haemoglobin concentration (HB) and a mean corpuscular volume (MCV) are the chosen CBC attributes. On the other hand, known types of haemoglobin from six ranges of retention time identified via high performance liquid chromatography (HPLC) are the chosen haemoglobin typing attributes. The stratified 10-fold cross-validation results indicate that the best classification performance with average accuracy of 93.23% (standard deviation = 1.67%) and 92.60% (standard deviation = 1.75%) is achieved when the naïve Bayes classifier and the multilayer perceptron are respectively applied to samples which have been pre-processed by attribute discretisation. The results also suggest that the HB attribute is redundant. Moreover, the achieved classification performance is significantly higher than that obtained using only haemoglobin typing attributes as classifier inputs. Subsequently, the naïve Bayes classifier and the multilayer perceptron are applied to an additional data set in a clinical trial which respectively results in accuracy of 99.39% and 99.71%. These results suggest that a combination of CBC and haemoglobin typing analysis with a naïve Bayes classifier or a multilayer perceptron is highly suitable for automatic thalassaemia screening.

1. Introduction

Thalassaemia is a genetic disease that has a high prevalence in Thailand [1]. It is a result of an abnormality in genes that govern the formation of a protein called globin, which is a major component of haemoglobin (Hb). Since each red blood cell contains approximately 300 million molecules of haemoglobin, a modification of globin structure affects the cell structure and functionality. This subsequently leads to the reduction in the life span of a red blood cell [2]. The globin protein contains two components: α-globin and β-globin. The α-globin and β-globin synthesis is regulated by genes on chromosomes 16 and 11, respectively. Since the transmission mode of abnormal genes is autosomal recessive, a person must have two copies of a recessive gene on the same chromosome to develop the disease. In general, blood characteristics are analysed during the course of disease diagnosis. A complete blood count (CBC) and haemoglobin typing are the primary screening tests for a laboratory diagnosis of thalassaemia. Nonetheless, there is still a limitation in the data analysis due to a large number of candidate blood characteristics. Moreover, there are many types of thalassaemia and thalassaemia trait. (Persons with thalassaemia trait do not have the disease but inherit genes the cause the disease.) As a result, manual diagnosis is needed to be carried out by trained professionals [3–6].

Early attempts to develop an automatic diagnostic tool involve CBC data analysis using image processing [7], statistical [8] and clustering techniques [9]. Later, the research interest has shifted

* Corresponding author at: Department of Electrical Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok, 1518 Piboolsongkram Road, Bangsue, Bangkok 10800, Thailand. Tel.: +66 2 9132500x8410; fax: +66 2 5856149.
E-mail addresses: n.chaiyaratana@gmail.com, nchl@kmutnb.ac.th (N. Chaiyaratana).

1746-8094/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bspc.2011.03.007

to the use of expert systems where both rule-based [10–12] and hybrid neural network/rule-based systems [13] have been successfully implemented for clinical trials. Nonetheless, these tools broadly differentiate between a wide range of blood disorders including various types of anaemia. As a result, many subsequent research works focus on the development of a diagnostic tool that only differentiates between thalassaemic patients, persons with thalassaemia traits and normal subjects. These works cover the implementation of a multilayer perceptron [14–16], a *k*-nearest neighbour technique [15], a support vector machine [15] and genetic programming [16] as thalassaemic diagnostic tools. Among these machine learning techniques, the multilayer perceptron emerges as the most suitable tool for the thalassaemia classification problem in Thailand [16] which covers higher varieties of haemoglobinopathies than other countries [1]. The multilayer perceptron is capable of handling a problem with 13 classes of thalassaemia abnormality and two classes of normal subjects with and without iron deficiency. However, the classification accuracy during the clinical trial that covers 300 samples is only 81.6% [16].

A significant improvement in thalassaemia classification accuracy has been achieved through the application of machine learning techniques in conjunction with haemoglobin typing inputs. The techniques that have been applied to the problem include a C4.5 decision tree, a random forest and a multilayer perceptron. The C4.5 decision tree is proven to be the most suitable technique where the classification accuracy during a clinical trial involving 1000 samples from 13 classes of thalassaemia abnormality and a normal subject class is 93.1% [17]. The improvement in classification accuracy can be achieved because CBC and haemoglobin typing data represents different aspects of blood characteristics. CBC information is useful for the diagnosis of various types of anaemia [10] while haemoglobin typing information can confirm the haemoglobinopathies [18]. Nonetheless, haemoglobin typing alone is insufficient for the classification between certain types of thalassaemia abnormality. For instance, haemoglobin typing characteristics cannot be used to differentiate between a person with α-thalassaemia 1 trait, a person with α-thalassaemia 2 trait and a normal subject [19].

Since both CBC and haemoglobin typing data are usually available for the laboratory diagnosis of thalassaemia, an attempt to develop an automated diagnostic tool that takes both forms of data should be carried out. Consequently, this should lead to an improvement of classification accuracy. In this article, the possibility of using CBC and haemoglobin typing data in thalassaemia classification by machine learning is investigated. The choices of machine learning technique include a multilayer perceptron, a C4.5 decision tree and a naïve Bayes classifier. The first two techniques are chosen because they are proven to be suitable in the early investigations involving CBC [16] and haemoglobin typing inputs [17], respectively. In contrast, a naïve Bayes classifier is selected due to its classification efficacy and implementation simplicity [20]. As a result, it is common to provide classification accuracy achieved by a naïve Bayes classifier as a comparison baseline [20,21].

With the availability of CBC and haemoglobin typing data and selected choices of machine learning technique, an investigation can be conducted as follows. Firstly, the data is pre-processed via input attribute discretisation. Informative attributes identified in the previous works [16,17] are discretised since it has been reported that a proper discretisation of continuous-valued attributes can significantly improve the classification accuracy of both multilayer perceptron and C4.5 decision tree [17]. The attributes are thus discretised via an information-theoretic technique proposed by Fayyad and Irani [22]. As a result, the data with discrete-valued attributes is available for classifier benchmarking. Since the information contained within the CBC attributes and necessary for the classification may overlap with that contained within the haemoglobin typing attributes, redundant attributes that can be removed without affecting the classification accuracy are also identified during classifier benchmarking. Correlation analysis via symmetrical uncertainty measurement [23] is subsequently performed on the attributes necessary for the classification. After the classifier performance evaluation is completed, the best classifier together with the pruned attribute set is chosen for a clinical trial involving a separate data set. This independent data set contains more samples than those in the early clinical trials by Wongseree et al. [16] and Piroonratana et al. [17]. Finally, classification analysis of the clinical trial results is carried out to determine the feasibility of the chosen classifier and pruned attribute set. Every step in the procedure described above is illustrated in Fig. 1 and is implemented using a WEKA package [24].

The organisation of this article is as follows. In Section 2, materials and methods are explained. These include the description of CBC and haemoglobin typing data, an information-theoretic attribute discretisation technique, symmetrical uncertainty, a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron. The results from attribute discretisation, classifier benchmarking together with redundant attribute identification and a clinical trial are then discussed in Section 3. Finally, the conclusions are given in Section 4.

2. Materials and methods

2.1. CBC and haemoglobin typing data sets

The data sets for thalassaemia classification consist of various blood characteristics obtained through CBC and haemoglobin typing. In this study, the data sets consist of eight input attributes. Details of these attributes are given in Table 1. The first two attributes are obtained through CBC: a haemoglobin concentration (HB) and a mean corpuscular volume (MCV). These two attributes are chosen since they are proven to be highly informative in the early studies [15,16]. The last six attributes are obtained through haemoglobin typing. Multiple haemoglobin typing attributes, characterised through the use of high performance liquid chromatography (HPLC) [25], are necessary for classification since a blood specimen generally contains more than one type of haemoglobin. As a result, many types of thalassaemia abnormality can be identified through the difference in proportion of haemoglobin contents [19,26,27]. The haemoglobin typing attributes are extracted from elution chromatograms [28]. Typical elution chromatograms of two different specimens are illustrated in Fig. 2. The first chromatogram shows that the specimen is mostly made up from Hb A_0. If the MCV value obtained via CBC is greater than 75 fL, this specimen is most likely to be taken from a normal person. Otherwise, this specimen must be obtained from a person with α-thalassaemia 1 trait. The second chromatogram indicates that the specimen consists of Hb A_0 and Hb E. This means that the specimen is taken from a person with Hb E trait. It is noticeable that some types of thalassaemia abnormality can be diagnosed via haemoglobin typing analysis alone while the diagnosis of other types requires both CBC and haemoglobin typing information. Fig. 2 also illustrates that different types of haemoglobin are detectable in the form of elution peaks at different retention times. Hence, a chromatogram can be divided into multiple sections where each section occupies a non-overlapping range of retention time. Consequently, each chromatogram section represents a unique attribute for the classification where the percentage of haemoglobin in the elution profile corresponds to the attribute value. The last six attributes and the associated types of haemoglobin are also summarised in Table 1. It is noticed that two attributes representing the elution profiles which occupy the retention time between 161 and 199 s

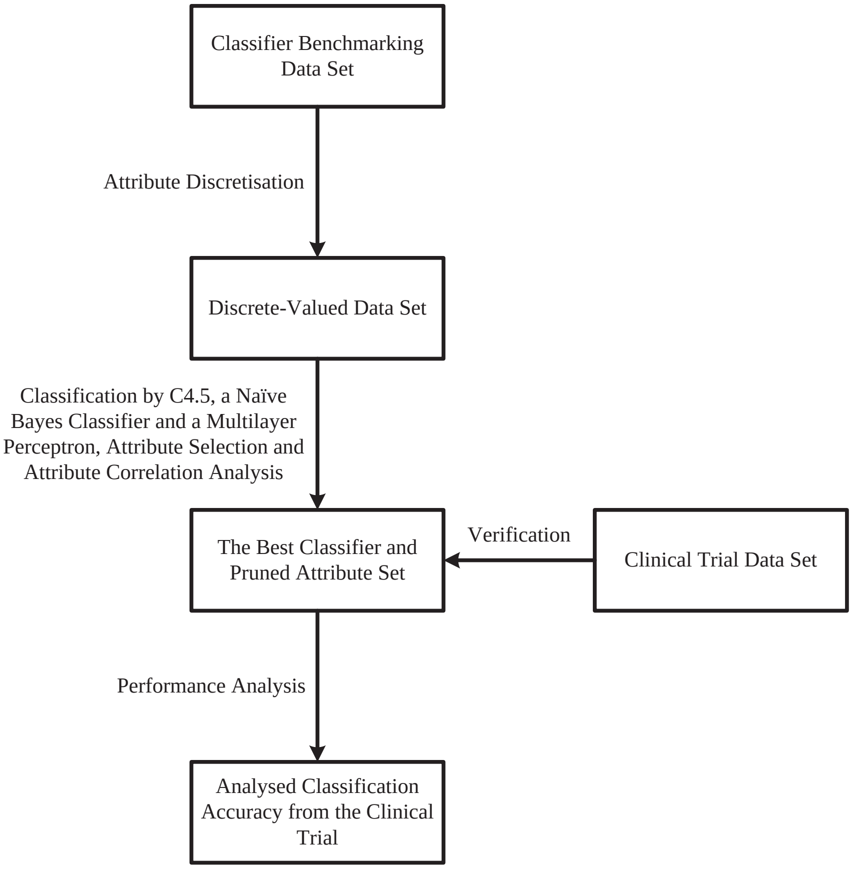

Classifier Benchmarking Data Set
Attribute Discretisation
Discrete-Valued Data Set
Classification by C4.5, a Naïve Bayes Classifier and a Multilayer Perceptron, Attribute Selection and Attribute Correlation Analysis
The Best Classifier and Pruned Attribute Set
Verification
Clinical Trial Data Set
Performance Analysis
Analysed Classification Accuracy from the Clinical Trial

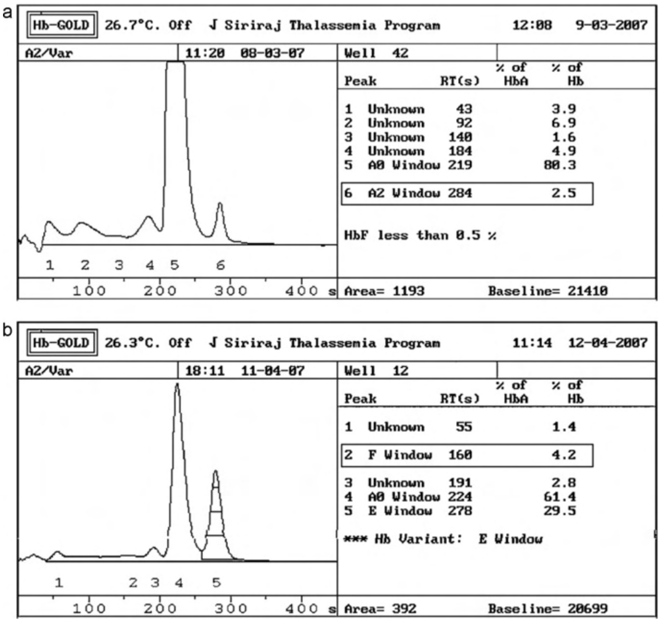
Fig. 1. Schematic diagram for the methodology employed in the study.

and between 231 and 249 s are not needed. These two attributes correspond to unknown types of haemoglobin and are proven to be uninformative for the classification task [17].

Two confirmed diagnosis data sets are acquired for this study. The first data set is created for the evaluation of classifier performance while the second set is used in a clinical trial. The data set for the classifier evaluation consists of 1402 samples which represent the majority of blood specimens from adults that need to be screened for thalassaemia. On the other hand, the data set for the clinical trial contains 8054 samples and is at least eight times larger than those from the previous studies [16,17]. The clinical trial data set represents a typical distribution of specimens which are submitted for screening during a fixed time period. Both data sets are

Table 1
Input features or attributes for thalassaemia classification. The first two attributes are CBC attributes while the last six attributes represent different types of haemoglobin.

Attribute	Attribute name	Description (measurement unit)
1	HB	Haemoglobin concentration (gram/decilitre, g/dL)
2	MCV	Mean corpuscular volume (femtolitre, fL)
3	Hb Bart's	Percentage of haemoglobin at retention time 0–68 s (%)
4	Hb A_{1C}/Hb F	Percentage of haemoglobin at retention time 69–160 s (%)
5	Hb A_0	Percentage of haemoglobin at retention time 200–230 s (%)
6	Hb E	Percentage of haemoglobin at retention time 250–280 s (%)
7	Hb A_2	Percentage of haemoglobin at retention time 281–289 s (%)
8	Hb D/Hb S/ Hb Constant Spring/ Hb C	Percentage of haemoglobin at retention time 290–320 s (%)



```
Hb-GOLD  26.7°C. Off  √ Siriraj Thalassemia Program      12:08   9-03-2007
A2/Var              11:20  08-03-07   Well  42
                                                            % of    % of
                                      Peak          RT(s)   HbA     Hb
                                      1  Unknown     43              3.9
                                      2  Unknown     92              6.9
                                      3  Unknown    140              1.6
                                      4  Unknown    184              4.9
                                      5  A0 Window  219             80.3
                                      6  A2 Window  284              2.5
                                      HbF less than 0.5 %
                                      Area= 1193          Baseline= 21410
```

```
Hb-GOLD  26.3°C. Off  √ Siriraj Thalassemia Program      11:14   12-04-2007
A2/Var              18:11  11-04-07   Well  12
                                                            % of    % of
                                      Peak          RT(s)   HbA     Hb
                                      1  Unknown     55              1.4
                                      2  F Window   160              4.2
                                      3  Unknown    191              2.8
                                      4  A0 Window  224             61.4
                                      5  E Window   278             29.5
                                      *** Hb Variant:  E Window
                                      Area= 392           Baseline= 20699
```

Fig. 2. Elution chromatograms of (a) either a normal specimen or a specimen from a person with α-thalassaemia 1 trait depending on the MCV value and (b) a specimen from a person with Hb E trait that are obtained from an Hb Gold HPLC system. RT(s) represents the retention time in seconds for each fraction of elute. % of Hb represents the percentage of haemoglobin in the elution peak.

Table 2
Two data sets for thalassaemia classification. The first set contains 1402 samples and is used for classifier benchmarking. The second set consists of 8054 samples which are collected for a clinical trial.

Class	Description	Category	Number of samples	
			Classifier benchmarking	Clinical trial
1	Normal subject	Normal	78	436
2	α-Thalassaemia 2 trait	Trait	150	978
3	Hb Constant Spring trait	Trait	23	10
4	Hb E trait	Trait	548	4292
5	Hb E trait + Hb Constrant Spring trait	Trait	6	1
6	Hb E trait + abnormal haemoglobin	Trait	5	0
7	α-Thalassaemia 1 trait	Trait	180	1008
8	α-Thalassaemia 1 trait + Hb E trait	Trait	116	431
9	Homozygous Hb E	Trait	105	571
10	β-Thalassaemia trait	Trait	74	267
11	HPFH	Trait	9	1
12	Abnormal haemoglobin	N/A	11	1
13	Hb H disease	Disease	39	27
14	Hb H-Constant Spring disease	Disease	5	0
15	Homozygous Hb Constant Spring	Disease	8	0
16	EA Bart's disease	Disease	11	1
17	β^0-Thalassaemia/Hb E	Disease	15	22
18	β^+-Thalassaemia/Hb E	Disease	9	3
19	Homozygous β-thalassaemia	Disease	10	5
	Total		1402	8054

collected from Siriraj Hospital, Bangkok, Thailand during 1 January 2007 and 31 December 2008. The data acquisition has been conducted in accordance with the Faculty of Medicine Siriraj Hospital Ethics Committee's guideline and in accordance with the Helsinki Declaration. In addition, informed consent has been obtained from all individuals. The description of both data sets is summarised in Table 2. The samples are made up from seven groups of thalassaemic patients, ten groups of persons with thalassaemia trait, one group of persons with abnormal haemoglobin and one group of normal subjects. Some types of thalassaemia abnormality in the data set for classifier benchmarking are not presented in the specimens collected for the clinical trial. In other words, the classifier benchmarking data set covers more types of thalassaemia abnormality than those presented in the data set for clinical trial. This is carried out to increase the possibility of using the chosen classifier in other data sets without or with minimal additional classifier training.

2.2. Attribute discretisation

The thalassaemia data sets are pre-processed by means of attribute discretisation. The discretisation technique selected for this study is developed by Fayyad and Irani [22]. The technique, which has been successfully applied in an early investigation into thalassaemia classification [17], is an information-theoretic technique that employs entropy-based splitting and minimum description length stopping criteria. A chosen cut point within the range of each attribute value is ensured to lie at a boundary between two classes. A new cut point is introduced recursively to each sample subset and is accepted if a significant information gain—the difference between the information values before and after the split—is achieved. The class entropy of a sample set S, which consists of samples from $|C|$ classes, is defined as

$$Ent(S) = -\sum_{c \in C} p(c)\ \log_2 p(c) \tag{1}$$

where p denotes the probability and c is a class. A cut point T, which is performed on an attribute A of the sample set S, creates a partition that has the class information entropy

$$E(A, T, S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2) \tag{2}$$

where S_1 and S_2 are the sample subsets of S and $S_1 + S_2 = S$. The cut point T is accepted according to the minimum description length stopping criterion if and only if

$$Gain(A, T, S) > \frac{\log_2(|S| - 1)}{|S|} + \frac{\log_2(3^{|C|} - 2) - [|C|Ent(S) - |C_1|Ent(S_1) - |C_2|Ent(S_2)]}{|S|} \tag{3}$$

where $|C_1|$ and $|C_2|$ are the numbers of classes in the subsets S_1 and S_2, respectively and $Gain(A, T, S)$ is the information gain of the cut point, which is defined as

$$Gain(A, T, S) = Ent(S) - \frac{|S_1|}{|S|} Ent(S_1) - \frac{|S_2|}{|S|} Ent(S_2). \tag{4}$$

2.3. Symmetrical uncertainty

Symmetrical uncertainty is an information-theoretic measure discussed by Press et al. [23] that defines a correlation between two discrete-valued variables. Consider a sample set in which each sample is described by m discrete-valued attributes $A_1, \ldots, A_m$. The entropy of an attribute A_i before and after observing an attribute A_j is respectively given by

$$H(A_i) = -\sum_{a_i \in A_i} p(a_i)\ \log_2 p(a_i) \tag{5}$$

and

$$H(A_i|A_j) = -\sum_{a_j \in A_j} p(a_j) \sum_{a_i \in A_i} p(a_i|a_j) \log_2 p(a_i|a_j) \tag{6}$$

where a_i is a value of the attribute A_i and a_j is a value of the attribute A_j. The degree of correlation between attributes A_i and A_j can be estimated via symmetrical uncertainty (SU) which is defined by

$$\begin{aligned} SU &= 2 \times \left[\frac{H(A_i) - H(A_i|A_j)}{H(A_i) + H(A_j)}\right] \\ &= 2 \times \left[\frac{H(A_j) - H(A_j|A_i)}{H(A_i) + H(A_j)}\right] \\ &= 2 \times \left[\frac{H(A_i) + H(A_j) - H(A_i, A_j)}{H(A_i) + H(A_j)}\right] \end{aligned}. \tag{7}$$

The value range of symmetrical uncertainty is [0,1]. An *SU* value close to zero indicates a weak correlation while an *SU* value close to one indicates a strong correlation [29].

2.4. C4.5 decision tree

A C4.5 decision tree is one of the most widely used inductive inference tools [30]. The tree is generally constructed in a top-down manner. The construction begins at the root node where each attribute is evaluated using a statistical test to determine how well it can classify the training samples. The best attribute is chosen as the test at the root node of the tree. A descendant of the root node is then created for either each possible value of this attribute if it is a discrete-valued attribute or each possible discretised interval of this attribute if it is a continuous-valued attribute. Next, the training samples are sorted to the appropriate descendant node. The process is repeated using the training samples associated with each descendant node to select the best attribute for testing at that point in the tree. This forms a greedy search for a decision tree, in which the algorithm never backtracks to reconsider earlier node choices. Although it is possible to add a new node to the tree until all samples that are assigned to one node belong to the same class, the tree is not allowed to grow to its maximum depth. A node is only introduced to the tree only when there are a sufficient number of samples left from sorting. After the complete tree is constructed, a tree pruning is usually carried out to avoid data over-fitting.

A statistical test used in C4.5 for assigning an attribute to each node in the tree also employs an entropy-based measure. The assigned attribute is the one with the highest information gain ratio among attributes available at that tree construction point. The information gain ratio *GainRatio*(*A*, *S*) of an attribute *A* relative to the sample set *S* is defined as

$$GainRatio(A, S) = \frac{Gain(A, S)}{SplitInformation(A, S)} \quad (8)$$

where

$$Gain(A, S) = Ent(S) - \sum_{a \in A} \frac{|S_a|}{|S|} Ent(S_a) \quad (9)$$

and

$$SplitInformation(A, S) = -\sum_{a \in A} \frac{|S_a|}{|S|} \log_2 \frac{|S_a|}{|S|}. \quad (10)$$

S_a is the subset of S for which the attribute A has the value a. Obviously, the information gain ratio can be calculated straightaway for discrete-valued attributes. In contrast, continuous-valued attributes are needed to be discretised prior to the information gain ratio calculation.

2.5. Naïve Bayes classifier

A naïve Bayes classifier is a classification system in which the class prediction is based on the application of Bayes theorem [31]. Consider a set of training samples where each sample is made up from m discrete-valued attributes and a class from a finite set C. The naïve Bayes classifier can probabilistically predict the class of an unknown sample using the available training sample set to calculate the most probable output. The most probable class c_{NB} of an unknown sample with the conjunction $a_1, a_2, \ldots, a_m$ is given by

$$c_{NB} = \arg\max_{c \in C} p(c|a_1, a_2, \ldots, a_m). \quad (11)$$

With the use of Bayes theorem, this expression can be rewritten as

$$\begin{aligned} c_{NB} &= \arg\max_{c \in C} \frac{p(a_1, a_2, \ldots, a_m|c)p(c)}{p(a_1, a_2, \ldots, a_m)} \\ &= \arg\max_{c \in C} p(a_1, a_2, \ldots, a_m|c)p(c) \end{aligned}. \quad (12)$$

The naïve Bayes classifier functions by assuming that the attributes are conditionally independent given the class. In other words, given the class of the sample the probability of observing the conjunction $a_1, a_2, \ldots, a_m$ is the product of the probability of observing each attribute:

$$p(a_1, a_2, \ldots, a_m|c) = \prod_i p(a_i|c). \quad (13)$$

Substituting Eq. (13) into Eq. (12), the most probable class as predicted by the naïve Bayes classifier is

$$c_{NB} = \arg\max_{c \in C} p(c) \prod_i p(a_i|c). \quad (14)$$

A Laplace estimate [32] is used to calculate $p(a_i|c)$, that is

$$p(a_i|c) = \frac{|S_{a_i|c}| + 1}{|S_c| + |A_i|} \quad (15)$$

where $|S_{a_i|c}|$ is the number of samples from the class c in which the ith attribute (A_i) has the value a_i, $|S_c|$ is the number of samples from the class c and $|A_i|$ is the number of possible values for the attribute A_i.

2.6. Multilayer perceptron

A neural network is an interconnected group of artificial neurons that uses a computational model for information processing. The neural network selected for this study is a multilayer perceptron [33]. The model of a neuron shown in Fig. 3(a) indicates that q input signals are received by the neuron. These inputs are weighted and summed together. The threshold, which is treated as an extra connection weight, is then applied to the weighted-sum result. Thus, the linear combiner output (z) or input to the activation function is given by

$$z = \sum_i w_i u_i \quad (16)$$

where u_i is the ith input to the neuron and w_i is the connection weight for the for the input u_i. In addition, $u_0 = -1$ and w_0 is the threshold. The neuron output ($h(z)$) is the output from the activation function and is denoted by

$$h(z) = \frac{1}{1 + \exp(-z)}. \quad (17)$$

As a result, the output signal from each neuron is limited by a logistic sigmoid function. The neuron model described above is used throughout the multilayer feed-forward network illustrated in Fig. 3(b). The multilayer perceptron is implemented in this manner because the early studies indicate that a multilayer perceptron of this form works well with both CBC [16] and haemoglobin typing inputs [17].

Since the multilayer perceptron is used as a classifier, the number of network inputs is equal to the number of attributes while the number of network outputs is equal to the number of classes. Each output neuron represents one of possible classes with the highest-valued output taken as the network prediction. This is often referred to as a 1-of-n output encoding technique [31]. In this study, the network has a single hidden layer, which is a layer of neurons that receive attributes as inputs and send signals to output neurons. The number of neurons in the hidden layer is a design parameter

Please cite this article in press as: D. Setsirichok, et al., Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening, Biomed. Signal Process. Control (2011), doi:10.1016/j.bspc.2011.03.007

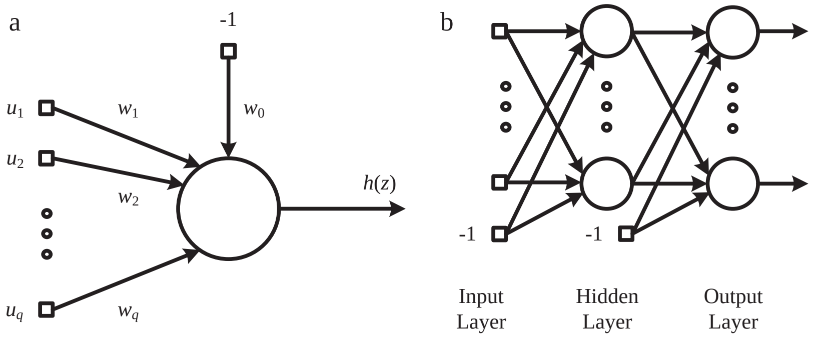


Fig. 3. Schematic diagram of a multilayer perceptron: (a) computational model of a neuron and (b) feed-forward network with one hidden layer.

and can generally be set using a heuristic rule. The number of hidden nodes is set to $\lfloor(\text{number of attributes} + \text{number of classes})/2\rfloor$ in this study [24].

The multilayer perceptron is trained by a back-propagation algorithm. The algorithm employs a sample-by-sample updating rule for adjusting connection weights. A training sample is presented to the network during the iteration. The signal is fed in a forward manner through the network until the network output is obtained. The error between actual and target network outputs is then calculated and used to adjust the connection weights. The adjustment procedure, which is based on a gradient descent method, is first applied to connection weights in the output layer. Next, connection weights in the hidden layer are adjusted. The iteration is completed when all connection weights have been adjusted.

Table 3
Discrete intervals of the attributes.

Attribute	Attribute name	Interval	Range
1	HB	1	HB $\leq$ 8.65
		2	8.65 < HB $\leq$ 10.65
		3	10.65 < HB $\leq$ 14.55
		4	HB > 14.55
2	MCV	1	MCV $\leq$ 56.45
		2	56.45 < MCV $\leq$ 61.35
		3	61.35 < MCV $\leq$ 64.05
		4	64.05 < MCV $\leq$ 69.95
		5	69.95 < MCV $\leq$ 72.15
		6	72.15 < MCV $\leq$ 74.95
		7	74.95 < MCV $\leq$ 78.75
		8	78.75 < MCV $\leq$ 79.95
		9	79.95 < MCV $\leq$ 84.05
		10	84.05 < MCV $\leq$ 87.45
		11	MCV > 87.45
3	Hb Bart's	1	% of Hb $\leq$ 1.70
		2	% of Hb > 1.70
4	Hb A_{1C}/Hb F	1	% of Hb $\leq$ 0.58
		2	0.58 < % of Hb $\leq$ 3.35
		3	3.35 < % of Hb $\leq$ 10.85
		4	10.85 < % of Hb $\leq$ 74.35
		5	74.35 < % of Hb $\leq$ 83.60
		6	% of Hb > 83.60
5	Hb A_0	1	% of Hb $\leq$ 1.05
		2	1.05 < % of Hb $\leq$ 49.10
		3	49.10 < % of Hb $\leq$ 52.75
		4	52.75 < % of Hb $\leq$ 62.85
		5	62.85 < % of Hb $\leq$ 65.15
		6	65.15 < % of Hb $\leq$ 67.65
		7	67.65 < % of Hb $\leq$ 72.35
		8	72.35 < % of Hb $\leq$ 77.05
		9	77.05 < % of Hb $\leq$ 83.25
		10	83.25 < % of Hb $\leq$ 83.35
		11	% of Hb > 83.35
6	Hb E	1	% of Hb $\leq$ 7.70
		2	7.70 < % of Hb $\leq$ 18.65
		3	18.65 < % of Hb $\leq$ 24.95
		4	24.95 < % of Hb $\leq$ 40.75
		5	40.75 < % of Hb $\leq$ 59.80
		6	59.80 < % of Hb $\leq$ 75.80
		7	% of Hb > 75.80
7	Hb A_2	1	% of Hb $\leq$ 0.05
		2	0.05 < % of Hb $\leq$ 1.05
		3	1.05 < % of Hb $\leq$ 2.15
		4	2.15 < % of Hb $\leq$ 2.55
		5	2.55 < % of Hb $\leq$ 3.55
		6	3.55 < % of Hb $\leq$ 11.40
		7	11.40 < % of Hb $\leq$ 20.10
		8	% of Hb > 20.10
8	Hb D/Hb S/Hb Constant Spring/Hb C	1	% of Hb $\leq$ 0.05
		2	% of Hb > 0.05

3. Results and discussions

3.1. Attribute discretisation

The original CBC and haemoglobin typing data contains eight attributes as given in Table 1. The attributes in classifier evaluation samples are discretised using the information-theoretic technique [22] described in Section 2.2. The discrete intervals of each attribute are illustrated in Table 3. The discrete intervals for attributes 3, 4, 6 and 8, which are haemoglobin typing attributes, are similar to those reported in Piroonratana et al. [17] while more discrete intervals for attributes 5 and 7 are introduced. This implies that the discrete intervals for some attributes from the previous study by Piroonratana et al. [17] are sufficient for the classification task in the current study. However, new discrete intervals for the other attributes are also required to accommodate more types of thalassaemia. Furthermore, a cut point at MCV = 74.95 fL is introduced to attribute 2. This conforms to the expert rule explained earlier for the differentiation between normal subjects and persons with α-thalassaemia 1 trait.

3.2. Classifier performance evaluation

Three candidate classifiers—a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron—are benchmarked using the classifier evaluation data in stratified 10-fold cross-validation experiments [34]. The training of a multilayer perceptron is terminated prior to the occurrence of data over-fitting. The suitable number of training epochs is chosen using a training and validation approach [31]. Basically, a validation data set is employed to detect the training epoch where the training error continues to decrease while the validation error begins to increase. This is the point where data over-fitting occurs. In this study, 10% of data samples are used as the validation samples. In contrast to a multilayer perceptron,

Table 4
Summarised classification performance of the C4.5 decision tree, naïve Bayes classifier and multilayer perceptron. The results are averaged over 30 runs of stratified 10-fold cross-validation involving 1402 samples. The numbers in the brackets are standard deviations.

Attribute set	Accuracy (%)		
	C4.5 decision tree	Naïve Bayes classifier	Multilayer perceptron
Complete	89.88 (1.74)	92.77 (1.74)	92.34 (1.69)
Without HB	89.30 (1.82)	93.23 (1.67)	92.60 (1.75)
Without MCV	79.80 (2.58)	80.49 (2.34)	77.04 (2.65)
Without both HB and MCV	77.98 (2.23)	79.27 (2.13)	76.31 (2.44)

C4.5 has a built-in mechanism for data over-fitting avoidance. This is achieved via a rule post-pruning strategy [30].

The classification performance of C4.5, a naïve Bayes classifier and a multilayer perceptron on the classifier evaluation data obtained from stratified 10-fold cross-validation is summarised in Table 4. The results from each classifier are tallied from 30 runs where new cross-validation folds are generated for each run. The results obtained by excluding CBC attributes from classifier inputs are also given for comparison purposes. The exclusion of haemoglobin typing attributes from classifier inputs lead to unacceptably low classification accuracy; these results are not shown. It is noticeable that both naïve Bayes classifier and multilayer perceptron have the highest classification accuracy. Moreover, the use of six haemoglobin typing attributes in conjunction with the MCV attribute as classifier inputs is proven to be sufficient while the HB attribute appears to be a redundant attribute. The symmetrical uncertainty analysis of seven attributes necessary for the classification in Table 5 reveals that only the Hb A_0, Hb E and Hb A_2 attributes are moderately correlated among one another while the remaining attributes are uncorrelated.

Among three classifiers, it can be clearly seen that both naïve Bayes classifier and multilayer perceptron have higher classification performance than C4.5 when all attributes are used and when only HB is not used as an input (t-test's p-value < 0.0001). Moreover, the performance difference between naïve Bayes classifier and multilayer perceptron in both cases is statistically insignificant ($p > 0.05$). However, when both HB and MCV are excluded from classifier inputs the classification accuracy of C4.5 is significantly better than that of the multilayer perceptron ($p < 0.01$). In the early work by Wongseree et al. [16], which involves the application of a multilayer perceptron and a genetic programming based decision tree to the classification problem with CBC attributes, the multilayer perceptron is proven to be the best classifier. In contrast, C4.5 is proven to be the best classifier in comparison to a multilayer perceptron and a random forest in the classification problem with haemoglobin typing attributes [17]. Based on the evidence from the early and present studies, it is possible to deduce that C4.5 is more suitable to the problem when only haemoglobin typing attributes are considered as inputs while a multilayer perceptron is more suitable to the problem that involves CBC attributes.

Table 5
A symmetrical uncertainty (SU) analysis of six haemoglobin typing attributes and the MCV attribute. An *SU* value close to zero denotes a weak correlation while an *SU* value close to one denotes a strong correlation. The description of each attribute has been given in Table 1.

Attribute	2	3	4	5	6	7	8
2	1.0000	0.0202	0.0196	0.0992	0.1360	0.0430	0.0031
3		1.0000	0.0043	0.0298	0.0543	0.0713	0.0606
4			1.0000	0.0921	0.0870	0.0473	0.0053
5				1.0000	0.4960	0.3130	0.0157
6					1.0000	0.4950	0.0231
7						1.0000	0.0202
8							1.0000

As mentioned earlier, it is hypothesised that using both CBC and haemoglobin typing attributes as inputs to the classifiers should lead to an improvement in the classification accuracy. The classification performance of each classifier clearly supports this hypothesis where the difference between the classification accuracy obtained using all attributes and that using only haemoglobin typing attributes is statistically significant ($p < 0.0001$). Nonetheless, MCV is the only necessary CBC attribute for the classification. This is deduced from the results which clearly indicate that the difference between the classification accuracy obtained using all attributes and that with the exclusion of HB attribute is not statistically significant ($p > 0.05$). In addition, the removal of MCV from the classifier inputs also leads to a significant degradation of classification performance ($p < 0.0001$). Further analysis of the classification accuracy reveals that the increase in classification error mainly stems from an inability to differentiate between normal, α-thalassaemia 1 trait and α-thalassaemia 2 trait samples. This conforms to the early explanation regarding the limitation of using haemoglobin typing attributes as the sole inputs for the classification task involving these three classes [17,19]. Although the early studies by Amendolia et al. [15] and Wongseree et al. [16] suggest that both HB and MCV are informative CBC attributes, it is most likely that the use of haemoglobin typing attributes is sufficient to satisfy the need for using the HB attribute.

It is observed that the redundancy of the HB attribute can only be identified via the inspection of the variation in classification accuracy after changing the attribute combination, which is a simple form of wrapper attribute selection [35]. An attempt to perform attribute selection by other techniques including a correlation-based feature selection technique [20] and a ReliefF technique [36] fails to identify this redundancy. These techniques are considered because the attributes selected by the techniques are not transformed, which makes the clinical interpretation of results a straightforward task. Moreover, the correlation-based feature selection technique is proven to be capable of identifying informative haemoglobin typing attributes in the early study by Piroonratana et al. [17].

Detailed classification performance of all three classifiers obtained without the HB input is subsequently analysed. The results from an example run of the stratified 10-fold cross-validation given in Tables 6–8 indicate that the classification accuracy of each classifier for the classes with a small number of samples is low. Nonetheless, most of the misclassified samples from these classes are identified as samples from classes which are closely related to the true classes. For instance, mixed Hb E and Hb Constant Spring trait samples as well as mixed Hb E and abnormal haemoglobin samples are classified as Hb E trait, mixed α-thalassaemia 1 and Hb E trait and homozygous Hb E samples. Similarly, samples from persons with Hb H-Constant Spring disease are misclassified as homozygous Hb Constant Spring samples and vice versa. The results from other runs and from stratified 10-fold cross-validation with other settings for attribute combination also have a similar trend and hence are not shown.

Based on the above discussion of classification results, it can be concluded that the necessary attributes for this study are six haemoglobin typing attributes and the MCV attribute while the suitable classifiers for use with these attributes are a naïve Bayes classifier and a multilayer perceptron. Moreover, a multilayer perceptron is equally suitable to a naïve Bayes classifier in terms of the actual implementation for clinical trials. This is because the storage of trained connection weights for a multilayer perceptron requires a similar amount of space to that of probability values for a naïve Bayes classifier. In lieu of these reasons, a naïve Bayes classifier and a multilayer perceptron are the chosen classifiers for a clinical trial.

Please cite this article in press as: D. Setsirichok, et al., Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening, Biomed. Signal Process. Control (2011), doi:10.1016/j.bspc.2011.03.007

Table 6
Detailed classification performance of the C4.5 decision tree from one run of stratified 10-fold cross-validation without the HB attribute. The classification accuracy is 89.90%, which sufficiently represents the average accuracy of 89.30% (standard deviation = 1.82%) obtained from 30 runs. The description of each class has been given in Table 2.

Actual class	Identified class (number of samples)																			Class acc. (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
1	78																			100.00
2		146	1				3													97.33
3	1	4	13				5													56.52
4	1	3		538			1	4	1											98.18
5				2	1			3												16.67
6				3				1									1			0.00
7		3		2			169	1		2			3							93.89
8					1			115												99.14
9									103								2			98.10
10	8	7					12			44	3									59.46
11	3	1					1			1	1	2								11.11
12	1	2					4	1		1	1							1		0.00
13	1	2					12						24							61.54
14			2				2						1							0.00
15	1		5										2							0.00
16						1	1	1								8				72.73
17				3					1								11			73.33
18								2										7		77.78
19	4	1								4									1	10.00

Table 7
Detailed classification performance of the naïve Bayes classifier from one run of stratified 10-fold cross-validation without the HB attribute. The classification accuracy is 93.37%, which sufficiently represents the average accuracy of 93.23% (standard deviation = 1.67%) obtained from 30 runs. The description of each class has been given in Table 2.

Actual class	Identified class (number of samples)																			Class acc. (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
1	74									2	2									94.87
2		145					3			1			1							96.67
3		2	21																	91.30
4	1	1		540				3	1			2								98.54
5				2				4												0.00
6				2				1	2											0.00
7		3		1			169	1		4			2							93.89
8				6				110												94.83
9									104								1			99.05
10	3	2					1			63	5									85.14
11	3						1			1	4									44.44
12	1	1		2			2	1			2						1	1		0.00
13			1				5						33							84.62
14													1	2	2					40.00
15			3											2	3					37.50
16								1								10				90.91
17									1								14			93.33
18								2										7		77.78
19																			10	100.00

Table 8
Detailed classification performance of the multilayer perceptron from one run of stratified 10-fold cross-validation without the HB attribute. The classification accuracy is 92.65%, which sufficiently represents the average accuracy of 92.60% (standard deviation = 1.75%) obtained from 30 runs. The description of each class has been given in Table 2.

Actual class	Identified class (number of samples)																			Class acc. (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
1	73			1						3	1									93.59
2		142	1				4			1		1	1							94.67
3		1	20												2					86.96
4	1			538		1		4	1	1		2								98.18
5			1	1	2			2												33.33
6				2				1	1			1								0.00
7		3		2			166	2		2	1	1	3							92.22
8				1	2			111								1		1		95.69
9									103								2			98.10
10	1	2					3			63	5									85.14
11										4	3	2								33.33
12	1	1		1			2	1		1	1	1					1	1		9.09
13		1					3						34		1					87.18
14		1											2		2					0.00
15			4												4					50.00
16								1								10				90.91
17									2			1					12			80.00
18								1										8		88.89
19										1									9	90.00

Table 9
Detailed classification performance of the naïve Bayes classifier from the clinical trial involving 8054 samples. The classification accuracy is 99.39%. The description of each class has been given in Table 2.

Actual class	Identified class (number of samples)																			Class acc. (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
1	436																			100.00
2		977					1													99.90
3			10																	100.00
4				4291					1											99.98
5				1																0.00
6																				
7							1008													100.00
8				37				394												91.42
9									570								1			99.82
10		7								260										97.38
11											1									100.00
12																	1			0.00
13													27							100.00
14																				
15																				
16																1				100.00
17																	22			100.00
18																		3		100.00
19																			5	100.00

Table 10
Detailed classification performance of the multilayer perceptron from the clinical trial involving 8054 samples. The classification accuracy is 99.71%. The description of each class has been given in Table 2.

Actual class	Identified class (number of samples)																			Class acc. (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
1	436																			100.00
2		976					1			1										99.80
3			10																	100.00
4				4289		2			1											99.93
5		1																		0.00
6																				
7	1						1007													99.90
8				13				418												96.98
9									571											100.00
10										267										100.00
11											1									100.00
12												1								100.00
13													27							100.00
14																				
15																				
16																1				100.00
17									3								19			86.36
18																		3		100.00
19																			5	100.00

3.3. Clinical trial

In the previous sub-section, both naïve Bayes classifier and multilayer perceptron with six haemoglobin attributes and the MCV attribute are proven to be the best approach for thalassaemia classification. The naïve Bayes classifier and the multilayer perceptron are subsequently used in a clinical trial involving 8054 samples. The distribution of classes within the clinical trial data set has been given in Table 2. A naïve Bayes classifier and a multilayer perceptron, which are trained with classifier evaluation samples, are applied to the clinical trial data set where the classification accuracy of 99.39% and 99.71% are respectively achieved. Detailed classification performance of both classifiers is given in Tables 9 and 10. It is noticeable that nearly all misclassified samples from both classifiers stem from mixed α-thalassaemia 1 and Hb E trait samples. They are classified as Hb E trait samples, which belong to the class that are closely related to the true class. This indicates that both naïve Bayes classifier and multilayer perceptron are highly suitable to the present classification problem.

4. Conclusions

In this article, a thalassaemia classification problem in Thailand is investigated. The aim is to identify whether the human subject is a person with abnormal haemoglobin, a person with thalassaemia trait, a thalassaemic patient or a normal person using complete blood count (CBC) and haemoglobin typing data. The data sets contain eight attributes and 19 classes. The first two attributes are CBC attributes: a haemoglobin concentration (HB) and a mean corpuscular volume (MCV). On the other hand, the last six attributes reflect the percentages of haemoglobin at a specific range of retention time. In other words, these attributes represent different types of haemoglobin. The study is divided into two main parts: classifier and attribute subset selection and a clinical trial. Candidate classifiers for the task include a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron. The experiment involving stratified 10-fold cross-validation reveals that both naïve Bayes classifier and multilayer perceptron are the most suitable classifier for the data that has been pre-processed by attribute discretisation. The experiment results also suggest that using both CBC and haemoglobin typing attributes as classifier inputs can significantly improve the classification accuracy over that achieved using only haemoglobin typing attributes. Furthermore, the experiment results indicate that HB is a redundant attribute for this study. The naïve Bayes classifier and the multilayer perceptron are subsequently applied to an additional data set in a clinical trial. The analysis of classification errors from the trial indicates that most of the misclassified samples are identified as samples from classes which are closely related to the true classes. This helps emphasise the suitability of a naïve Bayes classifier and a multilayer perceptron as an automated thalassaemic classification tool.

Acknowledgements

The authors are extremely grateful to two anonymous reviewers and Prof. Robert Allen for their valuable comments and suggestions, which have contributed a lot towards improving the content and presentation of this article. The authors acknowledge Dr. Sissades Tongsima for his assistance on capturing chromatogram images from the Hb Gold HPLC system. Damrongrit Setsirichok, Theera Piroonratana and Waranyu Wongseree were supported by the Thailand Research Fund (TRF) through the Royal Golden Jubilee Ph.D. Programme (Grant Nos. PHD/1.E.KN.51/A.1, PHD/1.E.KN.50/A.1 and PHD/1.E.KN.49/A.1, respectively). Touchpong Usavanarong was supported by the Faculty of Engineering

Please cite this article in press as: D. Setsirichok, et al., Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening, Biomed. Signal Process. Control (2011), doi:10.1016/j.bspc.2011.03.007

of the King Mongkut's University of Technology North Bangkok. Chanin Limwongse was supported by the Mahidol Research Grant. Nachol Chaiyaratana was supported by the Thailand Research Fund, Office of the Higher Education Commission and Faculty of Engineering of the King Mongkut's University of Technology North Bangkok.

References

[1] S. Fucharoen, P. Winichagoon, Hemoglobinopathies in Southeast Asia: molecular biology and clinical medicine, Hemoglobin 21 (1997) 299–319.
[2] D.J. Weatherall, J.B. Clegg, The Thalassemia Syndromes, fourth ed., Blackwell Science, Malden, MA, 2001.
[3] C.V. Jimenez, J. Minchinela, J. Ros, New indices from the H*2 analyser improve differentiation between heterozygous beta or delta beta thalassaemia and iron-deficiency anaemia, Clinical and Laboratory Haematology 17 (1995) 151–155.
[4] A. Demir, N. Yarali, T. Fisgin, F. Duru, A. Kara, Most reliable indices in differentiation between thalassemia trait and iron deficiency anemia, Pediatrics International 44 (2002) 612–616.
[5] G. Ntaios, A. Chatzinikolaou, Z. Saouli, F. Girtovitis, M. Tsapanidou, G. Kaiafa, Z. Kontoninas, A. Nikolaidou, C. Savopoulos, I. Pidonia, S. Alexiou-Daniel, Discrimination indices as screening tests for β-thalassemic trait, Annals of Hematology 86 (2007) 487–491.
[6] O. Sripichai, W. Makarasara, T. Munkongdee, C. Kumkhaek, I. Nuchprayoon, A. Chuansumrit, S. Chuncharunee, N. Chantrakoon, P. Boonmongkol, P. Winichagoon, S. Fucharoen, A scoring system for the classification of β-thalassemia/Hb E disease severity, American Journal of Hematology 83 (2008) 482–484.
[7] P.R. Lund, R.D. Barnes, Automated classification of anaemia using image analysis, The Lancet 300 (1972) 463–464.
[8] R.L. Engle, B.J. Flehinger, S. Allen, R. Friedman, M. Lipkin, B.J. Davis, L.L. Leveridge, HEME: a computer aid to diagnosis of hematologic disease, Bulletin of the New York Academy of Medicine 52 (1976) 584–600.
[9] G. Barosi, M. Cazzola, C. Berzuini, S. Quaglini, M. Stefanelli, Classification of anemia on the basis of ferrokinetic parameters, British Journal of Haematology 61 (1985) 357–370.
[10] S. Quaglini, M. Stefanelli, G. Barosi, A. Berzuini, ANEMIA: an expert consultation system, Computers and Biomedical Research 19 (1986) 13–27.
[11] S. Quaglini, M. Stefanelli, G. Barosi, A. Berzuini, A performance evaluation of the expert system ANEMIA, Computers and Biomedical Research 21 (1988) 307–323.
[12] G. Lanzola, M. Stefanelli, G. Barosi, L. Magnani, NEOANEMIA: a knowledge-based system emulating diagnostic reasoning, Computers and Biomedical Research 23 (1990) 560–582.
[13] N.I. Birndorf, J.O. Pentecost, J.R. Coakley, K.A. Spackman, An expert system to diagnose anemia and report results directly on hematology forms, Computers and Biomedical Research 29 (1996) 16–26.
[14] S.R. Amendolia, A. Brunetti, P. Carta, G. Cossu, M.L. Ganadu, B. Golosio, G.M. Mura, M.G. Pirastru, A real-time classification system of thalassemic pathologies based on artificial neural networks, Medical Decision Making 22 (2002) 18–26.
[15] S.R. Amendolia, G. Cossu, M.L. Ganadu, B. Golosio, G.L. Masala, G.M. Mura, A comparative study of k-nearest neighbour, support vector machine and multi-layer perceptron for thalassemia screening, Chemometrics and Intelligent Laboratory Systems 69 (2003) 13–20.
[16] W. Wongseree, N. Chaiyaratana, K. Vichittumaros, P. Winichagoon, S. Fucharoen, Thalassaemia classification by neural networks and genetic programming, Information Sciences 177 (2007) 771–786.
[17] T. Piroonratana, W. Wongseree, A. Assawamakin, N. Paulkhaolarn, C. Kanjanakorn, M. Sirikong, W. Thongnoppakhun, C. Limwongse, N. Chaiyaratana, Classification of haemoglobin typing chromatograms by neural networks and decision trees for thalassaemia screening, Chemometrics and Intelligent Laboratory Systems 99 (2009) 101–110.
[18] S. Sirichotiyakul, R. Saetung, T. Sanguansermsri, Prenatal diagnosis of beta-thalassemia/Hb E by hemoglobin typing compared to DNA analysis, Hemoglobin 33 (2009) 17–23.
[19] J.M. Old, Screening and genetic diagnosis of haemoglobin disorders, Blood Reviews 17 (2003) 43–53.
[20] M.A. Hall, G. Holmes, Benchmarking attribute selection techniques for discrete class data mining, IEEE Transactions on Knowledge and Data Engineering 15 (2003) 1437–1447.
[21] K. Polat, S. Günes, Automated identification of diseases related to lymph system from lymphography data using artificial immune recognition system with fuzzy resource allocation mechanism (fuzzy-AIRS), Biomedical Signal Processing and Control 1 (2006) 253–260.
[22] U.M. Fayyad, K.B. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, in: R. Bajcsy (Ed.), Proceedings of the 13th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, San Mateo, CA, 1993, pp. 1022–1027.
[23] W.H. Press, B.P. Flannery, S.A. Teukolski, W.T. Vetterling, Numerical Recipes in C: The Art of Scientific Computing, Cambridge University Press, Cambridge, UK, 1988.
[24] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, second ed., Morgan Kaufmann, San Francisco, CA, 2005.
[25] G.M. Clarke, T.N. Higgins, Laboratory investigation of hemoglobinopathies and thalassemias: review and update, Clinical Chemistry 46 (2000) 1284–1290.
[26] C.N. Ou, C.L. Rognerud, Diagnosis of hemoglobinopathies: electrophoresis vs. HPLC, Clinica Chimica Acta 313 (2001) 187–194.
[27] R.B. Colah, R. Surve, P. Sawant, E. D'Souza, K. Italia, S. Phanasgaonkar, A.H. Nadkarni, A.C. Gorakshakar, HPLC studies in hemoglobinopathies, Indian Journal of Pediatrics 74 (2007) 657–662.
[28] A. Joutovsky, J. Hadzi-Nesic, M.A. Nardi, HPLC retention time as a diagnostic tool for hemoglobin variants and hemoglobinopathies: a study of 60,000 samples in a clinical diagnostic laboratory, Clinical Chemistry 50 (2004) 1736–1747.
[29] D. Pierrakos, G. Paliouras, Personalizing web directories with the aid of web usage data, IEEE Transactions on Knowledge and Data Engineering 22 (2010) 1331–1344.
[30] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.
[31] T.M. Mitchell, Machine Learning, McGraw-Hill, Singapore, 1997.
[32] B. Cestnik, Estimating probabilities: a crucial task in machine learning, in: L. Aiello (Ed.), Proceedings of the 9th European Conference on Artificial Intelligence, Pitman, London, UK, 1990, pp. 147–149.
[33] D.E. Rumelhart, J.L. McClelland (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1, MIT Press, Cambridge, MA, 1986.
[34] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: C.S. Mellish (Ed.), Proceedings of the 14th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, San Mateo, CA, 1995, pp. 1137–1143.
[35] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1997) 273–324.
[36] M. Robnik-Šikonja, I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, Machine Learning 53 (2003) 23–69.

Please cite this article in press as: D. Setsirichok, et al., Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening, Biomed. Signal Process. Control (2011), doi:10.1016/j.bspc.2011.03.007

References

Amendolia, S. R., Brunetti, A., Carta, P., Cossu, G., Ganadu, M. L., Golosio, B., Mura, G. M. and Pirastru, M. G. (2002). A real-time classification system of thalassemic pathologies based on artificial neural networks. *Medical Decision Making*, *22*, 18-26.

Amendolia, S. R., Cossu, G., Ganadu, M. L., Golosio, B., Masala, G. L. and Mura, G. M. (2003). A comparative study of *k*-nearest neighbour, support vector machine and multi-layer perceptron for thalassemia screening. *Chemometrics and Intelligent Laboratory Systems*, *69*, 13-20.

Barosi, G., Cazzola, M., Berzuini, C., Quaglini, S. and Stefanelli, M. (1985). Classification of anemia on the basis of ferrokinetic parameters. *British Journal of Haematology*, *61*, 357-370.

Birndorf, N. I., Pentecost, J. O., Coakley, J. R. and Spackman, K. A. (1996). An expert system to diagnose anemia and report results directly on hematology forms. *Computers and Biomedical Research*, *29*, 16-26.

Carter, K. W., McCaskie, P. A. and Palmer, L. J. (2006). JLIN: a java based linkage disequilibrium plotter. *BMC Bioinformatics*, *7*, 60.

Cestnik, B. (1990). Estimating probabilities: a crucial task in machine learning. In L. Aiello (Ed.), *Proceedings of the 9th European Conference on Artificial Intelligence* (pp. 147-149). London, UK: Pitman.

Cho, Y. M., Ritchie, M. D., Moore, J. H., Park, J. Y., Lee, K. U., Shin, H. D., Lee, H. K. and Park, K. S. (2004). Multifactor-dimensionality reduction shows a two-locus interaction associated with type 2 diabetes mellitus. *Diabetologia*, *47*, 549-554.

Clarke, G. M. and Higgins, T. N. (2000). Laboratory investigation of hemoglobinopathies and thalassemias: review and update. *Clinical Chemistry*, *46*, 1284-1290.

Colah, R. B., Surve, R., Sawant, P., D′Souza, E., Italia, K., Phanasgaonkar, S., Nadkarni, A. H. and Gorakshakar, A. C. (2007). HPLC studies in hemoglobinopathies. *Indian Journal of Pediatrics*, *74*, 657-662.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, *11*, 2463-2468.

Culverhouse, R., Suarez, B. K., Lin, J. and Reich, T. (2002). A perspective on epistasis: limits of models displaying no main effect. *American Journal of Human Genetics*, *70*, 461-471.

Demir, A., Yarali, N., Fisgin, T., Duru, F. and Kara, A. (2002). Most reliable indices in differentiation between thalassemia trait and iron deficiency anemia. *Pediatrics International*, *44*, 612-616.

Dudek, S. M., Motsinger, A. A., Velez, D. R., Williams, S. M. and Ritchie, M. D. (2006). Data simulation software for whole-genome association and other studies in human genetics. In R. B. Altman, A. K. Dunker, L. Hunter, T. Murray and T. E. Klein (Eds.), *Proceedings of the Pacific Symposium on Biocomputing 2006* (pp. 499-510). Singapore: World Scientific.

Engle, R. L., Flehinger, B. J., Allen, S., Friedman, R., Lipkin, M., Davis, B. J. and Leveridge, L. L. (1976). HEME: a computer aid to diagnosis of hematologic disease. *Bulletin of the New York Academy of Medicine*, *52*, 584-600.

Epstein, M. P. and Satten, G. A. (2003). Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics*, *73*, 1316-1329.

Fayyad, U. M. and Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In R. Bajcsy (Ed.), *Proceedings of the 13th International Joint Conference on Artificial Intelligence* (pp. 1022-1027). San Mateo, CA: Morgan Kaufmann.

Fiorito, M., Torrente, I., De Cosmo, S., Guida, V., Colosimo, A., Prudente, S., Flex, E., Menghini, R., Miccoli, R., Penno, G., Pellegrini, F., Tassi, V., Federici, M., Trischitta, V. and Dallapiccola, B. (2007). Interaction of *DIO2* T92A and *PPARγ2* P12A polymorphisms in the modulation of metabolic syndrome. *Obesity*, *15*, 2889-2895.

Fucharoen, S. and Winichagoon P. (1997). Hemoglobinopathies in Southeast Asia: molecular biology and clinical medicine. *Hemoglobin*, *21*, 299-319.

Gayán, J., González-Pérez, A., Bermudo, F., Sáez, M. E., Royo, J. L., Quintas, A., Galan, J. J., Morón, F. J., Ramirez-Lorca, R., Real, L. M. and Ruiz, A. (2008). A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis. *BMC Genomics*, *9*, 360.

Gloria-Bottini, F., Magrini, A., Antonacci, E., La Torre, M., Di Renzo, L., De Lorenzo, A., Bergamaschi, A. and Bottini, E. (2007). Phosphoglucomutase genetic polymorphism and body mass. *American Journal of Medical Sciences*, *334*, 421-425.

Hall, M. A. and Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, *15*, 1437-1447.

Hallgrímsdóttir, I. B. and Yuster, D. S. (2008). A complete classification of epistatic two-locus models. *BMC Genetics*, *9*, 17.

Hanson, R. L., Ehm, M. G., Pettitt, D. J., Prochazka, M., Thompson, D. B., Timberlake, D., Foroud, T., Kobes, S., Baier, L., Burns, D. K., Almasy, L., Blangero, J., Garvey, W. T., Bennett, P. H. and Knowler, W. C. (1998). An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians. *American Journal of Human Genetics*, *63*, 1130-1138.

Heidema, A. G., Boer, J. M. A., Nagelkerke, N., Mariman, E. C. M., van der A, D. L. and Feskens, E. J. M. (2006). The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. *BMC Genetics*, *7*, 23.

Heidema, A. G., Feskens, E. J. M., Doevendans, P. A. F. M., Ruven, H. J. T., van Houwelingen, H. C., Mariman, E. C. M. and Boer, J. M. A. (2007). Analysis of multiple SNPs in genetic association studies: comparison of three multi-locus methods to prioritize and select SNPs. *Genetic Epidemiology, 31*, 910-921.

Hoh, J., Wille, A. and Ott, J. (2001). Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Research, 11*, 2115-2119.

Hsieh, C. H., Liang, K. H., Hung, Y. J., Huang, L. C., Pei, D., Liao, Y. T., Kuo, S. W., Bey, M. S. J., Chen, J. L. and Chen, E. Y. (2006). Analysis of epistasis for diabetic nephropathy among type 2 diabetic patients. *Human Molecular Genetics*, *15*, 2701-2708.

Jakulin A. and Bratko, I. (2003). Analyzing attribute dependencies. *Lecture Notes in Artificial Intelligence, 2838*, 229-240.

Jakulin, A., Bratko, I., Smrke, D., Demšar, J. and Zupan, B. (2003). Attribute interactions in medical data analysis. *Lecture Notes in Artificial Intelligence*, *2780*, 229-238.

Jimenez, C. V., Minchinela, J. and Ros, J. (1995). New indices from the H*2 analyser improve differentiation between heterozygous beta or delta beta thalassaemia and iron-deficiency anaemia. *Clinical and Laboratory Haematology*, *17,* 151-155.

Joutovsky, A., Hadzi-Nesic, J. and Nardi, M. A. (2004). HPLC retention time as a diagnostic tool for hemoglobin variants and hemoglobinopathies: a study of 60,000 samples in a clinical diagnostic laboratory. *Clinical Chemistry*, *50,* 1736-1747.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *28,* 27-30.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, *36,* D480-D484.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, *34,* D354-D357.

Kohavi R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In C. S. Mellish (Ed.), *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 1137-1143). San Mateo, CA: Morgan Kaufmann.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97,* 273-324.

Lanzola, G., Stefanelli, M., Barosi, G. and Magnani, L. (1990). NEOANEMIA: a knowledge-based system emulating diagnostic reasoning. *Computers and Biomedical Research*, *23*, 560-582.

Leak, T. S., Mychaleckyj, J. C., Smith, S. G., Keene, K. L., Gordon, C. J., Hicks, P. J., Freedman, B. I., Bowden, D. W. and Sale, M. M. (2008). Evaluation of a SNP map of 6q24-27 confirms diabetic nephropathy loci and identifies novel associations in type 2 diabetes patients with nephropathy from an African-American population. *Human Genetics*, *124*, 63-71.

Li, W. and Reich, J. (2000). A complete enumeration and classification of two-locus disease models. *Human Heredity*, *50*, 334-349.

Lund, P. R. and Barnes, R. D. (1972). Automated classification of anaemia using image analysis. *The Lancet*, *300*, 463-464.

March, R. E., Putt, W., Hollyoake, M., Ives, J. H., Lovegrove, J. U., Hopkinson, D. A., Edwards, Y. H. and Whitehouse, D. B. (1993). The classical human phosphoglucomutase (*PGM1*) isozyme polymorphism is generated by intragenic recombination. *Proceedings of the National Academy of Sciences of the United States of America*, *90*, 10730-10733.

Marchini, J., Donnelly, P. and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, *37*, 413-417.

Mitchell, T. M. (1997). *Machine Learning*. Singapore: McGraw-Hill.

Moore J. H. and White, B. C. (2007). Tuning ReliefF for genome-wide genetic analysis. *Lecture Notes in Computer Science*, *4447*, 166-175.

Moore, J. H., Gilbert, J. C., Tsai, C. T., Chiang, F. T., Holden, T., Barney, N. and White, B. C. (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology*, *241*, 252-261.

Motsinger, A. A., Ritchie, M. D. and Reif, D. M. (2007). Novel methods for detecting epistasis in pharmacogenomics studies. *Pharmacogenomics*, *8*, 1229-1241.

Musani, S. K., Shriner, D., Liu, N., Feng, R., Coffey, C. S., Yi, N., Tiwari, H. K. and Allison, D. B. (2007). Detection of gene × gene interactions in genome-wide association studies of human population data. *Human Heredity*, *63*, 67-84.

Neuman, R. J. and Rice, J. P. (1992). Two-locus models of disease. *Genetic Epidemiology*, *9*, 347-365.

Ntaios, G., Chatzinikolaou, A., Saouli, Z., Girtovitis, F., Tsapanidou, M., Kaiafa, G., Kontoninas, Z., Nikolaidou, A., Savopoulos, C., Pidonia, I. and Alexiou-Daniel, S. (2007). Discrimination indices as screening tests for β-thalassemic trait. *Annals of Hematology*, *86*, 487-491.

Old, J. M. (2003). Screening and genetic diagnosis of haemoglobin disorders. *Blood Reviews*, *17*, 43-53.

Ou, C. N. and Rognerud, C. L. (2001). Diagnosis of hemoglobinopathies: electrophoresis vs. HPLC. *Clinica Chimica Acta*, *313*, 187-194.

Pierrakos, D. and Paliouras, G. (2010). Personalizing web directories with the aid of web usage data. *IEEE Transactions on Knowledge and Data Engineering*, *22*, 1331-1344.

Piroonratana, T., Wongseree, W., Assawamakin, A., Paulkhaolarn, N., Kanjanakorn, C., Sirikong, M., Thongnoppakhun, W., Limwongse, C. and Chaiyaratana, N. (2009). Classification of haemoglobin typing chromatograms by neural networks and decision trees for thalassaemia screening. *Chemometrics and Intelligent Laboratory Systems*, *99*, 101-110.

Polat, K. and Günes, S. (2006). Automated identification of diseases related to lymph system from lymphography data using artificial immune recognition system with fuzzy resource allocation mechanism (fuzzy-AIRS). *Biomedical Signal Processing and Control*, *1*, 253-260.

Press, W. H., Flannery, B. P., Teukolski, S. A. and Vetterling, W. T. (1988). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge, UK: Cambridge University Press.

Qi, L., van Dam, R. M., Asselbergs, F. W. and Hu, F. B. (2007). Gene-gene interactions between *HNF4A* and *KCNJ11* in predicting type 2 diabetes in women. *Diabetic Medicine*, *24*, 1187-1191.

Quaglini, S., Stefanelli, M., Barosi, G. and Berzuini, A. (1986). ANEMIA: an expert consultation system. *Computers and Biomedical Research, 19,* 13-27.

Quaglini, S., Stefanelli, M., Barosi, G. and Berzuini, A. (1988). A performance evaluation of the expert system ANEMIA. *Computers and Biomedical Research, 21*, 307-323.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science, 273,* 1516-1517.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F. and Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Human Genetics, 69,* 138-147.

Robnik-Šikonja, M. and Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning, 53,* 23-69.

Rumelhart, D. E. and McClelland, J. L. (Eds.) (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1. Cambridge, MA: MIT Press.

Sale, M. M., Freedman, B. I., Langefeld, C. D., Williams, A. H., Hicks, P. J., Colicigno, C. J., Beck, S. R., Brown, W. M., Rich, S. S. and Bowden, D. W. (2004). A genome-wide scan for type 2 diabetes in African-American families reveals evidence for a locus on chromosome 6q. *Diabetes*, *53*, 830-837.

Schork, N. J., Boehnke, M., Terwilliger, J. D. and Ott, J. (1993). Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. *American Journal of Human Genetics*, *53,* 1127-1136.

Sirichotiyakul, S., Saetung, R. and Sanguansermsri, T. (2009). Prenatal diagnosis of beta-thalassemia/Hb E by hemoglobin typing compared to DNA analysis. *Hemoglobin*, *33,* 17-23.

Spencer, N., Hopkinson, D. A. and Harris, H. (1964). Phosphoglucomutase polymorphism in man. *Nature*, *204*, 742-745.

Sripichai, O., Makarasara, W., Munkongdee, T., Kumkhaek, C., Nuchprayoon, I., Chuansumrit, A., Chuncharunee, S., Chantrakoon, N., Boonmongkol, P., Winichagoon, P. and Fucharoen, S. (2008). A scoring system for the classification of β-thalassemia/Hb E disease severity. *American Journal of Hematology*, *83*, 482-484.

Thameem, F., Wolford, J. K., Wang, J., German, M. S., Bogardus, C. and Prochazka, M. (2002). Cloning, expression and genomic structure of human *LMX1A*, and variant screening in Pima Indians. *Gene*, *290*, 217-225.

The GAIN Collaborative Research Group. (2007). New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nature Genetics*, *39*, 1045-1051.

The Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*, 661-678.

Watanabe, I., Tomita, A., Shimizu, M., Sugawara, M., Yasumo, H., Koishi, R., Takahashi, T., Miyoshi, K., Nakamura, K., Izumi, T., Matsushita, Y., Furukawa, H., Haruyama, H. and Koga, T. (2003). A study to survey susceptible genetic factors responsible for troglitazone-associated hepatotoxicity in Japanese patients with type 2 diabetes mellitus. *Clinical Pharmacology and Therapeutics*, *73*, 435-455.

Weatherall, D. J. and Clegg, J. B. (2001). *The Thalassemia Syndromes* (4th ed.). Malden, MA: Blackwell Science.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). San Francisco, CA: Morgan Kaufmann.

Wongseree, W., Chaiyaratana, N., Vichittumaros, K., Winichagoon, P. and Fucharoen, S. (2007). Thalassaemia classification by neural networks and genetic programming. *Information Sciences*, *177*, 771-786.

Yu, W., Gwinn, M., Clyne, M., Yesupriya, A. and Khoury, M. J. (2008). A navigator for human genome epidemiology. *Nature Genetics*, *40*, 124-125.

Zeggini, E., Scott, L. J., Saxena, R. and Voight, B. F. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics*, *40*, 638-645.