รูปแบบ Abstract (บทคัดย่อ)

Project Code : TRG5780062

Project Title : Development of genetic variation database and bioinformatics tools that promote standard nomenclature and data sharing for Human Variome Project

Investigator : Supakit Prueksaaroon, Faculty of Engineering, Thammasat University
(ชื่อนักวิจัย)

E-mail Address : psupakit@engr.tu.ac.th

Project Period : 2 ปี
(ระยะเวลาโครงการ)

Abstract (บทคัดย่อ)

ในยุคหลังของการพัฒนาด้านพันธุกรรม หลายประเทศได้เกิดความร่วมมือในการแข่งปันองค์ความรู้ด้านพันธุกรรมมนุษย์ในอันที่จะสร้างศูนย์กลางการจัดเก็บข้อมูลจีโนมร่วมกัน อาทิเช่นโครงการ HapMap ที่ได้มีการพัฒนากลไกการค้นหา และการหาความสัมพันธ์ของข้อมูลพันธุกรรมในยีน การค้นหาสาเหตุของการกลายพันธุ์ ซึ่งปัญหาของข้อมูลดังกล่าวคือมีความกระจัดกระจาย และไม่สามารถแบ่งปันได้อย่างมีประสิทธิภาพ โครงการ Human Variome Project (HVP) โดย Prof. Richard Cotton ได้พลักดันการสร้างมาตรฐานของข้อมูลที่รองรับการพัฒนาเป็นฐานข้อมูลความหลากหลายทางชีวภาพ และการกลายพันธุ์ เพื่อให้มีการใช้งานในหลายประเทศ ซึ่งประเทศในภูมิภาคอาเซียน ประเทศไทยมีความพร้อมในการพัฒนาเป็น โครงการความหลากหลายข้อมูลทางชีวภาพ ได้นำเสนอการกำหนดรูปแบบของมาตรฐานข้อมูลความหลากหลายทางชีวภาพ ที่รอบรับกับมาตรฐานของฐานข้อมูลยีน ฐานข้อมูลการกลายพันธุ์ ที่มีการใช้ในประเทศสมาชิก โดยในหลายประเทศในภูมิภาคเอเชียตะวันออกเฉียงใต้ ประเทศไทย ได้เป็นหนึ่งในประเทศสมาชิกของโครงการ ในระดับศูนย์ข้อมูลประเทศ โดยในระดับศูนย์ข้อมูลประเทศ มีการสร้างคลังในการจัดเก็บ และแลกเปลี่ยนข้อมูล ภายใต้มาตรฐานการแบ่งปันข้อมูลกลางระหว่างประเทศสมาชิก ซึ่งคลังจัดเก็บข้อมูลจะต้องรองรับปริมาณข้อมูลขนาดใหญ่ และมีประสิทธิภาพการทำงานที่ดี โดยระบบไฟล์ซิสเต็ม Hadoop รองรับการจัดเก็บไฟล์ข้อมูลแบบกระจาย มีความสามารถในการจัดการไฟล์ข้อมูลขนาดใหญ่ ในงานวิจัยนี้ ดำเนินการพัฒนาระบบจัดเก็บข้อมูลสำหรับการแบ่งปันในโครงการความหลากหลายทางชีวภาพบนเทคโนโลยีของ Hadoop Federation ซึ่งสนับสนุนการเข้าถึงข้อมูลจากหลายศูนย์ข้อมูลประเทศได้พร้อมกัน ผ่านกลไกการพิสูจน์ตัวตนด้วยบัญชีผู้ใช้ส่วนกลาง โดยผู้ดูแลระบบของแต่ละประเทศเป็นผู้ควบคุมสิทธิ

การเข้าถึงข้อมูล โดยผลการทดสอบแสดงให้เห็นถึงประสิทธิภาพในการแบ่งปันข้อมูลภายในประเทศ และระหว่างประเทศสมาชิก ผลของงานวิจัยนี้นำไปสู่เครื่องมือในการแบ่งปันข้อมูลความหลากหลายทางพันธุกรรมมนุษย์ระหว่างประเทศสมาชิก

คำสำคัญ : HVP, ความหลากหลายทางพันธุ์กรรมมนุษย์, การกลายพันธุ์, Hadoop

During the post-genomic era, many multinational collaborative projects utilizing knowledge of human genome were initiated to construct a central repository storing important annotated genome information, e.g., international HapMap project etc. With the advent of high throughput sequencing and genotyping technologies, mapping a disease gene or discovering the causative mutations can be achieved with much less effort. However, such gene/mutation-disease information is scattered and cannot be shared efficiently. The Human Variome Project (HVP) led by Prof. Richard Cotton was conceived with the intention to standardize and share the information. This can conceptually be achieved by setting up a global (virtual) gene-disease database comprising many ethnic-specific mutation databases from all participating countries (country nodes). Among several countries in Southeast Asian, Thailand already had a large collection of human mutations in this region. As a country node, we need to create a repository that can *host* and *exchange* human genetic variations in the *standard nomenclature* representation with other participating countries. The repository storages might have support large amount of data and acquire good performance. The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. HDFS provides high throughput access and is suitable for large-scale data sets. In this work, we develop shared data repository using Hadoop federation. The users in each country node can access data from any registered servers from their own country by using federation account. The local administrator has controlled the accessing permission. We performed experiment to evaluate the access time between local nodes, and remote country nodes. The contribution of this work seeks for the support to accomplish the requirement of a country node by curating human variation-disease information and providing tool to exchange such data with other HVP country nodes.

Keywords : HVP, Genetic variation, Mutation, Standard mutation nomenclature, Hadoop

1. Introduction

The Human Variome Project (http://www.humanvariomeproject.org) was established with a vision to provide a complete knowledge about genetic diseases by means of sharing human variations that are associated with diseases. Tackling this problem by creating a single database where everyone should submit their variation data to be not practical due to varying requirements from different environments. The HVP project, hence, focuses on fostering a development of ethnic-specific in each country. These nodes should be able to interoperate with other nodes so that the underlying variation data can be shared. At the early stage, several working groups have been set up to address the two main goals 1) collecting specific gene-disease and 2) encouraging different countries to collect country-specific variations[1,2]. The collected information must comply with the mutation standard nomenclature[3] so that the data from different nodes can be shared. Such an interoperability standard was proposed by Patrinos et al.[4] for every country node to follow that also include