Database System for Biological Data Mining

Harvey H, Kriangsak T.

February 22, 2020

Introduction

Biological data has been made available on many online databases due to advances in data collection: whole genome sequences or high throughput sequencing, to name a few. This allows biologists to aggregate the data of not only humans but other species of interest. In this class project, we will work on constructing a database system for one kind of biological data: Single Nucleotide Polymorphisms (SNPs).

Single nucleotide polymorphisms, frequently called SNPs (pronounced "snips"), are the most common type of genetic variation among people. Each SNP represents a difference in a single DNA building block, called a nucleotide. For example, a SNP may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in a certain stretch of DNA.

SNPs occur normally throughout a person's DNA. They occur almost once in every 1,000 nucleotides on average, which means there are roughly 4 to 5 million SNPs in a person's genome. These variations may be unique or occur in many individuals; scientists have found more than 100 million SNPs in populations around the world. Most commonly, these variations are found in the DNA between genes. They can act as biological markers, helping scientists locate genes that are associated with disease. When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease by affecting the gene's function.

Most SNPs have no effect on health or development. Some of these genetic differences, however, have proven to be very important in the study of human health. Researchers have found SNPs that may help predict an individual's response to certain drugs, susceptibility to environmental factors such as toxins, and risk of developing particular diseases. SNPs can also be used

to track the inheritance of disease genes within families. Future studies will work to identify SNPs associated with complex diseases such as heart disease, diabetes, and cancer. (I will put citation in the official paper)

As the data becomes available and there shall be kinds of pattern one my try to draw from conducing some experiments. Our goal is to try to use some simple data mining algorithms to interpret the data.

Remark: The data for this project is obtained form UniProt and dbSNP oneline databases.

Objectives and Solutions

In this class project, we are trying to construct a database such that it is optimal for potential queries in SNPs interpretations. To be precise, as this is the data of data mining and machine learning, we are aiming to organize our database such that it allows faster better functionality for data mining problems. The list of the what we will try to provide is as follows:

- As a database shall have, we will provide basic functionality including: create, search, update and delete functions.
- What are some of patterns of amino acid chances that are the most frequent (using Apriori and Sequential Pattern Data Mining algorithms.)
- And Which of those causes diseases and which does not.
- After obtaining the frequency table, we would try to further our understanding of the data by trying to come up with pattern drawn using associative rules.
- In optimizing queries for the above techniques, we will adopt the idea of indexing so that it gives a faster search engine on our database.
- we will also try to allow visualization obtained from mining algorithms.
- If the time would not be the constraint, we will try to develop a (nice) front-end application for interactive users.