CPSC545 Introduction to Data Mining

by *Prof. Martin Schultz &*
Prof. Mark Gerstein

Student Name: Yu Kor Hugo Lam
Student ID : 904907866
Due Date : March 27, 2007

Proposed Final Project

Introduction

Pseudogenes are "dead" genes that originate from functional known genes. They do not either code for any protein or express in the cell. There are two main types of pseudogenes, processed pseudogenes and duplicated pseudogenes. Processed pseudogenes are pseudogenes that are generated by retrotransposition. Duplicated pseudogenes are pseudogenes that are generated by gene duplication and mutation. Since each pseudogene has a parent gene and each gene should have a protein family, each pseudogene can also be classified into different protein families, what we call pseudofams, according to their parent genes though they are not protein coding.

Prediction

Classifying different pseudogenes into different pseudofams yields different numbers of pseudogenes and genes for each of those families. Since a protein family represents a group of evolutionarily related proteins, pseudogenes in the same pseudofam also share the same evolutionary origin. As pseudogenes do not have protein coding ability, they can evolve under very little selection pressure. As a result, they have a higher degree of freedom to evolve over time comparing to normal genes. For example, a pseudofam can have more than a thousand of pseudogenes while it only has about 10 genes; or a pseudofam can have no pseudogene but more than 20 genes. So given a set of pseudofams, it is interesting to see if the number of pseudogenes and genes in one species could formulate a model to predict whether there are any pseudogenes in the same pseudofams in another species (closely or distantly related).

Method

1. Generate pseudogene datasets using the Pseudogene Pipeline at Gerstein Lab
2. Classify pseudogenes into different pseudofams based on their parent genes
3. Count the number of total pseudogenes, processed pseudogenes, duplicated pseudogenes, normal genes, and etc in each pseudofam
4. Use data mining techniques such as Support Vector Machines with different kernels to predict the aforementioned
5. Compare and analyze the results from the prediction