CPSC545 Introduction to Data Mining
by *Prof. Martin Schultz &*
Prof. Mark Gerstein

Student Name: Yu Kor Hugo Lam
Student ID : 904907866
Due Date : May 7, 2007

Final Project Report

Introduction

Pseudogenes are "dead" genes that apparently have no functions. They originate from genes that once functional. Due to certain kinds of mutation and sequence rearrangement, these genes are disabled and cannot be transcribed or translated into any functional molecules. There are two main types of pseudogenes, they are 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.

Retrotransposition is a process that reverse transcribes an mRNA, or a portion of it, back into the DNA. Normally, the mRNA transcribed from the DNA will be translated into a protein. However, there are also times that it will be retrotransposed into the DNA. Because the mRNA being retrotransposed is usually mature, having the introns spliced out, a Poly-A tail, and most importantly lacking a promoter, it becomes non-functional even though it is integrated back into the DNA. As a result, it becomes a processed pseudogene.

Duplicated pseudogene arises from a different mechanism, gene duplication, which is common and important in the evolution of genomes. Gene duplication usually occurs as a result of an error in homologous recombination. Since the duplicated gene oftentimes has no selective pressure, its mutation happens faster and much more freely if comparing to a single-copy gene. When a mutation disables the function of the duplicate, it becomes a non-functional pseudogene, which may have introns and promoter, but usually does not have much impact.

Although pseudogenes are not functional, or not functioning like a normal gene does, studying pseudogenes is not trivial. For instances, it gives us hints about life histories like a fossil record; some of them seem to be actively transcribed and may involve in calibrating the genome; some can be converted back into a functional gene through mutations; and it improves gene annotation.

Therefore, researchers have been developing different techniques to identify, classify and annotate pseudogenes. PseudoPipe developed at Gerstein Lab at Yale is a case in point. It is an automated pseudogene identification pipeline that involves using BLAST to rapidly cross reference potential parent proteins against the intergenic regions of the genome, processing the resulting hits, and classifying the pseudogenes based on a combination of criteria including intron-exon structure, existence of stop codons, frameshifts and etc.

Goal

PseudoPipe classifies pseudogenes as Processed Pseudogene, Duplicated Pseudogene, and Pseudogene Fragment. Basically, if a pseudogene has more than one exon, it will be classified as duplicated pseudogene. If it has one exon and spans 70% the parent or more, it will be classified as processed pseudogene. Otherwise, if it spans less than 70%, it will be classified as pseudogene fragment, which means it cannot be determined as either a processed pseudogene or a duplicated pseudogene.

The goal of this project is first to reproduce the classification model to see if the data conform to these criteria, second to identify any clusters among the pseudogenes, and then to train a model to predict the types of those pseudogene fragments.

Method

Data Source

The pseudogene dataset was retrieved from PseudoPipe (http://pgenes.gersteinlab.org:9917/) and the data corresponding to processed pseudogene, duplicated pseudogene, and pseudogene fragment were extracted. The attributes of the data are as follows:

Chromosome	Pgene chromosome	NumDels	Number of deletions
ChrStart	Start of pgene	NumShifts	Number of shifts
ChrEnd	End of pgene	NumStops	Number of stops
Strand	Strand on chromosome	Expect	Expectation score
Parent	Gene parent	PctIdent	Percent identical
QueryStart	Start of matched parent region	PolyA	Poly A signal
QueryEnd	End of matched parent region	Disable	Disablement class
QueryLength	Length of parent matched	NumExons	Number of exons
Frac	Fraction of parent spanned	Class	Pgene classification
NumIns	Number of insertions		

Tools

Weka was used to carry out the data mining tasks. It is an open-source software with a collection of machine learning algorithms for data mining. It has a GUI to apply the algorithms directly to a dataset and a set of API for writing custom code. It also contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

R, a free software environment for statistical computing and graphics, was used to process the data generated from Weka and to generate those graphs that are not available in Weka. And ScatterPlot3D, a library package in R which plots a three dimensional (3D) point cloud, was used to plot the clustering result in 3D.

Decision Tree

Decision tree, *a.k.a*. classification tree and reduction tree, was used to reproduce the pseudogene classification model. In a decision tree in data mining, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. It is a supervised approach to classification and its implementation in Weka is called J48. The algorithm can be summarized as follows:

1. Create a node and choose an attribute that best separates the data.
2. Create a branch for each value or each set of values of the node's attribute
3. Assign the data to different branches according to the attribute values of the branches
4. For each branch,
    a. if all members have the same class, terminate and label it with that class
    b. if it has no further distinguishing attributes, label it with the major class in it
5. For each branch that has not been labeled, repeat the process

K-means

K-means, following the attribute selection by Principle Component Analysis (PCA) which will be discussed later, was used to cluster the pseudogenes. It is one of the simplest unsupervised learning algorithms that solve clustering problem. It clusters the objects based on their attributes into k groups. The idea is to find the centers of the k clusters. However, it assumes that k is known in advance. Following is the summary of the algorithm:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids
2. Assign each object to the group that has the closest centroid
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Principle Component Analysis

A common weakness of clustering is that it depends pretty much on visualizing the obtained clusters. However, visualizing high dimensional data directly is technically impossible. PCA, principle component analysis, can transform the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. Therefore, the PCA procedure was applied to the pseudogene data as an attribute selection process. Then the k-means algorithm was applied to the data in the new feature space. The algorithm of PCA is as follows:

1. Subtract the mean from each of the data dimensions. The mean subtracted is the average across each dimension.
2. Calculate the covariance matrix. That is all possible covariance values between all the different dimensions.
3. Calculate the eigenvectors and eigenvalues of the covariance matrix, which is also a square matrix.
4. Order the eigenvectors by their eigenvalues, from highest to lowest. This gives the components in order of significance.
5. Derive the new data set by multiplying the eigenvectors with the original data set.

Bayesian Network

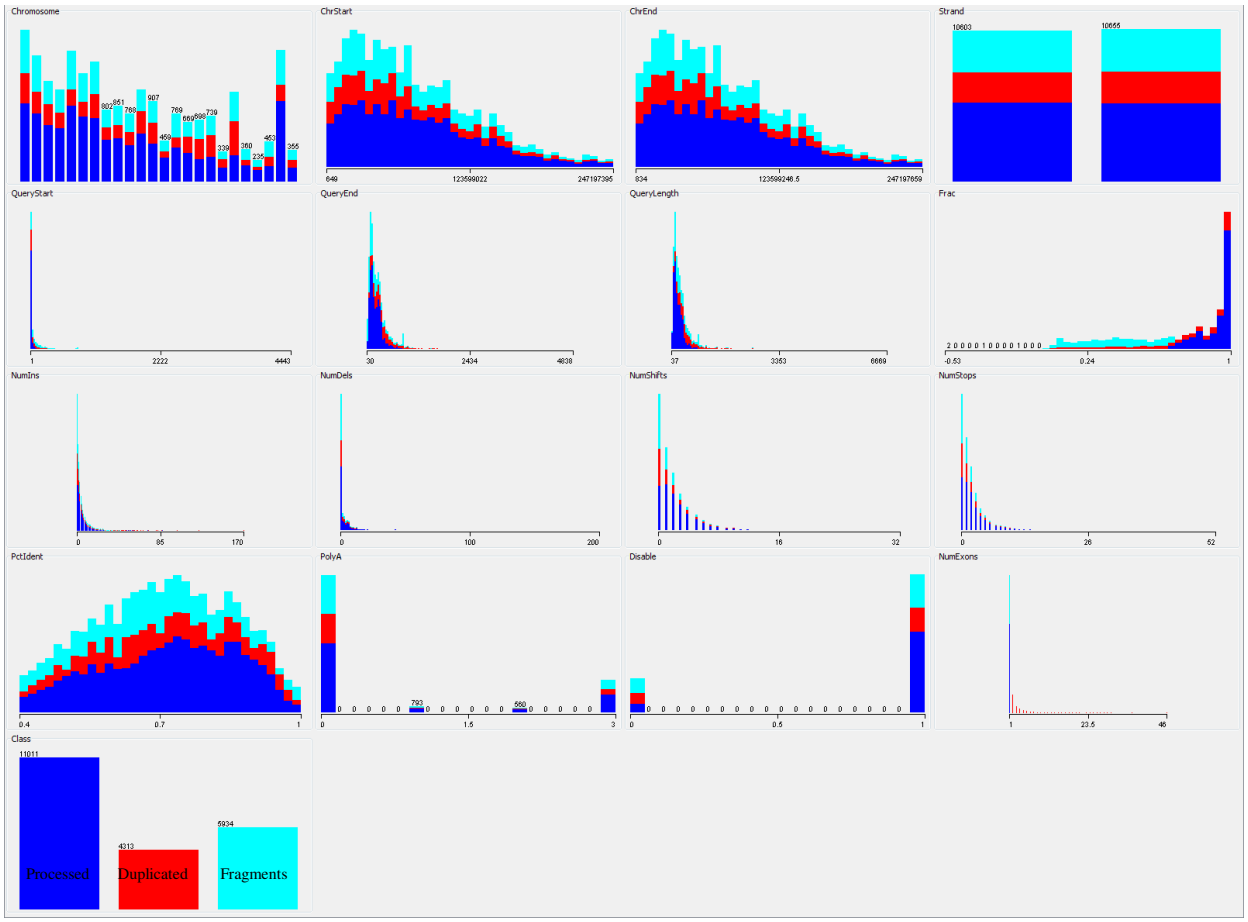
A Bayesian Network is a directed acyclic graph that contains nodes and arcs in which the nodes represent the variables and the arcs encode the conditional dependencies. Probabilistic inference could be carried out by computing the posterior distribution of variables given the evidence. It was chosen to create a probabilistic model to predict the types of the pseudogene fragments.

Results

Getting and preprocessing the data

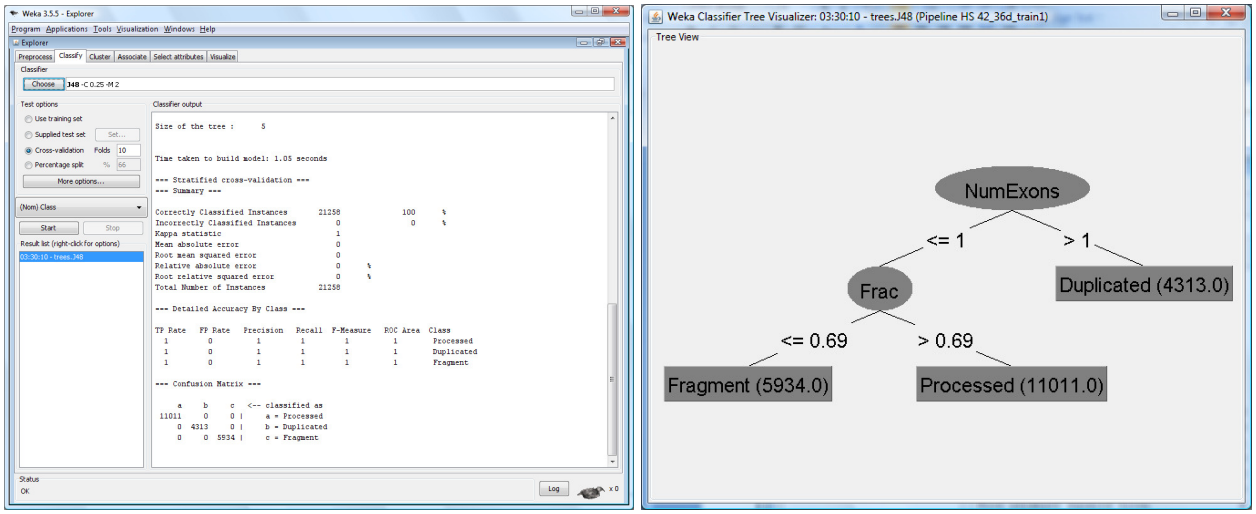
The pseudogene data set, which contains 27,763 records, was generated by the PseudoPipe. Data corresponding to processed pseudogene, duplicated pseudogene, and pseudogene fragment were extracted, which resulted in a data set having 21,258 records. Records of the pseudogene fragments were removed for training the Bayes Net. Moreover, a set of attributes were selected for PCA and Bayes Net. Details will be discussed later. Following shows a portion of the data and the distribution of the data:

Chromosome	ChrStart	ChrEnd	Strand	Parent	QueryStart	QueryEnd	QueryLength	Frac	NumIns	NumDels	NumShifts	NumStops	Expect	PctIdent	PolyA	Disable	NumExons	BasicClass
1	484	639	-	ENSP00000332932	5	56	194	2.700000e-01	0	0	0	0	1.000000e-08	7.120000e-01	0	0	1	FP
1	817	1367	+	ENSP00000371947	1	181	181	1.000000e+00	4	0	1	1	2.000000e-24	9.290000e-01	0	D	1	PSSDD0
1	2845	3533	+	ENSP00000228264	821	970	970	1.500000e-01	0	4	0	0	1.000000e-28	7.800000e-01	0	0	3	DUP
1	3887	4009	+	ENSP00000295199	1	41	41	1.000000e+00	0	0	0	0	4.000000e-17	9.760000e-01	0	0	1	GENE-SINGLE
1	42315	43196	+	ENSP00000373977	7	307	315	9.600000e-01	8	9	7	5	6.000000e-55	6.260000e-01	0	D	1	PSSDD0



Reproducing the pseudogene classification model

As mentioned in the method section, decision tree was used to reproduce the pseudogene classification model. Except for the "class" attribute which specifies the type of the pseudogenes, all other attributes, such as number of insertion, number of deletion, and number of frameshifts, were used to build the decision tree model. The whole data set was used as a training set and a 10-fold cross validation was used to validate the model. Following shows the result:



The result shows that there were 21,258 instances correctly classified and 0 instance incorrectly classified. The resulted decision tree has a final size of five and consists of two nodes and three leaves.

The root node is "NumExons", which is the number of exons. When the number of exons is greater than 1, the instance will be classified as "Duplicated", which is Duplicated Pseudogene. When the number of exons is less than or equal to 1, it comes to a child decision node "Frac", which is the fraction of parent spanned by the pseudogene. If the instance spans 70% or more its parent, it will be classified as "Processed", which is Processed Pseudogene. Otherwise, if it spans less than 70% of its parent, it will then be classified as "Fragment", which is Pseudogene Fragment.

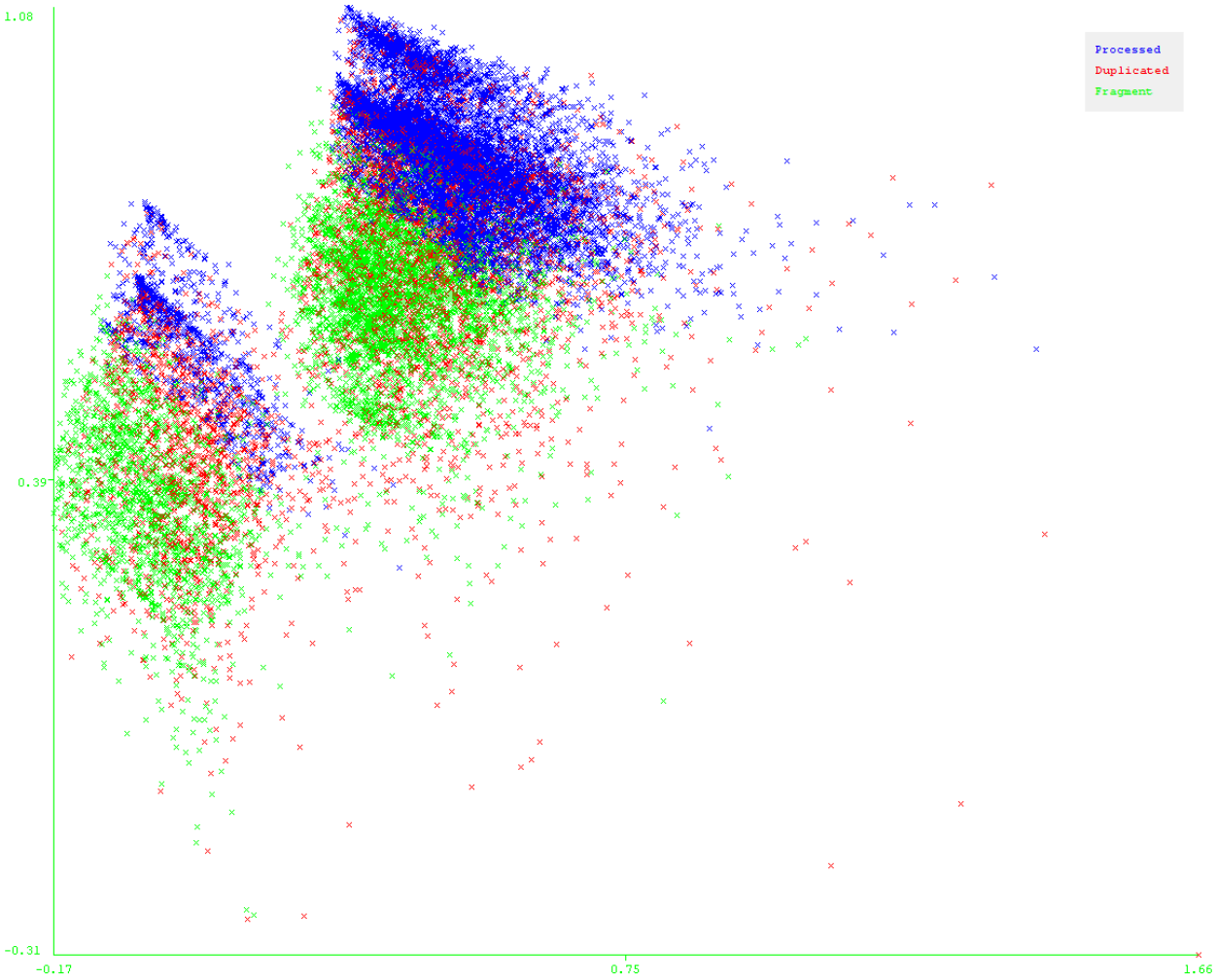
The decision tree generated on the data successfully reproduced the pseudogene classification model which we already discussed in the previous section. The tree is clean and accurately represents the model underlying the data.

Selecting attributes for clustering

Principle Components Analysis was used to select attributes for the clustering because it enhances visualization of the high dimensional data and the order of the components, each contains a set of the attributes, to certain extent represents their significance.

Before the analysis, a subset of the attributes was chosen to reduce the dimension of the data and so the complexity of the analysis. The chosen attributes were QueryLength, Frac, NumIns, NumDels, NumShifts, NumStops, PcIdent, PolyA, Disable, and NumExons.

After the analysis, there were 9 principle components generated and all of them were selected to transform the original data. The figure below shows the plot of the second principle component against the first principle component:



Clustering in the new feature space

After the attribute selection process, the original data was transformed by the principle components and a K-means clustering was performed in the new feature space. By visualizing the previous plot of the second principle component against the first principle component, it can be seen that there could be 2 clusters in the feature space. As a result, k was set to 2 in the clustering. Following shows the result of the clustering:

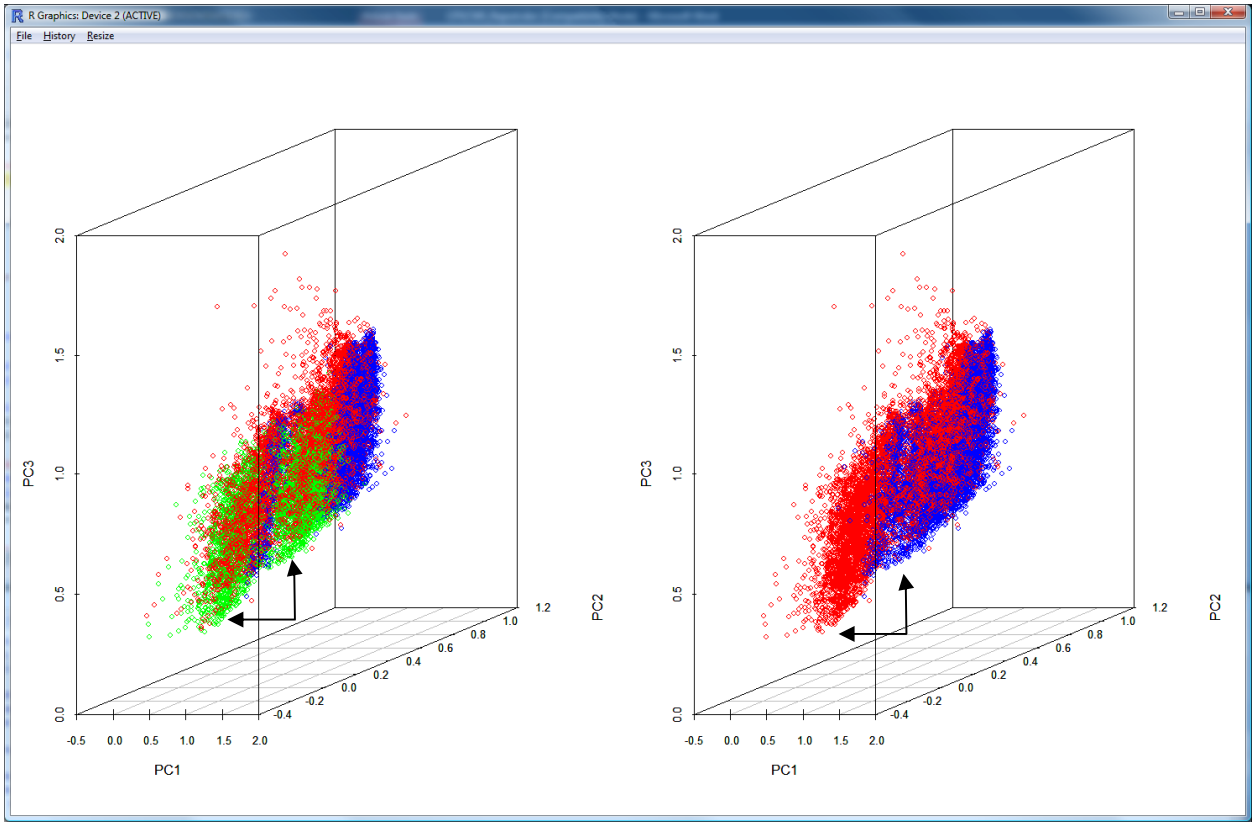


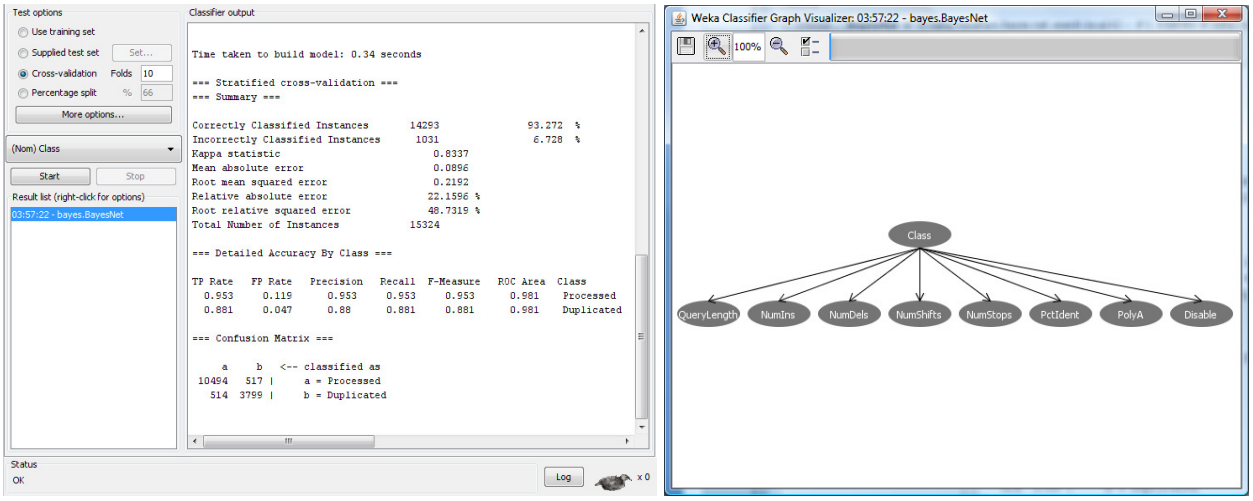
Figure on the left shows the data of the processed pseudogenes (blue), duplicated pseudogenes (red), and pseudogene fragments (green) plotted in 3D by the first three principle components. Figure on the right shows the data of cluster 1 (red) and cluster 2 (blue). From the patterns shown in the figures, it is very likely that cluster 1 represents duplicated pseudogenes and cluster 2 represents processed pseudogenes. However, these two clusters do not seem to represent the clusters visualized (see the arrows). And by removing the attribute Disable, these two visualized clusters disappeared. So they probably represent a group of pseudogenes having disablement and a group without disablement. And after removing the attribute PolyA, which probably accounted for 4 clusters in several other PCA plots, no observable clusters could be identified.

Building a prediction model

The previous result showed that the clusters identified by K-means could be potentially used to predict the types of pseudogene fragments. However, a supervised learning technique may be more suitable to build such a model, given that our classification of processed pseudogenes and duplicated pseudogenes is significant.

Bayesian Network was chosen to build a probabilistic prediction model for pseudogene fragments. The training data set was the original data set with the data of pseudogene fragments being removed. The attributes used in building the model were same as the attributes used in PCA, except that the NumExons and Frac attributes were removed since the information of these two attributes in the pseudogene fragments is weak.

Following shows the Bayes Net trained by the data set aforementioned:



The model has a 10-fold cross validation and the result shows that it correctly classified 14,293 instances (93.3%) and incorrectly classified 1,031 instances (6.7%). It has a 0.95 precision and recall, a true positive rate of 0.953, and a false positive rate of 0.119 for classifying processed pseudogenes. And for duplicated pseudogenes, it has 0.88 precision and recall, a true positive rate of 0.881 and a false positive rate of 0.047.

The prediction model for pseudogene fragments based on the Bayesian Network and the pseudogene data set shows a relatively high accuracy in classifying processed pseudogenes and duplicated pseduogenes.

Conclusion

The decision tree model successfully reproduced the pseudogene classification model from the date set generated by the PseudoPipe system. The K-means clustering on the data transformed into principle components identified two clusters, which could be labeled as processed pseudogene and duplicated pseudogene. However, no observable and unknown clusters have been identified in this investigation. And by using Bayesian Network, a highly accurate prediction model for predicting the types of pseudogene fragments has been built.

References

Literatures

- Gerstein, M & Zheng, D. The real life of pseudogenes. Sci Am 295: 48-55 (2006).
- Zhang, Z, Carriero, N , Zheng, D, Karro, J, Harrison, PM & Gerstein, M. PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics 22: 1437-9 (2006).

Websites

- http://pgenes.gersteinlab.org:9917/
- http://grb.mnsu.edu/grbts/doc/manual/J48_Decision_Trees.html
- http://csnet.otago.ac.nz/cosc453/
- http://dataminingresearch.blogspot.com/search/label/PCA
- http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/kmeans.html
- http://en.wikipedia.org/wiki/Karhunen-Lo%C3%A8ve_transform
- http://en.wikipedia.org/wiki/Bayesian_network
- http://www.cs.waikato.ac.nz/ml/weka/
- http://www.r-project.org/