Clustering Pseudogenes

USING K-MEANS WITH PCA

by Hugo Y. K. Lam

Introduction

SECTION ONE

Pseudogene

- "false" genes , which look like real genes but have no apparent function
- First recognized and dubbed pseudogenes during the late 1970s, when early gene hunters began trying to pinpoint the chromosomal locations associated with production of important molecules

Gerstein, M & Zheng, D. The real life of pseudogenes. Sci Am 295: 48-55 (2006).

Pseudogene

- For example, while seeking the gene responsible for making betaglobin (a key component of the hemoglobin protein that transports oxygen through the bloodstream) scientists identified a DNA sequence that looked like a globin gene but could not possibly give rise to a protein.
- Essential functional parts of the gene's anatomy were disabled by mutations, making it impossible for cellular machinery to translate the gene into a useful molecule.

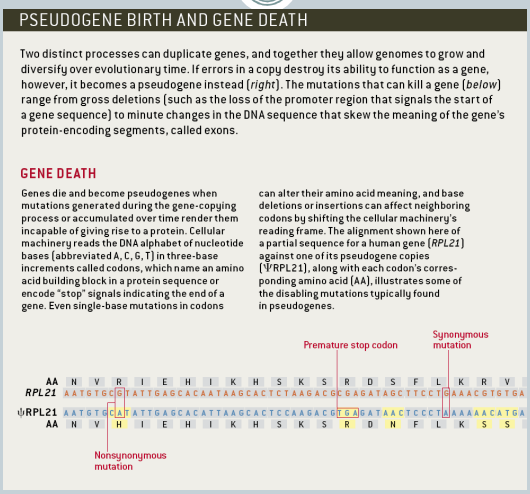
Gerstein, M & Zheng, D. The real life of pseudogenes. Sci Am 295: 48-55 (2006).

Pseudogenome

- Pseudogenes are the molecular remains of broken genes, which are unable to function because of lethal injury to their structures.
- The great majority of pseudogenes are damaged copies of working genes and serve as genetic fossils that offer insight into gene evolution and genome dynamics.
- Identifying pseudogenes involves intensive data mining to locate gene-like sequences and analysis to establish whether they function.
- Recent evidence of activity among pseudogenes, and their potential resurrection, suggests some are not entirely dead after all

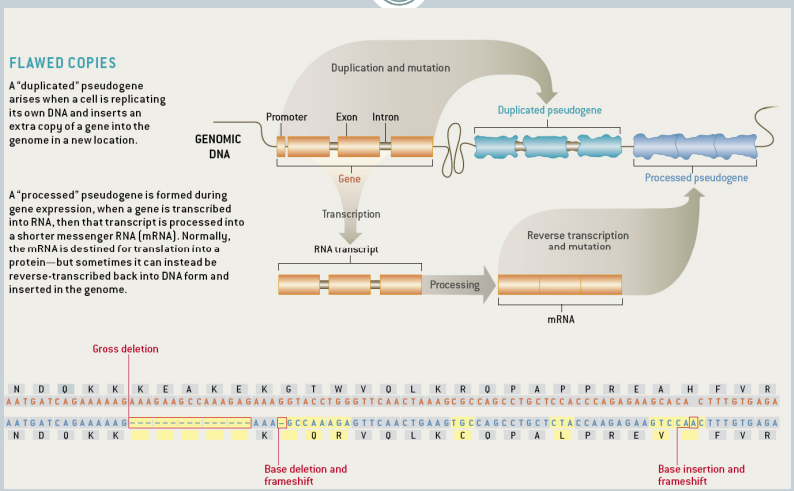
Gerstein, M & Zheng, D. The real life of pseudogenes. Sci Am 295: 48-55 (2006).

Birth of Pseudogene

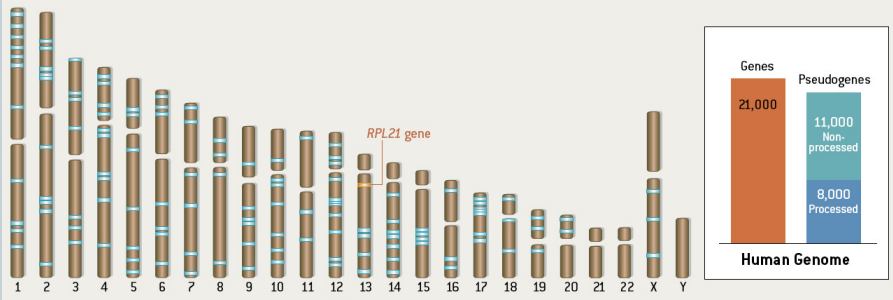


Gerstein, M & Zheng, D. The real life of pseudogenes. Sci Am 295: 48-55 (2006).

Birth of Pseudogene



Human Pseudogenome



PSEUDOGENE DESCENDANTS (*blue*) of the ribosomal protein gene *RPL21* (*orange*) are scattered across the human chromosomal landscape. Overall distribution of pseudogenes in the human genome appears to be completely random, although some local genome regions tend to contain more pseudogenes. Those DNA regions may be analogous to certain geochemical environments that better preserve mineral fossils. Identification of genes and pseudogenes is an ongoing process, but to date more than 19,000 pseudogenes have been identified in the human genome—only slightly less than the current tally of around 21,000 human genes (*inset*). About 8,000 of our pseudogenes are processed; the rest include duplicated pseudogenes and other nonprocessed subcategories.

Gerstein, M & Zheng, D. The real life of pseudogenes. Sci Am 295: 48-55 (2006).

Pseudogenomics

- Hints about Life Histories
  - Often, genes involved in an organism's response to its environment are subject to extensive duplication and diversification over time, leading to large gene families, such as the olfactory receptor repertoire.
  - Many dead-on-arrival pseudogene copies are an immediate byproduct of this process.

Gerstein, M & Zheng, D. The real life of pseudogenes. Sci Am 295: 48-55 (2006).

Pseudogenomics

- But the subsequent death of additional duplicates, which gives rise to new pseudogenes, is also frequently connected to changes in an organism's environment or its circumstances.
- Consequently, differences in the pseudogenes of animals offer hints about their diverse life histories that are not as easily detected in comparisons of working genes, which are strongly constrained by their function.

Gerstein, M & Zheng, D. The real life of pseudogenes. Sci Am 295: 48-55 (2006).

Pseudogenomics

- May not be totally dead genes
  - A few pseudogenes appear to be better preserved than one would expect if their sequences were drifting neutrally.
  - Recent experiments by Thomas Gingeras of Affymetrix and by Michael Snyder of Yale University have found that a significant fraction of the intergenic regions in the human genome are actively transcribed.

Gerstein, M & Zheng, D. The real life of pseudogenes. Sci Am 295: 48-55 (2006).

Pseudogenomics

- Pseudogene conversion
  - Some evidence also exists for the possibility of pseudogene resurrection — a dead gene turning back into a living one that makes a functional protein product.
  - Careful sequence comparisons have shown that one cow gene for a ribonuclease enzyme was a pseudogene for much of its history but appears to have been reactivated during recent evolutionary time

Gerstein, M & Zheng, D. The real life of pseudogenes. Sci Am 295: 48-55 (2006).

Pseudogenomics

- Improves gene annotation
  - Early efforts to catalogue pseudogenes were largely driven by the need to distinguish them from true genes when annotating genome sequences.
  - Identifying pseudogenes is not as straightforward as recognizing genes
  - Establishing a suspected pseudogene's inability to function is more challenging.

Gerstein, M & Zheng, D. The real life of pseudogenes. Sci Am 295: 48-55 (2006).

PseudoPipe

- An automated pseudogene identification pipeline
  - identify all the regions in the genome that share sequence similarity with any known protein, using BLAST
  - resolve the paternity ambiguity of the pseudogenes, i.e. determine among the paralogous query proteins which one most likely gave rise to the pseudogene.

Zhang, Z, Carriero, N , Zheng, D, Karro, J, Harrison, PM & Gerstein, M. PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics 22: 1437-9 (2006).

Pseudogene Classification

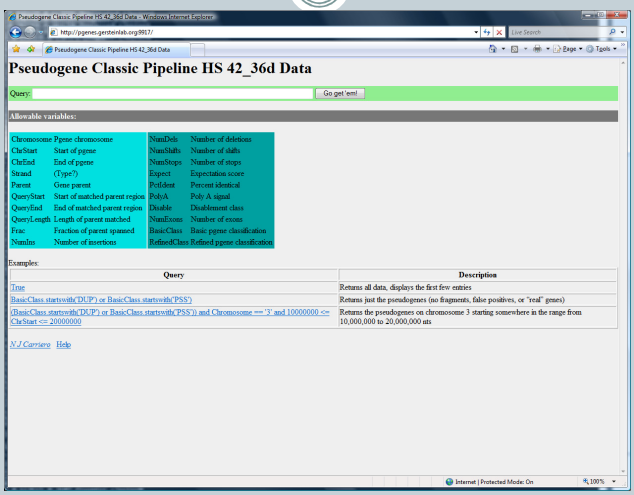
- Processed Pseudogenes
  - Lack introns, have small flanking direct repeats and a 3' polyadenine tail
- Duplicated Pseudogenes
  - Distinguished from processed pseudogenes by a combination of these features, with the emphasis on the evidence of ancient introns
- Pseudogene Fragments
  - are protein/chromosome homologies that have high sequence similarity, but are too decayed to be reliably assessed as processed or duplicated (i.e. <70% coverage of the parent gene)

Zhang, Z, Carriero, N , Zheng, D, Karro, J, Harrison, PM & Gerstein, M. PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics 22: 1437-9 (2006).

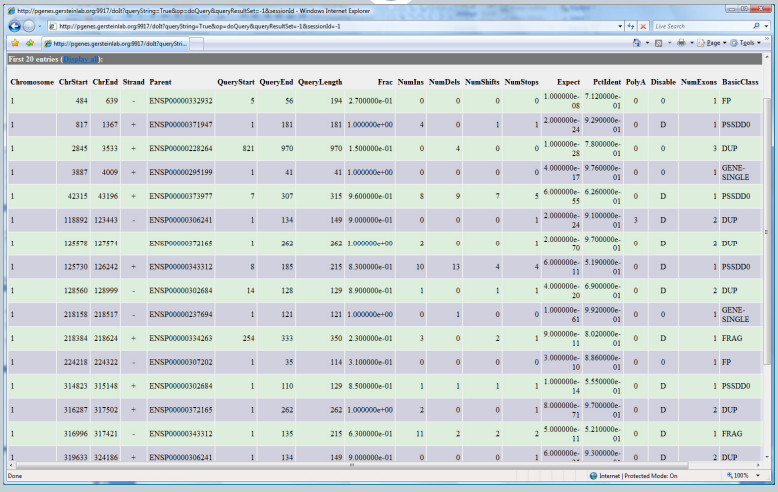
Pseudogene Data Set

SECTION TWO

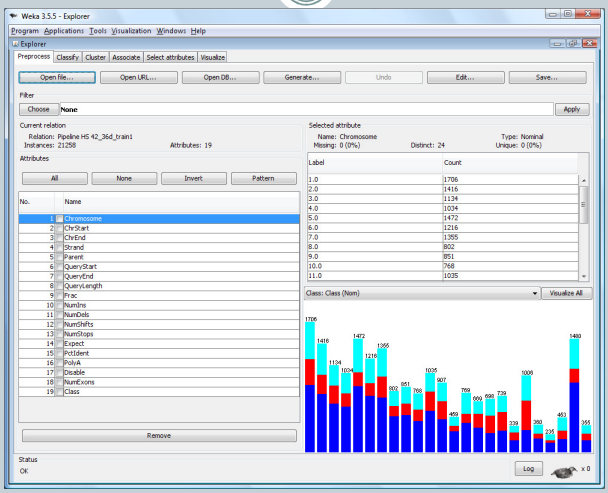
Pseudogene Dataset



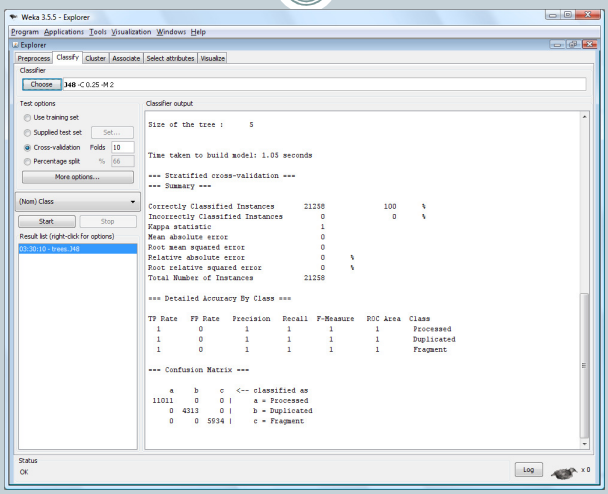
Pseudogene Dataset



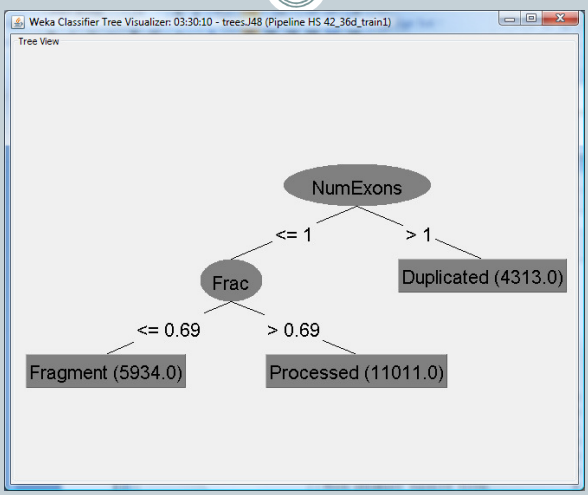
Weka: Dataming Tool



Reproduce the Model



The Pseudogene Classifier



Question

- Is there any way to classify the pseudogenes into processed or duplicated pseudogenes even they are pseudogene fragments that do not have enough information to make a significant conclusion just based on the classification model?

Solution

- Use data mining techniques
  - Cluster the data into two groups
  - Label the clusters as processed and duplicated pseudogenes
  - Remap the pseudogene fragments to the clusters
- Clustering Technique
  - K-means

The Clustering

SECTION THREE

K-means

- K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem.
- The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters)
- The main idea is to define k centroids, one for each cluster

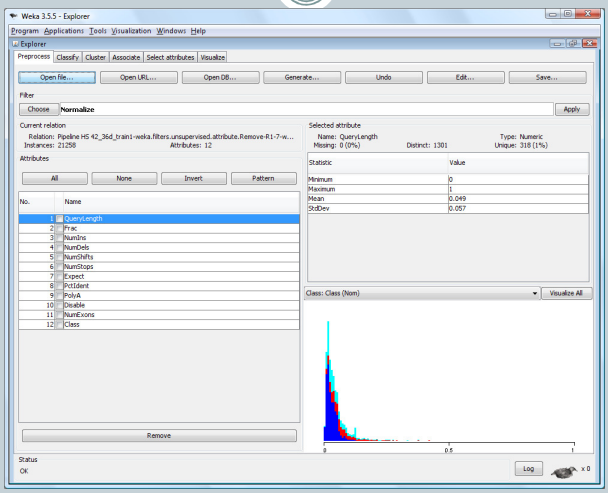
http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/kmeans.html

K-means Algorithm

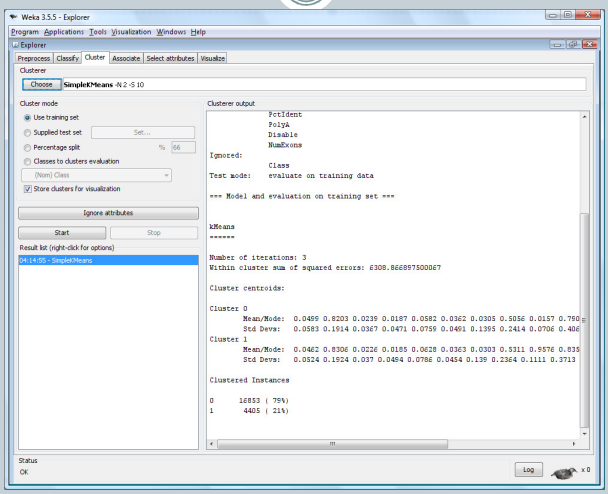
1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 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.

http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/kmeans.html

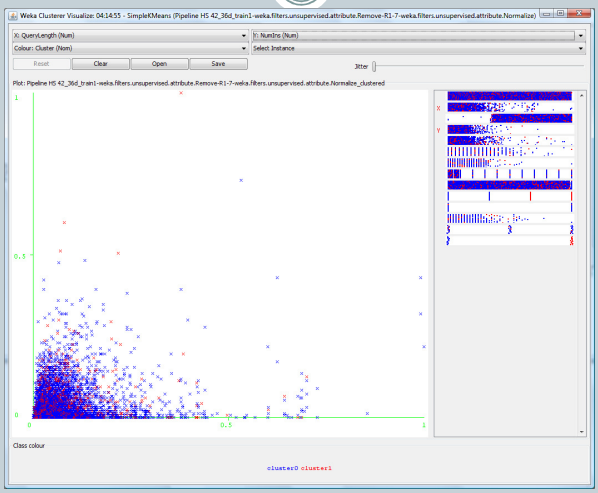
Data Pre-Processing



K-means Clustering



Clustering Result



K-means with PCA

- A weakness, which is common to clustering in general, concerns the visualization of the obtained clusters
- A possible solution is to preprocess the data using PCA
  - the PCA procedure is applied to the data. Using the principal components the data is mapped into the new feature space
  - the k-means algorithm is applied to the data in the feature space. The final objective is to be better able to distinguish the different clusters

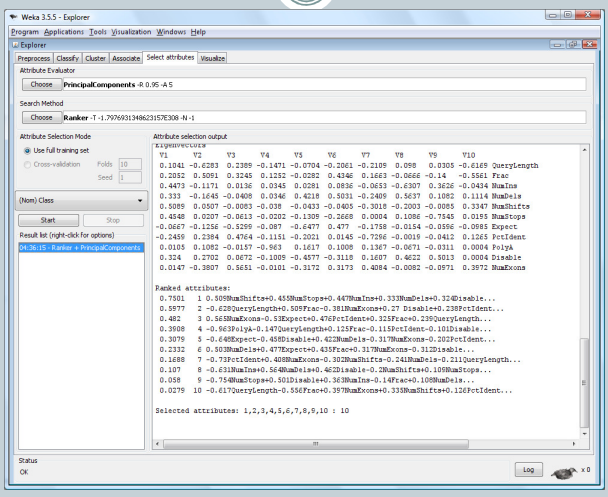
http://dataminingresearch.blogspot.com/search/label/PCA

Principle Component Analysis

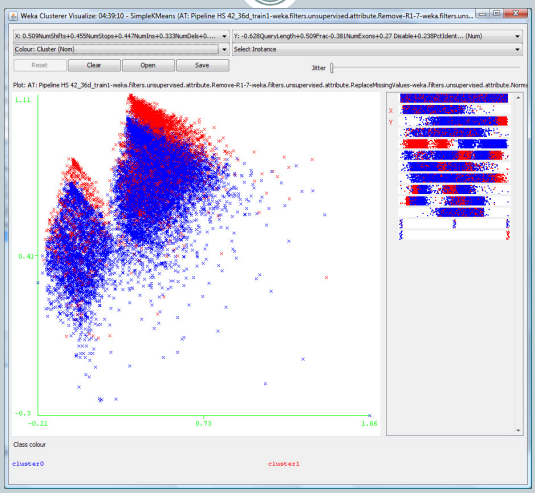
- PCA transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.
- PCA can be used for dimensionality reduction in a data set while retaining those characteristics of the data set that contribute most to its variance, by keeping lower-order principal components and ignoring higher-order ones. Such low-order components often contain the "most important" aspects of the data.

http://en.wikipedia.org/wiki/Karhunen-Lo%C3%A8ve_transform

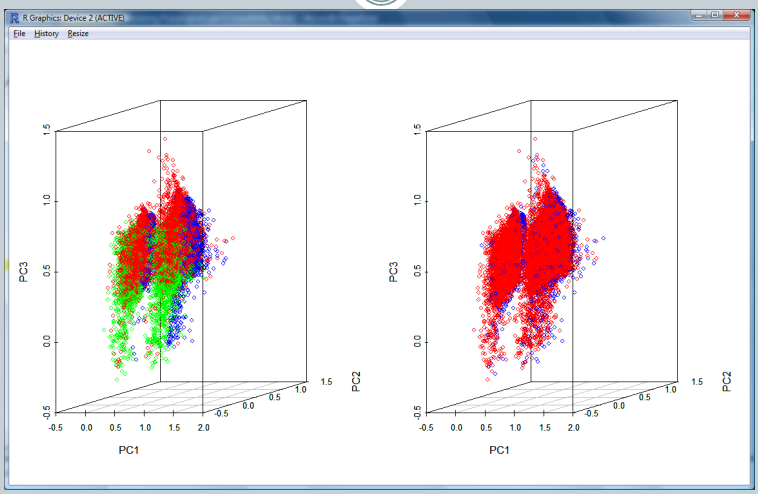
PCA Attribute Selection



Clustering PCA Result



A 3D View of the Clusters



Work to do

- Try more different parameters and clustering algorithms
- Relabel the pseudogene fragments according to the clusters
- Calculate the error rate
  - Use the training data set with just the processed and duplicated pseudogenes as a test set
  - Use the test set to evaluate the clustering
  - Calculate the error rate and plot the ROC curve