12-2015

APPLY DATA CLUSTERING TO GENE EXPRESSION DATA

Abdullah Jameel Abualhamayl Mr.
California State University, San Bernardino, 004776875@coyote.csusb.edu

APPLY DATA CLUSTERING TO GENE EXPRESSION DATA

A Project

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Computer Science

---

by

Abdullah Jameel Abualhamayl

December 2015

APPLY DATA ANALYSIS TECHNIQUE TO GENE EXPRESSION DATA

A Project

Presented to the

Faculty of

California State University,

San Bernardino

---

by

Abdullah Jameel Abualhamayl

December 2015

Approved by:

Dr. Haiyan Qiao, Adviser, Computer Science

Dr. Owen J. Murphy, Committee Member

Dr. George M. Georgiou, Committee Member

ABSTRACT

Data clustering plays an important role in effective analysis of gene expression. Although DNA microarray technology facilitates expression monitoring, several challenges arise when dealing with gene expression datasets. Some of these challenges are the enormous number of genes, the dimensionality of the data, and the change of data over time. The genetic groups which are biologically interlinked can be identified through clustering. This project aims to clarify the steps to apply clustering analysis of genes involved in a published dataset. The methodology for this project includes the selection of the dataset representation, the selection of gene datasets, Similarity Matrix Selection, the selection of clustering algorithm, and analysis tool. R language with the focus of Kmeans, fpc, hclust, and heatmap3 packages in R is used in this project as an analysis tool. Different clustering algorithms are used on Spellman dataset to illustrate how genes are grouped together in clusters which help to understand our genetic behaviors.

ACKNOWLEDGEMENTS

I would like to thank all people who supported this work. Special thanks to Dr. Haiyan Qiao for help and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION TO GENE EXPRESSION ANALYSIS

Overview

The need for analyzing data becomes so important day after day. It can be defined as the procedures of studying and inspecting data in order to extract and highlight any important information that can lead us to proper conclusions which help in supporting decision making [1]. It can be used in science, business, and social science domains. In business, many companies consider data analysis as a crucial part in streamline operations, business decisions, customer engagement, and improve production [2]. In biology, and particularly in gene data, the necessity of analyzing gene data becomes more important because it can find cures for human beings when we analyze the data in genes. Unfortunately, dealing with gene expression data is so complicated due to many reasons such as; the gigantic number of genes in the dataset and the difficulty of biological systems [3].

One important technique for gene expression analysis is clustering. It does prove its effectiveness on analyzing many gene expression data on several studies [4]. Clustering in general implies unsupervised techniques for multivariate data analysis which put them into groups on the basis of similarity score. It can be considered as a one stage for data analysis. Cluster analysis is preferred for the comprehension of expression level of multiple genes simultaneously through

a microarray data. The functions of unknown genes can be determined from clues obtained from gene clusters that have similar expression level in different samples. Determination of expression levels of multiple genes manifest pathological pathways through which they induce disease and it facilitates defining new subclasses of a disease [5]. Clustering technique reduce the dataset by providing a visual gene expression pattern. Depending on the type of multivariate dataset, the clustering technique can be selected from the library already established for this purpose. Several clustering algorithms can be applied on gene expression dataset such as; hierarchical clustering, K-means clustering, and fuzzy clustering.

Significance

Gene expression analysis has considerable importance in medical sciences. The work of our biological system is still a mystery. We need to understand the gene expression data and what it does imply. By analyzing gene expression, we can unravel functions of unknown genes and their participation in cellular processes. Through this, we can prevent diseases and save more lives.

Another advantage of analyzing gene expression is that it provides information regarding gene expression regulation, translated product interaction with other genes or genes encoding it. In different cells, expression levels of a particular gene differ significantly and this information can also be obtained

through clustering of gene expression datasets. Above all, it determines alterations in gene expression in normal and diseased tissues.

CHAPTER TWO

INTRODUCTION TO R LANGUAGE

Using R Language in Data Analysis

One of the most useful programming language in terms of data mining and clustering gene expression dataset is R language. It is a scripting language that allows users to contribute in its enormous repository. R language has proved its capabilities as a powerful analysis tool [6]. It has so many advantages besides it is an open source programing language such as; the Comprehensive R Archive Network CRAN packages which is a repository that has over than 7000 contributed packages from its active users; moreover, it can integrate with other programming languages such as; Java, C++, and Python. Another advantage is the graphical capability for R. All of these are in free software which enlarges the popularity of R community. Due to these advantages, new idea and technology appears first in R.

R Packages

In R, a package refers to a collection of compiled code and functions that made to serve an intended purpose. As it is mentioned before, R has around 7000 packages. Those packages can be downloaded from the CRAN repository. Since active users are prone to make mistakes more than professionals, every new contributed package is tested on several servers and operating systems.

When installing R, the language comes with several packages that are ready to use. Others can be installed from the CRAN repository by adding a line of code following this format:

install.packages("package name", dep = TRUE)

More information about installing R langue and the packages can be found on: https://cran.r-project.org/

It is worthy to mention that the National Center for Biotechnology Information NCBI provides an enormous amount of gene expression datasets. Researchers in biomedicine and biotechnology use the NCBI website when they look for a dataset in this field. All those datasets are available online in their website.

CHAPTER THREE

CLUSTERING GENE EXPRESSION DATA

Difficulties in Clustering Gene Expression Data

Gene expression datasets are Big Data which is a well-known term for datasets that very complex or large which make it difficult to process using traditional software techniques.

Table 1. A Sample of Spellman Dataset

time	40	50	60	70	80
YAL001C	-0.07	-0.23	-0.1	0.03	-0.04
YAL014C	0.215	0.09	0.025	-0.04	-0.04
YAL016W	0.15	0.15	0.22	0.29	-0.1
YAL020C	-0.35	-0.28	-0.215	-0.15	0.16
YAL022C	-0.415	-0.59	-0.58	-0.57	-0.09
YAL036C	0.54	0.33	0.215	0.1	-0.27
YAL038W	-0.625	-0.6	-0.4	-0.2	-0.13
YAL039C	0.05	-0.24	-0.19	-0.14	-1.22

The Table 1 represents a sample of Spellman dataset which is an example for gene dataset. The original dataset has 4381 genes (rows) and 23 conditions (columns). In general, Dealing with gene expression dataset is so complicated due to many reasons such as; the enormous number of genes as presented. Moreover, each gene has several conditions and it changes with time; therefore, the numbers are not fixed (known as time series data). In addition, the complexity of our biological system makes the data sometimes exposed to modify in order to get better clustering. The dataset that has been extracted from

the DNA microarray normally contains outliers and missing values. Therefore, pre-processing dataset is fundamental for the purposes of getting clear clustering results. The problems that occur while pre-processing gene expression dataset have become an interesting subject for researchers. Some methods start pre-processing dataset by removing genes and conditions that have no effect on overall samples. Others apply technique on the matrix to set the mean of each row equal to zero and the variance to one. Above all, pre-processing the dataset must be accomplished before clustering.

There are two different methods of cluster analyses: Supervised and unsupervised analysis [7]. In supervised clustering, usually we train a clustering algorithm in order to present particular clustering patterns.

In this project, the unsupervised analysis will be applied. This type of clustering usually doesn't involve any other variable, which is regarded as external variable for obtained gene expression data. It can be used to group genes with similar function together.

Steps to Cluster Gene Expression Data

The steps include Selection of the dataset representation, Selection of gene datasets, Similarity Matrix Selection, clustering algorithm, and analysis tool.

Selection of the Dataset Representation

Generally, in terms of gene expression clustering, there are three types of

data representation [3]: the first type is gene-based clustering. In this method, genes are regarded as objects, whereas samples are considered as features. By far, this method is the most used among the other three methods. Table 2 illustrates this method.

Table 2. Gene-based Clustering

Time/Genes	YAL001C	YAL014C	YAL016W	YAL020C	YAL022C	YAL036C	YAL038W	YAL039C
40	-0.07	0.215	0.15	-0.35	-0.415	0.54	-0.625	0.05
50	-0.23	0.09	0.15	-0.28	-0.59	0.33	-0.6	-0.24
60	-0.1	0.025	0.22	-0.215	-0.58	0.215	-0.4	-0.19
70	0.03	-0.04	0.29	-0.15	-0.57	0.1	-0.2	-0.14
80	-0.04	-0.04	-0.1	0.16	-0.09	-0.27	-0.13	-1.22

In Table 2, each row represents the gene name with different values that change over time. The columns represent the conditions which are the value of the gene at a particular time. We choose this type of dataset representation if the focus of study is genes and how they related to each other.

The second type is sample-based clustering which considers genes as feature and samples as objects. This type is preferable if the goal of study is to compare conditions not genes.

Table 3. Sample-based Clustering

Time/Genes	YAL001C	YAL014C	YAL016W	YAL020C	YAL022C	YAL036C	YAL038W	YAL039C
40	-0.07	0.215	0.15	-0.35	-0.415	0.54	-0.625	0.05
50	-0.23	0.09	0.15	-0.28	-0.59	0.33	-0.6	-0.24
60	-0.1	0.025	0.22	-0.215	-0.58	0.215	-0.4	-0.19
70	0.03	-0.04	0.29	-0.15	-0.57	0.1	-0.2	-0.14
80	-0.04	-0.04	-0.1	0.16	-0.09	-0.27	-0.13	-1.22

In Sample-based clustering, each row represents the value of the gene at a particular time whereas, the columns denote the gene names with different values over time.

The third type is called subspace clustering. This is a modern technique that focuses on subsets formed by genes and samples. The clustering will be performed on only some subsets of the whole gene expression dataset. In this technique, genes and samples can be treated as objects or features. Applying such technique requires deep knowledge of the dataset to determine which genes and samples are appropriate for particular experiment. For instance, we can study only a subset of genes and conditions that presented in Table 1.

Table 4. Subspace Clustering

Time/Genes	50	60	70
YAL016W	0.15	0.22	0.29
YAL020C	-0.28	-0.215	-0.15
YAL022C	-0.59	-0.58	-0.57
YAL036C	0.33	0.215	0.1

In general, a careful consideration should be made before selecting the data representation. It is essential to determine what the purpose of the study is and which data representation serves this purpose. A prior knowledge of genes and conditions is required some time which skills that a computer scientist might lack.

Selection of Gene Dataset

A lot of gene expression datasets need to be pre-processed to remove outliers, missing values, and any unreliable data before applying analysis. Neglecting per-processing the dataset could lead to misleading results. Genes having uncertain values should be removed from the dataset as they produce low signals when compared with noise. It's not appropriate to exclude distances which are not completely accurate. Genes that produce a signal that lies in the background noise range result in increasing noise to clustering; therefore, it is desirable to exclude such signals. In certain type of clustering, experts in gene expression dataset may choose to exclude some genes, conditions, or even both (subspace clustering). Doing that requires deep knowledge in the dataset they opt to work on. By excluding outliers and selecting an appropriate number of effective genes and conditions, we can acquire better and more reliable clustering than selecting the whole dataset.

In this project, I aim to apply different data clustering on Spellman dataset [4]. A small part of the original dataset is presented in Table1.

Similarity Matrix Selection

We use the similarity matrix to compute proximity which is provided to clustering algorithm for instance similarity or dissimilarity scores and distances between two genes or samples. To define these parameters we need the following conditions [8]:

d (i, j) > 0

This implies that the distance between two parameters should be positive or zero

d (i, i)=0

This implies the distance between any point and itself is zero

d (i, j) = d (j, i)

This implies that the distance between i and j is the same between j and i

Majority of cluster programs have a computational model that measures distances. Mainly, there are three widely used methods which are: Euclidean distance, Manhattan distance, and Pearson's Correlation .The first one is the Euclidean distance. It is very common to use this matric. It can be measured through the following equation:

$$d(i,j) = \sqrt{\sum_{k=1}^{n} (Xik - Xjk)^2}$$

In some cases, for clustering gene expression data, standardization of the expression is performed prior to further calculation through the following equation:

$$dSE(i,j) = \sqrt{\sum_{k=1}^{n} \frac{1}{s_p^2}[(Xik - Xjk) * (Xik - Xjk)]}$$

Where s_p^2 represents the pooled variance which is a way to estimate variance among different populations that have different mean. The pooled variance s_p^2 can be calculated in indexed populations as follow:

$$S_P^2 = \frac{\sum_{i=1}^{K}(n_i - 1)s_i^2}{\sum_{i=1}^{K}(n_i - 1)} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2}{n_1 + n_2 + \cdots + n_k - k}$$

Alternatively, we can use the range instead of the pool variance as follow:

$$dSER(i,j) = \sqrt{\sum_{k=1}^{n} \frac{1}{R^2}[(Xik - Xjk) * (Xik - Xjk)]}$$

Where R indicates range.

Another famous matric is the Manhattan distance. It is defined as the sum of linear distances. It is also known as 'city block' distance. More specifically:

$$dman = \sum_{k=1}^{n} |(Xik - Xjk)|$$

All the method mentioned previously fall in class of Minskowski Distances. Mathematically expressed as follows:

$$d(i,j) = \left[\sum_{k=1}^{n} |Xik - Xjk|^{q}\right]^{\frac{1}{q}}$$

Correlation matrix can also be used for the calculation. In this method of proximity, similarity scores are considered and it is employed when expression profiles show a comparative measure of expression levels. In such instances, a distance is measured based on similarity scores of different genes. One of the common types of correlation is Pearson's Correlation, which can be calculated as follows:

$$p(i,j) = \frac{\sum_{k=1}^{n}[(Xik - \bar{X}i) * (Xik - \bar{X}j)]}{[\sum_{k=1}^{n}(Xik - \bar{X}i)^2 * \sum_{k=1}^{n}(Xik - \bar{X}j)^2]^{\frac{1}{2}}}$$

Where $-1 \leq p(i,j) \leq 1$

Transformations can be represented by two forms. One form is the following:

$$d(i,j) = \frac{\left(1 + p(i,j)\right)}{2}$$

Another form for Transformations Pearson's Correlation is:

$$d(i,j) = |p(i,j)|$$

Based on the form we used, we can convert similarity to dissimilarities through following expression:

$$\mathrm{dis}(i,j) = \frac{\left(1 - s(i,j)\right)}{2}$$

$$\mathrm{dis}(i,j) = 1 - |s(i,j)|$$

Clustering Algorithm

There are different clustering algorithms which are chosen on the basis of cluster analysis applications. For example, filtering, agglomerative hierarchical clustering, divisive hierarchical clustering, Ward's method and Partitioning method. All these algorithms perform clustering on the basis of similarity scores and remove elements which are entirely dissimilar. It is notable that using different clustering methods lead to different results. The widely used algorithms for biological samples are agglomerative hierarchical clustering. Partitioning involves specified number of objects in a given set of clusters.

In terms of the clustering number, several ways might be applied to verify the appropriateness of any given clustering. For example, the homogeneity quantity looks for several values of K first, then, it searches for a leveling off. The formula below determines the average distance between each observation and its center:

$$H_{avg,k} = \frac{1}{N_{g,k}} \sum_{i=1}^{N_{g,i}} d\left(X_{i,k}\ C(X_{i,k})\right)$$

Where C is the number of given cluster.

Another way is to use the separation quantity, in this method, the weighted average distances between the clusters are calculated by:

$$S_{avg} = \frac{1}{\sum_{k=1} N_k N_i} \left(\sum_{k=1} N_k N_i d(C_k, C_i) \right)$$

Analysis Tool

In order to cluster gene expression data, we need a powerful software tool. It is extremely difficult to apply data clustering to gene expression dataset manually. In this project, I used R language which One of the most useful programming language in terms of data mining and clustering gene expression dataset as an analysis tool. R language has been introduced in chapter two on this dissertation.

With the intention of clustering Spellman dataset, I aim to use R language four packages which are: Kmeans, fpc, hclust, and heatmap3 packages. Kmeans and hclust packages are already built-in R, whereas fpc and heatmap3 are required to be installed through the CRAN repository.

Apply Data Clustering on a Gene Expression Dataset

Clustering has proved to be effective for analyzing a gene expression dataset [4]. Here. I will apply the steps I introduced previously to cluster gene expression dataset. I will assume that we want to use gene-based clustering where genes are objects and samples are features. A Spellman dataset will be used here for this experiment. It has been introduced previously in this project. In short, the dataset has 4381 genes and 23 conditions.

The K-means package that is built in R language uses the algorithm of [9]. This algorithm uses the Euclidean distance as a matrix system to calculate the distance between genes. Figure 1 explains the whole process.

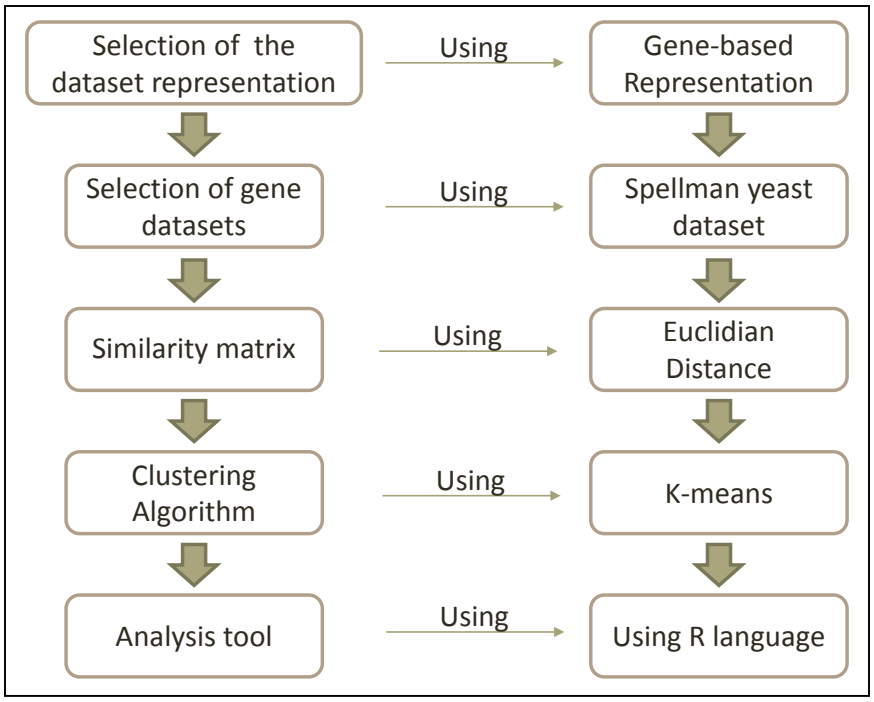


Figure 1. The Process to Cluster Gene Expression Dataset Using K-means Algorithm.

With the intention of visualizing the clustering result using K-means algorithm, I installed fpc package which allows us to cluster objects using K-means algorithm and plot it. Figure 2 shows the result of clustering the whole Spellman dataset.

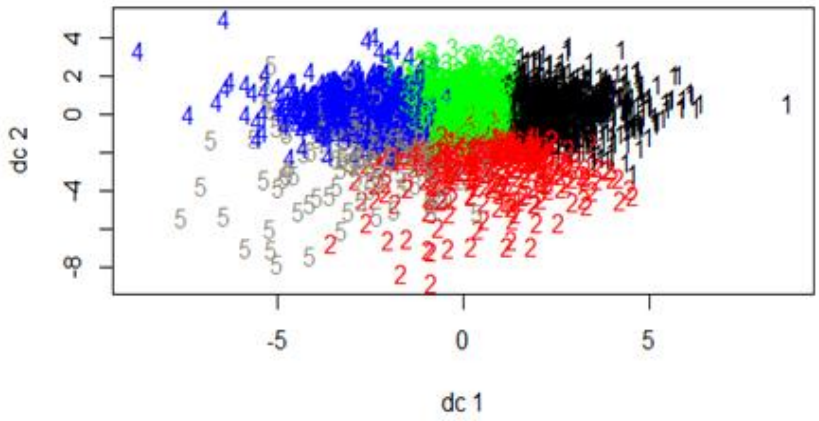


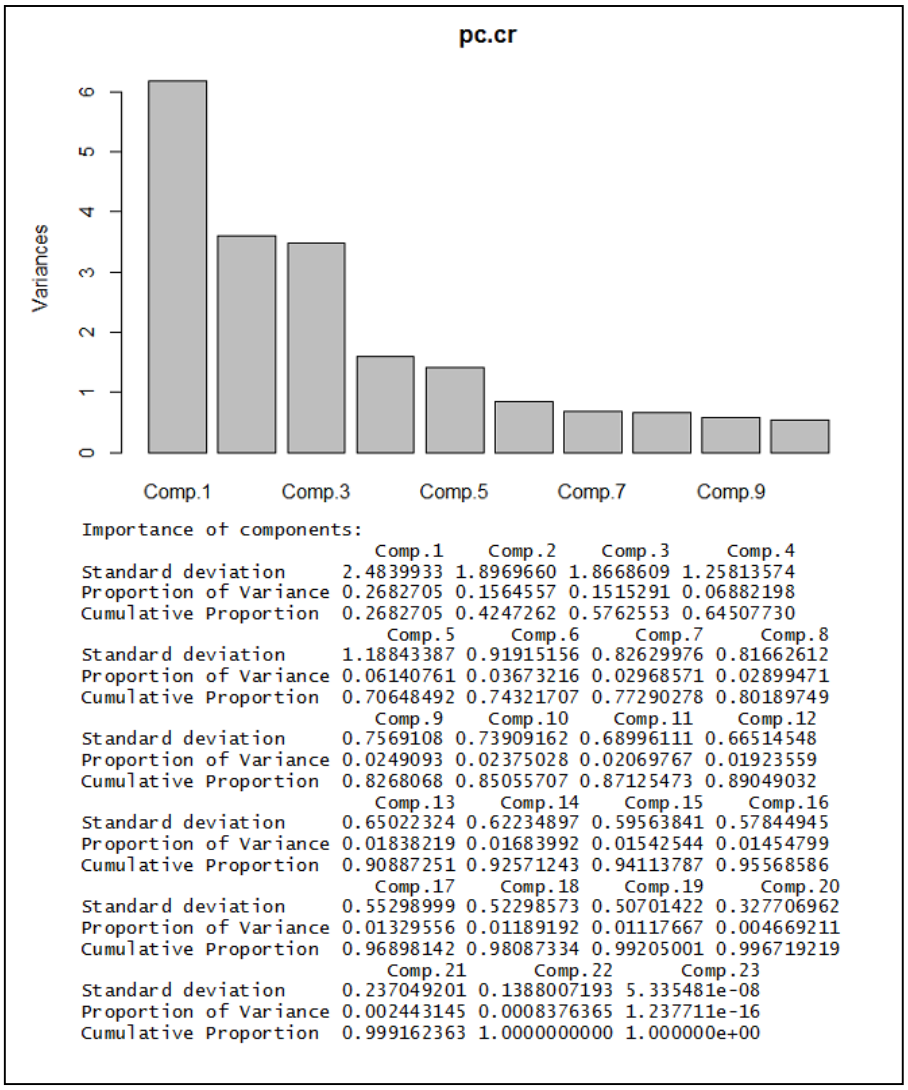
Figure 2. Clustering Spellman Dataset Using Kmeans and fpc Packages.

In Figure 2, I applied k-means clustering algorithm on Spellman dataset using Kmeans package and fpc package. K-means algorithm allows us to partition the genes into K clusters. The Euclidean distance formula is used in kmeans package. I used fpc package visualizes the clustering result as it is presented in Figure 2.

A gene can be assigned to a certain cluster in the first stage. And once the reevaluation for all genes is done, that gene might be grouped in a different cluster. One drawback for K-means algorithm is the lack of capability to handle outliers.

In order to reduce the dimensionality of Spellman dataset, kmeans package uses a principal components analysis (PCA) [10]. Using the PCA helps to plot the whole dataset as a two or three dimensional space. The fpc package plots the first two principal componant for each gene. One drawback for the PCA technique is that it does not perform well if it is applied on high dimensional data

such as Spellman dataset. Figure 3 shows the importance of each principle component with a summary.



Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	2.4839933	1.8969660	1.8668609	1.25813574
Proportion of Variance	0.2682705	0.1564557	0.1515291	0.06882198
Cumulative Proportion	0.2682705	0.4247262	0.5762553	0.64507730

	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	1.18843387	0.91915156	0.82629976	0.81662612
Proportion of Variance	0.06140761	0.03673216	0.02968571	0.02899471
Cumulative Proportion	0.70648492	0.74321707	0.77290278	0.80189749

	Comp.9	Comp.10	Comp.11	Comp.12
Standard deviation	0.7569108	0.73909162	0.68996111	0.66514548
Proportion of Variance	0.0249093	0.02375028	0.02069767	0.01923559
Cumulative Proportion	0.8268068	0.85055707	0.87125473	0.89049032

	Comp.13	Comp.14	Comp.15	Comp.16
Standard deviation	0.65022324	0.62234897	0.59563841	0.57844945
Proportion of Variance	0.01838219	0.01683992	0.01542544	0.01454799
Cumulative Proportion	0.90887251	0.92571243	0.94113787	0.95568586

	Comp.17	Comp.18	Comp.19	Comp.20
Standard deviation	0.55298999	0.52298573	0.50701422	0.327706962
Proportion of Variance	0.01329556	0.01189192	0.01117667	0.004669211
Cumulative Proportion	0.96898142	0.98087334	0.99205001	0.996719219

	Comp.21	Comp.22	Comp.23
Standard deviation	0.237049201	0.1388007193	5.335481e-08
Proportion of Variance	0.002443145	0.0008376365	1.237711e-16
Cumulative Proportion	0.999162363	1.0000000000	1.000000e+00

Figure 3. The Importance of Each Principle Component with a Summary.

According to Figure 3, Spellman dataset lose some information whenever we reduce the dimension of the dataset. This information might be valuable. The first two principle component explains 42.75% of the point variability. We need to maintain this percentage to be as larger as possible. It is worth to mention that using the first 8 principle component expresses 80.18% of the data. However, it is difficult, if not impossible, to plot an eight-dimensional dataset.

Although all genes on Figure 2 are so close to each other, overlapping clustering is not supported in kmeans package. Each gene is supposed to have only one center. Each number on Figure 2 representing the first two principal axes for each gene which in total are 4831 genes. Using fpc package to draw the k-means result of Spellman dataset makes the data appears overlapped. That happens because the plot is two-dimensional space.

In an attempt to illustrate Figure 2, I plotted Figure 4 and Figure 5 which are Spellman dataset in three-dimensional space instead of two-dimensional space and the result confirms that there is no an overlap among the data. It is worth to mention that using the first three principle components illustrates 57.90% of the data which is an expectable rate compared to using the first two components.

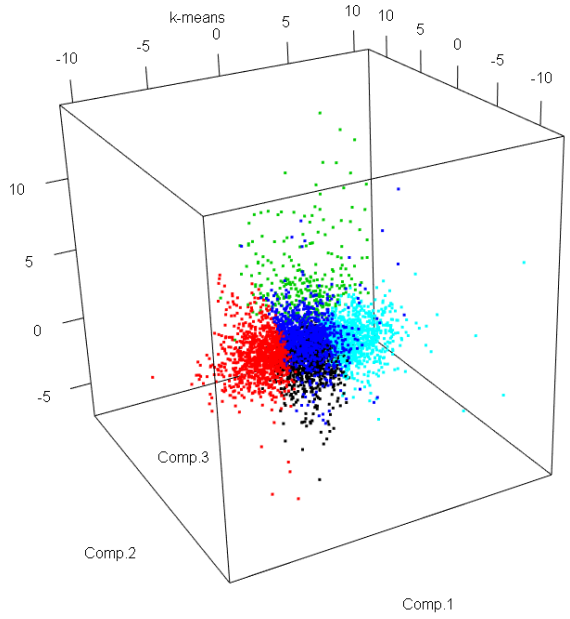


Figure 4. Plotting the Kmeans Package Result in 3D-Part A.

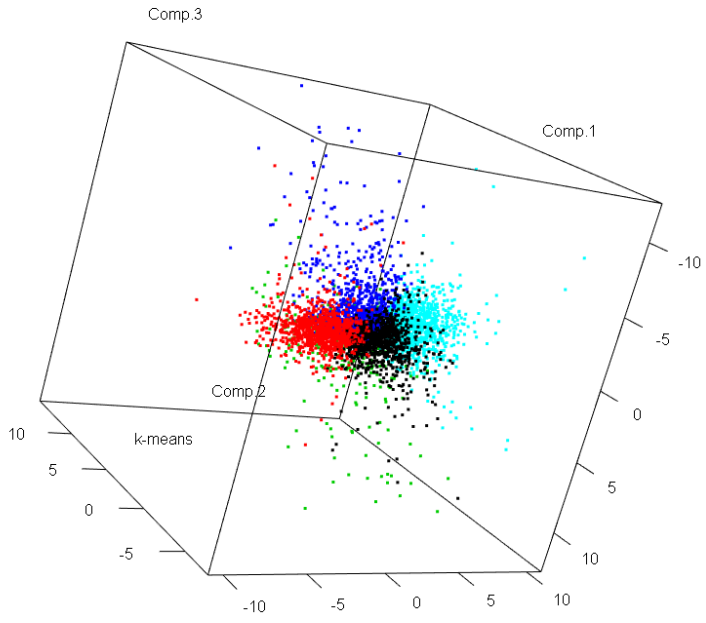


Figure 5. Plotting the Kmeans Package Result in 3D-Part B.

Using kmeans package and fpc package deliver two-dimensional plot. However, drawing clear lines to distinct the regions of each cluster in is impossible in Spellman dataset. By using fpc package, the conductor of cluster analysis would receive general suggestions to the possible relations among genes in a cluster and the general structure, but if the focused of the study is to get a visual vision of each cluster, using Kmeans and fpc backage would not give a clear clustering suggestions. Another disadvantage for using these packages on Spellman is the high rate of losing data by selecting the first two principle components.

The third package I applied on Spellman dataset is hclust package. This package allows us to apply hierarchical clustering on the selected dataset.

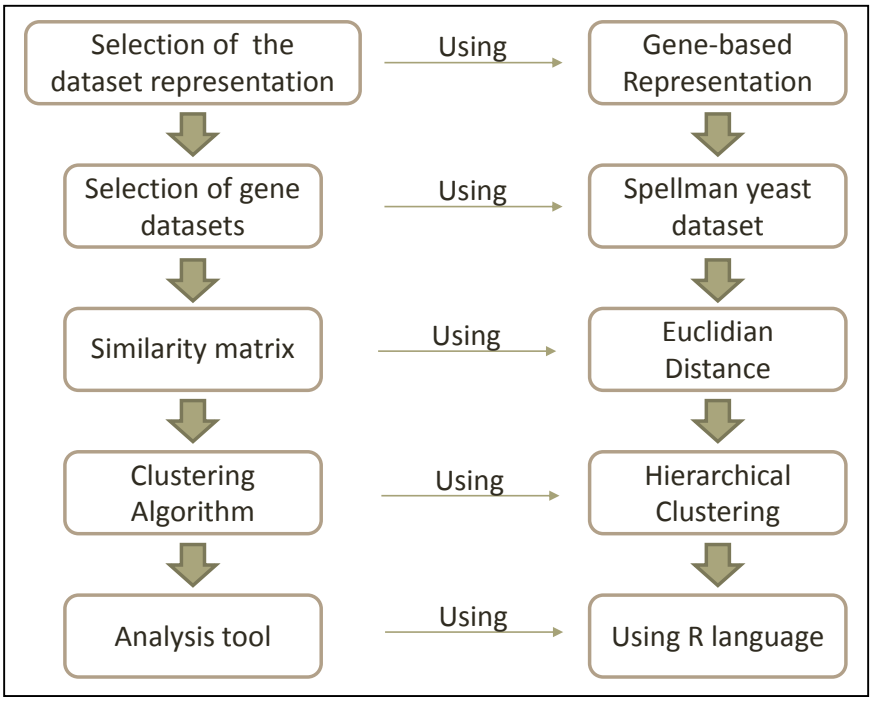


Figure 6. The Process to Cluster Gene Expression Dataset Using Hierarchical Clustering.

Using the hierarchical clustering helps to determine the appropriate number of the K clusters. I will illustrate that later on in this report. Hierarchical clustering can handle outliers which is an advantage that it has over K-mean algorithm.

Inside the package, there are 8 implemented algorithms which are: single linkage, complete linkage, average linkage, ward D, mcquitty, median, centroid, and ward D2. All these algorithms present similar result [11]. In Figure 7, ward D2 is used. Ward D2 uses Euclidean distance formula to compute the dissimilarity among genes in cluster. Using this package, Genes are assigned to

its own cluster first. Then, the most two similar clusters are merged into a new cluster. The distance matrix is updated after forming any cluster. This step goes on until all clusters are agglomerated into one cluster.

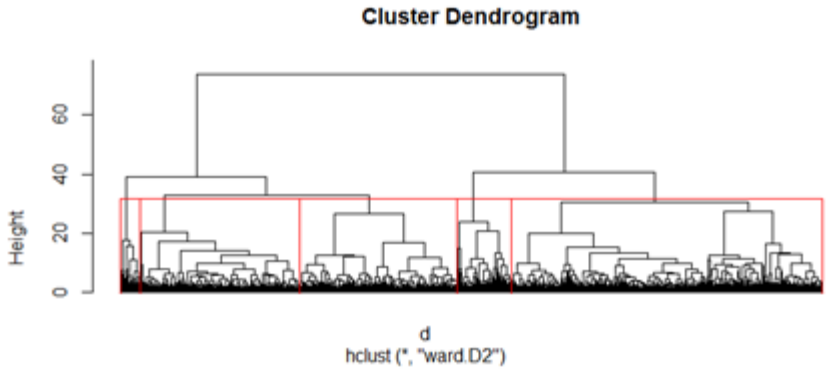


Figure 7. Clustering Spellman Dataset Using hclust Package.

In the diagram, the y-axis indicates how closely the genes were when they grouped together in clusters. Consequently, the heights of tree branches increase from the end of leaves which is zero to the last merging in Figure 7 which is around 76.

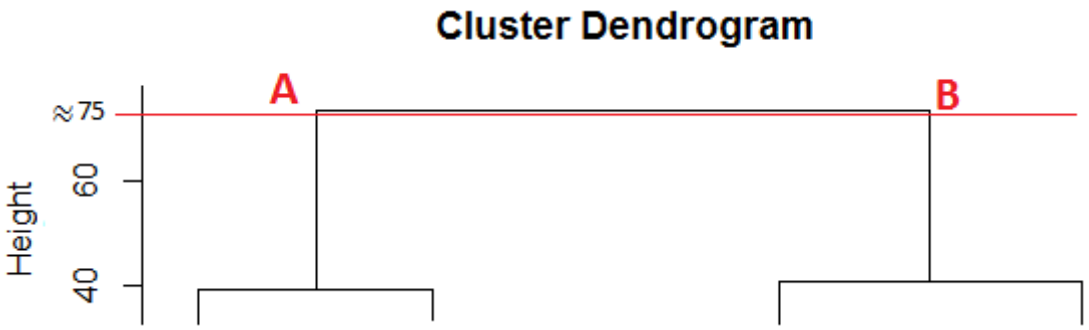


Figure 8. The Upper Part of the Cluster Dendogram.

To illustrate the idea, in Figure 8, the approximate height of 75 is crossing two lines form the x-axis (A and B). All genes below that are grouped together in two clusters (A and B). Clustering might be ineffective at this height.

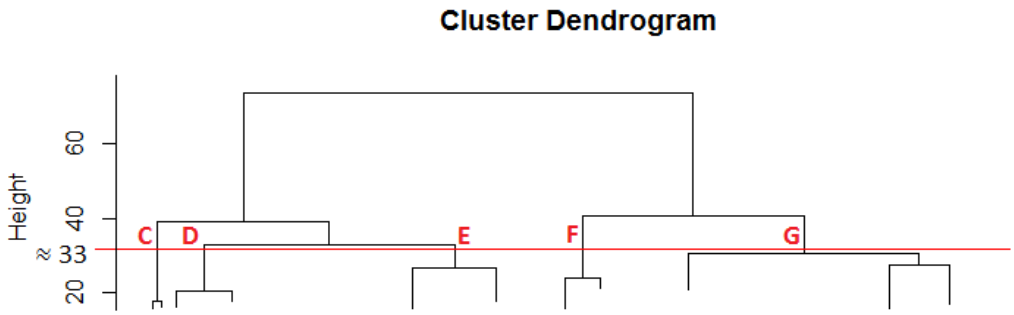


Figure 9. Determining the Number of K Cluster Based on Cluster Dendrogram.

In Figure 9, we can notice that at the approximate height of 33, the line is crossing five lines of x-axis (C, D, E, F, and G). All genes are close to each other's at that height. Consequently, the overall suggestion of Figure 9 is that five clusters is an appropriate number to group Spellman dataset. Those five clusters can be seen on red in Figure 7. Due to the high number of genes, further interpretation for the dataset becomes slightly impossible. It is worth pointing out that all branches that join a cluster in relatively high height are suspected to be outliers. For instance, the cluster on the leftmost side of dendrogram (marked as X in Figure 10) merged relatively in high height compared to the others. Consequently, excluding those genes possibly help to get better clustering result.

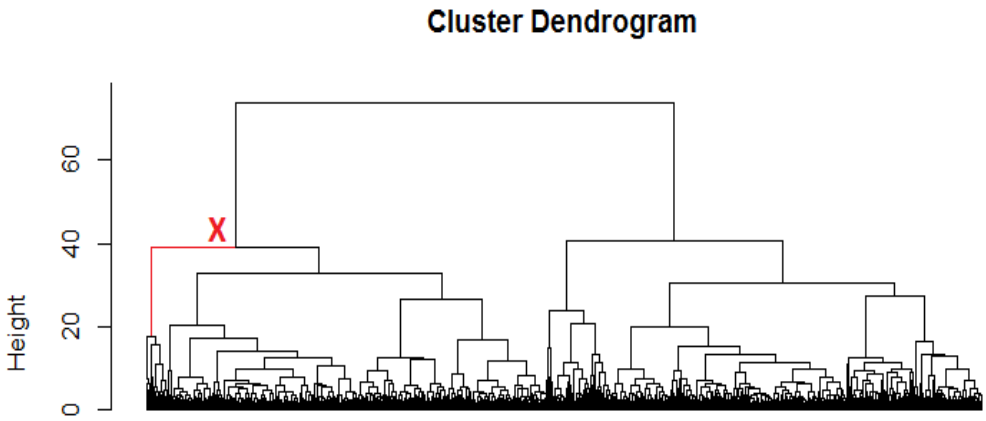


Figure 10. Potential Outliers in Cluster Dendogram.

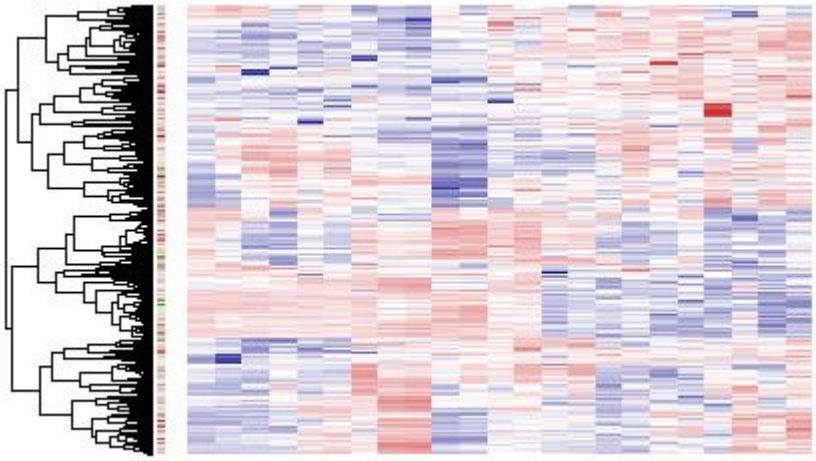


Figure 11. Using heatmap3 Package on Spellman Dataset.

The fourth package I used is the heatmap3 package. Heat map is a transformation for the numerical data in dataset to a colorful image. Since human brain recognizes colors easily compared to numbers, it is quit helpful to see the whole dataset in colors. It gives researchers to get a first impression of the

dataset. In clustering analysis, heat maps are one of the best visual tools for dense data points. In Figure 11, heat maps are easily identified when there is a very high concentration of activity. This process is sometimes called the hotspot analysis. The figure indicates that all areas in red are having similar value which impels significant cluster such as in upper right area of the figure. The same applies for the blue areas of the figure. White areas neutral values, whereas the dark blue or red areas indicates outliers. Nowadays, cluster heat map is widely used in social sciences to generate biological graphs [12]. In addition, they are the most popular representation as they compact large amount of information into a small space so bring out coherence in data.

CHAPTER FOUR

CONCLUSIONS

Applying data analysis to gene expression datasets plays an important role in medical sciences. Understanding more about genes and its participation in cellular processes can help us to prevent diseases and save lives. Clustering is an effective approach to analysis gene expression datasets. Several clustering algorithms can be applied on gene expression dataset such as; K-means clustering, hierarchical clustering, and fuzzy clustering. The enormous number of genes and its conditions supports the need of a powerful software tool such as R language to apply the clustering. R language has become one of the most powerful language in terms of data mining. Its CRAN repository has over 7000 packages which assist the work for researchers. In this project, four packages are used which are Kmeans, fpc, hclust, and heatmap3 packages. The graphical capability for R language is outstanding as well. Being free software is another advantage for R users. As a result, new technology appears first in R language.

Gene expression datasets are very complex which make it difficult to process using traditional software techniques. It is considered a time series data, which is a data that changes over time. Pre-processing the dataset to get rid of unreliable values is important to get better clustering results. In this project, a Spellman dataset is used as an example for a gene expression dataset. It has 4381 genes and 23 conditions.

There are several steps to cluster gene expression datasets which are: Selection of the dataset representation, selection of gene dataset, Similarity Matrix Selection, clustering algorithm, and analysis tool. There are three ways to represent the data: gene-based clustering, sample-based clustering, and subspace clustering. Each category defines which to consider as objects and which to consider as features. It is up to the researcher to determine the type of clustering base in the way that serves the purpose of the study.

Using fpc, hclust, and heatmap3 packages can visualize the result of clustering the dataset in R. Using fpc package to plot the result of K-means algorithm provides a two-dimensional plot which would not give clear clustering suggestions. Plotting Spellman dataset in a three-dimensional plot provides an acceptable rate of the data loss. However, using the first eight principle component save more than 80% of the data. Better suggestions regard the dataset can be obtained using hclust package. It helps to determine the appropriate number of clusters. Additionally, it assists in terms of detecting and excluding potential outliers. Drawing heat map using heatmap3 package is quit useful and gives a first impression for the whole dataset. In general, those packages assist the researchers by giving overall suggestions upon the selected dataset. However, there is a clear gap between computer science and biology. Although clustering is well-known for a computer scientist, the pre-knowledge of the gene expression datasets and how they relate to each other is still a mystery for computer scientists, and vice versa. More investigations and studies have to

be done towards a comprehensive understanding of clustering and gene expression data.

REFERENCES

[1] P. Bihani, and S.T. Patil, "A Comparative Study of Data Analysis Techniques," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),* vol.3, no.2, pp. 95-101, 2014

[2] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer, “Enterprise data analysis and visualization: An interview study,” *IEEE Trans. Visualization and Computer Graphics (TVCG),* vol.18, no.12, pp. 2917-2926, 2012.

[3] K.Y. Yeung, D.R. Haynor, and W.L. Ruzzo, “Validating Clustering for Gene Expression Data,” *Bioinformatics,* vol. 17, pp. 309-318, 2001.

[4] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, pp. 14863-14868, 1998.

[5] D.K. Slonim, “From Patterns to Pathways: Gene Expression Data Analysis Comes of Age,” *Nature Genetics Supplement,*vol. 32, pp.502-508, 2002.

[6] C. Kim, J. Hosking, and J. Grundy, "Model Driven Design and Implementation of Statistical Surveys," in *47th Hawaii International Conference on System Sciences*, Hawaii, HI, 2007, pp. 285c-285c.4

[7] A. Brazma and J. Vilo, "Gene Expression Data Analysis," *FEBS Letters,* vol. 480, pp. 17-24, 2000.

[8] M. Reimers. (2015, Februry 15). *Exploratory Analysis. An Opinionated Guide to Microarray Data Analysis*[Online]. Available: http://www.people.vcu.edu /~mreimers/OGMDA/exploratory.analysis.html.

[9] J. A. Hartigan and M. A. Wong, “Algorithm as 136: A k-means clustering algorithm,” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, pp. 100–108, 1979.

[10] M.A. Peeples. (2011, November 10). *R Script for K-Means Cluster Analysis* [Online]. Available: http://www.mattpeeples.net/kmeans.html.

[11] T. Galili, "dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering," *Bioinformatics*, 2015, doi: 10.1093/bioinformatics/btv428.

[12] L. Wilkinson and M. Friendly, The history of the cluster heat map *The American Statistician,* vol. 63, pp. 179-184, 2009