Identifying The Best Metrics To Find The Best Quality Clusters Of Genes From Gene Expression Data

By

Joydhriti Choudhury
15301125
Tanzima Rahman Roshni
15301111
Md. Tawhidul Islam Chowdhury
16101321
Raihanoor Reza Rayon
18141021

A thesis submitted to the Department of Computer Science and Engineering in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
April 2019

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name Signature:

Joydhriti Choudhury 15301125	Tanzima Rahman Roshni 15301111
Md. Tawhidul Islam Chowdhury 16101321	Raihanoor Reza Rayon 18141021

Approval

The thesis titled "Identifying the best metrics to find the best quality clusters of genes from gene expression data" submitted by

1. Joydhriti Choudhury (15301125)
2. Tanzima Rahman Roshni (15301111)
3. Md. Tawhidul Islam Chowdhury (16101321)
4. Raihanoor Reza Rayon (18141021)

of Spring, 2019 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on 25th April,2019.

Examining Committee:

Co-Supervisor:
(Member)

Rasif Ajwad
Lecturer
Department of Computer Science and Engineering
Brac University

Supervisor:
(Member)

Dr. Md. Abdul Mottalib
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Md. Abdul Mottalib
Professor
Department of Computer Science and Engineering
Brac University

Abstract

Microarray data is used to create groups of similar genes based on their phenotypic attributes. Information extracted from these groups of gene can be applied to pathway analysis, disease predictions, target identification in drug design and many other important applications and functionalities in biology. However, how to determine a distance metric to measure the similarities among genes has always been a great challenge. In our work, we have studied sixteen combination of distance-linkage combinational metrics and tried to find the groups of similar genes based on their expression level by building phylogenetic tree. Furthermore, to validate our findings we have evaluate the output of the same trails on three different datasets. Our work suggests that, Maximum distance metric with the combination of Average linkage metrics gives the optimal quality while grouping similar genes together by building a phylogenetic tree.

Keywords: Bioinformatics, Microarray, Gene Expression, Phylogenetic Tree, Hierarchical Clustering, Distance Metric, Linkage Method.

Acknowledgement

Firstly, We would like to thank our Supervisor Professor Md. Abdul Mottalib PhD for giving us the honour to complete our thesis under him. We would also like to thank our Co-Supervisor Rasif Ajwad, Lecturer of Brac University, Computer Science and Engineering department. He has been there as a guardian and showed us the correct path through the whole journey. We would also like to show our gratitude to Faisal Bin Ashraf, Lecturer of Brac University, Computer Science and Engineering department for inspiring us with his leadership and holding the torch when the night was darkest. Moreover, we would like to thank Brac University Computer Science and Engineering department for providing us with the facilities and equipment to conduct this research. Last but not the least, we would like to thank all of the working member from Research Lab unit for their unconditional professionalism and staying with us late hours even when their shifts were over.

Table of Contents

List of Tables

List of Figures

List of Acronyms

RNA	Ribonucleic acid
rRNA	Ribosomal RNA
DNA	Deoxyribonucleic acid
tRNA	Transfer RNA
mRNA	Messenger RNA
SLINK	Single LINK
TrEMBL	Translated EMBL Nucleotide Sequence Data Library
HCPM	Health Care Practice Management
DHC	Dioxin health certificate
CSV	CommaSeparated Values
ARFF	Attribute Relation File Format
H5N1	Hemagglutinin Type 5 and Neuraminidase Type 1
SRBCT	Small Round Blue-Cell Tumor

Glossary

DNA	Deoxyribonucleic Acid. Basic building block of most living beings. Carries genotypic and phenotypic information from ancestor to descendant.
RNA	Ribonucleic Acid. Single strained nucleic acid. Plays an important role in proteins synthesis. It also carries ancestral information in organisms don't content DNA.
Central Dogma	A biological process in which information is extracted from DNA and converted into protein.
Phylogenetic Tree	A structure to represent relationship among a group of biological data.
Clustering	Clustering means grouping a set of similar data together. There are two types of clustering. 1) Hierarchical clustering, and 2) Non-hierarchical clustering.

Chapter 1

Introduction

Recent advantages in computational technology has been a blessing from heaven for mankind. On one hand, using these technologies tons of data is extracted every day from various sources previously thought impossible to extract from. Genome sequence of different species, different image databases with leveled image, financial and economic databases are some example of these data extraction. On the other hand, many technique and algorithms have been invented to deal with these huge amount of data. Various machine learning technique such as neural networks, hierarchical and non-hierarchical clustering enables computers to understand real world scenario. With the help of these computational technologies, computers are now leaving incredible impacts on image processing, data mining, bioinformatics, natural language processing and many other sectors.

Like many other sectors, bioinformatics has also benefited from the advancement of machine learning techniques. Both supervised and unsupervised machine learning technique have been used for cancer cell detection, target identification in drug design, pathway analysis etc. In addition, microarray dataset and various types of sequential dataset (rRNA sequence, protein sequence, DNA sequence etc.) has played insignificant role into the advancement of bioinformatics as a new frontier of scientific study and research.

Phylogenetic tree is a popular structure to represent biological data. This tree structure sorts object into an ancestral-descendant manner. One can easily trace down a common ancestor of two different species by following the branches of a phylogenetic tree. With the advancement of machine learning technique, it has become easier to create a phylogenetic tree regardless of the data type. Researchers use hierarchical clustering, one kind of unsupervised learning, to create phylogenetic tree. Generally, hierarchical clustering technique groups similar kind of data together into different clusters and then calculate the similarity among those clusters and connect them. At the end, we get a phylogenetic tree where each leaf correspond to separate objects and branches corresponds to the relationship between each of the objects.

As bioinformatics deals with disease prediction and drug creation, it is very important for these phylogenetic tree to be as much accurate as possible. However, very small amount of research have been conducted on this aspect. In our work we have tried to find an answer of this question. We have tried to discover one single methodology which provided the most accurate phylogenetic tree.

We have discussed about few of the basic concepts of biology and define few terminology further down this section to make it easier for the reader to understand our methodologies. In section 2, we have explained previous works of famous researchers that have motivated us to conduct research. In Section 3, we have described each of

our three datasets and explained all of our sixteen methodology. Reader will have a clear understanding on our research by reading this section. In section 4, we have presented our result and discussed on them elaborately. In last section, we have presented a summary and future work possibilities in this line.

1.1 Basic of DNA

Deoxyribonucleic Acid (DNA) is the basic building block of almost every living organism. It is mostly found in cell nucleus and a small amount is also found in mitochondria. Each DNA consists of four basic chemical bases, i) Adenine (A) , ii) Guanine (G), iii) Cytosine(C) & iv) Thymine (T). This four Alphabets are used to represent DNA numerically.

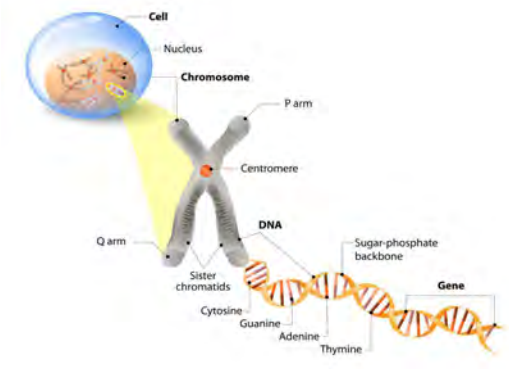

Figure 1.1: Cell, DNA and Gene.

DNA is represented as a string of A,G,C,T. Different combinations of these four letters represent different genome, Figure (1.1).

Genes are the fragments of DNA sequence which determines physical and functional nature of a species. Gene can be of hundred to million DNA bases in length. Gene makes us who we are and what our behavior will be. We humans have two copies of gene. One from our each parent. These gene determines every aspects of us e.g. our eye color, height, hair color, emotions etc. Gene can be of different length. Every gene occupies a portion of DNA of a species. Genes can evolve which leads to different DNA sequence and the result is different phenotype.

1.2 Gene Expression Data

Gene expression data [1] is one kind of a data storage popular for storing experimental biological data. It uses a special kind of technology called microarray thus known as microarray data as well. Microarray is a glass slide, divided into thousands of spot. Each spot is designated for one single gene. When researchers want to extract data from an entity (suppose mRNA) they create two sample of it. One sample is known as natural sample and the other one is known as control sample. In the natural sample,

researchers let the gene stay as it is without combining it with some chemical. On contrary, they mix some chemical or keep it in some keep control sample in some artificial environment. After that they add two different fluorescent for these two separate samples (for example red dye for the natural sample and green dye for control sample) and wash the on to microarray slide. After doing that, they use a laser and excite the array to measure the abundance of RNA. If the microarray turns into red that means the RNA from natural sample are in abundance. If the microarray turns into green then the control sample will be in abundance. The microarray turns yellow then both of the sample will be equally abundant and black color will mean that none of them achieved abundance. Researchers examine gene transcription level of an organism under different conditions at various development stages and creates gene expression profile. These profiles can explain the functionality of each gene from a genome sequence.
Gene expression data generally represents a two dimensional array where each raw represents a gene and each column represents a sample (a sample can be a tissue, time variance or different development condition and stages). Each cell of the array containing a numeric value represents the expression level of a single gene in a single sample. We can see a gene expression data of six genes under six different experimental sample in table. [1.1]

	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6
Gene 1	-1.2	-2.1	-3	-1.5	1.8	2.9
Gene 2	2.7	0.2	-1.1	1.6	-2.2	-1.7
Gene 3	-2.5	1.5	-0.1	-1.1	-1	0.1
Gene 4	2.9	2.6	2.5	-2.3	-0.1	-2.3
Gene 5	0.1		2.6	2.2	2.7	-2.1
Gene 6	-2.9	-1.9	-2.4	-0.1	-1.9	2.9

Table 1.1: Gene Expression Data.

Gene expression data can also be represented in color. In Figure 1.2, red means abundance of natural sample and green means abundance of control sample. Black means none of them are abundance and grey means missing data.

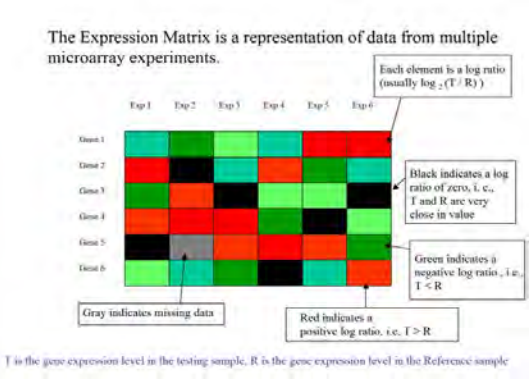

Figure 1.2: Gene Expression Matrix represented with color.

1.3 RNA

RNA stands for ribonucleic acid. It is a particle with long chains of nucleotides. Combining a nitrogenous base, a ribose sugar, and a phosphate contains nucleotide. It is a single stranded long molecule. Ribonucleic acid (RNA) has the bases adenine (A), cytosine (C), guanine (G), and uracil (U).

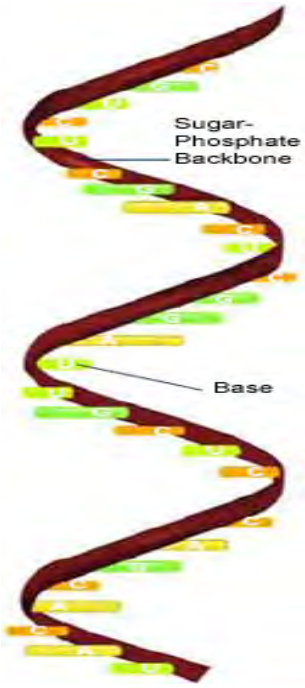

Figure 1.3: RNA.

1.4 Central Dogma

Central dogma is a process which describes how information flows from DNA all the way down to protein. In other words, it describes the process that creates protein from DNA. This process consists two parts,
i) Transcription
ii) Translation

In transcription, an enzyme called RNA polymerase, synthesizes an RNA copy of a segment of DNA. On the other hand, in translation RNA deletes the unnecessary part and by the help of ribosome creates protein. Transcription and translation happen differently in prokaryotic and eukaryotic cells. In prokaryotic cells, transcription and translation occur simultaneously. When transcription process is creating an RNA copy of a segment of DNA, translation process starts converting the created RNA to protein. This happens because of the absence of nucleus in prokaryotic cells. On the contrary, as eukaryotic cells contain nucleus, transcription process occurs inside the nucleus and translation process occurs in cytoplasm. In every cell, not all of the gene are expressed equally. Some of the gene are expressed better than the others. As a result only the expressed genes are deciphered into RNA. Protein coding gene are generally divided into three parts.
i) Promotor
ii) Coding sequence
iii) terminator
Promotor is the base-pair which indicates from where the transcription starts. On the other hand, coding sequence is the original part which gets deciphered into RNA. Finally, terminator indicates the unnecessary segment of the DNA and determine the end of transcription process for that genome sequence.

1.5 Phylogenetic Tree

Phylogenetic tree [2] is a structure which represents different organisms into ancestor-descendant manner based on their phenotypic characteristic. From the beginning of the biology, scientists wanted to place every known organism on a single phylogenetic tree under one single ancestor. To accomplish this task, they used fossil records and the phenotypic characteristics of every known organism of their time. However, after the emergence of sequential molecule data, genome sequence has replaced phenotypic characteristic as the criteria of phylogenetic tree. Scientists now have genome sequence data of every living organism and have created a universal phylogenetic tree [3]. Furthermore, with the help of computational technology, it has become possible to create a phylogenetic tree of every gene from a single organism. In that phylogenetic tree, genes that are placed closely to one another tends to show similar characteristic. Lots of research are currently going on to find methodology that will be able to create phylogenetic tree more efficiently and accurately.

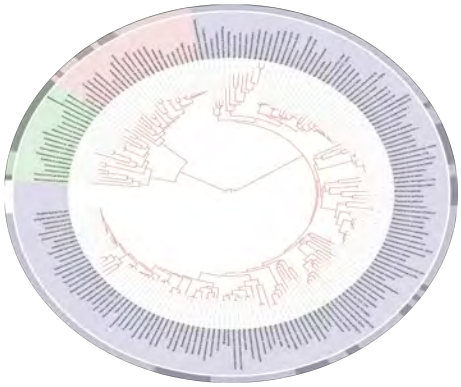

Figure 1.4: Phylogenetic TREE OF LIFE.

1.6 Clustering

Clustering is one kind of unsupervised machine learning technique, which groups similar kinds of data into a cluster. Thus dividing the whole dataset into several clusters, where each cluster is separated from the other. Clustering techniques plays a very important role in data mining. While clustering computer is presented with unleveled data. Using different clustering technique, it determines the characteristic of every single element from the dataset and try put it in a group where the other elements show same kinds of characteristics. As all the elements from a cluster displays same kind of characteristics, one can easily predict about the behavior of an unknown element if the common behavior of the cluster it belongs to is known. For example, if we want to predict some of the behavior of a cow and we know the common behaviors of the mammal cluster as well as that cow belongs to mammal cluster, we will be able to accurately predict some of the behavior of cow. In bioinformatics, clustering also plays an important role. It helps in pathway analysis, disease prediction, cancer cell detection, drug prediction etc.

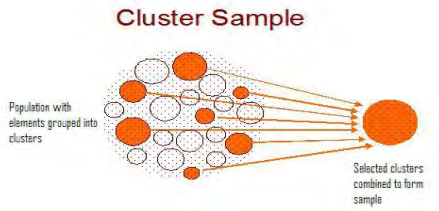

Figure 1.5: Clustering sample.

In this figure (1.5), we can see that, orange circles displays a different kind of characteristics from white circles. A clustering technology will notice this dissimilarity.

It will separate all of the orange circle from white circle and put them together in a single cluster (we can call it orange circle).
There are two types of clustering technique exists in machine learning.figure (1.6)
1) Non-Hierarchical Clustering.
2) Hierarchical clustering

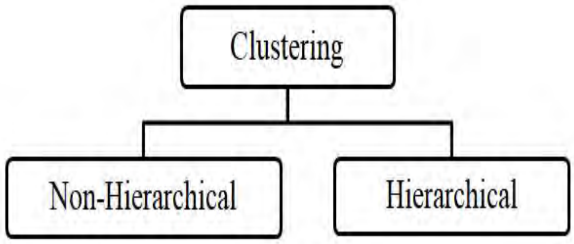

Figure 1.6: Clustering.

1.6.1 Non-Hierarchical Technique

Non-Hierarchical clustering divides the dataset into separate groups but don't create any relationship among the groups. For example, if there is a dataset consisting combination of ducks, hens and peacocks, non-hierarchical clustering technique will divide the dataset into three different group named ducks, hens and peacocks. However, it will not establish any relation among those clusters. Some of the popular non-hierarchical clustering technique used in bioinformatics are
i) K-Means,
ii) Nearest Neighbour

1.6.1.1 K-means : K-means [4] grouping is a sort of unsupervised realizing, which is utilized when it has unlabeled information. The objective of this calculation is to, to breakthrough radical in the data with the quantity of gatherings spoke to by the variable K. K-implies grouping is a sort of unsupervised realizing, which is utilized when it has unlabeled information. Data points are clustering based law of similarity. The main idea is to define bunches in the information, with the quantity of gatherings spoke to by the variable K.In figure (1.7)

$$J = \sum_{j-1}^{k} \sum_{i-1}^{n} \mid\mid x_i - c_j \mid\mid \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

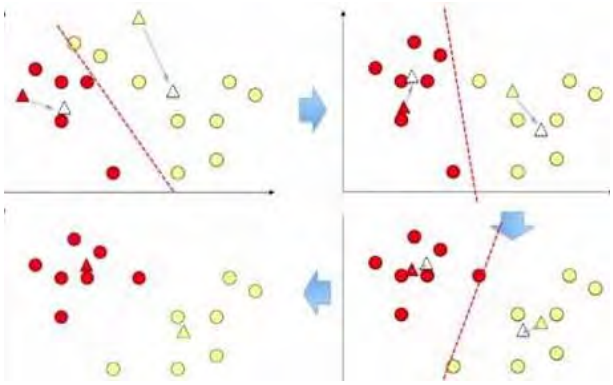

Figure 1.7: K-means Clustering.

1.6.1.2 Nearest Neighbour Method : Nearest neighbor methods [5], for example, the Jarvis-Patrick strategy, appoint mixes to an indistinguishable group from some number of their closest neighbors. User characterized parameters decide what number of closest neighbors should be considered, and the fundamental level of nearest between nearest neighbor records. In figure (1.8)

Given a bounded example of X1,...,Xn ϵ Rd , the number K of bunches to develop, and a number m ϵ N with K $\leq$ m$\leq$ n, randomly pick a subset of m "seed focuses" Xs1 ,...,Xsm . Allot every single other datum focuses to their nearest seed focuses, that is for all j = 1,...,m characterize the set Zj as the subset of information focuses whose closest seed point is Xsj . At the end of the day, the sets Z1,...,Zm are the Voronoi cells actuated by the seeds Xs1 ,...,Xsm . At that point consider all segments of Xn which are consistent on every one of the sets Z1,...,Zm. All the more formally, for given seeds we characterize the set Fn as the arrangement everything being equal.

$$F_n := \{f : X \longrightarrow 1, \ldots. \text{ K} \mid \forall_j = 1, \ldots.m : \forall_z, z^{'} \epsilon \text{ Z}_j : f(z) = f(z^{'})\} \ldots\ldots(2)$$

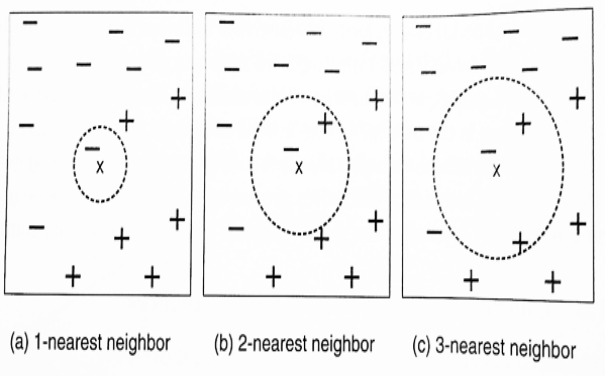

Figure 1.8: Nearest Neighbor Algorithm.

1.6.2 Hierarchical technique

Hierarchical clustering is a strategy for analyzing of cluster to exhibit groups in chain of hierarchy way. A large portion of the commonplace methods are not ready to make groups for alteration subsequent to consolidating or part process. Thus, if the consolidating procedures of items have issues, it may deliver the poor quality of cluster. Hierarchical cluster includes making cluster that have a foreordained requesting higher to lower. There are two sorts of Hierarchical clustering. Divisive and Agglomerative.

1.6.2.1 Divisive : Divisive [6] calculations start with simply just a single cluster that contains all sample information. The single bunch parts into at least 2 groups that have higher difference between them until the point when the quantity of groups winds up number of tests or as indicated by the user. IN AHC Algorithm is a repeat condition used to compute the uniqueness between a bunch Ck and a group framed by blending two other group Ci and Ci' .

$$\mathrm{D}_l w(C_k,\ C_i \cup C_i^{'}) = \alpha_i \mathrm{D}_l w(C_k,C_i)+\alpha_i \mathrm{D}_l w(C_k,\ C_i)+\beta D_l w(C_i,\ C_i^{'})+\gamma |D_l w(C_l w,C_i)\text{-}D_l w(C_k,C_i^{'})| \quad (3)$$

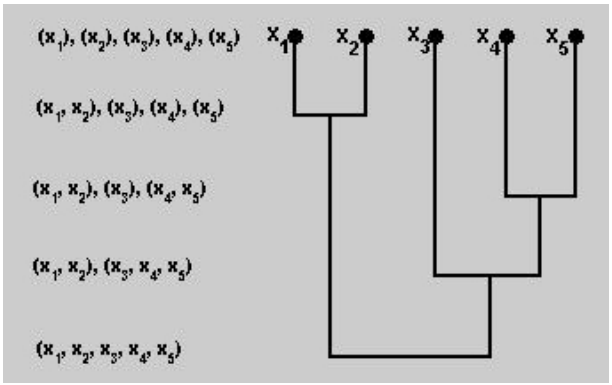

Figure 1.9: Divisive Algorithm.

In figure (1.9), it begins with the whole set of x1,x2,x3,x4 and x5 and proceed to divide into successfully smaller cluster.It is a Top Down approach.

1.6.2.2 Agglomerative : Agglomerative [7] calculation works by gathering the information one by one based on the closest separation measure of all the pairwise remove between the information point.Again separate between the information point is recalculated utilizing these accessible techniques. Some of them are:
1) single-closest separation or single linkage
2) finish most remote separation or finish linkage
3) normal separation or normal linkage
4) centroid separate
5) ward's technique - whole of squared euclidean separation is limited

An extremely helpful plan, in difference terms, which grasps all the various leveled techniques said up until now, is the Lance - Williams difference refresh equation. On the off chance that focuses (objects) I and j are agglomerated into group $i \cup j$, at that point we should basically indicate the new disparity between the group and every single other point (protests or bunches) The equation is

$$d(i \cup \text{j, k}) = \alpha\text{id(i, k)} + \alpha\text{jd(j, k)} + \beta\text{d(i, j)} + \gamma \mid \text{d(i, k) - d(j, k)} \mid \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

where αi, αj, βd, and γ characterize the agglomerative model.

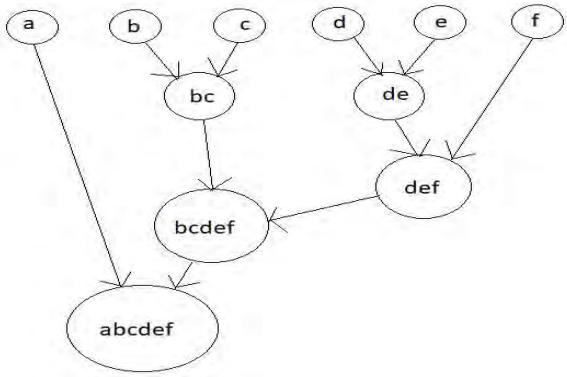

Figure 1.10: Agglomerative Algorithm.

In figure (1.10),a,b,c,d,e and f begins with each element as a separate cluster and merge them into successfully a large cluster. It is a bottom up approach.So in this figure (1.11), it is visible that Agglomerative Algorithm and Divisive Algorithm are opposite to each other.

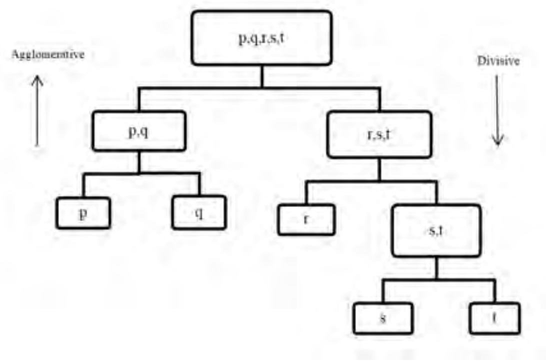

Figure 1.11: Compare Agglomerative and Divisive Algorithm.

1.7 Distance Methods

Distance methods are the mathematical calculations for calculating similarity or difference between two or more data points in multiple dimension space.

1.7.1 Minkowski Distance

Minkowski Distance [34] is a distance operate during a generalized metric which incorporates Euclidian distance and Manhattan distance during a generalized kind. This distances square measured by the Minkowski metric of totally different orders between two objects with three variables wherever displayed in a passing arrangement with x-, y- and z-axes. The unfolded cube shows the approach the totally different orders of the Minkowski metric calculate the space between the two points. Minkowski distance is generalized version of both Euclidean and Manhattan distance algorithms.

$$d = \sqrt[p]{\sum_{i=1}^{n} (a_i - b_i)^p} \quad \text{.......} (5)$$

where, a and b are the different data points. When p is 2, it works as Euclidean, and it works as Manhattan when p is 1.

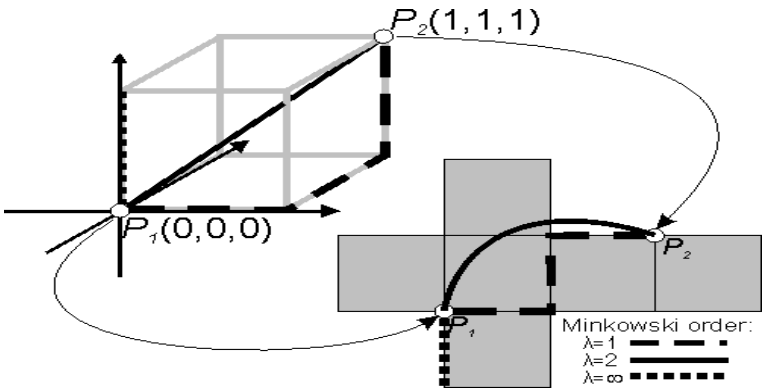

Figure 1.12: Minkowski Distance.

1.7.2 Euclidean Distance

The distance between two points outlined because the root of the add of the squares of the variations between the corresponding coordinates of the points [8].The geometric distance between two points in either the plane or third-dimensional area measures the length of a section connecting the two points. As an example, in two-dimensional elementary geometry, the euclidian distance between two points a = (a, a) and b = (b, b) is outlined as:

$$d = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \quad \text{.....(6)}$$

where a and b are two instances, each having n number of features.

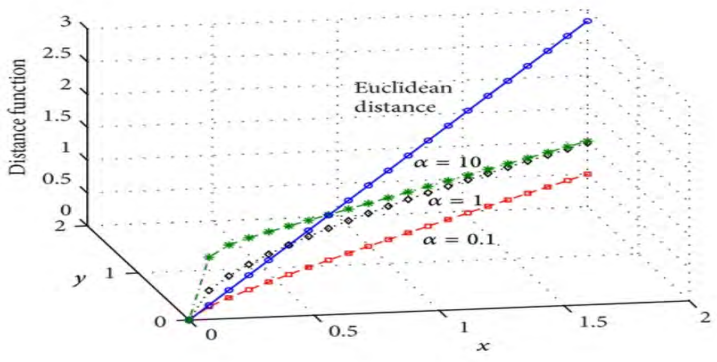

Figure 1.13: Euclidean Distance.

1.7.3 Manhattan Distance

The Manhattan distance [9] between combine two vectors or points measured on axes at right angles.For dispute, given two points p1 and p2 in a two-dimensional plane at (x1, y1) and (x2, y2) severally.It is the space between two points wherever summing absolutely the modulation of their various co ordinates calculated over the dimention of the vectors.

$$d = \sum_{i=1}^{n}(|a_i\text{-}b_i|) \quad (7)$$

where a and b are two instances, each having n number of features.

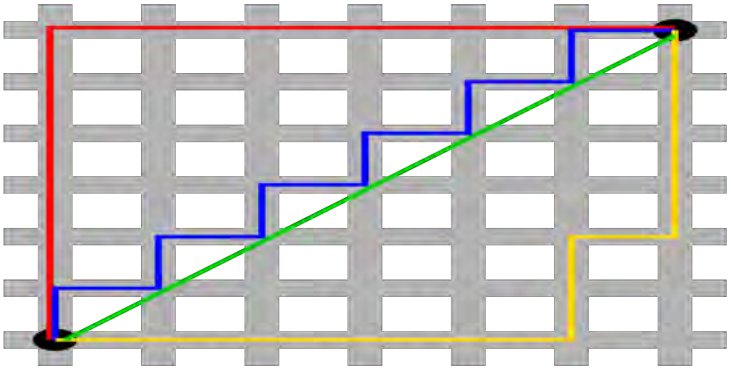

Figure 1.14: Manhattan Distance.

1.7.4 Hamming Distance

The Hamming Distance could be a measurement communicating the gap between a couple of items by the quantity of bungles among their sets of factors which are utilized for string and bitwise examinations, however, can jointly be helpful for numerical variables. Though the distance is also a metric, here bestowed version permits to outline a threshold.Variables having associate absolute distinction below the threshold area thought-about as equal.Victimization values larger than zero for this threshold and the Triangle distinction can be profound for a few calculated distances.Victimization zero as thresholds below area unit not outlined in the metric Hamming distance.

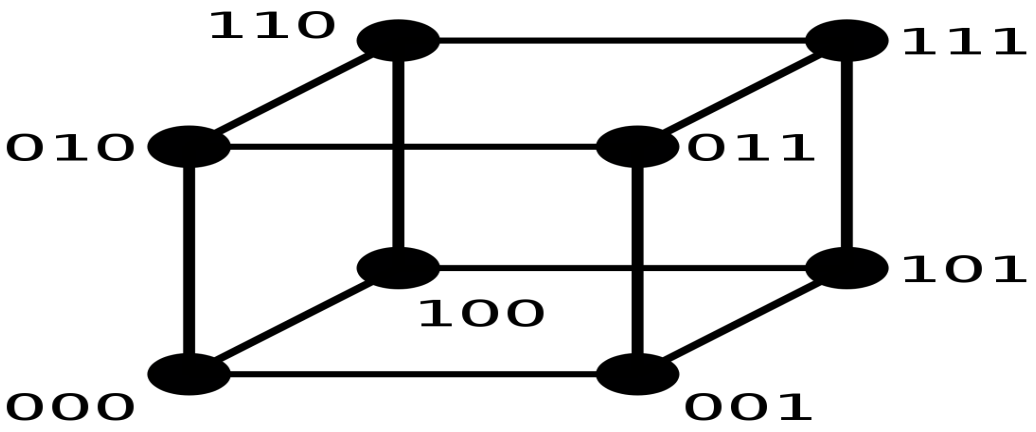

Figure 1.15: Hamming Distance.

1.7.5 Maximum Distance

Maximum distance [10] work out on a vector aperture place where the distance between two vectors is the greatest of their variation on their coordinate . It is also called Chebyshev distance. It gives the total lot differences between co ordinates of a pair of objects. This horizon source can be used for both ordinal and quantitative variables. It is also called chessboard range, as in the round of chess the base number of moves required by a ruler to move from one square on a chessboard to an alternate equivalents the Chebyshev extent between the focuses of the squares.

$$d = max((a_1 - b_1), (a_2 - b_2)) \quad (8)$$

where a and b are two instances, each having 2 features.

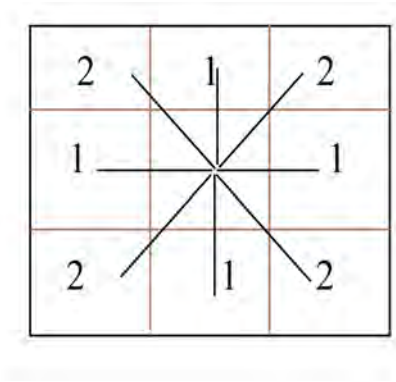

Figure 1.16: Maximum Distance.

1.8 Linkage Methods

Linkage methods are the way of linking two or more clusters to each other. Linkage methods are used to create hierarchical tree of given data by calculating relative distance among one cluster to another.

1.8.1 Single Linkage Method

A Single linkage method [11] is a group of clusters where in each step two cluster in different group combines to each other which has the nearest pair of elements and they are not belonging in the same cluster.in this method, it considers the smallest linkage criterion based on pairwise dissimilarities between elements of cluster 1 and cluster 2.

$$d = min(p, q) \quad (9)$$

where, p is an element of cluster P and q is an element of cluster Q.

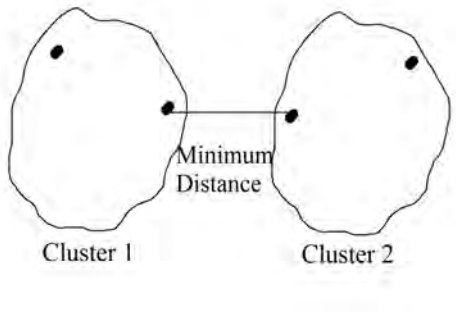

Figure 1.17: Single Linkage.

1.8.2 Complete Linkage Method

A Complete linkage method [12] is a process where the elements of two cluster are sequentially merged which has the smallest maximum pairwise distance. It computes pairwise non-similarities between the elements of cluster 1 and 2 considering the largest value to produce a compact cluster.

$$d = max(p, q) \quad (10)$$

where, p is an element of cluster P and q is an element of cluster Q.

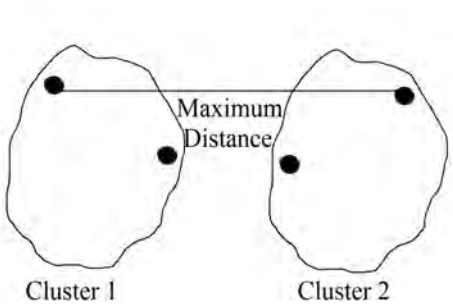

Figure 1.18: Complete Linkage.

1.8.3 Average Linkage Method

An Average linkage method [12] defined the distance between two clusters as an average between all pairs of elements where each pair is one of the objects from each group. In this method considering average, it calculates all pairwise dissimilarities between elements of cluster 1 and 2.

$$d = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} dist(a_i,b_j)}{mn} \quad \text{..... (11)}$$

where, m and n is the number of instances in cluster a and b.

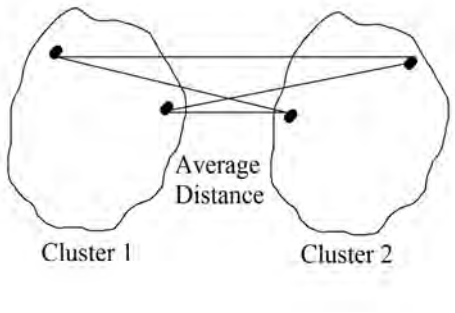

Figure 1.19: Average Linkage.

1.8.4 Centroid Method

The Centroid method measured the distance of similarities or dissimilarities among clusters which has two mean vectors which compute dissimilarity between the centroid of cluster 1 and 2.

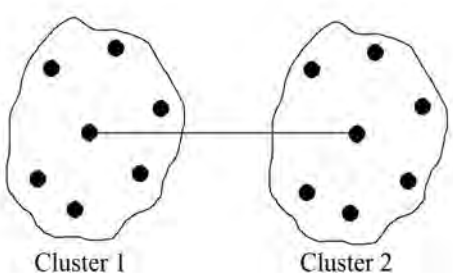

Figure 1.20: Centroid Method.

1.8.5 Ward's Method

The Wards method [13] is computationally intensive which makes a new cluster to minimize variance in each step of clustering where selecting the pair of clusters to merge based on the optimal value of an equitable function providing the smallest increase in total error sum of squares.

$$d_{ij} = d(X_i, X_j) = (|\mathrm{X}_i - X_j|)^2 \quad (12)$$

where, X_i is the value of X^th element in i cluster and X_j is the value of X^th element in j cluster.

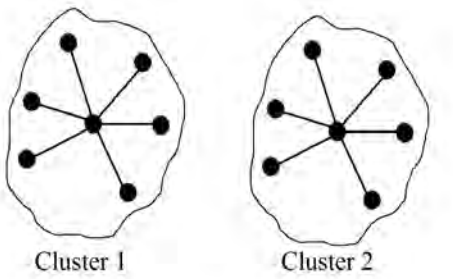

Figure 1.21: Ward's Method.

Chapter 2

Literature Review

From the days of the Charles Darwin, phylogenetic tree has been considered as one of the center figure of the biology. Ability to sort different entities based on their evolutionary characteristics and create an ancestor-descendant relationship among those entities made phylogenetic tree one of most favorite tools to represent biological data. Charles Darwin, in his famous book "The Origin of Species", produced one of the first phylogenetic tree representation in the history of biology. However, lack of biological data forced Darwin to leave many blank spaces on that evolutionary tree.

In the beginning researchers used fossil records to create phylogenetic tree. Those trees were full of error as only a small portion of the whole ecology ever had any chance to be fossilized. A huge number of species remained unnoticed from those phylogenetic trees. The scenario changes with the arrival of molecular sequences. Zuckerkandl and Pauling [14] in their paper "Molecules as Documents of Evolutionary History" shifted the basis of phylogenetic inference from cellular characteristic to molecular sequences. From there on, researchers have used rRNA sequencing to create phylogenetic tree.

From here question rises that, why should we choose rRNA sequencing rather other tRNA or mRNA sequencing? The reason is that, rRNA is present in every single bacteria [15]. Furthermore, the functionality of rRNA has not change over the time. As a result all the changes in the sequence actually means the evolution of the organism over the time [15]. In addition, rRNA is also large enough for informatics purpose [15]. All of these reasons made rRNA sequencing a perfect contender to build phylogenetic tree.

As time passed by, researchers were able to extract sequences from many other molecules (such as protein sequence, aminoacyl tRNA synthetases [16]). They used these sequences to create phylogenetic trees. To their surprise these trees contradicted with rRNA based phylogenetic trees [16]. Some of the arguments arises because rRNA trees are demanding that Archaea and Bacteria are closer to each other because they have more metabolic genes common with each other than the Eukaryotes [16]. This closeness is completely based on the number of genes rather than phylogenetic analysis. Some researchers raised question on the prokaryotic root of the rRNA based phylogenetic tree. Now reserchers are suggesting that, we should take Eukaryotes as the root rather than prokaryotes while creating universal phylogenetic tree [17]. Another objection raised against rRNA based phylogenetic tree is that, these phylogenetic trees are completely ignoring one of most the important events of evolutionary process named horizontal gene transfer. Horizontal gene transfer occurs when a bacteria receives DNA from another bacteria which is not its ancestor [18]. As a result,

the DNA sequence of the first bacteria changes. Horizontal gene transformation is a common process in evolution and in the earlier stages of life it played a key role shaping microbial sequences. rRNA phylogenetic trees, however, seems to be unable to address such an important evolutionary process.

One of the major problems researchers faced in their early days while creating phylogenetic tree is to calculate how closely two genes are related to each other. With the advancement of microbiology, they had a huge amount of data but were unable to extract information from them because of human limitations. Computer science solved this major conflict. With the emergence of machine learning, computers were able to learn real world scenario and work million times faster than any human being. Bioinformatics researchers used this new technology to extract information from huge dataset. The clustering techniques, one kind of unsupervised machine learning technique, to create phylogenetic trees. Many clustering techniques arrived to group similar looking data together. Hierarchical clustering, a special type of clustering technique where similar clusters merged into a bigger group, used to form complex phylogenetic trees from dataset containing thousands of genes.

Researcher used different types of distance methods to measure the distance between two genes. Some of the popular distance methods are, Minkowski distance method, Manhattan distance method, Euclidean distance method, Chebyshev distance method, Hamming distance method etc. All of these distance methods calculate distance between two genes using different mathematical equations. For example, if the dataset is time variant dataset, where expression of a gene is recorded as numeric value on different time scale, the distance methods will calculate the similarity of a pair of gene depending on the numeric values of every time scale. If the dataset is a sequence dataset, where a gene is stored as a genome sequence, the distance method will calculate the distance of a pair by comparing each element of sequence. After calculating the distance between each pair of genes on a level, it will merge the closest ones together. After a distance method is finished running over a dataset, the dataset gets divided into several separate clusters. Each of those clusters contains genes closely related to each other. To connect these groups we need linkage method. Researchers use different kinds of linkage methods. Some of them are single linkage method, complete linkage method, average linkage method etc. With the help of linkage method every separate clusters get connected and ultimately creates a complete hierarchical tree.

Search for an appropriate hierarchical clustering technique dates back to 1960s. Stephen C. Johnson in his paper Hierarchical Clustering Scheme [19] tried to discover two hierarchical clustering methods. One method will try to form optimally connected clusters and the other method will try to form clusters, which are optimally connected. As mentioned before, he faced problems with computational capabilities of his time. R.Sibson in his paper SLINK: An optimally efficient algorithm for the single-

link cluster method [20] discussed about creating clusters using single linkage method. On the other hand, inspired by Sibson's work, D.Defays proposed a methodology to create cluster using complete linkage method in his paper An efficient algorithm for complete linkage method [21]. Furthermore, in his paper, A Survey of Recent Advances in Hierarchical Clustering Algorithm [22], F.Murtage, proposed a framework which decreases the complexity of clustering method from O(N2) to O(NlogN). In 1993, Edgardo A.Ferran and Bernard Pflugfelder, described a method of hierarchical clustering based on statistic and artificial intelligent [23]. They used protein sequence data as input and reduced time complexity previously known. In 1999, Fin Drablos used hierarchical cluster analysis to identify non-polar interactions in protein [24]. Moreover, Antje Krause, Jens Stoye and Martin Vingron tried to cluster protein sequence in large scale using single linkage method [25]. They used their methodology on a set of sequence of all know protein sequence data from Swiss-Prot Rel 41 and TrEMBL Rel 43. On the other hand, Marcilio CP de Souto, Ivan G Costa, Daniel SA de Araujo,Teresa B Ludermir and Alexander Schliep in their research paper Clustering cancer gene expression data: a comparative study[26] explained that K-means clustering generates better result than hierarchical clustering while considering cancer tissue. However, they also mentioned that, clustering cancer gene expression data is very much different from clustering gene expression data. In cancer gene expression data, one has to cluster thousands of cancer tissues and every tissue is described by hundreds of genes. On contrary, in gene expression data thousands of gene are described using a limited number of expressions, as a result K-Means performs better in cancer tissue expression data. Additionally, programs such as, HCPM[27], Jerarca[28], ClUSTAG[29], DHC[30] are developed to create hierarchical clustering trees.

Although lots of research has been conducted on how to reduce time complexity of various clustering method a very few has been conducted on calculating the accuracies of those clustering technique. Following Moore's law our computers are getting faster every year. We believe that, we are in a position of history, where we can take our eyes off from how much time a method take to create cluster and focus on how accurately those clusters are created using that method. In addition, bioinformatics research domain also include various kind of disease prediction, (cancer prediction, small blue cell tumor) as well as which types of drags should be used in which condition. Because of this reasons we believe that, researcher should give higher preference in accuracy rather than in time complexity. In our research paper, we have proposed sixteen different distance-linkage methodologies to create hierarchical phylogenetic tree. We have chosen three different datasets to justify our process and have found out one methodology which produced optimal result on each dataset.

Chapter 3

Methodology

Our method uses different distance method with different linkage method over multiple dataset. This works on different dataset and finds the optimal algorithm for that dataset.

3.1 Defination

In this section , some definitions are introduced which will help to describe our algorithm clearly.

Dendogram : Representation of a hierarchical tree.

Average Silhouette Width : Represents how much similar an object is in its own cluster.This is the distance within the cluster, it shows how compact a cluster really is. It can also be called the radius of a cluster. It is measured from the distance between the farthest position of the nodes within the cluster.

Distance Within Cluster : Represents the distance of a particular cluster from all other clusters.This is the distance between each cluster in the whole dendrogram. It is usually measured from the centroid of one cluster to the centroid of another cluster.

Fitness Function : We have used a fitness function to determine the quality of our clustering; the function we used in our research is a ratio between Average Silhouette Width and Distance Within Cluster. The equation and its description is discussed section III.

3.2 Proposed Methodology

Step 1

At first we import a microarray dataset in the system which is a 2D(Two Dimention) matrix containing gene expression data of different genes at different time stamps. These datasets can be in CSV(Comma Separated Values) or ARFF(Attribute Relation File Format). For example, two of our dataset “SRBCT” is ARFF file format and “Spellman” is a CSV file. Different library is needed in order to read these file. With appropriate library we can extract data from these files and proceed to the next step.

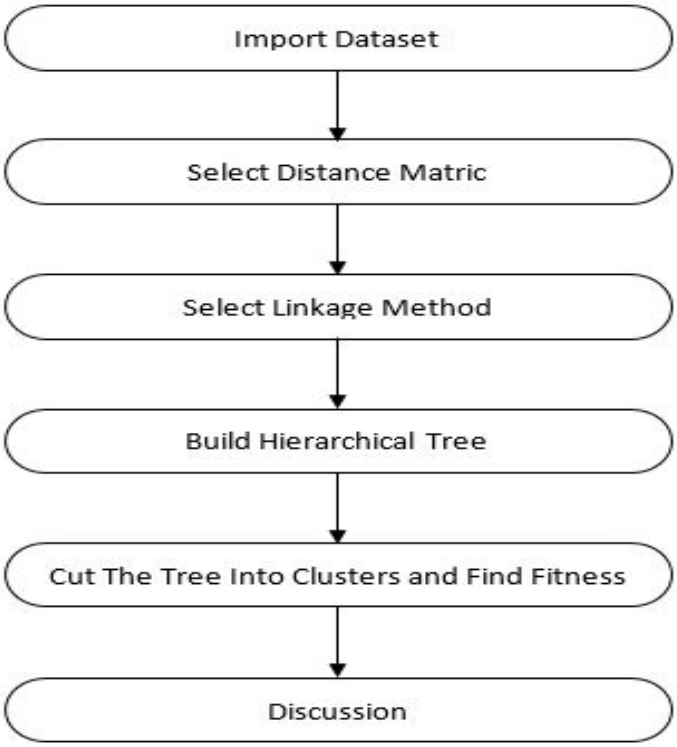

Figure 3.1: Flow diagram of our proposed method.

Step 2

After the dataset has been imported, we run different distance measuring algorithm over the dataset. These distance measuring algorithms takes timestamp value of genes expression data from the 2D array and calculates distance from gene to gene. We mainly used four distance measuring algorithm which are Euclidian, Manhattan, Minkowski and Maximum. All these method represents different calculation. Though other methods can also be introduced.

Step 3

After calculating distance between genes, we need to make cluster of genes to build the hierarchy and here the linkage method takes place. We have used a linkage method to connect those separated groups until all the groups fall under a single tree. A linkage method calculates the distance between one group of data to another group of data, and connects two groups comparing their distance value. After connecting a group pair on a level, it considers the group pair as a single group and calculates the distance between that group and the other groups the same way mentioned above. Thus after connecting different groups of data on different levels we get a hierarchical phylogenetic tree. Some of the popular linkage methods are Single, Average, Com-

plete, Ward etc.

Step 4

After using a linkage method, we get a hierarchically clustered tree, where the tree leaves indicate each gene from the data set. The branches of the hierarchical tree shows how much dissimilar the genes are in terms of their expression level, and draw the dendogram of the tree. In the dendogram, similar genes are branched closely and dissimilar genes are branched farther away.

Step 5

From the plot we can see the tree height and where most of the clustering are done. We cut the tree at different height which produces different groups of data. Now we can decide which height gives more accurate clusters. Cluster accurary is evaluated with the help of "Distance Within Cluster" and "Average Silhouette Width" value. This two value helps us to understand state of the clusters. With this two value we can see which distance method with the combination of linkage method gives the optimal output. After that, we calculate the fitness value from "Distance Within Cluster" and "Average Silhouette Width" which lets us decide the tree cut height, which is best for a specific distance and a particular linkage method. Formula for fitness calculation is:

$$r = \max\left(\frac{s_1}{w_1}, \frac{s_2}{w_2}, \frac{s_3}{w_3} \ldots\ldots\ldots\ldots \frac{s_n}{w_n}\right) \quad (13)$$

Here, S_i (i=1,2,3,...n) represents "Average Silhouette Width" value and W_i (i=1,2,3,...n) represents "Distance Within Cluster" value, where "i" represents each iteration of "s" and "w". For every other distance method, we run form step 2 to step 5 and continue the process.

Step 6

After running step 2, 3, 4 and 5 for different distance and linkage combination, we get total of sixteen fitness value for each dataset.

Example

To understand the methodology better, let us consider that we have some microarray data (Table 3.1). There we have four timestamp value of expression level of four genes.

Names	Time 1	Time 2	Time 3	Time 4
Gene 1	0.2	0.4	0.08	0.06
Gene 2	0.25	0.32	0.73	0.45
Gene 3	0.55	0.63	0.18	0.19
Gene 4	0.28	1.37	0.35	0.21

Table 3.1: Expression level of four gene with four timestamp value.

For instance, we will use Euclidean Distance for calculating distance between expression level of different genes. For "Time 1" timestamp we will take the squared difference of expression value of "Gene 1" and "Gene 2". In the same way, we will calculate for all the timestamp and after that we will calculate the root of sum of all timestamp. This is how we calculate Euclidean distance between "Gene 1" and "Gene 2". Follow the same procedure for calculating distence between all the genes. Calculated values are shown in Table 3.2.

Names	Gene 1	Gene 2	Gene 3	Gene 4
Gene 1	0	0.76	0.45	1.02
Gene 2	0.76	0	0.75	1.14
Gene 3	0.45	0.75	0	0.80
Gene 4	1.02	1.14	0.80	0

Table 3.2: Euclidean Distance value from gene to gene.

As we have all the distance value, now we can use a linkage method in order to form a hierarchical tree from all these data. Here we are using "Single" linkage method. Single linkage starts building the tree by linking two closest genes first. In Table 2 we can see the lowest value is 0.47 which is the distance between "Gene 1" and "Gene 3". Therefore Single linkage method will make cluster of those two genes.

Names	{Gene 1, Gene 3}	Gene 2	Gene 4
{Gene 1, Gene 3}	0	0.75	0.80
Gene 2	0.75	0	1.14
Gene 4	0.80	1.14	0

Table 3.3: 1st iteration of Single Linkage Method.

Notice here that distance of "Gene 2" and "Gene 4" from cluster "Gene 1, Gene 3" is calculated by taking the lowest distance value of "Gene 2" and "Gene 4" from the clustered genes individually. For "Gene 2" the lowest distance value available is 0.75 and this will be the updated value for "Gene 2" in Table 3. Same procedure is followed for "Gene 4". This process will continue until we are left with last two cluster.

Name	{{Gene 1, Gene 3}, Gene 2}	Gene 4
{{Gene 1, Gene 3}, Gene 2}	0	0.80
Gene 4	0.80	0

Table 3.4: 2nd iteration of Single Linkage Method.

From this table now we have the hierarchy of genes. Now we can make a dendogram from this data and visualize how the genes are connected.

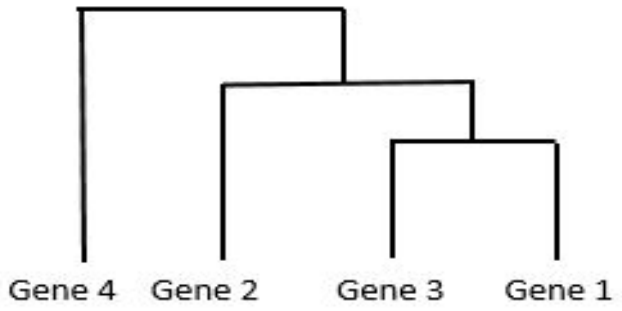

Figure 3.2: Dendogram by generating table.

As we have the hierarchical tree, now we cut the tree in different height and we get clusters of genes. Then we compute the fitness of the clustering by calculating "Average Silhouette" and "Within Cluster" values. Average Silhouette value ranges from -1 to +1. The closer the value to +1 indicates better cluster. And the Within Cluster value ranges from 0 to +. The closer the value to 0 is better. In this process after getting all sixteen combination we can compare the results.

Chapter 4

Result and Discussion

We have implemented our methodology on 4 distance methods (Euclidean, Manhattan, Maximum and Minkowski) using 4 linkage methods (Average, Complete, Single and Ward). After building the phylogenetic tree we cut on different height so accurately that it gives the perfect result to find best quality clusters. So our overall experiment process contains 16 Different expression to show Different cluster and experiments on three Different data set.

4.1 Dataset Description

In our methodology we have used three datasets. SRBCT, InfluenzaH5N1 and Spellman. Description of the dataset is given below.

Dataset A : SRBCT

Description:

Data domain: Small Round Blue Cell Tumors
Row: 83 sample count
Column: 2308 number of genes

Summary:

The small round blue cell tumors (SRBCTs) are 4 different childhood tumors named so because of their similar appearance on routine histology, which makes correct clinical diagnosis extremely challenging. The expression data of the genes are taken as sample counts on different times stamps [31].

Dataset B : Influenza H5N1

Description:

Data domain: Influenza virus H5N1 infection of U251 astrocyte cell line: time course
Row: 24 transcripts
Column: 18 timepoints

Summary:

Analysis of U251 astrocyte cells infected with the influenza H5N1 virus for up to 24 hours. Results provide insight into the immune response of astrocytes to H5N1 infection. The values of the data represent the intensity transformed count of the U251 astrocyte cells [32].

Dataset C : Spellman

Description:

Data domain: Saccharomyces cerevisiae cell cultures
Row: 4381 transcripts
Column: 23 timepoints

Summary:

The Spellman dataset provides the gene expression data measured (on a custom platform) in Saccharomyces cerevisiae cell cultures that have been synchronized at different points of the cell cycle by using a temperature-sensitive mutation (cdc15-2), which arrestes cells late in mitosis at the restrictive temperature (it can cause heat-shock). These expression values are represented in the dataset [33].

4.2 Experimental Analysis

By using Different distance methods and Different linkage methods we get best quality clusters on three different data set.In Maximum distance method along with complete linkage on data set SRBCT we implemented our methodology and found the hierarchical tree. As whole cluster image is too much compact so we have tried to show as much as possible clear picture by clipping the image.A small fraction of the tree has been shown in Fig. 4.1.

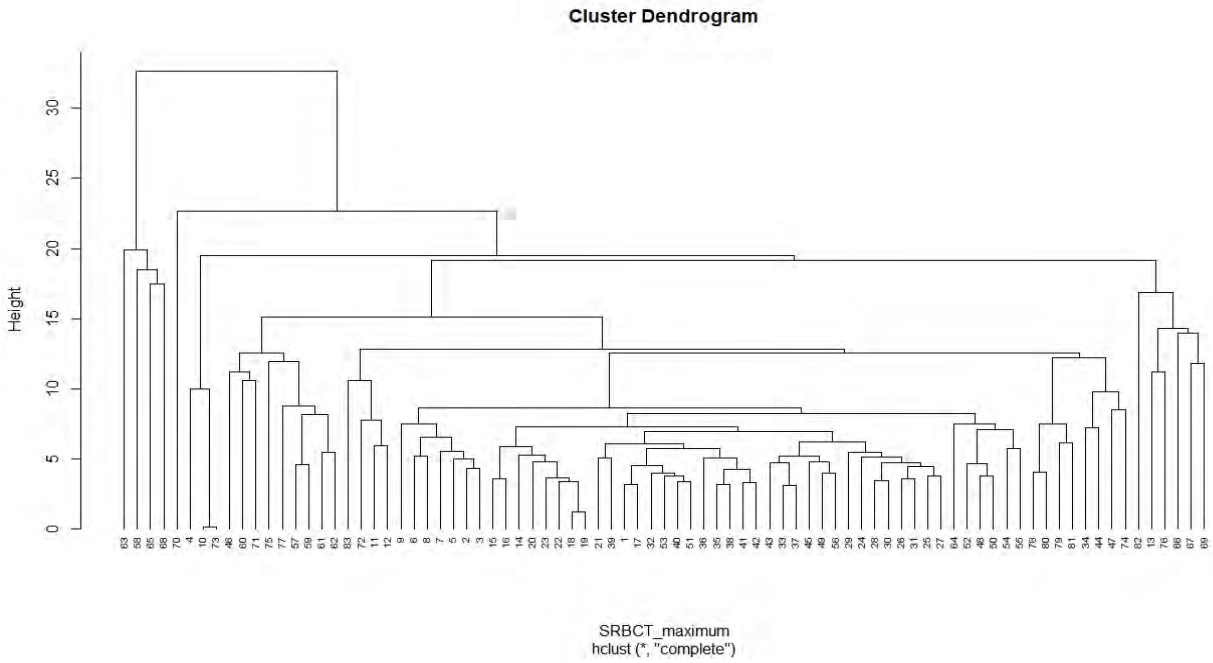

Figure 4.1: Fraction of the tree generated from Data set SRBCT using Maximum distance method and Complete linkage method.

We cut the tree at different height which gives us a set of clusters where we have measured fitness by calculating Average Silhouette width value and Distance within cluster value for each height of the cut of cluster. Then we gather those values and plot a graph after normalizing the values Fig. 4.2.

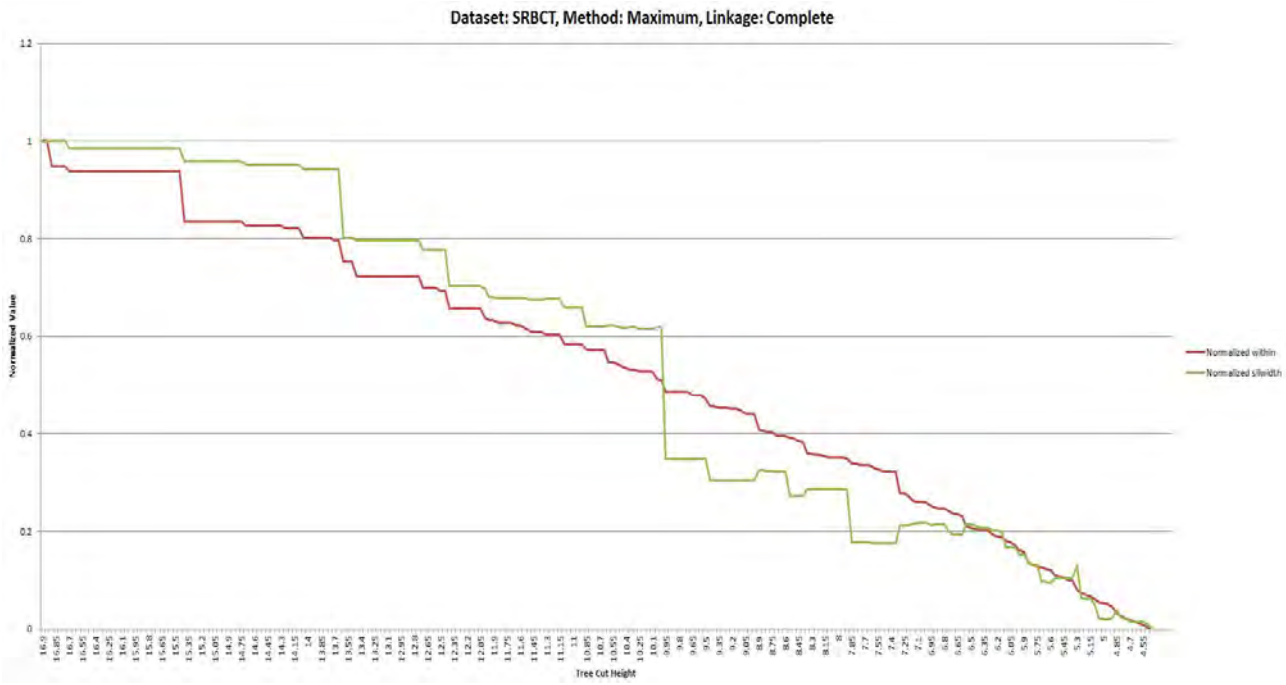

Figure 4.2: Cluster fitness after cutting at different height on SRBCT Data set.

Analyzing this tree we get the best fitness value 3.7256 x 10-5 at a cutting point of 5.3 in the tree using fitness function method.

Again,We have used maximum distance method combined with single linkage on data set InfluenzaH5N1, and we get the hierarchical tree Fig. 4.3. In order to get the best fitted clusters we have also cut the tree at different heights. Then again we have measured the Average Silhouette width value and Distance within cluster for finding the best clusters in this trial using fitness function method.

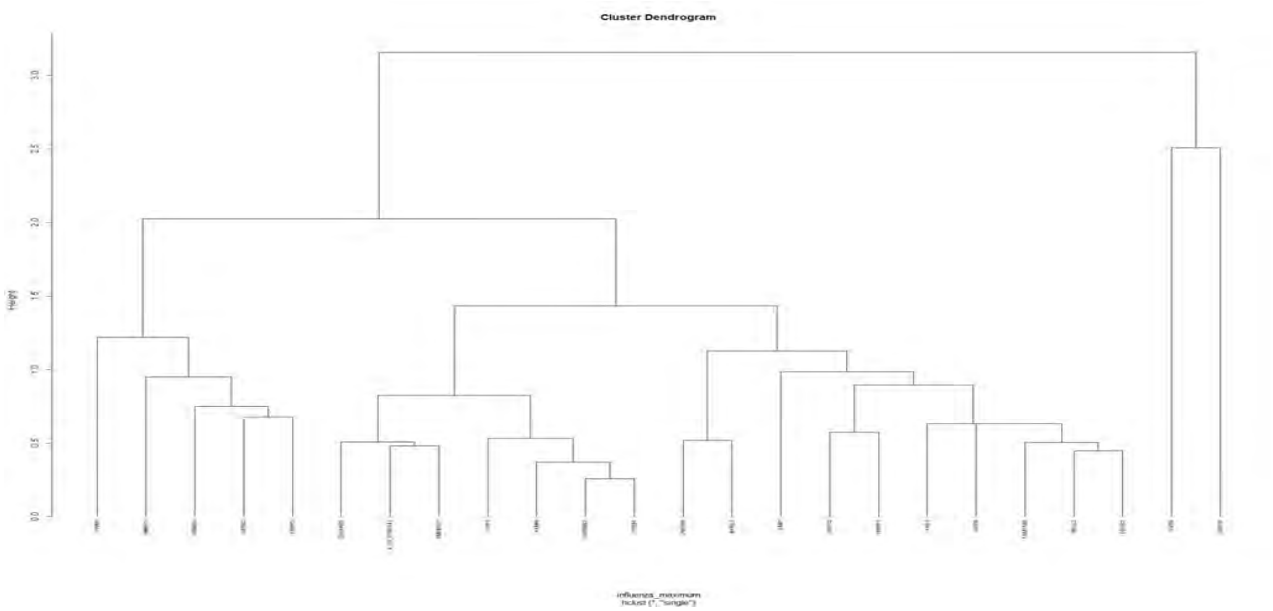

Figure 4.3: Fraction of the tree generated from Data set InfluenzaH5N1 using Maximum distance method and Single linkage method.

Fig. 4.4 displays the graphs of different fitness values at different cut points after normalizing the values. In this trial, we get 1.0346 on 0.35 height, which is giving the best result i.e., the best quality clusters.

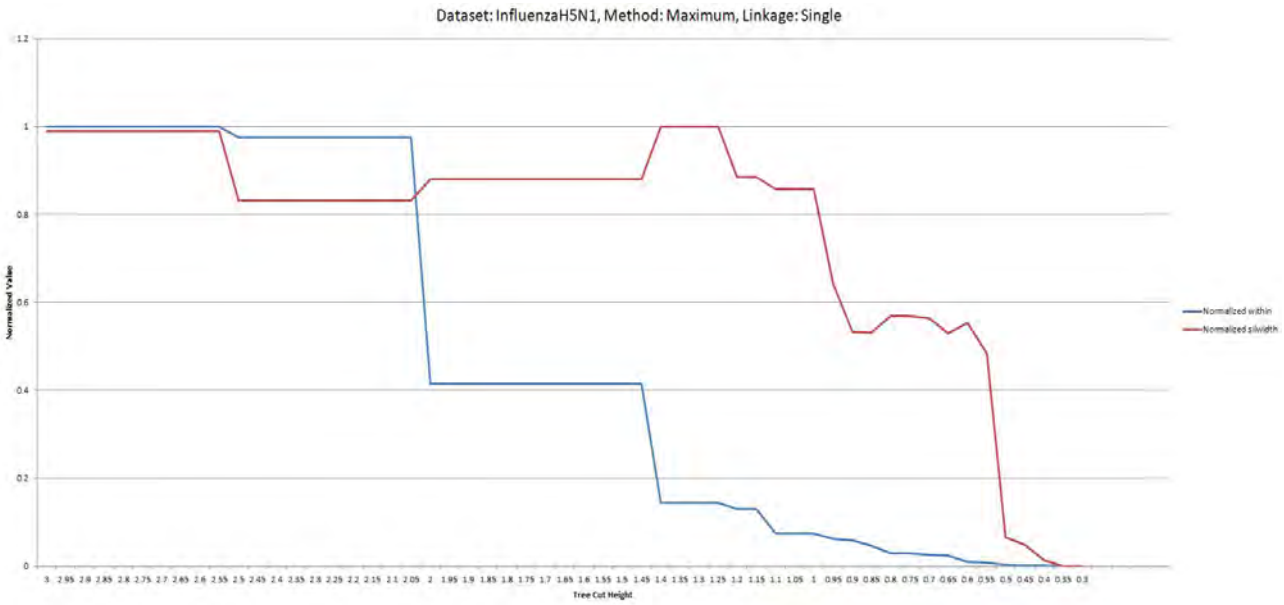

Figure 4.4: Cluster fitness after cutting at different height on InfluenzaH5N1 Data set.

Moreoevr,In Spellman dataset we implemented our methodology and found the hierarchical tree using Maximum distance method along with average linkage. As whole cluster image is too much compact so we have tried to show as much as possible clear picture by clipping the image.A small fraction of the tree has been shown in Fig. 4.5.

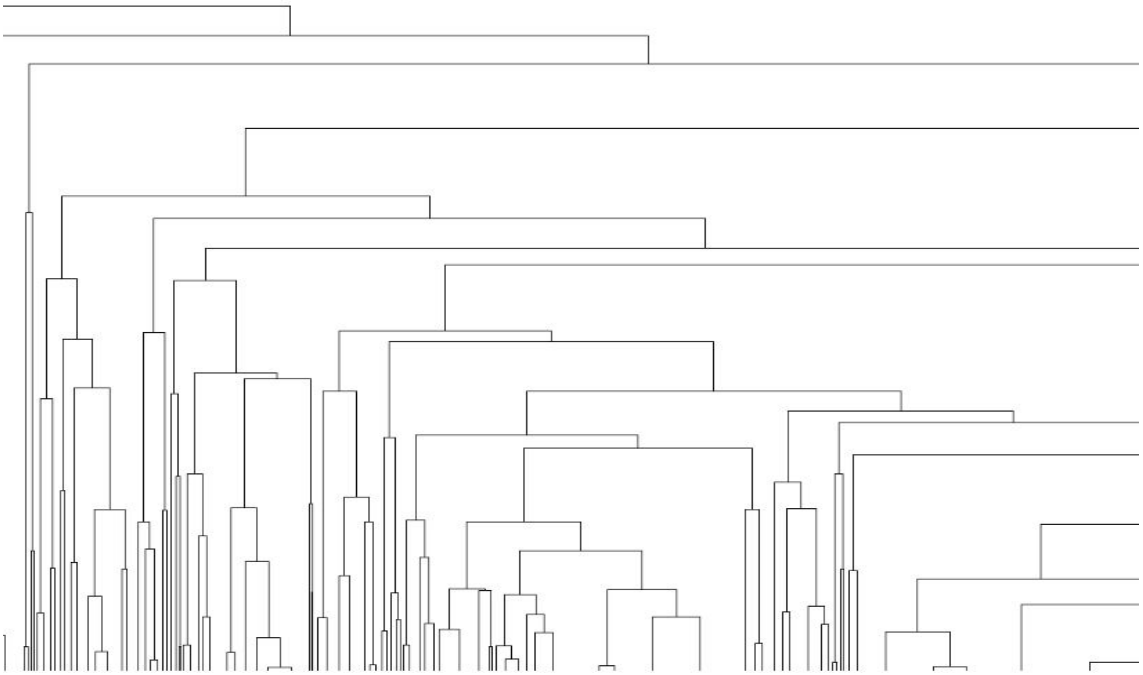

Figure 4.5: Fraction of the tree generated from Data set Spellman using Maximum distance method and Average linkage method.

Furthermore,We cut the tree at a different height which gives us a lot of groups of clusters where we have estimated fitness by calculating Average Silhouette width value and Distance within cluster value for every stature of the cut of clusters. At that point we assemble those qualities and plot a graph in the wake of normalizing the values.

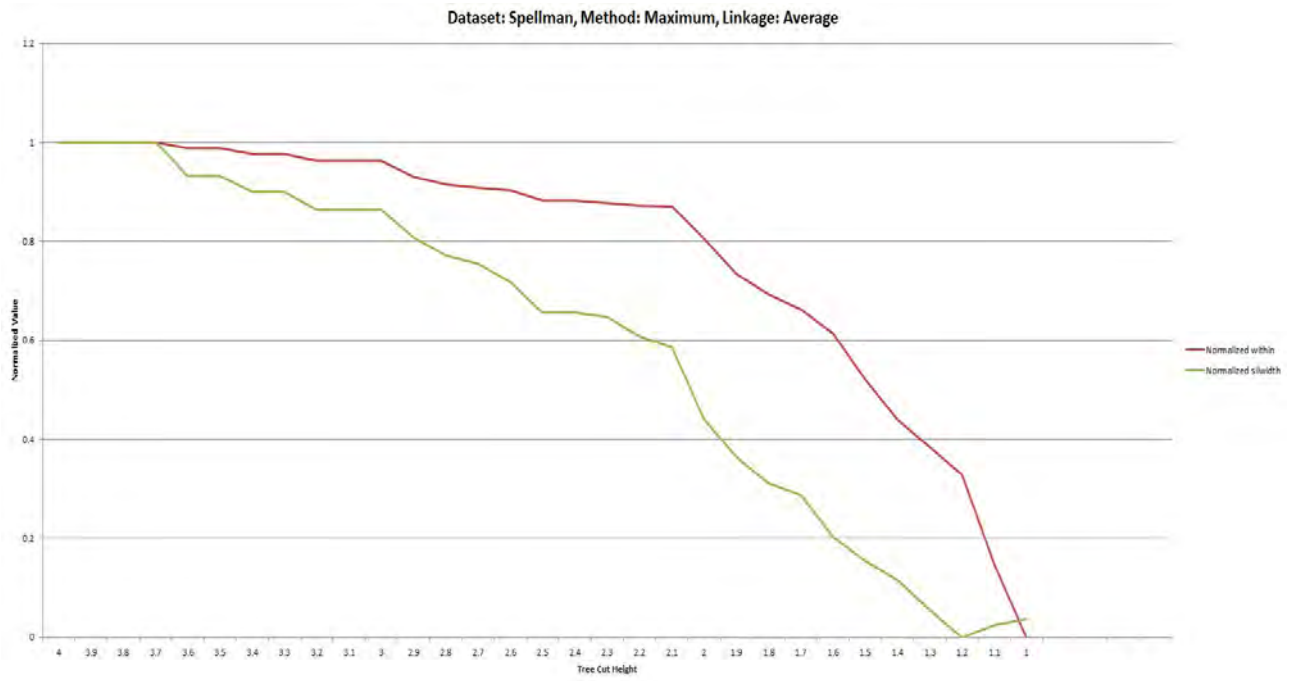

Figure 4.6: Cluster fitness after cutting at different height on Spellman Data set.

Analyzing this tree on spellman dataset we get the best fitness value 1.618 x 10-4 at a cutting point of 3.7 in the tree using fitness function method.

We have gone through all the 16 preliminaries for three different data sets, of which figures from three trials have been reported here. Additionally, the entire outcomes we get from all the trials of our experimentation are given in Table 4.1 and Table 4.4. Here it is appearing the Average Silhouette Width value and Distance Within Cluster which we have used on 3 data sets, 4 distance metrics and 4 linkage methods. Apart from this we have reported here all the clusters and graphs picture in the Appendix Section.In the Appendix section , Appendix A contains total 48 clusters images for the three data sets as well as in Appendix B it contains 48 graph images same for those clusters which we have discussed in section 4.2 of Experimental Analysis.

4.3 Experimental Results

The whole results we get from all of the trial of our experimentation are provided in Table 4.1 and Table 4.2. Here it is showing the Average Silhouette Width value and Distance Within Cluster which we have used on 3 data sets,combined with 4 distance metrics and 4 linkage methods.

Dataset	Distance/Linkage	Average		Complex	
		Average Silhouette Width	Distance Within Cluster	Average Silhouette Width	Distance Within Cluster
SRBCT	Euclidean	0.2311	34332.47	0.2759	58496.52
	Manhattan	0.2482	1636685	0.6282	6073842
	Maximum	0.2695	6109.289	0.2777	7454.639
	Minkowski	0.2311	34332.47	0.2759	58496.52
Influenza H5N1	Euclidean	0.0346	0.1485	0.0346	0.1485
	Manhattan	0.0374	1.6178	0.0374	1.6178
	Maximum	0.0346	0.0334	0.0346	0.0334
	Minkowski	0.0346	0.1485	0.0346	0.1485
Spellman	Euclidean	0.5113	16961.94	0.3282	17110.32
	Manhattan	0.6497	268345	0.3766	248568.9
	Maximum	0.6888	4256.024	0.6448	4236.655
	Minkowski	0.5113	16961.94	0.1040	14094.44

Table 4.1: Experimental results for all combination of distance metric and linkage method.

Dataset	Distance/Linkage	Single		Ward	
		Average Silhouette Width	Distance Within Cluster	Average Silhouette Width	Distance Within Cluster
SRBCT	Euclidean	0.2521	58097.16	0.0082	95928.06
	Manhattan	0.1695	2057250	0.01358	3481927
	Maximum	0.5247	17475.73	0.0338	12643.9
	Minkowski	0.2086	49793.96	0.1554	121354.2
Influenza H5N1	Euclidean	0.0346	0.14855	0.0346	0.1485
	Manhattan	0.0374	1.6178	0.0374	1.6178
	Maximum	0.0346	0.0334	0.0346	0.0334
	Minkowski	0.0346	0.1485	0.0346	0.1485
Spellman	Euclidean	0.5435	17696.39	0.1309	14319.56
	Manhattan	0.5877	269365.2	0.0684	179041.1
	Maximum	0.6276	4211.773	0.0487	3300.416
	Minkowski	0.3222	17080.06	0.0573	12511.37

Table 4.2: Experimental results for all combination of distance metric and linkage method.

4.4 Discussion

We have experimented our methodology on 3 data sets applying 4 distance metrics with 4 linkage methods. In Each of the Combination of distance we get Average Silhouette width and Distance within cluster for each of the combination of distance metrics and linkage methods are shown in Table 4.1 and Table 4.2.
In data set SRBCT, by implementing our methodology, we get average silhouette width and distance within cluster and found best ratio between average silhouette width and distance within cluster. We get the best fitness, which is 4.4129 x 10-5 by using maximum distance and average linkage. The worst fitness is 3.902 x 10-9 whenever we have used Manhattan distance and ward linkage. In order to do that, we get best cluster whenever cutting at height 3.55 and worst quality cluster after cutting at height 379.

In data set InfluenzaH5N1, by implementing our methodology we get average silhouette width and distance within cluster and found best ratio between average silhouette width and distance within cluster. We get the best fitness, which is 1.0346 by using Maximum distance and average linkage, complete linkage, single linkage and ward linkage. The worst fitness is 0.0231, whenever we have used Manhattan Distance and average linkage, complete linkage, single linkage and ward linkage. In order to do that, we get best cluster whenever cutting at height 0.4 and 0.35, worst quality cluster after cutting at height 3, 3.2 and 3.4.

In data set Spellman, by implementing our methodology we get average silhouette

width and distance within cluster and found best ratio between average silhouette width and distance within cluster. We get the best fitness, which is 1.6 x 10-4 by using Maximum distance and average linkage. The worst fitness is 3.82 x 10-7, whenever we have used Manhattan Distance and ward linkage. In order to do that, we get best cluster whenever cutting at height 3.7 and worst quality cluster after cutting at height 900.

Chapter 5

Conclusion

5.1 Summary of Work

Everyday bioinformatics researchers are facing heavy bombardment of biological data. These data are coming in all shapes and sizes. Some of them are sequential data. Protein sequences, genome sequences, RNA sequences, DNA sequences are some of the popular example of sequential data. Others are gene expression data, expression of genes under various circumstances are stored in those datasets. All of these datasets needs a proper representation technique. Thus enters phylogenetic tree. It is one of the most popular method in biological data representation, where all the elements of a dataset get connected with each other into ancestor-descendant fashion. Similar type of data stay closer to one another and dissimilar data placed as far as possible from one another. However, to implement a phylogenetic tree, elements of the datasets needs to get hierarchical clustered first. With the advent of machine learning techniques, there are many agglomerative hierarchical clustering technique available in our time. We have picked sixteen distance-linkage method combination and run them on three different datasets. From our research combination of maximum-average has emerged as optimal solution to create accurate phylogenetic tree.

5.2 Future Work

After studying lots of research paper on this topic and conducting a research by ourselves, we believe that there are so much scope left to research further on this topic. Firstly, we can take more distance methods into consideration and measure their accuracy. Secondly, combination of non-hierarchical and hierarchical clustering can be used in creating phylogenetic tree and measure accuracy and time complexity. Furthermore, we can apply our already achieved knowledge into disease prediction and calculate how well it can predict. For example, we can use our methodology in cancer cell prediction and try to measure its accuracy. To sum up, there are still many area left in this research domain where our findings can face tough criticism and we can look for a better answer.

References

1. Brazma, A., Vilo,J. (2000). Gene expression data analysis. FEBS letters, 480(1), 17-24.

2. Gouy, M., Guindon,S. Gascuel,O.(2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular biology and evolution , 27(2) , 221–224.

3. Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. Science, 284(5423), 2124-2128.

4. Hartigan J.A, Wong M.A .(1979) .Algorithm AS 136: A K-Means Clustering Algorithm.Journal of the Royal Statistical Society. Series C (Applied Statistics) .28(1) ,100–108.

5. Leif E. Peterson .(2009) .K-nearest neighbor. Scholarpedia journal.4(2). 1883.

6. Lance G. N. Williams W. T.(1967) .A General Theory of Classificatory Sorting Strategies:1. Hierarchical Systems.The Computer Journal.9(4).373–380.

7. Chidananda,K and Gowda,G.Krishna .1978 .Agglomerative clustering using the concept of mutual nearest neighbourhood.Pattern Recognition.10(2).105–112.

8. Anton, H., 1994, Elementary Linear Algebra (7th ed.), John Wiley Sons, pp. 170-171.

9. Craw S. 2017 Manhattan Distance. In: Sammut C., Webb G.I.(eds) Encyclopedia of Machine Learning and Data Mining.Springer, Boston, MA

10. Cyrus. D. Cantrell 2000. Modern Mathematical Methods for Physicists and Engineers. Cambridge University Press.

11. Everitt B 2011. Cluster analysis. Chichester, West Sussex, U.K: Wiley.

12. Pop, M., and Salzberg, S. L. 2008. Bioinformatics challenges of new sequencing technology. Trends in genetics, 24(3), 142-149.

13. Joe H. Ward Jr. 1963 Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association, 58:301, 236-244

14. Zuckerkandl, E., Pauling, L. (1965). Molecules as documents of evolutionary history. Journal of theoretical biology, 8(2), 357-366.

15. Janda, J. M., Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. Journal of clinical microbiology, 45(9), 2761-2764.

16. Woese, C. R. (2000). Interpreting the universal phylogenetic tree. Proceedings of the National Academy of Sciences, 97(15), 8392-8396.

17. Forterre, P., Philippe, H. (1999). Where is the root of the universal tree of life?. Bioessays, 21(10), 871-879.

18. Jain, R., Rivera, M. C., Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. Proceedings of the National Academy of Sciences, 96(7), 3801-3806.

19. Johnson, S. C. (1967). Hierarchical clustering schemes. Psychometrika, 32(3), 241-254.

20. Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single-link cluster method. The computer journal, 16(1), 30-34.

21. Defays, D. (1977). An efficient algorithm for a complete link method. The Computer Journal, 20(4), 364-366.

22. Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. The Computer Journal, 26(4), 354-359.

23. Ferrán, E. A., Pflugfelder, B. (1993). A hybrid method to cluster protein sequences based on statistics and artificial neural networks. Bioinformatics, 9(6), 671-680.

24. Drabl s, F. (1999). Clustering of non-polar contacts in proteins. Bioinformatics (Oxford, England), 15(6), 501-509.

25. Krause, A., Stoye, J., Vingron, M. (2005). Large scale hierarchical clustering of protein sequences. BMC bioinformatics, 6(1), 15.

26. de Souto, M. C., Costa, I. G., de Araujo, D. S., Ludermir, T. B., Schliep, A. (2008). Clustering cancer gene expression data: a comparative study. BMC bioinformatics, 9(1), 497.

27. Gront, D., Kolinski, A. (2005). HCPM—program for hierarchical clustering of protein models. Bioinformatics, 21(14), 3179-3180.

28. Aldecoa, R., Marín, I. (2010). Jerarca: Efficient analysis of complex networks using hierarchical clustering. PloS one, 5(7), e11585.

29. Ao, S. I., Yip, K., Ng, M., Cheung, D., Fong, P. Y., Melhado, I., Sham, P. C. (2004). CLUSTAG: hierarchical clustering and graph methods for selecting tag SNPs. Bioinformatics, 21(8), 1735-1736.

30. Jiang, D., Pei, J., Zhang, A. (2003, March). DHC: a density-based hierarchical clustering method for time series gene expression data. In Third IEEE Symposium on Bioinformatics and Bioengineering, 2003. Proceedings. (pp. 393-400). IEEE.

31. Zhu, Z., Ong, Y. S., Dash, M. (2007). Markov blanket-embedded genetic algorithm for gene selection. Pattern Recognition, 40(11), 3236-3248.

32. Lin X, Wang R, Zhang J, Sun X et al. 2015. Insights into Human Astrocyte Response to H5N1 Infection by Microarray Analysis. Viruses.

33. P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, B. Futcher. (1998) Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. Mol. Biol. Cell, 9:12 3273–3297.

34. Ichino, M., Yaguchi, H. (1994). Generalized Minkowski metrics for mixed feature-type data analysis. IEEE Transactions on Systems, Man, and Cybernetics, 24(4), 698-708.

Appendix A.

Appendix B.

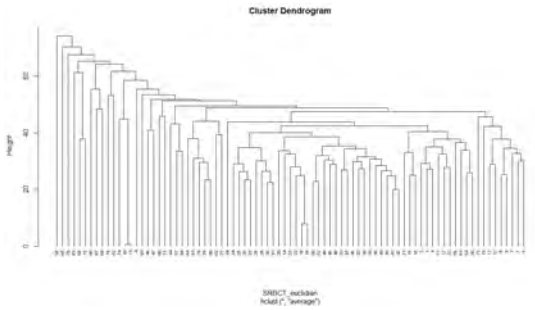

Figure 5.1: SRBCT Euclidean Average.

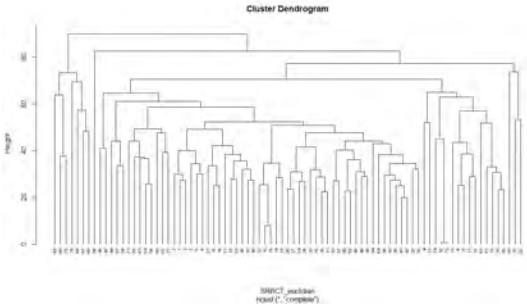

Figure 5.2: SRBCT Euclidean Complete.

Figure 5.3: SRBCT Euclidean Single.

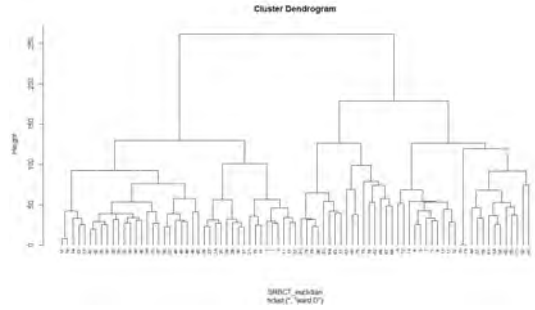

Figure 5.4: SRBCT Euclidean Ward.

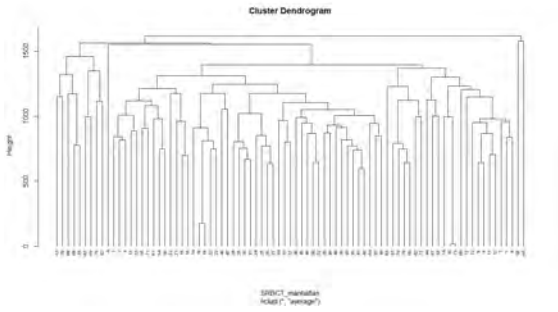

Figure 5.5: SRBCT Manhattan Average.

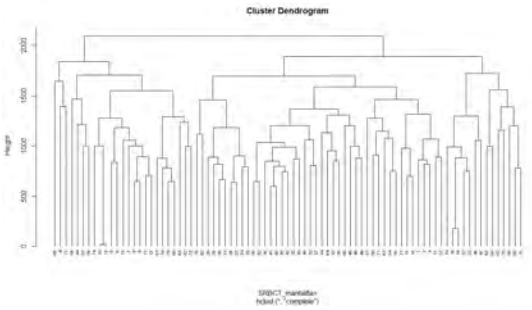

Figure 5.6: SRBCT Manhattan Complete.

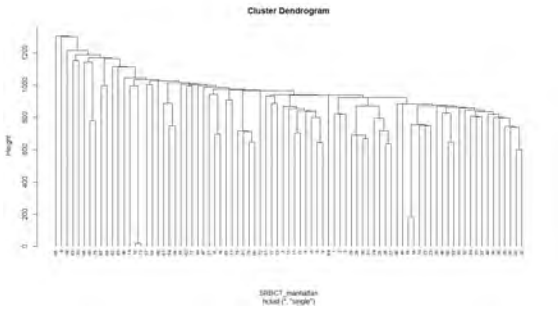

Figure 5.7: SRBCT Manhattan Single.

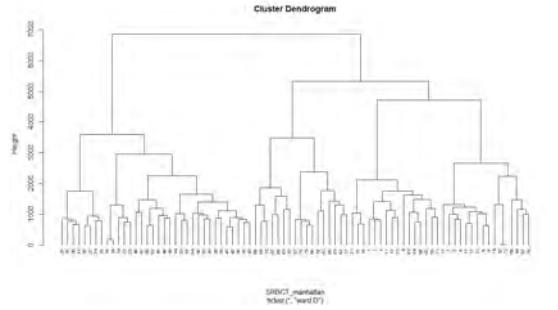

Figure 5.8: SRBCT Manhattan Ward.

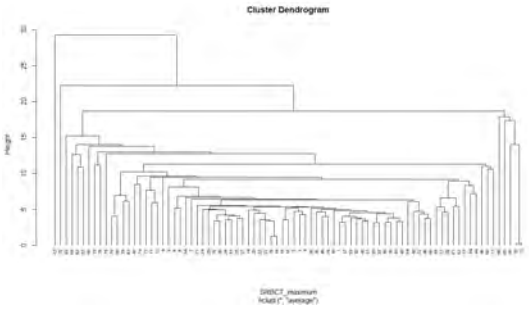

Figure 5.9: SRBCT Maximum Average.

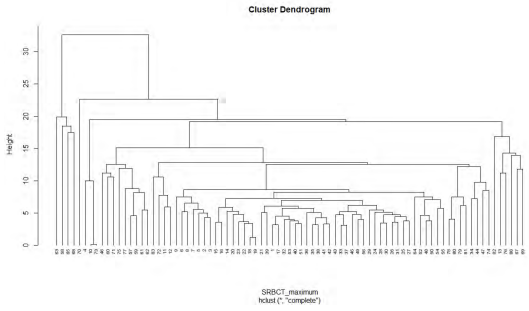

Figure 5.10: SRBCT Maximum Complete.

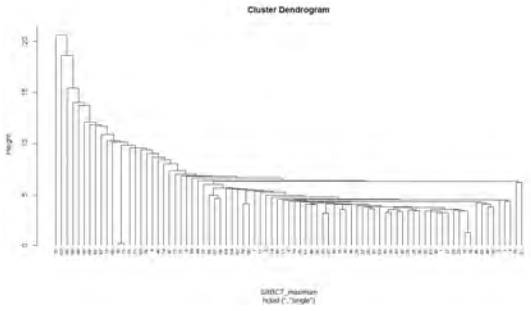

Figure 5.11: SRBCT Maximum Single.

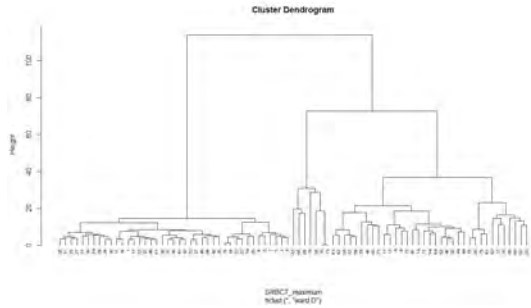

Figure 5.12: SRBCT Maximum Ward.

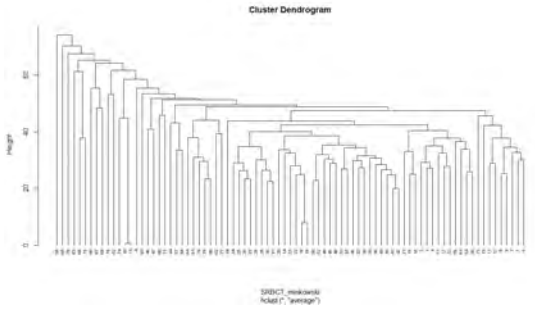

Figure 5.13: SRBCT Minkowski Average.

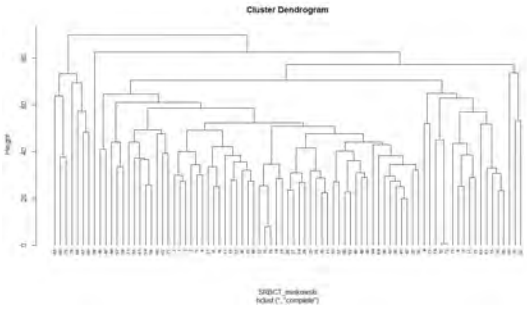

Figure 5.14: SRBCT Minkowski Complete.

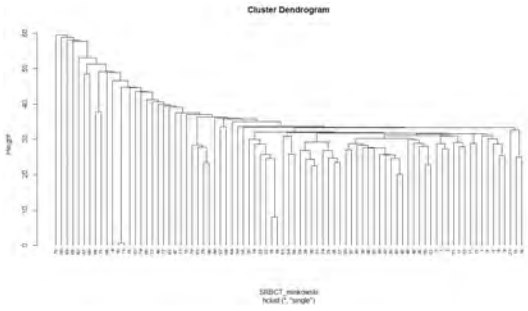

Figure 5.15: SRBCT Minkowski Single.

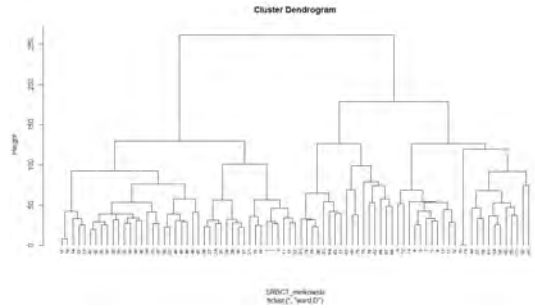

Figure 5.16: SRBCT Minkowski Ward.

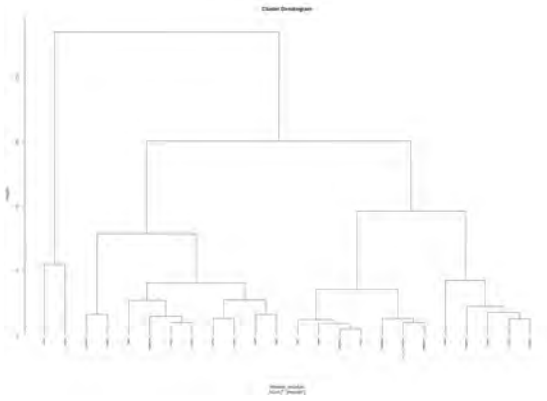

Figure 5.17: InfluenzaH5N1 Euclidean Average.

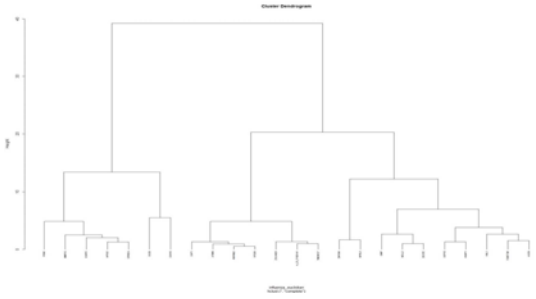

Figure 5.18: InfluenzaH5N1 Euclidean Complete.

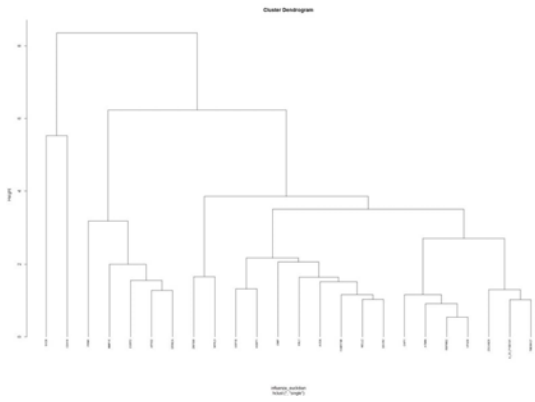

Figure 5.19: InfluenzaH5N1 Euclidean Single.

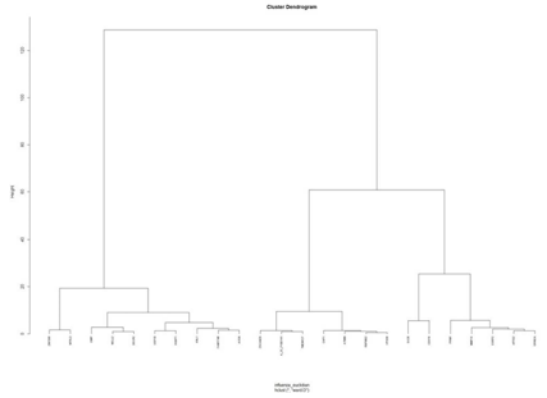

Figure 5.20: InfluenzaH5N1 Euclidean Ward.

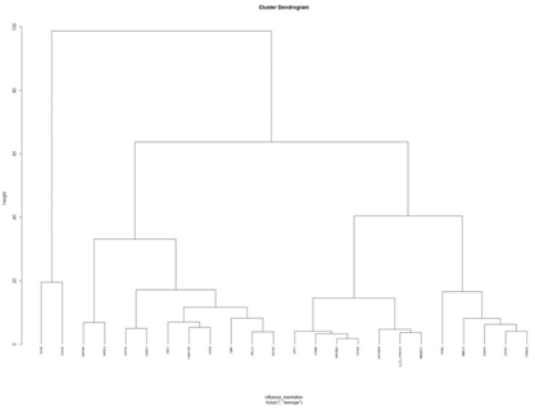

Figure 5.21: InfluenzaH5N1 Manhattan Average.

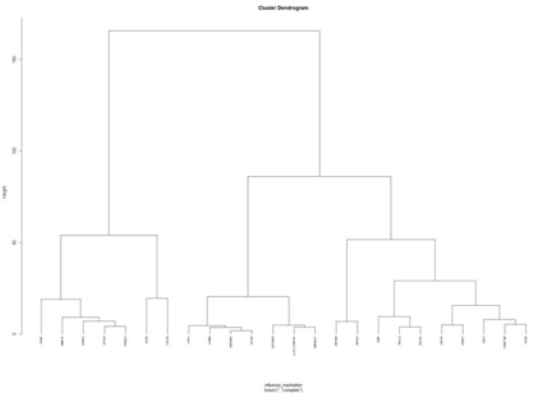

Figure 5.22: InfluenzaH5N1 Manhattan Complete.

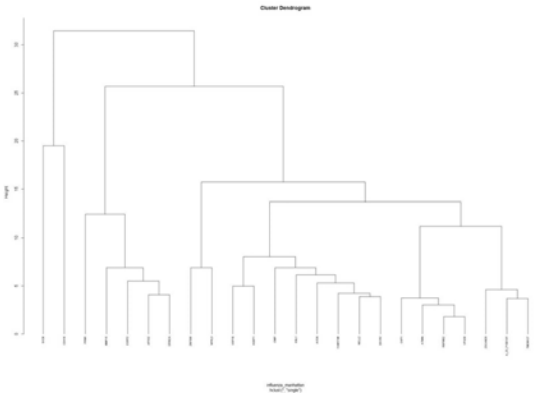

Figure 5.23: InfluenzaH5N1 Manhattan Single.

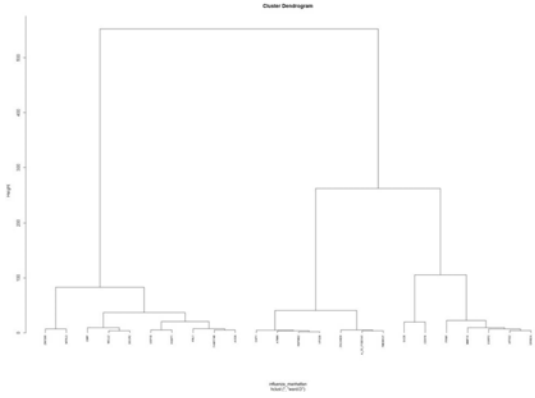

Figure 5.24: InfluenzaH5N1 Manhattan Ward.

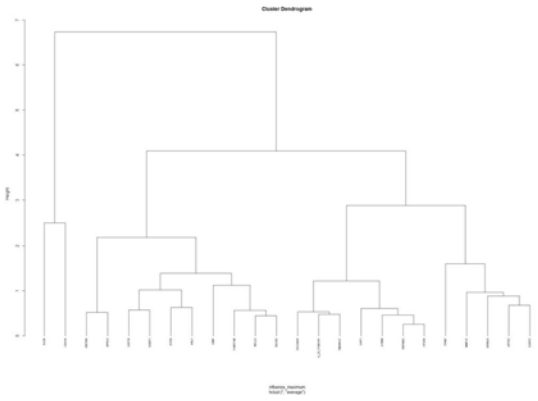

Figure 5.25: InfluenzaH5N1 Maximum Average.

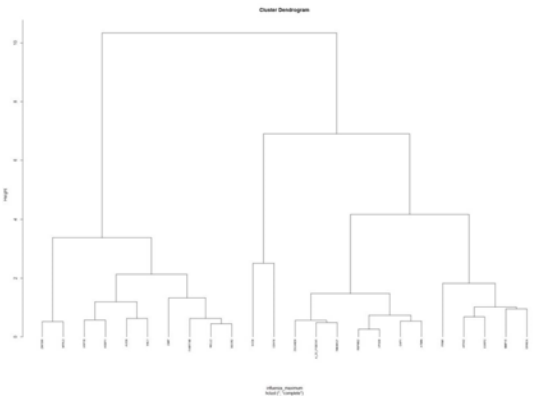

Figure 5.26: InfluenzaH5N1 Maximum Complete.

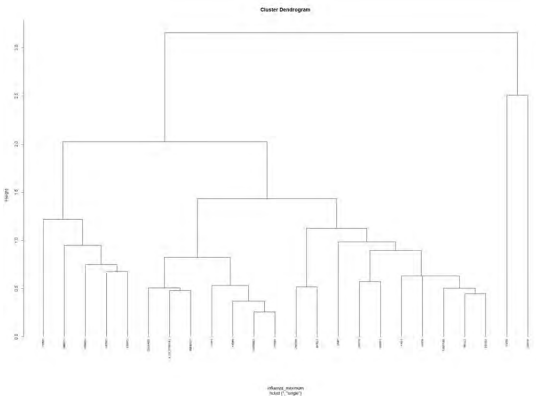

Figure 5.27: InfluenzaH5N1 Maximum Single.

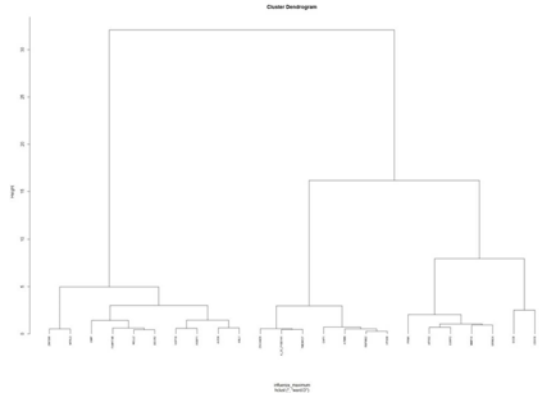

Figure 5.28: InfluenzaH5N1 Maximum Ward.

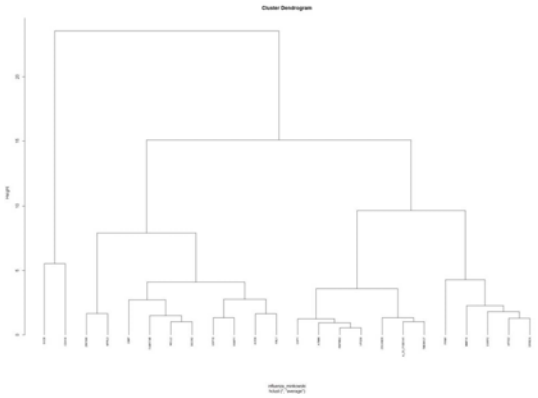

Figure 5.29: InfluenzaH5N1 Minkowski Average.

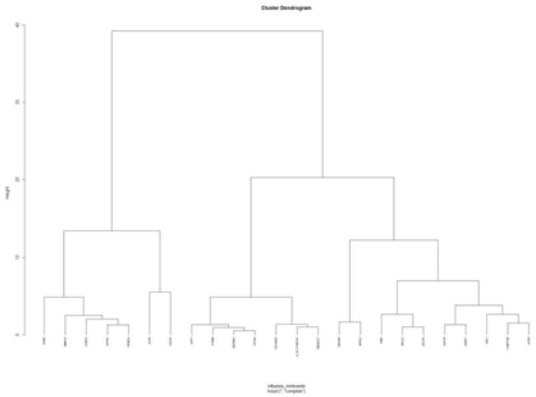

Figure 5.30: InfluenzaH5N1 Minkowski Complete.

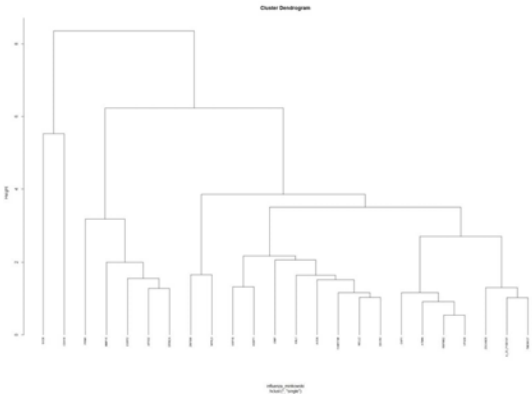

Figure 5.31: InfluenzaH5N1 Minkowski Single.

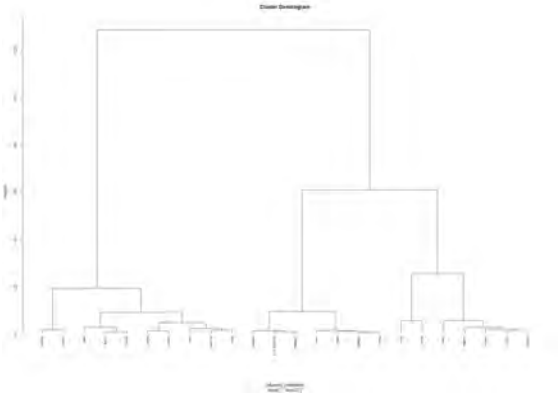

Figure 5.32: InfluenzaH5N1 Minkowski Ward.

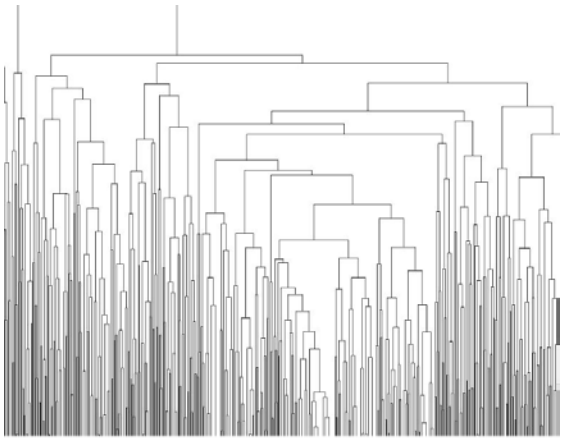

Figure 5.33: Spellman Euclidean Average (Clipped).

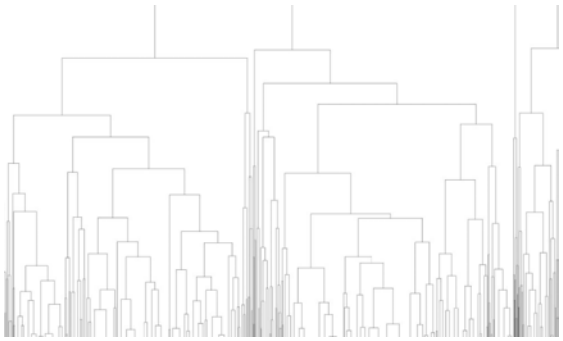

Figure 5.34: Spellman Euclidean Complete (Clipped).

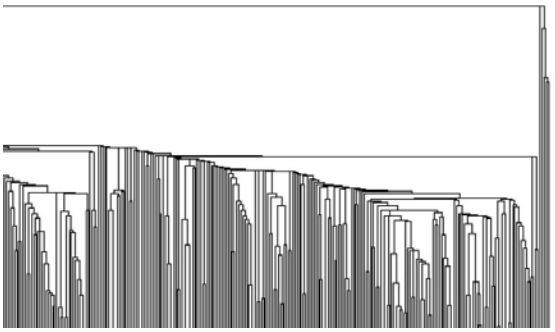

Figure 5.35: Spellman Euclidean Single (Clipped).

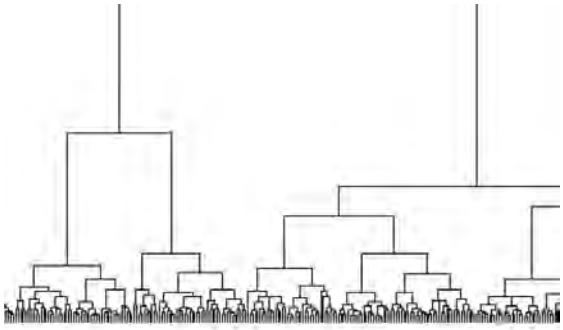

Figure 5.36: Spellman Euclidean Ward (Clipped).

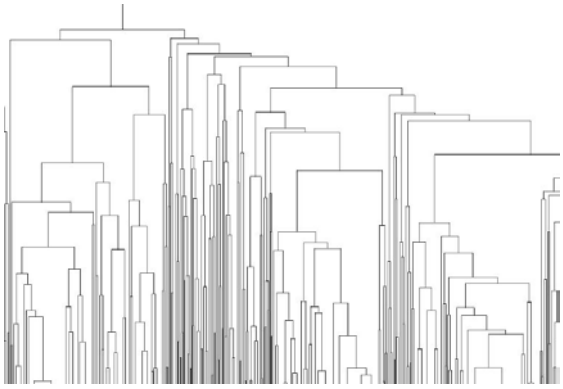

Figure 5.37: Spellman Manhattan Average (Clipped).

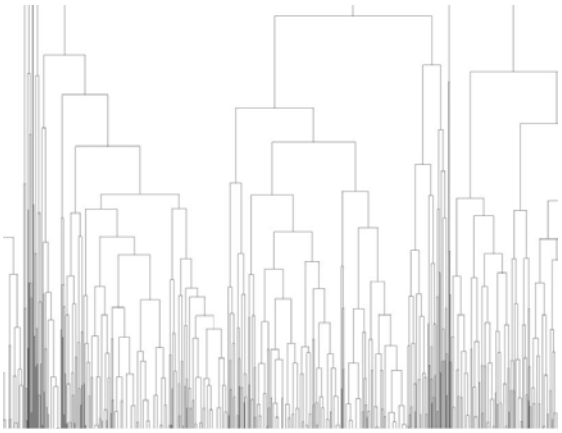

Figure 5.38: Spellman Manhattan Complete (Clipped).

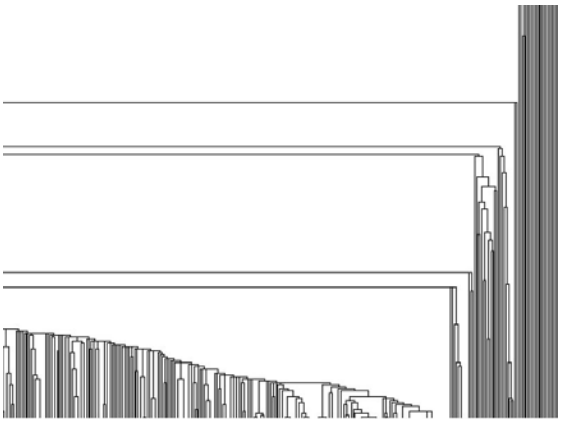

Figure 5.39: Spellman Manhattan Single (Clipped).

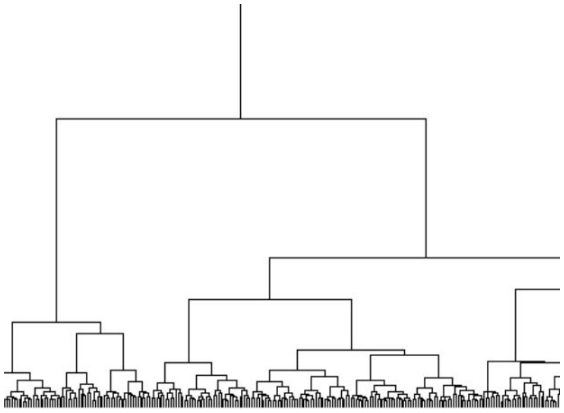

Figure 5.40: Spellman Manhattan Ward (Clipped).

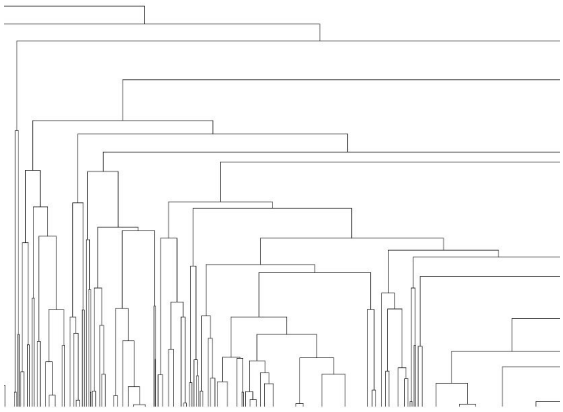

Figure 5.41: Spellman Maximum Average (Clipped).

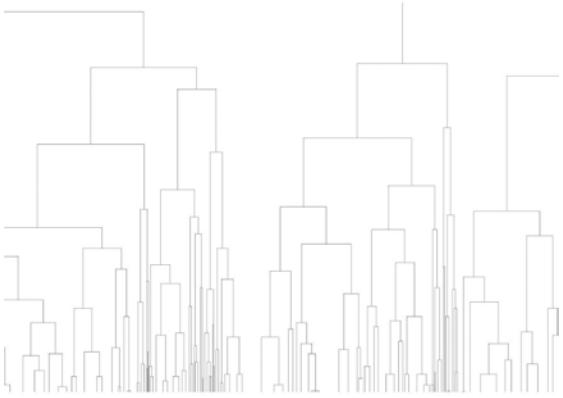

Figure 5.42: Spellman Maximum Complete (Clipped).

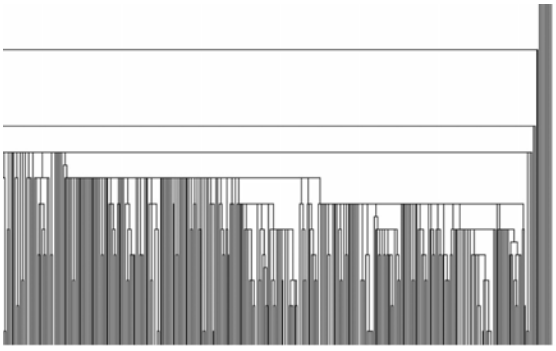

Figure 5.43: Spellman Maximum Single (Clipped).

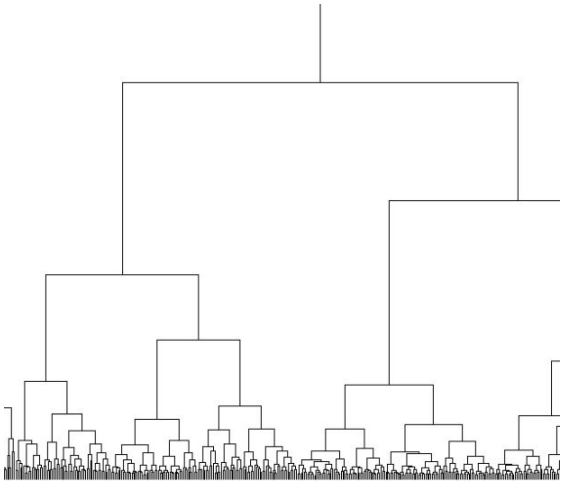

Figure 5.44: Spellman Maximum Ward (Clipped).

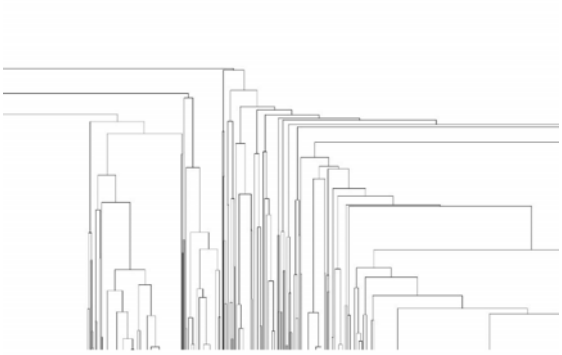

Figure 5.45: Spellman Minkowski Average (Clipped).

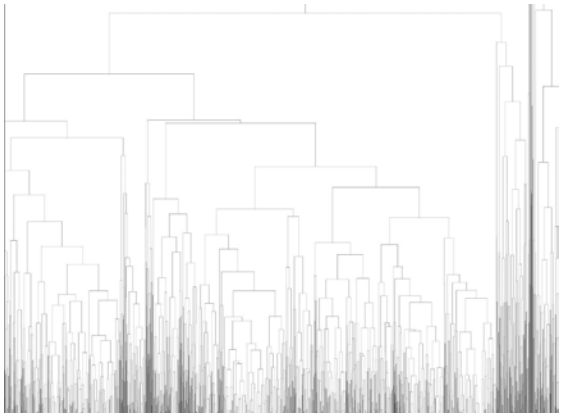

Figure 5.46: Spellman Minkowski Complete (Clipped).

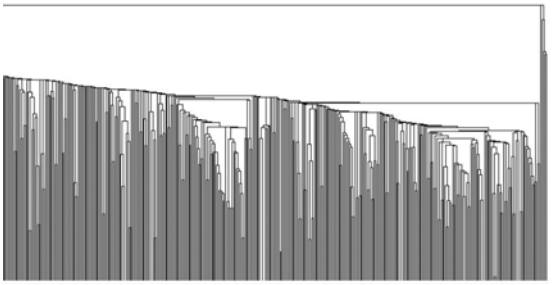

Figure 5.47: Spellman Minkowski Single (Clipped).

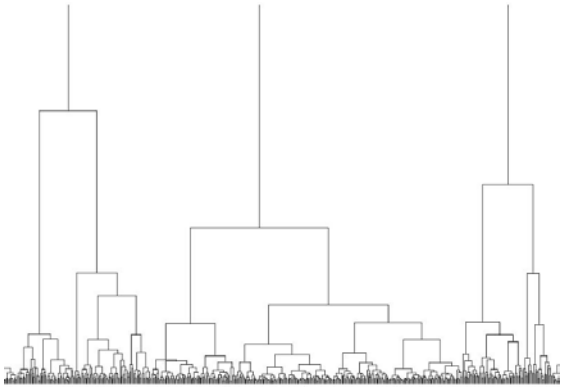

Figure 5.48: Spellman Minkowski Ward (Clipped).

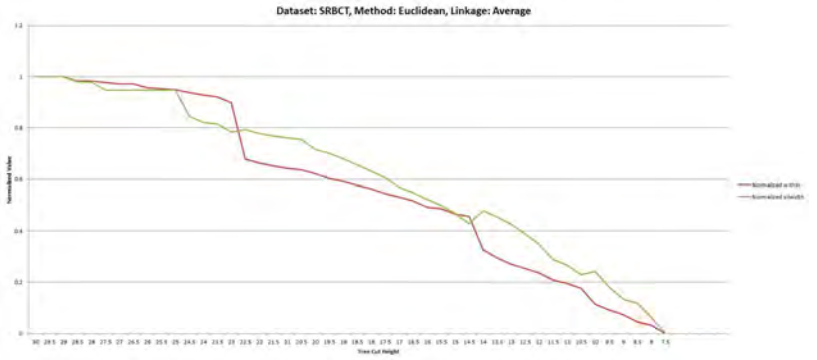

Figure 5.49: SRBCT Euclidean Average (Normalized).

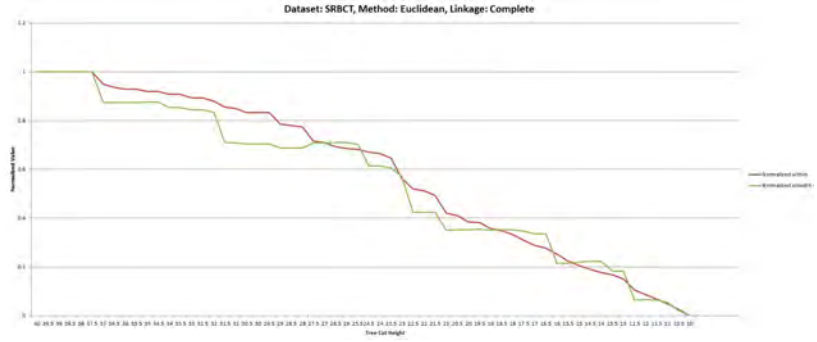

Figure 5.50: SRBCT Euclidean Complete (Normalized).

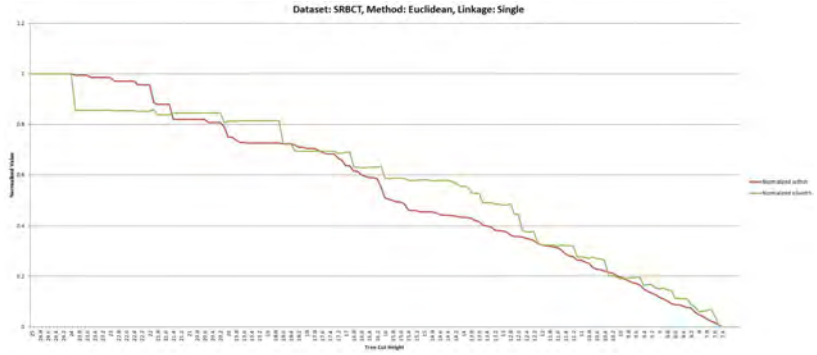

Figure 5.51: SRBCT Euclidean Single (Normalized).

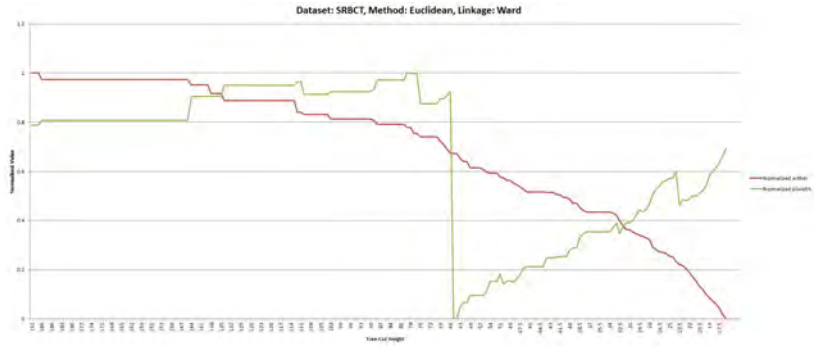

Figure 5.52: SRBCT Euclidean Ward (Normalized).

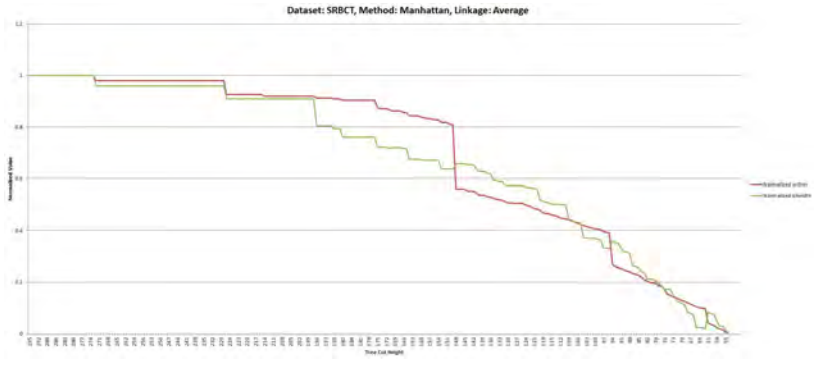

Figure 5.53: SRBCT Manhattan Average (Normalized).

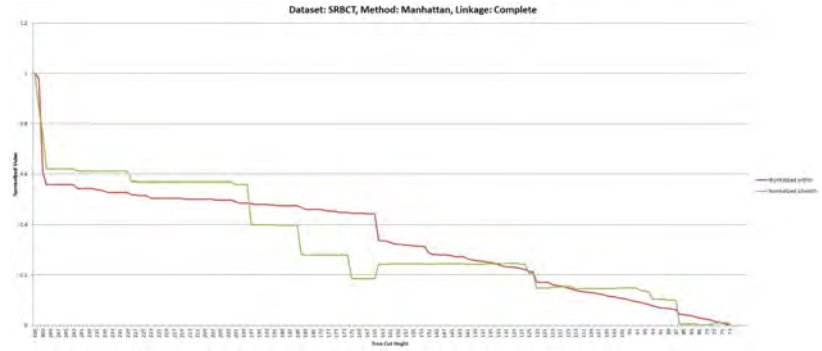

Figure 5.54: SRBCT Manhattan Complete (Normalized).

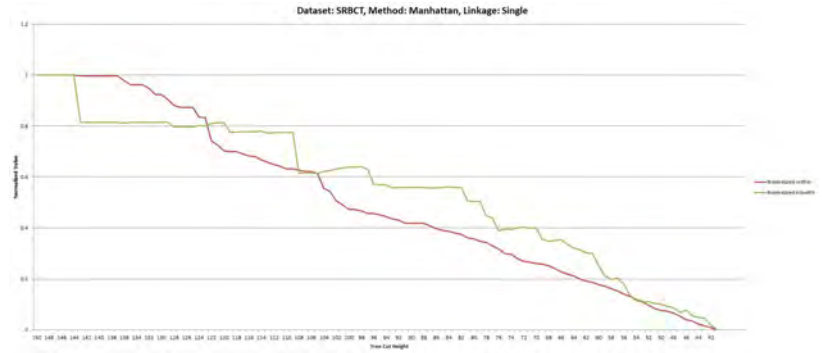

Figure 5.55: SRBCT Manhattan Single (Normalized).

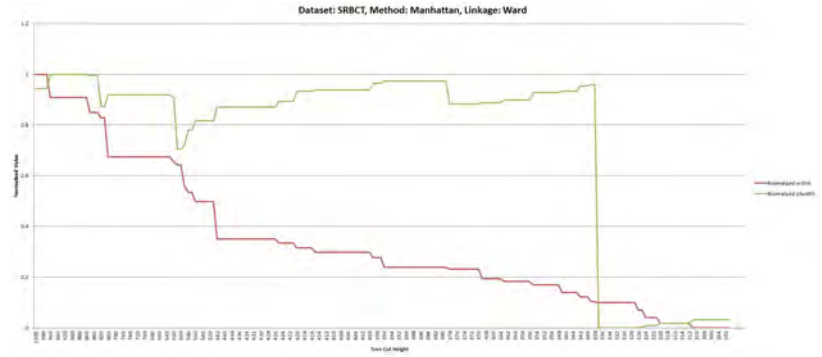

Figure 5.56: SRBCT Manhattan Ward (Normalized).

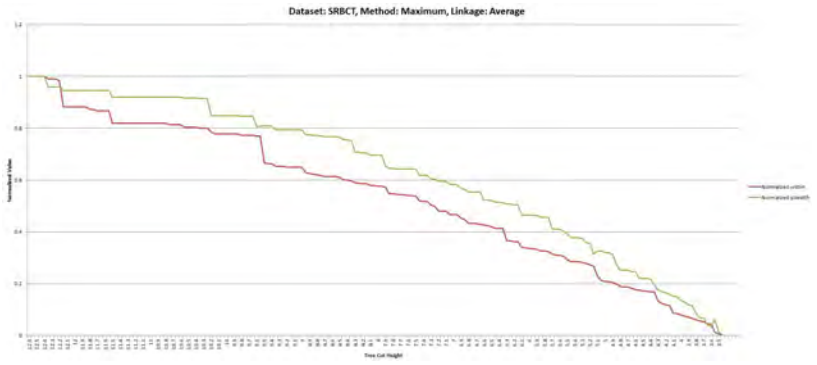

Figure 5.57: SRBCT Maximum Average (Normalized).

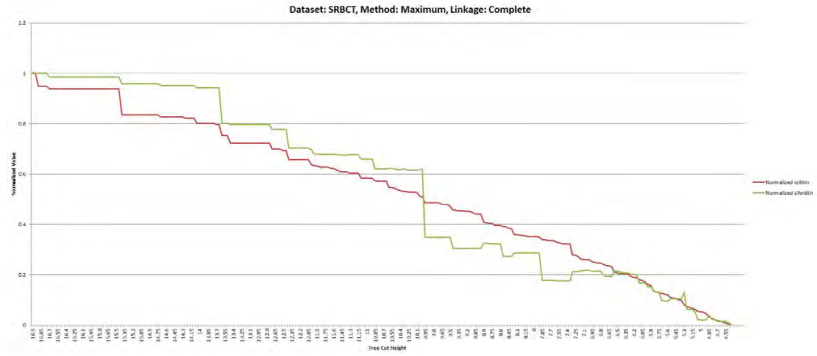

Figure 5.58: SRBCT Maximum Complete (Normalized).

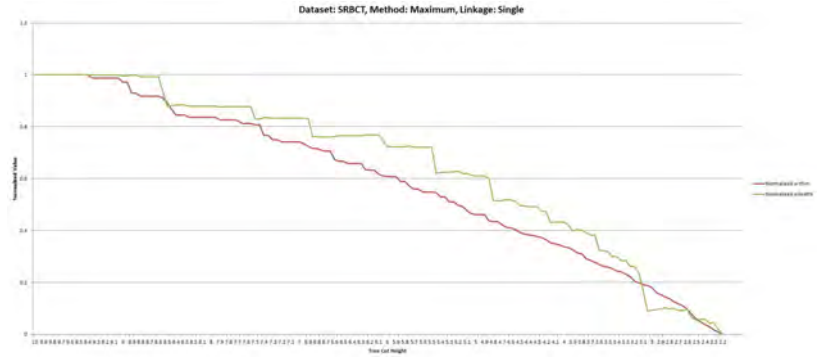

Figure 5.59: SRBCT Maximum Single (Normalized).

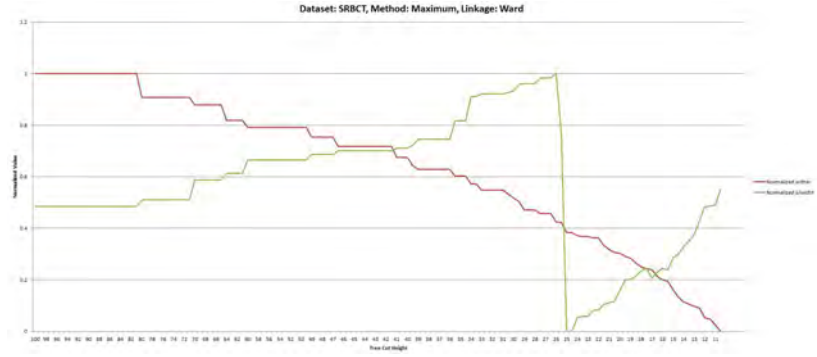

Figure 5.60: SRBCT Maximum Ward (Normalized).

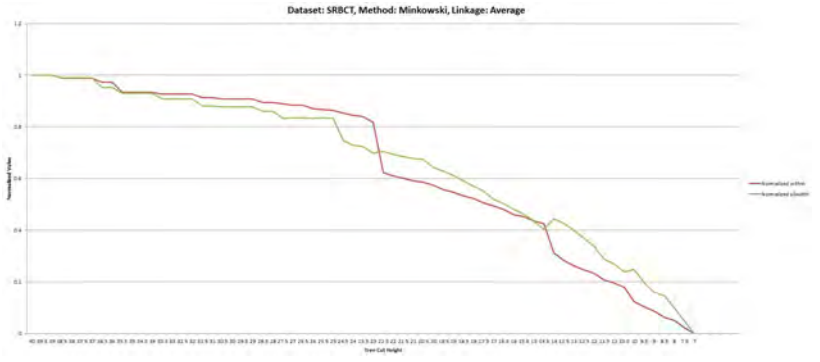

Figure 5.61: SRBCT Minkowski Average (Normalized).

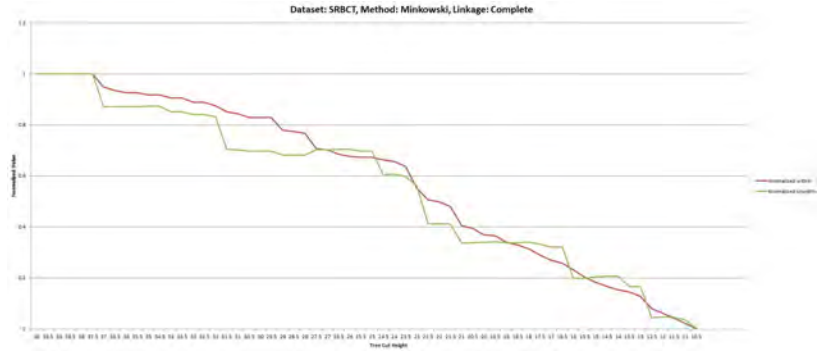

Figure 5.62: SRBCT Minkowski Complete (Normalized).

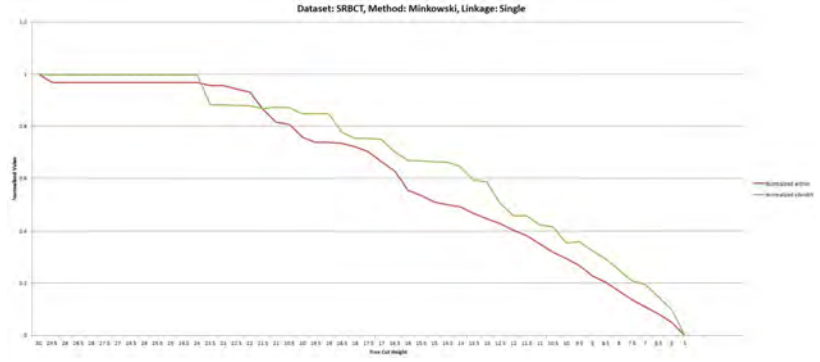

Figure 5.63: SRBCT Minkowski Single (Normalized).

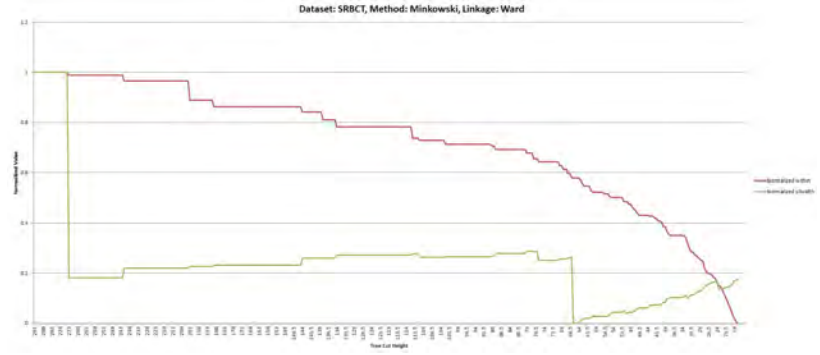

Figure 5.64: SRBCT Minkowski Ward (Normalized).

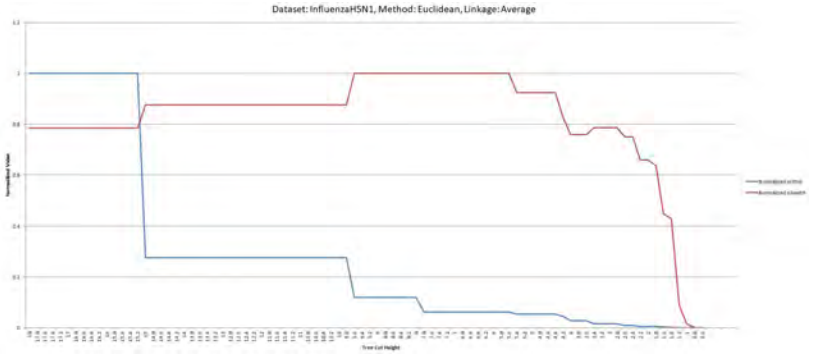

Figure 5.65: InfluenzaH5N1 Euclidean Average (Normalized).

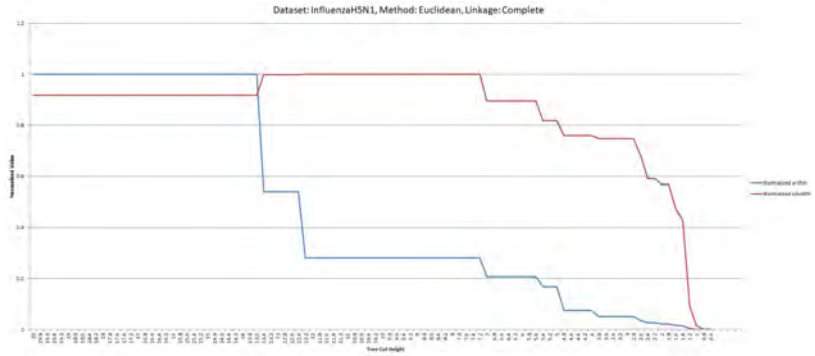

Figure 5.66: InfluenzaH5N1 Euclidean Complete (Normalized).

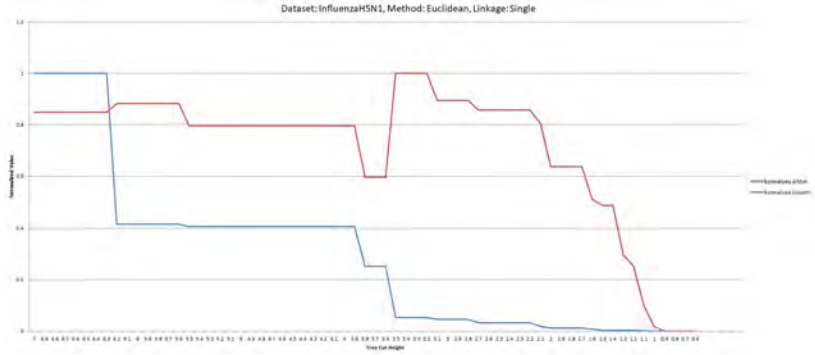

Figure 5.67: InfluenzaH5N1 Euclidean Single (Normalized).

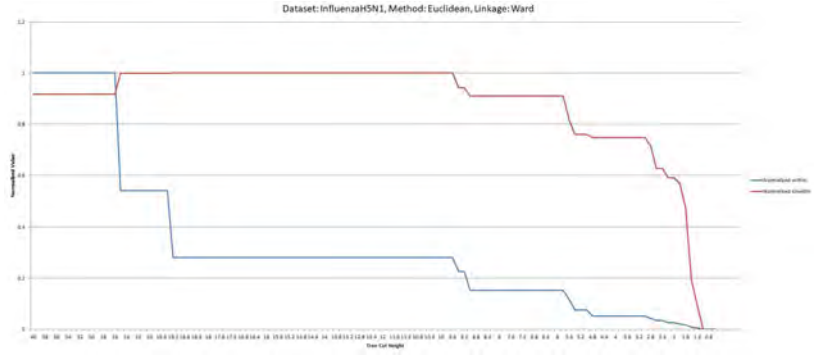

Figure 5.68: InfluenzaH5N1 Euclidean Ward (Normalized).

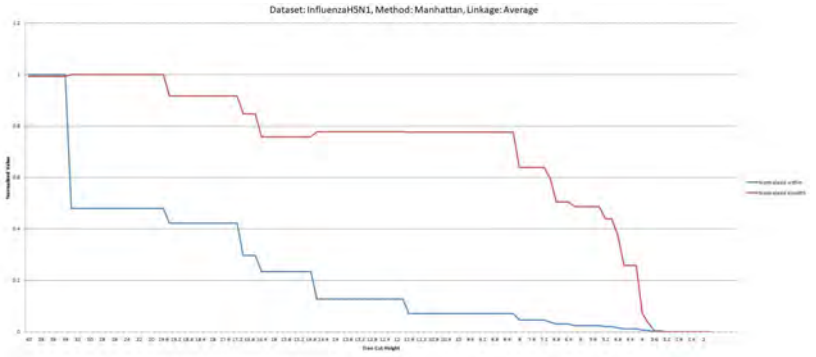

Figure 5.69: InfluenzaH5N1 Manhattan Average (Normalized).

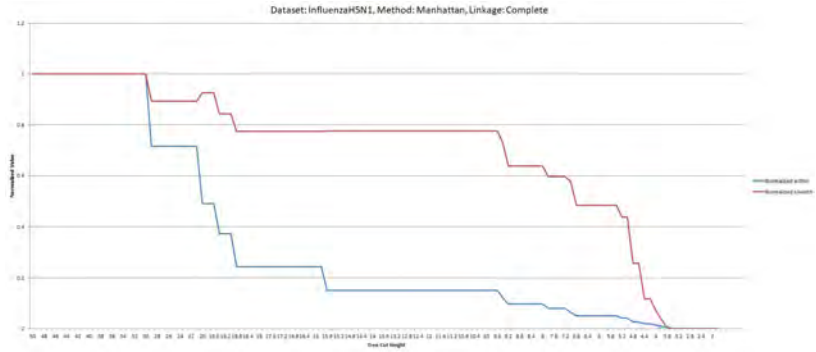

Figure 5.70: InfluenzaH5N1 Manhattan Complete (Normalized).

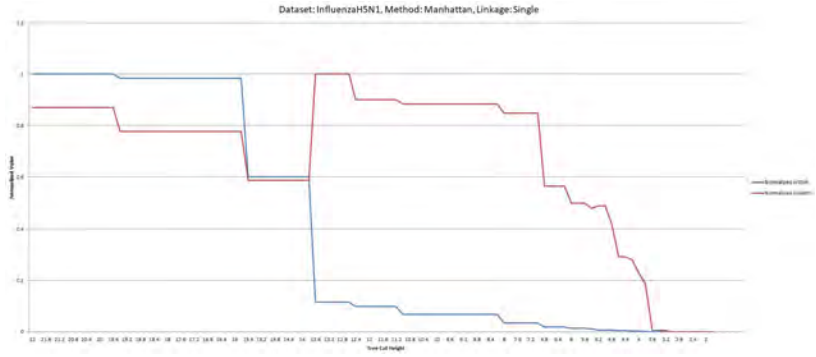

Figure 5.71: InfluenzaH5N1 Manhattan Single (Normalized).

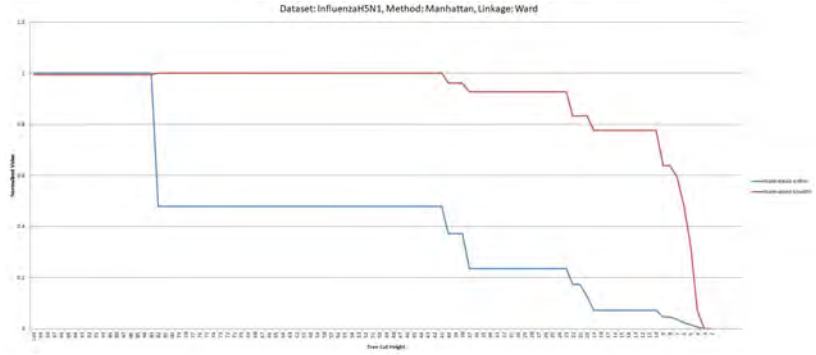

Figure 5.72: InfluenzaH5N1 Manhattan Ward (Normalized).

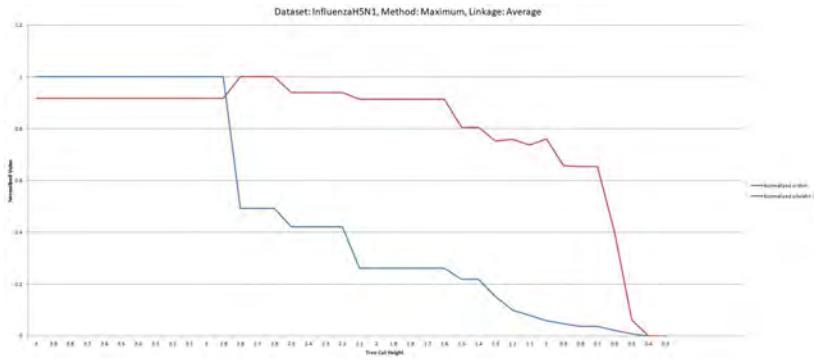

Figure 5.73: InfluenzaH5N1 Maximum Average (Normalized).

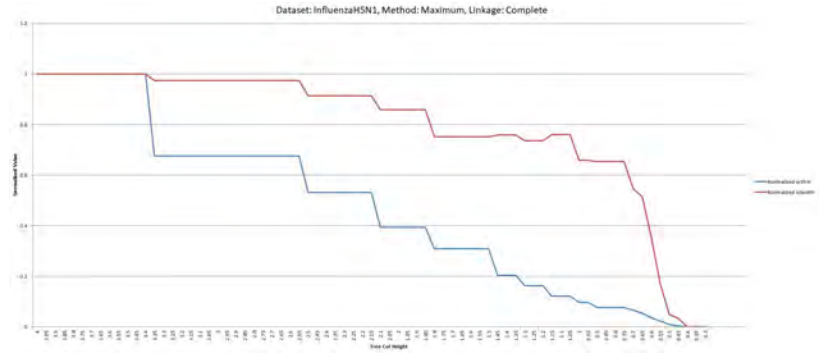

Figure 5.74: InfluenzaH5N1 Maximum Complete (Normalized).

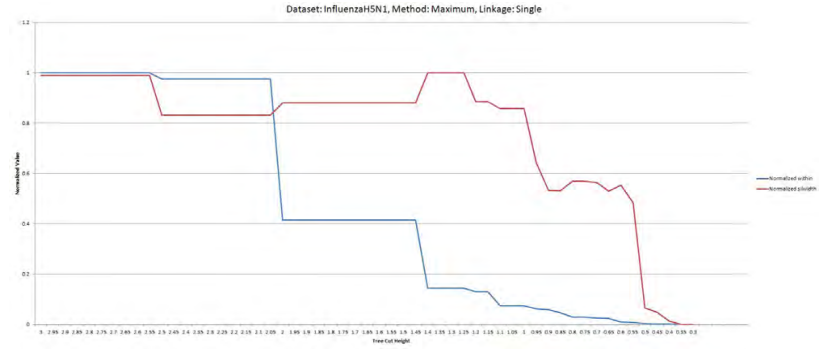

Figure 5.75: InfluenzaH5N1 Maximum Single (Normalized).

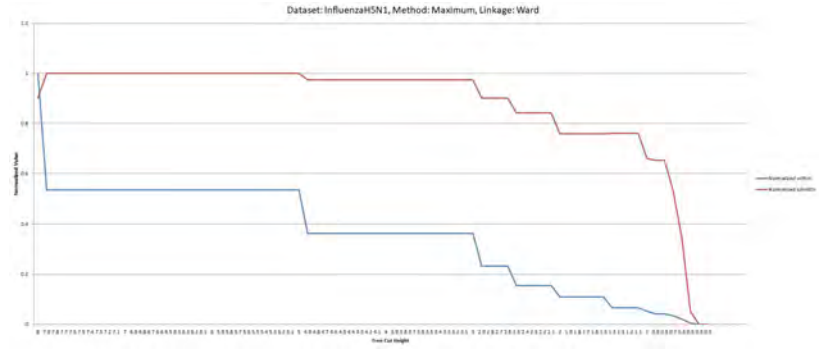

Figure 5.76: InfluenzaH5N1 Maximum Ward (Normalized).

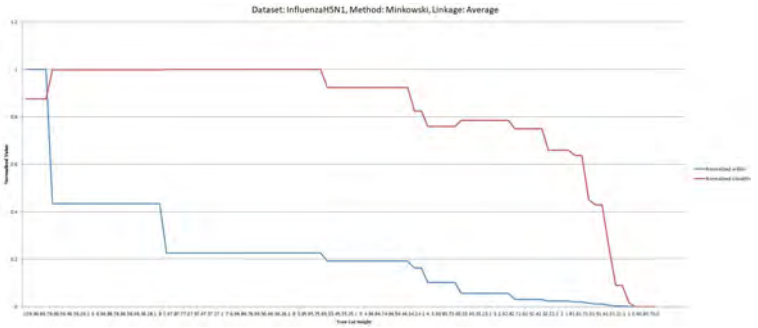

Figure 5.77: InfluenzaH5N1 Minkowski Average (Normalized).

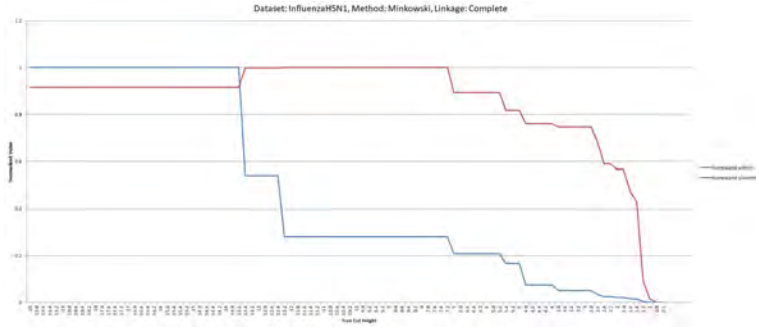

Figure 5.78: InfluenzaH5N1 Minkowski Complete (Normalized).

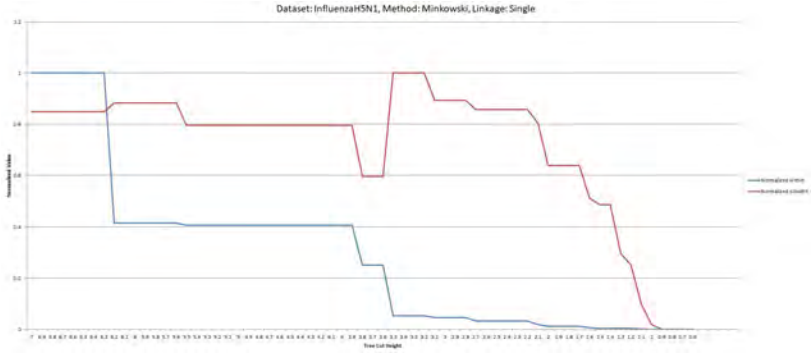

Figure 5.79: InfluenzaH5N1 Minkowski Single (Normalized).

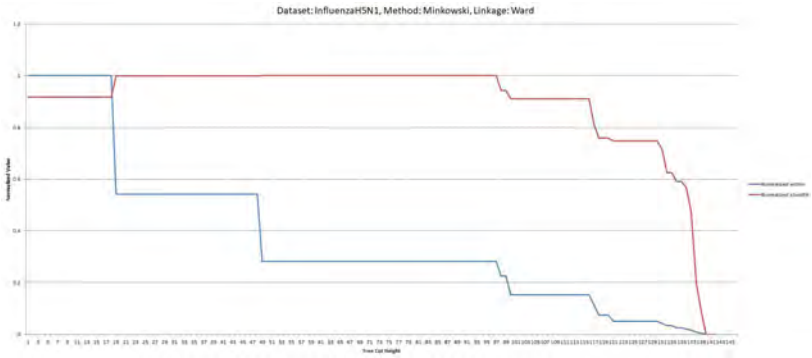

Figure 5.80: InfluenzaH5N1 Minkowski Ward (Normalized).

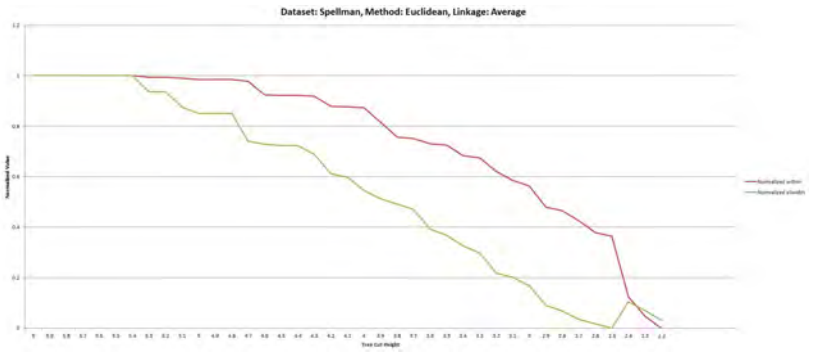

Figure 5.81: Spellman Euclidean Average (Normalized).

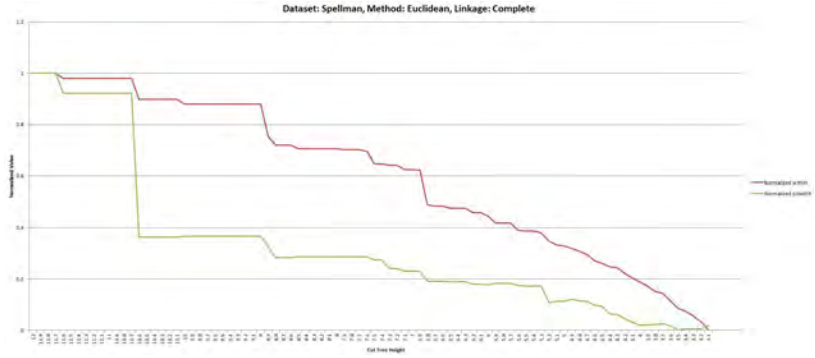

Figure 5.82: Spellman Euclidean Complete (Normalized).

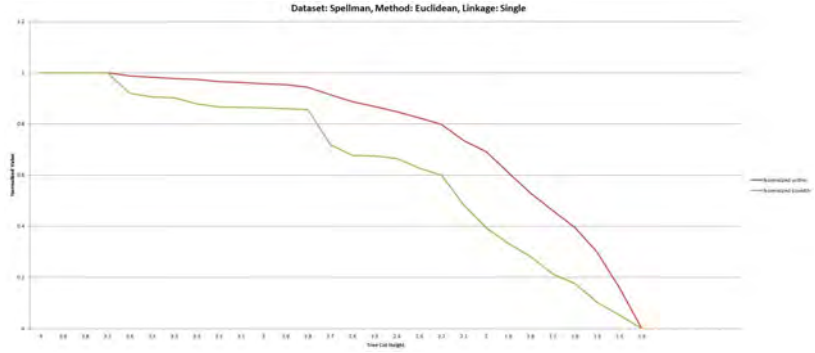

Figure 5.83: Spellman Euclidean Single (Normalized).

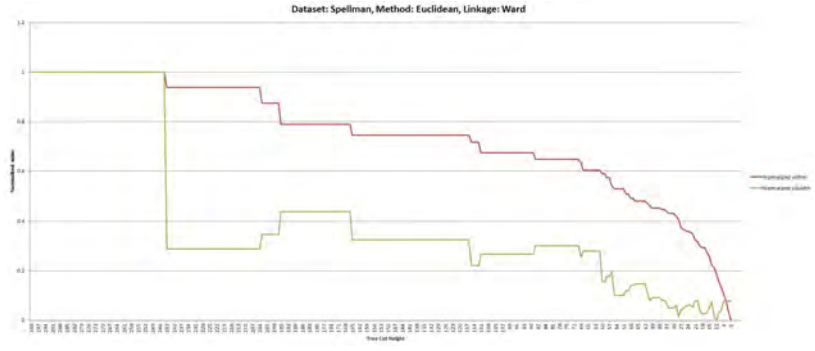

Figure 5.84: Spellman Euclidean Ward (Normalized).

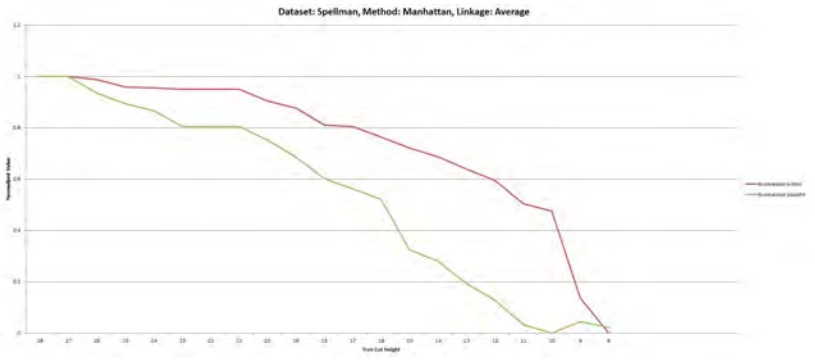

Figure 5.85: Spellman Manhattan Average (Normalized).

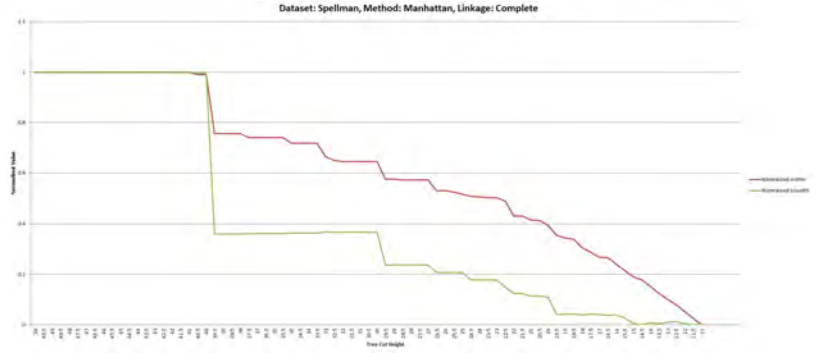

Figure 5.86: Spellman Manhattan Complete (Normalized).

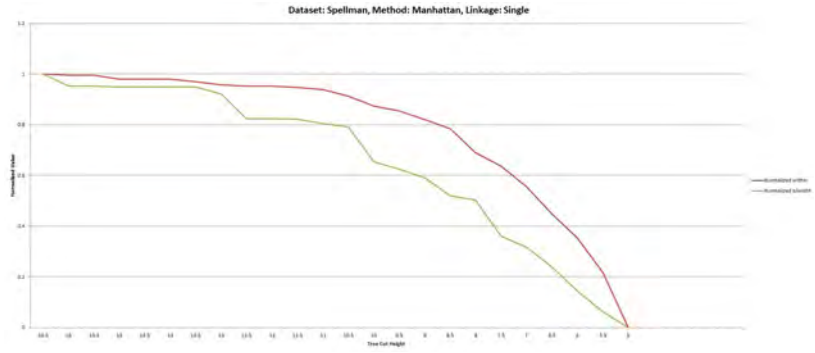

Figure 5.87: Spellman Manhattan Single (Normalized).

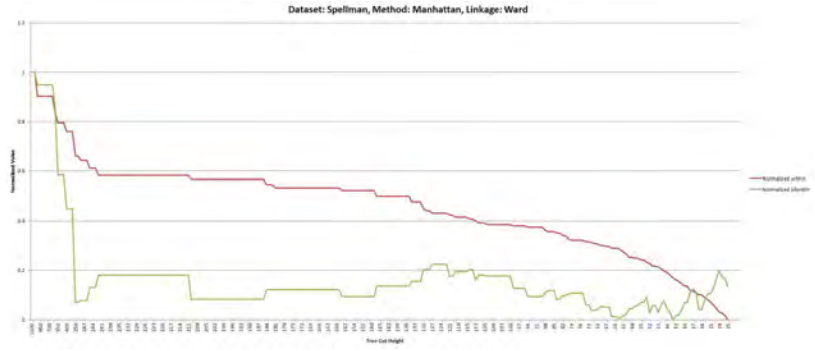

Figure 5.88: Spellman Manhattan Ward (Normalized).

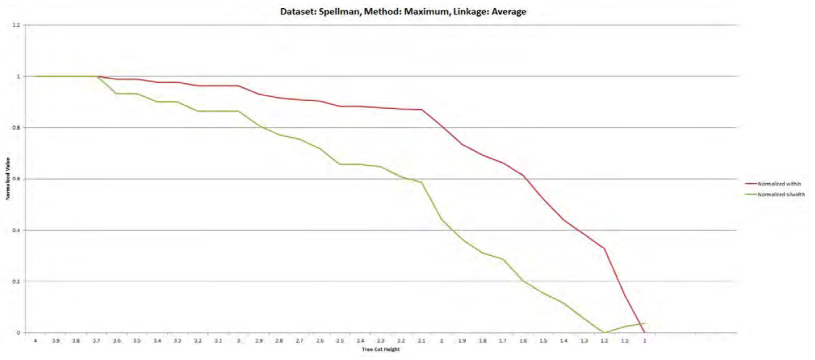

Figure 5.89: Spellman Maximum Average (Normalized).

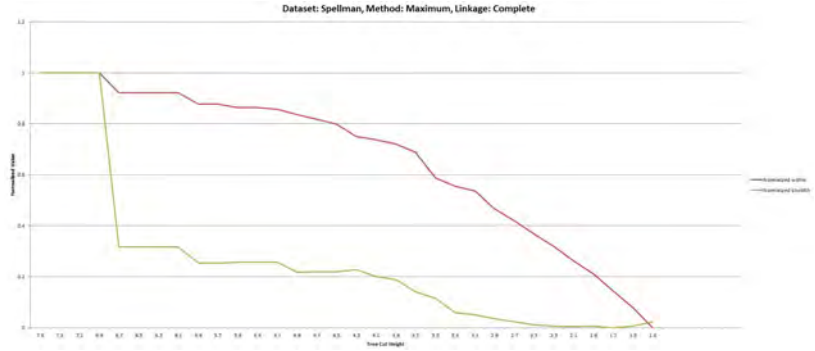

Figure 5.90: Spellman Maximum Complete (Normalized).

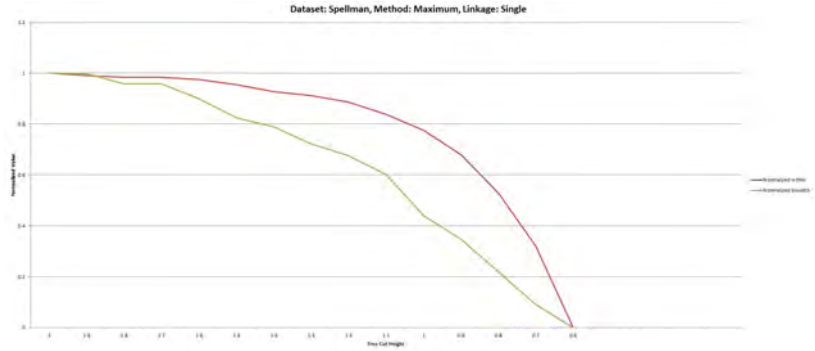

Figure 5.91: Spellman Maximum Single (Normalized).

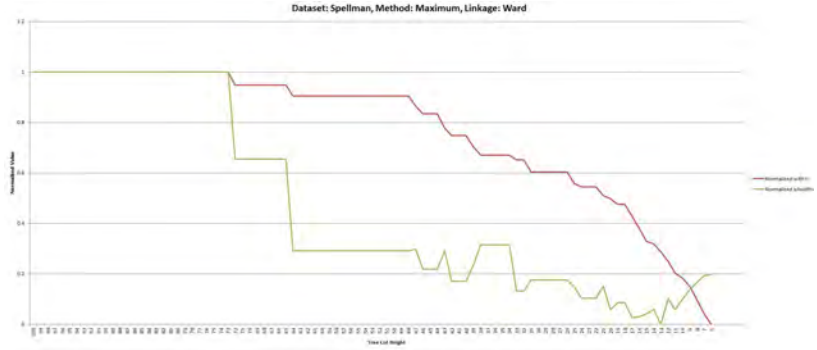

Figure 5.92: Spellman Maximum Ward (Normalized).

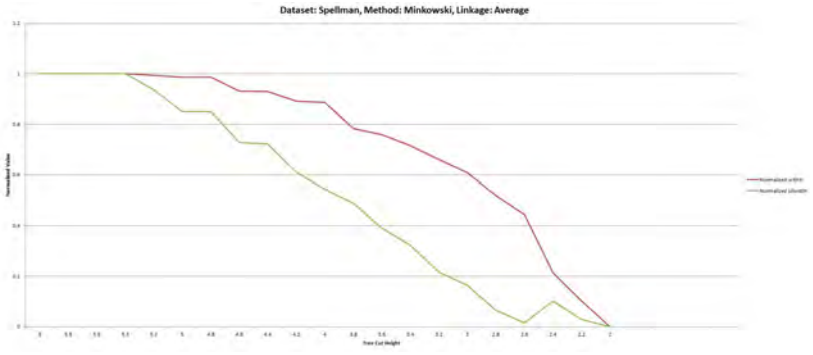

Figure 5.93: Spellman Minkowski Average (Normalized).

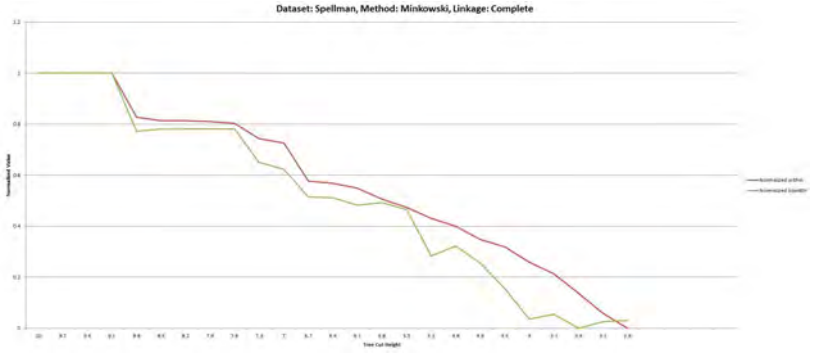

Figure 5.94: Spellman Minkowski Complete (Normalized).

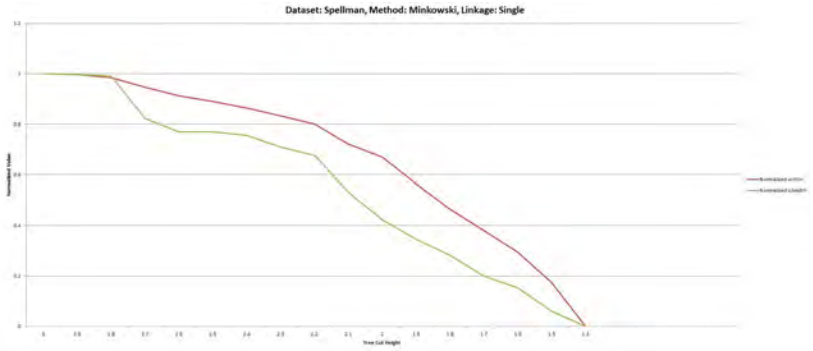

Figure 5.95: Spellman Minkowski Single (Normalized).

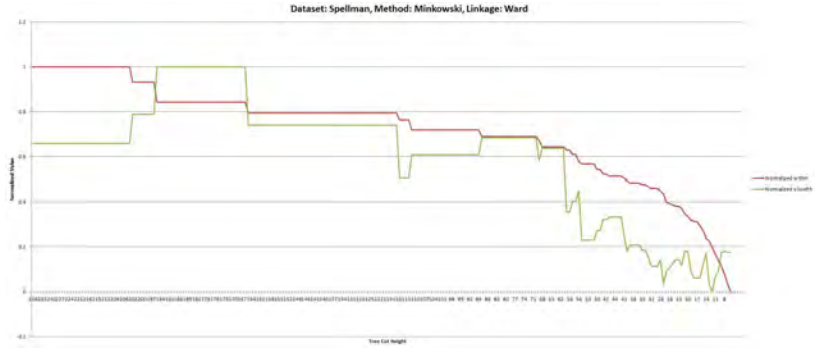

Figure 5.96: Spellman Minkowski Ward (Normalized).