Islamic University of Technology (IUT)

Department of Computer Science and Engineering (CSE)

Advanced Agglomerative Clustering Technique for Phylogenetic Classification Using Manhattan Distance

Authors

Redwan Karim Sony - 124401

&

Raihan Islam Arnob - 124424

Supervisor	**Co-Supervisor**
Prof. Dr. M.A. Mottalib	Rafsanjany Kushol
Head, Department of CSE	Lecturer, Department of CSE

A thesis submitted to the Department of CSE

in partial fulfillment of the requirements for the degree of B.Sc.

Engineering in CSE

Academic Year: 2015-16

November - 2016

Declaration of Authorship

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by Raihan Islam Arnob and Md. Redwan Karim Sony under the supervision of Professor Dr. M.A. Mottalib, Head of the Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh. It is also declared that neither of this thesis nor any part of this thesis has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Authors:

Raihan Islam Arnob
Student ID - 124424

Redwan Karim Sony
Student ID - 124401

Co-supervisor:

Rafsanjany Kushol
Lecturer
Department of Computer Science and Engineering
Islamic University of Technology (IUT)

Supervisor:

Dr. M.A. Mottalib

Head of the Department

Department of Computer Science and Engineering

Islamic University of Technology (IUT)

Acknowledgement

We would like to express our grateful appreciation for **Professor Dr. M.A. Mottalib**, Head of the Department, Department of Computer Science & Engineering, IUT for being our adviser and mentor. His motivation, suggestions and insights for this research have been invaluable. Without his support and proper guidance this research would never have been possible. His valuable opinion, time and input provided throughout the thesis work, from first phase of thesis topics introduction, subject selection, proposing algorithm, modification till the project implementation and finalization which helped us to do our thesis work in proper way. We are really grateful to him.

We are also grateful to **Mrs. Lipi Akter**, PhD Student, Department of Computer Science & Engineering, IUT and **Rafsanjany Kushol**, Lecturer, Department of Computer Science & Engineering, IUT for their valuable inspection and suggestions on our proposal of advanced agglomerative clustering technique.

Abstract

Trivial agglomerative hierarchical clustering technique has very high computational complexity O(n^3) and require more number of iterations to prepare the phylogenetic classification. To resolve this, we propose to use the **A**dvanced **A**gglomerative **C**lustering **T**echnique (AACT) using Manhattan Distance. The proposed method mainly aims to identify $\boldsymbol{m}$ number of distinct clusters over vast dataset with lower complexity and thus reducing the time complexity than existing methods. The proposed technique AACT consists of three phases. In the first phase, it partitions $\boldsymbol{l}$ number of clusters over the input dataset based on the traditional K-Means technique. In second phase, the proposed technique computes centroid over each individual clusters from the result of K-Means clustering and selects the representative gene closet to the centroid. In the final phase, the proposed method uses SLINK technique for tracing $\boldsymbol{m}$ distinct cluster over $\boldsymbol{l}$ clusters. Experiment result shows that the proposed technique is by far very fast than traditional agglomerative approach and even faster than some techniques that were developed recently for faster analysis.

Contents

1 Introduction

1.1 Overview

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data. Bioinformatics is both an umbrella term for the body of biological studies that use computer programming as part of their methodology, as well as a reference to specific analysis ”pipelines” that are repeatedly used, particularly in the field of genomics. Common uses of bioinformatics include the identification of candidate genes and nucleotides (SNPs). Often, such identification is made with the aim of better understanding the genetic basis of disease, unique adaptations, desirable properties (esp. in agricultural species), or differences between populations. In a less formal way, bioinformatics also tries to understand the organizational principles within nucleic acid and protein sequences.

Bioinformatics has become an important part of many areas of biology. In experimental molecular biology, bioinformatics techniques such as image and signal processing allow, extraction of useful results from large amounts of raw data. In the field of genetics and genomics, it aids in sequencing and annotating genomes and their observed mutations. It plays a role in the text mining of biological literature and the development of biological and gene ontologies to organize and query biological data. It also plays a role in the analysis of gene and protein expression and regulation. Bio-informatics tools aid in the comparison of genetic and genomic data and more generally in the understanding of evolutionary aspects of molecular biology. At a more integrative level, it helps analyze and catalogue the biological pathways and networks that are an important part of systems biology. In structural biology, it aids in the simulation and modeling of DNA, RNA, and protein structures as well as molecular interactions.

Metagenomics[1] is a new field of study that provides a deeper insight into the microbial world compared to the traditional single genome sequencing technologies. Traditional methods for studying individual genomes are well developed. However, they are not appropriate for studying microbial samples from the environment because traditional methods rely upon cultivated clonal cultures while more than 99% of bacteria are unknown and cannot be cultivated and isolated[2]. Metagenomics use technologies that sequence uncultured bacterial genomes in an environment sample directly[3] and thus makes it possible to study organisms which cannot be isolated or are difficult to grow in lab. It provides hope for a better understanding of natural microbial diversity as well as their roles and interactions. It also creates new opportunities for medicine, biotechnology, agricultural studies and ecology.

Recent development in phylogenetic classification has resulted in methods like improved agglomerative clustering technique [4] which takes into account the power of K-means clustering to shrink the vast data set and later on use the hierarchical clustering to derive the final result. In short, it takes the output of K-means clustering and puts it as input of agglomerative clustering technique to obtain the final result.

1.2 Problem Statement

The primary goals of metagenomics are to describe the populations of microorganisms and to identify their roles in the environment. Ideally, we want to identify complete genome sequences of all organisms present in the sample. However, metagenomic data is very complex, containing a large number of sequences of reads from many species. The number of species and their abundances levels are unknown. The assembly of a single genome is already a difficult problem, complicated by repeats and sequencing errors which may lead to high fragmentation of contigs and mis-assembly. In a metagenomic data, in addition to repeats within individual genomes, genomes of closely related species may also share homologous sequences, which could lead to even more complex repeat patterns that are very difficult to resolve. A lot of research has been done for assembling single genomes[5, 6, 7, 8]. But due to the lack of research on metagenomic assemblers, assemblers designed for individual genomes are routinely used in metagenomic projects [9, 10]. It has been shown that these assemblers may lead not only to mis-assembly, but also severe fragmentation of contigs [11]. A plausible approach is to improve the performance of such assemblers is to separate reads from different organisms present in the dataset before the assembly.

1.3 Motivation & Scopes

Microorganism can be found in almost every environment of the earth's biosphere and are responsible for numerous biological activities including carbon and nitrogen cycling [12], organic containment remediation[9, 13, 14] and human health and disease. Many human disorders, such as type 2 diabetes (T2D), obesity, dental cavities, cancer and some immune-related diseases are known to be related with a single or group of microorganisms[5, 6, 7, 8, 11, 15, 16]. In addition, different strains within the same species may have completely different impacts on human health, such as Escherichia Coli, which is highly virulent E. coli strain, whereas most other strains in this same species are non-pathogenic. Thus characterization and identification of microbial strains/species in the environment and individual human hosts is of crucial importance to reveal human-microbial interactions, especially for patients with microbial-mediated disorders.

Although different technologies have been developed, the characterization and identification of known microorganisms at strain/species levels remain challenging, mainly due to lack of high resolution tools and the extremely diverse nature of microbial communities. Currently, the most commonly used approach to characterize and identify microorganisms in complex environment is to sequence 16S ribosomal RNA (rRNA) gene amplicons using universally conserved primers [17]. However, owing to the high similarity of 16S rRNA gene sequences among different microorganisms, this approach can only confidently identify microorganisms at high taxonomic levels but not at the species/strain level, although species identification had been attempted in a few studies with less complex communities [14, 15]. Even at the genus level, resolution problems with 16S rRNA gene sequences have been reported by many investigators [16]. Therefore, it is necessary to use other molecular markers to identify and characterize microorganisms at the strain/species level in complex environments.

To simply put the motivation of our work, we can say-

- Firstly, Metagenomics has become a major issue in bioinformatics.
- Secondly, 99% of microorganism present in many natural environments are not readily cultivable and therefore are not assessable.
- Thirdly, novel genes are high potential for use in pharmaceutical products or production processes and those genes can be identified clearly from metagenome.
- Finally, metagenome study is increasing research scope in bioinformatics.

1.4 Research Challenges

Metagenomes contain a large amount of data and this data are totally unstructured as we saw those data are collected directly from environment. So, for processing metagenome and finally finding valuable information from those we will be needing an efficient algorithm. Again, as the data of metagenome are unstructured, so it is an unsupervised learning. As a result, for processing these data we need clustering.

Many computational tools have been developed for separating reads from different species or groups of related species. Some of the tools also estimate the abundance levels and genome size of species. These tools are usually classified as similarity based methods is to analyze the taxonomic content of a sample. Small scale approaches involving 16S rRNAs and 18S rRNAs are commonly used to determine evolutionary relationships by analyzing fragments that contain marker genes and comparing them with known marker genes. These methods take advantage of small number of fragments containing marker genes and require reads to have at least 1000 bps(base pairs). Two other tools handle a larger number of fragments: MEGAN and CARMA. MEGAN aligns reads to databases of known sequences using BLAST and assigns reads to taxa by lowest common ancestor approach. CARMA performs phylogenetic classification of unassembled reads using all Pfam domains and protein families as phylogenetic markers. These two methods work for very short reads. However, a large fraction of sequences may remain unclassified by these methods because of the absence of closely related sequences in the databases.

1.5 Thesis Outline

In Chapter 1 we have discussed our study in a precise and concise manner. Chapter 2 deals with the necessary literature review for our study and there development so far. In Chapter 3 we have stated the skeleton of our proposed method, proposed algorithm and also the flowchart to provide a detail insight of the working procedure of our proposed method **A**dvanced **A**gglomerative **C**lustering **T**echnique(AACT) using Manhattan Distance. Chapter 4 shows the results and comparative analysis of successful implementation of our proposed method. The final segment of this study contains all the references and credits used.

2 Literature Review

2.1 Metagenomics

Metagenomics is the study of multiple genomes i.e. metagenomes are taken directly from the environment. While the traditional methods, in which organisms were cultured in predetermined media under laboratory conditions, were able to produce a diversity profile; the missed the vast majority of biodiversity present in the environment. Recently, Kevin Chen and Lior Pachter (Researchers at the University of California, Berkeley) defined metagenomics as "the application of modern genomics techniques to the study of the communities of microbial organism directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species." [1]

Metagenomics is currently the only way to study genetic diversity present in the viral communities as they do not contain any universal phylogenetic marker (like 16S RNA for bacteria) which are typically used to culture bacterial organisms. Culturing the host and then infecting them with specific viruses or viral DNA obtained from the environment in the laboratory condition is not yet streamlined. However, the Metagenome sample obtained from the environment directly represent communities of population as opposed to isolated populations and thus, Metagenomics may help reveal information about how the population co-evolve. One key step in understanding our micro-biota is to identify lineages that have co-evolved with humans (or with mammals in general), and to identify the genomic consequences of this co-evolution. Co-evolution between a host and a beneficial symbiont, or a pathogen, is defined as reciprocal adaptation of each lineage in response to the other. For example, genetic changes that increase production of a metabolite by an intestinal bacterium may trigger selection of changes in the host genome that promote uptake or prevent synthesis of that metabolite.

2.2 Advancement is Sequencing Technologies

With the advent of powerful and economic next generation sequencing technologies such as Sanger sequencing or massively parallel pyrosequencing, metagenomics has become more popular. Sanger sequencing is based on chain termination with di-oxy-nucleotides whereas pyrosequencing is based on sequencing by synthesis method, i.e. the idea is to detect pyrosulphate release when nucleotide is incorporated. Sanger sequences are longer 750 base pairs (bp) than pyroseqencing techniques, specially 454 products reads of length 100 to 200bp. 454 titanium series products reads of length 400 to 500bp. Advantages of pyrosequencing over Sanger sequencing include a 10 much lower per base cost and no requirement for cloning. These generate sequence trace files from which base calling is done.

2.3 Bioinformatics Pipeline for Metagenome Processing

Once the raw reads are obtained, the data need to be processed and analyzed to see what story is hidden in it. The following procedures are followed

2.3.1 Sequence Processing

Processing of both, the genomic and metagenomic sequence data, follow common steps like preprocessing the sequence reads, assembly, Gene Prediction and Annotation. However, the main difference between genomes and metagenomes is that the former has a fixed end point like one or more completed chromosomes. However, in the case of metagenomes, we just get draft assemblies and may be sometimes almost complete genome of dominant populations.

2.3.1.1 Preprocessing of Sequence Reads

This is a very important step in metagenome processing. It involves base calling of raw data, removal of low complexity reads, removal of containment sequences, and removal of outliers, i.e. reads with very short length. Base calling involves identifying DNA bases from the DNA sequences from the trace files. The most commonly used base calling tool is phred. Phred assigns a quality value q to each called base based on the per base-error probability, p by using the following formula: q = -10 x $\log_{10}$(p). The other tool which is used in many other researches is Prinseq. Prinseq is a web as well as standalone tool that allows us to filter, trim, and reformat the metagenome data. It removes low quality reads based on quality scores obtained from phred to avoid complications in assemblies an downstream analysis. In trims poly-A/T tails, repeats of A's and T's at the end of the sequences because it can result in false positive during the similarity searches, since they have a good alignment with low complexity regions or sequences with tails. It removes sequences with a lot of ambiguous bases i.e. sequences with high number of Ns. A position in the sequence where a base cannot be identified is replaced by the letter N which means it is an ambiguous base. For removing low complexity reads, it calculates the sequence complexity using with both DUST and Entropy approach. DUST is the heuristic used to mask low complexity regions during BLAST search. DUST computes scores based on 11 how often different triplets occur in the sequences and are scaled from o to 100 and higher scores imply lower complexity. In case of Entropy approach, entropy values of trinuleotides in the sequence is computed and scores are scaled from 0 to 100 where lower entropy mean low complexity.

2.3.1.2 Assembly

Assembly is the process of combining reads based on similarity to obtain contiguous DNA segments called contigs. There are challenges in assembling metagenome as there could be problems like co-assembly of reads coming from different species because of non-uniform species distribution. This can happen if there are high sequence similarity between reads coming from closely related species. There are many publicly available assembly programs like Phrap, Celera Assembler, Newbler but these were all designed for assembling genomes from isolates and not for metagenomes which comprise of multiple species with read coverage that is non uniform. Therefore, their performances vary significantly. To mitigate these problems for de novo assembly, we need to pass our data through more than one assembler so that it helps solving miss-assembly of the largest contigs. To further strengthen our assembly, we can perform multiple assemblies by tweaking parameters for a particular assembler. To be absolutely sure of our assembly so that problems do not percolate further downstream analysis, we can perform manual inspection using scaffolding programs like ScaffVis or visualization programs like Consed. Comparative assemblies are easier to work with; where reference genome or fully sequenced genome is passed to assembler along with the metagenome. AMOS is an assembler that performs comparative assembly.

2.3.1.3 Gene Prediction and Annotation

The process of identifying protein coding genes and RNA sequences is known as gene prediction. There are two ways of performing gene calling: one is evidence-based and the other is ab initio gene prediction. The evidence-based method is based on BLAST similarity search to find homologs against a database of previously found genes. The ab initio gene prediction method allows gene identification based on intrinsic features of the DNA sequence to differentiate between coding region of a sequence form non-coding regions. This method is useful to identify those genes that do not hove homologs to existing database sequences, and to find novel genes. For the ab initio method, there are many gene-prediction tools, some or which requires training data sets(fgenes) while some are 12 self-trained on the target sequence (MetaGene, Glimmer, Genemark). MetaGene is the prokaryotic gene prediction tool developed specially for metagenomes. The program does not require training data set and it estimates di-codon frequency from the GC content of a given sequence. In case of complete genomes, both the ways of gene prediction are employed and the hits to genes in the database acts as training sets. In case of unassembled pyrosequencing reads and high complexity metagenomes, evidence based gene prediction is the only method used because of the fragmented nature and short read length of these data seta; as pointed out by Mavromatis[5]. Even in the case of less complex communities, it is better to perform gene prediction on both reads and contigs because reads from less abundant organisms remain unassembled and these reads may contain important functionality. The most commonly used tool to predict RNA genes like tRNA and rRNA is tRNA scan[7]. Finally, to assign protein function to metagenome data, protein sequences are compared to the database of protein family sequences like TIGFAM, and COGs.[5]

2.3.2 Data Analysis

Depending on the metagenome, there are different data analysis method. The most common analysis method are composition analysis on contigs, reclassification of reads after preprocessing, and binning. Next we cover the topic of binning. The process of associating sequence data to the contributing species is known as binning. The highly reliable binning in assembly, as reads coming from same species are assembled together. This is not the case in metagenome data sets as there are chances of co-assembly. Two most common ways to bin are based on sequence similarity and sequence composition. In case of sequence similarity, we compare our metagenome data using tools such as BLAST and MEGAN (Huson et al 2007), a metagenome analyzer to separate metagenome fragments based on phylogenetic groups. If the suitable maker genes are present, then assignment of fragments based on taxonomic group is feasible. However, in case of absence of maker genes for your metagenome, the other optional approach is to use (G+C) content along with phylogenetic information to separate fragments. The other binning method, based on sequence composition is entirely different as it makes use of oligonucleotide frequencies which 13 supposedly are distinct and help separate different genomes. The word length can range from 1 to 8, with longer words giving better resolution but are expensive computationally. Therefore, typical word length range from 3 to 6 bases long. This method is so far the best method. As pointed out by Teeling, in their experiment on 9054 genomic fragments generated from 118 complete bacterial genomes the scores and results obtained using tetra-nucleotide analysis were far superior compared to GC content binning method. The standalone tool available online for tetra-nucleotide analysis is called TETRA (Teeling et al 2004). TETRA computes z-scores from the divergence between observed versus expected tetra-nucleotide frequencies. To compute observed values, it counts frequencies of all 44 = 256 possible tetra-nucleotides for DNA sequences (both forward and reverse strand). To compute expected values, it counts expected frequencies for each tetra-nucleotide "by means of a maximal-order Markov model from the sequences di- and tri-nucleotide composition".

2.4 Clustering

2.4.1 What is Clustering?

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this king, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way".

A good clustering has minimum intra-cluster distance and maximum inter-cluster distance. This way a clustering performance evaluation can be achieved which will give us insights on how well the data are clustered. A cluster is therefore a collection of objects which are "similar" among them and are "dissimilar" to the objects belonging to other clusters. We can show this with a simple graphical example:

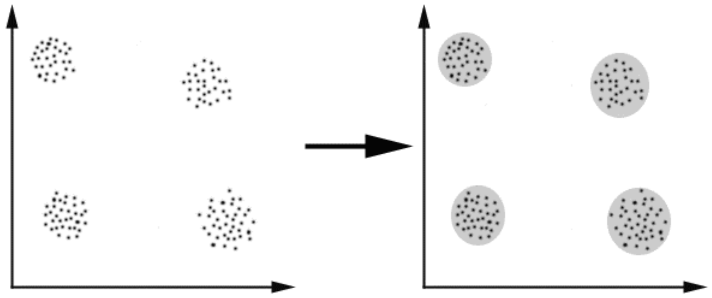


Figure 1: Clustering

In this case we easily identify the four clusters into which the data can be divided; the similarity criterion is the distance: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called distance based clustering.

Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defies a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

2.4.2 Goals of Clustering

So, the goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? In can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homologous groups (data reduction), in finding "natural clusters" and describe their own properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outliers detection). A good cluster provides maximum inter-cluster distance between the elements of two different clusters and minimum intra-cluster distance among the members of same cluster.

2.5 Clustering Methods

One of the goals of microarray data analysis is to cluster genes or samples with similar expression profiles together, to make meaningful biological inference about the set of genes or samples. Clustering is one of the unsupervised approaches to classify data into groups of genes or samples with similar patterns that are characteristic to the group. Clustering methods can be hierarchical (grouping objects into clusters and specifying relationships among objects in a cluster, resembling a phylogenetic tree) or non-hierarchical (grouping into clusters without specifying relationships between objects in a cluster) as schematically represented.

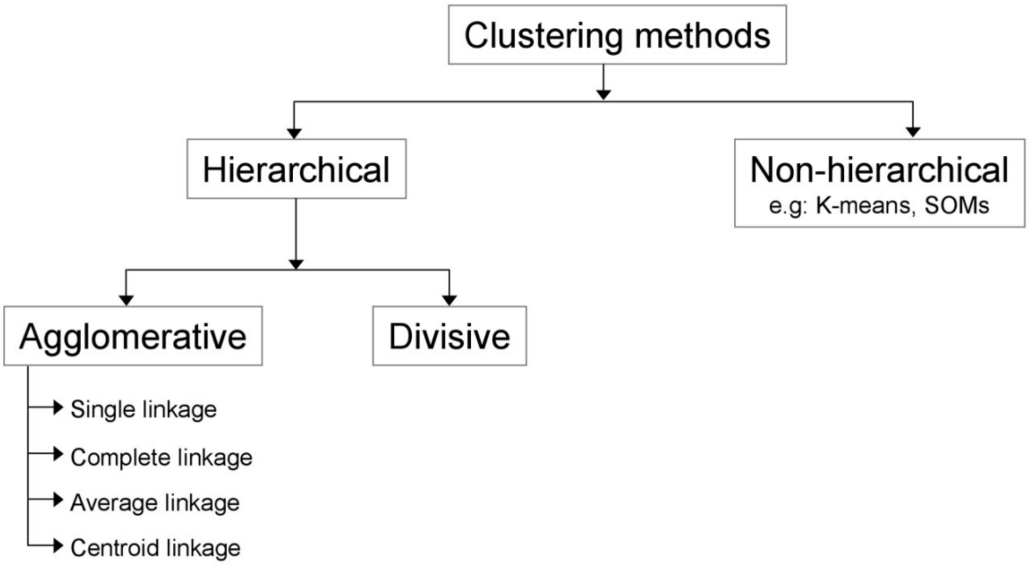


Figure 2: Clustering Methods

2.5.1 Hierarchical clustering

Hierarchical clustering may be agglomerative (starting with the assumption that each object is a cluster and grouping similar objects into bigger clusters i.e. bottom up approach) or divisive (starting from grouping all objects into one cluster and subsequently breaking the big cluster into smaller clusters with similar properties i.e. top down approach).

2.5.1.1 Hierarchical clustering: Agglomerative

In the case of a hierarchical agglomerative clustering, the objects are successively fused until all the objects are included. For a hierarchical agglomerative clustering procedure, each object is considered as a cluster. The first step is the calculation of pairwise distance measures for the objects to be clustered. Based on the pairwise distances between them, objects that are similar to each other are grouped into clusters. After this, pairwise distances between the clusters are re-calculated, and clusters that are similar are grouped together in an iterative manner until all the objects are included into a single cluster. This information can be represented as a dendrogram, where the distance from the branch point is indicative of the distance between the two clusters or objects.

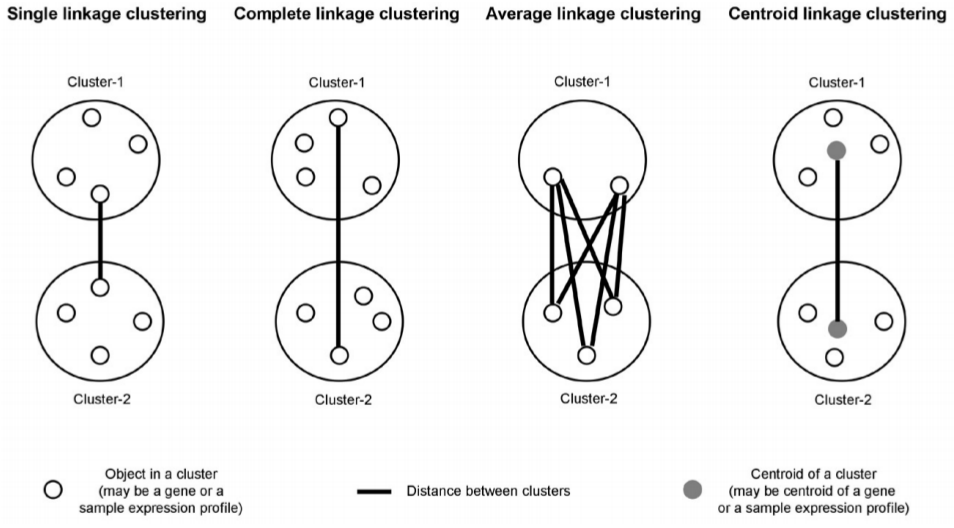

Single linkage clustering
Complete linkage clustering
Average linkage clustering
Centroid linkage clustering
Cluster-1
Cluster-1
Cluster-1
Cluster-1
Cluster-2
Cluster-2
Cluster-2
Cluster-2
Object in a cluster (may be a gene or a sample expression profile)
Distance between clusters
Centroid of a cluster (may be centroid of a gene or a sample expression profile)

Figure 3: Linkage Methods

- Single linkage clustering (Minimum distance): In single linkage clustering, distance between two clusters is calculated as the minimum distance between all possible pairs of objects, one from each cluster. This method has an advantage that it is insensitive to outliers. This method is also known as the nearest neighbour linkage.

- Complete linkage clustering (Maximum distance): In complete linkage clustering, distance between two clusters is calculated as the maximum distance between all possible pairs of objects, one from each cluster. The disadvantage of this method is that it is sensitive to outliers. This method is also known as the farthest Neighbour linkage.

- Average linkage clustering: In average linkage clustering, distance between two clusters is calculated as the average of distances between all possible pairs of objects in the two clusters.

- Centroid linkage clustering: In centroid linkage clustering, an average expression profile (called a centroid) is calculated in two steps. First, the mean in each dimension of the expression profiles is calculated for all objects in a cluster. Then, distance between the clusters is measured as the distance between the average expression profiles of the two clusters.

2.5.2 Non-hierarchical clustering

One of the major criticisms of hierarchical clustering is that there is no compelling evidence that a hierarchical structure best suits grouping of the expression profiles. An alternative to this method is a non-hierarchical clustering, which requires predetermination of the number of clusters. Non-hierarchical clustering then groups existing objects into these predefined clusters rather than organizing them into a hierarchical structure

3 Proposed Method

3.1 Skeleton of Proposed Method

Initially DNA is extracted from the environment directly and it is known as metagenomics. Metagenomes are manipulated using enzyme called “Restriction Endonucleases”. After that a library of metagenomics is constructed and finally DNA analysis is performed.

3.1.1 Sequence Analysis

The metagenome or the DNA sequence generally consists of a large number of nucleotides. A new approach to analysing gene functions has emerged. DNA arrays allow one to analyse the expression levels (amount of mRNA produced in the cell) of many genes under different time points and conditions to reveal which genes are switched on and switched off in the cell. The outcome of the study is an n by m expression matrix, I with the n rows corresponding to genes, and the m columns corresponding to different time points & conditions. The expression matrix I represents intensities of hybridization signals as provided by a DNA array. In reality, expression matrices usually represent transformed and normalized intensities rather than the raw intensities obtained as a result of a DNA array experiment.

Clustering algorithms group genes with similar expression patters into clusters with the hope that these clusters correspond to group of functionally related genes. To cluster the expression data, the n by m expression matrix is often transformed into an n by n distance matrix d where d_{ij} reflects how similar the expression patters of genes i and j are.

3.1.2 Clustering and Cluster Manipulation

We run the K-Means Algorithm on the values of the distance matrix and cluster the different values. We find a marker gene for each of the cluster from which the distances of other members of that cluster is minimum. Then considering each marker gene as single entity we apply S-LINK Agglomerative Hierarchical Clustering to produce the phylogenetic tree.

3.2 Proposed Algorithm

The **A**dvanced **A**gglomerative **C**lustering **T**echnique(AACT) using Manhattan Distance works in the following way.

3.2.1 K-means Stage

In this stage the traditional k-mean technique is applied and identified *l* distinct clusters over the input dataset *X*. Generally, the traditional k-means technique consists of three steps. In the first step, to fix the *l* centroids values $\bar{K} = \{\bar{K}_0,..,\bar{K}_\mathrm{l}\}$ over the input dataset *X* as defined $X = \{\mathrm{X}_0,.., \mathrm{X}_\mathrm{n}\}$, where *X* represents input dataset, n denotes number of objects that belong to input dataset *X* and $\bar{K}$ represents the number of centroid values identified in *X*. In the second step, it maps the *l* clusters in $\bar{K}$ over the input dataset *X* through the process of measuring Manhattan distance between dataset *X* and *l* centroid values as defined in the equation(1).

$$C_\mathrm{j} = Min\{D(X_\mathrm{i}, \bar{K}_\mathrm{j}) \ \mid \ \forall X_\mathrm{i} \in X, \ \forall \bar{K}_\mathrm{j} \in \bar{K}_\mathrm{l}\} \tag{1}$$

where *D(*X_i*,* $\bar{K}_j$*)* represents the Manhattan distance between i^{th} object in *X* and j^{th} centroid in $\bar{K}$ and is defined as equation (2)

$$D(X_\mathrm{i}, \bar{K}_\mathrm{j}) = \{(X_\mathrm{i} - \bar{K}_\mathrm{j}) \ \mid \ \forall X_\mathrm{i} \in X, \ \forall \bar{K}_\mathrm{j} \in \bar{K}\} \tag{2}$$

where X_i denotes the dataset *X* and $\bar{K}_j$ is centroid value of j^{th} cluster. In the next step, it partitions the input dataset *X* into *l* distinct clusters $C = \{\ C_0,...,C_l\}$ in $\bar{K}_j$ as defined in equation (3)

$$\bar{K}_\mathrm{j} = \{\frac{1}{N_\mathrm{j}} \sum_{l=0}^{n_\mathrm{j}} C_\mathrm{jl} \ \mid \ \forall C_\mathrm{jl} \in C_\mathrm{j}, \ \forall C_\mathrm{j} \in C\} \tag{3}$$

Where C_{ij} represents the i^{th} object in the j^{th} cluster that belongs to the C. Repeat the steps from step 2 to step 3 until the result of the current iteration equal to previous iteration. This modified K-means algorithm is described in the below subsection.

3.2.2 Algorithm for K-means Clustering

Input: $X = \{\mathrm{X_0}, .., \mathrm{X_n}\}$
Output: l-clusters $= \{\mathrm{C_0}, \mathrm{C_1}, .., \mathrm{C_l}\}$

Begin

1. Fix the l centroids values $\bar{K} = \{\bar{K}_0, .., \bar{K}_\mathrm{n}\}$ over the input dataset X.
2. Map l clusters in $\bar{K}$ over the input dataset X by using the equation (1) and (2).
3. Partition the input dataset X into l distinct clusters $C = \{\mathrm{C_0}, .., \mathrm{C}_l\}$ using the equation(3).

End

3.2.3 S-LINK Stage

In this stage, the S-LINK (Single Linkage) technique is applied to identify *'m'* clusters over the result of k-means technique *C*. *S-LINK* technique consists of four steps. In the first step, it computes centroid over each individual clusters in the result of k-means, *C* for *i=0,1,....,l* using the equation(4).

$$\bar{C} = \sum_{i=0}^{l} \sum_{j=0}^{n_\mathrm{i}} C_\mathrm{ij} \tag{4}$$

Where C_{ij} denotes the j^{th} object in the i^{th} cluster. l denotes the number of clusters and n_i denotes number of objects in the i^{th} cluster. In the second step, it constructs the distance matrix D_{ij} over the result of $\bar{C}$ based on Manhattan distance and is defined in equation (5).

$$D_{\mathrm{ij}} = \{_{\mathrm{i=0,1,...k\text{-}1;\ j=\ i+1,...k}}\ d(\bar{C}_{\mathrm{i}}, \bar{C}_{\mathrm{j}})\ \ |\ \ \forall\bar{C}_{\mathrm{i}} \in \bar{C},\ \forall\bar{C}_{\mathrm{j}} \in \bar{C}, \} \tag{5}$$

Where $d(\bar{C}_{\mathrm{i}}$,$\bar{C}_{\mathrm{j}})$ represents the distance between i^{th} and j^{th} cluster belonging in $\bar{C}$ and is defined in equation(6).

$$d(\bar{C}_{\mathrm{i}}, \bar{C}_{\mathrm{j}}) = |\bar{C}_{\mathrm{i}} - \bar{C}_{\mathrm{j}}| \tag{6}$$

If i^{th} and j^{th} cluster are containing more than one objects, then compute the distance of set of objects then compute the distance of set of object pairs between i^{th} and j^{th} clusters and then consider the minimum distance of object pair as a distance of i^{th} and j^{th} cluster as defined in equaion (7)

$$d(\bar{C}_{\mathrm{i}}, \bar{C}_{\mathrm{j}}) = min\{d(\bar{C}_{\mathrm{i}}, \bar{C}_{\mathrm{j}})\} \tag{7}$$

where $\bar{C}_{\mathrm{i}}$,$\bar{C}_{\mathrm{j}}$ denotes object pairs of i^{th} and j^{th} clusters and $\bar{C}$. In the fourth step, it finds the closest cluster pair with minimum distance Δd over the distance matrix D_{ij} as defined in equation(8).

$$\Delta d = min\{D_{\mathrm{ij}}\ \ |\ \ \forall D_{\mathrm{ij}} \in D\} \tag{8}$$

In the next step, merge the closest cluster pair $(\bar{C}_{\mathrm{i}}, \bar{C}_{\mathrm{j}})$ into a single cluster $\bar{C}_{ij}$. Then delete the j^{th} and compute the centroid of new cluster $\bar{C}_{i}$. Repeat the step two, until the number of iterations is satisfying *(l-m)* where m is the number of clusters. This modified S-LINK algorithm is described in the below section.

3.2.4 Algorithm for Agglomerative Clustering

Input: $C = \{C_{0},...,C_{l\text{-}1}\}$

Output: $G = \{G_{0},...,G_{l\text{-}1}\}$

Begin

1. Compute centroid over the each individual clusters in the result of K-means, C for i=*0,1,...,l-1* using the equation(4).

2. Construct distance matrix D_{ij} over the result of $\bar{C}$ based on Manhattan distance in equation (5) and (6).

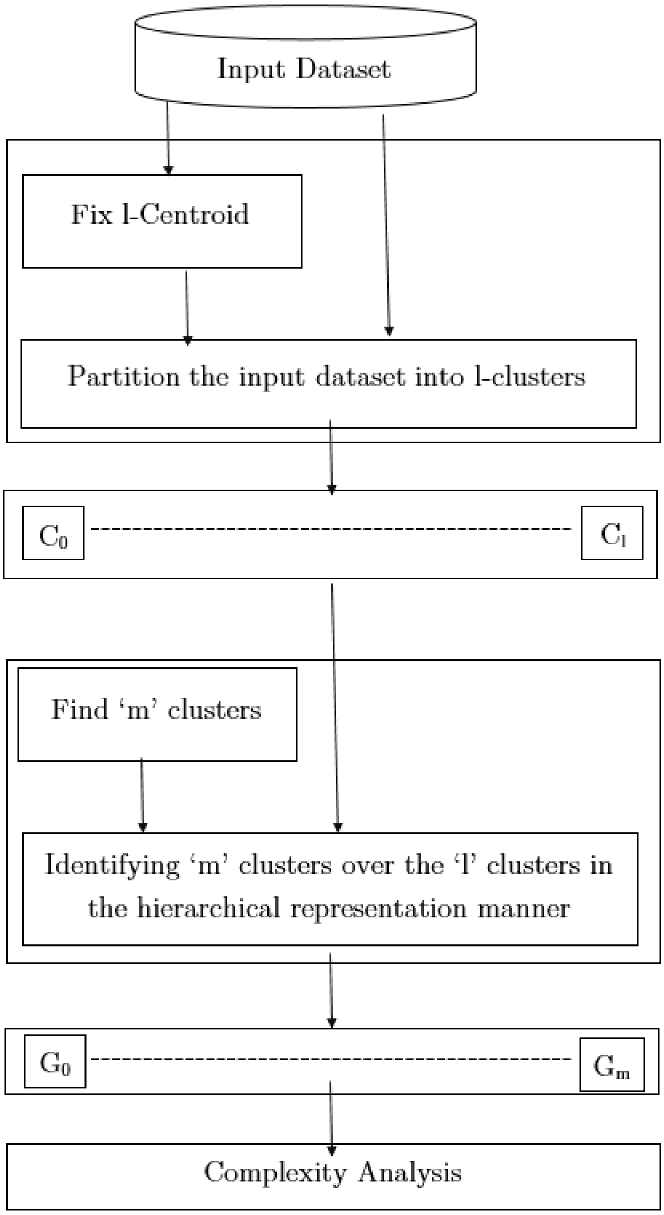
3. If i^{th} and j^{th} clusters are containing more than one objects then compute the distance of set of object pairs between i^{th} and j^{th} clusters and consider the minimum distance of object pairs as a distance of i^{th} and j^{th} cluster using equation(7).

4. Find the closest cluster pair with minimum distance Δd over the distance matrix D_{ij} using the equation(8).

5. Merge the closest pair *($\bar{C}_i$,$\bar{C}_j$)* into single cluster $\bar{C}_{ij}$. Delete the j^{th} cluster and compute centroid of new cluster $\bar{C}_i$. Repeat the steps, until the number of iterations is satisfying *(l-m)*.

End

3.3 Flowchart



Input Dataset
Fix l-Centroid
Partition the input dataset into l-clusters
C0
Cl
Find 'm' clusters
Identifying 'm' clusters over the 'l' clusters in the hierarchical representation manner
G0
Gm
Complexity Analysis

Figure 4: Flowchart

3.4 Simulation Technique

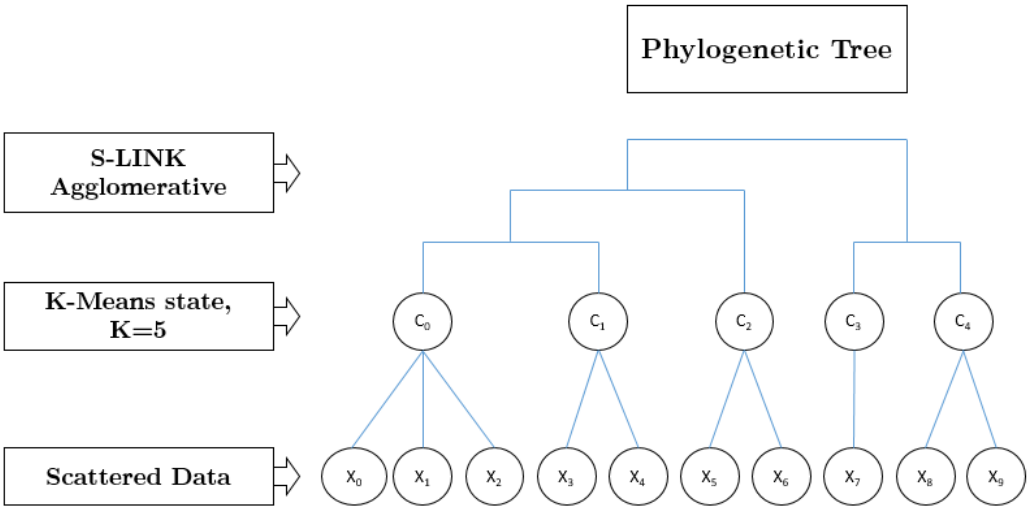


Figure 5: Simulation Technique

Here let us consider $X_0,\dots,\ X_9$ are the preliminary dataset which is found out by exponential of the intensity values in the dataset. In the first step, the number of clusters is specified and $C_0,\dots,\ X_5$ are the random clusters of K-means method. Then rest of the data is assigned to the clusters based on the closest distance.

Then from each of the clusters of K-means, one representative data is selected from each of the cluster. Now, in agglomerative stage, each pair of the closest cluster is merged together in each of the iteration and ultimately they merge to one cluster. If we want to get finally m number of clusters, then we have to stop m iterations before the iteration loop ends. Thus we get the phylogenetic tree.

4 Results & Discussion

4.1 Complexity Analysis

The proposed **A**dvanced **A**gglomerative **C**lustering **T**echnique(AACT) is better suitable to identify m distinct clusters over the large dataset with lesser computational complexity and finite number of iterations. In stage one *(k-means)* dataset of size n is reduced to l distinct number of clusters with l number of iterations and computational complexity of *O(nl)* where, n is the size of dataset and l is distinct number of clusters. In second stage *(S-LINK)* l distinct number of clusters obtained from stage one *k-means* is reduced to m number of groups with *(l-m)* number of iterations and computational complexity of $O(nl+l^2)$, where nl is the computational complexity of stage one and l^2 is the computational complexity obtained due to the construction of distance matrix D_{ij} in the S-LINK stage.

In this method, in case of using the distance function for distance matrix generation, using Manhattan distance rather than the Euclidian distance gives a consistent improvement in performance. Here also Minkowski distance can be used but they will give higher computational complexity due to calculation of higher power distance and subsequent higher square root.

4.2 Experimental Result

4.2.1 Dataset Description

- Title: Influenza virus H5N1 infection of U251 astrocyte cell line: time course
- Organism: *Homo sapiens*
- Platform: GPL6480: Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (Probe Name version)
- Citation: Lin X, Wang R, Zhang J, Sun X et al[18]
- Sample Count: 6

- Value type: Intensity transformed count
- Published Date: 04-01-2016
- 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.

4.2.2 Test Bench

- Processor: Intel Core i7-5820K @ 3.3 Ghz
- Chipset : Intel X99 Express Chipset
- Ram: 8 GB @ 2400 Mhz
- Platform: *MATLAB* 2016 64 bit

4.2.3 Time Comparison

In this section we compare among the trivial agglomerative method, most recently developed Improved Agglomerative Clustering Technique and our proposed method. As we can see from the table that our proposed method works faster than the *IACT* as Manhattan Distance was used to calculate distance among clusters instead of Euclidean Distance, which reduced the time required to compute the phylogenetic classification. So, our proposed method works faster than the *IACT*.

Sample Count	IACT (sec)	AACT (Proposed) (sec)	Trivial Agglomerative (sec)
4000	20.4776	19.063	40.7660
8000	40.5365	38.4779	80.8364
12000	62.3473	56.6029	120.7839
16000	80.9419	76.1390	163.9884
20000	103.3372	97.3996	203.4975
24000	124.1642	115.2550	243.6152

Here we see from *Figure 5* that our proposed method is far better than trivial approach since we have shrinked the large dataset using K-means clustering which performs fast but is not a hierarchical approach. So, next we use S-Linkage agglomerative method to generate the hierarchical classification for this case our phylogenetic classification.

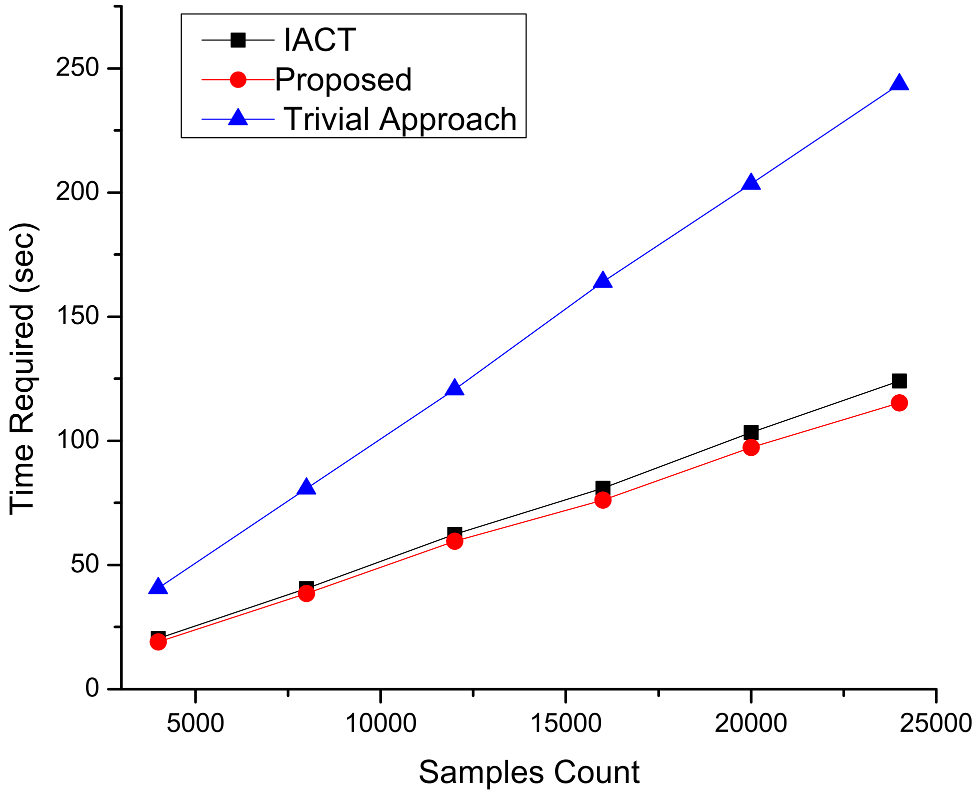


Figure 6: Time Comparison

4.2.4 Cluster Comparison

The cluster comparison between the Single Link Agglomerative hierarchical clustering and Complete Link Agglomerative hierarchical clustering is given below. Comparing the numbers we get that they perform similar but the proposed method performs better with low time consumption. The comparison is given in the following *Table 1: Cluster Comparison.*

In both the approach of $AACT$ & $IACT$ there are two main phases which are K-Means and Agglomerative (Bottom Up Approach). In both methods we select same points for initialization to achieve similar test environment. So for K-Means, in both methods, the number of clusters to be created is twenty. These twenty clusters have one representative gene each which next participates in agglomerative phase and gives us the phylogenetic tree. For agglomeratve phase the number of clusters to be created is ten. So finally we end up with ten different hierarchical classes.

Here the proposed method uses the Manhattan distance for calculating distance between two instances of data and the other one $IACT$ uses the Euclidian distance for calculating the distance between the two instances of data object. The Manhattan distance is based on absolute value distance, as opposed to squared error (Euclidean distance) distance. In practice, both of them provides nearly similar results most of the time. Absolute value distance should give more robust results, whereas Euclidean would be influenced by unusual values.

This is a multivariate technique, and ”distance” between two points involves aggregating the distances between each variable. So if two points are close on most variables, but more discrepant on one of them, Euclidean distance will exaggerate that discrepancy, whereas Manhattan distance will shrug it off, being more influenced by the closeness of the other variables.

AACT with Single Link (Proposed)				IACT with Complete Link			
K-Means		Agglomerative		K-Means		Agglomerative	
Cluster No.	No of Members	Cluster No.	No of Members	Cluster No.	No of Members	Cluster No.	No of Members
C1	2749	G1	2749	C1	3073	G1	9513
C2	617	G2	27626	C2	2342	G2	19007
C3	2685	G3	87	C3	2290	G3	2290
C4	87	G4	881	C4	88	G4	1591
C5	881	G5	2784	C5	942	G5	1300
C6	1050	G6	534	C6	1591	G6	1646
C7	1098	G7	180	C7	1300	G7	1344
C8	2784	G8	476	C8	3021	G8	1017
C9	1534	G9	481	C9	1682	G9	1089
C10	2180	G10	202	C10	2453	G10	1603
C11	2476			C11	2617		
C12	481			C12	516		
C13	2059			C13	1646		
C14	10412			C14	7749		
C15	202			C15	214		
C16	1363			C16	1344		
C17	1782			C17	1017		
C18	1704			C18	1689		
C19	1405			C19	1603		
C20	3451			C20	3823		

Table 1: Cluster Comparison

5 Conclusion and Future Work

The size of dataset in Metagenomics is humorously large. In order to manipulate this vast and increasing dataset we need very efficient algorithms. Next, each of the cluster of **A**dvnaced **A**gglomerative **C**lustering **T**echnique should be annotated from the databank of *NCBI (National Center for Biotechnology Information).*

So far all the clustering algorithms runs in sequential execution but for further improvement these algorithms can be optimized for parallel execution. Even in this age of distributed computing, this system can easily be optimized for distributed systems.

References

[1] David Koslicki, *, Simon Foucart and Gail Rosen Mathematical Biosciences Institute, the Ohio State Unviersity, Columbus, OH 43201, USA and Department of Mathematics and Department of Electrical and Computer Engineering, Drexel Unviersity, Philadelphia, PA, 19104, USA, Advance Access Publication June 20, 2013.

[2] Genivaldo Gueiros Z. Silva, Daneil A. Cuevas, Bas E. Dutilh and Robert A. Edwards, Computational Science Research Center, San Diego State Unviersity, San Diego, CA, USA, Department of Computer Science, San Diego State University, san Diego, CA, USA, Accepted 21 May 2014, Published 5 June 2014.

[3] Turnbaugh, P.J., Hamady, M., Yatsunenko T. Cantarel, B.L, Duncan A, Ley R.E., Sogin, M.L., Jones, Roe, B.A. Affouritit, J.P. et al.(2009). A core gut microbiome in obese an dlean twins. Nature, 457, 480-484.

[4] Shreedhar Kumar S, Jithender M, Chaithra B, Md. Sharif Nawaz, Anushree D. ”Improved Agglomerative Clustering Technique for Large Datasets.”

[5] D.L. Wheeler, T. Barreett, D. A. Benson, and et al. “Database resources of the National Center for Biotechnology Information”, Nucleic Acid Research, vol. 35, January 2007.

[6] D. A. Benson, I. Karsch-Mizrachi, D.J. Lipman, and et al, Gene Bank, “Nucleic Acids Research, vol. 37, pp. D26 31, Januray 2009.

[7] Qichao Tul, Zhili Hel, * andZhou 1,2,3,* [1]Department of Microbiology and Plant Biology, Institute for Environmental Genomics, University of Oklahoma, Norman, OK 73072, USA, [2]Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA and [3]State Key Joint Laboratory of Environmental Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, China. Published online 12 February 2014

[8] Ley, R.E. (2010) Obesity and the human microbiome. Curr. Opin. Gastroenterol. 26, 5-11

[9] N. Diaz L. Krause, A. Goesmann and et.al. TACOA- Taxonomic Classification of environmental genomic fragments using kernalized nearest neighbor approach, “BMC Bioinformatics, vol. 10, no 1. Pp. 56+, 2009.

[10] Y. W. Wu and Y. Ye A novel abundance-based algorithm for binning metagenomic sequences using l-tuples, “In proceedings of the 14th annual international conference RECOMB’10, pp.535 549, Springer, 2010.

[11] Larsen, N., Vogensen, F.K., van den Berg, F.W.J, Nielson, D.S., Andersen, et al. Gut microbiota in human adults with type2 diabetes differs from non-diabetic adults. PLoS one, 5, e9085

[12] David Koslicki$_1$, *, Simon Foucart$_2$ and Gail Rosen$_3$ ^{1}Mathematical Biosciences Institute, the Ohio State Unviersity, Columbus, OH 43201, USA and ^{2}Department of Mathematics and ^{3}Department of Electrical and Computer Engineering, Drexel Unviersity, Philadelphia, PA, 19104, USA, Advance Access Publication June 20, 2013.

[13] Genivaldo Gueiros Z. Silva, Daneil A. Cuevas, Bas E. Dutilh and Robert A. Edwards, Computational Science Research Center, San Diego State Unviersity, San Diego, CA, USA, Department of Computer Science, San Diego State University, san Diego, CA, USA, Accepted 21 May 2014, Published 5 June

[14] Turnbaugh, P.J., Hamady, M., Yatsunenko T. Cantarel, B.L, Duncan A, Ley R.E., Sogin, M.L., Jones, Roe, B.A. Affouritit, J.P. et al.(2009). A core gut microbiome in obese an dlean twins. Nature, 457, 480-484.

[15] S.D. Bently and J. Parkhill, Comparative genomic structure of prokaryotes, “ Annual Review of Genetics, vol 38, pp 771791, December 2004.”

[16] Y. W. Wu and Y. Ye A novel abundance-based algorithm for binning metagenomic sequences using l-tuples, “In proceedings of the 14th annual international conference RECOMB’10, pp.535 549, Springer, 2010.

[17] M. Wendall and R. Waterman, "Generalized gap model for bacterial artificial chromosome clone ngerprint mapping and shotgun sequencing", Genome Research, vol. 12. No 1, p. 19431949, 2002.

[18] Lin X, Wang R, Zhang J, Sun X et al. Insights into Human Astrocyte Response to H_5N_1 Infection by Microarray Analysis. *Viruses* 2015 May 22