# Algorithm for Clustering Gene Expression Data with Outliers Using Minimum Spanning Tree

**S. John Peter**

Department of Computer Science and Research Center, St. Xavier's College, Palayamkottai, Tamil Nadu, India

**Abstract:** ***Microarrays enable biologists to study genome-wide patterns of gene expression in any given cell type at any given time and under any given set of conditions. Identifying group of genes that manifest similar expression pattern is important in the analysis of gene expression in time series data. In this paper multidimensional gene expression data is represented using Minimum Spanning Tree (MST). A key property of this representation is that each cluster of the expression data corresponds to one sub tree of the Minimum Spanning Tree, which converts a multidimensional clustering problem to a tree partitioning problem. Each node represents one gene, and every edge is associated with a certain level of pheromone intensity, densities and the co-expression level between two genes. MST-based clustering method is presented for finding cluster in gene expression time series data using new dissimilarity measure namely DMk. It is effective in classifying DNA sequences with similar biological characteristics and discovering the relationship among the sequences.***

**Keywords:** Minimum Spanning Tree, Clustering, core edge, sub tree, outliers, Gene expression data

## 1. Introduction

Microarray technology has become one of the indispensible tools that many biologists use to monitor genome wide expression levels of genes in a given organism. Microarrays simultaneously measure the expression levels of thousands of genes in a single experiment. The results of microarray experiment are often organized as gene expression matrices whose rows represent genes and columns represent various environmental conditions or samples such as tissues [10]. The entries of these matrices give a numeric representation of the expression/activity of a particular gene under a given experimental condition. Application of microarrays range from, the study of gene expression in yeast under different environmental stress conditions to the comparisons of gene expression profiles for tumors from cancer patients. In addition to the enormous scientific potential of DNA microarrays to help in understanding gene regulation and interactions, microarrays have important applications in pharmaceutical and clinical research. By comparing gene expression in normal and disease cells, microarrays may be used to identify disease genes and targets for therapeutic drugs [22].

Clustering is a fundamental technique in exploratory data analysis and pattern discovery, aiming at extracting underlying cluster structures. Cluster analysis is concerned with multivariate techniques that can be used to create groups amongst the observations, where there is no a *priori* information regarding underlying group structure. The amount of genetic data is growing faster than the rate at which it can be analyzed. Clustering techniques provide a viable solution for handling and analyzing such rapidly growing genetic data. Clustering algorithms partition sequences into different biologically meaningful groups, facilitating therefore the prediction of functions of genes [26]. When a new gene is assigned to a cluster, the biological function of this cluster can be attributed to this gene with high confidence. On the other hand, clustering gene sequences into groups may also help with analyzing evolutionary relationships among the sequences in a cluster. Many clustering algorithms have already been used for cluster analysis of gene expression data, such as hierarchical clustering, K-Means clustering, Self Organizing Map algorithm (SOMs). Fuzzy based clustering algorithm, clustering algorithm based on artificial neural network algorithms are widely applied in analysis of gene expression data clustering. Lixin Tang [23] used combination of K-Means and genetic algorithm to analyze gene expression data. Fangxiang Wu [13] used a hybrid algorithm GWKMA of weighted K-means and genetic algorithms to analyze gene expression data. However these algorithms depend on the choice of initial cluster centers, because quite different patterns of genes will be resulted with different initial parameter values. To solve this problem many improved and optimized algorithms namely clustering algorithm based on sampling [17], density-based clustering for gene expression [6], graph based algorithm [4], k-means algorithm based on partition have been proposed.

An outlier is an observation of data that deviates from other observations so much that it arouses suspicious that was generated by a different mechanism from the most part of data. Outlier may be erroneous or real in the following sense. Real outliers are observations whose actual values are very different than those observed for rest of the data and violate plausible relationship among variables [25]. Erroneous outliers are observations that are distorted due to misreporting errors in the data collection process. Outliers of either type may exert undue influence on the result of data analysis. So they should be identified using reliable detection methods prior to performing data analysis [16], [21].

A framework for representing a set of multi-dimensional data as a Minimum Spanning Tree (MST), a concept from the graph theory was developed. A tree is a simple structure for

representing binary relationships and any connected component of a tree is called a subtree. Through this MST representation, we can convert a multi-dimensional clustering problem to a tree partitioning problem, i.e., finding a particular set of tree edges ("long" edges from either local or global point of view) and then cutting them. Representing a set of multi-dimensional data points as a simple tree structure will clearly lose some of the inter-data relationship. However we have rigorously demonstrated that no essential information is lost for the purpose of clustering. This is achieved through a rigorous proof that each cluster corresponds to one subtree, which does not overlap the representing subtree of any other cluster. Hence a clustering problem is equivalent to a problem of identifying these subtrees through solving a tree partitioning problem [6].

Genes that have similar expression profiles are identified using an appropriate similarity measure. Some commonly used distance metrics are Euclidean distance, Pearson's correlation coefficient, Jackknife correlation and Spearman's rank-order correlation coefficient. In general, Euclidean distance and Pearson's correlation coefficient are widely used as distance measures for clustering gene expression data [27]. However, Euclidean distance measure is not effective in reflecting functional similarities as well as interdependence among values. It can only account for closeness in values. Pearson's correlation coefficient accounts for the overall shape of genes, but is not robust to outliers. The approach presented in this paper used a new alignment-free distance measure based on k-tuples, DMk [30] (Distance Measure based on k-tuples) and Minimum Spanning Tree based clustering algorithm aims to speed up the clustering process by using the alignment free similarity measures and is able to produce clustering result. We have applied Minimum Spanning Tree (MST) based clustering algorithm with new similarity measure namely Distance Measure based on k-tuples DMk in clustering gene sequences on performing phylogenetic analysis [8].

This paper is structured as follows: In section 2 we review some of the existing works on various clustering algorithms used for gene expression analysis. In Section 3 we propose a new algorithm **MSTCG** for gene expression analysis with new similarity measure namely Distance Measure based on k-tuples (DMk). In section 4 the outcomes of the algorithm is discussed. Finally in conclusion we compared our results with the results of existing methods.

## 2. Related Work

The function of genes can be easily understood with the help of data mining techniques. Clustering identifies subsets of genes that behave in similar manner in a particular time. Genes in the same cluster have similar expression pattern. Grouping similar expression pattern and extract useful patterns from noisy datasets are two major challenges for clustering gene expression data. Gene expression data clustering techniques can be categorized as partitioning, hierarchical, density-based, model based and graph based [27].

Partitional methods: Given a database of *n* objects or data tuples, a partitioning method constructs *k* partitions of the data, where each partition represents a cluster k ≤ n. that is, it classifies the data into *k* groups, which together satisfy the following requirements: (1) each group must contain atleast one object and (2) each object must belong to exactly one group. The two popular partitional clustering algorithms are (1) k-means algorithm, where each cluster is represented by the mean value of the objects in the cluster, and (2) the k-medoids algorithm, where each cluster is represented by one of the objects located near the centre of the cluster. Specifying the number of clusters in advance is difficult in the case of gene expression data. Partitional methods are restricted to lower dimensional data. But gene expression data sets may have high dimensionality [15].

Hierarchical methods: A hierarchical method creates a hierarchical decomposition of the given set of data objects. It has two approaches. One is agglomerative approach, also called the bottom-up approach, which starts with each object forming a separate group. It successively merges the objects or groups close to one another, until all of the groups are merged into one or until a termination condition holds. Another one is divisive approach, also called the top-down approach, which starts with all the objects in the same cluster. In each successive iteration, a cluster is split into smaller clusters, until eventually each object in one cluster, or until a termination condition holds. Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm [12] adopts an agglomerative method to represent the clustered data set graphically. Deterministic-Annealing Algorithm (DAA) splits the genes through divisive approach for clustering [1]. However small change in the data set may greatly change the cluster structure. The computational complexity of hierarchical clustering is very high.

Density-based methods: Clustering methods have been developed based on the notion of density. The general idea is to continue growing the given cluster as long as the density in the "neighborhood" exceeds some threshold; that is, for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise and discover clusters of arbitrary shape. Density based hierarchical clustering (DHC) algorithm to identify co-expressed gene pattern was proposed in [18]. Density based approach identify clusters of arbitrary shapes even in the presence of noise. However density based clustering approach for gene expression data suffer from computational complexity and also depends on input parameters which needs domain knowledge.

Model-based methods hypothesize a model for each of the clusters and find the best fit of that model to each other. A model-based algorithm may locate clusters by constructing a

density function that reflects the spatial distribution of the data points. It also leads to a way of automatically determining the number of clusters based on standard statistics, taking "noise" or outliers into account and thus yielding robust clustering methods. Expectation Maximization (EM) algorithm finds good values for input parameters iteratively [9]. Model based approach gives an estimated probability that an object belong to particular cluster. Two different clusters may have gene with high correlation. The model based approach assumes that the data set fits into specific distribution which is not always possible in a gene expression data set.

In graph-based clustering algorithms, graphs are built as combination of objects, features or both, as the nodes and edges. The graph is partitioned by using graph theoretic algorithms. Cluster Identification via Connectivity Kernels (CLICK) [28] is suitable for subspace and high dimensional data clustering. Cluster Affinity Search Techniques (CAST) by [6] takes as input the pairwise similarities between genes and an affinity threshold. It does not require a user-defined number of clusters and handles outliers efficiently. But, it faces difficulty in determining a good threshold value. To overcome this problem E-CAST [5] calculates the threshold value dynamically based on the similarity values of the objects that are yet to be clustered.

Clustering by minimal spanning tree can be viewed as a hierarchical clustering algorithm which follows the divisive approach. Clustering Algorithm based on minimum and maximum spanning tree were extensively studied. Avis [3] found an $O\ (n^2\ log^2\ n)$ algorithm for the min-max diameter-2 clustering problem. Asano, Bhattacharya, Keil and Yao [2] later gave optimal O $(n\ log\ n)$ algorithm using maximum spanning trees for minimizing the maximum diameter of a bipartition. The problem becomes NP-complete when the number of partitions is beyond two. Asano, Bhattacharya, Keil and Yao also considered the clustering problem in which the goal to maximize the minimum inter-cluster distance.

The MST clustering algorithm has been widely used in practice. Xu (Ying), Olman and Xu (Dong) [32] use MST as multidimensional gene expression data. They point out that MST- based clustering algorithm does not assume that data points are grouped around centers or separated by regular geometric curve. Thus the shape of the cluster boundary has little impact on the performance of the algorithm.

## 3. MST Algorithm for Gene Data Set

A gene is a molecular unit of heredity of a living organism. It is widely accepted by the scientific community as a name given to some stretches of DNA and RNA that code for a polypeptide or for an RNA chain that has a function in the organism, though there still are controversies about what plays the role of the genetic material. Living beings depends on genes, as they specify all proteins and functional RNA chains. Genes hold the information to build and maintain an organism's cell and pass genetic traits to offspring. All organisms have many genes corresponding to various biological traits, some of which are immediately visible, such as eye color or number of limbs, and some of which are not, such as blood type, increased risk for specific diseases, or the thousands of basic biochemical processes that comprise life. A gene sequence is a succession of four symbols {A, C, G, T}. Because the similarity between the genes of two species indicates their evolutionary relationship, it is used in many clustering algorithms. The goal of sequence clustering is to partition biological sequences into meaningful/functional groups according to the similarity information, which is calculated using either an alignment- based method or an alignment-free method [8].

The traditional approach for clustering DNA sequences requires all-by-all comparisons from alignment [11, 29, 31]. Given two sequences: $S_1$ = AGCACACA and $S_2$ = ACACAGTA, $S_1^p$ and $S_2^p$ are used to represent the $p^{th}$ characters in $S_1$ and $S_2$, respectively.

The alignment score for $(S_1, S_2)$ is given by

$$Sim_{Score(S1,\ S2)} = \sum_{p=1}^{l} E(S_1^p, S_2^p)$$

where E is the cost of an alignment operation: deletion, substitution or insertion. However this distance measure relies on sequence alignment. Since sequence alignment suffers in computational aspect with regard to large biological databases, clustering methods relying on sequence alignment have difficulties in dealing with the large gene data. An alignment-free similarity measure helps to avoid the computational complexity of multiple sequence alignment for similarity computation [8].

A Minimal Spanning Tree (MST) is a weighted connected graph where the sum of the weights is minimal. A graph $G$ is a pair $(V, E)$, where $V$ is a finite set of the elements, called vertices, and $E$ is a collection of unordered pairs of $V$. An element of $E$, called edge, is $e_{i,j} = (v_i, v_j)$, where $v_i, v_j \in V$. In a weighted graph a weight function $w$ is defined, which function determines a weight $w_{i,j}$ for each edge$(v_i,v_j)$. The complete graph $K_N$ on a set of $N$ vertices is the graph that has all the N(N-1)/2 possible edges. Creating the Minimal Spanning Tree means searching the $G' = (V, E')$ the connected subgraph of $G$, where $E' \subset E$ and the cost is minimum.

The cost of MST is computed as

$$\sum_{e \in E'} w(e)$$

where $w(e)$ denotes the weight of the edge $e \in E$. In a graph $G$, where the number of the vertices is $N$, MST has exactly $N - 1$ edges.

A Minimal Spanning Tree can be efficiently computed in $O(N^2)$ time using either Prim's or Krushkal's algorithm [20]. A Minimal Spanning Tree can be used in clustering. Let $V=\{x_1, x_2 \ldots x_N\}$ be a set of the data with $N$ distinct objects which are to be distributed in different clusters. $x_i$ denotes the *i-th* object, which consists $n$ measured variables, grouped into an $n$-dimensional column vector $x_i=[x_{1,i}, x_{2,i}, \ldots x_{n,i}]^T x_i \in R^n$. Let $d_{i,j} = d(x_i, x_j)$ be the distance, defined between $x_i$ and $x_j$. The distance can be computed in different ways namely Euclidean distance, Manhattan distance, Mahalanobis distance, mutual neighbor distance, etc.
Removing edges from the MST leads to a collection of connected sub-graphs of $G$, which can be considered as clusters. Finding and eliminating such edges from **MST**, leads to the best clustering result. Such edges are called inconsistent edges [20].

Here an MST is used to represent a set of gene expression data and their significant inter-data relationship. The weight between two nodes (two genes) is calculated using DMk distance measure. There are also other ways to measure the distance between two gene expression profiles such as Euclidean distance, correlation distance and mahalanobis distance.

For a given data set, let $T(V, E)$ be Minimum Spanning Tree, on the set of points in the data set, where $V$ is set of points (also called vertices) and $E$ be the set of undirected edges connecting a pairs of vertices in $V$.

Some of the definition needed for our **MSTCG** algorithm is defined as,

Definition 1(***Core Edge***): Given an MST $T(V, E)$, *core edge* $CE(E)$ *is* defined as the edge which is having minimum length.

$$CE(E) = min(E) \quad (1)$$

Definition 2 (***Average Edge Weight***): Given an MST $T(V, E)$, *average edge weight* $\hat{W}$ is defined as the ratio between sum of the weight of the edges and total number of edges.

$$\hat{W}(E) = \sum_{i=1}^{n} |e_i| / n \quad (2)$$

Geometric notion of centrality are closely linked to facility location problem. The distance matrix $D$ can computed rather efficiently using Dijkstra's algorithm with time complexity $O(|V|^2 \ln |V|)$ [20].

The *eccentricity* of a vertex $x$ in $G$ and radius ρ (G), respectively are defined as

$$e(x) = \max_{y \in V} d(x, y) \text{ and } \rho(G) = \min_{x \in V} e(x)$$

The *center* of $G$ is the set

$$C(G) = \{x \in V \mid e(x) = \rho(G)\}$$

Based on the above definitions, the process of construction of sub trees (clusters) from the given data set represented in the form of an MST $T(V, E)$ is described as follows:

1. Search for *core edge* $CE_1$ (E) from the MST. It will be the base for first sub tree or cluster denoted by $C_1$. For each vertex $V_i$ in the sub tree, add the edges and the associated vertices $V_j$ from the MST $T(V, E)$ in to the sub tree, if the *distance* $(V_i, V_j)$ is less than the *average edge weight*.
2. Do step 2 until no edges can be added into the sub tree.
3. Search for next core edge $CE_2$ and then construct another sub tree $C_2$. In this way series of sub trees (clusters) are generated as $C_1, C_2, C_3 \ldots C_n$.
4. If any of the vertices not added to any of the clusters $C_1, C_2, C_3 \ldots C_n$ are considered as outliers, which can be determined based on distance from center of the MST $T$.

The construction of sub tree or cluster begins with *core edge* in the MST $T(V, E)$, then the vertices of the sub tree will be removed from the data set (MST), so each cluster is constructed in the highest density region in the existing data set. In other words, a series of clusters are automatically generated from high density region to low density region. The clustering process will stop growing, if the distance of the new edge greater than the average edge weight of the MST $T(V, E)$. This condition is used to guarantee the approximate evenly distribution of the vertices in each of the clusters or sub tree.

### 3.1 MSTCG Clustering Algorithm

Given a Data set S, our method first converts the data set into Dissimilarity Matrix (DM) then starts by constructing Minimum Spanning Tree (MST) from the Dissimilarity Matrix. The weight of the edge in the MST $T$ is Distance Measure based on k-tuples (DMk) distance between the two end points [20].

Our **MSTCG** algorithm mainly consists of the following steps;

- Representation of data points in the form of Dissimilarity Matrix (*DM*)
- Construction of MST T*(V, E)* using *DM*
- Generating subtree (clusters) using T*(V, E)*
- Detecting Outliers if any

Algorithm: MSTCG ( )

Input: Dissimilarity Matrix (DM)
Output: *n* Clusters with outliers

Let *n* be the number of clusters
Let $C_n$ be the sub tree or cluster
Let O be the outliers

1. Construct an MST *T(V, E)* from Dissimilarity Matrix (DM)
2. Compute the *average edge weight* ($\hat{W}$) of the edges from MST T*(V, E)*
3. Find center C of the MST *T (V, E)* using eccentricity of points
4. $n = 1$, $C_n = \Phi$, *Visited[V] = 0*, $O = \Phi$
5. **Repeat**
6. *(a, b)* = min (E) // finding core edge
7. **While**(*distance (a ,b)*$< \hat{W}$) *do*
8. $C_n = C_n \cup \{ (a, b)\}$ // new cluster formation begins
9. Visited[a] = *n*; Visited[b] = *n*
10. **For** each vertex $V_i$ in $C_n$ do
11. **For** each vertex $V_j$ linked with $V_i$ do
12. **If** *distance* $(V_i, V_j) < \hat{W}$ **and** Visited[$V_j$] = 0 **then**
13. $C_n = C_n \cup \{ (V_i, V_j) \}$; Visited[$V_j$] = *n*
// cluster or sub tree $C_n$ growing
14. Remove all the clustered vertices and edges from *T*
15. $n = n+1$
16. **Until** all the vertices are visited in *T (V, E)*
17. **For** each vertex *V*isited [*V*] = 0 do
18. **If** *distance* $(V_i, C) > \hat{W}$ **or** *distance* $(V_i, C_i) > \hat{W}$ **then** $O = O \cup \{V_i\}$
19. **Return** *n* number of clusters with outliers

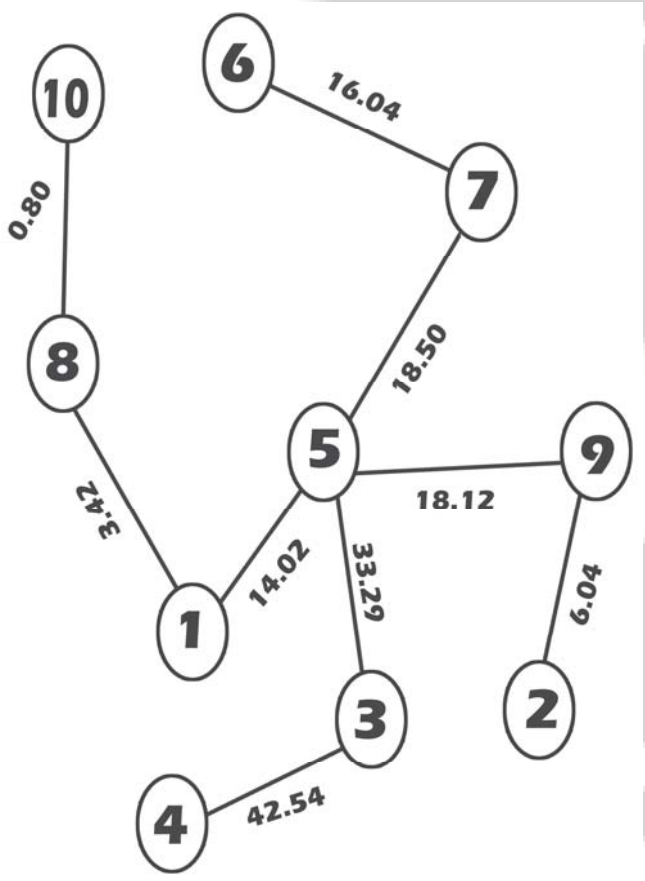


**Figure 1:** MST *T (V, E)* with Center node as 3

| *Vertex(v)* | 1 | 2 | **3** | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| *Eccentricity(v)* | 89.95 | 94.29 | **67.83** | 110.37 | 75.83 | 110.37 | 94.33 | 93.27 | 100.33 | 94.07 |
| *Visited(v)* | 0 | 0 | **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| *Vertex (v)* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| *Visited(v)* | 1 | 2 | **0** | **0** | 1 | 3 | 3 | 1 | 2 | 1 |

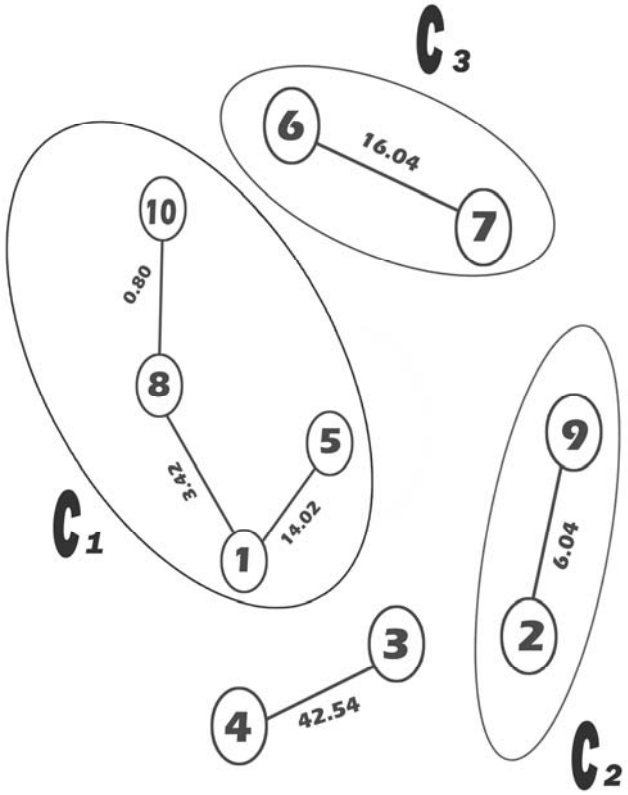


**Figure 2:** Three sub trees or clusters $C_1$, $C_2$, $C_3$ from the MST

Our **MSTCG** algorithm constructs MST *T (V, E)* from set of point *S* (line 1) is shown in fig 1. *Average edge weight* of the edges in MST is computed (line 2). For the given MST T*(V, E) Average edge weight* ($\hat{W}$) computed as 16.974. Using the eccentricity of points, the central vertex is computed as vertex 3 at line 3. Initially all the vertices in the MST are marked as unvisited, which are represented as Visited [V] = 0 at line 4. Next the *core edge* is identified (line 6) to form a new sub tree (line 8). Here the c*ore edge* is considered as a new initial cluster or sub tree. From this initial sub tree or cluster, cluster growth begins at line 8. The vertices linked with the *core edge* are marked as Visited vertex (line 9) is shown in the fig 2. For each vertex in the sub tree, add all the linked edges and vertices into the sub tree which satisfy the threshold value of *average edge weight (*line 13). The vertices or nodes which are included in the sub tree are marked as Visited vertex (line 13) and also removed from the MST (line 14). If the threshold

condition is not satisfied then current cluster growth comes to an end, then the cluster number is incremented by 1 (line 15) and new sub tree or cluster will be created. The lines 6 through 15 in the algorithm are repeated until the entire vertex in the MST is marked as visited or the entire vertex in the MST are removed. The cluster $C_1$ = {1, 5, 8, 10}, C2 = {2, 9}, $C_3$ = {6, 7}. In this way the series of sub trees or clusters $C_1$, $C_2$, $C_3$…$C_n$ are created as shown in the fig 2. The vertices or nodes 3 and 4 are marked as unvisited (shown in fig 2) which are far away from the clusters $C_1$, $C_2$, $C_3$, and also the vertex 4 is very far away from the center node 3 are considered as outliers which are determined at lines 17-18.

## 4. Results and Discussion

### 4.1 Dissimilarity Matrix Representation

The data points in most of the graph based clustering algorithm can be represented in the form of Dissimilarity Matrix (**DM**). It consists the distance values between the data points represented as lower or upper triangular matrix. In this paper alignment-free distance measure based on k-tuples, DMk (Distance Measure based on k-tuples) is used [8]. The selected data set includes the full β-globin gene sequences of 10 species reported by Feng et al. [14]. The similarity/dissimilarity matrices for the full sequences of β-globin gene of the 10 species using DMk are shown in Table 1. Smaller the distance between the species, more similar the two sequences are.

The **DM** for the full β-globin gene sequences of 10 species is shown in Table 1.

**Table 1:** Dissimilarity Matrix

| *Species* | *1 Human* | *2 Goat* | *3 Opossum* | *4 Gallus* | *5 Lemur* | *6 Mouse* | *7 Rat* | *8 Gorilla* | *9 Bovine* | *10 Chimpanzee* |
|---|---|---|---|---|---|---|---|---|---|---|
| *1.Human* | 0 | 22.95 | 37.65 | 111.47 | 14.02 | 35.21 | 20.68 | 3.42 | 25.07 | 3.54 |
| *2.Goat* | | 0 | 41.22 | 65.70 | 18.80 | 35.05 | 33.93 | 32.36 | 6.04 | 33.05 |
| *3.Opossum* | | | 0 | 42.54 | 33.29 | 64.03 | 51.64 | 46.35 | 40.41 | 49.73 |
| *4.Gallus* | | | | 0 | 90.93 | 80.07 | 95.26 | 121.09 | 61.69 | 122.65 |
| *5.Lemur* | | | | | 0 | 21.39 | 18.50 | 17.19 | 18.12 | 18.74 |
| *6.Mouse* | | | | | | 0 | 16.04 | 33.64 | 27.60 | 37.59 |
| *7.Rat* | | | | | | | 0 | 17.69 | 30.53 | 20.58 |
| *8.Gorilla* | | | | | | | | 0 | 33.66 | 0.80 |
| *9.Bovine* | | | | | | | | | 0 | 35.46 |
| *10.Chimpanzee* | | | | | | | | | | 0 |

**Table 2:** Minimum Spanning Tree Edges

| *Edge* | *Euclidean Distance/Weight* |
|---|---|
| {8, 10} | 0.80 |
| {1, 8} | 3.42 |
| {2, 9} | 6.04 |
| {1, 5} | 14.02 |
| {6, 7} | 16.04 |
| {5, 9} | 18.12 |
| {5, 7} | 18.50 |
| {5, 3} | 33.29 |
| {3, 4} | 42.54 |

From the fig 2 the most similar species pairs are identified as gorilla-chimpanzee, human-chimpanzee and human-gorilla, which are expected from their evolutionary relationship as shown in cluster $C_1$. Another cluster $C_2$ in fig 2 shows that goat-bovine have less similarity. Yet another cluster $C_3$ contains the species pair Rat-Mouse having a lesser similarity than the previous two clusters $C_1$ and $C_2$. The gallus is separated from rest of the species which coincides with the fact that gallus is the only non-mammalian species among these 10 species. It is also found that distance (dissimilarity) of

opossum is very far from the remaining mammals. Thus the gallus and opossum species are considered as outliers who are consistent with the biological morphology. The quality of the constructed clusters shows the quality of the Dissimilarity Matrix (DM) and the method of abstracting information from DNA sequences.

DMk measures the similarity between DNA sequences more effectively than any other distance measure including the k-tuple distance [8]. This is because DMk measures the distance between DNA sequences based on sequence structure and composition. From the outcomes through evaluation on gene families and constructing phylogenetic trees of full gene sequences of 10 species, we find that DMk gives more competitive results compared to the k-tuple distance.

## 5. Conclusion

Our **MSTCG** clustering algorithm does not requires domain knowledge of the given problem. Our algorithm finds series of clusters $C_1$, $C_2$, $C_3$…$C_n$. These clusters ensure guaranteed intra-cluster similarity. Our algorithm does not require the users to select and try various parameters combinations in order to getthe desired output. The key feature of our **MSTCG** algorithm is it fuses the advantages of both *density* and *graph-based* clustering approaches. Our algorithm finds clusters and outliers with less computational time. Our results are compared with UPMGA algorithm with both the distance measure namely k-tuple and DMk[8]. Our **MSTCG** algorithm gives better results than the existing methods. The running time of the algorithm is also less compared with the existing algorithm. The clustering performance of different clustering methods is the result of a combination of factors, including the types of sequence distance used for clustering and the choice of clustering algorithms In the future we will explore and test our proposed clustering algorithm in various different domains.

## References

[1] Alon.U, Barkai.N, Notterman.D, Gish.K, Ybarra.S, Mack.D and Levine.A. “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucloetide array”, Proceedings of National Academy of sciences USA, vol. 96, No. 12, pp. 6745-6750.

[2] T. Asano, B. Bhattacharya, M.Keil and F.Yao. “Clustering Algorithms based on minimum and maximum spanning trees”. *In Proceedings of the 4th Annual Symposium on Computational Geometry*,Pages 252-257, 1988.

[3] D. Avis “Diameter partitioning.” *Discrete and Computational Geometr*, 1:265-276, 1986.

[4] Bandyopadhyay, s.et al., An improved algorithm for clustering gene expression data, Bioinformatics, vol. 23(21), pp. 2859-2865, 2007.

[5] Bellaachia.A, Protnoy.D, Chen.Y and Elkahloun.A, “E-CAST: a data mining algorithm for gene expression data”, BIOKDD02: Workshop on Data Mining in Bioinformatics (with SIGKDD02 Conference), pp.49, 2002.

[6] Ben-Dor.A, Shamir. R and Yakhini.Z. “Clustering gene expression patterns”, Journal of Computational Biology, vol.6, Nos. 3-4, pp. 281-297, 1999.

[7] Cao Y, Janke A, Waddell PJ, Westerman M, Takenaka O, Murata S, Okada N, Paabo S, Hasegawa M: “Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders”, JMol Evol, 47(3)307-322, 1998.

[8] Dan Wei, Qingshan Jiang, Yanjie Wei and Shengrui Wang. “A novel hierarchical clustering algorithm for gene sequences”, BMC Bioinformatics 13:172 pp 1-15, 2012.

[9] Dempster.A, Laird.N and Rubin.D,“Maximum likelihood from incomplete data via the EM algorithm”, Journal of the Royal Statistical Society, vol.39, No. 1, pp 1-38, 1977.

[10] J.L. DeRisi, V.R. Iyer and P.O. Brown, “Exploring the metabolic and genetic control of gene expression on a genomic scale”, Science, 278(5338):680-686,1997

[11] Durbin R, Eddy SR, Krogh A, Mitchison G, “Biological Sequence Analysis probabilistic models of proteins and nucleic acids. Cambridge”, Cambridge University Press, 1998

[12] Eisen. M, Spellman.P, Brown.P and Botstein.D, “Cluster analysis and display of genome-wide expression patterns”, Proceedings of National Academy of Sciences USA, vol.95, pp 14863-14868, 1998.

[13] Fangxiang Wu.A, “Genetic Weighted K-means Algorithm for Clustering Gene Expression Data”, BMC Bioinformatics, 2008, 9(6): 68-75

[14] Feng J, Hu Y, Wan P, Zhang A, Zhao W, “New method for comparing DNA primary sequences based on a discrimination measure”,. J Theor Biol, 266(4): 703-707, 2010.

[15] Han, J. and Kamber, M. “Data Mining: Concepts and Techniques”, The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series editor Morgan Kaufmann Publishers, 550 pages, 2000.

[16] Hautamaki, V. Karkkainen, I. and Franti, P. “Outlier Detection Using k- Nearest Neighbor Graph”. In Proceedings of the International Conference on Pattern Recognition, Volume 3 pages 430 – 433, Cambridge, UK, 2004.

[17] Herrero, J. et al., “A hierarchical unsupervised growing neural network for clustering gene expression patterns, Bioinformatics”, vol. 17,pp. 126136, 2001.

[18] Jiang.D and Zhang. A, “DHC: a density-based hierarchical clustering method for time series gene expression data”, Proceedings of BIBLE2003: 3rd IEEE International Symposium and Bioinformatics and Bioengineering, Bethesda, Maryland, pp. 393-400,2003.

[19] Jin, W. Tung, A. and Han, J. “Mining Top-n Local Outliers in Large Databases”. In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 293 – 298, 2001.

[20] S. John Peter, S.P. Victor, "A Novel Algorithm for Meta similarity clusters using Minimum spanning tree". International Journal of computer science and Network Security. Vol.10 No.2 pp. 254 – 259, 2010.
[21] S. John Peter, "Detection of Outliers and Hubs Using Minimum Spanning Tree Based on Analytical perspective of Degree Numbers", Journal of Discrete Mathematical Sciences and Cryptography Vol 14, No. 5, pp 475- 488, 2011.
[22] Julia Ponomarenko, Tatyana Merkulova, Galina Orlova, Olef Fokin, Elena Gorshkov, Mikhail Ponomarenko, "Mining DNA sequences to predict sites which mutations cause genetic diseases", Institute of Cytology and Genetics, 10 Lavrentyev Ave., 630090 Novosibirsk, Russia. Knowledge-Based Systems 15, 225-233, 2002.
[23] Lixin Tang, Zihou Yang, "Improve the K-means algorithm using genetic algorithms",J. Mathematical Statistics and Applied Probability, 12940:350-356, 1997.
[24] Page RD: TreeView: "an application to display phylogenetic trees on personal computers", Bioinformatics 12:357-358, 1996.
[25] S.Papadimitriou, H.Kitawaga, P.Gibbons and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral", *Proc. Of the International Conference on Data Engineering* ,pp.315-326, 2003.
[26] Pierre Baldi and Soren Brunak." Bioinformatics: The machine learning approach".2001.
[27] Rosy Das, Jugal Kalita, Dhruba K. Bhattacharyya, "A new approach for clustering gene expression time series data". Int. Journal of Bioinformatics Research and Applications, vol. 5, No. 3, pp 310-328, 2009.
[28] Sharan and Shamir.R. "CLICK: a clustering algorithm with application to gene expression analysis", Proc. Eighth Int. Conf. on Intelligent Systems for Molecular Biology, AAAI Press, pp. 307-316, 2000
[29] Waterman MS: "Introduction to Computational Biology: Maps, Sequences and Genomes", Lodnon: Chapman and Hall, 1995.
[30] Wei D, Jiang Q, "A DNA Sequence Distance Measure Approach for Phylogenetic Tree Construction", In 5th IEEE International Conference on Bio-Inspired Computing: Theories and Applications: 23-26 Sept. Changsha, 204-212, 2010.
[31] White.J.R, Navlakha S., Nagarajan N., Ghodsi M., Kingsford C., Pop M, "Alignment and Clustering of Phylogenetic markers-implications for microbial diversity studies", BMC bioinformatics, 11:152, 2010
[32] Y.Xu, V.Olman and D.Xu. "Minimum spanning trees for gene expression data clustering", *Genome Informatics*, 12:24-33, 2001

## Author Profile

**S. John Peter** is working as Assistant professor in Computer Science, St.Xavier's college (Autonomous), Palayamkottai, Tirunelveli. He earned his M.Sc degree from Bharadhidasan University, Trichirappalli. He also earned his M.Phil from Bharadhidasan University, Trichirappalli. He also obtained his Ph.D degree in Computer Science from Manonmaniam Sundranar University, Tirunelveli. He has published research papers on clustering algorithm in various national and international Journals.