An Empirical Study of Partition based Clustering Algorithms and its Effectiveness over Gene Expression Data

Pallab Borah, Rituporna Dowerah, Koyel Mandal, Rosy Sarmah

ABSTRACT-Computational approaches such as clustering has found a lot of use in the analysis of Gene Expression Data. In this work, we attempt to study and compare four Partition based clustering algorithms (K-means, SOM, SOTA, and QT Clustering) over two Gene Expression datasets. Proximity measures play a major role in cluster analysis. In this paper we also study four different proximity measures (Pearson's correlation coefficient, Cosine similarity, Kendall tau rank correlation coefficient, and Spearman's rank correlation coefficient). Our experimental analysis show that among all the algorithms Self-Organizing Tree (SOTA) Algorithm has performed consistently and among the proximity measures, Pearson's Correlation Coefficient and Cosine Correlation has performed significantly well over the yeast dataset. We also test our results on the blood cancer data of Homo sapiens and found that highly enriched clusters contain genes responsible for blood cancer.

Keywords: *Partition based Clustering, Gene Expression Data, Proximity measures, p-value.*

1. INTRODUCTION

Gene expression is a mechanism by which a gene expresses itself in the phenotype of an organism. The Gene Expression Data (GED) is the output of microarray experiments and is in the form of a matrix with rows corresponding to genes and columns representingdifferent samples. These matrices have to be further analyzed in order to extract information about underlying biological processes. Genes having similar patterns of expression can be grouped together with the premise that they have similar cellular function. This can be done by a process called clustering. Clustering [1] is a process of classifying data objects into disjoint groups such that the objects in the same group or class have high similarity to each other, while data objects belonging to different groups are dissimilar. These disjoint classes or groups are known as clusters. Gene Expression Data clusteringare typically of three types:(i) Gene-based clustering that considerssamples as features and genes as data objects, (ii) Sample based clustering that treats genes as features and samples as data objects and (iii) Subspace clustering which considers both genes and samples symmetrically [1]. Gene-based clustering has many approaches and some of them are mentioned below. Partition based clustering algorithm constructs k partitions of data where each partition represents a cluster, i.e. it classifies the data into k groups such that each group contains at least one object and each object belongs to exactly one group [1]. Hierarchical clustering algorithm [1] generates a hierarchical series of nested clusters that can be represented by a tree called dendrogram. These types of algorithms are classified into two types: Agglomerative algorithms, which is a bottom-up approach and Divisive algorithms, which is also known as top-down approach. Graph theoretical approaches [1] are presented in terms of graphs. This type of clustering techniques converts the problems of dataset clustering into graph theoretical problems. Density-based clustering algorithms use order preservation ranking and regulation information in order to identify relevant clusters in GED. It clusters GED with high accuracy and is also found to be robust to outliers [2]. Clustering can be considered the most important unsupervised learning since it deals with finding a structure in a collection of unlabeled data.

For clustering algorithms, proximity measures play an important role. Choosing an appropriate proximity measure is of great importance to achieve satisfactory clustering results. Proximity measure, also called similarity measure is a method to compute similarity between data objects. In this work, we study four different proximity measures and compare among them. The four measures are as follows: Pearson's correlation coefficient (PCC), Spearman's rank correlation coefficient [1], Cosine similarity measure and Kendall's tau rank correlation coefficient. PCC measures the similarity of the changes in the expression levels of two profiles (patterns). It measures the strength of linear relationship between two patterns [3]. Spearman's correlation coefficient [1] measures the strength of the monotonic relationship between paired data. Cosine similarity measure [4] is a classic measure which computes the normalized dot product of two attributes. It gives the cosine of the angle between two attributes. Therefore, it is a measure of orientation and not magnitude. Kendall's tau coefficient [5] measures the ordinal association between two measured quantities.

In this paper, we will empirically study the performance of four partition based clustering algorithms using four proximity measures. From our study we will choose the best algorithm among them and use it for further analysis on human blood cancer data. We will finally try to predict genes responsible for blood cancer.

Cancer is the growth of uncontrolled genes anywhere in the body. There are many different types of cancer. For example, Breast cancer, Lung cancer, Skin cancer, etc. Blood

cancer [6] is a type of cancer that is initiated in blood forming tissue such as bone marrow, or in the cells of the immune system. There are three categories of blood cancer: Leukemia, Lymphoma and Myeloma. Most of these cancers are initiated in the bone marrow. A stem cell transplant as a part of blood cancer treatment may be the only best chance for a cure. A Gene Biomarker (GB) [6] is a DNA (or RNA) sequence that shows normal biological processes, pathogenic processes and response to therapeutic intervention. A biomarker reflects the expression, function and the regulation of a gene. Therefore, it is a biological parameter. Biomarker measurements help in the development and evaluation of novel therapies by comparing clinical responses to the effects of interventions of molecular and clinical pathways. This helps researchers to understand the differences in clinical responses [7].

The rest of the paper is as follows: In section 2, we discuss about the related work, i.e. the four algorithms that we have considered. Section 3 gives the results of our experimental analyses comparing the four algorithms and the four proximity measures. The subsections of section 3 give the descriptions of the datasets on which the analyses were carried out, as well as the results separately. Also in subsection 3.4 we mention the biomarkers detected from the human blood cancer dataset. Finally we conclude in section 4.

2. RELATED WORK

In this section, we discuss the four partition based algorithms: K-means algorithm, QT clustering, Self-Organizing Map and Self-Organizing Tree Algorithm. As we mentioned in Section 1, partition based algorithms partition the dataset into k clusters such that $k \leq n$, where n is the total number of data objects.

2.1 K-MEANS ALGORITHM

K-means [8] is an algorithm that partitions an n-dimensional dataset into k disjoint sets on the basis of a sample. The resultant clusters appear to be efficient in the sense of within-class variance. The concept of k-means generalizes the ordinary sample mean. This algorithm is fast and simple. The time complexity of k-means is $O(l * k * n)$ where l is the number of iterations and k is the number of clusters.The main idea of K-means Algorithm is to define k-centers, one for each cluster. These centers should be placed in an intelligent way because different locations compute different results. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroid as barycenter of the clusters resulting from the previous step. After we have these k new centroid, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more [8].

Jiang et al. [1] states in their paper that K-means algorithm typically converges into small number of iterations. There are several drawbacks of this algorithm. First, k is unknown in advance and has to be specified by the user. To find the optimal number of clusters, users repeatedly run the algorithm with different k and compare the results. In this work, we have used 3 methods namely the elbow method, the gap static method and the average silhouette method to find the optimum number of k. Second, since k-means forces each gene into a cluster, as a result the algorithm may be sensitive to noise [1].

2.2 QT CLUSTERING

The Quality Threshold cluster algorithm [9] was developed to find large clusters that have a quality guarantee. This algorithm requires specifying a threshold diameter and the minimum number of elements in each cluster.The algorithm works as follows: Starting with the first data object (gene), a candidate cluster is formed. The genes which have highest similarity to it, are added to the cluster, iteratively. Each iteration adds the gene that minimizes the increase in cluster diameter. The process continues until no further genes can be added to the cluster without surpassing the threshold diameter. A second cluster is formed by repeating the above process. The whole process is terminated when the largest remaining cluster has fewer than the prespecified minimum number of elements [9].This algorithm can be used to find cell-cycle regulated genes. Also it can initiate a cluster with a specific gene of interest, which is more effective than clustering the whole dataset and choosing the cluster that contains the specific gene. However, it has some drawbacks. First, it is computationally intensive and also time consuming. The computational time is increased due to increase in minimum cluster size, the number of genes on the selected gene list and also due to decrease in the minimum correlation. Second, QT Clustering is more expensive than the K-means algorithm. However, it overcomes the weaknesses of K-means, which means it doesn't require specifying the number of clusters [9].

2.3 SELF-ORGANIZING MAP

The Self-Organizing Map (SOM) [10,11] was developed based on a single layered neural network by Kohonen. In this process, the data objects (genes) that are presented at the input and output neurons are arranged in a simple neighborhood structure called lattice. The neurons are associated with a reference vector, and eachdata point is mapped to neurons with the closest reference vector. In the process of running the algorithm, the data objects act as training samples. After the training, clusters are identified by mapping all the data points to the output neurons [1]. SOM generates a map of high dimensional data set in 2D or 3D where the similar clusters are placed near to each other. The approach of SOM is more robust than k-means to the outliers. However, the users need to input the number of clusters and the grid structure of the neuron map. Furthermore, SOM is not effective in the case when data set contains irrelevant data points, such as genes with invariant patterns. In this case, SOM produces an output in which the majority of clusters are populated by this type of data [1].

2.4 SELF-ORGANIZING TREE ALGORITHM

The Self-Organizing Tree Algorithm (SOTA) [12] is a type of unsupervised, growing, self-organizing neural networkthat expands itself depending on the taxonomic relationship that exists among the sequences being classified. It was introduced mainly to construct phylogenetic trees from biological sequences, based on the principles of Kohonen's Self-Organizing Maps and on Fritzke's growing cell structure [12]. The initial system of SOTA consists of two external cells that are connected by an internal cell. These cells are initialized randomly with numbers ranging from 0 to 1. The size of each vector is same. On the basis of Kohonen's model, the input space is defined by the experimental input data, whereas the output space consists of a set of nodes, arranged according to certain topologies, usually two-dimensional grids. The algorithm maps the input space onto the smaller output space, which in result produces a reduction in the complexity of the analyzed data set. SOTA gives an output in the form of a binary tree topology, which in turn incorporates the principles of the growing cell structures algorithm of Fritzke. A series of nodes that are arranged in a binary tree are adapted to the characteristics of input data. The growing of the output nodes can be stopped at a desirable taxonomic level, or they can grow until every gene in the input data set is classified [13]. One of the advantages of SOTA is that its topology is a hierarchical tree. SOTA depends on the total size of the cells which implies the time complexity of SOTA to be a linear function. As a result, SOTA proves to be a promising algorithm for classification of large datasets.Herrero et al. [13] used SOTA to analyze gene expression data and the result obtained was similar to the results of hierarchical clustering with the robustness and accuracy of a neural network.

3. RESULTS

In this section, we will give the descriptions of the two datasets that we have used and all the different results starting from the optimal number of k.

3.1 DATASET DESCRIPTION

Table 1: Dataset description

	DATASET-1	DATASET-2
Organism	Saccharomyces Cerevisiae	Homo Sapiens
Data Details	Yeast Sporulation	Blood Cancer
Source	http://anirbanmukhopadhyay.50webs.com/data.html [14]	Armstrong 2002 V2 [15]
#Genes	474	2194
#Samples	7	72

3.2 RESULTS FOR THE INPUT OF K-MEANS AND SOM

As mentioned earlier in Section 2, we have found out the optimal number of k using Elbow method [16] that looks at the total Within-cluster Sum of Squares (WSS) as a function of the number of clusters, Average silhouette method [16] computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximizes the average silhouette over a range of possible values for k and Gap static method [16] that compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data.

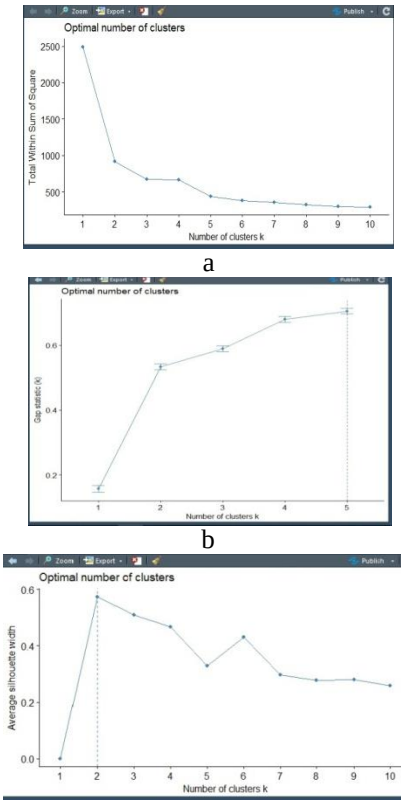


Figure 1: Computation of optimal number of k using (a) Elbow method (b) Gap static method (c) Silhouette method

- Elbow method: 5 clusters solution suggested.
- Silhouette method: 2 clusters solution suggested.
- Gap static method: 5 clusters solution suggested.

According to these observations, it is possible to define k=5 as the optimal number of clusters in the data.

For finding the number of neurons in case of Self-Organizing Map, we have used the equation $M = 5\sqrt{N}$, where M is the number of neurons, which is an integer close to the result of the right hand side of the equation, and N is the number of observations.We found M=13.44 and approximated it to 13.

3.3 RESULTS ON YEAST DATA

We have run the four partition based algorithms with the four proximity measures on the yeast dataset, using the tool MeV [17]. Multiple Experiment Viewer [17] is a cloud-based application that supports analysis, visualization and stratification of large genomic data, particularly for RNASeq and microarray data. After that,we used the tool FuncAssociate [18] for generating the p-values of the genes of

the clusters. FuncAssociate [18] is a web-based tool to help researchers use Gene Ontology attributes to characterize large sets of genes derived from experiments. We have used ClusterJudge [19] to calculate z-scores for the yeast dataset. ClusterJudge [19] is a tool that is used to judge quality of clustering methods performed elsewhere on some entities using mutual information. This tool supports only yeast datasets.

3.3.1 P-VALUE ANALYSIS

P values measure probability of finding the number of genes involved in a given Gene Ontology (GO) term within a cluster [18]. A low p-value indicates that the genes belonging to the enriched functional categories are biologically significant in the corresponding clusters. We have referred to significantly enriched clusters as those, whose p-values are above 5% significance level. The percentage of significantly enriched clusters is calculated by the formula below:

$$\%\ \text{ofsignificantlyenrichedclusters} = \frac{\text{No. ofclustersabove 5\% significancelevel}}{\text{Totalno. ofclusters}} \times 100\%$$

Table 2: Percentage of significantly enriched clusters

ALGORITHMS	PROXIMITY MEASURES	PERCENTAGE OF SIGNIFICANT CLUSTERS
K-Means	Cosine Similarity	100%
	Kendall's Tau coefficient	100%
	Pearson's correlation coefficient	100%
	Spearman's correlation coefficient	100%
QT Cluster	Cosine Similarity	100%
	Kendall's Tau coefficient	100%
	Pearson's correlation coefficient	100%
	Spearman's correlation coefficient	100%
Self-Organizing Map	Cosine Similarity	92.3%
	Kendall's Tau coefficient	92.3%
	Pearson's correlation coefficient	92.3%
	Spearman's correlation coefficient	92.3%
SOTA	Cosine Similarity	100%
	Kendall's Tau coefficient	80%
	Pearson's correlation coefficient	100%
	Spearman's correlation coefficient	80%

Table 3 gives the p-values of clusters obtained by K-means, QT clustering, SOM and SOTA with Cosine Similarity, Kendall's tau, Pearson's coefficient and Spearman's coefficient as the proximity measures, with p-value below $3 \times e^{-10}$. To restrict the size of the article, we have mentioned only the highest two values of each cluster.

Table 3: P-values obtained by using K-means, QT clustering, SOM and SOTA with Cosine similarity, Kendall's tau coefficient, Pearson's correlation coefficient and spearman's coefficient

Algorithm	Proximity Measures	Clusters	P value	GO id	GO Categories
K-means Clustering	Cosine Similarity	C1	1.77E-16	GO:0006090	pyruvate metabolic process
			1.51E-15	GO:0046496	nicotinamide nucleotide metabolic process
			1.93E-15	GO:0019362	pyridine nucleotide metabolic process
			4.67E-15	GO:0006006	glucose metabolic process
			7.83E-15	GO:0046031	ADP metabolic process
		C2	4.39E-50	GO:0002181	cytoplasmic translation
			7.12E-42	GO:0030529	intracellular ribonucleoprotein complex
			7.12E-42	GO:1990904	ribonucleoprotein complex
		C3	1.47E-29	GO:0051321	meiotic cell cycle
			2.32E-25	GO:0007131	reciprocal meiotic recombination
			2.32E-25	GO:0035825	reciprocal DNA recombination
		C4	3.69E-42	GO:0048646	anatomical structure formation involved in morphogenesis
			4.16E-38	GO:0043934	Sporulation
	Kendall's tau coefficient	C1	5.62E-21	GO:0048646	anatomical structure formation involved in morphogenesis
			2.63E-18	GO:0043934	Sporulation
			9.74E-18	GO:1903046	meiotic cell cycle process
		C2	1.29E-15	GO:0006094	Gluconeogenesis
			1.29E-15	GO:0019319	hexose biosynthetic process
			1.75E-15	GO:0006006	glucose metabolic process
			2.77E-15	GO:0006090	pyruvate metabolic process
			3.26E-15	GO:0046364	monosaccharide biosynthetic process
			8.14E-14	GO:0046031	ADP metabolic process
		C3	1.25E-31	GO:1903046	meiotic cell cycle process
			2.00E-30	GO:0044702	single organism reproductive process
		C4	1.26E-35	GO:0002181	cytoplasmic translation
			7.89E-34	GO:0030529	intracellular ribonucleoprotein complex
			7.89E-34	GO:1990904	ribonucleoprotein complex
		C5	4.91E-16	GO:0010927	cellular component assembly involved in morphogenesis
			7.09E-16	GO:0030476	ascospore wall assembly
			7.09E-16	GO:0042244	spore wall assembly
			9.00E-16	GO:0071940	fungal-type cell wall assembly
			1.14E-15	GO:0070726	cell wall assembly

			2.62E-15	GO:0048646	anatomical structure formation involved in morphogenesis
			8.81E-15	GO:0003006	developmental process involved in reproduction
	Pearson's Correlation Coefficient	C1	1.77E-16	GO:0006090	pyruvate metabolic process
			1.51E-15	GO:0046496	nicotinamide nucleotide metabolic process
			1.93E-15	GO:0019362	pyridine nucleotide metabolic process
			4.67E-15	GO:0006006	glucose metabolic process
			7.83E-15	GO:0046031	ADP metabolic process
		C2	4.39E-50	GO:0002181	cytoplasmic translation
			7.12E-42	GO:0030529	intracellular ribonucleoprotein complex
			7.12E-42	GO:1990904	ribonucleoprotein complex
		C3	1.47E-29	GO:0051321	meiotic cell cycle
			2.32E-25	GO:0007131	reciprocal meiotic recombination
			2.32E-25	GO:0035825	reciprocal DNA recombination
		C4	3.69E-42	GO:0048646	anatomical structure formation involved in morphogenesis
			4.16E-38	GO:0043934	Sporulation
	Spearman's correlation coefficient	C1	1.49E-35	GO:1903046	meiotic cell cycle process
			3.30E-34	GO:0044702	single organism reproductive process
		C2	9.63E-12	GO:0006090	pyruvate metabolic process
			1.67E-11	GO:0032787	monocarboxylic acid metabolic process
			2.67E-11	GO:0019752	carboxylic acid metabolic process
			7.25E-11	GO:0043436	oxoacid metabolic process
			7.67E-11	GO:0006082	organic acid metabolic process
			9.73E-11	GO:0006094	Gluconeogenesis
			9.73E-11	GO:0019319	hexose biosynthetic process
		C3	2.59E-48	GO:0002181	cytoplasmic translation
			2.73E-39	GO:0005840	Ribosome
			3.05E-39	GO:0003735	structural constituent of ribosome
			5.63E-39	GO:0030529	intracellular ribonucleoprotein complex
			5.63E-39	GO:1990904	ribonucleoprotein complex
			9.06E-39	GO:0044445	cytosolic part
		C4	1.85E-27	GO:0048646	anatomical structure formation involved in morphogenesis
			3.52E-25	GO:0003006	developmental process involved in reproduction
QT Clustering	Cosine Similarity	C1	3.80E-55	GO:1903046	meiotic cell cycle process
			1.66E-52	GO:0044702	single organism reproductive process

		C2	1.76E-46	GO:0002181	cytoplasmic translation
			1.68E-37	GO:0005840	Ribosome
			1.69E-37	GO:0003735	structural constituent of ribosome
			4.49E-37	GO:0044445	cytosolic part
		C3	8.09E-12	GO:0006094	Gluconeogenesis
			8.09E-12	GO:0019319	hexose biosynthetic process
			1.44E-11	GO:0046364	monosaccharide biosynthetic process
	Kendall's tau	C1	2.99E-56	GO:1903046	meiotic cell cycle process
			1.82E-53	GO:0044702	single organism reproductive process
		C2	8.79E-47	GO:0002181	cytoplasmic translation
			1.57E-37	GO:0044445	cytosolic part
			1.16E-36	GO:0003735	structural constituent of ribosome
		C3	5.48E-20	GO:0030684	Preribosome
			9.43E-17	GO:0022613	ribonucleoprotein complex biogenesis
	Pearson's correlation coefficient	C1	1.01E-51	GO:1903046	meiotic cell cycle process
			1.73E-49	GO:0044702	single organism reproductive process
		C2	2.34E-50	GO:0002181	cytoplasmic translation
			2.74E-41	GO:0005840	Ribosome
			3.46E-41	GO:0003735	structural constituent of ribosome
			4.59E-41	GO:0030529	intracellular ribonucleoprotein complex
			4.59E-41	GO:1990904	ribonucleoprotein complex
		C3	9.94E-12	GO:0006094	Gluconeogenesis
			9.94E-12	GO:0019319	hexose biosynthetic process
			1.52E-11	GO:0006090	pyruvate metabolic process
			1.77E-11	GO:0046364	monosaccharide biosynthetic process
			9.60E-11	GO:0046496	nicotinamide nucleotide metabolic process
	Spearman's correlation coefficient	C1	8.57E-56	GO:1903046	meiotic cell cycle process
			3.14E-53	GO:0044702	single organism reproductive process
		C2	4.81E-52	GO:0002181	cytoplasmic translation
			6.09E-43	GO:0005840	Ribosome
			8.55E-43	GO:0003735	structural constituent of ribosome
		C3	6.30E-13	GO:0006090	pyruvate metabolic process
			1.34E-11	GO:0006094	Gluconeogenesis
			1.34E-11	GO:0019319	hexose biosynthetic process
			1.62E-11	GO:0006006	glucose metabolic process
			2.39E-11	GO:0046364	monosaccharide biosynthetic process
			5.82E-11	GO:0046031	ADP metabolic process

			8.23E-11	GO:0009135	purine nucleoside diphosphate metabolic process
			8.23E-11	GO:0009179	purine ribonucleosidediphosphate metabolic process
			8.23E-11	GO:0009185	ribonucleosidediphosphate metabolic process
Self-Organizing Map	Cosine Similarity	C1	1.00E-12	GO:0046031	ADP metabolic process
			1.37E-12	GO:0009135	purine nucleoside diphosphate metabolic process
			1.37E-12	GO:0009179	purine ribonucleosidediphosphate metabolic process
			1.37E-12	GO:0009185	ribonucleosidediphosphate metabolic process
			2.65E-12	GO:0046496	nicotinamide nucleotide metabolic process
			3.04E-12	GO:0019362	pyridine nucleotide metabolic process
			3.20E-12	GO:0009132	nucleoside diphosphate metabolic process
			6.91E-12	GO:0006006	glucose metabolic process
			8.78E-12	GO:0006090	pyruvate metabolic process
			1.38E-11	GO:0072524	pyridine-containing compound metabolic process
			1.73E-11	GO:0006733	oxidoreduction coenzyme metabolic process
			1.95E-11	GO:0006096	glycolytic process
			1.95E-11	GO:0006757	ATP generation from ADP
			2.82E-11	GO:0006732	coenzyme metabolic process
			3.68E-11	GO:0006165	nucleoside diphosphate phosphorylation
			3.72E-11	GO:0006163	purine nucleotide metabolic process
		C2	7.75E-20	GO:0002181	cytoplasmic translation
			4.56E-16	GO:0044445	cytosolic part
			9.24E-16	GO:0003735	structural constituent of ribosome
		C3	1.09E-31	GO:0002181	cytoplasmic translation
			8.45E-28	GO:0003735	structural constituent of ribosome
		C4	8.36E-23	GO:0030684	Preribosome
			8.50E-18	GO:0022613	ribonucleoprotein complex biogenesis
		C5	3.55E-24	GO:0048646	anatomical structure formation involved in morphogenesis
			2.58E-23	GO:0003006	developmental process involved in reproduction
			2.58E-23	GO:0043934	Sporulation
		C6	3.76E-13	GO:0048646	anatomical structure formation involved in morphogenesis
			1.02E-11	GO:0030154	cell differentiation

			1.02E-11	GO:0030435	sporulation resulting in formation of a cellular spore
			1.12E-11	GO:0043934	Sporulation
			1.28E-11	GO:0005628	prospore membrane
			3.25E-11	GO:0044767	single-organism developmental process
			3.44E-11	GO:0032502	developmental process
		C7	4.49E-27	GO:0051321	meiotic cell cycle
			6.67E-23	GO:0007131	reciprocal meiotic recombination
			6.67E-23	GO:0035825	reciprocal DNA recombination
	Kendall's tau	C1	1.25E-13	GO:0046031	ADP metabolic process
			1.70E-13	GO:0009135	purine nucleoside diphosphate metabolic process
			1.70E-13	GO:0009179	purine ribonucleosidediphosphate metabolic process
			1.70E-13	GO:0009185	ribonucleosidediphosphate metabolic process
			4.01E-13	GO:0009132	nucleoside diphosphate metabolic process
			8.67E-13	GO:0006006	glucose metabolic process
			1.10E-12	GO:0006090	pyruvate metabolic process
		C2	1.26E-30	GO:0002181	cytoplasmic translation
			1.64E-27	GO:0044445	cytosolic part
			4.69E-27	GO:0003735	structural constituent of ribosome
			8.62E-27	GO:0044391	ribosomal subunit
		C3	2.19E-24	GO:0002181	cytoplasmic translation
			4.96E-22	GO:0022625	cytosolic large ribosomal subunit
			8.87E-22	GO:0003735	structural constituent of ribosome
		C4	8.01E-23	GO:0030684	Preribosome
			1.02E-20	GO:0022613	ribonucleoprotein complex biogenesis
		C5	9.47E-12	GO:0048646	anatomical structure formation involved in morphogenesis
			2.78E-11	GO:0030154	cell differentiation
			2.78E-11	GO:0030435	sporulation resulting in formation of a cellular spore
			3.01E-11	GO:0003006	developmental process involved in reproduction
			3.01E-11	GO:0043934	Sporulation
			4.13E-11	GO:1903046	meiotic cell cycle process
		C6	3.32E-12	GO:0048646	anatomical structure formation involved in morphogenesis
			6.88E-11	GO:0044702	single organism reproductive process
		C7	1.41E-27	GO:0051321	meiotic cell cycle

			3.15E-27	GO:0007131	reciprocal meiotic recombination
			3.15E-27	GO:0035825	reciprocal DNA recombination
			1.04E-25	GO:0044702	single organism reproductive process
			1.08E-25	GO:1903046	meiotic cell cycle process
	Pearson's correlation coefficient	C1	4.01E-20	GO:0051321	meiotic cell cycle
			4.84E-17	GO:0007131	reciprocal meiotic recombination
			4.84E-17	GO:0035825	reciprocal DNA recombination
		C2	7.56E-11	GO:0007131	reciprocal meiotic recombination
			7.56E-11	GO:0035825	reciprocal DNA recombination
			1.35E-10	GO:0051321	meiotic cell cycle
		C3	5.00E-17	GO:0043934	Sporulation
			6.77E-16	GO:0048646	anatomical structure formation involved in morphogenesis
		C4	6.69E-19	GO:0048646	anatomical structure formation involved in morphogenesis
			1.34E-15	GO:0044767	single-organism developmental process
			1.46E-15	GO:0032502	developmental process
			2.26E-15	GO:0030154	cell differentiation
			2.26E-15	GO:0030435	sporulation resulting in formation of a cellular spore
			2.57E-15	GO:0003006	developmental process involved in reproduction
			2.57E-15	GO:0043934	Sporulation
			7.53E-15	GO:0030476	ascospore wall assembly
			7.53E-15	GO:0042244	spore wall assembly
		C5	7.34E-24	GO:0030684	Preribosome
			1.41E-20	GO:0030529	intracellular ribonucleoprotein complex
			1.41E-20	GO:1990904	ribonucleoprotein complex
		C6	3.00E-39	GO:0002181	cytoplasmic translation
			3.03E-35	GO:0044445	cytosolic part
		C7	6.91E-12	GO:0006006	glucose metabolic process
			8.64E-11	GO:0046031	ADP metabolic process
	Spearman's correlation coefficient	C1	3.78E-11	GO:0046031	ADP metabolic process
			4.72E-11	GO:0009135	purine nucleoside diphosphate metabolic process
			4.72E-11	GO:0009179	purine ribonucleosidediphosphate metabolic process
			4.72E-11	GO:0009185	ribonucleosidediphosphate metabolic process
			8.76E-11	GO:0009132	nucleoside diphosphate metabolic process

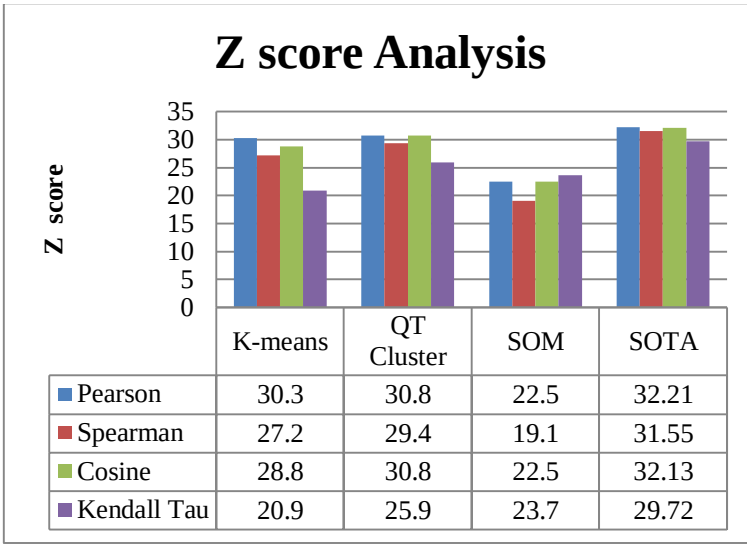
			1.53E-10	GO:0006006	glucose metabolic process
			1.83E-10	GO:0006090	pyruvate metabolic process
		C2	3.93E-47	GO:0002181	cytoplasmic translation
			6.47E-39	GO:0003735	structural constituent of ribosome
		C3	4.33E-13	GO:0030684	Preribosome
			1.48E-10	GO:0022613	ribonucleoprotein complex biogenesis
		C4	3.96E-17	GO:0048646	anatomical structure formation involved in morphogenesis
			8.68E-17	GO:0030154	cell differentiation
			8.68E-17	GO:0030435	sporulation resulting in formation of a cellular spore
			9.71E-17	GO:0043934	Sporulation
			2.95E-15	GO:0003006	developmental process involved in reproduction
			3.68E-15	GO:0048869	cellular developmental process
			3.84E-15	GO:0030476	ascospore wall assembly
			3.84E-15	GO:0042244	spore wall assembly
			3.87E-15	GO:0044767	single-organism developmental process
			4.13E-15	GO:0032502	developmental process
			4.87E-15	GO:0071940	fungal-type cell wall assembly
			6.15E-15	GO:0070726	cell wall assembly
		C5	3.15E-27	GO:0007131	reciprocal meiotic recombination
			3.15E-27	GO:0035825	reciprocal DNA recombination
			6.62E-26	GO:0051321	meiotic cell cycle
Self-Organizing Tree Algorithm	Cosine similarity	C1	4.32E-28	GO:0051321	meiotic cell cycle
			1.11E-26	GO:0007131	reciprocal meiotic recombination
			1.11E-26	GO:0035825	reciprocal DNA recombination
		C2	9.14E-42	GO:0048646	anatomical structure formation involved in morphogenesis
			9.17E-38	GO:0043934	Sporulation
		C3	1.34E-16	GO:0006090	pyruvate metabolic process
			1.08E-15	GO:0046496	nicotinamide nucleotide metabolic process
			1.38E-15	GO:0019362	pyridine nucleotide metabolic process
			3.61E-15	GO:0006006	glucose metabolic process
			6.18E-15	GO:0046031	ADP metabolic process
		C4	1.06E-50	GO:0002181	cytoplasmic translation
			2.15E-41	GO:0030529	intracellular ribonucleoprotein complex
			2.15E-41	GO:1990904	ribonucleoprotein complex
	Kendall's	C1	5.37E-61	GO:1903046	meiotic cell cycle process

	tau		8.03E-57	GO:0044702	single organism reproductive process
		C2	2.80E-13	GO:0006006	glucose metabolic process
			5.26E-13	GO:0006094	Gluconeogenesis
			5.26E-13	GO:0019319	hexose biosynthetic process
			9.40E-13	GO:0046364	monosaccharide biosynthetic process
			1.57E-11	GO:0019318	hexose metabolic process
			5.94E-11	GO:0005996	monosaccharide metabolic process
			8.55E-11	GO:0046031	ADP metabolic process
		C3	2.00E-47	GO:0002181	cytoplasmic translation
			5.71E-39	GO:0030529	intracellular ribonucleoprotein complex
			5.71E-39	GO:1990904	ribonucleoprotein complex
	Pearson's correlation coefficient	C1	4.32E-28	GO:0051321	meiotic cell cycle
			1.11E-26	GO:0007131	reciprocal meiotic recombination
			1.11E-26	GO:0035825	reciprocal DNA recombination
		C2	9.14E-42	GO:0048646	anatomical structure formation involved in morphogenesis
			9.17E-38	GO:0043934	Sporulation
		C3	1.34E-16	GO:0006090	pyruvate metabolic process
			1.08E-15	GO:0046496	nicotinamide nucleotide metabolic process
			1.38E-15	GO:0019362	pyridine nucleotide metabolic process
			3.61E-15	GO:0006006	glucose metabolic process
			6.18E-15	GO:0046031	ADP metabolic process
		C4	1.06E-50	GO:0002181	cytoplasmic translation
			2.15E-41	GO:0030529	intracellular ribonucleoprotein complex
			2.15E-41	GO:1990904	ribonucleoprotein complex
	Spearman's correlation coefficient	C1	8.47E-62	GO:1903046	meiotic cell cycle process
			1.05E-57	GO:0044702	single organism reproductive process
		C2	4.39E-14	GO:0006006	glucose metabolic process
			1.21E-13	GO:0006094	Gluconeogenesis
			1.21E-13	GO:0019319	hexose biosynthetic process
			2.17E-13	GO:0046364	monosaccharide biosynthetic process
		C3	2.00E-47	GO:0002181	cytoplasmic translation
			5.71E-39	GO:0030529	intracellular ribonucleoprotein complex
			5.71E-39	GO:1990904	ribonucleoprotein complex

3.3.2 Z SCORE ANALYSIS

The Z score is a test of statistical significance that helps to decide whether or not to reject the null hypothesis. Z scores are measures of standard deviation [19]. Higher z-scores indicate that the clustering results are more significantly related to the gene function [19].

Z score Analysis				
	K-means	QT Cluster	SOM	SOTA
Pearson	30.3	30.8	22.5	32.21
Spearman	27.2	29.4	19.1	31.55
Cosine	28.8	30.8	22.5	32.13
Kendall Tau	20.9	25.9	23.7	29.72



3.4 RESULTS ON HUMAN BLOOD CANCER DATA

From the above analysis over the yeast dataset in subsection 3.3, it is found that among all the algorithms Self-Organizing Tree Algorithm has performed consistently and among the proximity measures, Pearson's Correlation Coefficient and Cosine Correlation has performed significantly well over its counterparts.. However, from literature survey, we have found that Pearson's correlation coefficient has been widely used over gene expression data and has been proved to be better. Therefore, we run the SOTA using Pearson's similarity measure over the cancer dataset of Homo sapiens and found that the significantly enriched clusters obtained from SOTA gave gene biomarkers (genes responsible for cancer). Table 4 reports the results of the significantly enriched clusters obtained by SOTA.

Table 4: P values of the significant clusters obtained by SOTA using Pearson's correlation coefficient over the cancer dataset.

Clusters	P-value	GO numbers	GO categories
C1	2.13E-17	GO:0000786	Nucleosome
	4.59E-17	GO:0044815	DNA packaging complex
	4.46E-15	GO:0032993	protein-DNA complex
C2	2.78E-14	GO:0007165	signal transduction
	3.53E-13	GO:0007166	cell surface receptor signaling pathway
	5.48E-12	GO:0050789	regulation of biological process
	3.65E-11	GO:0048583	regulation of response to stimulus
	1.23E-10	GO:0050794	regulation of cellular process
	1.50E-10	GO:0006928	movement of cell or subcellular component
	1.73E-10	GO:0065007	biological regulation
	2.37E-10	GO:0004672	protein kinase activity
C3	1.40E-11	GO:0002376	immune system process
	1.60E-10	GO:0043299	leukocyte degranulation

	2.75E-10	GO:0045055	regulated secretory pathway
C4	6.79E-11	GO:0032502	developmental process
	1.49E-10	GO:0048519	negative regulation of biological process
	2.30E-10	GO:0048856	anatomical structure development
C5	2.68E-14	GO:0009605	response to external stimulus
	7.12E-13	GO:0009607	response to biotic stimulus
	7.98E-13	GO:0043207	response to external biotic stimulus
	7.36E-12	GO:0051704	multi-organism process
	2.27E-11	GO:0048583	regulation of response to stimulus
	2.68E-11	GO:0019221	cytokine-mediated signaling pathway
	6.70E-11	GO:0001775	cell activation
	2.02E-10	GO:0042127	regulation of cell proliferation

3.5 BIOMARKER DETECTION

The total number of clusters generated by the SOTA using Pearson's correlation coefficient over the blood cancer dataset is 11. Out of which 10 were significantly enriched clusters. As we have considered our p-value threshold for highly significant clusters as $3 \times e^{-10}$, we have got 5 clusters to be highly significant. We have constructed a network for the genes in each cluster using cytoscape 3.3.0 [20] and GeneMANIA plug in 3.4.1 [21]. From the constructed network with the help of Network Analyzer 3.3.2 [20] we have computed the degree of each of the nodes. We then took the top 10 highest degree genes to determine whether they are biomarkers of blood cancer or not. The validations of the biomarkers are being done by the GeneCards [22] and depicted in Table 5.

Table 5: The list of causal genes

GENE NAME	ENTREZ GENE ID	DEGREE	SOURCE
APP	351	239	GeneCards
TGFBR3	7049	214	GeneCards
TCF4	6925	212	GeneCards
CHEK1	1111	206	GeneCards
ITGAM	3684	230	GeneCards
CD14	929	100	GeneCards
SERPINA1	5265	202	GeneCards
S100A11	6282	133	GeneCards
MUC1	4582	236	GeneCards
ANXA3	306	225	GeneCards
CXCL12	6387	217	GeneCards
MMP9	4318	209	GeneCards
BMP2	650	195	GeneCards
ESR2	2100	8	GeneCards

4. CONCLUSION AND FUTURE WORK

From our experimental analysis over the yeast dataset, it is found that among all the algorithms Self-Organizing Tree Algorithm has performed consistently and among the proximity measures, Pearson's Correlation Coefficient and Cosine Correlation has performed significantly well over its

counterparts.. However, from literature survey, we have found that Pearson's correlation coefficient has been widely used over gene expression data and has been proved to be better.Therefore, we implement the SOTA using Pearson's similarity measure over the cancer dataset of Homo sapiens and found that the significantly enriched clusters obtained from SOTA gave gene biomarkers (genes responsible for disease).It has been observed that clusters below 5% significance level did not generate any biomarkers.In the blood cancer dataset of homo sapiens it has been seen that from the significantly enriched clusters detected by SOTA, a total of 14 gene biomarkers were found. The biomarkers have been validated by gene cards and literature. In this work, we attempted to give a way to analyze clustering algorithms and how they can be used in the detection of causal genes which in return would help in predicting cancer in patients. This would help in timely detection, prognosis and treatment of cancer.Here, we have taken into account only four partition based algorithms and four proximity measures. However, there is possibility of getting better results if the study is conducted taking more clustering algorithms and proximity measures into account.

REFERENCES

[1] Jiang, D., Tang, C. and Zhang, A., 2004. Cluster analysis for gene expression data: A survey. IEEE Transactions on knowledge and data engineering, 16(11), pp.1370-1386.

[2] Das, R., Bhattacharyya, D.K. and Kalita, J.K., 2010. Clustering gene expression data using an effective dissimilarity measure. International Journal of Computational BioScience (Special Issue), 1(1), pp.55-68.

[3] Son, Y.S. and Baek, J., 2008. A modified correlation coefficient based similarity measure for clustering time-course gene expression data. Pattern Recognition Letters, 29(3), pp.232-242.

[4] Ye, J., 2011. Cosine similarity measures for intuitionistic fuzzy sets and their applications. Mathematical and Computer Modelling, 53(1-2), pp.91-97.

[5] Kendall, M.G., (1955). *Rank Correlation Methods*. New York: Hafner Publishing Co.

[6] Mandal K., Sarmah R., Bhattacharyya D.K., 2018. Biomarker Identification for cancer Disease using Biclustering approach: An empirical study. IEEE/ACM transactions in computational biology and bioinformatics, 2018.

[7] Novelli, G., Ciccacci, C., Borgiani, P., Amati, M.P. and Abadie, E., 2008. Genetic tests and genomic biomarkers: regulation, qualification and validation. Clinical cases in mineral and bone metabolism, 5(2), p.149.

[8] MacQueen, J., 1967, June. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

[9] Heyer, L.J., Kruglyak, S. and Yooseph, S., 1999. Exploring expression data: identification and analysis of coexpressedgenes. Genome research, 9(11), pp.1106-1115.

[10] Kohonen, Teuvo; Honkela, Timo (2007). KohonenNetwor". Scholarpedia.

[11] Kohonen, Teuvo (1982). "Self-Organized Formation of Topologically Correct Feature Maps". Biological Cybernetics. 43 (1): 59–69.

[12]Dopazo, J., Carazo, J. M. (1997). Phylogenetic Reconstruction Using an Unsupervised Growing Neural Network That Adopts the Topology of a Phylogenetic Tree. *Journal of Molecular Evolution*. 44, pp. 226-233.

https://doi.org/10.1007/PL00006139

[13] Herrero, J., Valencia, A. and Dopazo, J., 2001. A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics, 17(2), pp.126-136.

[14] http://anirbanmukhopadhyay.50webs.com/data.html

[15] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," Nature Genetics, vol. 30, no. 1, pp. 41–47, 2002.

[16] http://www.sthda.com/english/articles/29-cluster-validation-essentials/96-determining-the-optimalnumber-of-clusters-3-must-know-methods/#average-silhouette-method

[17] Howe E., Holton K., Nair S., Schlauch D., Sinha R., Quackenbush J. (2010) MeV: MultiExperiment Viewer. In: Ochs M., Casagrande J., Davuluri R. (eds) Biomedical Informatics for Cancer Research. Springer, Boston, MA

[18] Berriz, G.F., King, O.D., Bryant, B., Sander, C. and Roth, F.P., 2003. Characterizing gene sets with FuncAssociate. Bioinformatics, 19(18), pp.2502-2504.

[19] Gibbons, F.D. and Roth F.P., (2002) Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation. Genome Research, vol. 12, pp1574-1581.

[20] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Research 2003 Nov; 13(11):2498-504

[21] Montojo J, Zuberi K, Rodriguez H, Bader GD, Morris Q. GeneMANIA: Fast gene network construction and function prediction for Cytoscape. *F1000Research*. 2014;3:153. doi:10.12688/f1000research.4572.1.

[22] Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D (April 1997). "GeneCards: integrating information about genes, proteins and diseases". Trends Genet. 13 (4): 163.