INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

# PATTERN SIMILARITY BASED CLUSTERING OF GENE EXPRESSION PATTERNS

**Irene Maria[1], Mr.Mathew Kurian[2]**

*[1]PG student, Department of Computer Science and Engineering, Karunya University, Tamilnadu,*
*m3irene@gmail.com*
*[2] Assistant Professor, Department of Computer Science and Engineering, Karunya University.*

## Abstract

For the identification of classes of same characteristics or similar objects among a set of objects, clustering can be used effectively. The definition of similarity can be different in one clustering model to another. The concept of similarity is often based on metrics as Manhattan distance, Euclidean distance, Pearson correlation coefficient or any other measures depending on the model which is used for clustering. Similar objects must have values which are close in at least any set of the dimensions. Clustering is a used as an unsupervised data analysis approach in machine learning in the field of data mining .Computation of pairwise distance in advance becomes a common requirement amongst many existing clustering methods. This makes it computationally expensive and difficult to manage with huge data sets used as in the field of bioinformatics. In the pattern similarity cluster model two objects can be told as similar if they show a pattern which is robust on a subset of the existing dimensions. The new similarity concept models a large variety of applications like as in the field of bioinformatics. As in DNA microarray analysis, the expression levels of two or more genes may increase and decrease synchronously according to the responses from the environmental incentives. The magnitude of their expression levels may not be close, but the patterns they exhibit can be more over same. Discovery of such clusters of genes is important in revealing significant information about gene regulatory networks. There are other applications that can also benefit from the new model, because it is able to capture not only the closeness of values but also the closeness of patterns showed by the any object present. Clustering methods have been applied to gene expression data sets in order to group genes sharing common or similar expression profiles into separate efficient groups. In such analyses, designing an appropriate (dis)similarity measure is critical. It is expected to be especially efficient when the shape of expression profile is vital in determining the gene relationship, yet the expression magnitude should also be taken into account for to some extent.

**Keywords**: *Gene expression, microarray technology, clustering, pattern similarity.*

## 1. Introduction

Cluster analysis (CA) partitions points of a data set into groups, so that data points within a group are more similar to each other than to the points in different groups. The use of clustering technique constitutes often a first step in a data mining process to reveal natural structures and identify underlying patterns in the data. Clustering is applied in many fields and leads an important role in a broad range of applications which include data mining. Applications of clustering deals with large datasets and data which possess many attributes, where there is the need of simplification or concise summaries.

It can provide an idea about the structure of the data. Objects can be clustered on the basis of a similarity measure. Conventional similarity measures are situation free and property-independent in nature. A recent developed way [1] to overcome such problems is the similarity measure. Consider the scenario where similarity (A, B) =f (A, B, E, C) where A and B are the objects that are being compared and E and C are the environment and a set of predefined concepts respectively. A knowledge-based approach to partition objects is discussed. An interesting subset of concepts called conceptual transformers, which is an abstraction from real life, can be taken into account.

There are many clustering methods in bioinformatics which have been applied for analysis of the gene expression data. Among them, most clustering models are distance based clustering which rely on measures like Euclidean distance, Manhattan distance and Cosine distance. But in many cases, these similarity functions are not always [2] suitable or precise in capturing the connections among the genes or the conditions. In reality, errors are obvious in biological experiments and perfect pattern matching in microarray data may not occur even among known co-ordinately structured genes.

The paper is organized as follows. In section 2, some basic concepts about clustering in the gene expressions are introduced. Section 3 presents the challenges faced by the proposed pattern similarity method. Finally, in section 4 give the conclusion of the paper.

## 2. Preliminaries

## 2.1 Clustering of gene expressions

Clustering can be used for pattern analysis as an effective tool as it has a large range of applications like mining in large data warehouse environments, dynamic routing, and text classification. In bioinformatics, the popular use of clustering is done where in gene expression analysis the expression profiles of genes are analysed across different biological conditions are clustered for pattern recognition. Genes that fall into the same cluster can be assumed to have some common functional properties [3], such as acting under control of the same regulatory mechanism. Another example of clustering in bioinformatics is the analysis of protein sequences to allocate a reputed function.

### *2.1.1Traditional Similarity Models*

Clustering when applied in high dimensional spaces or areas is frequently understood as difficult because theoretical outcomes examine the importance of nearby matching in high dimensional spaces of the data sets. Some research work [4] has focused on discovering clusters embedded in the subspaces of high dimensional data sets. Such a problem is known as subspace clustering. One of the well-known clustering algorithms [5] which are capable of finding clusters in subspaces is CLIQUE. CLIQUE is a method of clustering which works on density-and grid-based techniques. This method can discretize the data space into non-covering rectangular cells by partitioning each dimension into a fixed number of bins of equal dimension. A bin is dense if the fraction of total data points [6] contained in the bin is greater than a particular threshold. The method [4] finds solid cells in lower dimensional spaces and combines them to form clusters in higher dimensional spaces .The PROCLUS [7] and the ORCLUS [8] algorithms find projected clusters based on the representative cluster centres in a set of cluster dimensions. Another clustering approach, Fascicles [9], finds subgroups of data that share similar values in a subset of dimensions.

### *2.1.2 Distance Measures*

The most common distance measures or similarity measures for analysing gene expression data are the Euclidean distance and Pearson correlation coefficient, which are simple and easy in their implementation. But, in some situations, these both measures could be inappropriate to explore the true genetic relationship as Pearson correlation can be excessively sensitive to the outline topology of an expression curve and Euclidean distance cares only about the extent of changes. Euclidean distance or Euclidean metric is the "ordinary" distance between two points that one would measure with a ruler, and is given by the Pythagorean formula. The Pearson product-moment correlation coefficient is a measure of the direct connection or requirement between two variables X and Y, giving a value between +1 and −1 inclusive, where 1 is a total positive correlation, 0 is no correlation, and −1 is a total negative correlation for the values available [10].

### *2.1.3 Previous Work*

Clustering has been widely applied in bioinformatics in the past years. Hierarchical clustering techniques can be used for representing protein sequence family relationships. Eisen et al. [11] applied a variant of the hierarchical average-linkage clustering algorithm to identify groups of co-regulated genome-wide expression Patterns. Lowenstein et al. [12] applied agglomerative clustering to protein sequences that automatically builds a comprehensive evolutionary driven hierarchy of proteins from sequence alone. Frey et al. [13] applied an affinity propagation clustering technique to detect putative exons comprising genes from mouse chromosomes. They all claim lower computational cost in comparison to other algorithms, but they do not include the cost of pairwise similarity computations. Since this is the most expensive stage in large scale problems [3], the claimed advantage is exaggerated .Thus the focus is on reducing the computational complexity arising from this cost. Clustering can be said as one of the initial steps in data analysis of high-throughput expression measurements.

In most situations, microarray technology depends on the hybridization characteristics of nucleic acids to monitor DNA or RNA richness on a genomic scale. These have transformed the study of genome by allowing scholars and researchers to study the expression of thousands of genes concurrently for the first time. Thus the ability to simultaneously study thousands of genes under a host of differing conditions presents a huge challenge in the arena of computational science and data mining.

A wide range of data mining techniques like hierarchical clustering, K-means, Self-Organizing Maps have been implemented and used successfully on the analysis of high-dimensional gene expression data. Since gene expression profiles are programmed in real sequences, these algorithms may be intended to group gene expression vectors that are sufficiently close to each other. This may be done based on which certain distance or similarity measurement [14] that is applied. Even then, most algorithms suffer from several prominent shortcomings. Algorithms such as K-means and Self-Organizing Maps need the predefinition of the number of clusters which is usually unknown in advance. Hierarchical clustering [15] suffers from absence of robustness, no uniqueness. In case of large data set, it is also not practical to do clustering. Also the idea of forcing each gene into a specific cluster appears as a significant problem of these implementations.

## 2.2 Discovering the Distinct Patterns in Expression Profiles of Genes

The traditional analysis method of gene expression profiles uses clustering as a technique which finds groups of expressed genes that possess similar expression patterns [15] in their dataset. However, for large datasets, the clustering approach is time consuming and appears to be very difficult. Thus the idea of discovering the patterns that are distinct in gene expression profiles can be taken into account in such situations. Since the patterns shown by the gene expressions disclose their normalised mechanisms, it is important to find all different patterns existing in the dataset when there is a little knowledge available about the data. This approach [15] is iteratively done by selecting out pairs of gene expression patterns which have the largest dissimilarities in their expressions. This method can also be used as the pre-processing step for the initialization of the centres for clustering methods, like K-means or any other clustering techniques.

The concept of discovering the distinct patterns helps to discover the typical gene expression patterns which may show different variations depending on their underlying properties. They may represent the whole gene expression profiles of the data. Since genes show different patterns in the biological process, [16] the patterns they show are important for the analysis of their biological functions. There may be the many genes present that may share similar expression patterns also. Whatsoever, the number of different patterns and similar patterns is completely unidentified in prior. If there is the discovery of the different patterns existing in the dataset it will become much easier to find groups of genes with similar patterns also. These patterns shared by a group of genes are important [15] for the analysis of their biological functions .So instead of grouping genes into clusters with similar patterns, finding the collection of distinct gene expression patterns which are typical in the whole data set of genes would be more appropriate. And the distinct patterns can be regarded as to represent other genes with similar expression pattern. In another context, it is also called as prototype finding. And the distinct patterns found can be further used for constructing gene regulatory network which would be more useful in the analysis of the genes.

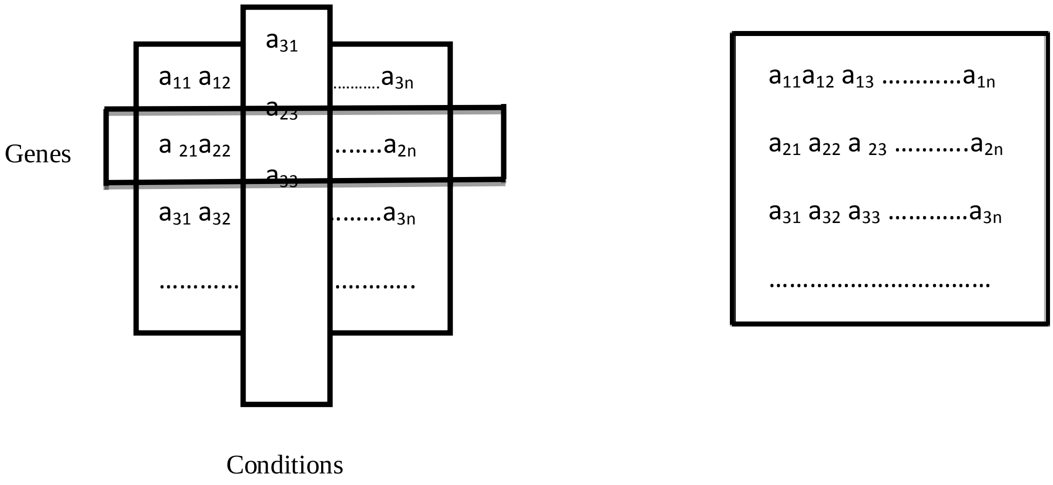
For this, a new clustering approach, named as the Pattern Similarity Clustering [4], can be used for capturing not only the nearness of objects but also the similarity of the patterns exhibited by the objects in the data. This

Cluster model can be told as a simplification of subspace clustering. But, it finds a much wider range of applications, including in DNA array analysis and in collaborative filtering, where pattern similarities among a set of objects carry very important meanings. Thus an effective depth-first algorithm needs to be used to mine the clusters. When comparing with the other cluster methods, this approach mines multiple clusters simultaneously, also detects overlapping clusters, and is resilient to outliers. It is also deterministic in discovering all clusters that are qualified, while the other approaches provide only an approximate answer.

### *2.2.1 Gene expression profiles*

The term ‘Gene expression profiling ‘can be defined as the measurement of the activity or the expression of thousands of genes at once, which will in order can create a global picture of the cellular function of that genes. Thus these profiles can, discriminate between the cells that are actively dividing, or show how the cells react to a particular treatment or behaviour. Expression profiling is a rational step after the sequencing a genome the order tells us what the cell could perhaps do, while the expression profile tells us what it is really doing at a point in time. It can be good to identify meaningfully regulated genes first and then find patterns by associating the list of significant genes to sets of genes known to share certain associations. For a given type of cell, the group of genes whose combined expression pattern is exceptionally characteristic to a given condition establishes the gene signature of this condition. Thus gene signature can be used to select a group of patients at a specific state of a disease with accurateness that enables the selection of treatments ideally. Gene Set Enrichment Analysis (GSEA)[17],[18] and other methods take benefit of this kind of logic but uses more sophisticated statistics, because the constituent genes in real processes shows more complex behaviour.

Gene expression profiles could be represented by an n m numerical data matrix G = (g (i, j)).Each row of G stands for a gene and each column stands for a condition like different time points or samples. Element g (i j) stands for the expression of gene i under condition j. There are usually thousands of genes compared to dozens of conditions. It is believed that genes with related functions tend to have similar expression patterns. Similarities of gene expression patterns could be represented by a n, n matrix A = a (i,j) as Fig.2.2.1 shows. We call it the similarity matrix as it shows the similarity of gene expression profiles. And a (i,j) stands for the similarity between gene i and gene j.



**(a)**

Original gene expression profiles
where each row stands for a gene
and each column stands for condition

**(b)**

Similarity matrix for the genes

Fig 2.2.1 Gene Expression Profiles

### *2.2.2 Discovering Distinct Patterns by considering the Largest Dissimilarities*

The pair of genes which have the lowest similarities among all the genes is found. They can be taken as the distinct patterns and delete the genes which have high similarities with them. There are two thresholds, $\Theta_1$ and $\Theta_2$ to be defined. Only when the similarity between two genes is smaller than $\Theta_1$ the two genes can be regarded as distinct patterns. And only when two genes have similarity higher than $\Theta_2$ we regard them as similar patterns. The following are the main steps [15].

Step 1. Compute the similarity matrix A for all genes.

Step 2. Pick up the minimum item a (i,j) from the similarity matrix A. If a (i, j) < $\Theta_1$, add gene i and j to the collection of distinct patterns. If a (i, j) > $\Theta_1$, stop the algorithm and output the distinct patterns.

Step 3. Update the similarity matrix as following, If a (i,p )> $\Theta_2$ for any a(i,p)€ A, mark gene p as neighbour of gene i. If a(j,q) > $\Theta_2$ for any a(j,q)€ A, mark gene q as neighbour of gene j. Update matrix A by deleting all rows and columns corresponding to the neighbours of gene i and gene j. Then go back to step 2 if A is not empty.

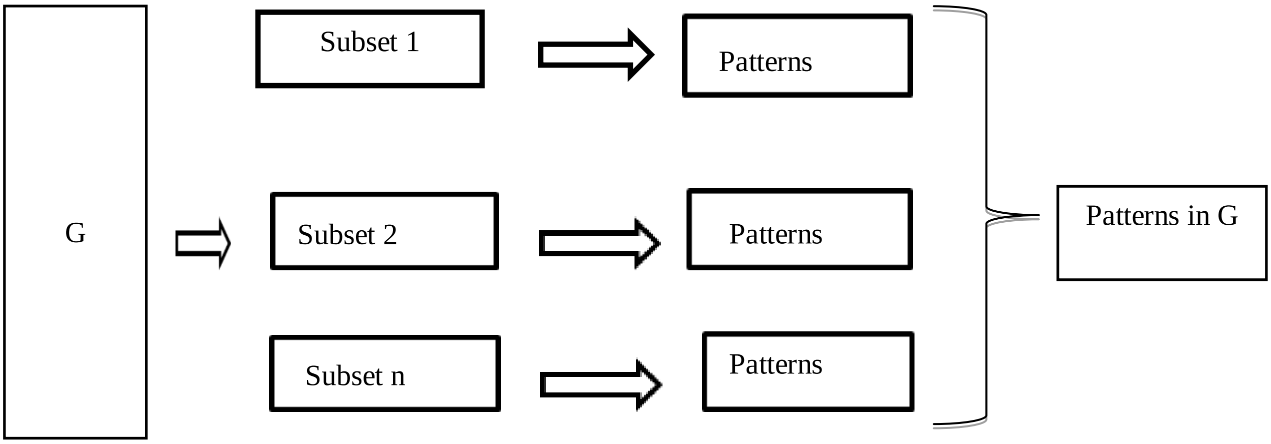


Fig.2.2.2 Handling Large Datasets by Divide and Conquer

## 3. Challenges

There are many challenges for the implementation of pattern similarity approach. First, identifying subspace clusters in high-dimensional data sets is a difficult task because of the problem caused by dimensionality. Real life data sets in DNA array study or collaborative filtering can have hundreds of attributes. Secondly, a new similarity model is needed as traditional distance functions may not be able to capture the pattern similarity [4] among the objects. For in some case, objects may not be close if measured by distance functions such as Euclidean, Manhattan, or Cosine. The task is different from pattern discovery in time series data and need not to be confused with pattern discovery in time series data, such as trending analysis in stock closing prices.

The biggest challenge in the task is in subspace clustering objects can contribute towards the cluster in any subcategory of the data columns and the number of data columns in actual life applications, such as DNA array analysis, in which data is large which is usually in the hundreds or even thousands.

Building a “depth first” clustering algorithm can be pointed as another challenge [4]. When many clustering algorithms finds clusters in lower dimensions initially and then merge them to develop clusters in advanced dimensions, the proposed approach first generate clusters in the highest dimensions, and then find low dimensional clusters which are not previously covered[15] by the high dimensional clusters .This approach is useful because  the clusters that span a large number of columns are usually more of interest, and it avoids forming clusters which are part of other clusters. The approach is also more effectual because the grouping of low dimensional clusters to form high dimensional ones is usually very expensive.

## 4. Conclusion

The similarity cluster model can be effective as it finds a large variety of applications including management of scientific biological data, such as in the DNA microarray, and in various other fields. In the datasets which are under exploration, the distance among the objects may not be close in any subspace, but they can still be understandable in being unstable or scaling patterns, which are not captured by traditional subspace clustering algorithms. The gene expression patterns can be frequently of great interest in DNA microarray analysis. However, the expression levels of two genes may rise and fall synchronously in response to a set of environmental stimuli which are under DNA microarray technology. The magnitude of their expression levels may not be close every time, but they may be capable of exhibiting the patterns that are very much alike. Discovering such clusters of genes is very important and essential in as it can be able to reveal the significant connections in gene regulatory networks. For large scale gene expression data with little advance knowledge; it is difficult and could be aimless on beginning any further analysis. Usually there will be a lot of genes sharing similar expression patterns. So reducing the number of significant genes would be great help for constructing gene regulatory networks. Also, very large dataset can be handled through a divide and conquer scheme by which we apply this technique on each partition of the original dataset respectively. The effectiveness of the algorithm is tested and the results are evaluated.

## References

[1]. B. Shekar, M. Narasimha Murty and G. Krishna, Pattern clustering: an artificial intelligence approach.
[2]. Xiangsheng Chen, Jiuyong Li1, Grant Daggard, and Xiaodi Huang, Finding Similar Patterns in Microarray Data.
[3]. Bassam Farran, Amirthalingam Ramanan, and Mahesan Niranjan, Sequential Hierarchical Pattern Clustering, School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, United Kingdom.
[4]. Haixun Wang Wei Wang Jiong Yang Philip S. Yu, Clustering by Pattern Similarity in Large Data Sets ,IBM T. J. Watson Research Center,30 Saw Mill River Road,Hawthorne, NY 10532.
[5] K.Fukunaga, IntroductiontoStatisticalPatternRecognition, Academic Press, 1990.
[6] Jiong Yang, Wei Wang, Haixun Wang, and Philip Yu, δ-clusters: Capturing Subspace Correlation in a Large Data Set,IBM T. J. Watson Research Centerfjiyang, ww1, haixun, psyug@us.ibm.com
[7]. M. Ester, H. Kriegel, J. Sander, and X. Xu., A density-based algorithm for discovering clusters in large spatial databases with noise,. In SIGKDD, pages 226–231, 1996
[8]. D. H. Fisher, Knowledge acquisition via incremental conceptual clustering, in Machine Learning, 1987.
[9]. H. V. Jagadish, J. Madar, and R. Ng., Semantic compression and pattern extraction with fascicles,. In VLDB, pages 186–196, 1999.
[10]http://en.wikipedia.org/wiki/Euclidean_distance,http://en.wikipedia.org/wiki/Pearson_product,moment_correlation_coefficient.
[11] Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D, Cluster analysis and display of genome-wide expression patterns, Proceedings of the National Academy of Sciences USA 95(25), 14863–14868 (1998)
[12] Loewenstein, Y, Portugaly, E., Fromer, M., Linial, Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space. Bioinformatics 24(13), 41–49 (2008).
[13] Frey, B.J., Dueck, Clustering by Passing Messages between Data Point, Science AAAS 315, 972–976 (2007)
[14].Miao Li, Biao Wang, Zohreh Momeni, and Faramarz Valafar, Pattern Recognition Techniques in Microarray Data Analysis ,Department of Computer Science,San Diego State University,San Diego, California, USA,
[15] Li Teng, and Laiwan Chan, Discovering Distinct Patterns in Gene Expression Profile, Department of Computer Science and Engineering The Chinese University of Hong Kong, Hong, Kong.
[16] J Yang, W Wang, H Wang, P S Yu, Clusters: Capturing Subspace Correlation (2002).
[17]http://en.wikipedia.org/wiki/Gene_expression_profiling#cite_note-6.
[18] http://www.wikidoc.org/index.php/Expression_profiling.
[19] C. H. Cheng, A. W. Fu, and Y. Zhang,Entropy-based subspace clustering for mining numerical data ,In SIGKDD, pages 84–93, 1999.)

### A Brief Author Biography

***Irene Maria*** –Irene Maria is currently pursuing her M.Tech in Computer Science and Engineering from the Department of Computer Science in Karunya University, Tamilnadu, India. She received her Bachelor's degree from Anna University, Tirunelveli in Computer Science and Engineering.

***Mathew Kurian*** – Mathew Kurian, finished his M.E in computer science and engineering from Jadavpur University, Kolkata and currently he is working as Assistant Professor in Department of Computer Science and Engineering in Karunya University. Previously, he worked as Software Engineer with Aricent Technologies.He is currently doing his PhD Degree in Data Mining.