# Optimization of Clustering Algorithms for Gene Expression Data Analysis using Distance Measures

Angela Makolo
Computer Science Department,
University of Ibadan,
Ibadan, Nigeria

Taiwo Adigun
Computer and Information Sciences Department,
Covenant University
Ota, Nigeria

## ABSTRACT

Clustering is one of the fundamental processes of analyzing gene expression data, basically by comparing gene expression profiles or sample expression profiles. Comparing expression profiles requires a measure apart from the actual clustering algorithm to quantify how similar or dissimilar the objects under consideration are. Various clustering algorithms have been used to analyze gene expression data. Some of these algorithms reported the incorporation of similarity measures like Euclidean Distance, Pearson Correlation and mutual information for their performance. This work considered different reported clustering algorithms for gene expression data analyses and the importance of different similarity measures for optimizing these clustering algorithms. To this end, no clustering technique in all the works investigated has been applied directly on gene expression data. It is observed that the output (distance matrix) of similarity or dissimilarity measures plays the role of input to clustering techniques, and those that did not use any of the popular proximity measures applied one or two approaches such as Constrained Coherency (CoCo), Silhouette coefficient measurement, and normalization and discretization, to refine gene expression data for improved cluster quality by speeding up the learning phase, reduction of computational space and handling of noise effectively.

## Keywords

Clustering Algorithms, Proximity Measure, Gene Expression Data, Distance Matrix, Microarray.

## 1. INTRODUCTION

Analysis of gene expression data can be classified into either a supervised or unsupervised machine learning process. The supervised process involves classification or prediction of the data from a known subset of gene expression data used as the training data [1], while the unsupervised approach groups similar genes into the same cluster based on proximity measure. Clustering approach is a powerful model that detects patterns or relationships in expression data.

The emergence of high throughput genomic data via technologies like microarray and next generation sequencing has made it possible to generate the expression levels of thousands of gene simultaneously under various experimental conditions [2],[25]. This has caused new problems to be raised, which calls for the need for large scale pre- and post-processing data analysis, and need for a coherent data management. Microarray technology has been mostly used on several occasions for analyses of gene expression data [15],[18],[24]. The goal is to analyze the data to determine the patterns among the genes, which leads to a better understanding of processes in the cell and therefore represents a significant step towards modeling of cell behaviours. Gene expression data are usually represented by a matrix called **expression matrix,** which could also be annotated**.** The columns represent experimental conditions or time points and these serve as the features of the dataset. Rows of the expression matrix represent the genes under consideration across all experimental conditions/time points. In the matrix, element $G_{ij}$ represents expression level of gene *i* under experimental condition *j*, whereas the expression levels of a gene across different experimental conditions are called **gene expression profile** while **sample expression profile** refers to the expression levels of all genes under an experimental condition.

Clustering techniques are explicitly or implicitly based on quantitative measures of dissimilarity between the objects of interest, and in gene expression analysis, the key concept is to compare gene expression in two or more cell/tissue types where the gene expression are assessed by measuring the number of RNA transcripts in a cell/tissue sample. Several clustering algorithms have been applied successfully to analyze gene expression data [4]-[7],[9],[13],[15] and several others have been reported to have incorporated different similarity/distance measures to optimize the analyses of the gene expression data [1],[3],[8],[12],[14],[16]-[22],[24]. Applying clustering approach in analyzing gene expression data often involve calculation of distances or similarities among the objects of the expression profiles [1], [11]. And in the case of selection of a clustering algorithm itself, choosing the right distance to be employed between the expression elements is probably one of the most difficult questions.

The motivation of this investigation is to produce clarifications to some pertinent issues of distance measures relating to the analyses of gene expression data. The first is to known if the output of distance measurement on expression data gives a meaningful result. And if it does, can the process itself be regarded as machine learning or a data mining technique? Besides, by blending distance measures with clustering algorithms, at what point does the distance measure come into play? , Does the distance measure determine the clustering algorithm to be used or the clustering algorithms determine the distance measure to be used? Also, is it not impractical to analyze gene expression data using clustering technique without any distance measure, if it is, what then is /are the effects of distance measures on clustering algorithms?

The rest of this paper is organized as follows: in section 2, we present and describe the basic concepts like the gene expression data, and distance measures. Review of clustering algorithms on expression data is presented in section 3. Section 4 gives the conclusion and suggests future research directions of this concept.

# 2. BACKGROUND

## 2.1 Gene Expression Data

Gene expression data provides the whole transcriptome analysis where thousands of genes are studied at the same time. Microarray and RNA sequencing are the two methods for a large-scale gene expression profiling currently in use. Microarray and RNA-seq are different technologies [25] but both can monitor expression levels of thousands of gene simultaneously. Our focus is on microarray source because quite a number of published papers have referred more to microarray data than RNA-seq data [4,5,8]. Microarray also known as DNA microarray or DNA chips is a technology that is used to measure the level of mRNA in a particular cell or tissue for many genes at once. The goal of many microarray experiments is to identify genes that are differentially transcribed with respect to different biological conditions of cell cultures or tissue sample. Obtaining gene expression data from microarray experiment involves 3 major processes known as image processing, transformation and normalization. Following the processes performed during microarray experiments, there are 3 kinds of data that are generated, which are curated in different online databases like Gene Expression Ominibus.

i. Image Data: The scanned image of the microarray chip.
ii. Expression Data: the normalized version of image scanned. It is given as a sequence of numbers in an n x m matrix that represents the expression of gene for a set of samples.
iii. Annotation Data: the additional information (metadata) that are appended to the expression data. It consists of textual descriptors that help to interpret the detected gene expression levels like functions of the genes and details of the samples (i.e disease state or normal state).

The expression data are usually presented in an expression matrix. Each column represents all the gene expression levels from a single experiment, and each row represents the expression of a gene across all experiments, samples or time points. Each element is a log ratio that is defined as log2(T/R), where T is the gene expression level in the testing sample and R is the gene expression level in the reference sample.

## 2.2 Distance Measures

In a multivariate analysis, calculation of similarities or dissimilarities among a set of items is the fundamental approach of having an optimal result. Distance is also termed dissimilarity. Although there are important differences between similarities and dissimilarities, the two sets of measures are sometimes referred to as distances. Distances show a measure of dependencies between two random variables and a small distance is equivalent to a large similarity. The function (metric) of similarities or dissimilarity satisfies 3 major attributes. For distance function, d for all sequences x and y:

i. Non-negativity: d(x,y) ≥ 0
ii. Symmetry: d(x,y) = d(y,x)
iii. Reflexivity: d(x,y) = 0, if and only if x = y

For similarity function, s for all sequences of x and y

i. Non-negativity: s(x,y) ≥ 0
ii. Symmetry: s(x,y) = s(y,x)
iii. Reflexivity: s(x,y) = 0, if and only if x = y
iv. S(x,y) increases linearly as x and y are more and more similar.

A distance between items in a multi-dimensional space is interpreted as measuring distance between two probability distributions, where the input is in form of a high dimensional data matrix. Gene expression data is one of high dimensional data matrices and a filter is applied to depopulate certain areas of the space before clusters are sought [11]. There are two different methods for quantifying the similarity and dissimilarity of gene expression profiles, the analysis of microarray data is either to find the similarities or dissimilarities of genes, or to find the similarities or dissimilarities of samples. The purpose of a measure of similarity or dissimilarity is to compare two lists of numbers (i.e vectors) and computes a single number which evaluates their similarity. An important basis for classification of distance measures is the standardization effect, which determines the distance measure to be used considering whether the sample data are from the same scale or not. There are distance measures that are only appropriate for data measured on the same scale where no adjustment is made for differences or variations in the samples (i.e. Euclidean Distance). Some others are used to standardize and compute similarities of the data to take care of differences in the rank ordering.

There are many distance measures, a selection of the most commonly used and most popular measures especially for gene expression data are described below.

### *2.2.1 Euclidean Distance*

The Euclidean distance or Euclidean metric is the "ordinary" (i.e. straight-line) distance between two points in Euclidean space. The basis of many measures of similarity and dissimilarity is Euclidean distance. The distance between vectors X and Y is defined as shown in equation 1.

$$d(x,y) = \sqrt{\sum_i^n (x_i - y_i)^2} \qquad (1)$$

Euclidean distance is only appropriate for data measured on the same scale. Euclidean distance is most often used to compare profiles of respondents across variables. For example, suppose our data consist of demographic information on a sample of individuals, arranged as a respondent-by-variable matrix. Each row of the matrix is a vector of m numbers, where m is the number of variables. We can evaluate the similarity (or, in this case, the distance) between any pair of rows.

**Table 1: Sample Gene Expression matrix. Each value is the expression level of the genes in different samples.**

| | **Exp *1*** | **Exp *2*** | **Exp *3*** | **---** | **Exp *n*** |
|---|---|---|---|---|---|
| **Gene *1*** | -1.28 | 0.77 | 3.42 | --- | -2.60 |
| **Gene *2*** | 0.12 | 1.00 | 4.01 | --- | 1.06 |
| **Gene *3*** | -2.07 | 1.92 | 3.00 | --- | -3.04 |
| **Gene *4*** | 1.77 | -2.67 | 2.97 | --- | 2.01 |
| -<br>-<br>- | -<br>-<br>- | -<br>-<br>- | -<br>-<br>- | -<br>-<br>- | -<br>-<br>- |
| **Gene *n*** | 2.01 | -1.97 | 1.86 | --- | 3.67 |

### *2.2.2 Pearson Correlation*

It is a measure of the linear correlation between two variables X and Y, giving a value between +1 and −1 inclusive, where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation. It is widely used in the sciences as a measure of the degree of linear dependence between two variables. Equation 2 shows the Pearson correlation coefficient equation.

$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (2)$$

### *2.2.3 Mutual Information*

The mutual information (MI) of two random variables is a measure of the variables' mutual dependence. Formally, the mutual information of two discrete random variables X and Y can be defined as in equation 3.

$$MI(U,V) = \sum_{i=1}^{R}\sum_{j=1}^{C} P(i,j)\log\frac{P(i,j)}{P(i)P(j)} \qquad (3)$$

## 2.3 Distance Matrix

Please use a 9-point Times Roman font, or other Roman font with serifs, as close as possible in appearance to Times Roman in which these guidelines have been set. The goal is to have a 9-point text, as you see here. Please use sans-serif or non-proportional fonts only for special purposes, such as distinguishing source code text. If Times Roman is not available, try the font named Computer Modern Roman. On a Macintosh, use the font named Times. Right margins should be justified, not ragged.

**Table 2: Format of a distance matrix. The element at the ith row and jth column is the distance between the ith and jth genes**

| | **Gene *1*** | **Gene *2*** | **Gene *3*** | **Gene *4*** | --- | **Gene *n*** |
|---|---|---|---|---|---|---|
| **Gene *1*** | $d_{11}$ | $d_{12}$ | $d_{13}$ | $d_{14}$ | --- | $d_{1n}$ |
| **Gene *2*** | $d_{21}$ | $d_{22}$ | $d_{23}$ | $d_{24}$ | --- | $d_{2n}$ |
| **Gene *3*** | $d_{31}$ | $d_{32}$ | $d_{33}$ | $d_{34}$ | --- | $d_{3n}$ |
| **Gene *4*** | $d_{41}$ | $d_{42}$ | $d_{43}$ | $d_{44}$ | --- | $d_{4n}$ |
| -<br>-<br>- | -<br>-<br>- | -<br>-<br>- | -<br>-<br>- | -<br>-<br>- | -<br>-<br>- | -<br>-<br>- |
| **Gene *n*** | $d_{n1}$ | $d_{n2}$ | $d_{n3}$ | $d_{n4}$ | --- | $d_{nn}$ |

# 3. PERFORMANCE ANALYSIS

The purpose of this investigation is to analyze and gather relevant information from the performances of different clustering algorithms that have been applied on gene expression dataset based on the use of proximity measures. The metrics are used to reduce the computational space and to identify local similar regions in gene expression profiles [2]. Quite a number of these algorithms included proximity measures in their approaches while few others are free from the use of proximity measures [9]. : In section 3.1, we first considered the clustering algorithms for analysis of gene expression data without any peculiar similarity measure and how they overcome the challenges of reducing computational space and finding local similar region in gene expression profiles. Later, we investigate the clustering algorithms for analysis of gene expression data that are reported with distance measures.

## 3.1 Clustering Algorithms for Gene Expression Data

The peculiarity and format of gene expression data highly require a distance or similarity measure to be applied before attempt is made to find the clusters among the genes. However, some reported clustering algorithms are devoid of these popular proximity measures and have been used to analyze gene expression dataset successfully. In [9], a clustering technique called GenClus was developed to analyze an incremental gene expression dataset. It does not use any proximity measure during the gene clustering but based on density-based approach. Instead of any similarity or distance measure, Genclus uses two steps called normalization and discretization. The gene expression is normalized to have mean 0 and standard deviation 1 (regulation information), discretization is then performed on the normalized data where clustering is thereafter run on discretized data. The same approach is used by Chandrasekhar et al. in [5] so as to overcome the challenge of computational space. However, density-based clustering techniques suffer from high computational complexity with increase in dimensionality and input parameter dependency [9].

A Constrained Coherency (CoCo), a data-driven approach was employed in [4] to serve as a similarity measurement before the clustering is performed. CoCo measures the pairwise relationship between genes via their decomposed components. EAGMFI algorithm developed in [7] employed Silhouette coefficient measurement for automatic evaluation of initial seed of centroids to depopulate the computational space before the actual clusters are found. Das et al. in [13] presented two clustering methods called Density-Based Approach (DGC) and the second is Frequent Itemset Mining Approach (FINN), but both methods use a novel dissimilarity measure. The proximity between any two genes gi and gj is given by a function defined as D(gi, gj ) where D is any proximity measure like Euclidean distance, Pearson's correlation, etc [13]. Moreover, in place of any popular proximity measure, Hestilow and Huang in [15] developed a clustering technique called Variational Bayes Expectation Maximization (VBEM) algorithm using a time-difference expression. The method consists of three steps; the first two steps address the challenge of high computational space of gene expression data where the last step clusters the data from the first two steps. The first step rescales the expression data, the second step captures the signal shape information by calculating the first-order time difference.

## 3.2 Clustering Algorithms for Gene Expression Data with Distance Measures

Bryan in [11] investigated the problems in gene clustering based on gene expression data and pointed out that genome-wide collections of expression trajectories often lack natural clustering structure, prior to ad hoc gene filtering. The filtering is basically to depopulate certain areas of the space before clusters are sought, and similarity measures are suggested to perform the filtering. Both correlation coefficient and Euclidean distance were computed from repeated measurements to produce pairwise similarities to improve the clustering of gene expression data in [14]. Clustering array data with repeated measurements with improve cluster quality was made possible because of variability-weighted approach applied on proximity measures. A new clustering method called CLARITY (Clustering with Local shApe-based similaRITY) for the analysis of microarray time course experiments was developed in [16]. The similarity measure is a local shape-based similarity measure defined by the Spearman rank correlation (SRC).

Metabolic networks were constructed by Hanisch in [17] using a graph distance function which combines information from expression data and biological networks, and the gene expression measurements was a correlation-based distance function. In [19], three different clustering techniques were compared against three different proximity measures on Comparative Genomic Hybridization (CGH) dataset of a population of cancer patient samples. The three pairwise distance/similarity measures are raw, cosine and sim, while the three clustering algorithms are bottom-up, top-down and k-means. The distance-based methods are novel but effectively exploit correlations between consecutive genomic intervals. Raw distance compares the value or status of each genomic interval separately; Sim distance merges contiguous aberrations of the same type into segments and counts the number of common segments between the given two samples, while the third measure, segment-based cosine similarity maps segments to vectors in a high dimensional space. It computes the distance between two vectors as the cosine of the angle between them [19]. A dimensionality reduction approach called Locality Preserving Projection (LPP) was proposed by Salome and Suresh in [24]. The LPP procedure for dimensionality reduction consists of three steps, namely, (1) generation of Distance matrix based on the Euclidean distance (2) determining adjacency matrix and (3) Calculating dimensionality reduced matrix [24]. Clustering of the data resulted from the previous step is done by Fuzzy C-Means (FCM) and later compared with k-means technique.

Mutual Information (MI) measure was compared with the well-known Euclidean distance and Pearson correlation coefficient in [20], while EM (Expectation Maximization) algorithm which provides the statistical frame work to model the cluster structure of gene expression data in [8]. The correlation coefficient was used to compute pairwise similarities of genes of different dataset in a systematic framework for assessing the results of clustering algorithms developed in [12].

**Table 3: Clustering Algorithms for Gene Expression Data with and without Distance Measures**

| S/N | Clustering Model | Popular Similarity Measure | Non- popular Similarity Measure | Similarity Measure Approach Used |
|---|---|---|---|---|
| 1 | GenClus [9] | No | Yes | Normalization and Discretization. |
| 2 | K-Means algorithm<br>hybridised with Cluster Centre Initialization Algorithm (CCIA) [5] | No | Yes | Normalization and Discretization |
| 3 | Dynamic Clustering [4] | No | Yes | Constrained Coherency (CoCo) function |
| 4 | Enhanced Automatic Generations<br>of Merge Factor for ISODATA (EAGMFI) algorithm [7] | No | Yes | Silhouette coefficient measurement |
| 5 | Density-Based Approach (DGC) and Frequent Itemset Mining Approach (FINN) [13] | No | Yes | Novel approach,<br>$D(g_i, g_j)$ function |
| 6 | Variational Bayes Expectation Maximization (VBEM) algorithm [15] | No | Yes | Similarity measures incorporated in the model |
| 7 | Combinatorial<br>Clustering Algorithm [11] | Yes | No | Filtering |
| 8 | Empirical Analysis of Clustering Algorithms [14] | Yes | No | Correlation coefficient and Euclidean distance |
| 9 | Local shApe-based similaRITY (CLARITY) algorithm [16] | Yes | No | Spearman rank correlation (SRC). |
| 10 | A Graph Distance Function with Hierarchical Clustering [17] | Yes | No | Correlation-based distance function |
| 11 | Bottom-Up, Top-Down And K-Means clustering algorithms [19] | No | Yes | A novel pairwise distance/similarity measures are raw, cosine and sim |
| 12 | Fuzzy C-Means (FCM) algorithm [24] | Yes | No | Locality Preserving Projection (LPP) with Euclidean distance |
| 13 | Empirical Analysis of Clustering Algorithms [20] | Yes | No | Mutual Information (MI) |

# 4. CONCLUSION AND SUGGESTIONS

It is observed that computing distance of gene expression data is a major and important preprocessing step before clustering, since it affects clustering results by speeding up the learning phase [2],[5]. Although, the level of performance of clustering techniques differ, the level of refinement and speed of proximity measure for dimensionality or computational space reduction used ultimately determines the overall performance of the clustering technique employed.

The Basic Similarity Measures are commonly used in gene expression analysis include the Euclidean distance and the Pearson correlation [16]. The output of the similarity or dissimilarity measures serves as the input to any clustering technique intended to be used. Therefore, attention should be paid to the selection of a proper distance measure for analyzing the clustering of gene expression data [20].

Though, some clustering techniques for analyzing gene expression data did not use the popular distance measures, most of the approaches used in place of normal proximity measures handle noise effectively. In overall interest of having better clusters from the analysis of gene expression data, an approach for the raw data refinement must be put in place and the result of the refinement itself cannot be considered meaningful for the analysis, therefore, the proximity measure approaches cannot be considered as any data mining or machine learning techniques.

## 5. REFERENCES

[1]. Brian T. (2006).An approach for clustering gene expression data with error information. BMC Bioinformatics, 7:17, doi:10.1186/1471-2105-7-17.

[2]. Pirim, H., Ekşioğlu, B., Perkins, A. and Yüceer, C. (2012) Clustering of High Throughput Gene Expression Data. Comput Oper Res., 39(12): 3046–3061. doi:10.1016/j.cor.2012.03.008.

[3]. Ernst, J., Nau, G.J.and Bar-Joseph, Z. (2005) Clustering short time series gene expression data. BIOINFORMATICS, Vol. 21 Suppl. 1, pages i159–i168, doi:10.1093/bioinformatics/bti1022

[4]. An, L. and Doerge, R. W. (2012) Dynamic Clustering of Gene Expression. International Scholarly Research Network ISRN Bioinformatics, Volume 2012, Article ID 537217, doi:10.5402/2012/537217

[5]. Chandrasekhar, T., Thangavel, K. and Elayaraja, E. (2011) Effective Clustering Algorithms for Gene Expression Data. International Journal of Computer Applications (0975 – 8887), Volume 32– No.4.

[6]. Sturn, A., Quackenbush, J. and Trajanoski, Z. () Genesis: cluster analysis of microarray data. BIOINFORMATICS APPLICATIONS NOTE, Vol. 18 no. 1, Pages 207–208.

[7]. Chandrasekhar, T., Thangavel, K. and Elayaraja, E. (2011) Performance Analysis of Enhanced Clustering Algorithm for Gene Expression Data. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, ISSN (Online): 1694-0814.

[8]. Valarmathie, P., Srinath, M.V., Ravichandran, T. and Dinakaran, K (2009).Hybrid Fuzzy C-Means Clustering Technique for Gene Expression Data. International Journal of Research and Reviews in Applied Sciences, ISSN: 2076-734X, EISSN: 2076-7366, Volume 1, Issue 1.

[9]. Sarmah, S. and Bhattacharyya, D.K. (2010) An Effective Technique for Clustering Incremental Gene Expression data. IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 3.

[10]. Yeung, K. Y. and Ruzzo, W. L. (2001) Principal component analysis for clustering gene expression data. BIOINFORMATICS, Vol. 17 no. 9, Pages 763–774.

[11]. Bryan, J. (2004) Problems in gene clustering based on gene expression data. Journal of Multivariate Analysis 90, 44–66.

[12]. Yeung, K. Y., Haynor, D. R. and Ruzzo, W. L. (2001) Validating clustering for gene expression data. BIOINFORMATICS, Vol. 17 no. 4, Pages 309–318.

[13]. Das, R., Bhattacharyya, D. K. and Kalita, J. K. (2010) CLUSTERING GENE EXPRESSION DATA USING AN EFFECTIVE DISSIMILARITY MEASURE. International Journal of Computational Bioscience, Vol. 1, No. 1.

[14]. Yeung, K.Y., Medvedovic, M. and Bumgarner, R.E. (2003) Clustering gene-expression data with repeated measurements. Genome Biology, Volume 4, Issue 5, Article R34.

[15]. Hestilow, T.J. and Huang, Y. (2009) Clustering of Gene Expression Data Based on Shape Similarity. EURASIP Journal on Bioinformatics and Systems Biology, Volume 2009, Article ID 195712, doi:10.1155/2009/195712.

[16]. Balasubramaniyan, R., Hullermeier, E., Weskamp, N. and Kamper, J. (2004) Clustering of Gene Expression Data Using a Local Shape-Based Similarity Measure. Bioinformatics © Oxford University Press.

[17]. Hanisch, D., Zien, A., Zimmer, R. and Lengauer, T. (2002) Co-clustering of biological networks and gene expression data. BIOINFORMATICS, Vol. 18 Suppl. 1, Pages S145–S154.

[18]. Souto, M., Costa, I., Araujo, D., Ludermir, T. and Schliep, A. (2008) Clustering cancer gene expression data: a comparative study. BMC Bioinformatics, 9:497, doi:10.1186/1471-2105-9-497.

[19]. Liu, J., Mohammed, J., Carter, J., Ranka, S., Kahveci., T. and Baudis, M. (2006) Distance-based clustering of CGH data. BIOINFORMATICS, Vol. 22 no. 16, pages 1971–1978, doi:10.1093/bioinformatics/btl185.

[20]. Priness, I., Maimon, O and Ben-Gal, I. (2007) Evaluation of gene-expression clustering via mutual information distance measure. BMC Bioinformatics, 8:111, doi:10.1186/1471-2105-8-111.

[21]. Jaskowiak, P., Campello, R. and Costa, I. (2014) On the selection of appropriate distances for gene expression data clustering. BMC Bioinformatics, 15(Suppl 2):S2.

[22]. Glazko, G. and Mushegian, A. (2010) Measuring gene expression divergence: the distance to keep. Biology Direct, 5:51.

[23]. Ray, S.S., Bandyopadhyay, S. and Pal, S.K. (2007) New Distance Measure for Microarray Gene Expressions using Linear Dynamic Range of Photo Multiplier Tube. Proceedings of the International Conference on Computing: Theory and Applications (ICCTA'07), 0-7695-2770-1/07.

[24]. Salome, J. and Suresh, R. M. (2012) Efficient Clustering for Gene Expression Data. International Journal of Computer Applications (0975 – 888), Volume 47– No.5.

[25]. Adigun, T., Makolo, A. and Fatumo, S. (2015) Input Dataset Survey of In-Silico Tools for Inference and Visualization of Gene Regulatory Networks (GRN). Computational Biology and Bioinformatics. Vol. 3, No. 6, pp. 81-87. doi: 10.11648/j.cbb.20150306.11.