Asian Journal of Applied Research (AJAR) 2015

E-ISSN: 2395-1702 P-ISSN: 2395-0382
Volume 01- Issue 01-, pp-01-10

Research Paper Open Access

SUPERVISED MULTI ATTRIBUTE GENE EXPRESSION DATA FOCUSING ON CANCER THERAPEUTICS

S.Gayathri [1], Ms.B.Rubadevi [2],
PG-Scholar,Computer Science and Engineering, Arunai Engineering College, Tiruvannamalai,
Assistant Professor,Faculty of Computer Science Department,Arunai Engineering College, Tiruvannamalai,

gayathrisivabalan92@gmail.com, ruparuba@gmail.com

***Abstract*—**Cancer research - One of the major research areas in the field of medical. Pointed out the exact tumor types provides an optimized solution for better treatment and toxicity minimization due to medicines to the patients. To get a clear picture on the insight of a problem, a clear cancer classification analysis system needs to be pictured followed by a systematic approach to analyze global gene expression which provides an optimized solution for the identified problem area. Molecular diagnostics provides a promising option of systematic human cancer categorization, but these tests are not widely functional because characteristic molecular markers for most solid tumor save yet to be identified. Recently, DNA microarray-based tumor gene expression profiles have been used for cancer diagnosis. Existing system focussed in ranging from old nearest neighbor analysis to support vector machine manipulation for the learning portion of the classification model. We don't have a clear picture of supervised classifier which can manage knowledge attributes coming two different knowledge streams. The input from multiple source, create an ontological store, cluster the data with attribute match association rule and followed by classification with the knowledge acquired.

***Keywords*—**Association rules, cancer, clustering, data mining, gene expression data, next generation sequencing.

1. Introduction

The prominence of DNA microarray technology [1] is the aptitude to be used to simultaneously monitor and study the expression levels of genes, the relationship between genes, their functions and classifying genes or samples that perform in a parallel or synchronized manner during imperative biological processes. Cancer is a major cause of all the natural mortalities and morbidities throughout the world. The deaths are caused due to cancer. Cancer is an abnormal and uncontrollable growth of cells in the body that turn malignant. Various types of cancer have been recognized, namely, breast cancer, lung cancer, brain cancer, cervical cancer, kidney cancer, liver cancer, Hodgkin's lymphoma, non-Hodgkin's lymphoma, ovarian cancer, skin cancer, thyroid cancer, uterine cancer, and testicular cancer.

Gene expression data, both at the transcript level and at the protein level, can be a precious tool in the understanding of genes, genetic networks, and cellular states. The eminence of DNA microarray technology [1] is the aptitude to be used to simultaneously monitor and study the expression levels of thousands of genes, the

relationship between genes, their functions and classifying genes or samples that perform in a parallel or synchronized manner during imperative biological processes.

DNA Methylation is the process of addition of a methyl group to the gene. A methyl group is fairly significant to organic chemistry and consists of one carbon atom bonded to three hydrogen atoms (CH3). DNA Methylation can modify the gene expression, diminishing it or making it gaudier. DNA methylation exhibits direct interception with the binding sites of particular transcription factors to their promoter. Also, they are involved with the direct binding of specific transcripts to the methylated DNA.

1.1The Human Epigenome Project

The Human Epigenome Project (HEP) was initiated to identify and catalogue Methylation Variable Positions (MVPs) in the human genome [32]. Similar to the profile of the Human Genome Project (HGP), HEP is also a private / public collaboration.

1.2Post- processing phase

Co-clustering techniques envisage distance functions and cluster quality measures for integrating data models and making them indexed. They generally group the gene expression data with similar expression patterns, i.e., co-expressed genes [12]. It also focuses on selecting and eliminating ambiguous and redundant rules. Both of these approaches' efficacies are increased significantly when worked in tandem.

A. Clustering

Clustering is an exceptional preference for initial data analysis and data mining processes. To perceive and identify appealing patterns of expression across multiple genes and experiments, reveal natural structures and reduce high-dimensional array data clustering must be ascertained to allow easier management of data set.

2. Related work

Erfaneh Naghieh and Yonghong Peng [1] After genome sequencing, DNA microarray analysis has become the most widely used functional genomics approach in the bioinformatics field. The enormous amount of unparalleled qualities of genome-wide data produced by the DNA Microarray research. Clustering is the process of grouping data objects into set of disjoint classes called clusters so that objects within a class are highly similar with one another and dissimilar with the objects in other course. It is currently the remote most worn method for gene expression analysis which provides a divide-and–conquer strategy to extract meaningful information from expression profile. Chad Creighton and Samir Hanash [2] Global gene appearance profile, both at the transcript level and the protein level, and it can be a valuable tool in the understanding of genes and cellular states. As larger gene expression data sets suit available, data mining techniques are applied to identify patterns of concern in the data. Association rules are used widely in the area of market basket analysis. Items in gene expression data can include genes that are highly uttered or reserved.Daxin Jiang, Chun Tang, and Aidong Zhang [3] The large number of genes and the complexity of biological networks greatly increase the challenges of comprehending and interpreting the resulting of data. The clustering techniques, which is crucial in the data mining process to reveal natural structures and identify interesting patterns in the underlying data. Cluster analysis seeks to panel a known data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups.

Mrs. Smita R. Londhe, Prof. Rupali A. Mahajan [4] Mining high utility item sets from a transactional database refers to the discovery of item sets with high utility like

profits. They gain the problem of producing a large number of candidate item sets. Such a large number of nominee item sets degrade the mining performance in terms of execution time and space requirement. The main objective of utility mining is to identify the item sets with highest utilities, by allowing for profit, quantity and cost or other user preferences. They have many applications in website click business promotion in chain hypermarkets, online e-commerce management, mobile commerce environment planning.

Andrew B. Goryachev, Pascale F. Macgregor [5] The use of DNA microarrays for the analysis of complex biological samples is becoming a ordinary part of biomedical explore. One of the most commonly used methods compares the relative abundance of mRNA in two different samples by probing a single DNA microarray all together. The outcome of this analysis are presented in the form of a model describing the relationship between the measured fluorescent intensities and the concentration of mRNA transcript. We developed and tested several algorithms for inference of the model parameters for the microarray data. [5]

3. Proposed algorithm

The procedures for cluster analysis are the feature selection, cluster algorithm selection, cluster validation and result interpretation. [2] The intimately connected steps of cluster analysis with feedback pathways is shown in the following figure.

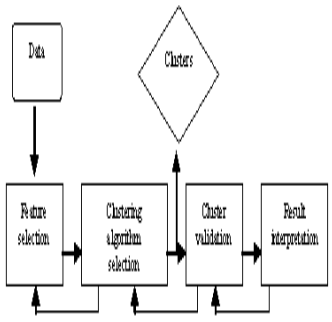


Fig.1.

3.1.Cluster algorithm design or selection

Different clustering algorithms and methods have been developed to improve the preceding ones, unraveling the problems and fit for specific fields [3]. There is no absolute clustering method that can be universally used to solve all problems. So in order to select or generate a suitable clustering strategy, it is vital to investigate the features of the problem. Patterns are grouped according to whether they resemble each other. Once a proximity measure is chosen, the structure of a clustering criterion function makes the partition of clusters an optimizing problem.

3.2.Clustering Techniques

Many diverse clustering techniques have extensively been under development [3]. The most widely used techniques in analysis of gene expression data which are applied in the early stages and proven to be useful are Hierarchical clustering [7], K-means clustering [8] and Self-organized maps (SOM) [9].

3.3.Hierarchical clustering

Hierarchical clustering [7] is the first and most common clustering method applied to gene expression data which is developed on the basis of a single layered neural network. A hierarchical series of nested clusters are generated by grouping genes with similar patterns of expression. Hierarchical clustering calculates all pairs-wise distance relationships between genes and experiments to merge pairs of values that are most similar to the formation of a node. These methods are either agglomerative algorithms (bottom-up approach) which joins clusters in a hierarchical manner or the more rapid dividing algorithms (top-down approach) which splits clusters hierarchically.

3.4.K-means Clustering

K-means clustering [8] is a simple and fast method used commonly due to its

straight forward implementation and small number of iterations. This algorithm divides the data set into k disjoint subsets. An estimation of the number of clusters (*k*) is made by the user and calculated as an input where the computer randomly assigns each gene to one of the k clusters. The drawbacks of this method are the lack of prior knowledge of the number of gene clusters in a gene expression data which results in the changing of results in the altering of results in successive runs since the initial clusters are selected randomly and the quality of the attained clustering has to be assessed.

3.5.Self-Organized Maps Clustering (SOM)

SOM [9] is a reasonably fast and easy to implement method, scalable to large data sets. It is intimately related to multidimensional scaling and its objective is to represent all data points in the source space by points in a target space where the distance and proximity relationships are preserved. At the input, the data objects are presented and output neurons are organized with a sample neighborhood grid structure.

3.6. Supervise and Un Supervised clustering

Supervised methods are used for analysis when trying to classify objects into known classes and finding genes that is mainly applicable to label classification [6]. Unsupervised sample-based clustering mines through data, congregating into a precise partition of the samples and a set of informative genes extract in relevant information without the presence of a teacher signal. Unsupervised approach is more complex than supervised. Common unsupervised methods include hierarchical clustering, k means clustering and self-organized maps etc.

Clustering gene expression data can be categorized into the three groups, 1) gene-based, 2) sample-based and 3) subspace clustering as both genes and samples is required to be clustered significantly.

3.7. Gene-based clustering

The gene-based clustering intends to group together co-expressed genes which indicate co-function and co-regulation which reveals the natural data structures [11]. Genes are treated as the object, while the samples are the features. Clustering algorithms for gene expression data should be competent of extracting useful information from a high level of background noise. A good clustering algorithm should depend as little as possible on prior knowledge also provide graphical representation of the cluster structure other than partitioning the data.

3.8. Sample-based clustering

Samples are generally related to various disease or drug effects within a gene expression matrix. Only a small subset of genes whose expression levels strongly correlate with the class distinction, rise and fall coherently and exhibiting fluctuation of a similar shape under a subset of conditions, called the informative gene that participates in any cellular process relevant. The remaining genes are regarded as noise in the data as they are irrelevant to the sample of interest.

3.9. Subspace clustering

The subset of features for different subspace cluster. Genes and samples are treated symmetrically, such that either genes or samples can be regarded as objects or features. A single gene may contribute in multiple pathways that may or may not be collective under all Conditions Subspace clustering [12] techniques confine coherence exhibit by the blocks within gene expression matrices. A block is a sub matrix defined by a subset of genes on a subset of samples.

3.10. Biclustering

Biclustering [13] performs simultaneous clustering on the row and column dimension of the data matrix where the gene exhibits

highly correlated activities for every condition instead of clustering these two dimensions separately, which distinct classes of clustering algorithms that perform simultaneous row-column clustering to identify submatrices, subgroups of genes and subgroups of conditions. Clustering derives a global model while Biclustering produces a local model.

3.11. Triclustering

Triclustering [14] is mined coherent clusters in three dimensional 3D gene expression datasets. It mines arbitrary positioned and overlapping clusters and depending on different parameter values which mines diverse variety of clusters, together with those with constant or similar values along each dimension as well as scaling and shifting expression patterns. Tricluster relies on graph-based approach to mine all valid clusters and merge/delete some clusters having large overlaps. Tri cluster can find significant tri clusters in the real microarray datasets.

4. Existing system:

Gene expression is the creation of a gene that results in a protein which tends to identify only the Gene manipulation for Cancer therapeutics. In our existing approach, identification of cancer by the gene expression have been implemented. The Genome of a differentiate Cell Contains all the Genes required to find the affected cells using Microarrays to Investigate the "Expression" of Thousands of Genes at a Time. Discredited gene terms can be used as descriptors of the specific states of gene for the cancer prediction analysis.

4.1.Existing algorithm:

1. Supervised Multi Attribute Clustering Algorithm:

A supervised Multi attributes clustering algorithm is used in our existing system in order to find the co-regulated clusters of genes whose collective expression is strongly associated with the sample categories of gene expressions.

4.2. Agglomerative Algorithm*:*

The algorithm forms cluster in a bottom-up manner,

1. Initially, put each gene expression data's in its own cluster.

2. Among all clusters, pick the two clusters with the smallest distance.

3. Replace these two gene hierarchical clusters with a new gene cluster, formed by merging the two original ones.

This kind of hierarchical **clustering** is called **agglomerative** because it merges cluster iteratively. There is also a divisive hierarchical **clustering**

4.3Proposed system:

- Predicting Cancer by analyzing and convert the gene expression is the proposed concept of our project, which leads to identifying and analyzing the cancer result set.
- Controlling Gene Activity from Gene to efficient Protein & Phenotype has also been analyzed in order to identify the cancer cells.
- In our proposed methodology, the experts documental DNA data mutilation (Gene expression segments) is a kind of binding site for proteins which make DNA inaccessible to be in a live state.

4.4 Proposed algorithm:

1. Semantics Sequence Structure Algorithm:

A semantic indexing algorithm which uses and relies on the domain specific temporal structure of the gene frame sequence for high-level gene expression data's.

2. Simultaneous Computing Big O Algorithm:

A simultaneous BigO algorithm is used to specify the performance of the gene expression, which will allow to find the individual gene's evidence for predicting cancer.

3. Ant Colony Optimization Algorithm:

Enable gene expression data's search for solutions collaboratively and effectively.

Example pseudo code:

- Procedure ACO_GeneExpression
- While (not_termination)
- Generate Solutions ()
- Daemon Actions ()
- Pheromone Update ()
- End while
- End procedure

II. SYSTEM DESIGN

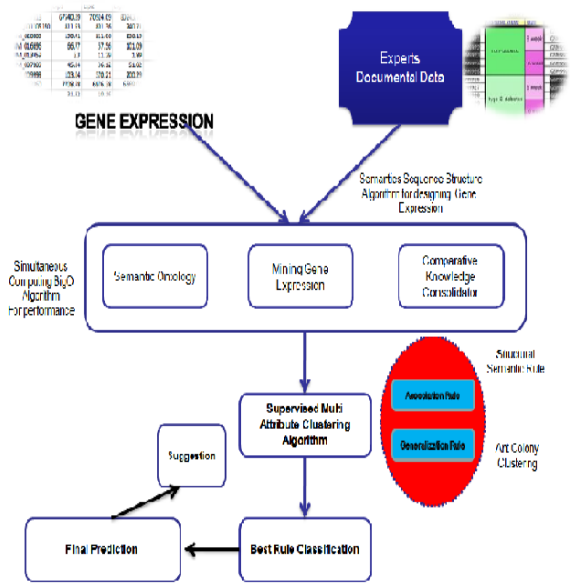


Fig.2. Architecture Diagram

A Semantic indexing algorithm which uses and relies on the domain specific temporal structure of the gene frame sequence forhighlevel gene expression data's. Simultaneous BigO algorithm is used to specify the performance of the gene expression, which will allow to find the individual gene's evidence for predicting cancer. Enable gene expression data's to search for solutions collaboratively and effectively.

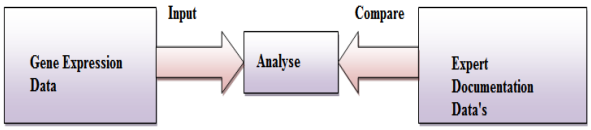
MODULES

Implementation is the stage of the project when the theoretical design is turned out into a working system.

1. Data Set Input Module
2. Gene Knowledge Extraction Module
3. Ontological Mapping Module
4. Gene Expression Design Module
5. Ontological Gene Association Module
6. Disease Prediction Module
7. Suggestor Module
8. Performance Evaluation Module

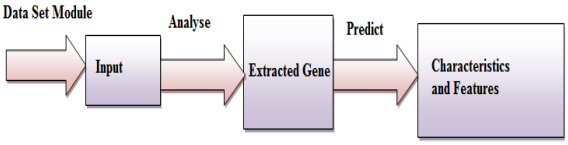
A. Data set input module:

Gene expression data set will be given as an input in order to analyze it with the expert documentation data. In this module we will be using a Semantics sequence structure algorithm for design gene expression which is nothing but the comparison of gene data's.



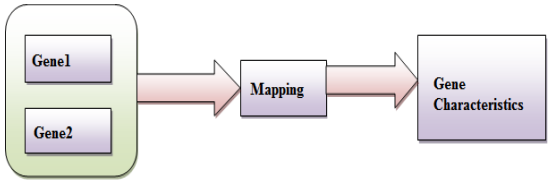
B. Gene knowledge extraction module:

In this module based on the analysis of the data set the input module's comparison, we will analyze the extracted gene to predict its characteristics and other features. Gene knowledge extraction module mainly focuses towards the performance of the individual gene expression data's.



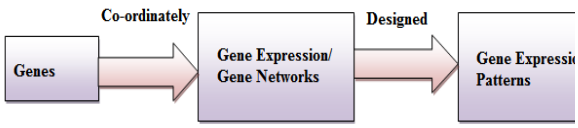
C. Ontological mapping module:

In this module, ontological mapping is done, which is nothing but the mapping of two different gene expression data's to find the difference in gene characteristics. The ontological mapping analysis will provide insights on the pragmatics of ontology mapping towards gene expression elaboration.



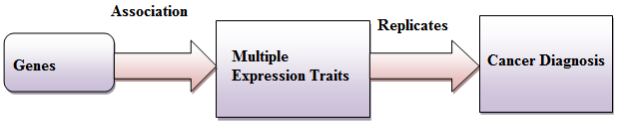
D. Gene expression design module:

In this module genes work co-ordinately as gene expression or gene networks in which it was designed to find gene expression patterns by grouping genes. Here the Genes whose expression is modulated by the genetic variants (different genes in the human body) will act as Trans-Regulated Gene Modules in Humans to identify the cancer.



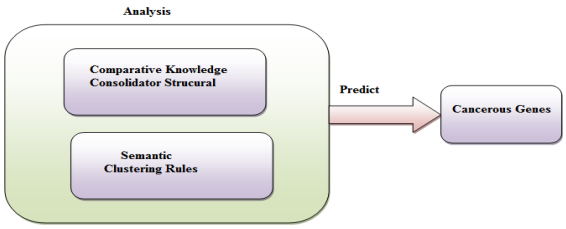
E. Ontological gene association module:

The association of the gene's with multiple expression traits was replicated in Cancer diagnosis (An analysis of cancer prediction in human genes), which is an independent study of Gene Ontology is implemented in this module to enrich the cancer diagnosis of genetic analysis.



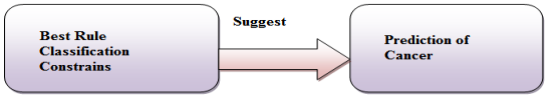
F. Disease prediction module:

Based on the analysis of the Comparative Knowledge Consolidator the structural and semantic clustering rules we have been predicting the cancerous genes.



G. Suggestor module:

In this module, under the best rule classification constrains, we suggest the final prediction of cancer.



H. Performance evaluation module:

In this module we will evaluate the performance and suggest the medicines that need to be taken up for the cancer diagnosis.

HEATMAP

The heatmap displays genome-wide data from transcriptome, protein, epigenetic,

mutation, RNA and PARADIGM pathway analysis studies and associated clinical information. The dataset name and the number of samples is displayed in the dataset header. Each dataset is collected of two Heatmaps : the left genomic heatmap and the right clinical feature heatmap.

Columns in the genomic heatmap represent individual probe that have been mapped to chromosome positions or gene names. All data are mapped to the human Mar. 2006 (NCBI36/hg18) assembly. Columns in the clinical heatmap represent clinical data associated with the genomic dataset with the names appearing at the bottom of the heatmap.

A. Heatmap key color:

Data values are represented with the subsequent default colors:

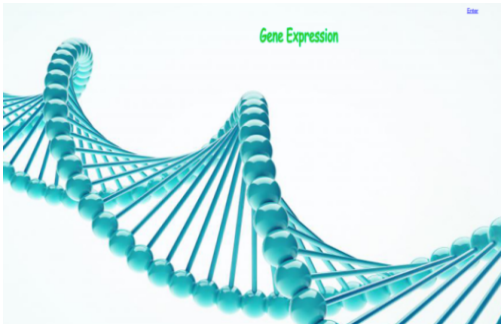
Heatmap	Data Type	Color	Data Value
genomic	gene expression	red	> 0
		green	< 0
		grey	no data
	other data types	red	> 0
		blue	< 0
		grey	no data
clinical	numerically continuous information (such as age)	yellow	> 0
		green	< 0
		grey	no data
	categorical information (such as treatment)	discrete colors	
		grey	no data

4.5. Discussion

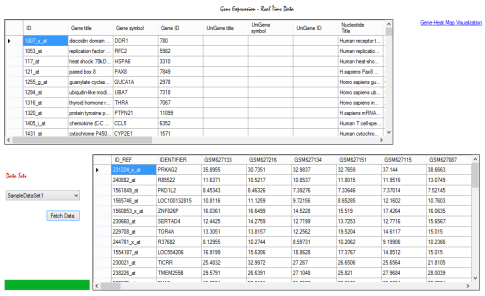
Predicting cancer by analyzing gene and converting the gene expression is the proposed concept of our project, which leads to identifying and analyzing the cancer result set. Controlling Gene Activity From Gene to Functional Protein & Phenotype has also been analyzed in order to identify the cancer cells. In our proposed methodology, the expert's documental DNA data Methylation (Gene expression segments) is a kind of binding site for proteins which make DNA inaccessible to be in alive state. Semantic Ontology based Mining Gene Expression analysis tends to compare the gene expression values by using the comparative Knowledge Consolidator. Supervised Multi Attribute Clustering Algorithm has been used to find the Best Rule Classification in the gene expression to find the Final Prediction of cancer disease.

4.6.Results

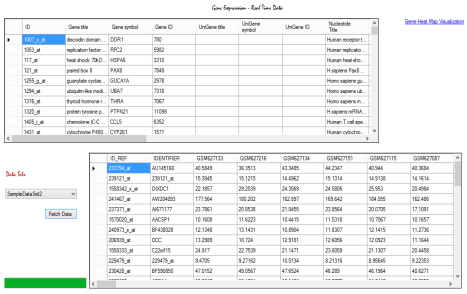
By allowing the monitoring of expression levels in cells for thousands of genes parallel, microarray experiment may lead to a more complete understanding of the molecular variations among tumors and hence to a finer and more informative classification. Therefore, the input data has to be concise and close to accurate to obtain the results of the same nature. The reasons liable can be time and economic constraints.



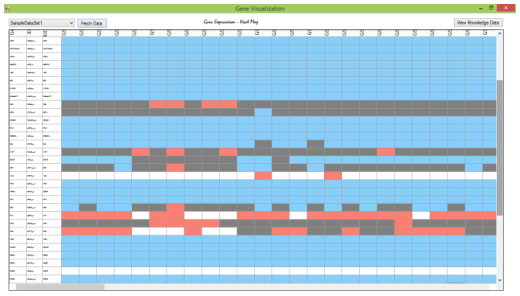
(a)



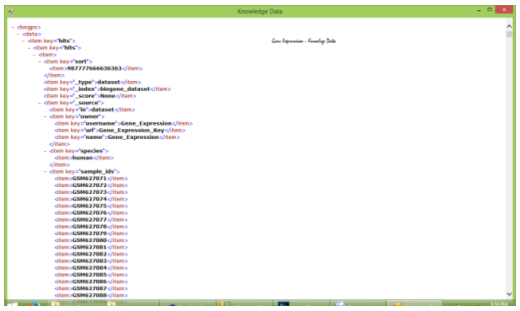
(b)



(c)



(d)



(e)

4.7.Conclusion

It is experimental that a reliable and precise classification of tumor is essential for successful diagnosis and treatment of cancer [3]. The future generations of sequencing can facilitate efficiency in deciphering faulty genes and validating diseases with less deception.

Reference

[1] Data and Statistics. World Health Organization, Geneva, Switzerland, 2006.

[2] PubMedHealth- U.S. Nat. Library Med., (2012).

[3] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," J. Amer. Statist. Assoc., vol. 97, no. 457, pp. 77–87, Mar. 2002.

[4] G.-M. Elizabeth and P. Giovanni, (2004, Dec.). "Clustering and classification methods for gene expression data analysis." Johns Hopkins Univ., Dept. of Biostatist. Working Papers. Working Paper 70.

[5] E. Shay, (2003, Jan.). "Microarray cluster analysis and applications".

[6] T. Kohonen, Self-Organising Maps. Berlin, Germany: Springer- Verlag, 1995.

[7] N. Pasquier, C. Pasquier, L. Brisson, and M. Collard, (2008). "Mining gene expression data using domain knowledge," Int. J. Softw. Informat, vol. 2, no. 2, pp. 215–231.

[8] N. Revathy and R. Amalraj, "Accurate cancer classification using expressions of very few genes," Int. J. Comput. Appl., vol. 14, no. 4, pp. 19–22, Jan. 2011.

[9] Y. Su, T. M. Murali, V. Pavlovic, M. Schaffer, and S. Kasif, (2003) "RankGene: Identification of diagnostic genes based on expression data," Bioinformatics, vol. 19, no. 12, pp. 1578–1579.

[10] K. Raza and A. Mishra, "A novel anticlustering filtering algorithm for the prediction of genes as a drug target," Amer. J. Biomed. Eng., vol. 2, no. 5, pp. 206–211, 2012.

[11] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey,"

IEEE Trans. Knowl. Data Eng., vol. 16, no. 11, pp. 1370–1386, Nov. 2004.

[12] D. A. Roff and R. Preziosi, "The estimation of the genetic correlation: The use of the jackknife," Heredity, vol. 73, pp. 544–548, 1994.

[13] T. Scharl and F. Leisch, "Jackknife distances for clustering timecourse gene expression data," in Proc. ASA Biometrics, 2006, p. 8.

[14] M. B. Eisen, T. P. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," Proc. Nat. Acad. Sci. USA, vol. 95, no. 25, pp. 14863–14868, Dec. 1998

[15] S. Tavazoie, D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," Nature Genetics, vol. 22, pp. 281–285, 1999.