Cancer Prediction using Mining Gene Expression Data

Mr. S.Sivakumar[1], D.Viji[2]
Assistant Professor[1], PG Student[2]
Computer Science Department
Adhiparasakthi Engineering College
Melmaruvathur – 603319.

ABSTRACT— *Cancer is a major cause of all natural mortalities and morbidities throughout the world.Pointed out the exact tumour types provides an optimized solution for the better treatment and toxicity minimization due to medicines on 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 analyse global gene expression which provides an optimized solution for the identified problem area. Molecular diagnostics provides a promising option of systematic human cancer classification, but these tests are not widely applied 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 neighbour analysis to support vector machine manipulation for the learning portion of the classification model. We don't have a clear picture of supervised classifier (Supervised Multi Attribute Clustering Algorithm) which can manage knowledge attributes coming two different knowledge streams. Our proposed system takes 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: *DNA Microarray, Gene expression, Ontology, supervised multi attribute clustering***.**

I. INTRODUCTION

Genes are pieces of DNA (deoxyribonucleic acid) inside each of our cells that instruct them how to make the proteins the body needs to function. DNA is the genetic "blueprint" found in each cell. Genes affect inherited traits passed on from a parent to a child, such as hair color, eye color, and height. They also affect whether a person is likely to develop certain diseases, such as cancer. Changes to these genes, called *mutations*, play an important role in the development of cancer. Mutations can cause a cell to make (or not make) proteins that affect how it grows and divides into new cells. Cancer is an abnormal and uncontrollable growth of cells in the body that turn malignant. This is not to be confused with tumors.

Even a tumor is an abnormal growth of cells. Notice that all the cancer cells are tumor but reverse is not possible. Cancer cells are can easily spread out. There are many causes of cancers: Drinking excess alcohol, Tobacco, Sunlight, Diet, Radiation, etc. Symptoms of cancer depend on the type and location of the cancer. For example, lung cancer can cause coughing, heavy breathing, chest pain, etc. Colon cancer often causes diarrhea, constipation, dysentery, and blood in the stool [13]. Some cancers may not have any symptoms at all. In certain cancers, such as pancreatic cancer, symptoms often do not start until the disease has reached an advanced stage.

Genetic testing is the process of using medical tests to look for changes (mutations) in a person's genes or chromosomes. The obvious benefit of genetic testing is the chance for a better understanding of our risk for a certain disease. Testing is not perfect, but it can often help us make decisions about our health. Genetic testing can cost a lot, and it can take several weeks to get the results. Using better technologies, tests are becoming more accurate and are able to look at more than one gene at a time.

II. DNA MICROARRAY ANALYSIS

DNA microarrays are one of the fastest growing technologies for genetic research. DNA microarrays are used to investigate cancer, for measuring changes in gene expression and learning how cells are respond to a disease or to a particular treatment. Even if microarrays represent a powerful source of biological information, using gene expression data to classify diseases on a molecular level for clinical diagnostic remains a challenging research problem.

Classifying microarray data poses several challenges to typical machine learning methods. In particular, microarray classification faces the “small N, large P” problem of statistical learning, where the number P of variables (gene expressions) is typically much larger than the number N of available samples[13]. The major aspects of the classifier design: the classification rule, the error estimation, and the feature selection. One of the main problems of traditional machine learning techniques concerns the ability of properly detecting false positives, i.e., samples erroneously assigned to a class even if they do not belong to the class library used to train the classifier. This misbehavior is clearly unacceptable since it would very likely lead to a misdiagnosis.

This micro array analysis model is very flexible, and it makes the implementation of classification, clustering. The classifier is not only able to correctly classify samples in the corresponding classes, but it is also able to correctly detect out-of-class samples, thus drastically reducing the false positive rate. cDNA microarrays models provided very good results.

III. DATAPROCESSING

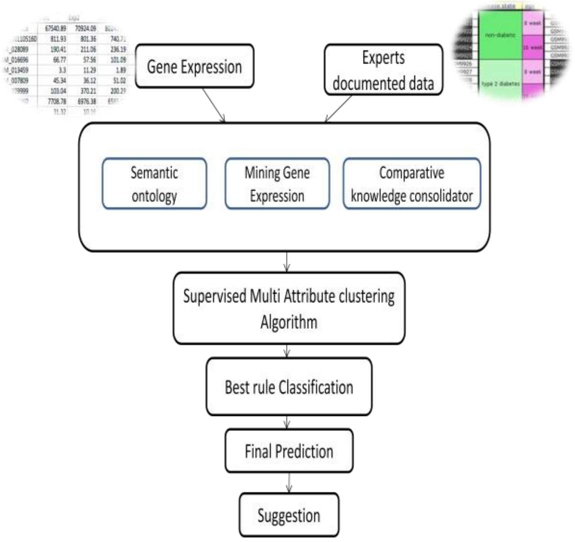

Figure 1 Data Processing for Mining Gene Expression Data

Figure 1 shows that, input is given as DNA Genetic Dataset and Experts documents to analyse the gene expression. Using these input data the following process takes place. They are:

1. **Semantic Ontology:** The process of grouping the genes based on semantic. In this process genes of similar categories are grouped together.
2. **Mining Gene Expression:** The process of mining the gene from the gene expression, where affected and unaffected genes are classified. Mining is the process of extracting particular information from collection of data.

A Comparative Knowledge Consolidator is used initially to extract its knowledge information for Semantic Ontology and Mining Gene Expression. At the end of these two processes we get affected gene form the whole set of gene expression. Clustering and classification is used further for the affected set of gene derived. This is done using Supervised Multi-Attribute Clustering algorithm.

Clustering Algorithm is based on Supervised and Unsupervised algorithm.

Unsupervised Clustering algorithm is much more complex than a Supervised manner, since no training set of samples can be utilized as a reference to guide informative gene selection. The following two challenges of unsupervised sample based clustering it make very hard to detect phenol types of samples and select informative genes.

Supervised Clustering algorithm are widely used by biologists to pick up the informative genes. The goal of informative gene selection step is to pick up those genes whose expression patterns can distinguish different pheno types of samples.

Then, the affected gene of clustered collection is then further classified using the Best Rule Classification. After filtering the affected gene by using various data mining techniques, now the Final Prediction Technique is used to predict the particular cancer more effectively. And finally suggestion will be given for a particular cancer depending upon the genetic information.

IV.GENE EXPRESSION DATA

A microarray experiment typically assesses a large number of DNA sequences (genes, cDNA clones, or expressed sequence tags) under multiple conditions. This cluster analysis of gene expression data without making a distinction among DNA sequences, which will uniformly be called "genes"[4]. Similarly, it is referred to all kinds of experimental conditions as "samples", if no confusion will be caused. A gene expression data set from a microarray experiment can be represented by a real-valued expression matrix $M = \{W_{ij} \mid 1 \le i \le n, 1 \le j \le m\}$as shown in Figure 2, where the rows ($G = \{g1\ldots gn\}$) form the expression patterns of genes, the columns (*S* = {*S1…Sm}*) represent the expression profiles of samples, and each cell is the measured expression level of gene *i* in sample *j*.

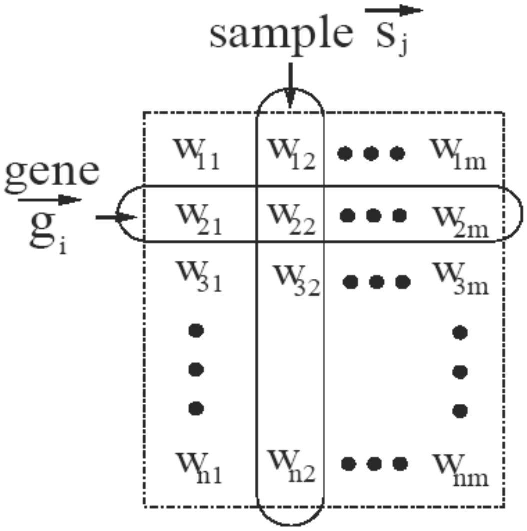
sample $\overrightarrow{s_j}$
gene $\overrightarrow{g_i}$
w_{11} w_{12} ●●● w_{1m}
w_{21} w_{22} ●●● w_{2m}
w_{31} w_{32} ●●● w_{3m}
w_{n1} w_{n2} ●●● w_{nm}

Figure 2 A Gene Expression Matrix

V. CATAGORIES OF GENE EXPRESSION CLUSTERING

One of the characteristics of gene expression data is that it is meaningful to cluster both **genes** and **samples.**

Gene expression can be analyzed into two ways gene-based clustering and sample based clustering.

A .Gene-Based Clustering

In such gene-based clustering, the genes are treated as the objects, while the samples are the features. On the other hand, the samples can be partitioned into homogeneous groups.

B. Sample-Based Clustering

Such sample-based clustering regards the samples as the objects and the genes as the features. The distinction of gene-based clustering and sample-based clustering is based on different characteristics of clustering tasks for gene expression data. Some clustering algorithms, such as K-means and hierarchical approaches, can be used both to group genes and to partition samples.

VI. VARIOUS CLUSTERING ALGORITHM

In this section, the problem of clustering genes based on their expression patterns. The purpose of gene-based clustering is to group together coexpressed genes which indicate cofunction and coregulation.

Challenges of Gene Clustering

First, cluster analysis is typically the first step in data mining and knowledge discovery. The purpose of clustering gene expression data is to reveal the natural data structures and gain some initial insights regarding data distribution. Therefore, a good clustering algorithm should depend as little as possible on prior knowledge, which is usually not available before cluster analysis.

Second, due to the complex procedures of microarray experiments, gene expression data often contains a huge amount of noise. Therefore, clustering algorithms for gene expression data should be capable of extracting useful information from a high level of background noise.

Third, our empirical study has demonstrated that gene expression data are often “highly connected” , and clusters may be highly intersected with each other or even embedded one in another . Therefore, algorithms for gene-based clustering should be able to effectively handle this situation.

Finally, users of microarray data may not only be interested in the clusters of genes, but also be interested in the relationship between the clusters and the relationship between the genes within the same cluster

A. K-Means

The K-Means algorithm is a typical partition-based clustering method. Given a pre-specified number K, the algorithm partitions the data set into K disjoint subsets which optimize the following objective function:

$$E = \sum_{i=1}^{K} \sum_{O \epsilon C_i} | O - \mu_i|^2$$

Here, O is a data object in cluster *Ci* and μ_i is the centroid (mean of objects) of *Ci*. Thus, the objective function E tries to minimize the sum of the squared distances of objects from their cluster centers [4].

Algorithm

1. The K-Means algorithm accepts the "number of clusters" to group data into, and the dataset to cluster the input values.

2. The K-Means algorithm then creates the first k initial clusters from the data set

3. The K-Means algorithm calculates the arithmetic mean of each cluster formed in the data set. The arithmetic mean is the mean of all the individual records in the cluster.

4. Next K-Means assigns each record in the dataset to only one of the initial clusters. Each record is assigned to the nearest cluster using proximity measure like Euclidean distance.
5. K-Means reassigns each record in the dataset to the most similar cluster and recalculates the arithmetic mean of the clusters in the dataset.
6. K-Means reassigns each record in the dataset to only one of the new clusters formed
7. The preceding steps are repeated until "stable clusters" are formed and the K-Means clustering is completed

The K-Means algorithm is simple and fast. The time complexity of K-Means is O(l*m*n), where *l* is the number of iterations and *k* is the number of clusters, *m* is the number of genes and *n* is the number of samples. Our empirical study has shown that the K-Means algorithm typically converges in a small number of iterations. However, it also has several drawbacks as a gene-based clustering algorithm.

First, the number of gene clusters in a gene expression data set is usually unknown in advance. To detect the optimal number of clusters, users usually run the algorithms repeatedly with different values of *k* and compare the clustering results. For a large gene expression data set which contains thousands of genes, this extensive parameter fine-tuning process may not be practical.

Second, gene expression data typically contain a huge amount of noise; however, the K-Means algorithm forces each gene into a cluster, which may cause the algorithm to be sensitive to noise.

B. SOM

The Self-Organizing Map (SOM) was developed by Kohonen [21], on the basis of a single layered neural network. The data objects are presented at the input, and the output neurons are organized with a simple neighborhood structure such as a two dimensional p*q grid. Each neuron of the neural network is associated with a reference vector, and each data point is "mapped" to the neuron with the "closest" reference vector.

In the process of running the algorithm, each data object acts as a training sample which directs the movement of the reference vectors towards the denser areas of the input vector space, so that those reference vectors are trained to fit the distributions of the input data set. When the training is complete, clusters are identified by mapping all data points to the output neurons.

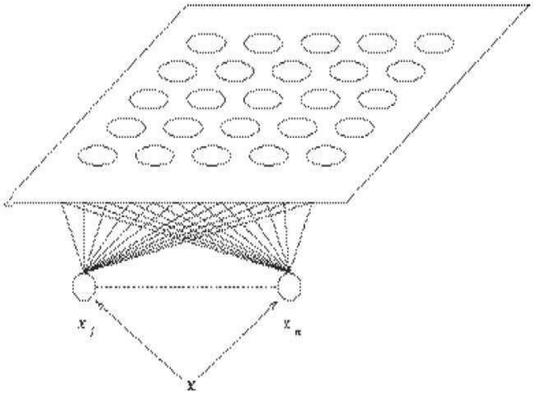

Figure 3 Schematic representation of a self-organizing map method

C. Hierarchical Clustering

Hierarchical clustering generates a hierarchical series of nested clusters which can be graphically represented by a tree, called *dendrogram*. The branches of a dendrogram not only record the formation of the clusters but also indicate the similarity between the clusters. By cutting the dendrogram at some level, we can obtain a specified number of clusters. By reordering the objects such that the branches of the corresponding dendrogram do not cross, the data set can be arranged with similar objects placed together. The hierarchical clustering scheme:
Let S={*Si,j*} is the input similarity matrix, where *Si,j* indicates similarity between two data objects based on Euclidean distance.

Algorithm:

1. Find a minimal entry s(i, j) in S, and merge clusters i and j.
2. Modify S by deleting rows and columns i, j and adding a new row *i* and column *j*, with their dissimilarities defined by:

$$s(k; i \cup j) = s(i \cup j, k) = \alpha_i s(k, i) + \alpha_j s(k, j) + \gamma|s(k, i) - s(k, j)|$$

3. If there is more than one cluster, then go to Step 1.
Common variants of this scheme, obtained for appropriate choices of the α- *s* and γ parameters, are the following:

$$\textit{singlelinkage} : s(k; i \cup j) = min = \{s(k; i); s(k; j)\}$$

$$\textit{completelinkage} : s(k; i \cup j) = max\{s(k, i); s(k, j)\}$$

$$\textit{averagelinkage} : s(k, i \cup j) = (n_i d(k, i) + n_j d(k; j))/(ni + nj),$$

where *ni* denotes the number of elements in cluster i.

Hierarchical clustering not only groups together genes with similar expression pattern but also provides a natural way to graphically represent the data set. The graphic representation allows users a thorough inspection of the whole data set and obtain an initial impression of the distribution of data. However, the conventional agglomerative approach suffers from a lack of robustness , i.e., a small perturbation of the data set may greatlychange the structure of the hierarchical dendrogram. Another drawback of the hierarchical approach is its high computational complexity. To construct a complete dendrogam (where each leaf node corresponds to one data object, and the root node corresponds to the whole data set), the clustering process should take *n*2¡*nn* merging (or splitting) steps. The time complexity for a typical agglomerative hierarchical algorithm is $O(n2logn)$. If a wrong decision is made in the initial steps, it can never be corrected in the subsequent steps.

Limitations Of Above Techniquqes

Clustering algorithms like K-Means, SOM, and hierarchical algorithms were studied and observed that K-Means requires number of clusters before clustering where it is not known earlier for gene expression data. For SOM, grid structure of the neuron map has to be mentioned earlier. The time complexity of hierarchical clustering is very high.

VII. SAMPLE BASED CLUSTERING

The goal of sample-based clustering is to find the phenotype structures or substructures of the samples. Previous studies have demonstrated that phenotypes of samples can be discriminated through only a small subset of genes whose expression levels strongly correlate with the class distinction. These genes are called informative genes. The remaining genes in the gene expression matrix are irrelevant to the division of samples of interest and thus are regarded as noise in the data set.

Although the conventional clustering methods, such as K-means, self-organizing maps (SOM), hierarchical clustering (HC), can be directly applied to cluster samples using all the genes as features, the signal-to-noise ratio (i.e., the number of informative genes versus that of irrelevant genes) is usually smaller than 1 : 10, which may seriously degrade the quality and reliability of clustering results . Thus, particular methods should be applied to identify informative genes and reduce gene dimensionality for clustering samples to detect their phenotypes.

The existing methods of selecting informative genes to cluster samples fall into two major categories: supervised analysis (clustering based on supervised informative gene selection) and unsupervised analysis (unsupervised clustering and informative gene selection).

Supervised sample-based clustering widely used in biologistics for selecting informative gens or pick up informative genes. A subset of samples is selected to form the training set. The goal of informative gene selection step is to pick out those genes whose expression patterns can distinguish different phenotypes of samples. the whole set of samples are clustered using only the informative genes as features. Since the feature volume is relatively small, conventional clustering algorithms, such as K-means or SOM, are usually applied to cluster samples. The future coming samples can also be classified based on the informative genes, thus the supervised methods can be used to solve sample classification problem.

Unsupervised sample-based clustering assumes no phenotype information being assigned to any sample. Unsupervised sample-based clustering is much more complex than a supervised manner since no training set of samples can be utilized as a reference to guide informative gene selection. unsupervised sample-based clustering make it very hard to detect phenotypes of samples and select informative genes. Since the number of samples is very limited while the volume of genes is very large, such data sets are very sparse in high-dimensional genes space.

Due to lack of drawback in unsupervised sample based clustering algorithm in this system we used supervised sample based clustering algorithm.

VIII BEST RULECLASSIFICATION

IF-THEN Rule:

Rule induction: is the process of extracting useful ‘if then’ rules from data based on statistical significance. A Rule based system constructs a set of if-then-rules.

IF conditions THEN conclusion:

This kind of rule consists of two parts. The rule antecedent (the IF part) contains one or more conditions about value of predictor attributes where as the rule consequent (THEN part) contains a prediction about the value of a goal attribute. An accurate prediction of the value of a goal attribute will improve decision-making process. IF-THEN prediction rules are very popular in data mining; they represent discovered knowledge at a high level of abstraction. Rule Induction Method has the potential to use retrieved cases for predictions [16]. Complex decision trees can be difficult to understand, for instance because information about one class is usually distributed throughout the tree.

IF conditions THEN conclusion

This kind of rule consists of two parts. The rule antecedent (the IF part) contains one or more conditions about value of predictor attributes where as the rule consequent (THEN part) contains a prediction about the value of a goal attribute. An accurate prediction of the value of a goal attribute will improve decision-making process. IF-THEN prediction rules are very popular in data mining; they represent discovered knowledge at a high level of abstraction.

In the health care system it can be applied as follows:
(Symptoms) (Previous--- history) → (Cause—of---disease).

XI CONCLUSION

This paper overcome the problem of traditional genetic testing and predicts particular cancer and giving suggestion for that type of cancer. The following problems raised in traditional testing Once the testing is complete, the lab reports the results in writing to the doctor or genetic counselor. You will then be given the results during another counseling session. This may not happen until several weeks after the samples are taken. The accuracy of the test and the meaning of the results will be discussed with you in detail. If the test result is *negative*, it means the gene mutation that was tested for is not present. The test result may be a "*false negative*". This means the test readsnegative but the mutation is actually there. Genetic testing can cost a lot, and it can take several weeks to get the results. But as better technologies are developed, tests are becoming more accurate and are able to look at more than one gene at a time. A gene expression data set typically contains thousands of genes.Recent DNA microarray technologies have made it possible to monitor transcription levels of tens of thousands of genes in parallel; also it will dramatically reduced false positive and false negative results it will provide better accurate results. Gene expression data can be clustered on both genes and samples. As a result, the clustering algorithms can be divided into two categories: gene-based clustering, sample- based clustering. Supervised multi attribute clustering algorithm will be effectively work compared with others. After clustering process Best rule classification used to predict the particular type of cancer accurately. We suggest the final prediction of cancer and suggest the medicine that needs to be taken up for the cancer diagnosis. Suggesting medicine is important thing since two different persons are affected by same cancer also we cannot suggest same medicine for those peoples because they get different genome or gene structure.Cancer prediction system can be furtherenhanced and expanded. It can also incorporate other datamining techniques, e.g., Time Series, Clustering andAssociation Rules. Continuous data can also be usedinstead of just categorical data. Another area is to use TextMining to mine the vast amount of unstructured dataavailable in healthcare databases. Another challenge wouldbe to integrate data mining and text mining.

REFERENCES

[1] Alfredo Benso, Stefano Di Carlo, and Gianfranco Politano" A cDNA Microarray Gene Expression Data Classifier for Clinical Diagnostics Based on Graph Theory" IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 8, No. 3, May/June 2011

[2] Anirban Mukhopadhyay∗, Senior Member, IEEE, Ujjwal Maulik, Senior Member, IEEE, and Sanghamitra Bandyopadhyay, Senior Member, IEEE, "An Interactive Approach to Multiobjective Clustering of Gene Expression Patterns" IEEE Transactions On Biomedical Engineering, Vol. 60, No. 1, January 2013

[3] Cosmin Lazar, Jonatan Taminau, Stijn Meganck, David Steenhoff, Alain Coletta, David Y. Weiss Solı´s, Colin Molter, Robin Duque, Hugues Bersini, and Ann Nowe "GENESHIFT: A Nonparametric Approach for Integrating Microarray Gene Expression Data Based on the Inner Product as a Distance Measure between the Distributions of Genes" IEEE/Acm Transactions On Computational Biology And Bioinformatics, Vol. 10, No. 2, March/April 2013

[4] Daxin Jiang, Chun Tang, and Aidong Zhang," Cluster Analysis for Gene Expression Data: A Survey" IEEE Transactions On Knowledge And Data Engineering, Vol. 16, No. 11, November 2004

[5] Elrasheid A.H. Kheirelseid, Nicola Miller, Kah Hoong Chang, Mary Nugent, Michael J. Kerin," Clinical applications of gene expression in colorectal cancer", Submitted Jan 24, 2013. Accepted for publication Feb 27, 2013. doi: 10.3978/j.issn.2078-6891.2013.010

[6] KM Fedorka1, WE Winterhalter1 and TA Mousseau2," The evolutionary genetics of sexual size dimorphism in the cricket Allonemobius socius" Heredity (2007) 99, 218–223 & 2007 Nature Publishing Group All rights reserved 0018-067X/07

[7] Kristopher L. Patton, David J. John, James L. Norris, Daniel R. Lewis, and Gloria K. Muday "Hierarchical Probabilistic Interaction Modeling for Multiple Gene Expression Replicates" IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 11, No. 2, March/April 2014

[8] Lipo Wang, Feng Chu, and Wei Xie "Accurate Cancer Classification Using Expressions of Very Few Genes" IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 4, No. 1, January-March 2007

[9] Robert S. H. Istepanian , Senior Member, IEEE, Ala Sungoor, and Jean-Christophe Nebel, Senior Member, IEEE " Comparative Analysis of Genomic Signal Processing for Microarray Data Clustering " IEEE Transactions On Nanobioscience, Vol. 10, No. 4, December 2011

[10] Ujjwal Maulik∗, Senior Member, IEEE, Anirban Mukhopadhyay∗, Senior Member, IEEE, and Debasis Chakraborty "Gene-Expression-Based Cancer Subtypes Prediction Through Feature Selection and Transductive SVM" Ieee Transactions On Biomedical Engineering, Vol. 60, No. 4, April 2013

[11] 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.

[12] 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.

[13] Shaurya Jauhari and S.A.M. Rizvi Mining Gene Expression Data Focusing Cancer Therapeutics: A Digest IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 11, No. 3, May/June 2014

[14] A. Ben-Dor, N. Friedman, and Z. Yakhini, "Class Discovery in Gene Expression Data," Proc. Fifth Ann. Int'l Conf. ComputationalMolecular Biology (RECOMB 2001), pp. 31-38, 2001.

[15] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering Gene Expression Patterns," J. Computational Biology, vol. 6, nos. 3/4, pp. 281-297, 1999.

[16] Harleen Kaur and Siri Krishan Wasan**,** Empirical Study on Applications of Data Mining Techniques in Healthcare, Journal of Computer Science 2 (2): 194-200, 2006ISSN 1549-3636.

[17] 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, [Online] Available: http://www.ijsi.org/1673-7288/2/215.pdf.

[18] 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, [Online] Avaialble: http://bioinformatics.oxfordjournals.org/content/19/12/1578.full.pdf.

[19] K. M Williams, "Statistical Methods for analysing microarray data: Detection of differentially expressed genes" Inst. Signal Process.,Tampere Univ. Technol. Tampere, Finland, Dep. Biology, Univ. York, York, U.K., 2004.

[20] B. Collard, "An ontology driven data mining process" Inst. TELECOM, TELECOM Bretagne, CNRS FRE 3167 LAB-STICC,Technopole Brest-Iroise, France & Univ. Nice Sophia Antipolis, France, 2008.

[21] T. Kohonen, Self-Organization and Associative Memory. Berlin: Spring-Verlag, 1984.