IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661 Volume 4, Issue 1 (Sep-Oct. 2012), PP 31-36
www.iosrjournals.org

An Overview on Gene Expression Analysis

Dr. R. Radha[1], P. Rajendiran[2]
[1](*Department of Computer Science, S. D .N. B. Vaishnave college of Women, chromepet, Chennai, Tamil nadu – India.)*
[2](*Department of Computer Science, Vidyaa Vikas Educational Institutions, Tiruchengode, Namakkal, Tamilnadu - India)*

Abstract: *Recent advances in DNA microarray technology, also known as gene chips, allow measuring the expression of thousands of genes in parallel under multiple experimental conditions [1]. This technology is having a significant impact on genomic studies. Disease diagnosis, drug discovery and toxicological research benefit from the microarray technology. Arrays are now widely used in basic biomedical research for mRNA expression profiling and are increasing being used to explore patterns of gene expression in clinical research.*
Keywords: *ANN, Classification, Clustering, Gene expression, Micro Array*

I. Introduction

Various approaches have recently been used in outcome prediction using gene expression data. It has been shown that specific patterns of gene expression occur during different biological states such as cell development and during normal physiological responses in tissues and cells. There are many data mining techniques which help to analyze the gene expression data [2]. The generation of quantitative expression patterns of many genes in parallel can be achieved by using techniques based on complementary DNA micro arrays [3], [4].

II. Gene Expression Data

A microarray experiment typically assesses a large number of DNA sequences (genes, CDNA clones, or expressed sequence tag [ESTs] under multiple conditions. These conditions may be a time series during a biological process (e.g: the yeast cell cycle) or a collection of different tissue samples [5]. The original gene expression matrix obtained from a scanning process contains noise, missing values, and systematic variations arising from the experimental procedure. Within a gene expression matrix, there are usually several particular macroscopic phenotypes of samples related to some diseases or drug effects. 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 [5].

A recent effort to understand how genes contribute to disease approaches the discovery of sub-classes of diffuse large B-cell lymphoma (DLBCL) by using expression analysis [6]. It has been shown that the discovery of sub-classes in DLBCL has not been successful by relying exclusively on morphological features [3]. Alizadeh et al [6] demonstrate that the molecular profile of a tumor obtained from CDNA microarrays can indeed be interpreted as a robust and clear picture of the tumor biology.

In [7], at Patrick Brown's lab at Stanford has used microarrays to measure gene expression levels for the entire yeast genome (approximately 6400 distinct CDNA sequences) during the diauxic shift (transition from sugar metabolism to ethanol metabolism) , sporulation and the entire cell cycle. These data sets are publicly available. The Brown lab also has an online guide to build your own arrayed and scanner. These micro arrays have been commercialized by Incyte pharmaceutical's microarray division(formerly Synteni). Incyte Gene expression Microarrays(GEMs) are available with templates from human, rat, mouse, plant and microbial genomes.

Different approaches have recently been used on outcome prediction using gene expression profiles. In the Cox proportional hazard regression method [8,9] genes most related to survival are first identified by a univariate Cox analysis, and a risk score is then defined as a linear weighted combination of the expression values of the identified genes[10,11].

Advances in techniques for high throughput data gathering, such as microarray and DNA sequencing machine have opened up new research avenues in genomics. Large-scale biological research such as genome projects are now producing enormous quantities of genomic data using these rapidly growing technologies. Transforming the massive data to useful biological knowledge is the present challenge. Different analysis tools are being developed in order to detect and understand the phenomena of gene regulation and physiological functions and assessing the quality of a genomic sequence [12].

With the wealth of gene expression data from microarray (such as high density oligonucleotide arrays and CDNA arrays) prediction, classification and clustering techniques are used for analysis and interpretation of the data. Some important recent applications are in molecular classification of acute leukemia(Golub et al.,1999,[14]), cluster analysis of tumor and normal colon tissues (Alon et al.,1999,[30]). Clustering and classification of human cancer cell lines (Ross et al.,2000,[66]). Diffuse Large B-cell lymphoma(DLBCL; Alizadesh et al., 2000, [6]), human mammary epithelial cells and breast cancer (Perou et al., 1999 [67], 2000 [68]) and skin cancer melanoma(Bittner et al.,2000[69]). These techniques have also helped to identify previously undetected sub types of cancer (Gloub et al.,1999[14];Alizadeh et al., 2000[6];Bittner et al.,2000[69];Perou et al.,2000[68]). The problem of 'prediction' may come in various forms of applications as well; the prediction of patient survival duration with germinal center B-like DLBCL compared to those with activated B-Like DLBCL using Kaplan Meier survival curves (Ross et al., 2000 [66]).

Gene expression data from DNA microarrays are characterized by many measured variables (genes) on only a few observations (experiments) although both the number of experiments and genes per experiments are growing rapidly [28].

Recent technical and analytical advances make it practical to quantitative the expression of thousands of genes in parallel using complementary DNA microarrays [3]. This mode of analysis has been used to observe gene expression variation in a variety of human tumors [29-30]. To apply this method to questions in normal and malignant lymphocyte biology, we designed a specialized microarray – the 'lympho chip'- by selecting genes that are preferentially expressed in lympho id cells and genes with known or suspected roles in processes important in immunology or cancer[31].

Due to recent advances in DNA microarray technology, is now feasible to obtain gene expression profiles of tissue samples at relatively low costs. Many scientists around the world use the advantages of this gene profiling to characterize complex biological circumstances and diseases microarray techniques that are used in genome wide gene expression and genome mutation analysis help scientists and physicians in understanding of the patho physiological mechanisms, in diagnoses and prognoses, and choosing treatment plans B. Transcriptional profiling is a tool that provides unique data about disease mechanisms, regulatory pathways, and gene function[3]. This technology not only allows comparison of gene profiles in normal and pathological tissues or cells, but also helps us establish interrelationships among genes, e.g.. Clustering of genes, coincident temporal pattern of expression, identify upstream and downstream targets of genes, understand mechanisms of disease at a molecular level, and define and validate novel drug targets[32].

A Comprehensive review of biological and technological aspects of microarray technology can be found in [33]. Ramaswamy et.al.[34] and Alizadeh et al., provide[35] a detailed discussion of the clinical implications of microarray in oncology. For excellent reviews on many different aspects of microarray technology, the reader is referred to the two special supplements [36-37]. References [38-39] provide an overview of gene expression data analysis. Topics covered include experimental design tissues, normalization, quality control, exploratory analysis (data visualization), and the problem of multiple testing for determining the differentially expressed genes. Aittoakallio et al [40] and Quackembush [41] underlined that the methods used to analyze the gene expression data can have a profound influence on the interpretation of the results and therefore a basic understanding of bio informatics tools is required for optimal experimental design and meaningful data analysis.

Availability of gene expression profiles of tissue samples from different diagnostic classes led to the application of many well- established pattern recognition / classification algorithms to these profiles, in an attempt to provide more accurate and automatic class prediction [35,39,6, 42].

Brazma et al.,[43] and Ball et al.,[44] discussed the importance of establishing a standard for recording and reporting microarray-based gene expression data and proposed a minimum information about a Microarray Experiment (MIAME) that describes the minimum information required to ensure that micro array data can be easily interpreted and that results desired from its analysis can be independently verified. Kuo et al.,[45] compared two high- throughput CDNA microarray technologies, stand ford type (i.e.. spotted) CDNA microarrays and Affymetrix oligonucleotide microarrays and showed that corresponding mRNA measurements from the two platforms showed poor correlation. Further their results suggest gene-specific, or more precisely, probe-specific factors influencing measurements differently in the two platforms, implying a poor prognosis for a broad utilization of gene expression measurements across platforms.

By measuring transcription levels of genes in an organism under various conditions, at different developmental stages and in different tissues, we can build up 'gene expression profiles' which characterize the dynamic functioning of each gene in the genome. We can imagine the expression data represented in a matrix with rows representing genes, columns representing samples and each cell containing a number characterizing the expression level of the particular gene in the particular sample we will call such a table a gene expression matrix. Building up a database of such matrices will help us to understand gene regulation, metabolic and signaling pathways, the genetic mechanisms of disease, and the response to drug treatments. For instance, if over expression of certain genes are correlated with a certain cancer, we can explore the conditions that affect

the expression of these genes and the genes that have similar expression profiles, we can also investigate which compound (particular drugs) lower the expression level of these genes[46].

III. Data mining classification technique for gene expression data

In data mining classification is one of the most important tasks. It maps the data into predefined targets. It is a supervised learning as targets are predefined. The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then the classifier is used to predict the group of new cases from the domain based on the value of other attributes.

The systematic classification of types of tumors is crucial to achieve advances in cancer treatment and research. It has been suggested that the specification of therapies according to tumor types differentiated by pathogenetic patterns may maximize the efficiency of the treatment and minimize toxicity on the patients [14, 6]. Several limitations about the conventional classification techniques based on morphological features of the tumor have been reported in the literature [15]. Moreover, by analyzing complex patterns defined by molecular markers, it has been demonstrated that there are subtypes of acute leukemia, prostate cancer and non-Hodgkin's-Lymphomas[14].

There are two useful tasks in cancer classification, prediction of classes and discovery of classes. The prediction task consists of the assignment of particular tumor samples to known types of cancer. The discovery task refers to the unsupervised identification of relevant groups of samples and the characterization of subtypes of cancer. Their research aims to implement a discovery task based on a global expression analysis approach.

Most approaches to the computational analysis of gene expression data are functionally significant classification of genes in unsupervised fashion and the discrimination of high risk patients from low risk ones. On the other hand, supervised learning techniques use training set to optimize the discrimination model. Artificial Neural Network (ANN) is one of the supervised methods and a powerful tool for accurately detecting causal relationships [13].

Tamayo et.at, have illustrated the value of Kohonen's self-organizing feature maps (SOFM) [16] to interpret gene expression patterns during yeast growth cycle and hematopoietic differentiation [17]. They identify predominant gene expression patterns in those biological processes that suggested, for instance, novel hypotheses about hematopoietic differentiation useful for the treatment of acute promyelocytic leukaemia. Similarly based on a SOFM, Golub et al. [14] approaches the problem of molecular classification of cancer.

Classification of biomedical data faces a special challenge because of the characteristics of the data: too few data examples with too many features. How to improve the classification performance or the generalization ability of a classifier in the biomedical domain becomes one of the active research areas. One approach is to build a fusion model to combine multiple classifiers together and result in a combined classifier which can achieve a better performance than any of its composing individual classifiers [12]. [18] proposed a sum classifier fusion model to combine multiple SVMs by applying the knowledge of fuzzy logic and genetic algorithms.

The most straight fordward classifier design approach is based on the concept of similarity. In this approach, the distance between the test patterns whose class is to be decided and the known representatives or prototypes of classes are measured. Given a training set and a similarity measure or metric, to decide for the class membership of a test sample, the k-nearest neighbors (k-NN) find the class membership of the k closet samples in the training set and take a majority vote. The k-NN classifier that assigns the test samples to the class of nearest observations in the training set is often used as a benchmark for other classifiers, since it always offers reasonable classification performance [47].

In the nearest mean classifier, the prototypes are the class means / centres or centroids. Tibshirani et al., [48] suggested an enhancement for the nearest centroid classifier, called Nearest Shrunken Centroids(NSC) (The NSC is also referred to as PAM. Prediction Analysis of Microarrays, due to the name of the associated paper and software). In NSC, weak components of the class- centroids are shrunk or deleted via soft-thresholding. The classification accuracy (expressed in terms of training test, and cross validation error rates) and the number of present (or undeleted) genes are plotted against a parameter called delta that adjusts the amount of shrinkage and an optimal value for delta is selected by examining the error rates shrinkage eliminates the information that does not contribute towards class prediction, i.e noise,. The contribution or strength of each class centroid to the classification is measured by a t- statistics, where the numerator is the difference between individual class means and the overall mean and the denominator is the pooled estimate of standard deviation inflated by a fudge factor.

Another popular classifier design approach is based on Artificial Neural Networks. NN consists of many interconnected processing elements, called neurons, resembling human brain's structure through different structures (varying number of layers and number of neurons per layer) linear or non linear transfer functions that the individual neurons use, and training paradigms during which the weights of the connections are adjusted or tuned, the NN can model / reveal complex relationship among inputs and outputs exemplified or embedded in the training data [32].

Other popular classifier design approaches include Fisher's Linear Discriminant Analysis (FLDA). The FLDA is both a class-predictor design and a feature extraction /selection approach, or expressed differently, FLDA is a classifier design approach with built in feature extraction / selection capability. A linear discriminant function is nothing but a special linear combination of the values of all the features that are used in classifier design.

In order to introduce other major classifier design approaches such as DLDA and DQDA that are frequently used in gene expression profile classification, we will briefly review the Bayesian decision theory. In this model based setting, the class conditional densities are assumed to have multivariate normal densities typically. In Classification and Regression Trees approach, Breimon et al., [49], used node impurity criteria such as entropy (information content) and Gini's index of diversity [51]. Some important features are selected and binary splits are formed on those features repeatedly. Each terminal features subset is associated with a class label. Dudoit et al.,[50] identified three main aspects for tree construction, selection of splits, decision to declare a node terminal or to continue splitting, and assignment of each terminal node to a class. Depending on how these topics are treated, many variations of tree are possible. Since the decisions / splits at nodes are binary, decision boundaries are parallel to the features axes, as such they are intrinsically suboptimal [47].

A common way to represent gene expression measurements does not only allow to directly combine microarray data sets, but also to readily apply the generated classifier on a new data set which is represented in the same manner. To this end [52, 53] proposed the method TSP (Top Scoring Pair) and [54] the generalized version kTSP(k-Top Scoring Pairs), classifiers which directly refer to the relative ranks, i.e the ordering of the actual gene expression value with in a profile kTSP was shown to perform as good as state –of- the –art algorithms while using a relatively small number of genes for classification.

Machine learning techniques such as neural networks are adequate for gene expression patterns and cancer classification analysis for their well-known pattern recognition and data organization capabilities [55],[56]. Advanced neural learning algorithms have not only improved the accuracy, reliability and efficiency of many medical pattern recognition systems, but they also show several advantages for the implementation of decision support systems in physiological genomics [57] [58].

IV. Data mining clustering technique for gene expression data

Clustering problems arise in many different applications such as data mining and knowledge discovery, data compression, pattern recognition and pattern classification in order to grouping similar genes in one cluster so that genes within the same cluster are similar to each other and different from genes in other cluster [19].

Clustering techniques have proven to be helpful to understand gene function, gene regulation, cellular processes, and sub types of cells. Genes with similar expression patterns (co expressed genes) can be clustered together with similar cellular function [5].

The purpose of clustering gene expression data is to reveal the nature structure inherent in the data. A good clustering algorithm should depend as little as possible on prior knowledge, for example requiring the predetermined number of cluster as an input parameter. Clustering algorithms for gene expression data should be capable of extracting useful information from noisy data. Gene expression data are often high connected and may have intersecting and embedded patterns [20]. Clustering algorithm which also provides some graphical representation of the cluster structure is much favored by biologists.

There are numerous clustering techniques presently available to cluster particularly the gene expression data such as hierarchical clustering technique which is a method used commonly by many people in early days. A common problem associated with this method is visualization of clustering results in terms of dendrogram which is difficult when a data set is large [21]. In the popular – k-means clustering method, the user was always uncertain to define the precise number of clusters. In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In some situations, the object may belong to more than one cluster, and associated with each element is a set membership level. Clustering may be either crisp (or) fuzzy [22]. Fuzzy clustering of microarray data has an advantage over crisp partitioning because of great amount of imprecision and uncertainty related with gene expression data [23].

Fuzzy c- means [24] and genetic algorithms (GA) [25],[26] have been used effectively in clustering gene expression data. The fuzzy c-means algorithm requires the number of clusters as an input parameter. The GA based algorithms have been found to detect biologically relevant clusters but are dependent on proper tuning of the input parameters.

[27] have presented a framework for the unsupervised analysis gene expression data. They developed an interrelated two-way clustering method which they applied on the gene expression matrices transformed from the new microarray data. This approach detects significant patterns within samples while dynamically selecting significant genes which manifest the conditions of actual empirical interest. Through iterative clustering the number of genes are reduced which improves the accuracy of sample class discovery. The method was proved effective by conducting experiment with two multiple sclerosis data sets and a leukemia data set. These

experiments indicate that this appears to be a promising approach for unsupervised sample clustering on gene array data sets.

The goal of clustering is to group together objects (gens or samples) with similar properties. This can also be viewed as the reduction of the dimensionality of the system. Clustering is not a new technique, many algorithms have been developed for it and many of these algorithms have been applied to analyze expression data. The hierarchical [59] and k-mean clustering algorithms [60] and [61] as well as self-organizing maps [62] have all been used for clustering expression profiles. Even a simple clustering algorithms based on binning (i.e. discrete `zing the expression profile space and clustering together the profiles that map into the same bin) has been shown to be useful for clustering genes and subsequent discovering of transcription factor binding sites[63]. More recently new algorithms have been developed specifically for gene expression profile clustering for instance based on finding approximate cliques in graphs [64].

Gene expression profile clustering does not necessarily require the full genome. For instance Iyer et.al.,[65] studied 8600 genes in human fibroblasts and obtained 10 distinct gene clusters each associated with genes with particular functional roles, such as signal transduction, coagulation, homeostasis, inflammation etc.

V. Conclusion

Gene expression profiling has great potential for accurate cancer diagnosis. In this paper, we have discussed different types of advances in techniques for high throughput data gathering such as microarrays and DNA sequencing machine that have opened up new research in genomics . Large-scale biological research such as genome projects are now producing enormous quantities of genomic data using these rapidly growing technologies. Different analysis tools are developed in order to detect and understand the phenomena of gene regulation and physiological functions and assessing the quality of a genomic sequence.

References

[1] M.B Eisen, P.T.Spellman,P.O.Brown, and D. Botstein, *Cluster analysis and display of genome- wide expression patterns*, (Proc,Natl.Acad.Sci. USA), *Vol.95*,pp.14863-8, (1998).

[2] P.J.Russel,*Fundamentals of genetics*, Second Edition, (San Francisco, Addison Wesly Longman Inc., 2000).

[3] M.Schena,D.Shalon, R.W.Davis and P.O.Brown, *Quantitative monitoring of gene expression patterns with a complementary DNA Micro Array Science*,270, 476-471, 1995.

[4] M.B.Eisen and P.O.Brown, *DNA arrays for analysis of gene expression*, Methods Enzymol.,303,179-205,(1999).

[5] 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, pp 1370-1384

[6] A.A.Alizadeh et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling Nature*, 403, 503-511,(2000).

[7] D'haeselecr, Shoudan Liang and Roland Somogyi, PsB99 Tutorial *Gene Expression Analysis and Modeling*.

[8] Cox, *D.R.Regression models and life-tables (with discussion)*,J.R.Stat Soc.,B34:184-220,(1972).

[9] Lunn,M., & McNeil,D.R.., *Applying Cox Regression to Competing Risks*, Biometrics 51: 524-532,(1995).

[10] Beer D.G.Kardia, S.L.,Huang, C.C.Giodano, T.J..,Levin, A.M.Misek, D.E.,Lin, L.,Chen.G.Gharib,T.G.,Thomas, D.G.,Lizyness, M.L.,Kuick, R.,Hayasaka, S.,Taylor, J.M.,Iannettoni, M.D.,Irringer, M.B&Hanash,S., *Gene Expression Profiles predict survival of patients with lung adenocarcinoma*, Nat. Med.,8(8): 816-823,2002.

[11] Rosenwald,A.et al, *The use of molecular profiling to predict survival after chemotherapy for Diffuse large-B cell lymphoma*, NEJM,346(25):1937-1947,2002.

[12] R. Radha., Gene Expression Analysis, *International Journal of Advanced Science and Technology*, Vol.33, August 2011.

[13] Khan, J. et al., *Classification and diagnostic Prediction of Cancers using gene expression profiling and artificial Networks*, Nat.Med., 7:673-679,2001.

[14] T.R.Golub,D.K.Slonim , P.Tamay, C.Huard, M.Gassembeek, J.P. Mesirov, H.Coller, M.L Loh, J.R.Downing, M.A.Caligiuri, C.D. Bloomfield and E.S. Lander, *Molecular classification of Cancer: class discovery and class prediction by Gene expression monitoring, science*, 286, 531-537, 1999.

[15] F.Azuaje, Interpretation of genome expression patterns: Computational challenges and opportunities, to be published by *IEEE Engineering in Medicine and Biology*, November 2000.

[16] T.Kohonem, *Self-organizing Maps*, (Heidelberg, Springer, 1995).

[17] P.Tamayo, D.Slonim, J.Mesirov, Qzhn, S.Kitareewan, E.Dmistrovsky, E.lander and T.R.Golub, *Interpreting Patterns of gene expression with self – organizing maps: methods and applications to hematopoietic differentiation*, The Proceedings of the National Academy of Sciences of U.S.A.,96,2907-2912,(1999).

[18]. Xiujuan chen, Yong Li, Robert Harrison, Yan-Qing Zhang, *Genetic fuzzy classification fusion of multiple SVMs for biomedical data* journal of intelligent & Fuzzy systems. Volume 18, issue 6, December 2007, IOS press Amsterdam.

[19]. Han, Kamber,*Data Mining Concepts and Techniques*, (Elsevier Publications, 2006).

[20]. D.jiang, J.pei, and A.Zhang, *DHC:* a density – based hierarchical clustering method for time series gene expression data. In Proceedings of BIBE2002,:*3rd IEEE International Symposium on Bio-informatics and Bio-Engineering*. Bethesda Maryland 2003, p.393.

[21]. Anil K.Jain and Richard C.Dubes, *Alogrithms for clustering data*, (Prentice Hall,New Jersey, 1988).

[22]. P. Valarmathie, Dr. MV. Srinath, Dr.T.Ravichandran, K.Dinakaran, 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, October 2009.

[23]. Anirban, Mukhopadhayay, Ujjuval Maulik and sanghamitra bandyopadhyay, *Efficient two stage fuzzy clustering of microarray gene expression data*, International Conference on information Technology (ICIT'06) , 2006 IEEE.

[24]. J.C. Bezdek, *Pattern Recognintion With Fuzzy Objective Function Algorithms*, (New York;Plenum Press, 1981).

[25]. S. Bandyopadhyay, A.Mukhopadhyay, and U.Maulik, "*An important algorithm for clustering gene expression data",Bioinformatics*, vol.23(21),pp. 2859-2865,2007.

[26]. U. Maulik, A.Mukhopadhyay, and S.Bandyopadhyay, "*Combining Pareto optimal clusters using supervised learning for identifying co-expressed genes*", BMC Bioinformatics. Vol.co(27),2009.

[27]. C. Tang and A.Zhang, 'Interrelated Two-Way Clustering and its Application on Gene Expression Data', Presented at *International Journal on Artificial Intelligence Tools*,2005, p.p.577-598.

[28]. Danh V.Nguyen[1] and David M. Rocke[2].*Tumor Classification by Partial least squares using Microarray Gene expression data.*, 1. Center for Image Processing and Integrated Computing and 2. Department Applied Science. University of California, Davis, CA 95616, USA, Received on November 23, 2000; revised on March 22, 2001; accepted on June 6, 2001.

[29]. Bubendrof. L. et al. Hormone Theraphy failure in human prostate Cancer: *analysis by complementary DNA and tissue Microarrays.* J.Natl Cancer Inst. Al., 1758-1764 (1999).

[30]. Alon, U et al. *Broad Patterns of Gene expression revealed by clustering analysis of tumor and normal colon tissues probed by Olignonucleotide arrays*. Proc.Natt.Acad Sci USA 96, 6745-6750(1999).

[31]. Alizadeh.A.et. al. The Lymphochip: *a specialized cDNA Microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes*. Cold Spring Harbor Symp. Quant-Biol.(in the press).

[32]. Mush H.Asyali, Dilek Colak, Omer Demirkaya and Mehmet S. Inan. *Gene Expression Profile Classification : A Review", Current Bioinformatics*, 2006, l, 55-73. Bentham Science Publishers Ltd.

[33]. Nguyen DV, Arpat AB, Wang N, Carroll RJ. DNA Microarray experiments : *biological and technological aspects Biometrics* 2002, 58; 701-17.

[34]. Ramaswamy S. Golub TR, *DNA Microarray in clinical oncology*, J.Clin Oncel 2002:20:1932-41.

[35]. Alizadeh AA, Ross DT, Perou CM, Rijin Mud, *Towards a novel classification of human malignancies based on gene expression patterns* J.pathol 2000:195;41-52.

[36]. *Nature Genetics, The chipping forecast*. Vol.(21) Supplement 1999. 1-60.

[37]. *Nature Genetics, The chipping forecast*. 2002: 461-552.

[38]. Loung YF, Cavalieri D. *Fundamentals of CDNA Microarray data analysis Trends Genetic* 2003;19-649-59.

[39]. Lu.Y.Han,J. *Cancer classification using gene expression data* Int Syst 2003;28:243-68.

[40]. Aittokallio T. Kurki M.Nevalainen O.Nikula T. West, A. Lahesmaa R. *Computational strategies for analyzing data in gene expression microarray experiments*. J.Bioinforma comput Bio, 2003;1;541-86.

[41]. Quacknenbush J. *Computational analysis of microarray data*. Nat. Rev. Genet 2001 : 2: 418-27.

[42]. Zhu J. Hastie T. *Classification of gene microarrays by penalized logistic regression*, Biostatistics;2004;5:427-43.

[43]. Brazma A. Hingamp P. Quackenbush J, et al., *Minimum information about a Microarray experiment (MIAME) – toward standards for microarray data*, NatGenet 2001:29:365-71.

[44]. Ball CA. Sherlock, G.Parkinson H et al, *standards for microarray data minimum information about a microarray data*, Science 2002, 298:539.

[45]. KuoWP. Jenssen Tk. Butte AJ, Ohno-Machado L.Kohane Is. *Analysis of matched mRNA measurements from two different micro array technologies, Bioinformatics* 2002, 18:405-12.

[46]. Gianni Cesareni ., Alvis Brazma, Jaak Vilo edited ,the *Gene expression data analysis*, European Molecular Biology laboratory, Outsation Hinxton – the European Bioinformatics Institute, Cambridge CB10, ISD, UK. Received 5 June 2000.

[47]. Jain A, Duin P, Mao J, statistical Pattern recognition : A review, *IEEE Transactions on PAMI 2000*;22:4-37.

[48]. Tibshirani R, Hastc, T. Narasimhan B. Chu G. *Diagnosis of multiple cancer types by Shrunken centroids of gene expression* ProcNAtl Acad Sci USA 2002: 99:6567-72.

[49]. Breiman L, Friedman JH, Olshen R, Stone C J. *Classification and Regression Trees*, Wadsworth, Bel mont, CA 1984.

[50]. Dudoit S, Fridlyand J, Speed TP *Comparison of discrimination methods for the classification of tumors using gene expression data*. J AM stat Assoc 2002;77-87.

[51]. Kulkarni SR, Lugosi G. Venkatesh SS learning Pattern Classification a survey. *IEEE Trans in from Theory 1998*;44:2178-206.

[52]. Geman D, d'Avignen C, Naiman DQ, Winslow RL; *Classifying gene expression Profiles from Pairwise mRNA Comparisons*. Stat Appl Genet Mol Biol 2004, 3: Article 19.

[53]. Xu.L, Tan Ac, Naiman DQ, Geman D, Winslow RL: *Robust Prostate Cancer marker genes emerge from direct integration of inter-study microarray data*. Bioinformatics 2005, 21(20):3905-11.

[54]. Tan AC, Naiman DQ, Xu,L, Winslow RL, Geman D; *simple decision rules for classifying human cancers from gene expression profiles*. Bioinformatics 2005, 21(20): 3896-904.

[55]. B. Ripley, *Pattern Recognition and Neural Networks* ,(Cambridge, England, Cambridge university press, 1996).

[56]. F. Azuaje, W. Dubitzky, P. Lopes, N. Black, K.Adamson, x. Wu and J. White, *Predicting Coronary disease risk based on short – Term RR Intervals Measurements; A Neural Network Approach, Artificial Intelligence in Medicine*, 15,275-298,1999.

[57]. F. Azuaje, Making Genome Expression Data Meaningful: Prediction and discovery of classes of cancer through a connectionist learning approach, to be published in the proceedings of the *IEEE symposium on Bioinformatics and Biomedical Engineering* (BIBE – 2000).

[58]. F. Azuaje, W. Dubitzky, N.Black and K.Adamson, discovering Relevance Knowledge in Data: *A Growing cell Structure Approach, IEEE Transactions on systems, Man and Cybernetics*, Part B, 30, 448-460,2000.

[59]. M. Eisen, P.T.Spellman, D. Botstein, P.O.Brown Proc. Natl. Acad. Sci. USA, 95(1998),pp. 14863-14867.

[60]. Hartigan, J.A.(1975) *Clustering Algorithms*, (John wiley and Sons, New York).

[61]. S. Tavazoie, D. Hughes, M.J.Campbell,R.J.Cho, G.M.Church, Nature Genet, 22(1999), pp.281 - 285 *View Record in Scopus / ci ted By in Scopus(1295)*.

[62]. P.Tamayo, D. Slonim , J. Mesirov, G.Zhu, S.Kitareewan, E.Dmitrovsky, E.Lander, T.Golub Proc, Natl.Acad. Sci.USA, 96(1999),pp. 2907-2912 *View Record in Scopus / cited By in Scopus (1702)*.

[63]. A. Brazma, I. Jonassen, J. Vilo, E. Ukkonen Genome Res., 8(1998), pp. 1202-1215. *View Record in Scopus / cited by in Scopus (169).*

[64]. Ben – Dor, A and Yakhini, Z. (1999) *Proceedings of the Third Annual International Conference on Computational Molecular Biology RECOMB – 1999*, pp. 33-42,(ACM Press , Lyon.)

[65]. V.R.Iyer, M.B.Eisen,D.T.Ross, G.Schuler, T. Moore, J.C.F.Lee, J.M.Trent, L.M.Staudt, J. Hudson Jr., M.S. Boguski, D. Lashkari, D. Shalon, D. Botstein, P.O.Brown Science, 283(1999), pp. 83-87.

[66]. Ross,D.T., (2000) *Systematic Variation in gene expression patterns in human cancer cell lines*, Nature Genet., 24, 227-235.

[67]. Perou,C.M., (1999) *Distinctive gene expression patterns in human mammary epithelial cells and breast cancer*, Proc. Natl Acad. Sci. USA, 96, 9112-9217.

[68]. Perou,C.M., (2000) *Molecular Portrait of human breast tumors*. Nature, 406, 747-752.

[69]. Bittner, M., (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 406, 536-540.