Application of Microarray Technology and Softcomputing in Cancer Biology : A Review

P.K.Vaishali vaishali5599@gmail.com
Department of Computer Science & Information Technology,
Jyothishmathi Institute of Technology & Science,
JNTU, Hyderabad, AP, INDIA.

Dr.A.Vinayababu dravinayababu@yahoo.com
Professor of CSE, Director of Admissions
JNTUH University,
Hyderabad, AP, INDIA.

Abstract

DNA microarray technology has emerged as a boon to the scientific community in understanding the growth and development of life as well as in widening their knowledge in exploring the genetic causes of anomalies occurring in the working of the human body. microarray technology makes biologists be capable of monitoring expression of thousands of genes in a single experiment on a small chip. Extracting useful knowledge and info from these microarray has attracted the attention of many biologists and computer scientists. Knowledge engineering has revolutionalized the way in which the medical data is being looked at. Soft computing is a branch of computer science capable of analyzing complex medical data. Advances in the area of microarray –based expression analysis have led to the promise of cancer diagnosis using new molecular based approaches. Many studies and methodologies have come up which analyszes the gene espression data by using the techniques in data mining such as feature selection, classification, clustering etc. emboiding the soft computing methods for more accuracy. This review is an attempt to look at the recent advances in cancer research with DNA microarray technology , data mining and soft computing techniques.

Keywords: DNA Microarray, Classification, Data Mining ,Soft Computing ,Gene Expression.

1. INTRODUCTION

Deoxyribonucleic acid (DNA) micro array technology provides tools for studying the expression levels of a large number of distinct genes simultaneously [9]. Micro array technology allows biologists to simultaneously measure the expressions of thousands of genes in a single experiment [8] [10] [11].

Gene expression data is widely used in disease analysis and cancer diagnosis [5]. Gene expression data from DNA micro arrays are characterized by many measured variables (genes) on only a few observations (experiments) although both the number of experiments and genes per experiment are growing rapidly [4] [6]. Gene expression data from DNA micro array can be characterized by many variables (genes), but with only a few observations (experiments). Prediction, classification, and clustering techniques are being used for analysis and interpretation of the data [1]. An important application of gene expression micro array data is classification of biological samples or prediction of clinical and other outcomes [2]. Micro array technology is to classify the tissue samples using their gene expression profiles as one of the several types (or subtypes) of cancer. Compared with the standard histopathological tests, the gene expression profiles measured through micro array technology provide accurate, reliable and objective cancer classification. The DNA micro array data for cancer classification consists of large number of genes (dimensions) compared to the number of samples or feature vectors [3] [7]. Classification analysis of micro array gene expression data has been widely used to uncover biological features and to distinguish closely related cell types that often appear in the diagnosis of cancer [37].Many researchers have developed and demonstrated different classification techniques for cancer classification based on micro array gene expression data. Feature selection techniques [12],[13] have been suggested before classification, which finds the top features that discriminate various classes. Kernel based techniques [14],[15] like SVM have already been used for binary disease classification problems. Gene selection[16] and neural networks[17] based classifications were also reported in microarray data analysis. soft computing has been successively used in bioinformatics thereby providing low cost, low ,better approximation and indeed good and more accurate solutions.

2. DNA MICROARRAY TECHNOLOGY

Although all of the cells in the human body contain the same genetic material, the same genes are not active in all of those cells. Studying which genes are active and which are inactive in different kinds of cells helps scientists understand more about how these cells function and about what happens when the genes in a cell don’t function properly. In the past scientists have only been able to conduct such genetic analyses on a few genes at once. With the development of DNA microarray technology, however, scientists can now examine thousands of genes at the same time, an advance that will help them determine the complex relationships between individual genes. The mountain of information that is the draft sequence of the human genome may be impressive, but without interpretation that is all it remains — a mass of data. Gene function is one of the key elements researchers want to extract from the sequence, and the DNA microarray is one of the most important tools at their disposal.Microarray technology will help researchers learn more about many different diseases—heart disease, mental illness, and infectious disease, to name only a few. One intense area of microarray research at the NIH is the study of cancer.In the past, scientists have classified different types of cancer based on the organs in which the tumors develop. With the help of microarray technology, however, they will be able to further classify these types of cancer based on the patterns of gene activity in the tumor cells and will then be able to design treatment strategies targeted directly to each specific type of cancer. Additionally, by examining the differences in gene activity between untreated and treated—radiated or oxygen-starved, for example—tumor cells, scientists can better understand how different types of cancer therapies affect tumors and can develop more effective treatments.
In short the usefulness of dna technology can be listed as
1.Can follow the activity of MANY genes at the same time.
2.Can get a lot of results fast
3.Can COMPARE the activity of many genes in diseased and healthy cells
4.Can categorize diseases into subgroups

Microarray Technology: Application in medical

Gene discovery: DNA Microarray technology helps in the identification of new genes, know about their functioning and expression levels under different conditions.

Disease Diagnosis: DNA Microarray technology helps researchers learn more about different diseases such as heart diseases, mental illness, infectious disease and especially the study of cancer. Until recently, different types of cancer have been classified on thedevelop. Now, with the evolution of microarray technology, it will be possible for the researchers to further classify the types of cancer on the basis of the patterns of gene activity in the tumor cells. This will tremendously help the pharmaceutical community to develop more effective drugs as the treatment strategies will be targeted directly to the specific type of cancer.

Drug Discovery: Microarray technology has extensive application in *Pharmacogenomics.* Pharmacogenomics is the study of correlations between therapeutic responses to drugs and the genetic profiles of the patients. Comparative analysis of the genes from a diseased and a normal cell will help the identification of the biochemical constitution of the proteins synthesized by the diseased genes. The researchers can use this information to synthesize drugs which combat with these proteins and reduce their effect.

Toxicological Research: Microarray technology provides a robust platform for the research of the impact of toxins on the cells and their passing on to the progeny. Toxicogenomics establishes correlation between responses to toxicants and the changes in the genetic profiles of the cells exposed to such toxicants

3. DATAMINING AND SOFT COMPUTING PARADIGM IN THE AREA OF GENE EXPRESSION IN CANCER DATA - RECENT RELATED RESEARCH : A REVIEW

There exists considerable literature on the application of different soft computing paradigm in the area of gene expression cancer data sets:One of the first landmark studies using microarray data to analyze tumor samples was done by Golub *et al.* [18] . This study on human acute leukemia showed that it was possible to use microarray data to distinguish between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without any previous knowledge. For the first time the potential of microarray data was shown by illustrating its use in discovering new classes using the previously introduced class discovery methods (i.e., unsupervised analysis) and, second, by using microarray data to assign tumors to known classes (i.e., supervised analysis).

Perou *et al.* [19] did a similar analysis using hierarchical clustering on breast cancer and found different groups of breast tumors.

Alon *et al.* [20] adopted a two-way clustering method whereby both genes and tumors were clustered. They showed that colon tumors and normal colon tissues were separated based on the microarray data. Also, they showed that co-regulated families of genes also clustered together.

Tibshirani *et al*. [21] built further on the development of supervised microarray data analysis methods by developing the nearest shrunken centroid method, also known as PAM. This technique not only allows predicting of classes, but also tries to limit the number of genes necessary to make the prediction. By limiting the number of genes, it is possible to develop cheaper methods to make a diagnostic test, such as smaller microarrays or quantitative PCR. After these studies research groups focused more on class.

Ahmad M. Sarhan [22] has presented a stomach cancer detection system based on Artificial Neural Network (ANN), and the Discrete Cosine Transform (DCT). The proposed system has extracted classification features from stomach micro arrays using the DCT. The features extracted from the DCT coefficients have been then applied to an ANN for classification (tumor or non-tumor). The micro array images used in this study have been obtained from the Stanford Medical Database (SMD). Simulation results have shown that the proposed system produces a very high success rate.

Bharathi *et al*. [23] have attempted to find the smallest set of genes that can ensure highly accurate classification of cancer from micro array data by using supervised machine learning algorithms. The significance of finding the minimum gene subset has been three fold:1) It has greatly reduced the computational burden and noise arising from irrelevant genes.2) It has simplified gene expression tests to include only a very small number of genes rather than thousands of genes, which could significantly bring down the cost for cancer testing. 3) It has called for further investigation into the possible biological relationship between these small numbers of genes and cancer development and treatment. Their simple yet very effective method has involved two steps. In the first step, they have chosen some important genes using a 2 way Analysis of Variance (ANOVA) ranking scheme. In the second step, they have tested the classification capability of all simple combinations of those important genes using a good classifier such as Support Vector Machines. Their approach has obtained very high accuracy with only two genes.

Bo Li *et al*. [24] have discussed that the gene expression data collected from DNA micro array are characterized by a large amount of variables (genes), but with only a small amount of observations (experiments). They have proposed a manifold learning method to map the gene expression data to a low dimensional space, and then explore the intrinsic structure of the features so as to classify the micro array data more accurately. The proposed algorithm could project the gene expression data into a subspace with high intra-class compactness and inter-class separability. Experimental results on six DNA micro array datasets have demonstrated that their method is efficient for discriminant feature extraction and gene expression data classification. Their work is a meaningful attempt to analyze micro array data using manifold learning method; there should be much room for the application of manifold learning to bioinformatics due to its performance.

Xiaosheng Wang *et al*. [25] have discussed that a gene selection is of vital importance in molecular classification of cancer using high-dimensional gene expression data. Because of the distinct characteristics inherent to specific cancerous gene expression profiles, developing flexible and robust feature selection methods has been extremely crucial. They have investigated the properties of one feature selection approach proposed in their previous work, which has been the generalization of the feature selection method based on the depended degree of attribute in rough sets. They have compared the feature selection method with the established methods with respect to the depended degree, chi-square, information gain, Relief-F and symmetric uncertainty, and analyzed its properties through a series of classification experiments. The results have revealed that their method is superior to the canonical depended degree of attribute based method in robustness and applicability. Moreover, their method has been comparable with the other four commonly used methods. More importantly, their method could exhibit the inherent classification difficulty with respect to different gene expression datasets, indicating the inherent biology of specific cancers.

Mallika et al. [26] have presented a novel method for improving classification performance for cancer classification with very few micro array Gene expression data. The method employs classification with individual gene ranking and gene subset ranking. For selection and classification, the proposed method has used the same classifier. The method has been applied to three publicly available cancer gene expression datasets from Lymphoma, Liver and Leukaemia datasets. Three different classifiers namely Support vector machines-one against all (SVM-OAA), K nearest neighbour (KNN) and Linear Discriminant analysis (LDA) have been tested and the results have indicated improvement in performance of SVM-OAA classifier with satisfactory results on all the three datasets when compared to the other two classifiers.

Chhanda Ray [27] has discussed DNA micro array gene expression patterns of several model organisms and provided a fascinating opportunity to explore important abnormal biological phenomena. The development of cancer has been a multi-step process in which several genes and other environmental and hormonal factors play an important role. They have proposed a new algorithm to analyze DNA micro array gene expression patterns efficiently for huge amount of DNA micro array data. For better visibility and understanding, experimental results of DNA micro array gene pattern analysis have been represented graphically. The shape of each graph corresponding to a DNA micro array gene expression pattern has been determined by using an eight-directional chain code sequence, which has been invariant to translation, scaling, and rotation. The cancer development has been identified based on the variations of DNA micro array gene expression patterns of the same organism by simultaneously monitoring the expressions of thousands of genes. At the end, classification of cancer genes has also been focused based on the distribution probability of codes of the eight-directional chain code sequences representing DNA micro array gene expression patterns and the experimental result has been provided.

While Chu *et al.* [31] [36]has used a five-genes set for 100% correct classification on the lymphoma data in the fuzzy NF framework, A dynamic fuzzy neural network, involving self-generation,parameter optimization, and rulebase simplification, is used [31][36] for the classification of cancer data such as lymphoma,liver cancer.

Banerjee *et al.* [35] [36]obtained a misclassification for just two samples from the test data using a two-genes set. In case of the leukemia data, a two-genes set is selected, whereas the colon data results in an eight-genes reduct size. An evolutionary rough feature selection algorithm [35][36] has been used for classifying microarray gene expression patterns. The effectiveness of the algorithm is demonstrated on three cancer datasets, viz., colon, lymphoma, and leukemia

S.Mitra et al. Has discussed An evolutionary rough *c*-means clustering algorithm applied to microarray gene expression data [28][36]. RSes are used to model the clusters in terms of upper and lower approximation.

S.Bicciato et.al. discussed An autoassociative neural network for *simultaneous* pattern identification, feature extraction, and classification of gene expression data [30] Results are demonstrated on leukemia and colon cancer20 datasets. The identification of gene subsets for classifying two-class

Modelling/data mining tasks	Result obtained	Reference

disease samples has been modeled as a multiobjective evolutionary optimization problem by K.Deb et.al. [32], involving minimization of gene subset size to achieve reliable and accurate classification based on their expression levels. Classification of acute leukemia, having highly similar appearance in gene expression data, has been made by combining a pair of classifiers trained with mutually exclusive features [29].

RSes have been applied mainly tomicroarray gene expression data, in mining tasks like classification [38], [33], H.Midelfart .et.al usedClassification rules (in *if–then* form) for extracting data from microarray data [38], using RSes with supervised learning. Gastric tumor classification in microarray data is made using rough set-based learning [33].

4. ASSEMBLANCE OF SOFT COMPUTING TECHNIQUES WITH MICROARRAY TECHNOLOGY IN CANCER BIOLOGY

The esemblence of soft computing techniques applied to Microarray Technology in Cancer Biology is described in table1. The table also gives the information of different datamining tasks involved.

5. CONCLUSION

Microarray technology technology has suppressed the conventional cancer diagnostic methods based on the morphological appearance of the cancerous cell which quiet often were misdiagnosed.

For more precision and effective results the emboiding of the different soft computing approaches is really recommendable.Based on the numerous publications investigating the use of DNA microarray technology ,data mining and soft computing to predict outcome in different cancer sites, this technology seems to be the most mature technology from all the omics.

Class Discovery methods.	Assigning tumors to known classes.	[34]
2-way Clustering	Both genes & tumors were clustered	[38]
Hirarchical clustering	Found different groups of Breast cancer.	[35]
Nearest Shrunken centroid method (PAM)	Limit on the number of genes necessary to prediction.	[39]
ANN & DCT	Very high success rate for classification of tumor and non tumors	[22]
Supervised machine learning	High accuracy with only two genes	[23]
Manifold learning method	Efficient discriminant feature extraction and gene expression data classification.	[24]
Rough sets ,feature selection	Superior in applicability and robustness.	[25]
Gene ranking and gene subset ranking	Improved classification performance	[26]
ANN,classification	Simultaneous pattern extraction,Leukemia classification	[29][30]
GA,classification	reliable and accurate classification based on their expression levels,minimization of gene subset size	[32]
NF,feature selection	Feature selection	[32]
Fuzzy NN(dynamic structure growing),feature selection. ANN,classifiers	Colon classification,Classification of acute leukemia, having highly similar appearance in gene expression data	[30][32]
RS+GA,clustering	effectiveness of the algorithm is demonstrated on three cancer datasets, viz., colon, lymphoma, and leukemia.	[28]
NF,self-generation, parameter optimization, and rulebase simplification,classification,feature selection.	Lymphoma classification.	[31]
NF,rule base simplification.	Classification of Small round blue cell tumor	[31]
NF,classification	Liver cancer100% correct classification on the lymphoma data	[31]
RS+GA,classification	Gastric tumor classification	[33]
GA(multi objective approach)	Lung cancer & mixed lineage leukemia more efficient resuls in gene selection as compared to single objective	[37]
GA ,SVM	Multiclass cancer categarization	[38]

TABLE 1: Use of Microarray Technology with Softcomputing in Cancer Research

REFERENCES

[1] Nguyen and Rocke, Classification of Acute Leukemia based on DNA Micro array Gene Expressions using Partial Least Squares, Kluwer Academic, Dordrecht, 2001

[2] Jian J. Dai, Linh Lieu, and David Rocke, "Dimension Reduction for Classification with Gene Expression Micro array Data", Statistical Applications in Genetics and Molecular Biology: Vol. 5, No. 1, 2006

[3] Alok Sharma and Kuldip K. Paliwal, "Cancer classification by gradient LDA technique using micro array gene expression data", Data & Knowledge Engineering, Vol. 66, pp. 338-347, 2008

[4] Chun-Hou Zheng, Bo Li, Lei Zhang and Hong-Qiang Wang, "Locally Linear Discriminant Embedding for Tumor Classification", In Proceedings of ICIC, pp.1093-1100, 2008

[5] Cheng-San Yang, Li-Yeh Chuang, Chao-Hsuan Ke and Cheng-Hong Yang, "A hybrid Feature Selection Method for Micro array Classification", International Journal of Computer Science, Vol. 35, No. 3, 2008

[6] Danh V. Nguyen, David M. Rocke, "Tumor Classification by Partial Least Squares Using Micro array Gene Expression Data", Bioinformatics, Vol. 18, No. 1, pp. 39-50, 2002

[7] Pengyi Yang and Zili Zhang, "An Embedded Two-Layer Feature Selection Approach for Microarray Data Analysis", IEEE Intelligent Informatics Bulletin, Vol.10, No.1, pp. 24-32, 2009

[8] Yuh-Jye Lee and Chia-Huang Chao, "A Data Mining Application to Leukemia Micro array Gene Expression Data Analysis", International Conference on Informatics, Cybernetics and Systems (ICICS), Kaohsiung, Taiwan, 2003

[9] James J. Chen and Chun-Houh Chen, "Micro array Gene Expression", Encyclopedia of Biopharmaceutical Statistics, 2nd Edition, Marcel Dekker, Inc., pp. 599-613, 2003

[10] Seeja and Shweta, "Microarray Data Classification Using Support Vector Machine", International Journal of Biometrics and Bioinformatics (IJBB), Vol. 5, No. 1, pp. 10-15, 2011

[11] Yee Hwa Yang and Natalie P. Thorne, "Normalization for Two-color cDNA Microarray Data", Science and Statistics: A Festschrift for Terry Speed, Vol. 40, pp. 403-418, 2003

[12] Fei Pana, Baoying Wanga, Xin Hub and William Perrizoa, "Comprehensive vertical sample-based KNN/LSVM classification for gene expression analysis", Journal of Biomedical Informatics, Vol. 37, pp. 240–248, 2004. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gassenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S., “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring”, Science, 286(15):531–537, 1999.

[13] Terrence S. Furey, Nello Cristianini, Nigel Duffy, David W. Bednarski, Michèl Schummer, and David Haussler, “Support vector machine classification and validation of cancer tissue samples using microarray expression data “, Bioinformatics6(10): 906-914 , 2000

[14] Zhang, X. and Ke, H.,” ALL/AML cancer classification by gene expression data using SVM and CSVM approach”, Genome Informatics, Universal Academy Press, pp. 237-239, 2000.

[15] Xin Zhao, Leo Wang-Kit Cheung, “Kernel-imbedded Gaussian processes for disease classification using microarray gene expression data”, BMC Bioinformatics.,8:67,2007.

[16] Wenlong Xu, Minghui Wang, Xianghua Zhang, Lirong Wang, Huanqing Feng," SDED: Anovel filter method for cancer-related gene selection", Bioinformation 2(7): 301-303,2008.

[17] D.P. Berrar, C.S. Downes, W. Dubitzky, "Multiclass Cancer Classification Using Gene Expression Profiling and Probabilistic Neural Networks", Pacific Symposium on Biocomputing 8:5-16, 2003.

[18] Golub TR, Slonim DK, Tamayo P, et al. Molecular classifi cation of cancer: class discovery and class prediction by gene expression monitoring. Science 1999 ; 286 : 531 -7

[19] Perou CM, SÃ¸rlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature 2000 406 : 747 -52

[20] Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 1999 ; 96 : 6745 -50

[21]. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 2002 ; 99 : 6567 -72

[22] Ahmad M. Sarhan, "Cancer Classification Based on Micro array Gene Expression Data Using DCT and ANN", Journal of Theoretical and Applied Information Technology, Vol. 6, No. 2, pp. 208-216, 2009

[23] Bharathi and Natarajan, "Cancer Classification of Bioinformatics data using ANOVA", International Journal of Computer Theory and Engineering, Vol. 2, No. 3, pp. 369-373, June 2010

[24] Bo Li, Chun-Hou Zheng, De-Shuang Huang, Lei Zhang and Kyungsook Han, “"Gene expression data classification using locally linear discriminant embedding", Computers in Biology and Medicine, Vol. 40, pp. 802–810, 2010

[25] Xiaosheng Wang and Osamu Gotoh, "A Robust Gene Selection Method for Micro array-based Cancer Classification", Journal of Cancer Informatics, Vol. 9, pp. 15-30, 2010

[26] Mallika and Saravanan, "An SVM based Classification Method for Cancer Data using Minimum Micro array Gene Expressions", World Academy of Science, Engineering and Technology, Vol. 62, No. 99, pp. 543-547, 2010

[27] Chhanda Ray, "Cancer Identification and Gene Classification using DNA Micro array Gene Expression Patterns", International Journal of Computer Science Issues, Vol. 8, Issue 2, pp. 155-160, March 2011.

[28] S. Mitra, "An evolutionary rough partitive clustering," *Pattern Recognit. Lett.*, vol. 25, pp. 1439–1449, 2004

[29] S. B. Cho and J. Ryu, "Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features," *Proc. IEEE*, vol. 90, no. 11, pp. 1744–1753, Nov. 2002.

[30] S. Bicciato,M. Pandin, G. Didon`e, andC.DiBello, "Pattern identification and classification in gene expression data using an autoassociative neural network model," *Biotechnol. Bioeng.*, vol. 81, pp. 594–606, 2003.

[31] F. Chu, W. Xie, and L. Wang, "Gene selection and cancer classification using a fuzzy neural network," in *Proc. 2004 Annu. Meet. North Amer. Fuzzy Information Processing Soc. (NAFIPS)*, vol. 2, pp. 555–559.

[32] K. Deb and A. Raji Reddy, “Reliable classification of two-class cancer data using evolutionary algorithms,” *BioSystems*, vol. 72, pp. 111–129, 2003.

[33] H. Midelfart, J. Komorowski, K. Nørsett, F. Yadetie, A. K. Sandvik,and A. Lægreid, “Learning rough set classifiers from gene expression and clinical data,” *Fundamenta Inf.*, vol. 53, pp. 155–183, 2002.

[34] M. E. Futschik, A. Reeve, and N. Kasabov, “Evolving connectionist systems for knowledge discovery from gene expression data of cancer tissue,” *Artif. Intell. Med.*, vol. 28, pp. 165–189, 2003.

[35] M. Banerjee, S. Mitra, and H. Banka, “Evolutionary-rough feature selection in gene expression data,” *IEEE Trans. Syst., Man, Cybern. C, Appl.*

[36] S.Mitra, YHaya.shi,Bioinformatics with softcomputing” IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 36, NO. 5, SEPTEMBER 2006.

[37] Fei Pana, Baoying Wanga, Xin Hub and William Perrizoa, "Comprehensive vertical sample-based KNN/LSVM classification for gene expression analysis", Journal of Biomedical Informatics, Vol. 37, pp. 240–248, 2004

[38] H. Midelfart, A. Lægreid, and J. Komorowski, *Classification of Gene Expression Data in an Ontology,* vol. 2199. Lecture Notes in Computer Science, Berlin, Germany: Springer-Verlag, 2001, pp. 186–194