# Microarray Based Disease Analysis Using Spatial EM and SVM Classification Algorithm

**Ms. S. Eniya[1], Mr. C. Saravanabhavan[2], Mrs. A. Kanimozhi[3]**
[1, 2, 3] Department of CSE
[1, 2, 3] Kongunadu College of Engineering & Technology

***Abstract-*** *Diseases classification using gene expression data is known to include the keys for addressing the fundamental harms relating to diagnosis and discovery. The recent introduction of DNA microarray technique has complete simultaneous monitoring of thousands of gene expressions possible. With this large quantity of gene expression data, researchers have started to discover the possibilities of disease classification using gene expression data. Quite a number of methods have been planned in recent years with hopeful results. But there are still a lot of issues which need to be address and understood. In order to gain insight into the disease classification difficulty, it is necessary to take a closer look at the problem, the proposed solutions and the associated issues all together. In this paper, we present a comprehensive clustering method and classification method such as Spatial Expectation Maximization, Support Vector classification and estimate them based on their calculation time, classification accuracy and ability to reveal biologically meaningful gene information. Based on our multiclass classification method to diagnosis the diseases and also find severity levels of diseases. Our experimental results show that classifier performance through graphs with improved accuracy.*

***Keywords***- Microarray data, Gene Expression, Clustering, Severity analysis, Classification

## I. INTRODUCTION

The recent initiation of microarray technologies has enabled biologists for the first time to concurrently monitor the activities of thousands of genes, constructing large quantities of complex data. Analysis of such data is becoming a main feature in the successful utilization of the microarray technology. Microarrays are tiny glass surfaces or chips, onto which microscopic amounts of DNA are attached in a grid layout. Each of the tiny spots of DNA relates to a single gene. The Structure of DNA is illustrated in fig 1.

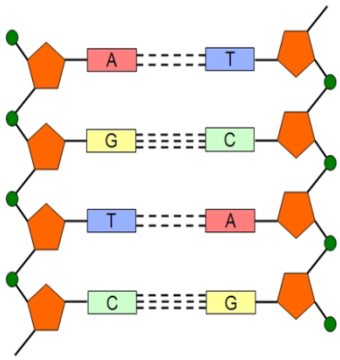


Fig 1: Structure of DNA

A, T, G, and C are the 'letters' of the DNA code and symbolize the chemicals adenine, thymine, guanine, and cytosine, respectively. These make up the nucleotide bases of DNA. Each gene's code merges these four chemicals in various ways to spell out three-letter 'words' that specify which amino acid is desired at every step in making a protein. The discovery of the genetic code ranks as one of the premiere events of biology and medicine. One of the most popular microarray applications is to compare gene expression levels in two dissimilar samples (e.g. healthy and diseased cells). RNA from the cells in the two different conditions are extracted and labeled with diverse fluorescent dyes (e.g. green for healthy and red for diseased cells). Both RNA are washed over the microarray. Gene patterns preferentially bind to their complementary sequences. The dyes allow measurement of the amount leap at each spot, in order to estimate the presence of genes. The microarray images are analyzed and the intensities considered. Finally, a gene expression matrix is obtained where rows correspond to genes and columns represent samples (i.e. unusual experimental conditions - stages, treatments, or tissues), and the numbers are the expression stage of the genes in the respective samples. In order to extract meaningful information from this data, data mining techniques are being employed. One goal in analyzing microarray data is to find genes which behave similarly over the course of a test by comparing rows in the expression matrix. These genes may be co-regulated or related in their function. Similar genes can be found by clustering methods. And gene patterns are classified by classification methods. These methods are used to predict diseases based predefined gene patterns and describe the patterns in following fig 2.

| 1st base | 2nd base | | | | | | | | 3rd base |
|---|---|---|---|---|---|---|---|---|---|
| | U | | C | | A | | G | | |
| U | UUU | (Phe/F) Phenylalanine | UCU | (Ser/S) Serine | UAU | (Tyr/Y) Tyrosine | UGU | (Cys/C) Cysteine | U |
| | UUC | | UCC | | UAC | | UGC | | C |
| | UUA | (Leu/L) Leucine | UCA | | UAA | Stop (Ochre) | UGA | Stop (Opal) | A |
| | UUG | | UCG | | UAG | Stop (*Amber*) | UGG | (Trp/W) Tryptophan | G |
| C | CUU | | CCU | (Pro/P) Proline | CAU | (His/H) Histidine | CGU | (Arg/R) Arginine | U |
| | CUC | | CCC | | CAC | | CGC | | C |
| | CUA | | CCA | | CAA | (Gln/Q) Glutamine | CGA | | A |
| | CUG | | CCG | | CAG | | CGG | | G |
| A | AUU | (Ile/I) Isoleucine | ACU | (Thr/T) Threonine | AAU | (Asn/N) Asparagine | AGU | (Ser/S) Serine | U |
| | AUC | | ACC | | AAC | | AGC | | C |
| | AUA | | ACA | | AAA | (Lys/K) Lysine | AGA | (Arg/R) Arginine | A |
| | AUG[A] | (Met/M) Methionine | ACG | | AAG | | AGG | | G |
| G | GUU | (Val/V) Valine | GCU | (Ala/A) Alanine | GAU | (Asp/D) Aspartic acid | GGU | (Gly/G) Glycine | U |
| | GUC | | GCC | | GAC | | GGC | | C |
| | GUA | | GCA | | GAA | (Glu/E) Glutamic acid | GGA | | A |
| | GUG | | GCG | | GAG | | GGG | | G |

Fig 2: Gene Expression.

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non-protein coding genes such as transfer RNA (tRNA) or small nuclear RNA (snRNA) genes, the product is a functional RNA. The process of gene expression is used by all known life - eukaryotes (including multi-cellular organisms), prokaryotes (bacteria and archaea), and utilized by viruses - to generate the macromolecular machinery for life.

Several steps in the gene expression process may be modulated, including the transcription, RNA splicing, translation, and post-translational modification of a protein. Gene regulation gives the cell control over structure and function, and is the basis for cellular differentiation, morphogenesis and the versatility and adaptability of any organism. Gene regulation may also serve as a substrate for evolutionary change, since control of the timing, location, and amount of gene expression can have a profound effect on the functions (actions) of the gene in a cell or in a multi-cellular organism.

In genetics, gene expression is the most fundamental level at which the genotype gives rise to the phenotype, i.e. observable trait. The genetic code stored in DNA is "interpreted" by gene expression, and the properties of the expression give rise to the organism's phenotype. Such phenotypes are often expressed by the synthesis of proteins that control the organism's shape, or that act as enzymes catalyzing specific metabolic pathways characterizing the organism.

## II. RELATED WORK

Wai-Ho Au,et.al. [1] presented an attribute clustering method which is able to group genes based on their interdependence so as to excavate meaningful patterns from the gene expression data. It could be used for gene grouping, selection and classification. The separation of a relational table into attribute subgroups permits a small number of attributes within or crosswise the groups to be selected for analysis. By clustering attributes, the search for dimension of a data mining algorithm is abridged. The reduction of search dimension is particularly important to data mining in gene expression data because such data typically contains of a huge number of genes (attributes) and a small number of gene expression profiles (tuples).The majority data mining algorithms are typically developed and optimized to balance to the number of tuples as a substitute of the number of attributes. The situation becomes even inferior when the number of attributes overwhelms the numeral of tuples, in which case, the likelihood of reporting patterns that are actually irrelevant due to chances becomes rather high.

Wolfgang Huber,et.al. [2] reviewed the methods utilized in processing and study of gene expression data generated using DNA microarrays. This type of research permits determining relative levels of mRNA abundance in a place of tissues or cell populations for thousands of genes simultaneously. Naturally, such an experiment needs computational and numerical analysis techniques. At the outset of the processing pipeline, the computational procedures are mostly determined by the knowledge and experimental setup that are used. Subsequently, as more consistent intensity values for genes emerge, pattern discovery methods arrive into play. The most striking peculiarity of this kind of data is that one usually obtains capacity for thousands of genes for only a much smaller number of conditions.

Marcel Dettling,et.al.[3] presented a promising innovative method for searching functional groups, each made up of only a few genes whose consensus expression profiles presents useful information for tissue discrimination. Due to the combinatorial difficulty when clustering thousands of genes rely on a greedy strategy. It optimizes an experiential objective function that quickly and competently measures the cluster's ability for phenotype discrimination. The output of our algorithm is thus potentially important for cancer type diagnosis. At the same time it is very accessible for interpretation, since the output consists of a very partial number of clusters, each summarizing the information of small

amount of genes. Thus, it may also expose insights into biological processes and give hints on explaining how the genome works.

Trevor Haste,et.al [4] addressed the problem of analyzing such data, then explain a statistical method, which they have called 'gene shaving'. The method recognizes subsets of genes with coherent expression patterns and large distinction across conditions. Gene shaving diverges from hierarchical clustering and other widely used methods for analyzing gene expression studies in that genes may belong to extra than one cluster and the clustering may be supervised by a result measure. The technique can be 'unsupervised', that is, the genes and models are treated as unlabeled, or partially or fully supervised by using known properties of the genes or samples to help in finding meaningful groupings. Illustrate the use of the gene shaving method to investigates gene expression measurements made on samples from patients with diffuse large B-cell lymphoma. The method classifies a small cluster of genes whose expression is highly predictive of survival.

Chris Ding,et.al [5] proposed a minimum redundancy maximum relevance (MRMR) feature selection framework. Genes selected via MRMR provide a more balanced coverage of the space and detain broader characteristics of phenotypes. They lead to significantly improved class predictions in extensive experiments on five gene expression data sets: NCI, Lymphoma, Lung, Leukemia and Colon. Improvements are observed consistently among four classification methods: Naïve Bayes, Linear discriminate analysis, Logistic regression and Support vector machines.

Hanchuan Peng,et.al [6] present a theoretical analysis showing that mRMR is equivalent to Max-Dependency for first-order feature selection, but is more efficient. Second, investigate how to combine mRMR with other feature selection methods into a two-stage selection algorithm. By doing this, Then show that the space of candidate features selected by mRMR is more characterizing. This property of mRMR facilitates the integration of other feature selection schemes to find a compact subset of superior features at very low cost. Third, through comprehensive experiments, compare mRMR, Max-Relevance, Max-Dependency, and the twostage feature selection algorithm, using three different classifiers and four data sets.

Roberto Battiti,et.al [7] investigated the application of the mutual information criterion to evaluate a set of candidate features and to select an informative subset to be used as input data for a neural network classifier. Because the mutual information measures arbitrary dependencies between random variables, it is suitable for assessing the "information content" of features in complex classification tasks, where methods bases on linear relations (like the correlation) are prone to mistakes. The fact that the mutual information is independent of the coordinates chosen permits a robust estimation. Nonetheless, the use of the mutual information for tasks characterized by high input dimensionality requires suitable approximations because of the prohibitive demands on computation and samples. An algorithm is proposed that is based on a "greedy" selection of the features and that takes both the mutual information with respect to the output class and with respect to the already-selected features into account.

D. Nguyen,et.al [8] proposed a novel analysis procedure for classifying (predicting) human tumor samples based on microarray gene expressions. This procedure involves dimension reduction using Partial Least Squares (PLS) and classification using Logistic Discrimination (LD) and Quadratic Discriminate Analysis (QDA). We compare PLS to the well known dimension reduction method of Principal Components Analysis (PCA). Under many circumstances PLS proves superior; we illustrate a condition when PCA particularly fails to predict well relative to PLS. The proposed methods were applied to five different microarray data sets involving various human tumor samples: (1) normal versus ovarian tumor; (2) Acute Myeloid Leukemia (AML) versus Acute Lymphoblastic Leukemia (ALL); (3) Diffuse Large B-cell Lymphoma (DLBCLL) versus B-cell Chronic Lymphocytic Leukemia (BCLL); (4) normal versus colon tumor; and (5) Non-Small-Cell-Lung-Carcinoma (NSCLC) versus renal samples. Stability of classification results and methods were further assessed by re-randomization studies.

## III. GENE CLUSTERING

**K means clustering:**

The main objective in cluster analysis is to group objects that are similar in one cluster and separate objects that are dissimilar by assigning them to different clusters. One of the most popular clustering methods is K-Means clustering algorithm. It classifies object to a pre-defined number of clusters, which is given by the user (assume K clusters). The idea is to choose random cluster centres, one for each cluster. These centres are preferred to be as far as possible from each other. In this algorithm mostly Euclidean distance is used to find distance between data points and centroids.

**EM algorithm:**

Given the microarray data and the current set of model parameters, the probability to associate a gene (or experiment) to every cluster is evaluated in the E step. Then, the M step finds the parameter setting that maximizes the likelihood of the complete data. The complete data refers to both the (observed) microarray data and the assignment of the genes (or experiments) to the clusters. The likelihood of the model increases as the two steps iterates, and convergence is guaranteed. The EM algorithm iterates between Expectation (E) steps and Maximization (M) steps. In the E step, hidden parameters are conditionally estimated from the data with the current estimated. In the M step, model parameters are estimated so as to maximize the likelihood of complete data given the estimated hidden parameters. When the EM algorithm converges, each data object is assigned to the component (cluster) with the maximum conditional probability

**Spatial EM algorithm:**

A gene-based clustering is used to group the gene patterns. Patterns are clustered based on genetic code transcriptions. The proposed methodology includes Spatial EM that can be used to calculate spatial mean and rank based scatter matrix to extract relevant patterns and further implement KNN (K- nearest neighbor classification) approach to diagnosis the diseases. An important finding is that the proposed semi supervised clustering algorithm is shown to be effective for recognizing biologically significant gene clusters with excellent predictive capability. Spatial-EM modifies the component estimates on each M-step by spatial median and rank covariance matrix to gain robustness at the cost of increasing computational burden and losing theoretical tractability. Pseudocode of the algorithm is described as:

Initialization t = 0, $\mu_j, \sum_j = I, \tau_j = \frac{1}{K}\ for\ \forall_j$
Do until $\tau_j^t coverage\ for\ all\ j$
For j=1 to K
E-Step: Calculate $T_{ji}^t$
M-Step: Update $\tau_j^{t+1}$
Define$w_{ji}^t$, Find $\mu_j^{t+1}$, Find$(\sum_j^{t+1})^{-1}\ and\ (\sum_j^{t+1})^{-1/2}$
End
t=t+1
End

In spatial algorithm can first calculate the maximum coverage of data and then initialize all variables and perform Expectation and Maximization steps as in EM algorithm. The EM iteration alternates between performing an expectation (E) step, which creates a function for the hope of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which figures parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to decide the distribution of the latent variables in the next E step. The EM algorithm proceeds from the observation that the following is a way to explain these two sets of equations numerically.

**Gene Classification:**

Microarray classification approaches based on machine learning algorithms applied to DNA microarray data have been shown to have statistical and medical relevance for a variety of diseases. One particular machine learning algorithm, Support Vector Machines (SVMs), has exposed promise in a variety of biological classification tasks, including gene expression microarrays. SVMs are powerful classification systems based on regularization techniques with excellent performance in many practical classification problems. The Support Vector Machine is rooted in statistical learning theory. It is different from the other classification method in the sense that SVM tries to maximize the separation between samples of two classes. Normally, only a subset of the data samples determines the decision hyper plane. Suppose the n data samples belong to two classes $\{(x_1, y_1), \dots, (x_n, y_n)\}, x_i \epsilon \Re^m and\ y_t$ =1 or -1. A support vector machine tries to find a hyper plane $w^T x + b = 0$ which satisfies

$$\boldsymbol{y_i} w^T x_i + b \geq 1 - \varepsilon_i, i = 1, \dots, n,$$

where $\varepsilon_i \geq 0, i = 1, \dots, n$ are slack variables. As the distance from a model to the hyper plane is inversely proportional to $w^T w$ a quadratic minimization problem is formulated as follows:

Minimize $w^T w + C \sum_{i=1}^n \varepsilon_i$
Subject to $\boldsymbol{y_i} w^T x_i + b \geq 1 - \varepsilon_i, i = 1, \dots, n,$

where C is a parameter to balance the generalization facility represented in the first term $w^T w$ and separation ability indication in the second term$\sum_{i=1}^n \varepsilon_i$. A smaller value of the first term corresponds to better generalization, while the fewer positive values of the slack variables in the second term correspond to fewer misclassifications on the training samples. When the later is equal to zero, the training samples are linearly separable and there is no misclassification.

## IV. RESULTS AND DISCUSSION

Experimental results can evaluate the performance of the system using Accuracy rate. The accuracy rate is calculated using true positive, true negative, false positive and false negative metrics. So the accuracy rate is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}$$

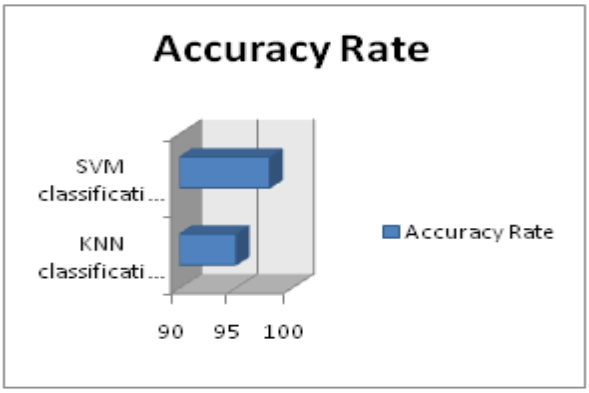


Fig 4: Performance evaluation

Proposed framework provide improved accuracy rate in disease classification and analyzed severity level of diseases.

## V. CONCLUSION

Microarray is an important tool for cancer classification at the molecular level. It monitors the expression levels of large number of genes in parallel. With large amount of expression data obtained through microarray experiments, suitable statistical and machine learning methods are needed to search for genes that are relevant to the identification of different types of disease tissues. In this paper, we have proposed a hybrid gene selection method, which combines a spatial EM methods and SVM classification to achieve high classification performance. Then provide severity level for each classified diseases.

## REFERENCES

[1] W.-H. Au, K.C.C. Chan, A.K.C. Wong, and Y. Wang, "Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 2, no. 2, pp. 83-101, Apr.-June 2005.

[2] Wolfgang Huber, Anja von Heydebreck, Martin Vingron, "Analysis of microarray gene expression data,"J. Statistical Physics, vol. 110, nos. 3-6, pp. 1117-1139, 2003.

[3] M. Dettling and P. Buhlmann, "Supervised Clustering of Genes," Genome Biology, vol. 3, no. 12, pp. 0069.1-0069.15, 2002.

[4] T. Hastie, R. Tibshirani, M.B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W.C. Chan, D. Botstein, and P. Brown, "'Gene Shaving' as a Method for Identifying Distinct Sets of Genes with Similar Expression Patterns," Genome Biology, vol. 1, no. 2, pp. 1-21, 2000.

[5] C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," J. Bioinformatics and Computational Biology, vol. 3, no. 2, pp. 185-205, 2005.

[6] D. Koller and M. Sahami, "Toward Optimal Feature Selection," Proc. Int'l Conf. Machine Learning, pp. 284-292. 1996.

[7] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," Artificial Intelligence, vol. 97, nos. 1/2, pp. 273-324, 1997.

[8] A.K. Jain and R.C. Dubes, Algorithms for Clustering Data. Prentice Hall, 1988.

[9] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification and Scene Analysis. John Wiley and Sons, 1999.

[10] D. Jiang, C. Tang, and A. Zhang, "Cluster Analysis for Gene Expression Data: A Survey," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 11, pp. 1370-1386, Nov. 2004.

[11] W.-H. Au, K.C.C. Chan, A.K.C. Wong, and Y. Wang, "Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 2, no. 2, pp. 83-101, Apr.-June 2005.

[12] A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G.C. Tseng, "Evaluation and Comparison of Gene Clustering Methods in Microarray Analysis," Bioinformatics, vol. 22, no. 19, pp. 2405-2412, 2006.

[13] M. Medvedovic and S. Sivaganesan, "Bayesian Infinite Mixture Model Based Clustering of Gene Expression

Profiles," Bioinformatics, vol. 18, no. 9, pp. 1194-1206, 2002.

[14] Y. Joo, J.G. Booth, Y. Namkoong, and G. Casella, "Model-Based Bayesian Clustering (MBBC)," Bioinformatics, vol. 24, no. 6, pp. 874-875, 2008.

[15] J. Herrero, A. Valencia, and J. Dopazo, "A Hierarchical Unsupervised Growing Neural Network for Clustering Gene Expression Patterns," Bioinformatics, vol. 17, pp. 126-136, 2001.

[16] W. Haiying, Z. Huiru, and A. Francisco, "Poisson-Based Self- Organizing Feature Maps and Hierarchical Clustering for Serial Analysis of Gene Expression Data," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 4, no. 2, pp. 163-175, Apr.- June 2007.

[17] L.J. Heyer, S. Kruglyak, and S. Yooseph, "Exploring Expression Data: Identification and Analysis of Coexpressed Genes," Genome Research, vol. 9, no. 11, pp. 1106-1115, 1999.

[18] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub, "Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation," Proc. Nat'l Academy of Science USA, vol. 96, no. 6, pp. 2907-2912, 1999.

[19] K.Y. Yeung and W.L. Ruzzo, "Principal Component Analysis for Clustering Gene Expression Data," Bioinformatics, vol. 17, no. 9, pp. 763-774, 2001.

[20] G.J. McLachlan, K.-A. Do, and C. Ambroise, Analyzing Microarray Gene Expression Data. Wiley-Interscience, 2004.