Computational analysis of DNA microarray data using datamining

R.R. SHELKE[1]* and V.M. DESHMUKH[2]

[1]Hari Om colony, Behind Hollywood colony, Kathora Road, VMV Post, Amravati - 444 604 (India)
[2]Department of Computer Science and Engineering,
Prof. Ram Meghe Institute of Technology and Research, Badnera, Amravati - 444 701 (India)

(Received: April 18, 2007; Accepted: May 15, 2007)

ABSTRACT

The traditional methods in molecular biology generally work on a "one gene in one experiment" basis which means that the throughput is very limited and the "whole picture" of gene function is hard to obtain. In the past several years, a new technology, called DNA microarray, has attracted tremendous interests among biologists.Researchers can now routinely investigate the biological molecular state of a cell measuring the simultaneous expression of tens of thousands of genes using DNA microarrays. In the present work , the datamining for such a DNA microarray data has been focused in greater details. To fulfill this need datamining technique has been developed. The 'cell slide images' have been processed to calculate different gene values in the particular slide which in turn maintain the database for cluster analysis and visualization which helps to compare database of DNA sample images for disease diagnosis . The diseases taken into consideration in the present work are cancer , hepatatis and diabetes.

Keywords: Datamining, DNA, Clustering and Visualisation.

INTRODUCTION

Datamining is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of datamining is prediction and predictive datamining is the most common type of datamining and one that has the most direct business applications,such as DNA microarray data analysis and visualization .

Microarrays are a revolutionary new technology with great potential to provide accurate medical diagnosis which helps to find the right treatment and cure for many diseases and provide a detailed genome-wide molecular portrait of cellular states. The work included in this issue is a good sample of second generation methodologies and techniques that is being used under development today. As it can be seen from the results , they are very promising and extend the possibilities of applying computational analysis and datamining to aid research in biology and medicine. The present work not only would like to close this introduction with a brief discussion to emphasize the large potential payoff of these analytical efforts but also pointing out the huge challenges ahead. There is little doubt about the potential of computational and statistical analysis of molecular probes to improve the understanding of the cell and the possibilities of molecular medicine. Finding new insights into the molecular basis of biological processes and searching for new drugs and treatments is a problem of high complexity and where the techniques of molecular biology has been applied for many decades. The process is analogous to a large search of a few molecular entities, connections or relationships in a large sea of possibilities. One important goal of current and future computational analysis methods short of reverse engineering the entire cell circuitry,which by the way it is still an intractable problem, should be to reduce that search and help to expose the most promising candidates (gene, proteins, drugs etc.) for further study.

DNA microarray experiments generate a substantial amount of information about global gene expression. Gene expression profiles can be represented as points in multi-dimensional space. It is essential to identify relevant groups of genes in biomedical research. Clustering is helpful in pattern recognition in gene expression profiles. Some clustering techniques have been introduced. A critical aspect in the analysis of gene expression data is identification of clusters of genes that have similar expression patterns. Clustering techniques transform a large matrix of expression levels for different genes in different conditions into a more organized and informative collection gene sets, which are expected to share similar biological properties. Clustering techniques are predominantly influential in tissue classification, function annotation, and other biomedical applications[1]. Clustering has been studied for a long time in statistical learning. There are many algorithms that was carried out in gene expression profiling, *i.e.*, hierarchical clustering method, k-means clustering, self-organizing map (SOM)[1], principle component analysis[1], fuzzy c-means clustering[2], CLICK[3], adaptive quality-based clustering[4], quantum clustering[5], mean-shift[5], bagged clustering[6] and Gustafson-Kessel method[7], etc. The fractal clustering algorithm in gene expression profiling is also given in detail[10]. It provides a very natural way of defining clusters that is not restricted to any particular cluster shape. This algorithm is based on the introduction of the concept of the fractal dimension in clusters. This method clusters points in such a way that data points in the same cluster are more self-affine among themselves than to the points in other clusters, although the clusters do not have to be ideal fractal themselves[8].

Now a days ,different diseases are arising which in turn causes an increment in death rate.Different blood testing are available for diagnosis. Though various blood testings or any other testings are available,diagnosis of some diseases is very difficult. Therefore, only blood testing or any other testing (eg.spinal cord fluid testing)is not sufficient. The DNA testing is one of the advanced technique in which affected DNA & normal DNA slides are compared for disease diagnosis. With the software developement, it became a very easy and beneficial task.In software development, slide images of palient's DNA and disease affected DNA can be processed and compared. Pixel values of these images stored in the database can be also compared for ultimate result . In the present work , datamining of DNA microarray data is performed in which ,slide images of patient's DNA sample and disease affected DNA sample are processed and compared. Database of the two samples are also compared so that exact result can be given . The result shows whether the patient is suffering from particular disease or not. In this work, slide images of three diseases viz. cancer, hepatitis, diabetic are considered.

Present work

A visual programming environment has been done for functional genomic data analysis. Its basic data processing units are called widgets. Each widget implements a task of data manipulation, analysis, model building or visualization. The advantage of widgets lies in their modularity. Widgets are connected through channels and communicate with each other by sending and receiving data. The output of one widget is used as an input for one or several other subsequent widgets. This property relieves the user from the need to design data structure which is one of the greatest obstacles for lay users. A collection of widgets and their communication channels is called a schema which is essentially a program designed by the user for a specific data analysis task. The programming process creating a schema with widgets and their connections has been done visually through an easy-to-use graphic interface. It is easily possible to maintain database to perform Clustering and Visualization of DNA microarray data and to compare database of disease DNA sample for disease diagnosis. It has been also compared slide images of patient DNA sample & disease DNA sample for disease diagnosis which helps to find proper treatment to the patient .The block diagram used in the Present work is as follows :

RESULTS AND DISCUSSION

There are several steps in the design & analysis of DNA microarray experiment. Those are 1. DNA type 2. Chip fabrication 3. sample preparation 4. readout 5. software. The advanced technique(Software) which give the exact result has been used . Once the software is developed, it can be used for different types of DNA . But only checking sample image is not sufficient for best result.It's database has also to be tested . For database testing datamining is the best way in the present work. Datamining, clustering & visualization of DNA Microarray data has been done for the proper output.

The patient's name whose DNA is to be tested is entered and the record is added and updated for next step .The microdata so added & recorded in the previous case has been viewed in further form along with visualization of the disease affected sample.

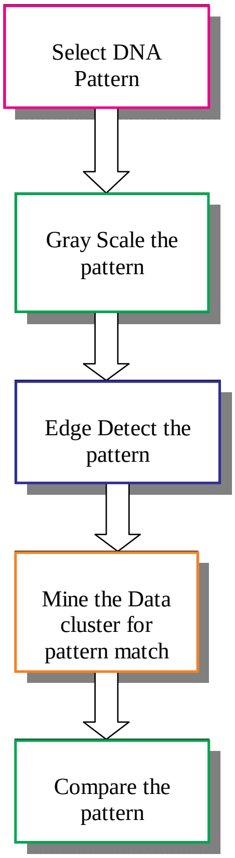
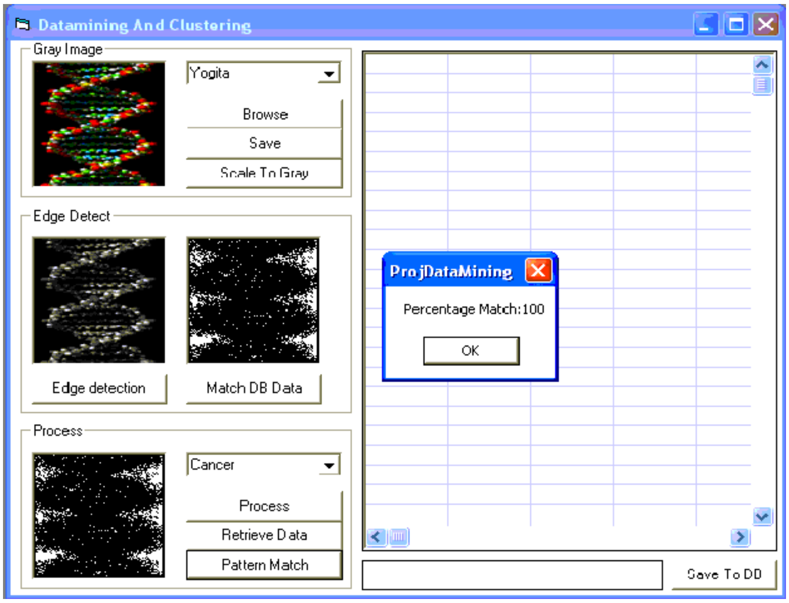


Fig. - 1: Block diagram

The main output form in the present work is as shown in the Fig. -2. In this step, the gray scaling, edge detection, processing, matching has been done.First, the DNA sample image of patient has been selected .The image so selected is gray scaled and edge detections has been done. After that the disease affected sample image is selected (cancer, diabetic or hepatatis). This image is then processed. In processing, gray scaling and edge detection has been done. The processed pattern is then compared with patients image pattern. In comparison, percentage of matching is given which finally gives the result whether the patient is suffering by particular disease or not. This is the ultimate result of the present work. Again, databases of two slides has been compared to see the output.The same procedure has been followed for diabetic and hepatitis also.

Conclusions

Microarrays are a revolutionary new technology with great potential to provide accurate medical diagnostics. The research work included in the present work is a good sample of second generation methodologies and techniques those are being used under development today. It can be seen from the results that microarrays are very promising and extend the possibilities of applying computational analysis and datamining to aid research in biology and medicine. In the present work, It is easily possible to maintain database to perform Clustering and Visualization of DNA



Percentage Match:100

Fig. - 2: Comparison of DNA sample and database

microarray data and to compare database of disease DNA sample for disease diagnosis. It has been also compared slide images of patient DNA sample & disease DNA sample for disease diagnosis which helps to find proper treatment to the patient. The diseases taken into the consideration are cancer,diabeties and hepatitis .

There are various possible applications of this concept which can be benefited to the nation as well as society. Following are some applications in brief.

Gene discovery

Particular genes can be discovered.

Disease diagnosis

Infected DNA & normal DNA can be compared and disease diagnosis is done properly.

Drug discovery:

Pharmacogenomics

Pharmacogenomics is the hybridization of functional genomics and molecular pharmacology. The goal of pharmacogenomics is to find correlations between therapeutic responses to drugs and the genetic profiles of patients. It is possible to detect appropriate drug for the patient so that the reaction of drug can be avoided.

Toxicological research: Toxicogenomics

Toxicogenomics is the hybridization of functional genomics and molecular toxicology. The goal of toxicogenomics is to find correlations between toxic responses to toxicants and changes in the genetic profiles of the objects exposed to such toxicants .

REFERENCES

1. J. Quanckenbush, Computational analysis of microarray data, *Nature Review Genetics*, **2,** 418-427, (2001).
2. D. Dembele, Kastner, P., Fuzzy C-means method for clustering microarray data,*Bioinformatics*, **19**, 973-980, (2003).
3. R. Sharan, Maron-Katz, A., and Shamir, R., Click and expander: a system for clustering and visualizing gene expression data, *Bioinformatics*, **19,** 1787-1799, (2003).
4. F. D. Smet, Mathys J., Marchal K., THijs G., Moor BT and Moreau Y, Adaptive quality-based clustering of gene expression profiles, *Bioinformatics*, **18**, 735-746, (2002).
5. D. Barash, Comaniciu,D., Meanshift clustering for DNA Microarray Analysis, *Proceeding of the 2004 IEEE Computational Systems Bioinformatics Conference*, (2004).
6. S. F. Dudoit, J., Bagging to improve the accuracy of a clustering procedure, *Bioinformatics*, **19**, (2003).
7. D.-W. Kim, Lee,K.H., Lee,D, Detecting clusters of different geometrical shapes in microarray gene expression data, *Bioinformatics*, *vol*. Advanced Access Published, pp. 1-11, (2005).
8. L. Liebovitch, and T. Toth, A Fast algorithm to determine fractal dimensions by box counting, *Physics Letters*, **141A**, (1989).
9. G. Mizuguchi, X. Shen, J. Landry , ATP-driven exchange of histone h2az variant catalyzed by swvl1 chromatin remodelling complex, *Science*, **303,** 343-348, (2004).
10. Lu Yong Wang, Fractal Custering for Microarray Data Analysis *Proceeding of the 2004 IEEE Computational Systems Bioinformatics Conference*, (2004).
11. Chipping Forecast, The Chipping Forecast. *Special Supplement. Nature Genet.* **21,** (1999).
12. The Chipping Forecast II. Special Supplement. *Nature Genet.* **32,** (2002).
13. M. Schena, et al ,Quantitative monitoring of gene expression patterns with a cDNA microarray. *Science., **270:*** 467-470 (1995).
14. J.L. DeRisi, et al., Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science,* **278:** *680*-686 (1997).
15. S. Chu, *et al.,* The transcriptional program of germ cell development in budding yeast. *Science,* **282:** 699-705 (1998).
16. V.R. Iyer, et al.,The transcriptional program in the response of human fibroblasts to serum. *Science,* **283:** 83-87, (1999)
17. J. DeRisi , et al., Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* Dec;**14**(4): 457-60 (1996).