International Journal of Computing Science and Information Technology, 2018
ISSN: 2278-9669, April 2018 (http://ijcsit.org)

Data Mining Tools For Analyzing Microarray

Sharmila Ravishankar [1].
Seshadripuram College,
Bangalore, India;

January 2018

Abstract

Microarray based analysis is now a consolidated methodology and has widespread use in areas such as diagnostics and drug target identification. Large-scale microarray studies are also becoming crucial to a new way of conceiving experimental biology. A main issue in microarray transcription is data analysis and mining. When microarrays became a methodology of general use, considerable effort was made to produce algorithms and methods for the identification of expressed genes. More recently, the focus has switched to algorithms and database development for microarray data mining. Furthermore, the evolution of microarray technology is allowing researchers to grasp the regulative nature of basic expression analysis with mRNA characteristics and with DNA characteristics, i.e. comparative genomic nucleotide structure. In this paper, we will review approaches used to detect expressed genes from microarray.

Keywords: Microarray transcription, Expressed Gene, Nucleotide

1. Introduction

Microarray is a collection of microscopic DNA spots attached to a solid surface. It is used to measure the expression levels of large numbers of genes simultaneously or regions of a genome. Each DNA spot contains Pico moles(10^{-12} moles) of a specific DNA sequence, known as probes(oligos). These can be a short section of a gene or other DNA element that are used to hybridize.

A microarray experiment consists of the following components: a set of *probes*, an *array* on which these probes are immobilised at specified locations, a *sample* containing a complex mixture of labelled biomolecules that can bind to the probes, and a *detector* that is able to measure the spatially resolved distribution of label after it has bound to the array. The probes are chosen such that they bind to specific sample molecules; for DNA arrays, this is ensured by the high sequence-specificity of the hybridization reaction between complementary DNA strands. The array is typically a glass slide or a nylon membrane. The sample molecules may be labelled through the incorporation of radioactive markers, such as 33P, or of fluorescent dyes, such as phycoerythrin, Cy3, or Cy5. After exposure of the array to the sample, the abundance of individual species of sample molecules can be quantified through the signal intensity at the matching probe sites.

In general there are five basic aspects of microarrays: a) coupling biomolecules to a platform; b) preparing samples for detection; c) hybridization; d) scanning; and e) analyzing the data.

Background:

Microarray is one such technology which enables the researchers to investigate and address issues which were once thought to be non-traceable by facilitating the simultaneous measurement of the expression levels of thousands of genes. A microarray is simply a glass slide on which DNA molecules are fixed on an ordered manner at specific locations called spots or probes [Fig: 1]. The spots are printed on the glass slide by different technologies such as photolithography to robot spotting. The DNA in a spot may either be complete copy of genomic DNA or short stretch of oligo-nucleotides that correspond to a gene.

Fig: 1 Affymetrix- Microarray

Using microarrays one can analyse the expression of many genes in a single reaction quickly and in an efficient manner. It has empowered the scientific community to understand the fundamental aspects underlining the growth and development of life as well as to explore the genetic causes of anomalies occurring in the functioning of the human body. The core principle behind microarrays is hybridization between two DNA strands, the property of complementary nucleic acid se-

quences to specifically pair with each other by forming hydrogen bonds between complementary nucleotide base pairs. However, with the generation of large amounts of microarray data, it has become increasingly important to address the challenges of data quality and standardization related to this technology. The recent advancement of the microarray technology has allowed for a very high resolution mapping of chromosomal aberrations with the use of their tiling array platform.

Computational data analysis tasks such as data mining which includes classification and clustering used to extract useful knowledge from microarray data [Fig :2]. In addition, relating gene expression data with other biological information; it will provide kind of biological discoveries such as transcription factor biding site analysis, pathway analysis, and protein- protein interaction network analysis. In the present paper focus was given on biologist's perspective to get knowledge about the several tools and programs available for microarray data mining tasks.

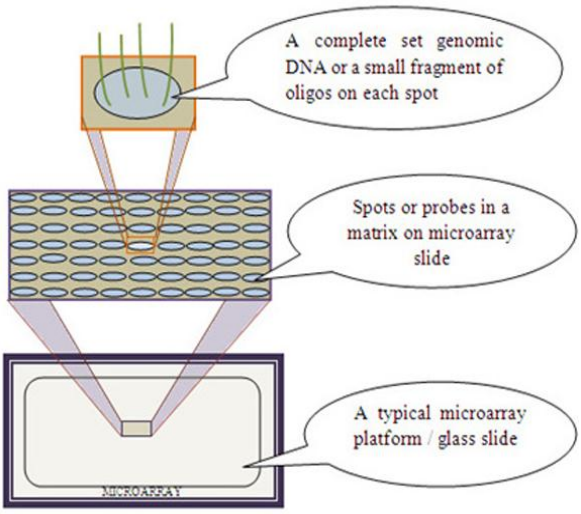

Figure: 2 A typical microarray platform

Microarray Data Analysis:

Microarray data sets are commonly very large, so it is extremely useful to reduce the dataset to those genes that are best distinguished between the two cases (e.g. normal vs. diseased). Such analyses produce a list of genes whose expression is considered to change and known as differentially expressed genes. Identification of differential gene expression is the first task of an in depth microarray analysis. There are two common methods for in depth microarray data analysis are:

1) Clustering 2) Classification

Clustering is one of the unsupervised approaches to classify data into groups of genes or samples with similar patterns that are characteristic to the group.

Classification is supervised learning and also known as class prediction or discriminant analysis. Generally, classification is a process of learning-from-examples. Given a set of pre-classified examples, the classifier learns to assign an unseen test case to one of the classes.

Identification of Differentially Expressed Genes:

Differentially expressed genes are the genes whose expression levels are significantly different between two groups of experiments. The genes are relevant for discovering potential drug targets and biomarkers. In the earlier stage, simple "fold change" approach was used to find differences under assumption that changes above some threshold, (For example, two-fold) were biologically significant. There are several univariate statistical methods were used later to determine either the expression or relative expression of a gene
from normalized microarray data, including t tests , modified t-test known as SAM , two-sample t tests , F-statistic and Bayesian models.

For more complex datasets with multiple classes, Analysis of Variance (ANOVA) techniques were used. Various software packages have been developed and available to identify changes in expression using the above statistical methods. The commonly used and freely available programs with its underlying algorithm are

S. No	Algorithm	Software Tools
1	Modified t-test known as SAM	SAM
2	Non-parametric t-test/ ANOVA	MeV
3	Student's t-test and Mann-Whitney test	iArray
4	Optimal Discovery Procedure	EDGE
5	Simple t-test or regularized t-tests	Cyber-T

Cluster Analysis:

Clustering is the most popular method currently used in the first step of gene expression data matrix analysis. It is used for finding co-regulated and functionally related groups. Clustering is particularly interesting in the cases when we have complete sets of an organism's genes. There are three common types of clustering methods:

1) Hierarchical clustering
2) K-means clustering and
3) Self-organizing maps.

Hierarchical clustering is a commonly used unsupervised technique that builds clusters of genes with similar patterns of expression. This is done by iteratively grouping together genes that are highly correlated in terms of their expression measurements, then continuing the process on the groups themselves. It is a method of cluster

analysis which seeks to build a hierarchy of clusters. A dendrogram represents all genes as leaves of a large, branching tree. The number and size of expression patterns within a data set can be estimated quickly, although the division of the tree into actual clusters is often performed visually. It generally falls into two categories:

1) Agglomerative and
2) Divisive.

Agglomerative is a bottom up approach where each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy. Divisive is a top down approach i.e., all observations start in one cluster and splits are performed recursively as one moves down the hierarchy.

K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. It is one of the simplest clustering techniques and it is commonly used in medical imaging and biometrics. The K-means clustering algorithm typically uses the Euclidean properties of the vector space. After the initial partitioning of the vector space into K parts, the algorithm calculates the center points in each subspace and adjusts the partition so that each vector is assigned to the cluster the center of which is the closest. This is repeated iteratively until either the partitioning stabilizes or the given number of iterations is exceeded.
A self-organizing map (SOM) is a neural network based non-hierarchal clustering approach. (SOMs) work in a manner similar to K-means clustering. The commonly used and freely available programs for clustering analysis are:

S. No	Algorithm	Software Tools
1	Hierarchical clustering, K-means clustering self-organizing maps	Cluster and Treeview
2	Hierarchical clustering, K-means clustering self-organizing maps	dChip
3	Hierarchical clustering, K-means clustering, Tree EASE, self organizing maps, & QT-clustering etc.	MeV
4	Hierarchical clustering, K-means clustering, and QT-clustering	MAGIC Tools
5	Bayesian clustering program on a-temporal expression data.	CAGED

S. No	Algorithm	Software Tools
1	Artificial Neural Networks, Decision trees, k Nearest Neighbors, Support Vector Machines, and many	weka
2	Artificial Neural Networks, Decision trees, k Nearest Neighbors, Support Vector Machines, and many	SAS
3	Artificial Neural Networks, Decision trees, k Nearest Neighbors, Support Vector Machines, and many	IBM/SPSS Clementine
4	Support Vector Machines	SVMlight
5	Support Vector Machines	LIBSVM

Classification:

Classification is also known as class prediction, discriminant analysis, or supervised learning. Given a set of pre-classified examples, (for example, different types of cancer classes such as AML and ALL) a classifier will find a rule that will allow to assign new samples to one of the above classes.

For classification task, one must have sufficient sample numbers to allow an algorithm to be trained known as training test and then to have it tested on an independent set of samples known as test set. Using normalized gene expression data as input vectors, classification rules can be built. There are a wide range of algorithms that can be used for classification are:

1) k Nearest Neighbors (kNN)
2) Artificial Neural Networks
3) Weighted voting and
4) Support vector machines (SVM).

The promising application of classification is in clinical diagnostics to find disease types and sub types. The general data mining and machine learning application tools are used for classification tasks are:

Knowledge Discovery with Microarray Data:

Classification, clustering and identification of differential genes can be considered as basic microarray data analysis tasks with gene expression profiles alone. However, Gene expression profiles can be linked to other external resources to make new discoveries and knowledge. Some of the common applications that addressed with gene expression data with other

biomedical information are :

1. Identification of transcription factor binding sites.
2. Protein-protein interaction network and pathway analysis.
3. Gene set enrichment analysis.

Identification of transcription factor binding sites:

The identification of functional elements such as transcription-factor binding sites (TFBS) on a whole-genome level is the next challenge for genome sciences and gene-regulation studies. Transcription factors act as critical molecular switches in the gene expression profiling. Transcription factors play a prominent role in transcription regulation; identifying and characterizing their binding sites is central to annotating genomic regulatory regions and understanding gene-regulatory networks.

Protein-protein interaction network and pathway analysis:

Protein-protein interactions (PPI) are useful tools for investigating the cellular functions of genes. It is a core of the entire interactomics system of any living cell. PPI improves our understanding of diseases and can provide the basis for new therapeutic approaches.

Gene set enrichment analysis:

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether a set of genes shows statistically significant and concordant differences between two biological states. The gene sets are defined based on prior biological knowledge, e.g., published information about biochemical pathways, located in the same cytogenetic band, sharing the same Gene Ontology category, or any user-defined set. The goal of GSEA is to determine whether members of a gene set tend to occur toward the top (or bottom) of the list, in which case the gene set is correlated with the phenotypic class distinction.

Conclusion:

DNA Microarray is a revolutionary technology and microarray experiments produce considerably more data than other techniques. Integrating gene expression data with other biomedical resources will provide new mechanistic or biological hypotheses. However, innovative statistical techniques and computing software are essential for the successful analysis of microarray data.

Microarrays are able to simultaneously monitor the expression levels of thousands of genes. Such gene expression information can be used in medicine for comparing clinically relevant groups (eg, healthy vs diseased), uncovering new subclasses of diseases, and predicting clinically important outcomes, such as the response to therapy and survival. However, the improved understanding that can be gained with this technology is critically dependent on the quality of the analytical tools employed. This review shows the current bioinformatics tools and the promising applications for analyzing data from microarray experiments. The various data analysis perspectives and soft wares mentioned in the paper will help the biological expertise as a good foundation for computational analysis of microarray data.

REFERENCES:

[1] *A Practical Approach to Microarray Data Analysis*, Kluwer Academic, Boston, Mass, USA, 2002, edited by D. Berrar, W.Dubitzky and M. Granzow.

[2] National Library of Medicine, "PubMed literature abstract database," http://www.ncbi.nlm.nih.gov/pubmed.

[3] M. Kathleen Kerr and Gary A. Churchill. Statistical design and the analysis of gene expression microarray data. *Genet. Res.*, 77:123–128, 2001.

[4] G.A. Churchill. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics*, 32 Suppl. 2:490–495, 2002.

[5] Yee Hwa Yang and Terence P. Speed. Design issues for cDNA microarray experiments. *Nat. Rev. Gen.*, 3:579–588, 2002.