BioImpacts, **2011**, 1(2), 75-86

http://bi.tbzmed.ac.ir/

A Glance at DNA Microarray Technology and Applications

Amir Ata Saei[1] and Yadollah Omidi[1,2]*

[1]*Research Center for Pharmaceutical Nanotechnology, Faculty of Pharmacy, Tabriz University of Medical Sciences, Tabriz, Iran*

[2]*Ovarian Cancer Research Center, School of Medicine, University of Pennsylvania, Philadelphia, USA*

ARTICLE INFO

Article Type:
Review Article

Article History:
Received: 27 June 2011
Revised: 13 July 2011
Accepted: 20 July 2011
ePublished: 04 Aug 2011

Keywords:
Microarray
Data Mining
Omics
Gene Expression Profiling

ABSTRACT

Introduction: Because of huge impacts of "OMICS" technologies in life sciences, many researchers aim to implement such high throughput approach to address cellular and/or molecular functions in response to any influential intervention in genomics, proteomics, or metabolomics levels. However, in many cases, use of such technologies often encounters some cybernetic difficulties in terms of knowledge extraction from a bunch of data using related softwares. In fact, there is little guidance upon data mining for novices. The main goal of this article is to provide a brief review on different steps of microarray data handling and mining for novices and at last to introduce different PC and/or web-based softwares that can be used in preprocessing and/or data mining of microarray data. ***Methods:*** To pursue such aim, recently published papers and microarray softwares were reviewed. ***Results:*** It was found that defining the true place of the genes in cell networks is the main phase in our understanding of programming and functioning of living cells. This can be obtained with global/selected gene expression profiling. ***Conclusion:*** Studying the regulation patterns of genes in groups, using clustering and classification methods helps us understand different pathways in the cell, their functions, regulations and the way one component in the system affects the other one. These networks can act as starting points for data mining and hypothesis generation, helping us reverse engineer.

Introduction

Proteins, the amazing molecules of nature are almost involved in any activity in the cells from production of energy and biosynthesis of all component macromolecules to the maintenance of cellular architecture, and the ability to act upon intra- and extracellular stimuli. Each cell within an organism contains DNA which is crucial to produce the entire repertoire of proteins to cover the needs of an organism. The human genome project has determined the sequences that make up the human genome (3 billion base pairs). The number of human genes is estimated to be 30,000 to 100,000. It is now well known that the complementary sequences of most mRNA molecules could be transcribed in any biological process.

Only a portion of these genes are expressed and turned into functional proteins. However, some of the genes expressed in a single cell are likely to be present in all cells because they serve routine functions necessary for maintaining life in all cells and are called "housekeeping" genes. Other proteins serve specialized functions and are only required in particularly differentiated cell types for example, heart cells or neurons. Each cell's function determines the genes that have to be expressed in that specific type of cell.

Activities of a cell are highly controlled by cellular networks or more clearly the protein concentration. When any kind of change is imposed to the cell system these cellular networks and regulatory mechanisms become active and thus can be more readily detected. Global knowledge or a fingerprint of the transcriptional state could provide a wealth of information useful to biologists. This knowledge can be used in prediction of unknown genes functions, identification of biomarkers, target discovery, accurate diagnostics, development of prognostic tests and disease sub-class determination. At the very least, comparison of gene expression patterns in normal and pathological cells could provide useful diagnostic information and help identify genes that would be reasonable targets for therapeutic intervention (Afshari *et al.* 1999; Bednar 2000; Chin and Kong 2002;

***Corresponding author:** Yadollah Omidi (PhD), Tel.: + 98 411 3367914, Fax: +98 411 3367929, E-mail**:** yomidi@tbzmed.ac.ir

Dixon 2002; Dudda-Subramanya *et al.* 2003). Schematic steps of DNA microarray technology is shown in Fig. 1.

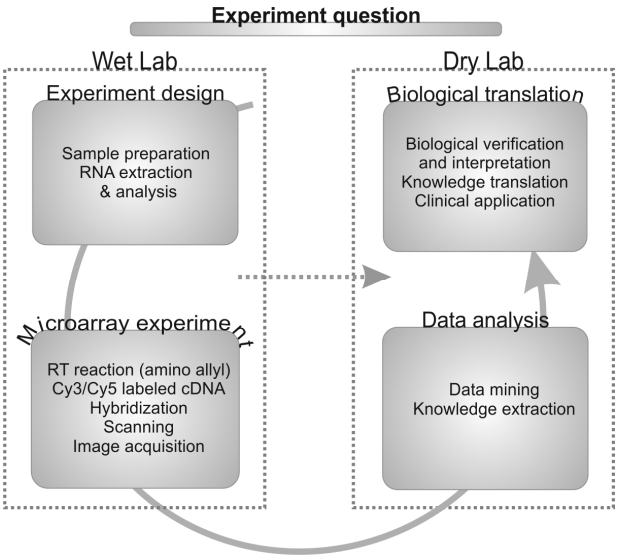

Fig. 1. Schematic steps of DNA microarray technology.

Microarray technology is a recent hybridization-based technique (gets back to 1990s) that allows simultaneous analysis and consequently estimation of an abundance of many nucleic acid species. Microarray has perhaps been so far, the most important revolution in functional genomics.

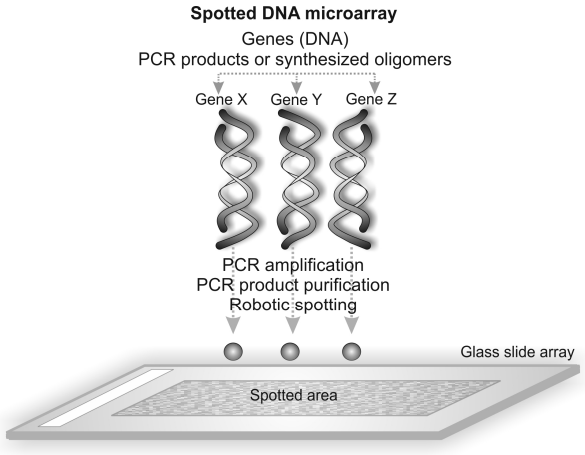

Fig. 2. Schematic illustration of spotted genes on a glass slide array. Glass slide arrays are produced by the robotic spotter that spots genes (e.g., PCR products, cDNAs, clone libraries or long oligonucleotides) onto coated glass slides. Each spot on the array represents a particular contiguous gene fragment, i.e. 40–70 nucleotides for oligonucleotide arrays, or several hundred nucleotides for PCR products.

As shown in Fig. 2, this technique involves robotic placement of individual, pure nucleic acid species on a glass surface. The entire complement of transcript mRNAs present in a particular cell type is extracted from cells and then a fluorophore-tagged cDNA representation of the extracted mRNAs is made *in vitro* by an enzymatic reaction termed *reverse transcription*. Then multiple fluorescently labeled nucleic acids are hybridized to the array, spots are detected and fluorophore-tagged hybrids are measured across the array with a scanning confocal microscope. The microarray technology is particularly useful for comparing the mRNAs from two cell types or two treatments.

Image capturing and analysis plus primary data extraction

Fluorophore-tagged representations of mRNA from two treatments, each tagged with a fluorophore emitting a different color light (usually green and red), are hybridized to the array of cDNAs and then fluorescence emission at the site of each immobilized cDNA is quantified and finally an image is produced. Measured fluorescent intensities ideally represent transcript levels in the sample. The main steps of the experimental approach of transcriptomic microarray are shown in Fig. 3 (panels A and B for wet and dry lab experiments, respectively).

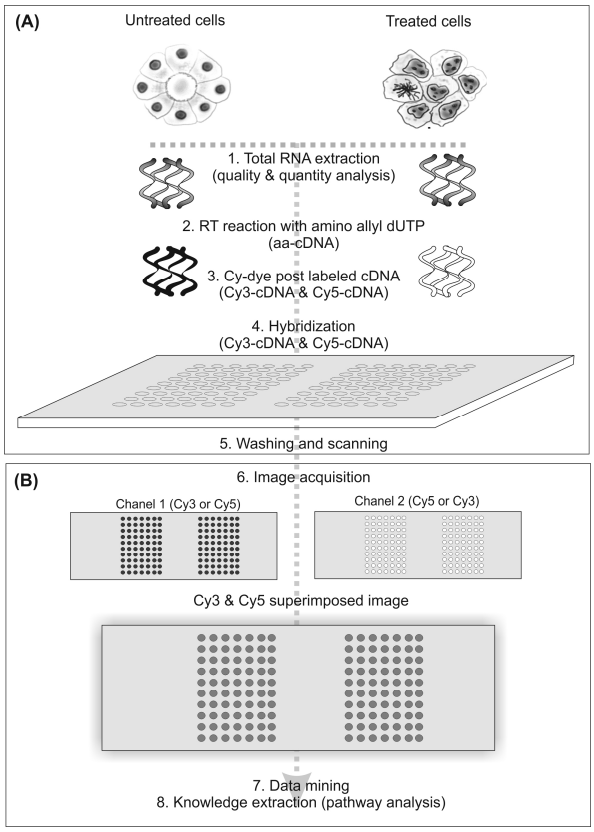

Fig. 3. Main steps of the experimental approach of transcriptomics DNA microarray for wet (A) and dry (B) lab experiments.

In single channel hybridization each slide is hybridized with a single biological sample labelled with a unique dye. Most new technologies follow this approach, e.g. Affymetrix, Agilent, Codelink. However in competitive hybridization each slide is hybridized with two biological samples each labelled with a different dye. Log ratios of the two color intensities ideally represent the relative abundance of the transcripts in one sample compared to the transcripts in the other one. The typical fluorescent images of hybridized cDNA microarray are shown in Fig. 4.

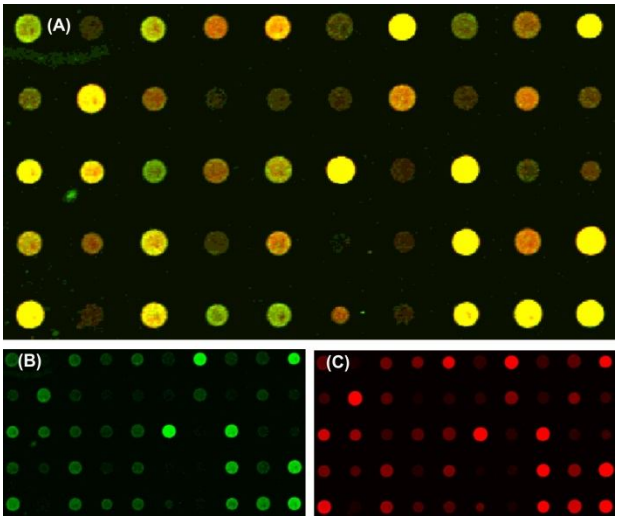

Fig. 4. Typical fluorescent images of hybridized cDNA microarray. A) Superimposed fluorescent image of Cy3-cDNA and Cy5-cDNA hybridization. B) Fluorescent images of Cy3 channel. C) Fluorescent images of Cy5 channel. Data are adapted with permission from (Barar *et al.* 2009).

Most manufacturers of microarray scanners provide their own software for image processing (Korn et al. 2004). Image analysis level of the experiment includes scanning of the image, spot recognition, gridding, segmentation and intensity extraction (plus background subtraction) respectively. Gridding is finding the true place of spots on the array and matching them with their corresponding IDs (Giannakeas *et al.* 2006; Lonardi and Luo 2004; Zacharia and Maroulis 2008). Segmentation is where the spots can be separated from the background. It defines the shape of each spot. Many methods are included in softwares for segmentation such as Fixed Circle Segmentation, Adaptive Circle Segmentation, Adaptive Shape Segmentation and Histogram Segmentation, with their names indicating their function and mechanism. The selection of the best method is dependent on the quality of the produced images, dominant shape of the spots and personal experience (Ahmed *et al.* 2004; Katzer *et al.* 2003; Lehmussola *et al.* 2006). The intensity of a spot in microarray needs to be corrected for the background intensities to reduce biases. A simple method called global correction is to subtract a constant from all spot intensity values. Another method is local correction which subtracts different values depending upon the location of a spot. A problem with these methods is when the background intensity is larger than the spot intensity. This results in a negative number and makes further analysis inappropriate (e.g., log transform). To address this issue, more sophisticated background correction methods have been proposed, such as a two-dimensional locally weighted linear regression (LOWESS) smoothing. By subtraction of the background, the intensity of the spots on the array can be measured.

Data mining

Fig. 5 represents various steps of DNA microarray data mining and its translation into clinical applications.

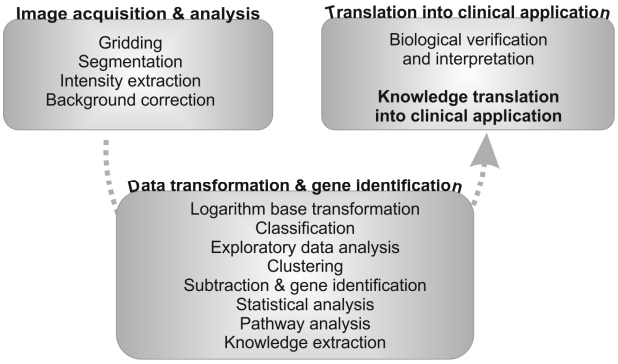

Fig. 5. Translation of DNA microarray data into clinical applications.

Normalization

Many sources of errors and inconsistencies may be involved in image processing. These include but of course not limited to irregularities in the array surface, variations in the laboratory processes, different DNA strands having different hybridization properties, different platforms being used in the process, different scanner settings, different amount of mRNA used, dye effect (different dyes have different efficiencies after all), random noise, and background effects. These inequalities necessitate normalization.

Normalization is the process that adjusts the individual hybridization intensities in a data matrix in order to balance them for the following data analysis. In other words, normalization is the process of correcting for bias within arrays and between arrays, prior to analysis (Kadanga *et al*. 2008). There are different normalization methods available. Global normalization, mean log centering, linear regression, LOWESS and rank invariant methods are mostly used (Chen 2003; Geller 2003; Quackenbush 2002). Among these methods, LOWESS analysis can remove intensity-dependent effects in the log2(ratio) values (Quackenbush 2002). LOWESS

normalization can detect systematic deviations in the R-I (ratio-intensity) plot. In this method, systematic variations are corrected by (1) a locally weighted linear regression as a function of the $\log_{10}$ (intensity) and (2) compensating the experimentally observed ratio with the best-fit average $\log_2$ (ratio). Table 1 represents important considerations in different steps of microarray data management.

Table 1. Considerations in different steps of microarray data management

Analysis step	Important considerations	References
Experimental design and implementation	Number of the replicates must be determined carefully Experimental errors should be avoided as much as possible The biological question behind the experiment should be defined carefully Information collection standards (MIAME) must be met	(Bolstad 2004; Churchill 2002; Foster 2002; Kerr 2003; Simon 2003)
Image acquisition and analysis	Image should be scanned at appropriate resolution Gridding step must be manually proofread Good choice of segmentation algorithm should be considered	(Istepanian 2003; Kadanga *et al.* 2008; Yang *et al.* 2002)
Data preprocessing and normalization	Poor quality spots and spots with intensity lower than the background plus two standard deviations should be discarded Log-transformation of the intensity ratio should be done LOWESS normalization is mostly used	(Cui 2003; Geller 2003; Quackenbush 2002)
Identification of differentially expressed genes	Use methods other than fixed threshold to infer significance Select a cut-off value for rejection of the null hypothesis that a gene is not differentially expressed	(Cui 2003; Gusnanto *et al.* 2007)
Dimension reduction	Use different methods to visualize the data from various perspectives	(Dai *et al*. 2006)
Supervised clustering or classification	Avoid over-training of the classifier Try more robust methods like neural networks Try to hold balance between the accuracy and generalizability Try different methods with different parameter settings to explore into the data	(Babyak 2004; Hawkins 2004; Jirapech-Umpai and Aitken 2005; Juan and Huang 2007; Khan *et al.* 2001)

After normalization the expression ratio can be calculated. The expression ratio is simply the normalized value of the expression level for a particular gene in the query sample divided by its normalized value for the control.

Then the ratio (T) for gene i can be written as:

$$T_i = \frac{R_i}{G_i}$$

where R and G represent the red (target) and green (reference) intensities .

The very basic preprocessing step is taking logarithm of each entry in gene expression data matrix in order to expand the dynamic range of gene expression signals. This is called log transformation.

$$T_i^{'} = \log_2\left(\frac{R_i}{G_i}\right)$$

Dealing with missing values

The gene expression data matrix may have missing values due to non-systematic inconsistencies such as pollution on the glass, image corruption during scanning, low resolution images, as well as systematic errors occurring in the microarray manufacturing process. Missing value estimation is important for at least two reasons. First, some popular analysis methods such as *principal component analysis* (PCA) require the complete data matrices to function. Second, most data mining methods can benefit from having accurate estimation of missing values. Model-based methods may be the most popular. Other common techniques include nearest neighbor methods, iterative analysis of variance methods, filling in least squares estimates (Bo *et al.* 2004; Kim *et al.* 2004), randomized inference, and likelihood-based approaches (Troyanskaya 2001). In the context of microarray, sometimes simple techniques such as replacing missing values with zeros or the average of the corresponding row or column are sufficient. However, these methods are not optimal because they do not consider problem-specific

information that may be useful for better estimation. More sophisticated approaches have been also proposed. For example, the KNNimpute algorithm aims at minimizing data modelling assumptions and takes advantage of the correlation structure of the gene expression data by using genes with expression profiles similar to the gene of interest. As an another example, the SVDimpute method exploits SVDs to estimate missing values by obtaining a set of mutually orthogonal expression patterns that can be linearly combined to approximate the expression of all genes in the data.

Identification of differentially expressed genes

All microarray experiments are carried out to find genes which are differentially expressed between two (or more) samples of cells (2000; Abiko *et al.* 2004; Acin *et al.* 2007; Caetano *et al.* 2004; De *et al.* 2004). This goal has two prerequisites. The first is to select a statistic which will rank the genes in order of evidence (significance) for differential expression, from strongest to weakest evidence. The second is to choose a critical-value for the ranking statistic above which any value is considered significant. Filtering unnecessary data has some advantages. First of all, not only elimination of the unchanged genes will help data mining procedures easier to handle but will also fasten them. The primary importance of ranking arises however, from the fact that only a limited number of genes can be followed up in a typical biological study due to limited resources. Methods used in finding differentially expressed genes are fixed cut-off threshold (usually 2 fold), unusual ratio, univariate statistical tests e.g. the t-test (Neely 2003) in case the samples are independent and F test or ANOVA (Pavlidis 2003) in case the number of conditions under study is more than two. All above methods assume that data follows normal distribution. When normal distribution criteria are not met, non-parametric tests like Kruskal-Wallis procedure (instead of one-way ANOVA) or Friedman procedure (instead of two-way ANOVA) are used.

With those tests that use *P values* Bonferonni correction is used to reduce the number of false discovery rate (FDR) by reducing the significance cut-off. *P value* is a popular cut-off and is defined to be the minimum false positive rate at which an observed statistic can be called significant. Genes with *P values* smaller than the set threshold are more probably significant (Cheng and Pounds 2007; Grant *et al.* 2005; Gusnanto *et al.* 2007; Pawitan *et al.* 2005; Tsai *et al.* 2003).

Higher level analysis of microarray data

Once differentially expressed genes have successfully been distinguished, high level analyses or data mining of microarray data begins. Data Mining is all about automating the process of searching for patterns in the data. In other words, it is an iterative process of discovery.

Dimension reduction

The complexity of most data analysis algorithms depends on the number of input dimensions, so reducing the number of genes or experimental conditions in a microarray data set is helpful for efficient analysis, as long as the reduced data set maintains important information in the original data (Bura and Pfeiffer 2003; Dai *et al.* 2006).

Dimensionality reduction algorithms can be classified into feature selection and feature extraction. *Feature selection* is to select k dimensions, out of the original d dimensions, that can best represent the original data set (Chen *et al.* 2007; Jirapech-Umpai and Aitken 2005). *Feature extraction* is to find a new set of k dimensions that are some combinations of the original d dimensions. The most popular feature extraction algorithms may be the linear projection method such as *principal component analysis* (PCA) for unsupervised learning (Li *et al*. 2008) and *linear discriminant analysis* (LDA) for supervised learning (Shen *et al*. 2006). PCA is also called *singular value decomposition* (SVD) depending on the context.

Other methods used in dimension reduction are Independent Component Analysis (Saidi *et al.* 2004; Zheng *et al.* 2008) (ICA), Correspondence Analysis (CA) (Fellenberg *et al.* 2001; Kishino and Waddell 2000; Tan *et al.* 2004) and Multidimensional Scaling (MDS).

Clustering and classification

When one has done multiple experiments, under different conditions -different patients, different time points, and etc- one can group the genes, which behave similarly and based on the pattern of the distinguishing genes, one can for example set boundaries between different subtypes of cancer. One can identify samples with similar expression level patterns or genes which are similar across samples. The main aim is to look for the most different features that should be the best at discriminating classes. Among different approaches used to pursue such aim, the "Euclidean distance clustering method" seems to be the commonest methodology. Fig. 6 represents the schematic illustration of Euclidean distance clustering method for expressed genes.

Supervised approaches are the analyses which are designed to determine the genes that fit a predetermined pattern. In the case of a supervised learning, one can use the annotation of either the gene or the sample, and create clusters of genes or samples in order to identify patterns that are characteristic for the cluster. In other words one can specify relationships among objects in supervised learning (Jirapech-Umpai and Aitken 2005). The main goal of supervised learning is data classification and subsequently prediction. Unlike supervised learning, unsupervised methods are used to characterize the components of a data set without the a

priori input or knowledge of a training signal; i.e. in the case of an unsupervised learning, the expression data is analyzed to identify patterns that can group genes or samples into clusters without the use of any form of annotation (Boutros and Okey 2005). However, annotation information may be taken into account at a later stage in unsupervised learning to make meaningful biological inferences (Redestig *et al*. 2007).

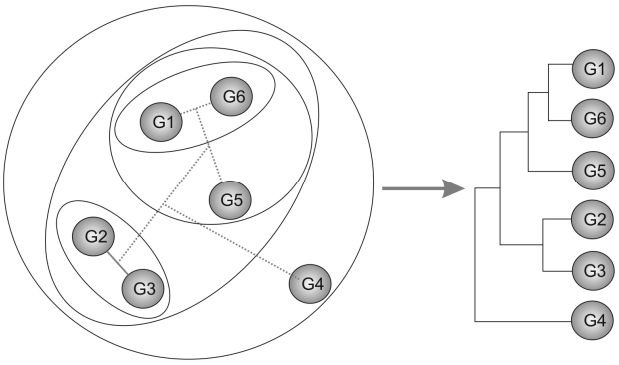

Fig. 6. Schematic illustration of Euclidean distance clustering of expressed genes (G).

The most commonly used popular supervised techniques are nearest neighbors (Mezghani *et al.* 2008; Shen and Hasegawa 2008; Shen and Chou 2005), support vector machines (Brown *et al.* 2000; Chu and Wang 2005; Furey *et al.* 2000) and neural networks (Khan *et al.* 2001; Lancashire *et al.* 2009; Linder *et al.* 2007; O'Neill and Song 2003; Ringner and Peterson 2003). The most common unsupervised techniques are hierarchical clustering (Chipman and Tibshirani 2006; Makretsov *et al.* 2004); k means clustering (Steinley 2006; Wu 2008), self-organizing maps (Covell *et al*. 2003), relevance networks and principal-components analysis (Liu *et al.* 2002; Wang and Gehan 2005).

The methods presented up until now are correlative methods. These methods cluster genes together according to the measure of correlation between them. Genes that are clustered together may *and only may* imply that they participate at the same biological process. However these methods are computationally cheap and one cannot infer the relationships between the genes. The basic questions in functional genomics are: "How is the expression of this specific gene affected by the expression of other genes in the cell?" and "Which other genes are under the influence of this gene?"

Reverse engineering of gene regulatory networks

Perhaps the most recent and the most important part in microarray data analysis is reverse engineering of gene regulatory networks for understanding the dynamics of gene expression. Pathway analysis towards functional enrichment can be fulfilled using two methods one of which is *time-series data* (Dewey 2002; Filkov *et al.* 2002; Klevecz *et al.* 2007; Maraziotis *et al.* 2007) and second one is *steady-state data of gene knockouts* (Rawool and Venkatesh 2007).

In the former approach the amount of expression of a certain gene at a certain time is a function of expression of the other genes at all previous time points. In the latter approach, the effects of deleting a certain gene on the expression of other genes are inspected and based on the regulation of the other genes; the function of that certain gene in regulation of the other genes is assessed. These methods still lack full applicability, because there is a need for more knowledge on sophisticated networks in the cells in order to identify the hidden role of different molecules in the circuitry of gene regulation.

Understanding the expression dynamics helps us infer innate complexities and phenomenological networks among genes. Defining the true place of the genes in cell networks is the main phase in our understanding of programming and functioning of living cells. Studying the regulation patterns of genes in groups, using clustering and classification methods helps us understand different pathways in the cell, their functions, regulations and the way one component in the system affects the other one. These networks can act as starting points for data mining and hypothesis generation, helping us reverse engineer.

So far various softwares have been used for image acquisition/processing and data mining. Table 2 represents some important softwares available for handling of microarray data. Of these softwares, some of them such as TM4 are freely available while some others such as ImaGen and GeneSight are commercially available. Among these tools, some deal with gene ontology which may help us towards better understanding of function genomics. For example, the Expression Analysis Systematic Explorer (EASE) developed by Database for Annotation, Visualization and Integrated Discovery (DAVID) Bioinformatics team, is a customizable, standalone, Windows(c) desktop software application that facilitates the biological interpretation of gene lists derived from the results of microarray, proteomic, and SAGE experiments.

Table 2. Softwares available for preprocessing and/or data mining of microarray data

Software	Methods available in the package	Advantages and/or features	Related URLs
ImaGene	Automated grid-finding, spot-finding, spot & array-level quality control, segmentation and normalization	ImaGene works with all scanners including Tecan, Agilent, GenePix (Axon), Perkin Elmer, Innopysis, and more; all kinds of arrays are supported (glass, filter, membrane, custom, or commercial).	Free trial version at http://www.biodiscovery.com/index/imagene
GeneMaths XT	Normalization, unsupervised learning, supervised learning, Analysis of variance and multivariate analysis. Analysis tools are also available for time- course experiments.	Complete and professional for data mining of microarray results Integrated error handling and hypothesis testing tools	http://www.applied-maths.com/genemaths/genemaths.htm
Avadis	Data analysis and visualization	Avadis has a built-in GO browser to view ontology hierarchies. Prebuilt Affymetrix workflows, Avadis is highly tuned to work with Affymetrix GeneChip® data.	http://avadis.strandgenomics.com/
(DAVID) Database for Annotation, Visualization and Integrated Discovery	Integrated solutions for the annotation and analysis of datasets	David can identify enriched biological themes, particularly GO terms, discover enriched functional-related gene groups, visualize genes on BioCarta& KEGG pathway maps, list interacting proteins, explore gene names in batch, link gene-disease associations and highlight protein functional domains and motifs.	http://david.abcc.ncifcrf.gov/
EASE (Expression Analysis Systematic Explorer)	EASE, developed by DAVID Bioinformatics team, is a customizable, standalone, Windows(c) desktop software application that facilitates the biological interpretation of gene lists derived from the results of microarray, proteomic, and SAGE experiments.	EASE provides statistical methods for discovering enriched biological themes within gene lists, generates gene annotation tables, and enables automated linking to online analysis tools.	
EGAN (Exploratory Gene Association Networks)	Visualizing and interpreting the results of high-throughput exploratory assays in an interactive hypergraph of genes, relationships (protein-protein interactions, literature co-occurrence, etc.) and meta-data (annotation, signaling pathways, etc.). EGAN provides comprehensive, automated enrichment analysis	Links to external web resources including more than 240 000 articles at PubMed, hypergeometric and GSEA-like enrichment statistics	
FunCluster	Detecting co-regulated biological processes involving	FunCluster's functional analysis relies on GO and KEGG annotations and is currently available for three organisms: Homo sapiens, Mus musculus and Saccharomyces cerevisiae.	Software can be downloaded from the FunCluster website, or from the worldwide mirrors of CRAN. FunCluster is provided freely under the GNU General Public License 2.0.
FunNet	A tool for exploring transcriptional interactions in gene expression datasets.	FunNet is provided both as a web-based tool and as a standalone R package.	http://corneliu.henegar.info/FunNet.htm, http://www.geneontology.org/GO.tools.microarray.shtml#funnet
GOALIE (Generalized Ontological Algorithmic Logical Invariants Extractor)	GOALIE is a tool for the construction of time-course dependent enrichments in pathway analysis.		http://bioinformatics.nyu.edu/Projects/GOALIE/ http://bioinformatics.nyu.edu/~marcoxa/work/GOALIE/
GeneSight™	Data analysis and data mining tool (GenePie™ visualization; 2-D scatter plots; interactive ratio histogram plotting, hierarchical, K-means, and neural network clustering, PCA, and time series. The confidence analyzer tool can use replicated gene expression data for identifying genes having true differential expression.)	GeneSight is fully integrated with BioDiscovery'sImaGene image analysis program. GeneSight can easily import array data contained in any text-based file format.	http://www.biodiscovery.com/genesight.asp
GeneSifter	Statistical framework with 15 advanced options		http://www.genesifter.net/web/
Expression Profiler	Clustering, pattern discovery, statistics (thru R), machine-learning algorithms and visualization.		http://www.ebi.ac.uk/expressionprofiler/

Software	Methods available in the package	Advantages and/or features	Related URLs
GenMAPP	Visualizing gene expression data on maps representing biological pathways and groupings of genes.	Integrated with GenMAPP are programs to perform a global analysis of gene expression or genomic data in the context of hundreds of pathway MAPPs and thousands of GO terms (MAPPFinder), import lists of genes/proteins to build new MAPPs (MAPPBuilder), and export archives of MAPPs and expression/genomic data to the web.	www.genmapp.org/
Bioconductor	Bioconductor comes with many packages that cover the very parts of gene expression data mining.	Different packages are available in the Bioconductor website. When released BioC 2.5, consisted of 352 packages. For more information refer to http://www.bioconductor.org	http://www.bioconductor.org/
SNOMAD (Standardization and NOrmalization of MicroArray Data)	Web Tools for the Standardization and Normalization of MicroArray Data. Useful mostly for paired microarray data.	Free and user-friendly software	http://pevsnerlab.kennedykrieger.org/snomadinput.html
RelNet	Relevance Networks	The software is written in Java and it runs under any operating system. It dynamically determines the latest names, symbols, functions, and genome position for each gene and includes these in the relevance networks output.	http://chip.org/relnet/
BASE (BioArray Software Environment)	Web-based microarray database and analysis platform		http://base.thep.lu.se/
Partek Genomics Suite	Advanced statistics and interactive data visualization specifically designed to extract biological signals from noisy data.		http://www.partek.com
TM4	The TM4 suite of tools consist of four major applications, Microarray Data Manager (MADAM), TIGR Spotfinder (image processing tool), Microarray Data Analysis System (MIDAS), and Multi-experiment Viewer (MeV), as well as a Minimal Information About a Microarray Experiment (MIAME)-compliant MySQL database.	MeV identifies patterns of gene expression and differentially expressed genes MADAM is a java-based application to load and retrieve microarray data to and from a database. TIGR Spotfinder is an image processing software. MIDAS is a microarray data quality filtering and normalization tool.	http://www.tm4.org/
BNArray	Constructing gene regulatory networks using Bayesian networks	BNArraycan handle microarray datasets with missing data.	http://www.cls.zju.edu.cn/binfo/BNArray/
ArrayPipe	Application features range from quality assessment of slides through various data visualizations to normalization and detection of differentially expressed genes.		http://www.pathogenomics.ca/arraypipe/ koch.pathogenomics.ca/cgi-bin/pub/arraypipe.pl
BRB array tools	Visualization and statistical analysis of microarray gene expression data	The software provides tools available for predictive classifier development and complete cross-validation. It offers links to genomic websites for gene annotation and analysis tools for pathway analysis.	http://linus.nci.nih.gov/BRB-ArrayTools.html
Vector Xpression™	Storing, managing, and analyzing (Normalizing, Identify differentially expressed genes and classification)		http://register.informaxinc.com/solutions/xpression/main.html http://www.informaxinc.com/downloads.html
Engene	Visualizing, preprocessing and clustering	Clustering analysis algorithms include k-means, HAC, fuzzy c-means, kernel c-means, SOMs and PCA	https://chirimoyo.ac.uma.es/engenet/
ExpressYourself	The software performs correction of the background array signal, normalization, scoring, combination of replicate experiments, filtering problematic regions of the array and quality assessment of hybridizations.	ExpressYourself investigates the quality of experiments by measuring hybridization consistency within single slides and across replicated experiments. The data quality step calculates the overall performance of experiments and highlights problematic array regions.	Freely available at http://bioinfo.mbb.yale.edu/expressyourself/
fCluster	Fuzzy clustering of microarray data		http://fuzzy.cs.uni-magdeburg.de/fcluster/
GEPAS (Gene Expression Pattern Analysis Suite)	Normalization, and preprocessing such as log transformation, replicate handling and missing value imputation. It supports hierarchical clustering and SOMs for data clustering.	On-line tutorials are available from main web server(http://bioinfo.cnio.es).	http://gepas.bioinfo.cnio.es
Genes@Work	Genes@Work is a pattern discovery and classification system.		http://www.research.ibm.com/FunGen/FGGenesAtWorkDoc.html
SilicoCyte	Automated image analysis, data annotation, analysis and visualization.	SilicoCyte supports integration with advanced visualization tools and LIMSs (LIMS).	http://www.cytogenomic.com/silicocyte.htm
Spotfire	Spotfire allows users to interactively mine, visualize, and analyze large sets of technical, multidimensional data.		http://spotfire.tibco.com/

Software	Methods available in the package	Advantages and/or features	Related URLs
STEM(Short Time-series Expression Miner)	Clustering, comparing, and visualizing short time series gene expression data (8 time points or fewer).		http://www.cs.cmu.edu/~jernst/stem/ Evaluation license at http://www.andrew.cmu.edu/user/zivbj/stemevaluationreg.html
Gene ARMADA (Automated Robust MicroArray Data Analysis)	Automated data import, noise correction and filtering, normalization, statistical selection of differentially expressed genes, clustering, classification and annotation	Besides being fully automated, Gene ARMADA incorporates numerous functionalities of the Statistics and Bioinformatics Toolboxes of MATLAB.	
ArrayXPath	Mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics (SVG).	ArrayXPath is empowered by integrating gene-pathway, disease-pathway, drug-pathway and pathway–pathway correlations with integrated GO, Medical Subject Headings and OMIM Morbid Map-based annotations.	http://www.snubi.org/software/ArrayXPath/
CARMAweb (Comprehensive R-based Microarray Analysis web service)	Data preprocessing (background correction, quality control and normalization), detection of differentially expressed genes, cluster analysis, dimension reduction and visualization, classification, and GO-term analysis.		CARMAweb is freely available at https://carmaweb.genome.tugraz.at.
GoMiner	GoMiner is a program for visualizing the genes on a list within the context of the structure of the GO.	Instead of analyzing microarray results with a gene-by-gene approach, GoMiner classifies the genes into biologically coherent categories and assesses these categories.	http://discover.nci.nih.gov/gominer/index.jsp
MARS (microarray analysis, retrieval, and storage system)	Image analysis, normalization, gene expression clustering, and mapping of gene expression data onto biological pathways		http://genome.tugraz.at/mars/mars_description.shtml
ChipInspector	Detecting differentially expressed genes and clustering.		http://www.genomatix.de/products/ChipInspector/index.html
dChip	Probe-level (e.g. Affymetrix platform) and high-level analysis of gene expression microarrays and SNP microarrays.	High-level analysis in dChip includes comparing samples, hierarchical clustering, viewing expression and SNP data along chromosome	http://www.biostat.harvard.edu/complab/dchip
L2L	L2L is a database of published microarray gene expression data, and a software tool for comparing that published data with a user's own microarray results.		http://depts.washington.edu/l2l/
KDE	Visualizations and normalizations (including RMA, Li&Wong methods), preprocessing, statistical analysis, unsupervised and supervised analysis procedures		http://www.inforsense.com/kde.html
SeqExpress	A number of clustering and analysis techniques; integrated gene expression and analysis result visualizations, integration with the Gene Expression Omnibus and an optional data sharing architecture	SeqExpress is free and runs under Windows.	www.seqexpress.com/

Ethical issues

None to be declared.

Conflict of interests

Authors declare no conflict of interest.

Acknowledgement

Authors are grateful to the Ministry of Health, Care and Medical Education for the financial support.

References

2000.Global Analysis of Differential Gene Expression Between Prostate Cancer and Normal Prostate Tissues Using CDNA Microarray of Open-Reading Frame Expressed Sequence Tags (ORESTES). *Prostate Cancer Prostatic Dis,* 3(S1), S36.

Abiko Y, Hiratsuka K, Kiyama-Kishikawa M, Tsushima K, Ohta M and Sasahara H. **2004**.Profiling of Differentially Expressed Genes in Human Gingival Epithelial Cells and Fibroblasts by DNA Microarray. *J Oral Sci,* 46(1), 19-24.

Acin S, Navarro MA, Perona JS, Surra JC, Guillen N, Arnal C *et al.* **2007**.Microarray Analysis of Hepatic Genes Differentially Expressed in the Presence of the Unsaponifiable Fraction of Olive Oil in Apolipoprotein E-Deficient Mice. *Br J Nutr,* 97(4), 628-638.

Afshari CA, Nuwaysir EF and Barrett JC. **1999**.Application of Complementary DNA Microarray Technology to Carcinogen Identification, Toxicology, and Drug Safety Evaluation. *Cancer Res,* 59(19), 4759-60.

Ahmed AA, Vias M, Iyer NG, Caldas C and Brenton JD. **2004**.Microarray Segmentation Methods Significantly Influence Data Precision. *Nucleic Acids Res,* 32(5), e50.

Babyak MA. **2004**.What You See May Not Be What You Get: a Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. *Psychosom Med,* 66(3), 411-421.

Barar J, Hamzeiy H, Mortazavi-Tabatabaei SA, Hashemi-Aghdam SE and Omidi Y. **2009**.Genomic Signature and

Toxicogenomics Comparison of Polycationic Gene Delivery Nanosystems in Human Alveolar Epithelial A549 Cells. *Daru,* 17(3), 139-147.

Bednár M. **2000**.DNA Microarray Technology and Application.). *Med Sci Monit,* 6(4), 796-800.

Bo TH, Dysvik B and Jonassen I. **2004**.LSimpute: Accurate Estimation of Missing Values in Microarray Data With Least Squares Methods. *Nucleic Acids Res,* 32(3), e34.

Bolstad BM, Collin F, Simpson KM, Irizarry RA, Speed TP. **2004**.Experimental Design and Low-Level Analysis of Microarray Data. *Int Rev Neurobiol,* 60:25-58.

Boutros PC and Okey AB. **2005**.Unsupervised Pattern Recognition: an Introduction to the Whys and Wherefores of Clustering Microarray Data. *Brief Bioinform,* 6(4), 331-343.

Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS *et al.* **2000**.Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines. *Proc Natl Acad Sci U S A,* 97(1), 262-267.

Bura E and Pfeiffer RM. **2003**.Graphical Methods for Class Prediction Using Dimension Reduction Techniques on DNA Microarray Data. *Bioinformatics,* 19(10), 1252-8.

Caetano AR, Johnson RK, Ford JJ and Pomp D. **2004**.Microarray Profiling for Differential Gene Expression in Ovaries and Ovarian Follicles of Pigs Selected for Increased Ovulation Rate. *Genetics,* 168(3), 1529-1537.

Chen YJ, Kodell R, Sistare F, Thompson KL, Morris S, Chen JJ. **2003**.Normalization Methods for Analysis of Microarray Gene-Expression Data. *J Biopharm Stat,*13(1):57-74.

Chen Z, Li J and Wei L. **2007**.A Multiple Kernel Support Vector Machine Scheme for Feature Selection and Rule Extraction From Gene Expression Data of Cancer Tissue. *Artif Intell Med,* 41(2), 161-175.

Cheng C and Pounds S. **2007**.False Discovery Rate Paradigms for Statistical Analyses of Microarray Gene Expression Data. *Bioinformation,* 1(10), 436-446.

Chin KV and Kong AN. **2002**.Application of DNA Microarrays in Pharmacogenomics and Toxicogenomics. *Pharm Res,* 19(12), 1773-1778.

Chipman H and Tibshirani R. **2006**.Hybrid Hierarchical Clustering With Applications to Microarray Data. *Biostatistics,* 7(2), 286-301.

Chu F and Wang L. **2005**.Applications of Support Vector Machines to Cancer Classification With Microarray Data. *Int J Neural Syst,* 15(6), 475-484.

Churchill GA. **2002**.Fundamentals of Experimental Design for CDNA Microarrays. *Nat Genet,* 32 Suppl, 490-5.

Covell DG, Wallqvist A, Rabow AA and Thanki N. **2003**.Molecular Classification of Cancer: Unsupervised Self-Organizing Map Analysis of Gene Expression Microarray Data. *Mol Cancer Ther,* 2(3), 317-332.

Cui X, Churchill GA. **2003**.Statistical Tests for Differential Expression in CDNA Microarray Experiments. *Genome Biol,* 4(4), 210. Epub 2003 Mar 17.

Dai JJ, Lieu L, Rocke D. **2006**.Dimension Reduction for Classification With Gene Expression Microarray Data. *Stat Appl Genet Mol Biol,* Article6. Epub 2006 Feb 24.

De K, Ghosh G, Datta M, Konar A, Bandyopadhyay J, Bandyopadhyay D *et al.* **2004**.Analysis of Differentially Expressed Genes in Hyperthyroid-Induced Hypertrophied Heart by CDNA Microarray. *J Endocrinol,* 182(2), 303-314.

Dewey TG. **2002**.From Microarrays to Networks: Mining Expression Time Series. *Drug Discovery Today,* 7(20), s170-s175.

Dixon B. **2002**.Microarray Technology: an Array of Applications That Is Far From Micro. *Biotechnol Adv,* 20(5-6), 361-362.

Dudda-Subramanya R, Lucchese G, Kanduc D and Sinha AA. **2003**.Clinical Applications of DNA Microarray Analysis. *J Exp Ther Oncol,* 3(6), 297-304.

Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD and Vingron M. **2001**.Correspondence Analysis Applied to Microarray Data. *Proc Natl Acad Sci U S A,* 98(19), 10781-10786.

Filkov V, Skiena S and Zhi J. **2002**.Analysis Techniques for Microarray Time-Series Data. *J Comput Biol,* 9(2), 317-330.

Foster WR, Huber RM. **2002**.Current Themes in Microarray Experimental Design and Analysis. *Drug Discov Today,* 7(5), 290-2.

Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M and Haussler D. **2000**.Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. *Bioinformatics,* 16(10), 906-914.

Geller SC, Gregg JP, Hagerman P, Rocke DM. **2003**.Transformation and Normalization of Oligonucleotide Microarray Data. *Bioinformatics,* 19(14),1817-23.

Giannakeas N, Fotiadis DI and Politou AS. **2006**.An Automated Method for Gridding in Microarray Images. *Conf Proc IEEE Eng Med Biol Soc,* 1, 5876-5879.

Grant GR, Liu J and Stoeckert CJ, Jr. **2005**.A Practical False Discovery Rate Approach to Identifying Patterns of Differential Expression in Microarray Data. *Bioinformatics,* 21(11), 2684-2690.

Gusnanto A, Calza S and Pawitan Y. **2007**.Identification of Differentially Expressed Genes and False Discovery Rate in Microarray Studies. *Curr Opin Lipidol,* 18(2), 187-193.

Hawkins DM. **2004**.The Problem of Overfitting. *J Chem Inf Comput Sci,* 44(1), 1-12.

Istepanian RS. **2003**.Microarray Image Processing: Current Status and Future Directions. *IEEE Trans Nanobioscience,* 2(4),173-5.

Jirapech-Umpai T and Aitken S. **2005**.Feature Selection and Classification for Microarray Data Analysis: Evolutionary Methods for Identifying Predictive Genes. *BMC Bioinformatics,* 6, 148.

Juan HF and Huang HC. **2007**.Bioinformatics: Microarray Data Clustering and Functional Classification. *Methods Mol Biol,* 382), 405-416.

Kadanga AK, Leroux C, Bonnet M, Chauvet S, Meunier B, Cassar-Malek I, Hocquette JF. **2008**.Image Analysis and Data Normalization Procedures Are Crucial for Microarray Analyses. *Gene Regul Syst Bio,* 17(2), 107-112.

Katzer M, Kummert F and Sagerer G. **2003**.Methods for Automatic Microarray Image Segmentation. *IEEE Trans Nanobioscience,* 2(4), 202-214.

Kerr MK. **2003**.Experimental Design to Make the Most of Microarray Studies. *Methods Mol Biol,* 224,137-47.

Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F *et al.* **2001**.Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks. *Nat Med,* 7(6), 673-679.

Kim H, Golub GH and Park H. **2004**.Missing Value Estimation for DNA Microarray Gene Expression Data: Local Least Squares Imputation. *Bioinformatics,* 21(2),187-98. Epub 2004 Aug 27.

Kishino H and Waddell PJ. **2000**.Correspondence Analysis of Genes and Tissue Types and Finding Genetic Links From Microarray Data. *Genome Inform Ser Workshop Genome Inform,* 11), 83-95.

Klevecz RR, Li CM and Bolen JL. **2007**.Signal Processing and the Design of Microarray Time-Series Experiments. *Methods Mol Biol,* 377, 75-94.

Korn EL, Habermann JK, Upender MB, Ried T and McShane LM. **2004**.Objective Method of Comparing DNA Microarray Image Analysis Systems. *Biotechniques,* 36(6), 960-967.

Lancashire LJ, Lemetre C and Ball GR. **2009**.An Introduction to Artificial Neural Networks in Bioinformatics--Application to Complex Microarray and Mass Spectrometry Datasets in Cancer Studies. *Brief Bioinform,* 10(3), 315-29. Epub 2009 Mar 23.

Lehmussola A, Ruusuvuori P and Yli-Harja O. **2006**.Evaluating the Performance of Microarray Segmentation Algorithms. *Bioinformatics,* 22(23), 2910-2917.

Li GZ, Bu HL, Yang MQ, Zeng XQ and Yang JY. **2008**.Selecting Subsets of Newly Extracted Features From PCA and PLS in Microarray Data Analysis. *BMC Genomics,* 9 Suppl 2, S24.

Linder R, Richards T and Wagner M. **2007**.Microarray Data Classified by Artificial Neural Networks. *Methods Mol Biol,* 382, 345-372.

Liu A, Zhang Y, Gehan E and Clarke R. **2002**.Block Principal Component Analysis With Application to Gene Microarray Data Classification. *Stat Med,* 21(22), 3465-3474.

Lonardi S and Luo Y. **2004**.Gridding and Compression of Microarray Images. *Proc IEEE Comput Syst Bioinform Conf,* 122-130.

Makretsov NA, Huntsman DG, Nielsen TO, Yorida E, Peacock M, Cheang MC *et al.* **2004**.Hierarchical Clustering Analysis of Tissue Microarray Immunostaining Data Identifies Prognostically Significant Groups of Breast Carcinoma. *Clin Cancer Res,* 10(18 Pt 1), 6143-6151.

Maraziotis IA, Dragomir A and Bezerianos A. **2007**.Gene Networks Reconstruction and Time-Series Prediction From Microarray Data Using Recurrent Neural Fuzzy Networks. *IET Syst Biol,* 1(1), 41-50.

Mezghani N, Husse S, Boivin K, Turcot K, Aissaoui R, Hagemeister N *et al.* **2008**.Automatic Classification of Asymptomatic and Osteoarthritis Knee Gait Patterns Using Kinematic Data Features and the Nearest Neighbor Classifier. *IEEE Trans Biomed Eng,* 55(3), 1230-1232.

Neely JG, Hartman JM, Forsen JW Jr, Wallace MS. **2003**.Tutorials in Clinical Research: VII. Understanding Comparative Statistics (Contrast)--Part B: Application of T-Test, Mann-Whitney U, and Chi-Square. *Laryngoscope,* 113(10), 1719-25.

O'Neill MC and Song L. **2003**.Neural Network Analysis of Lymphoma Microarray Data: Prognosis and Diagnosis Near-Perfect. *BMC Bioinformatics,* 4:13. Epub 2003 Apr 10.

Pavlidis P. **2003**.Using ANOVA for Gene Selection From Microarray Studies of the Nervous System. *Methods,* 31(4), 282-9.

Pawitan Y, Michiels S, Koscielny S, Gusnanto A and Ploner A. **2005**.False Discovery Rate, Sensitivity and Sample Size for Microarray Studies. *Bioinformatics,* 21(13), 3017-3024.

Quackenbush J. **2002**.Microarray Data Normalization and Transformation. *Nat Genet,* 32 Suppl, 496-501.

Rawool SB and Venkatesh KV. **2007**.Steady State Approach to Model Gene Regulatory Networks--Simulation of Microarray Experiments. *Biosystems,* 90(3), 636-655.

Redestig H, Repsilber D, Sohler F and Selbig J. **2007**.Integrating Functional Knowledge During Sample Clustering for Microarray Data Using Unsupervised Decision Trees. *Biom J,* 49(2), 214-229.

Ringner M and Peterson C. **2003**.Microarray-Based Cancer Diagnosis With Artificial Neural Networks. *Biotechniques,* Suppl, 30-5.

Saidi SA, Holland CM, Kreil DP, MacKay DJ, Charnock-Jones DS, Print CG *et al.* **2004**.Independent Component Analysis of Microarray Data in the Study of Endometrial Cancer. *Oncogene,* 23(39), 6677-6683.

Shen F and Hasegawa O. **2008**.A Fast Nearest Neighbor Classifier Based on Self-Organizing Incremental Neural Network. *Neural Netw,* 21(10), 1537-1547.

Shen H and Chou KC. **2005**.Using Optimized Evidence-Theoretic K-Nearest Neighbor Classifier and Pseudo-Amino Acid Composition to Predict Membrane Protein Types. *Biochem Biophys Res Commun,* 334(1), 288-292.

Shen R, Ghosh D, Chinnaiyan A and Meng Z. **2006**.Eigengene-Based Linear Discriminant Model for Tumor Classification Using Gene Expression Microarray Data. *Bioinformatics,* 22(21), 2635-2642.

Simon RM, Dobbin K. **2003**.Experimental Design of DNA Microarray Experiments. *Biotechniques,* Suppl, 16-21.

Steinley D. **2006**.K-Means Clustering: a Half-Century Synthesis. *Br J Math Stat Psychol,* 59(Pt 1), 1-34.

Tan Q, Brusgaard K, Kruse TA, Oakeley E, Hemmings B, Beck-Nielsen H *et al.* **2004**.Correspondence Analysis of Microarray Time-Course Data in Case-Control Design. *J Biomed Inform,* 37(5), 358-365.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. **2001**.Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics,* 17(6), 520-5.

Tsai CA, Hsueh HM and Chen JJ. **2003**.Estimation of False Discovery Rates in Multiple Testing: Application to Gene Microarray Data. *Biometrics,* 59(4), 1071-1081.

Wang A and Gehan EA. **2005**.Gene Selection for Microarray Data Analysis Using Principal Component Analysis. *Stat Med,* 24(13), 2069-2087.

Wu FX. **2008**.Genetic Weighted K-Means Algorithm for Clustering Large-Scale Gene Expression Data. *BMC Bioinformatics,* Suppl 6, S12.

Yang YH, Buckley MJ, Dudoit S and Speed TP. **2002**.Comparison of Methods for Image Analysis on CDNA Microarray Data. *Journal of Computational and Graphical Statistics,* 11(1), 108-136.

Zacharia E and Maroulis D. **2008**.An Original Genetic Approach to the Fully Automatic Gridding of Microarray Images. *IEEE Trans Med Imaging,* 27(6), 805-813.

Zheng CH, Huang DS, Kong XZ and Zhao XM. **2008**.Gene Expression Data Classification Using Consensus Independent Component Analysis. *Genomics Proteomics Bioinformatics,* 6(2), 74-82.

