MICROARRAY DATA ANALYSIS: CURRENT PRACTICES AND FUTURE DIRECTIONS

Faaiza Vaince[1], Jerry Bona[2], Hassan M Fathallah-Shaykh[1,2]

[1]Department of Neurological Sciences, Section of Neuro-Oncology, Rush University Medical Center, 1725 West Harrison Street, Chicago, IL, 60612, USA, [2]Department of Mathematics, Statistics, & Computer Science, The University of Illinois at Chicago, 851 S. Morgan Street, Chicago, IL 60607

Address correspondence to:

Hassan M Fathallah-Shaykh
Rush University Medical Center
1735 W. Harrison St, 3rd floor - Cohn Building
Chicago, IL 60612
USA
Tel: (312) 563-3563
Fax: (312) 563-3562
E-mail: hfathall@rush.edu

Abstract

Microarrays have become one of the leading technologies used for gene expression analysis and functional genomics in many biological fields. Potential applications of microarrays can facilitate advances in molecular biology, systems biology, functional genomics, clinical medicine, and pharmacogenomics. However, microarray data can also lead to inaccurate and irreproducible conclusions. Here, we present a critical review of current computational tools used for normalization, statistical analysis, cluster analysis, and mathematical modeling-based analysis. Despite the pitfalls and challenges that still encompass the computational analysis of microarray data, the use of this technology remains very promising. In our opinion, achieving the full potential of microarray technology requires additional theoretical advances.

Introduction

In the past few years, DNA microarrays have progressed to become one of the leading technologies used for gene expression analysis and functional genomics . Because of the potential to elucidate the behavior of tens of thousands of gene transcripts in various cellular and tissue states, microarrays have generated intense interest in many biological fields . Applications in clinical medicine include the study of gene function, regulation, and interaction particularly in comparing the molecular properties of normal to pathological states. Potential applications include the ability to predict clinical behavior, prognosis, and drug response from tissue samples .

At the same time, it must be acknowledged that microarrays can also be misleading in producing inaccurate and irreproducible conclusions. The challenge lies with improving the accuracy and efficiency of the analytical tools used to identify those candidate genes responsible for differentiating phenotypes. This task is hampered because current practices are complicated by various inconsistencies and discrepancies . The purpose of this review is to give an overview of the process that goes into DNA microarray analysis. In so doing, both the advancement and the pitfalls of current methods will be highlighted. The methods that will be discussed and compared are those that are involved in normalization, statistical analysis, mathematical modeling-based analysis, and cluster analysis. The hope is that from this review, one will not be dismayed at the challenges to microarray data analysis, but instead have a better understanding of the progress that has been made as well as the areas where improvement is still necessary to bring this highly promising technology to its full power.

Platform selection

Current platforms include spotting cDNA products of specific genes, in situ synthesized oligonucleotides, and the Affymetrix Gene Chip® . The concepts discussed below can similarly be applied to all array types when formatted appropriately.

Hybridization

Samples to be analyzed usually consist of cDNA prepared from the mRNA population of a particular tissue sample during a particular state, i.e. a healthy or diseased state. This cDNA is typically prepared using one of two fluorescent dyes Cy3 and Cy5 (both are utilized in dye swapping procedures) so that following hybridization, independent images are generated for the control and query samples from which a quantification and comparison can be generated based on the relative fluorescence intensities. Some protocols actually amplify an initial mRNA sample population via in vitro transcription, which produces an aRNA population that is then hybridized onto a plate for analysis. Commercially and freely available software packages can be used to provide this image processing. The assumption is that the binding intensities indicated are proportional to the relative quantity of mRNA, or gene expression, present for the particular genes being assayed.

Normalization

Sources of variations and heterogeneity are numerous in microarray experiments because of the multiplicity of the biological steps, reagents, and equipment used. These include: 1) variations in the quantity of mRNA present in the samples, 2) the efficiency of hybridization and washing, 3) variations in the labeling, detection, and measurement efficiencies of the fluorescent signals, and 4) any other undesired technical or systemic variances, such as in scanner or image processing, specific to the experiment . These experimental variations introduce noise that is heterogeneous between experiments. It is important to realize that normalization yields only at best a *partial correction* of the noise.

Typically, normalization algorithms will rescale the data to generate expression ratios that are then transformed and reported as log ratios, which allow for straightforward comparisons of down-regulation vs. up-regulation of genes. There are several different normalization methods that can be applied to DNA microarray data and they can also be applied across various platforms. In other words, normalization can be applied to data within a single or dual channel (two-color) slide, to data in paired-slides (as in dye-swapping experiments), and to data across multiple slides .

Control genes

A set of reference genes is usually selected for the process of normalization. There are a number of ways in which this is done. They include: (i) all genes in the array, (ii) housekeeping genes that are constantly expressed across cell lines, (iii) a control sector of genes, (iv) rank invariant genes . There are advantages and disadvantages to using each of these as a set of reference genes.

Using all the available genes as control genes gives the most versatility and stability for correcting both spatial bias/ (layout bias in spotted arrays) and intensity based biases in fluorescence detection. But, this is not a very good approach when analyzing samples containing high differential expression as with custom arrays or arrays where the general expression profile is very different from that of the query samples.

The use of predetermined or housekeeping genes (genes that, by experience of hypothesis, are constantly expressed regardless of cell state) as a control can be useful in situations featuring highly differentially expressed query genes . On the other hand, recent evidence counsels action when using housekeeping genes, as many of them are not actually so very static in their expression levels, especially in tumors .

A well-selected control set of genes may be effective in adjusting for intensity-based biases. Additionally, an advantage of using an MSP (microarray sample pool) titration series is that prior biological assumption is not required in application . On the other hand, it is clear that a control set is not as effective as using all the genes for providing accurate spatial normalization and for low intensity values. If the control set is large, their use in, for example, spike-in or dilution series experiments may again present cost problems .

The use of rank-invariant set of genes allows for intensity-dependent normalization and is a more conserved set , but may not span the entire intensity range and may fail in cases where the majority of genes are either up or down-regulated .

Finally, the method of nonlinear rank-dependent transformation does not require a control set. This method transforms the ranked expression ratios of one scanned image (Cy3) to model the other (Cy5) based on the slopes of the expression curves at specific ranks.

Evidently, the advantages and disadvantages of each approach must be carefully considered when selecting an option for a particular experiment. Algorithms for normalization include those that are based on total intensity normalization, regression analysis, and ratio statistics . These categories are discussed in more depth below.

Total Intensity/Ratios Normalization

The total intensity/ratios techniques rely on an assumption to formulate a factor that is utilized to re-scale the intensity ratio of every gene in the experiment. The assumption is that the amount of up-regulation of genes will be balanced by the amount of down-regulation of genes. The total intensity technique assumes that the sum of all intensities should be equivalent for both dyes when using two channels per gene. The total ratio technique assumes that the product of all expression ratios is equal to one . As such, all the data is rescaled to meet the presumed standard.

Normalization via Regression Analysis

Normalization techniques using regression models can be broken up into several symmetrically distributed categories. They all start with the assumption that the levels of up and down-regulation are symmetrically distributed, and that the set of genes under study will display an average expression ratio equal to one (or log ratio equal to zero). One technique is commonly referred to as global normalization. This approach utilizes the global median of log intensity ratios to fit a normalization curve, and the intensities are adjusted accordingly so that the average expression ratio is equal to one. This model would be inappropriate in arrays that are not expected to display a symmetrical distribution of up and down-regulated genes. Also, with this approach, the log ratios for expression levels are not dependent on intensity biases.

However, there are intensity dependent regression models and these can be categorized as either linear or nonlinear. The linear model uses the least square method to fit a curve. For example, a graph plotting the log ratios of the Cy5 dye intensities vs. the Cy3 dye intensities would be scattered along a straight line that is fitted for best slope using regression techniques. Again, the slope is expected to be one if the fluorescent dyes for the samples were equally labeled and detected. As such, all the data is rescaled for normalization so that the slope becomes one for the normalized value . In the nonlinear models, intensity measurements across all the genes in the array are not assumed to be uniformly organized along a straight line. Instead, the microarray data may be analyzed by local regression techniques, the most common of which is the LOWESS (locally weighted scatter plot smoothing) model . Most of the current regression algorithms are set on default parameter values, which may not always be corrected and optimized by the researcher, thereby compromising the effective reduction of systemic variation. However more of an effort is certainly being made now to unravel approaches that optimize these parameters . Recently, additional nonlinear local regression models have also been proposed. These models include those involving global curve-fitting or partial fitting using splines , signal distribution analysis , generalized cross-validation , and wavelet regression .

Finally, in addition to global and intensity-dependent regression models, spatiality-dependent models have recently been introduced and are steadily gaining popularity. This model takes into factor spatial variations in intensity across a microarray slide and can be performed in a smooth robust manner that relies on a medial spatial filter . The efficacy of this model is dependent on randomized placement of the highly differentially expressed spots , which is a challenge considering this requires advance knowledge that may not be available. Also, many recent models have proposed the incorporation of both spatiality and intensity-dependent normalization, which appear to give better results than either alone .

Normalization via Ratio-based Statistics

Essentially, these types of normalization algorithms quantify and determine the significance of the fluorescent signal ratios. Chen et al have refined their previous ratio-based method by considering the contribution of background intensity to the measured signal intensities . A new signal-to noise ratio is constructed. For high signal-to noise ratios, a constant coefficient of variation for the two channel intensities is assumed . An approximate probability density for the expression ratio is calculated, and again the mean expression ratio is normalized to one, and confidence limits are established to determine those genes that are differentially expressed. For low signal-to-noise ratios, Chen et al. have proposed the refined model that corrects for the greater relative contribution of background noise . A more generalized model of the original model proposed by Chen et al has also been suggested. With this model, the coefficient of variation is not dependent on the mean intensity . Instead the intensities are calibrated and incorporated into a statistical model whose parameters are estimated for maximum likelihood, and are based on replicate data.

Other recent examples include an intensity-dependent Bayesian normalization method which also corrects for errors in total intensity measurements , a robust two-way semi-linear model , a single-channel normalization method which utilizes principal component analysis , and methods that incorporate adjustments for systemic variations caused by background intensities and gene-specific dye biases . As can be seen there are many options to consider when choosing a normalization scheme for a given dataset, and the selection should be made cautiously with the experimental objective in mind. Variations and combinations of the aforementioned methods are often utilized to obtain more accurate and qualitative results. For example, the composite method proposed by Yang et al utilizes all the genes in the array as well as a novel set of controls (the MSP titration series) for normalization, and its local regression analysis accounts for both spatiality and intensity dependent biases .

While normalization is necessary to standardize the data, normalization alone is *not* effective in identifying those genes that are differentially expressed because it does not filter the noise. Furthermore, normalization does not suffice as a data mining technique from which specific gene-to-gene relationships and patterns of gene expression can be deduced. Specifically, normalization strategies are limited because: 1) noise is preponderant, or the overwhelming majority of measurements obtained by microarrays are false, and 2) the geometrical distribution of noise is neither symmetrical nor linear . As such, further data analysis is needed, the techniques of which are discussed below.

Data Analysis

The goals of data analysis are usually to discover: 1) those genes that are differentially expressed between samples/reference, and 2) "patterns of gene expression." Below, a variety of different data analysis practices are reviewed, including statistical analysis, mathematical-modeling based analysis, and cluster analysis.

Statistical Analysis

Statistical analysis of gene expression levels may be conducted in order to identify and define those genes that are differentially expressed. This process is commonly referred to as inference, which entails the ranking of genes in order to measure the degree of differentiation and assessing the statistical significance of the results . Inference can be performed by either permutation-based or model-based statistical methods.

Permutation Based Methods:

Under permutation-based methods, the parameters for differential expression are defined by a test statistic, the significance of which is assessed by comparison of the observed value to the null distribution (which is computed by simultaneous permutation of all the sample labels for all genes being analyzed).

A commonly used model includes the two-sample t-test where two conditions or samples are compared and a p-value is obtained. Different versions exist of the t-test depending on the size of both samples and the amount of variance. In the t-test model proposed by Dudoit et al. , the p-value is calculated by permutation. The test performs well when the distribution of gene expression is uni-modal and symmetrical, but is not so efficient with more realistic distributions that exhibit multiple modes . In a similar parametric approach, a regression model is utilized in robust statistical procedure that seeks to compare the expression profiles of individual genes between two sample groups . It does so by first correcting for heterogeneity using all expression values, and then finding Z scores for each gene as a ratio of mean difference between the two groups over the standard error for the corresponding gene. However, in order to translate the Z scores to significant p-values, an asymptotic normal distribution must be assumed. As such, this test is inefficient when smaller sample sizes are used.

Other methods include the Wilcoxan rank sum method, the empirical Bayes method, and the significance analysis of microarray method. The Wilcoxan rank sum method is a robust method that compares the rank sum of two groups of samples, but in so doing it may lose information about individual differences within the same group . Its utility is also limited to larger sample sizes where the chances are higher for determining levels of gene expression that are actually statistically significant. The empirical Bayes method as proposed by Newton et al estimates levels of gene expression changes within a simple hierarchical model . From this model, significant changes are identified by a posterior probability distribution for the actual differential expression and an empirical Bayes estimate of this expression level. The significance analysis of microarrays method assigns a score to each gene in attempt to find those scores that are statistically significant. The scores are based on changes in gene expression relative to the standard deviation of repeated measurements for the gene . Using permutations, an estimate for the false discovery rate of such measurements is calculated. Tusher et al. report an FDR of 12 % with this type of analysis. The problem with the above two nonparametric approaches is that in order to construct a null distribution in the measurements, an assumption is made based on a symmetrical distribution of random errors of measured gene expression . This can lead to large false discovery rates. Zhao and Pan propose a modified mixture model approach that attempts to overcome these problems by constructing alternative null and test statistics . Recently, a new statistical nonparametric method for identifying differential gene expression based on relative entropy has also been proposed . This method combines relative entropy with kernel density estimation to detect differentiation in gene expression. It also claims to be flexible in its application to distributions that are uni-modal or multi-modal, and also claims to yield novel results as compared with the current methods already in practice. The practicality of this method may be limited by its sensitivity to changing the controlling parameters, and its future utility has yet to be determined.

Model-Based Methods:

Model-based methods utilize a statistical model to calculate the parameters that define differential gene expression. These models display the mean expression and the standard error, and make assumptions about the statistical distribution of the noise or error in the data . The advantage of these models is that they take into consideration technical errors for each gene, and account for variation across technical replicates . The subsequent inferences that are made about individual and group-wise gene expression vary according to different methods. One example is to perform this inference by analysis of variance . This method constructs error bars via ANOVA methods, and essentially normalizes the data to account for variations due to any potential confounding factors. These model-based methods can be viewed as yet another normalization method, but we discuss them here since they also determine the statistical significance of any differential gene expression in the dataset.

Statistical Correction:

Upon performing statistical inference, statistical correction is often necessary for gene-to-gene comparisons that are subject to multiple rounds of statistical testing. If correction was not applied, there would be many expressed genes whose differential expression would mistaken be found to be statistically significant. One common approach that was used before was the Bonferroni correction. This correction requires that the p-value be multiplied by the number of tests conducted, so that the correction can control the family-wise error (FEW) . The family wise error represents the probability of finding a false-positive result for differential expression for individual genes. In another type of approach that used more often today, the Benjamini and Hochberg correction , the false discovery rate (FDR) is controlled. This correction procedure multiplies the univariate p-value by the number of genes, which is then divided by the rank of the p-value . Other alternative models for statistical correction are also available, including those proposed in the context of significance analysis of microarrays and the Bayesian hierarchical model . Unless smaller sample sizes are utilized, most of the results from statistical correction will yield drastically few significant findings for differential gene expression.

To summarize, a recurring problem with statistical analysis is that each method relies on it own set of assumptions. In general statistical models/tests perform well these assumptions are met, like when the distribution of gene expression is uni-modal and symmetrical. However, they are not so efficient with more realistic distributions . The efficacy of many of the statistical models is also limited to larger sample sizes wherein statistical significance can be more reliably determined, however the larger the sample size, the more statistical correction that is necessary for multiple testing. Also, comparisons between different methods yield inconsistent results for which method is superior , so one must be careful in choosing which approach would work best for their dataset. In fact, each microarray dataset has its unique noise distribution, which: 1) is heterogeneous between datasets, 2) is nonlinear and is not uni-modal, and 3) is rank-dependent .

Mathematical Modeling-Based Analysis

An alternative to statistical analysis is a newer method being applied to DNA microarray analysis. This mathematical modeling-based analysis is not based on probability theory. As mentioned above, since statistical significance is based on normal, symmetrical, and uni-modal distributions, there may be significant limitations when the datasets include hundreds of thousands of measurements and when a large percentage of the measurements are false. The significance of the "p-value" should be treated with caution in situations where sample size is very large, especially when artifacts constitute a large percentage of the data. For example, because recalling defective cars is too costly, a car manufacturer cannot afford a p-value of 0.5 or 0.05. Similarly, because noise constitutes a predominant majority of microarray data and because thorough validation is too costly to be considered feasible, a p-value of 0.05 is untenable, especially in the genome-scale profiling of tens of thousands of genes. The goals of mathematical modeling-based methods are to shed light on the nature and behavior of noise and to create systems that discover highly specific states of differential gene expression by effectively filtering technical noise . Specificity and sensitivity are computed from true negative datasets compare the same pool of brain RNA to itself (same-to-same), and spike-in experiments, respectively.

MASH is an algorithm that yields highly specific discovery of states of genetic expression . MASH includes two filters, F1 and F2, which are based on the rank-dependent slopes and consistency of replicate measurements, respectively. MASH specificity is a thousand-fold better, but its sensitivity is equal to other methods . Background-subtracted spot intensities are sorted in ascending order to assign a *Rank* to every spot. The curves of the log-transformed rank-sorted measurements are fitted to a non-linear mathematical equation that couples each curve to its unique set of 19 variables that reflect the slopes at specific ranks. The curve fitting is based on a systematic schema. A careful look at the function that is the outcome of the fitting reveals that it is in fact composed of a single type of element, namely the functions

$$^{\circ}(x) = 1/(1 + (^{\circ}x)^{a}$$

where ° > 0 and *a* are constants. The fully fitted curve is in fact just a polynomial in these type of elements and powers of the independent variable x. A careful reader may notice an absolute value in the formula presented, but it turns out that the parameter a_{15} is never such that this absolute value has no force. These elements are in common use in approximation theory in case they have integer powers (they are sometimes called Padé approximants) and the present approximation is just an extension of this common practice.

What is more interesting is that a dataset so large and complex can in fact be modeled so successfully by such a small number of parameters! The method used thus far does have an *ad homynum* aspect to it, but it suffices to establish the principle that these datasets do not have as many degrees of freedom as one might suppose at the outset. In this aspect, it is a bit like the currently popular methods for modeling turbulence, where carefully chosen basis elements can lead to a very good description of the flow using a relatively small number of parameters (see, for example Holmes *et al.*.

A recent study of the geometrical distribution of datasets in 3-D space reveals that:

1) noise constitutes the predominant majority of microarray measurements, 2) the geometry/distribution of the data in 3-D space is ran-dependent and unique to each datasets, and 3) the distribution of noise replicates the distribution of all data. Using these ideas, a new algorithm (Fs) improves the sensitivity of MASH without lowering its high specificity . In particular, Fs uses a filter to isolate some of the noise in the dataset to construct rank and noise-specific upper and lower bound surfaces that essentially eliminates technical noise. The specificity, sensitivity, and accuracy of Fs are 99.999%, 92%, and 100%, respectively .

Cluster Analysis

Following data analysis methods that may filter the noise within individual experiment, clustering algorithms can be used for multivariate analysis of multiple DNA microarray experiments as a means for classification and pattern discovery of gene expression. The goal of cluster analysis is to identify and group those genes that are similarly expressed across different experimental conditions. These algorithms organize genes according to their expression vectors, or their representative location in "expression space" . For example when a gene is studied across different samples, its expression profile is construed as an expression vector. These expression vectors can then be organized in rows in expression matrices alongside other genes (organized in subsequent rows).

Before cluster analysis can be applied, certain parameters must be resolved. Of particular importance is defining the distance metric that will be utilized in the clustering algorithm. This calculated measurement is necessary in order stipulate the distance required between any 2 expression vectors when placed in a cluster together. There are a variety of metric and semi-metric distance measures that are utilized . One of the most common is the Euclidean distance, which is a simple calculation of the distance between two coordinates on a graph. This metric works well when the variables are first standardized, but is not as effective on raw data for which the scales of graphical display can vary. Another popular distance metric utilized is the Pearson correlation, which is more effective for experiments conducted along a time-course.

Like normalization methods, cluster analysis can be applied by various methods that can be broken down into different categories . For one, clustering can be conducted by divisive or agglomerative methods. Divisive methods start with all the genes in one cluster. These genes are then subsequently split off into additional groups. Agglomerative techniques work in a reverse fashion by starting with single gene groups that are subsequently clustered into groups that contain more and more genes with each round of clustering. Cluster analysis can also be categorized as unsupervised and supervised. Unsupervised analysis consists of clustering groups into classes of unknown function. Supervised analysis takes previously unknown samples and clusters them into known classes. A known class is one for which biologically relevant information is already available. Examples of all the above categories of cluster analysis will be reviewed below.

Hierarchical Clustering:

One of the most widely used clustering techniques is an agglomerative model known as hierarchical clustering, which is relatively simple to apply and see visually distinct results . Under this model, single gene expression profiles are fused together on the basis of similarity (proximity in expression space) to form larger groups. These groups are consequently fused into a smaller number of groups until the process has been exhausted. The results are represented in a single hierarchical tree that can be diagramed with a dendrogram. The algorithms applied to group the genes can vary according to the distance metric that is chosen, and as such the results will vary as well. For example, in single-linkage clustering, the minimum distance is calculated between a pair of gene expression vectors. This method can be problematic in producing "chaining" results that fail to resolve relatively distinct clusters that are intercepted by intermediate "noise" points . Other examples of hierarchical algorithms include centroid clustering, Ward's method, complete-linkage clustering, and average-linking clustering. The problem with several of these hierarchical methods is that they are biased towards finding tightened 'spherical clusters' even when other shapes may be more appropriate . As a result, the clusters do not end up representing the expression of genes assigned to the clusters, and if one wrong assignment is made early on, it cannot be corrected .

K-means Clustering:

Another common clustering technique is one known as k-means clustering. In this method, genes are clustered into a predetermined number of clusters (set by the user). A series of computations then regroups the genes so that intra-cluster distances are minimized and inter-cluster dissimilarity is maximized. As with hierarchical clustering, various algorithms can apply k- means clustering . K-means clustering works best when previous knowledge of the system is utilized in order to help specify clusters as well as seed cases, or genes, within the clusters. However, unless effective 'recovery' computations are performed, these methods are typically not as effective as hierarchical methods .

Neural Network Clustering:

Neural network algorithms are modeled after the learning processes of cognitive science and the neurological functions of the brain. These networks are used to build a data 'training set' that provides the framework for predictions and classification of microarray data sets . The most common neural-network-based paradigm utilized for gene expression clustering is a divisive approach known as self-organizing maps . These maps compare gene expression vectors to reference vectors in order to assign the genes to clusters via an iterative process. These clusters are fitted into a predetermined (set by the user to establish the number of resultant clusters) geometric configuration such as a two dimensional rectangular or hexagonal grid. Like k-means clustering, this method is more effective when it is conducted in a semi-supervised fashion where previous knowledge of the system is utilized in constructing the initial seeds. However, unless the dimensionality is simplified in conjunction with another method such as principal component analysis as described below, its complex and multidimensional data can be hard to visualize.

Principal Component Analysis:

Principal component analysis is essentially a weighted type of clustering useful in simplifying multi-dimensional data by reducing redundant data . It does so by summarizing all the values of a gene expression vector into one number. With each computational step of analysis, it maximizes variability between samples so that the best separation of data can be easily visualized. As in other clustering methods, principal component analysis cannot positively define the genes designated to each cluster. Variations of this analysis have recently been proposed, such as total principal component regression, which takes into account the errors in both independent and dependent variables . These methods and their family of related techniques, such as factor analysis and principal coordinate analysis, are most effective when combined with other clustering methods such as the ones described above.

Supervised Cluster Analysis:

These methods differentiate genes into already specified categories and are often utilized in tumor diagnosis and drug discovery . One example of such a method is the support vector machine. An SVM is essentially a reference set containing those genes that are known to be positively co-expressed and those genes that are negative for this expression . Using this reference or 'training' set which is based on available biological information, the SVM sets out to distinguish other genes in the data set as either members or non-members of these predefined classes . The computations required in this method include kernel function measurements, which are often improperly applied to yield incorrect classifications. An effort must be made to choose the best kernel functions and parameters to minimize error. Also, it is important that classification results be cross-validated (i.e. by a Monte Carlo cross validation estimate) to predict the probability of misclassification .

In addition to the above models, other clustering algorithms exist and continue to be proposed, one such model is bootstrapping cluster analysis, which uses confidence intervals to estimate the differential expression of individual genes and assess the stability of cluster analysis results .

The major problems of clustering methods are they are biased to find clusters especially when noise is not only preponderant but also heterogeneous between datasets. Microarray experiments often cluster by technical variables like date, technician, batch, etc. In addition, it is very important to realize that in the presence of a preponderance of noise, most of the clusters are likely to be false especially when the clustering methods are supervised. If cluster techniques are used incorrectly, misleading interpretations will distort the original relationships amongst the genes in the dataset. Furthermore, because the large sets of data in microarrays are subject to heterogeneous dataset-dependent technical noise, and noise may not be reliably eradicated, one cannot find much accuracy in the groups of genes that are established by cluster analysis. In the presence of technical noise, relations discovered by clustering techniques may be erroneous and may explain the inability to replicate or generalize the results of a single or a few datasets. Another important source of instability stems from the assumption of accuracy for the fold changes between experiments. Clustering techniques compare samples based on the levels of genetic expression or fold changes of expression ratios. Unfortunately, the fold change values for gene expression across different samples do not always yield an accurate comparison. Thus, in the presence of unfiltered noise and in the absence of an accurate system for comparing fold changes, relations inferred from clustering methods should be interpreted cautiously. The merit of clustering results, especially in classifying diseases, can be promising but is highly dependent on the quality of the input data, i.e. how noise-free they are. Clustering is very useful when combined with filtering methods, as detailed below .

Applications

Systems Biology/Biological Chemistry

Single genes or molecules of the cell belong to rich networks of molecular interactions that include transcriptional regulation, signaling pathways, protein-protein, and protein-nucleic acid interactions. Recent evidence support the intuitive idea that based on the knowledge of the identity of differentially expressed genes and their states of genetic expression, one could draw inferences about gene-to-gene and gene-to-protein interactions and protein states that may eventually uncover complete molecular systems behind complex phenotypes. Examples include the discovery of balanced opposing molecular functions behind the phenotypes of meningiomas, and complex molecular systems that create the phenotypes of motility as well as resistance to endoplasmic reticulum and oxidative stress in cultured gliomas . This idea is illustrated in Figure 1. Highly specific discovery of transcriptionally regulated genes uncovers only a part of the puzzle of molecular systems; nonetheless, the other components that are not dependent on the up- or down regulation of genes may be inferred and tested experimentally to obtain a complete picture of the molecular system (the Mona Lisa).

Clinical Applications and Pharmacogenomics

The clinical applications of DNA microarrays are numerous. Potentially, they have utility in early and efficient diagnosis of disease, particularly cancer , and in diagnosis by detecting genes that can serve as markers . Also, microarrays can be used to detect genes that make individuals susceptible to certain diseases or discover molecules that make cells resistant to certain treatments or protective mechanisms . In so doing, prevention and treatment programs can be individually tailored for optimal effect on the healthy and patient population. Also the discovery of differential gene expression for predefined disease phenotype can serve as prognostic markers for individual patients . The clinical relevance of microarrays spreads across many medical disciplines such as oncology, infectious diseases, lung disease, cardiology, and pharmacogenomics .

Conclusions

Despite the pitfalls and challenges that still encompass the computational analysis of microarray data, the use of this technology remains promising amidst the advances that have been made in refining the analytic techniques. Achieving the full potential of microarray technology requires additional advances in mathematical theory and applications, and a more thorough understanding of the underlying biological systems and mechanisms involved in cellular processes.

REFERENCES

Aitman, T. J. (2001) DNA microarrays in medical practice. *BMJ* **323**, 611-615.

Alizadeh, A. A.; Eisen, M. B.; Davis, E.; Ma, C.; Lossos, I.; Rosenwald, A.; Boldrick, J.; Sabet H.; Tran T.; Yu X.; Powell JI.; Yang L.; Marti, G.; Moore, T.; Hudson, J., Jr.; Lu, L.; Lewis, D.; Tibshirani, R.; Sherlock, G.; Chan, WC.; Greiner, TC.; Weisenburger, DD.; Armittage, JO.; Warnke, R.; Levy, R.; Wilson, W.; Grever, MR.; Byrd, JC.; Botstein, D.; Brown, P. O. and Staudt, L. (2000) Distinct types of diffuse late B-cell lymphomas identified by gene expression profiling. *Nature* **403**, 503-511.

Arabie, P. and Hubert, L. J. (1996) Clustering and Classification. World Scientific Publishing Co. Pte. Ltd., Singapore.

Armstrong, N. J. and van de Wiel, M. A. (2004) Microarray data analysis: from hypotheses to conclusions using gene expression data. *Cell Oncol* **26**, 279-290.

Baird, D.; Johnstone, P. and Wilson, T. (2004) Normalization of microarray data using a spatial mixed model analysis which includes splines. *Bioinformatics* **20**, 3196-3205.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc B* **7**, 289-300.

Berger, J. A.; Hautaniemi, S.; Jarvinen, A. K.; Edgren, H.; Mitra, S. K. and Astola, J. (2004) Optimized LOWESS normalization parameter selection for DNA microarray data. *BMC Bioinformatics* **5**, 194.

Bloom, G.; Yang, I. V.; Boulware, D.; Kwong, K. Y,; Coppola, D.; Eschrich, S.; Quackenbush, J. and Yeatman, T. J. (2004) Multi-platform, multi-site, microarray -based human tumor classification. *Am J Pathol* **164**, 9-16.

Brown, M. P.; Grundy, W. N.; Lin, D.; Cristianini, N.; Sugnet, C. W.; Furey, T. S.; Ares, M. Jr, and Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A* **97**, 262-267.

Chen, Y.; Dougherty, E. R, and Bittner, M. L, (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J Biomed Optics* **2**, 364-374.

Chen, Y.; Kamat, V.; Dougherty, E. R.; Bittner, M. L.; Meltzer, P. S. and Trent, J. M. (2005) Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics* **18**, 1207-1215.

Chen, Y. J.; Kodell, R.; Sistare, F.; Thompson, K. L.; Morris, S. and Chen, J. J, (2005) Normalization methods for analysis of microarray gene-expression data. *J Biopharm Stat* **13**, 57-74.

Dudoit, S.; Yang, Y. H.; Callow, M. J. and Speed, T. P. (2000) Statistical methods for identifying differentially expressed genes in replicate microarray experiments. *Technical Report* **234**, 395-402.

Duggan, D. J.; Bittner, M.; Chen, Y.; Meltzer, P. and Trent, J. M. (1999) Expression profiling using cDNA microarrays. *Nat Genet.* **21**, 10-14.

Eisen, M. B.; Spellman, P. T.; Brown, P. O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **8,** 14863-14868

Everitt, B. S (1993) Cluster Analysis. Oxford University Press, New York.

Fathallah-Shaykh, H. M. (2005) Noise and rank-dependent geometrical filter improves sensitivity of highly specific discovery by microarrays. *Bioinformatics* **21**, 4255-4262.

Fathallah-Shaykh, H. M.; He, B.; Zhao, L.-J.; Engelhard, H.; Cerullo, L.; Lichtor, T.; Byrne, R.; Munoz, L.; Von Roenn, K.; Rosseau, G.; Glick, R.; Chen, S. and Khan, F. (2003) Genomic expression discovery predicts pathways and opposing functions behind phenotypes. *J Biol Chem* **278**, 23830-23833.

Fathallah-Shaykh, H. M. (2005a) Genomic Discovery reveals a molecular system for resistance to ER and oxidative stress in cultured glioma. *Arch Neurol* **62**, 233-236.

Fathallah-Shaykh, H. M. (2005b) Microarrays: applications and pitfalls. *Arch Neurol* **62**, 1669-1672.

Fathallah-Shaykh, H. M.; He, B.; Zhao, L.-J. and Badruddin, A. (2004) Mathematical algorithm for discovering states of expression from direct genetic comparison by microarrays. *Nucleic Acids Res* **32**, 3807-3814.

Fathallah-Shaykh, H. M.; Rigen, M.; Zhao, L.-J.; Bansal, K.; He, B.; Engelhard, H.; Cerullo, L.; Von Roenn, K.; Byrne, R.; Munoz, L.; Rosseau, G.; Glick, R.; Lichtor,

T. and DiSavino, E. (2002) Mathematical modeling of noise and discovery of genetic expression classes in gliomas *Oncogene* **21**, 7164-7174.

Fathallah-Shaykh, H. M. (2005c) Logical networks inferred from highly specific discovery of transcriptionally regulated genes predict protein states in cultured gliomas. *Biochem Biophys Res Comm* **336**, 1278-1284.

Furey, T. S.; Cristianini, N.; Duffy, N.; Bednarski, D. W.; Schummer, M. and Haussler, D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**, 906-914.

Futschik, M. and Crompton, T. (2004) Model selection and efficiency testing for normalization of cDNA microarray data. *Genome Biol* **5**, R60.

Grant, G. M.; Fortney, A.; Gorreta, F.; Estep, M.; Del Giacco, L.; Van Meter, A.; Christensen, A.; Appalla, L.; Naouar, C.; Jamison, C.; Al-Timimi, A.; Donovan, J.; Cooper, J.; Garrett, C. and Chandhoke, V. (2004) Microarrays in cancer research. *Anticancer Res* **24**, 441-448.

Holmes, P.; Lumley, J. and Berkooz, G. (1996) Turbulence structures, dynamical systems and symmetry (Cambridge Monographs on Mechanics).

Huber, W.; von Heydebreck, A.; Sultmann, H.; Poustka, A. and Vingro,n M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, S96-S104.

Ittrich, C. (2005) Normalization for two-channel microarray data. *Methods Inf Med* **44**, 418-422.

Jin, J. Y.; Almon, R. R.; DuBois, D. C. and Usko W. J. (2003) Modeling of corticosteroid pharmacogenomics in rat liver using gene microarrays. *J Pharmacol Exp Ther* **307**, 93-109.

Kerr, M. K. and Churchill, G. A. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc Natl Acad Sci U S A* **98**, 8961-8965.

Kerr, M. K.; Martin, M. and Churchill, G. A. (2000) Analysis of variance for gene expression microarray data. *J Comput Biol* **7**, 819-837.

Khan, J.; Wei, J. S.; Ringner, M.; Saal, L. H.; Ladanyi, M.; Westermann, F.; Berthold, F.; Schwab, M.; Antonescu, C. R.; Peterson, C. and Meltzer, P.S. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* **7**, 673-679.

Khimani, A. H.; Mhashilkar, A. M.; Mikulskis, A.; O'Malley, M.; Liao, J.; Golenko, E. E.; Mayer, P.; Chada, S.; Killian, J. B. and Lott, S. T. (2005) Housekeeping genes in cancer: normalization of array data. *Biotechniques* **38**, 739-745.

Kothapalli, R.; Yoder, S. J.; Mane, S. and Loughran, T. P. Jr. (2002) Microarray results: how accurate are they? *BMC Bioinformatics* **3**, 22-

Lipshutz, R. J.; Morris, D.; Chee, M.; Hubbell, E.; Kozal, M. J.; Shah, N.; Shen, N.; Yang, R. and Fodor, S. P. (1995) Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* **19**, 442-447.

Lockhart, D. J.; Dong, H.; Byrne, M. C.; Follettie, M. T.; Gallo, M. V.; Chee, M. S.;

Mittmann, M.; Wang, C.; Kobayashi, M.; Horton, H. and Brown, E. L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**, 1675-1680.

Martin-Magniette. M. L.; Aubert, J.; Cabannes, E. and Daudin, J. J. (2005) Evaluation of the gene-specific dye bias in cDNA microarray experiments. *Bioinformatics* **21**, 1995-2000.

Meloni, R.; Khalfallah, O. and Biguet, N. F. (2004) DNA microarrays and pharmacogenomics. *Pharmacol Res* **49**, 303-308.

Newton, M. A.; Kendziorski, C. M.; Richmond, C. S.; Blattner, F. R. and Tsui, K. W. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol* **6**, 37-52.

Newton, M.; Kendziorski, C.; Richmond, C.; Blattner, F. and Tsui, K. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data *J Comp Biol* **8**, 37-52.

Ntzani, E. E. and Ioannidis, J. P. (2003) Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet* **362**, 1439-1444.

Pan, W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **18**, 546-554.

Park, T.; Yi, S. G.; Kang, S. H.; Lee, S.; Lee, Y. S. and Simon, R. (2003) Evaluation of normalization methods for microarray data. *BMC Bioinformatics* **4**, 33-

Qin, L. X. and Kerr, K. F. (2004) Empirical evaluation of data transformations and ranking statistics for microarray analysis. *Nucleic Acids Res* **32**, 5471-5479.

Quackenbush. J. (2001) Computational analysis of microarray data. *Nat Rev Genet* **2**, 418-427.

Quackenbush, J. (2002) Microarray data normalization and transformation. *Nat Genet* **32**, 496-501.

Raychaudhuri, S.; Stuart, J. M. and Altman, R. B. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pacific Symposium on Biocomputing* 455-466.

Schena, M.; Shalon, D.; Davis, R. W. and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470.

Schena. M.: Shalon, D.: Heller, R.; Chai, A.; Brown, P. O. and Davis, R. W. (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci U S A* **93**, 10614-10619.

Simon, R. (2003) Using DNA microarrays for diagnostic and prognostic prediction. *Expert Rev Mol Diagn* **3**, 587-595.

Stoyanova, R.; Querec, T. D.; Brown, T. R. and Patriotis, C. (2004) Normalization of single-channel DNA array data by principal component analysis. *Bioinformatics* **20**, 1772-1784.

Tamayo, P.; Slonim, D.; Mesirov, J.; Zhu, Q.; Kitareewan, S.; Dmitrovsky, E.; Lander, E. S. and Golub, T. R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* **96**, 2907-2912.

Tan, P. K.; Downey, T. J.; Spitznagel, E. L. Jr.; Xu, P.; Fu, D.; Dimitrov, D. S.; Lempicki, R. A.; Raaka, B. M. and Cam, M. C. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* **31**, 5676-5684.

Tan, Y.; Shi, L.; Ton,g W. and Wang, C. (2005) Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data. *Nucleic Acids Res* **33**, 56-65.

Tarca,. A. L.; Cooke, J. E. and Mackay, J. (2005) A robust neural networks approach for spatial and intensity-dependent normalization of cDNA microarray data. *Bioinformatics* **21**, 2674-2683.

Thomas, J. G.; Olson, J. M.; Tapscott, S. J. and Zhao, L. P. (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res* **11**, 1227-1236.

Tseng, G. C.; Oh, M. K.; Rohlin, L.; Liao, J. C. and Wong, W. H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res* **29**, 2549-2557.

Tusher, V.; Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionozing radiation response *Proc Natl Acad Sci USA* **98**, 5116-5121.

Villeneuve. D. J. and Parissenti. A. M. (2004) The use of DNA microarravs to investigate the pharmacogenomics of drug response in living systems. *Curr Top Med Chem* **4**, 1329-1345.

Wadlow, R. and Ramaswamy, S. (2005) DNA microarrays in clinical cancer research. *Curr Mol Med* **5**, 111-120.

Wang, D.; Huang, J.; Xie, H.; Manzella, L. and Soares, M. B. (2005) A robust two-way semi-linear model for normalization of cDNA microarray data. *BMC Bioinformatics* **6**, 14-

Wang, J.; Ma, J. Z. and Li, M. D. (2004) Normalization of cDNA microarray data using wavelet regressions. *Comb Chem High Throughput Screen* **7**, 783-791.

Weinstein. J. N.: Mvers. T. G.: O'Connor. P. M.: Friend. S. H.: Fornace. A. J. Jr.: Kohn. K. W.; Fojo, T.; Bates, S. E.; Rubinstein, L. V.; Anderson, N. .L; Buolamwini, J. K.; van Osdol, W. W.; Monks, A. P.; Scudiero, D. A.; Sausville, E. A.; Zaharevitz, D. W.; Bunow, B.; Viswanadhan, V. N.; Johnson, G. S.; Wittes, R. E. and Paulll K. D. (1997) An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343-349.

Wilson, D. L.; Buckley, M. J.; Helliwell, C. A. and Wilson, I. W. (2003) New normalization methods for cDNA microarray data. *Bioinformatics* **19**, 1325-1332.

Wolfinger, R. D.; Gibson, G.; Wolfinger, E. D.; Bennett, L.; Hamadeh, H.; Bushel, P.; Afshari, C. and Paules, R. S. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol* **8**, 625-637.

Workman, C.; Jensen, L. J.; Jarmer, H.; Berka, R.; Gautier, L.; Nielser, H. B.; Saxild, H. H.: Nielsen. C.: Brunak. S. and Knudsen. S. (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol* **3**, research0048.

Yan, X.; Deng, M.; Fung, W. K. and Qian, M. (2005) Detecting differentially expressed genes by relative entropy. *J Theor Biol* **234**, 395-402.

Yang, I. V.; Chen, E.; Hasseman, J. P.; Liang, W.; Frank, B. C.; Wang, S.; Sharov, V.; Saeed, A. I.; White, J.; Li, J.; Lee, N. H.; Yeatman, T. J. and Quackenbush, J. (2002) Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol* **3**, research0062.

Yang, Y.; Dudoit, S.; Luu, P.; Lin, D.; Peng, V.; Ngai, J. and Speed, T. (2002) Normalization of cDNA microarrav data: a robust composite method addressina single and multiple slide systematic variation. *Nucleic Acids Res* **30**, e15.

Yoon, D.; Yi, S. G.; Kim, J. H. and Park, T. (2004) Two-stage normalization using background intensities in cDNA microarray data. *BMC Bioinformatics* **5**, 97.

Zhang, D.; Wells, M. T.; Smart, C. D. and Fry, W. E. (2005) Bayesian normalization and identification for differential gene expression data. *J Comput Biol* **12**, 391-406

Zhao, Y. and Pan, W. (2003) Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics* **19**, 1046-1054.

EGEND

Figure 1. A cartoon that illustrates one application of highly specific discovery by microarrays. A typical microarray data set is beset by the large amount of extraneous noise from which the real data must be mined for interpretation. In (a) noise is added to the painting of the Mona Lisa. (b) Highly specific and sensitive discovery by mathematical modeling-based data analysis filters noise and delivers highly accurate information on states of genetic expression. This information uncovers pieces of the biological picture or network. (c) By inference, current knowledge leads to re-arranging discovered states of genetic expression into a system. The remaining pieces of the picture can be deduced from hypotheses that can then be tested and validated by standard biological techniques until the complete biological picture is elucidated (d).