Statistical Analysis of DNA Microarray Data in Cancer Research

Jianqing Fan[1] and Yi Ren[2]

Abstract Microarray techniques have been widely used to monitor gene expression in many areas of biomedical research. They have been widely used for tumor diagnosis and classification, prediction of prognoses and treatment, and understanding of molecular mechanisms, biochemical pathways, and gene networks. Statistical methods are vital for these scientific endeavors. This article reviews recent developments of statistical methods for analyzing data from microarray experiments. Emphasis has been given to normalization of expression from multiple arrays, selecting significantly differentially expressed genes, tumor classifications, and gene expression pathways and networks.

Due to the advances in bioimaging technology, large-scale measurements of mRNA abundance have become widely available through microarray techniques. With advanced statistical techniques, microarray analyses enable simultaneous study of the entire genome in one experiment. Currently, four widely used array platforms are available: commercial Affymetrix GeneChip, oligonucleotide microarrays, cDNA microarrays, and customized microarrays, which have been widely applied to cancer research. They have had substantial effect on tumor diagnosis and classification, prediction of prognosis and response to therapy, and understanding of the molecular mechanisms of tumorigenesis and tumor development. Furthermore, gene expression profiling by microarray will further refine the future for individualized treatment for cancer patients based on the molecular classification of subtypes.

Proper statistical analysis is vital to the success of array use. What makes microarray data analysis differs from traditional statistics is the systematic biases inherent in the variations of experimental conditions and distinguishing features associated with the microarray outputs: high dimensionality (making simultaneous inferences on thousands of genes) and sparsity (only a small fraction of genes are statistically differentially expressed). These challenges have forged new collaborations between statisticians, computational biologists, and molecular and clinical investigators to develop more discriminatory statistical methods to address the issues arising from analysis of microarrays. These have resulted in development of many powerful and effective software packages using different and advanced statistical methods. "Bioconductor" (http://www.bioconductor.org/) is an open source and development software project for the analysis of genomic data to which many researchers contributed their updated statistical techniques.

Authors' Affiliations: [1]Statistics Lab, Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey and [2]Center for Hematology and Oncology Molecular Therapeutics, Taussig Cancer Center, Cleveland Clinic, Cleveland, Ohio

Received 4/27/06; revised 5/16/06; accepted 5/24/06.

Grant support: National Science Foundation grant DMS-0354223 and NIH grant R01-GM07261.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Jianqing Fan, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08540. Phone: 609-258-7924; Fax: 609-258-8115; E-mail: jqfan@princeton.edu.

doi:10.1158/1078-0432.CCR-06-1033

Process of Biomedical Studies Using Microarray

Experiments are first designed to answer biological questions, and then microarray experiments are conducted (Fig. 1). After that, expression profiles through accompanying computer software are extracted from scanned images. With raw data, gene expression with low quality is filtered using coefficients of variations or the intensities of scanned images (1–3) and systematic biases are removed via normalization techniques. After data being properly normalized, downstream statistical analysis is conducted, such as selecting significant genes and classifying different types of tumors. Microarray techniques are typically used as a screening tool, biological validation and interpretation should be done to further study the selected genes.

Preprocessing and Normalization for cDNA Microarrays

The quality of microarray data is paramount important for downstream statistical analysis. These include RNA quality, probe labeling, hybridization condition, washing, and signal and background detection in the scanning process. This is especially true when studying low-abundance RNA species due to the effect of background systematic biases, which include slide, block, and dye effects, and uneven hybridization and processing, among others. Small variations in these conditions can induce significant changes in gene expression, resulting in both false-negative and false-positive predictions. However, these variations and systematic biases in microarray data can be attenuated by proper control and adequate replication of the studies and statistical normalization.

After obtaining the expression profiles from both test and reference samples, we can compute the average of the log intensities from fluorescent dyes Cy3 and Cy5 channels for each gene (1). This average is often referred to as the intensity

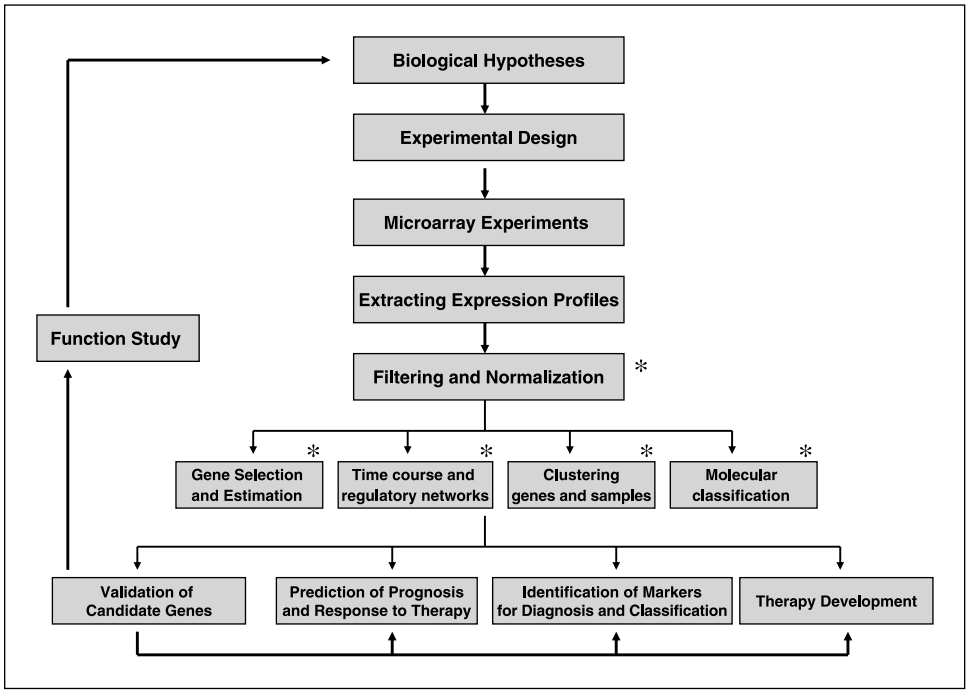

Fig. 1. Asterisks, schematic representation of microarray strategy and steps for statistical analysis (marked with *).

of a particular gene. Association with each gene is the log ratio between the expression profiles among the test and reference groups and its physical location on cDNA chips. The first step is to filter the data with low repeatability. This can be done statistically by controlling the coefficient of variation (smaller than certain threshold) and the intensity (bigger than certain level; refs. 2, 3). After preprocessing, systematic biases need to be removed to account for overall brightness of scanned images and experimental variations, such as block and dye effects. This step is essential for multiple array comparisons and downstream statistical analysis. It is collectively referred to as normalization.

There are several methods for normalization. The most primitive one is the global normalization, which makes every array have the same median intensity. The method is useful when there is no intensity effect and no block effect. However, ample statistical studies (1–6) reveal the presence of these effects. To account for these effects, Dudoit et al. (1) normalized data at each block and intensity, called "Lowess normalization" (1). The basic biological assumption is that the average of up-regulated expression profiles and down-regulated profiles is approximately the same at each intensity level. This assumption is not necessarily valid especially for customized arrays or treated cells. Tseng et al. (3) used an "invariant set of genes" as proxy of housekeeping genes and estimated the intensity effect based only on the data in the invariant set. To remove the aforementioned biological assumption, Fan et al. (2) introduced a semilinear in-slide model normalization technique, which takes advantages within-array replications. Based on ~100 replicated clones within an array, the intensity and block effects were estimated and removed from the expression profiles. The rationale is that the difference of expression profiles between replicated clones in an array reflects the systematic biases in addition to the random noise, and systematic biases of block and intensity effects can be extracted from these pairs of data. By using semilinear in-slide model, migration inhibitory factor targeting genes in neuroblastoma cells were selected and validated by real-time reverse transcription-PCR, whereas some of these genes were missed by ordinary normalization methods (2). Fan et al. (4) significantly widened the scope of applicability by creating 'synthetic' replications and aggregating information from other arrays. Other useful normalization methods include two-way semilinear model (5) and robust normalization (6).

Within-Array Replications

Within-array replications are not only powerful for normalization but also useful for validation whether data have been properly normalized. The basic idea is that the differences among within-array replications are purely random noises after the systematic biases are removed. When the noise levels are estimated for each individual gene and an array is properly normalized, the sum of the standardized square differences follows approximately a χ^2 distribution. This provides a simple and useful diagnostic test statistic to check if an array has been properly normalized. The test statistic can also be used as a criterion for selecting a normalization method for a given array—the one with smallest test statistic (most consistent replications) is the most preferable. The variances of associated differences can be estimated by using smoothing techniques (3) and empirical Bayes method (7). Details will be posted on the web as a forthcoming article. Within-array replications have also been used to improve the precision with which the genewise variances are estimated and thereby improve inference methods designed to identify differentially expressed genes (8).

Selecting Significant Genes

An important statistical question is to select differentially expressed genes between the test and reference samples or

more complex comparisons (9). The first step is to select a proper test statistic. It is usually a modification of a t test statistic, such as that in Significance Analysis of Microarrays (10), a modified one-sample (2) and two-sample t test in refs. 10, 11, F-statistic (7), or an empirical Bayes procedure (12, 13). A marked feature is the sparsity—only a small fraction of genes are differentially expressed. In choosing a test statistic, the procedures taking care of sparsity features are usually more powerful for microarray applications with increased sensitivity and specificity.

After a test statistic has been selected, the next step is to compute the P of the test statistic. Because of simultaneous inferences in the order of thousands, we need to select genes with associated Ps of order 10^{-3} or 10^{-4}. Looking up normal or t tables relies predominately on mathematical assumptions rather than on the data. It is not robust to the mathematical assumptions. Resampling techniques, such as permutation or bootstrapping, are frequently used. Balanced permutation and sieved permutation methods can be found in refs. 10, 11 for estimating Ps. Due to limited number of arrays, permutation does not provide enough resolution for computing Ps in an order of 10^{-3} or 10^{-4}. This is accomplished by the marginal aggregation method (2, 10, 14) via the assumption that the marginal distributions of test statistics are approximately the same under the null hypotheses. The assumption allows us to pull all permuted test statistics together to compute Ps, which improves the resolution of computing Ps by a factor of thousands or more.

With estimated Ps, a statistical task is then to find significantly differentially expressed genes. For simultaneous testing of hypotheses of thousands or tens of thousands, the probability of making at least one false discovery (positives) might not be relevant. Controlling the false discovery rate (15) or local false discovery rate (16) is far more relevant (15–17). There are many statistical papers on controlling different aspects of false discovery rate. Storey et al. (15) and Dudoit et al. (17) provided an overview of these statistical methods in genomic applications. These techniques require Ps to be estimated very accurately, usually in the order of 10^{-6}. Alternatively, one can choose a testing procedure, which picks significant genes when the test statistics exceed a certain critical value or their associated Ps are less than a threshold, and then estimate the false discovery rate (2, 10, 11). For example, suppose that there are 100 genes with estimated Ps < 0.001 among 15,000 genes, the expected number of false discovered genes is no larger than $0.001 \times 15{,}000 = 15$, giving an estimated false discovery rate of $15/100 = 15\%$ among 100 selected genes (2).

Tumor Classification and Clustering

An important application of microarray techniques is to find biomarkers for clinical and pathologic tumor classification. This is exemplified by the work of Inamura et al. (18). Inamura et al. analyzed tumor and normal lung samples by hierarchical clustering with the nonnegative matrix factorization approach and divided lung squamous cell carcinoma into two distinct subclasses, which showed the significant differences in clinical outcome and molecular characteristics (18).

Classification is also referred to as supervised learning. Many statistical learning techniques have been proposed for classification and clustering. These include tree and forest-based methods (19, 20), margin-based classifiers, such as boosting and supporting vector machine (20), and receiver operating characteristic regression (20). Using normalized microarray data as input vectors, classification rules can be built. Svrakic et al. (21) gave a comprehensive overview on genomic applications of these classification methods.

In tumor classification, we hope to select only tens of genes or biomarkers that have high discriminative power with low misclassification rate. This not only provides molecular and genomic understanding on how these genes are related to different classes of tumor but also reduces misclassification rates for prediction. A method of shrunken centroids of gene expression profiles has been proposed for selecting genes that are important for tumor classification (22). Statistical variable selection methods, such as penalized divergence for misclassification, can also be used for selecting important genes and other biomarkers (23).

Clustering, also called unsupervised learning, algorithms are frequently used to group genes with similar expression profiles (21). This facilitates our visualization of coexpressions of genes and also allows us to cluster arrays with similar expression patterns. An important component of clustering algorithms, such as the hierarchical and K-mean algorithms, is to define appropriate metrics in an input space (21, 22). The input vectors can be either the expression profiles across different arrays for grouping genes with similar expression patterns across different subjects or the expression profiles across different genes for clustering subjects with similar microarray data. Commonly used metrics include the Euclidean distance and Pearson correlation. With these metrics, the hierarchical and K-mean algorithms can be applied for clustering. This is often presented as dendrograms or color-coded representation of similarly expressed genes.

Time Course and Regulatory Networks

To monitor transient gene expression patterns, temporal progression of a disease, or response to a treatment, gene expression data are obtained from the same tissue or cells at different time points. An important statistical question is whether genes have been differentially expressed at certain time points after treatment. The Hotelling T^2 test can be used to check whether the expression profiles remain constant over time. This identifies genes that have expressed over the time course of experiments. At any time points, some genes are up-regulated or down-regulated and some keep unchanged. Patterns of expressions over time courses are important for understanding expression pathways. A further improvement is to account the uncertainty of measurements at each time point. One-sample t test statistic can be used to assess whether the expression of a gene is up-regulated or down-regulated or unchanged at each given time point. Then, the patterns of expression over time course can be statistically identified using a simple classification technique (21), which provides useful tools for understanding the regulatory process and biochemical pathways. Schulte et al. (24) showed differential gene expression patterns, including immediate early genes, "delayed" genes, and effector genes in TrkA- and TrkB-expressing neuroblastomas. This finding displayed the distinct regulation kinetics with regards to induction of immediate early genes and downstream molecular targets.

Microarray techniques have also been used to study the casual inferences about genetic variation, gene networks, interactions in regulatory processes, and biochemical pathways. They provide useful tools for understanding interactions, associations, and networks among genes (25).

Customized Arrays

Attractive advantages of customized microarrays are the ability to do focused array experiments on a smaller scale by limiting the number of genes being. This enables researchers to focus only on hundreds of genes of primary interest with more reliable measurements (larger DNA spots) and within-array replications. The selection biases and within-array replications in customized arrays provide an ideal platform for semilinear in-slide model normalization (3). The simplified image analysis and data analysis together with low cost give hope that this powerful technology will become a standard molecular biology tool for routine use in the clinical setting.

Analysis of Affymetrix Array Data

Affymetrix GeneChip arrays (Affymetrix, Santa Clara, CA) obtain the expression profile of a mRNA of a gene by the combined intensity information from probes in a probe set, which consists of 11 to 20 probe different 25-mer oligonucleotides, interrogating a different part of the sequence of a gene. Several techniques have been proposed for extracting expression profiles from the information at probe level. These include the detection signals prominently featured in Affymetrix GeneChip Operating Software, the model-based expression index (26), and the robust multichip average (27).

Normalization is needed to account for the overall brightness of scanned images and other experimental variations. Several methods have been proposed for normalizing data at probe level, and their effects on the analysis of gene expressions have been examined. The quantile normalization is frequently used for probe level normalization (28). However, most researchers use detection signals from MAS 5.0 as starting points for their investigation. The techniques for probe level normalization are not effective for data at detection signal level. Semilinear in-slide model has been effectively extended to normalize detection signals to account of intensity effect (11).

After data have been properly normalized, statistical techniques can be applied to select significant genes (11). The essential difference between data from cDNA and Affymetrix GeneChip is that the test and reference expression profiles are paired in cDNA arrays but not in Affymetrix arrays. Thus, two-sample test statistics should be used for selecting significantly expressed genes using Affymetrix arrays (11). Replications (at least two) should also be made for control sample to reduce measurement errors and to improve the sensitivity and specificity in the downstream statistical analysis. With the average expression profiles of control arrays, log ratios of expression profiles between the test and reference arrays for all genes can be computed (11). The downstream statistical analysis, such as tumor classification and biochemical expression pathways, can be analogously analyzed in the same way as the cDNA microarray.

Conclusion

Microarrays provide powerful tools for simultaneously monitoring mRNA expression in many areas of biomedical research. In these research endeavors, innovative statistical techniques and computing software are essential for the success of scientific investigations. Many of these software and procedures are available in the open source and development platform "bioconductor." This significantly facilitates the tools available for genomic and cancer research.

Acknowledgments

We thank Dr. Ernest Borden for critical reading of the article.

References

1. Dudoit Y, Yang Y, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stat Sin 2002;12:111–39.
2. Fan J, Tam P, Vande Woude G, Ren Y. Normalization and analysis of cDNA micro-arrays using within-array replications applied to neuroblastoma cell response to a cytokine. Proc Natl Acad Sci U S A 2004;101:1135–40.
3. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations, and assessment of gene effects. Nucleic Acids Res 2001;29:2549–57.
4. Fan J, Peng H, Huang T. Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency [with discussion]. J Am Stat Assoc 2005;100:781–813.
5. Huang J, Wang D, Zhang CH. A Two-way semi-linear model for normalization and analysis of cDNA microarray data. J Am Stat Assoc 2005;100:814–29.
6. Ma S, Kosorok MR, Huang J, Xie H, Manzella L, Soares MB. Robust semiparametric cDNA microarray normalization and significance analysis. Biometrics. In press 2006.
7. Cui X, Hwang JTG, Qiu J, Blades NJ, Churchill GA. Improved statistical tests for differential gene expression by shrinking variance components estimates. Biostatistics 2005;6:59–75.
8. Smyth GK, Michaud J, Scott HS. Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics 2005;21:2067–75.
9. Kerr MK, Churchill GA. Experimental design for gene expression microarrays. Biostatistics 2001;2:183–201.
10. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98:5116–21.
11. Fan J, Chen Y, Chan HM, Tam P, Ren Y. Removing intensity effects and identifying significant genes for Affymetrix arrays in MIF-suppressed neuroblastoma cells. Proc Natl Acad Sci U S A 2005;103:17751–6.
12. Lonnstedt I, Speed T. Replicated microarray data. Stat Sin 2002;12:31–46.
13. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004;3:Article 3.
14. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003;19:368–75.
15. Storey JD, Tibshirani R. Statistical significance for genome-wide studies. Proc Natl Aca Sci U S A 2003;100:9440–5.
16. Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J Am Stat Assoc 2004;99:96–104.
17. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Stat Sci 2003;18:71–103.
18. Inamura K, Fujiwara T, Hoshida Y, et al. Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization. Oncogene 2005;24:7105–13.
19. Zhang HP, Yu CY, Singer B. Cell and tumor classification using gene expression data: construction of forests. Proc Natl Acad Sci U S A 2003;100:4168–72.
20. Hastie TJ, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2001.
21. Svrakic NM, Nesic O, Dasu MRK, Herndon D,

Perez-Polo JR. Statistical approach to DNA chip analysis. Recent Prog Horm Res 2003;58:75–93.

22. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002; 99:6567–72.

23. Fan J, Li R. Statistical challenges with high dimensionality: feature selection in knowledge discovery. Proc Madrid Intl Congress Math. In press 2006.

24. Schulte J, Schramm A, Klein-Hitpass L, et al. Microarray analysis reveals differential gene expression patterns and regulation of single target genes contributing to the opposing phenotype of TrkA- and TrkB-expressing neuroblastomas. Oncogene 2005; 24:165–77.

25. Li H, Gui J. Gradient directed regulatioization for sparse Gaussian concentration graphs, with applications of inference of genetic networks. Biostatistics 2006;7:302–17.

26. Li C, Wong WH. Model based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A 2001;98: 31–6.

27. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 2003;31:e15.

28. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4: 249–64.