Published in final edited form as:
Curr Opin Biotechnol. 2008 February ; 19(1): 26–29.

Microarray Based Expression Profiling and Informatics

Richard Simon
National Cancer Institute 9000 Rockville Pike MSC 7434 Bethesda MD 20892 rsimon@nih.gov

Summary

Microarray based expression profiling is a powerful technology for studying biological mechanisms and for developing clinically valuable predictive classifiers. The high dimensional readout for each sample assayed makes it possible to do new kinds of studies but also increases the risks of misleading conclusions. We review here the current state-of-the-art for design and analysis of microarray based investigations.

Introduction

Microarray based gene expression profiling is a powerful technology that can be effectively used to (i) find genes whose expressions are correlated with a phenotype or (ii) find a classifier for predicting the phenotype of a sample. The first objective is often called *class comparison* in cases where the phenotype takes two or more categorical values. For example, one might look for the genes that are differentially expressed in cell lines containing a p53 mutation compared to other cell lines. The paradigm of finding genes correlated with a phenotype also includes problems where the phenotype is a quantitative measurement or even survival time of the patients whose tumors are being profiled. An example of the second kind of objective is prediction of whether a patient is likely to respond to a drug based on a pre-treatment expression profile of his or her tumor. Using a training set of expression profiles for patients who were treated with the drug and whose response is known, one can develop a predictive classifier for use with future patients.

Because gene expression profiling provides such a high-dimensional read-out for each specimen assayed, it offers both great opportunity for new kinds of investigation and great risk of error. If one compares expression profiles for each of 10,000 genes among samples for two classes of samples, even there are no genes that are really expressed differently in the classes there will on average be 500 false positive genes found statistically significantly differentially expressed between the classes at a $p<.05$ level. Clustering the expression profiles of the specimens using these 500 “significant” genes will generally produce two distinct clusters but the findings will be spurious [1]. Our objective here is to provide some guidance on the design and analysis of microarray expression profiles to biomedical scientists who are attempting to utilize this potentially powerful technology. The BRB-ArrayTools software contains the statistical methods we have found most appropriate and effective for the analysis of such studies [2]. The software is available at http://linus.nci.nih.gov/brb.

Study Design

Clear objectives are essential for the effective design of microarray studies. The objectives indicate the kinds of samples that should be included and the number of such samples. The

statistical power for identifying differentially expressed genes or for developing classifiers is generally determined by the number of *biological replicates* in each class. These are distinguished from *technical replicates* which are just repeat assays of the same RNA samples. Most commercial microarray platforms have reached a degree of reproducibility that technical replicates are of very limited value. Technical artifacts still exist, however, and so it is important to perform the assays in a manner that does not confound phenotype classes with assay performance. For example, in comparing expression of p53 mutant cell lines to p53 wild type cell lines, one should avoid assaying all the mutants with one set of reagents on one week and the wild type cell lines with a different set of reagents on another week. If a large number of samples are to be assayed, the phenotype classes should be intermixed in the group assayed at each time. Pooling samples is generally not advantageous [3]. When dual-label arrays are used, there are additional design issues to be addressed concerning whether to use a common reference RNA or to pair the samples from different classes for co-hybridization on each array. Dobbin et al. provide a thorough discussion of this issue [4]. Dobbin and Simon provide formulas and graphs for determining the number of experimental/biological replicates needed for class comparison problems [5] or for developing a predictive classifier.[6]

Finding Genes Whose Expression is Correlated with a Phenotype

In finding genes whose expression is correlated with a phenotype a key analysis objective is to limit the number of false positive findings. Many publications have used average fold change between the classes to identify differentially expressed genes. This approach, however, ignores variation of gene expression among samples within the same class, ignores the fact that the variation differs among genes, and does not provide any control on the number of false positive findings. The simplest approach to addressing these deficiencies is to use a simple statistical test, such as a t-test to evaluate differential expression separately for each gene. By using a stringent threshold of significance the number of false positive findings can be limited; a threshold of $p<.001$ results in 1 false positive gene per 1000 genes analyzed on average. If there are few samples per class, however, the statistical power of this approach will be poor because the estimates of within-class variation, made separately for each gene, will be very imprecise. Improved methods based on t or F statistics which borrow variance information among genes are recommended if there are less than 10 samples per class. [7,8] These methods are called regularized t-tests, random variance t-tests or empirical Bayes t-tests. They are based on the assumption that the within class variances for different genes come from the same distribution, but not that they are equal.

More sophisticated multivariate testing procedures can provide greater power than the regularized t-tests while controlling the number or proportion of false discoveries. If N genes are reported as differentially expressed among classes and m of those are false positives, then m/N is the false discovery proportion. The expected value of m/N is called the *false discovery rate*. When using univariate methods like the regularized t-test, one can compute a conservative estimate of the false discovery rate as $p_{[m]}N / m$ where $p_{[m]}$ denotes the m'th smallest p value among the N genes evaluated. [9] The widely used SAM method of Tusher et al. [10] is a multivariate method that controls the false discovery rate. The multivariate permutation test of Korn et al. [11] controls the probability that m/N exceeds a specified limit; it can also be used to control the probability than m exceeds a specified number. These methods take advantage of the correlation of expression among different genes and are effective even when there are relatively few samples per class. A comparison of methods for finding genes whose expression is correlated with phenotype was reported by Jeffery et al. [12]

Most of the methods used for finding genes whose expression is correlated with a phenotype can be used with categorical phenotypes, quantitative phenotypes or survival time phenotypes. The measure of correlation used for each gene varies depending on the type of phenotype of

interest. For categorical phenotypes the multivariate methods such as SAM and the multivariate test of Korn et al. is based on computing regularized t-tests for each gene. For survival time phenotypes p values from univariate proportional hazards regression analyses can be used.

With time course data the phenotype is time after an experimental intervention and the basic analysis is for the purpose of identifying genes whose expression is changing with time. Those genes can be identified in a manner that controls the number or proportion of false discoveries. Clustering those genes sorts them into sets showing similar patterns over time. One can also identify genes whose variation with time differs based on some other phenotype. [13] Supervised methods for analyzing time course data are available in specialized software [2, 14].

In the past investigators have generally first identified those genes whose expression is correlated with a phenotype and then used functional annotations to try to understand the interrelationships among the genes. The effectiveness of this post-hoc annotation of gene lists is limited by the statistical stringency necessary in creating the gene lists in order to limit the false discovery rate. More recently methods have become available that utilize annotation information prospectively in the identification of gene sets whose expression is correlated with phenotype. For any a priori specified set of genes, one tests either (i) whether the degree of correlation among phenotypes for the genes in the set is greater than one would expect for a random set of genes represented on the array; or (ii) whether any genes in the set have expression correlated with the phenotype. A number of statistical methods have been proposed for testing these hypotheses. [15-19] For example, BRB-ArrayTools includes gene lists for sets of genes with the same Gene Ontology annotation, sets of genes for each Kegg or Biocarta pathway, sets of genes for each Broad Institute signature, sets of genes that are targets of the same transcription factor, sets of genes that are putative targets of the same microRNA, and sets of genes that contain a common protein domain.

Class Prediction

Many prognostic factor studies are conducted using a convenience sample of available specimens from a heterogeneous group of patients who have received a variety of treatments. Showing that a new classifier is prognostic for such a mixed group often has uncertain therapeutic relevance. Predictive classifiers that identify which patients respond to specific treatments are often more valuable. In planning a study to develop a predictive classifier, considerable care should be given to selecting cases so that the result has potential therapeutic relevance. Very often this objective can be enhanced by selecting cases who participated in an appropriate clinical trial.

Numerous algorithms have been used effectively with DNA microarray data for class prediction. Many of the widely used classifiers combine the expression levels of the genes selected as informative for discrimination using a weighted linear function

$$l\left(\underline{x}\right) = \sum_{i \in G} w_i x_i \qquad (1)$$

where x_i denotes the log-ratio or log-signal for the i'th gene, w_i is the weight given to that gene, and the summation is over the set G of genes selected for inclusion in the classifier. For a two-class problem, there is also a threshold value d; a sample with expression profile defined by a vector $\underline{x}$ of values is predicted to be in class 1 or class 2 depending on whether $l(\underline{x})$ as computed from equation (1) is less than the threshold d or greater than d respectively. Many of the widely used classifiers are of the form shown in (1); they differ with regard to how the weights are determined.

Dudoit et al. [20,21] compared many classification algorithms and found that the simplest methods, diagonal linear discriminant analysis and nearest neighbor classification, usually performed as well or better than the more complex methods. Nearest neighbor methods are not of the linear form shown in (1). They are based on computing similarity of a sample available for classification to samples in a training set. Often Euclidean distance is used as the similarity measure, but is calculated with regard to the set of genes selected during training as being informative for distinguishing the classes. The PAM method of Tusher et al. is a popular form of nearest neighbor classification. [10] Ben-Dor et al. [22] also compared several methods and found that nearest neighbor classification generally performed as well or better than more complex methods. Similar results were found by Wessels et al. [23]. In addition, Wessels et al. and Lai et al. [24] found that simple gene selection strategies generally worked as well or much better than more complex multivariate strategies. The simple strategies generally select genes based on their univariate correlation with the class phenotype; e.g. using t-statistics. Multivariate methods attempt to identify sets of genes that work well together for classification. Because of the large numbers of candidate genes and the larger numbers of ways of combining the genes, few datasets are large enough to support multivariate gene selection without over-fitting in a manner that results in poor prediction for independent samples. Lai et al. point out serious biases in the way that many of the multivariate methods have been evaluated, resulting in unsubstantiated claims.

A cardinal principal for evaluating a predictive classifier is that the data used for testing the classifier should not be used in any way for building the classifier. The simple *split-sample* method achieves this by partitioning the study samples into two parts. The separation is often done randomly, with half to two-thirds of the cases used for developing the classifier and the remainder of the cases in the test set. The cases in the test set should not be used in any way, until a single completely specified model is developed using the training data. At that time, the classifier is applied to the cases in the test set. For example, with an expression profile classifier, the classifier is applied to the expression profiles of the cases in the test set and each of them are classified, as a responder or non-responder to the therapy. The patients in the test set have received the treatment in question and so one can count how many of those predictive classifications were correct and how many were incorrect. In using the split sample method properly, a single classifier should be defined on the training data. It is not valid to develop multiple classifiers and then use their performance on the test data to select among the classifiers.[25]

There are more complex forms of dividing the data into training and testing portions. These *cross-validation* or *re-sampling* methods utilize the data more efficiently than the simple division described above [26]. Cross-validation generally partitions the data into a large training set and a small test set. A classifier is developed on the training set and then applied to the cases in the test set to estimate the error rate. This is repeated for numerous training-test partitions and the prediction error estimates are averaged. In order to honor the key principal of not using the same data to both develop and evaluate a classifier, it is essential that for each training-test partition the data in the test set is not used in any way[27]. Hence a model should be developed from scratch in each training set. This means that multiple classifiers are developed in the process of doing cross-validation and those classifiers will in general involve different sets of genes. It is completely invalid to select the genes beforehand using all the data and then to just cross-validate the model building process for that restricted set of genes. Radmacher et al.[28] and Ambroise and McLachlan [29] demonstrated that such pre-selection results in severely biased estimates of prediction accuracy. In spite of this known severe bias, this error is made in many developmental classifier studies. It is also made in many biased reports touting the merits of new kinds of classifiers [24].

Conclusion

Gene expression profiling is a powerful tool for elucidating biological mechanisms and moving medicine toward a more predictive future. Effective use of this technology requires substantially increased emphasis on interdisciplinary collaboration for the design and analysis of studies. The current state of the literature with regard to analysis of microarray expression data is of serious concern[1]. The key limitation for effective use of this technology is not software engineering for managing large datasets. Development of high dimensional biotechnology also highlights the importance of new directions for the training of biomedical scientists.

REFERENCES

1. Dupuy A, Simon R. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. Journal of the National Cancer Institute 2007;99:147–157. [PubMed: 17227998]
2. Simon R, Lam A, Li MC, Ngan M, Menenzes S, Zhao Y. Analysis of gene expression data using BRB-ArrayTools. Cancer Informatics 2007;2:11–17.
3. Shih JH, Michalowska AM, Dobbin K, Ye Y, Qiu TH, Green JE. Effects of pooling mRNA in microarray class comparison. Bioinformatics 2004;20:3318–3325. [PubMed: 15247103]
4. Dobbin K, Shih J, Simon R. Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. Journal of the National Cancer Institute 2003;95:1362–1369. [PubMed: 13130111]
5. Dobbin K, Simon R. Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics 2005;6:27–38. [PubMed: 15618525]
6. Dobbin K, Simon R. Sample size planning for developing classifiers using high dimensional DNA expression data. Biostatistics 2007;8:101–117. [PubMed: 16613833]
7. Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. Proceedings of the National Academy of Science 2003;100:9991–9996.
8. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inference of gene changes. Bioinformatics 2001;17:509–519. [PubMed: 11395427]
9. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003;19:368–375. [PubMed: 12584122]
10. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Science 2001;98:5116–5121.
11. Korn E, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: Application to high-dimensional genomic data. Journal of Statistical Planning and Inference 2004;124:379–398.
12. Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics 2006;7:359. [PubMed: 16872483]
13. Storey JD, Xiao W, Leek JT, Tomkins RG, Davis RW. Significance analysis of time course microarray experiments. Proceedings of the National Academy of Science 2005;102:12837–12842.
14. Leek JT, Monsen E, Dabney AR, Storey JD. EDGE: extraction and analysis of differential gene expression. Bioinformatics 2006;22:507–508. [PubMed: 16357033]
15. Subramanian A, Tamayo P, Mootha VK. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Science 2005;102:15545–15550.
16. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. Proceedings of the National Academy of Science 2005;102:13544–13549.

17. Kong SW, Pu WT, Park PJ. A multivariate approach for integrating genome-wide expression data and biological knowledge. Bioinformatics 2006;22:2373–2380. [PubMed: 16877751]

18. Jiang Z, Gentleman R. Extensions to gene set enrichment. Bioinformatics 2007;23:306–313. [PubMed: 17127676]

19. Pavlidis P, Qin J, Arango V, Mann JJ, Sibille E. Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. Neurochemical Research 2004;29:1213–1222. [PubMed: 15176478]

20. Dudoit, S.; Fridlyand, J. Classification in microarray experiments.. In: Speed, T., editor. Statistical analysis of gene expression microarray data. Chapman & Hall/CRC; 2003. p. 93-158.

21. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for classification of tumors using gene expression data. Journal of American Statistical Association 2002;97:77–87.

22. Ben-Dor A, Bruhn L, Friedman N. Tissue classification with gene expression profiles. Journal of Computational Biology 2000;7:536–540.al. e

23. Wessels LFA, Reinders MJT, Hart AAM, Veenman CJ, Dai H, He YD, Veer LJvt. A protocol for building and evaluating predictors of disease state based on microarray data. Bioinformatics 2005;21:3755–3762. [PubMed: 15817694]

24. Lai C, Reinders MJT, Veer LJvt, Wessels LFA. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC Bioinformatics 2006;7:235. [PubMed: 16670007]

25. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 2006;7:91. [PubMed: 16504092]

26. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: A comparison of resampling methods. Bioinformatics 2005;21:3301–3307. [PubMed: 15905277]

27. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the analysis of DNA microarray data: Class prediction methods. Journal of the National Cancer Institute 2003;95:14–18. [PubMed: 12509396]

28. Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. Journal of Computational Biology 2002;9:505–511. [PubMed: 12162889]

29. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of the National Academy of Science 2002;99:6562–6566.