Translational Review

Practical Approaches to Analyzing Results of Microarray Experiments

Naftali Kaminski and Nir Friedman

Departments of Functional Genomics and Respiratory Medicine, Sheba Medical Center, Tel-Hashomer; and School of Computer Science and Engineering, Hebrew University, Jerusalem, Israel

Microarray technology is rapidly becoming a standard laboratory technique. The main challenges related to the successful implementation of the technology are analysis-related. In this article we provide a practically oriented review focusing on methods for analysis of large-scale gene expression data in the research laboratory. We describe the various common clustering methods and outline our approach to using them. We discuss methods for scoring genes for their relevance, focusing on the statistical meaning of microarray results, especially with regard to the problem of multiple testing. We also deal with the problem of adding biologic meaning to the results of microarray experiments and describe advanced tools that represent different but valid directions in providing automated solutions to this problem. The tools and approaches described and discussed here should provide the reader with a preliminary understanding of the analysis of the results of microarray experiments. The practical focus of this review should remove the mystery behind the analysis of microarray experiments, thus leading to more productive and efficient use of the technology.

Microarray technology is rapidly becoming a standard technique used in research laboratories all across the world. In essence, all the variants of the technology allow simultaneous profiling of the expression levels of tens of thousands of genes, potentially whole genomes in a single experiment (1–3). This unique power provides scientists with an opportunity to look at the transcriptional profile of biologic systems, processes, and diseases in an unbiased fashion. The relative ease (despite the prohibitive cost) of performing microarray experiments in molecular laboratory settings, combined with the potential power of the technology, have captured the imagination of scientists in academic and industry research institutes. This combination of ease of use with unforeseen power also appealed to administrators in the same institutes and in funding agencies, thus leading to rapid spread of the use of the technology. Many research groups in the academy and industry have implemented microarrays in multiple experimental settings with varying degrees of success. Microarrays have been successfully applied to almost every aspect of biomedical research (4–9). However, many more experiments remain uncompleted, or even worse, unavailable to the scientific community. Interestingly, instead of excitement at the results of experiments that utilize microarrays, scientists often express some confusion, a tendency to focus on what is already known, and a sense of weariness. These feelings arise from the tension between the relative ease of producing the results and the objective difficulty of dealing with the results. Provided that one has the funding and setting, it should not take one trained laboratory assistant more than a week to run 8–10 samples (from total RNA extraction to microarray hybridization). A month's worth of results from such a laboratory assistant will leave the investigator with the task of dealing with half a million or more data points. The objective difficulty resulting from the large quantities of information generated by the experiments is further complicated by the lack of simple and accepted approaches to analyzing large-scale gene expression data (or even visualizing the data). Additionally, the diversity of experimental designs and schemes adds to the confusion. Microarrays are used to classify diseases, to identify the effects of a stimulus *in vivo* or *in vitro*, to single out genes that may play a role in a specific disease or a specific biologic process, and to distinguish transcriptional programs that underlie such a process. The size of experimental groups and the design of the experiments also vary widely, from time-course experiments to cross-sectional studies, from single observations with no repeats to analysis of hundreds of samples. Naturally, these diverse experimental schemes pose diverse computational requirements—the analysis of an experiment designed discover a new class of a disease is different from an experiment designed to test the immediate targets of a known transcriptional activator. The wealth and complexity of information that characterizes results of microarray experiments has led to the suggestion that there may not be a single "best" analytic approach and that indeed the application of several analytical and computational approaches to a dataset may aid in the exposure of different and complementary aspects of the data (10, 11).

As this is a rapidly evolving field, an attempt to provide a comprehensive review of available computational tools and approaches may prove outdated even before publication. Thus, we aim to provide the reader with a practical approach to analysis of microarray results with examples of publicly available computational tools. We focus on the most commonly used analytic tools such as clustering methods, on methods that deal with the statistical aspects of the analysis of the results of microarray experiments, and on methods that we have been involved in developing. We also discuss some advanced methods that go "beyond" clustering that are particularly useful in the analysis of

complex experimental settings and as hypothesis-generating tools. We do not, however, deal with questions that may be more relevant to clinical research, such as class discovery and class prediction. These have been reviewed by us elsewhere (12, 13). The approaches outlined here do not deal with normalization issues, and assume that the data obtained from large-scale gene expression studies and used for the analysis is technically and biologically sound. The analytical tools described and discussed here, as well as the outlined approach, should provide the reader with a good understanding of what is possible in computational analysis of microarray data.

The Gene Funnel

Generally, one can describe the process of analyzing gene expression data as a funnel-shaped process. You start out with many genes, and by applying filters this number is gradually reduced. In our team, the first step in the analysis involves looking at the experiments and detecting outliers and technically inferior experiments. We have noticed that because of the wealth of information it is often very hard to identify a "bad" experiment if it managed to escape this preliminary screen. Once we are convinced that all arrays lack defects and are relatively comparable, we define a set of genes that we term "legal genes." These are genes that pass a certain threshold of expression in at least one of the experiments. Because we run our experiments on Affymetrix microarrays (Affymetrix, Santa Clara, CA), we use the calls provided by the Affymetrix software analysis suite for this initial filtration. The parameters that we use are the signal (the gene expression level) and the detection call (present or absent call) (14). We usually eliminate genes that do not have a present call in a fraction of the experiments (5% or 1 if less then 20 microarrays). We also set an expression level threshold that a gene needs to pass in a fraction of the experiments (5%) to be included in the dataset. To determine this threshold we hybridize the same sample on two microarrays and compare the expression levels (Figure 1A). As a general rule, the consistency between values obtained from running the same sample on two microarrays is intensity-dependent—that is, the lower the intensity is, the lower the agreement between the two microarrays. Often, a discernible threshold can be observed (Figure 1A). Over this threshold the consistency is indeed impressive and lies within the 2-fold range (Figure 1A), a phenomenon not observed when two different samples are compared (Figure 1B). This filtration process usually reduces the number of the genes in the dataset by a third to a half (depending on the sample number, the type of tissues being investigated, etc.). The next step is to define a set of "active genes." These are genes in which "something" happened to their expression level. There are many computational and statistical ways to define an active gene. We take a simple and straightforward approach. We start by converting the data to be in terms of expression ratios. This is the natural representation in two dye-competitive hybridization systems but not in Affymetrix microarrays, where the readings are absolute gene expression levels. What we often do is what we term "virtual two dye." Basically, we create a set of ratios by dividing the values of every gene by a

number. This number may be the geometric mean of the controls if we are dealing with distinct groups, the geometric mean of all the experiments, or of any other group of experiments. This transformation allows us to query the genes for their activity. We usually use relatively weak filters and ask for genes that changed at least 1.5-fold in every direction in at least 5% of the experiments (or 1 if there are fewer than 20 experiments). Genes that did not vary are excluded from the "active genes" dataset, but may serve later for controls in verification experiments. This process usually greatly reduces the number of genes in the dataset. It also allows for more stringent and specific queries, such as genes that changed in a certain direction in a certain subset of the experiments. The "active genes" dataset will be the dataset that we will use for all other analyses.

Clustering Methods

The most commonly used analysis tools are clustering methods (10). Clustering methods attempt to identify genes
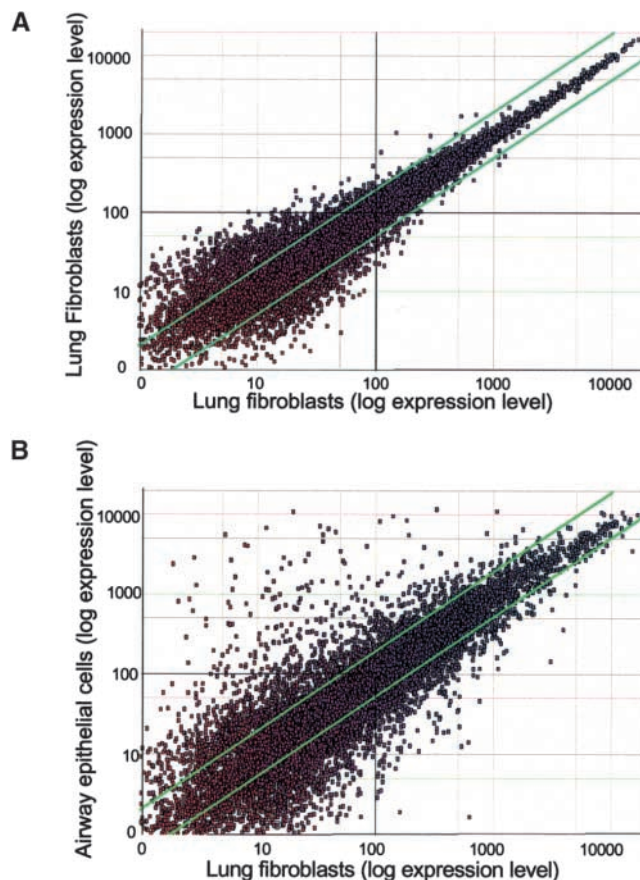

Figure 1. A *scatter plot* of gene expression levels of the same sample (a pool of seven plates of primary fibroblasts) run on two GeneChip U95A Arrays (Affymetrix) (*A*) and of two different samples (a pool of seven plates of lung fibroblasts and a pool of five normal human bronchial epithelial cells) (*B*). The *green oblique lines* correspond to 2-fold change. Pools were generated from equal amounts of labeled cRNA that was previously hybridized into individual arrays. The experimental design was previously described by us (23).

that behave similarly across a range of conditions or samples. The motivation to find such genes is driven by the assumption that genes that demonstrate similar patterns of expression share common characteristics, such as common regulatory elements, common functions, or (in the case of mixed tissue studies) common cellular origin. This assumption is supported by the successful application of clustering algorithms to the analysis of basic mechanisms of cellular function in yeast and mammalian systems (6, 9, 15, 16). The main advantage in clustering tools is that using them mitigates the inherent difficulty in becoming familiar with the results. By grouping the genes into clusters that behave similarly, these methods allow the investigator to browse the data in a less intimidating and chaotic atmosphere. Additionally, many of the methods are relatively easy to visualize, thus improving the accessibility of the biologically meaningful information that is in the data. Several clustering methods have been applied to gene expression data, hierarchical clustering (17) being the most popular, but others include k–means (18), deterministic annealing (19), self-organizing maps (20), combinatorial methods and graph theoretical approaches (21, 22), and super-paramagnetic clustering (16). The results of clustering algorithms are highly dependent on the input, i.e., on the data that is used for the analysis. Methods for gene filtering are commonly applied. It is important to remember and note that various schemes for selecting genes that are legal (with expression levels that pass a certain level) and active (that change compared with a certain threshold) may affect the results of the analysis, as will inclusion of multiple and diverse experiments. In general, we apply more than one clustering method to a dataset. We usually compare the results of several methods and of several gene filtering schemes before we decide what is the true signal in this dataset. Many times the decision will lie on the reproducibility of the cluster using various methods. This process, although tedious, allows us to gain confidence that the patterns we observe represent true biologic phenomena that are independent of the analysis method and, as previously stated, to become familiar with the data and with the true signal that characterizes it.

Hierarchical Clustering

This is probably the most popular clustering approach to large-scale gene expression data. The principles behind this approach are reasonably easy to understand and very intuitive to visualize. Tools for hierarchical clustering (Cluster) and visualization (Treeview) designed by Michael Eisen (17) were freely available to the research community from the very early stages of the introduction of microarray technology, making this clustering into a standard in the field. Basically, this is an agglomerative process in which single-member clusters are fused to bigger and bigger clusters. In somewhat more detail, the procedure starts by computing a pairwise distance matrix between all the genes, the distance matrix is explored for the nearest genes, and they are defined as a cluster. After a new cluster is formed by agglomeration of two clusters, the distance matrix is updated to reflect its distance from all other clusters. Then, the procedure searches for the nearest pair of clusters to agglomerate, and so on. This procedure leads to a hierarchical dendogram in

which multiple clusters are fused in nodes according to their similarity, finally resulting in a single hierarchical tree. There are several hierarchical clustering algorithms that differ in the way the distances are calculated. As mentioned earlier, Cluster and Treeview can be obtained from Michael Eisen's lab at http://rana.lbl.gov/EisenSoftware.htm.

k-Means Clustering

This is an iterative procedure that searches for clusters that are defined in terms of their "center" points or means. Once a set of cluster centers is defined, each gene is assigned to the cluster it is closest to. The clustering algorithm then adjusts the center of each cluster of genes to minimize the sum of distances of genes in each cluster to the center. This results in a new choice of cluster centers, and so we can reassign genes to clusters and repeat the process. These iterations are applied until convergence. This method has more of a "global" character than hierarchical approaches. It does not generate a hierarchical tree, but rather a predetermined number of clusters. k-Means clustering is especially useful in cases in which one knows how many distinct gene expression patterns to expect. The site previously mentioned also provides tools for k-means calculations.

Self-Organizing Maps

Self-organizing maps (SOMs) were introduced to the analysis of microarray data by Tamayo and coworkers (20). As in the k-means procedure, one assigns the data into a predetermined set of clusters. However, unlike k-means, what follows is an iterative process in which gene expression vectors in each cluster are "trained" to find the best distinctions between the different clusters. In other words, a partial structure is imposed on the data and then this structure is iteratively modified according to the data. SOM is superior when dealing with "messy" data that contains outliers and irrelevant parameters. We frequently apply SOM analysis to new datasets, as a means to browse the trends in the data and to detect outliers. Although many software packages contain SOM feature, we use GeneCluster that can be obtained from Whitehead Institute/MIT Center for Genome Research website at http://www-genome.wi.mit.edu/cancer/software/software.html.

Graph-Theoretic Clustering

Graph-theoretic clustering (CLICK) is an innovative clustering method that utilizes graph-theoretic and statistical techniques to identify tight groups of highly similar elements (kernels), which are likely to belong to the same true cluster (22). Several heuristic procedures are then used to expand the kernels into the full clustering. CLICK and a visualization tool (Expander) are available at http://www.math.tau.ac.il/~roded/click.html.

We do not have a real preference for using a specific clustering method, although each has its own computational merits. Most often, we apply more than one clustering method on a dataset and we feel more confident if clusters are reproduced using different methods.

Finding Relevant Genes

Relevant genes are defined as genes in which the expression level characterizes a specific experimental condition.

Usually, these are genes in which the expression levels differ significantly between different experimental conditions. The simplest way is to compare the expression level of a gene in the different conditions and look at the relative expression – "fold change" of the gene: the ratio of its expression level in one set of samples to its expression in another set. This can even be done between two samples. Naturally, the question immediately arises: what is the threshold of relative expression at which a gene is considered changed? Many published studies use a threshold of 2-fold change that is based on the earliest microarray works (1). However, the ratio itself is not a good predictor of the relevance of a gene. In an ongoing study of lung tumors in our lab, the fold ratio of 2 and more in a single tumor compared with its matching control did not predict the fold ratio of this same gene calculated from a comparison of 20

lung tumors to 7 normal tissues (Figure 2A). Impressively, there were genes with fold ratios of more than 8 in the single sample comparison that had changed in the opposite direction when averages were calculated. This observation was also true when we compared fold ratios of a single pair of samples (tumor/normal) to fold ratios of a pool of tumor RNAs and a pool of normal RNAs (Figure 2B), and in primary cell culture experiments (Figure 2C). Furthermore, the genes that had the highest fold ratios in the single-sample comparisons were not necessarily those that had significant P values in the multiple replicate analysis (Figures 1A, 1B, and 1C, *blue dots*). When we compared fold ratios calculated from the pools comparison with those obtained from the average comparison, we did observe that genes substantially changed in pooled samples comparison seemed to better overlap with genes with significant P val-
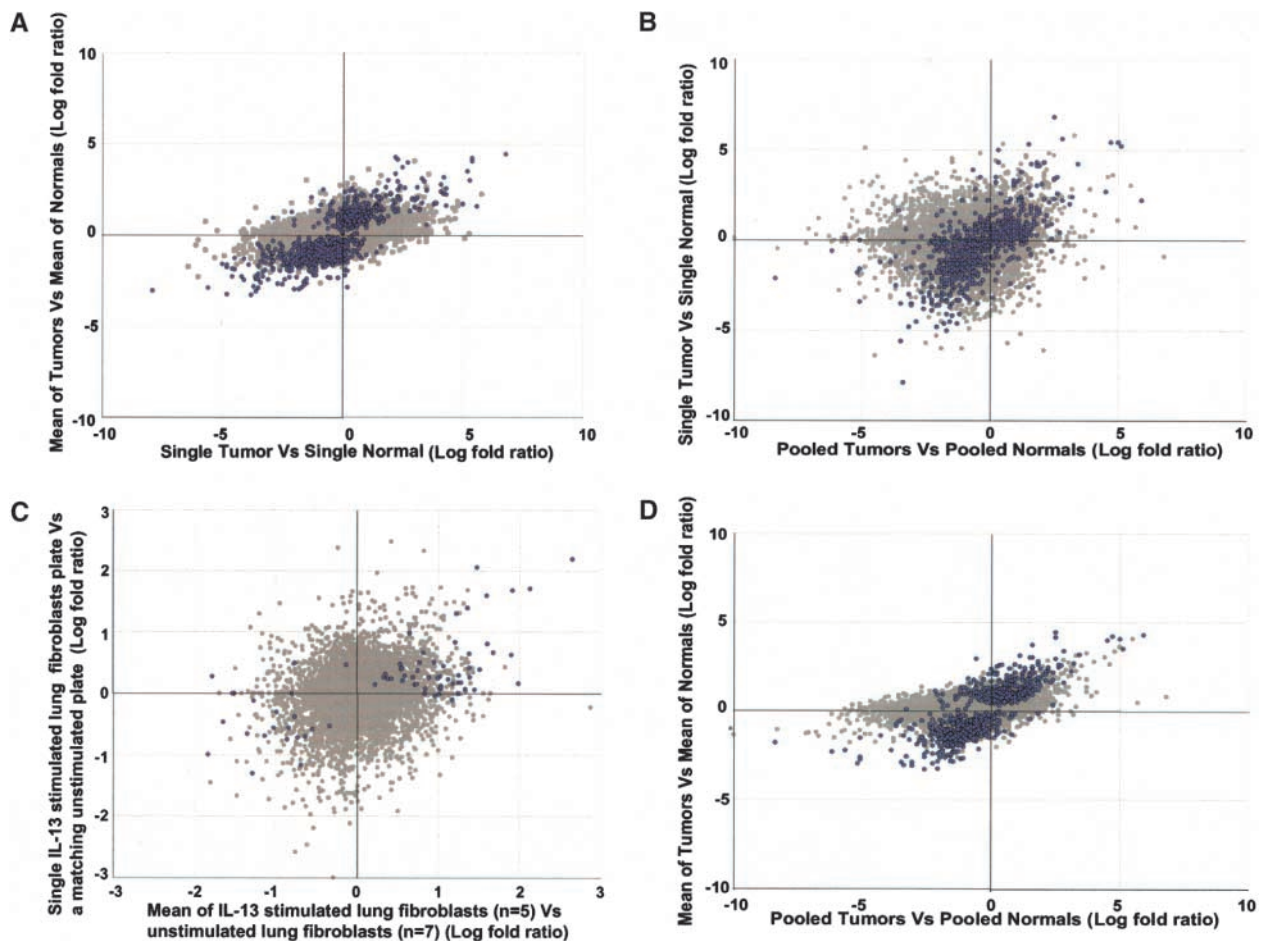

Figure 2. (*A*) Comparison of fold ratios obtained from comparing fold ratios obtained from the average of multiple samples (20 lung tumors compared with 7 normal lung tissues) with fold ratios obtained from a single comparison (one tumor compared with its matching normal). (*B*) Comparison of fold ratios obtained from a single comparison with fold ratios obtained from pooled (20 pooled tumors hybridized on one microarray compared with 7 pooled normal samples). (*C*) Comparison of fold ratios from average multiple cell culture samples (five plates of IL-13–stimulated primary lung fibroblasts compared with seven plates of untreated primary lung fibroblasts) with fold ratios obtained from a single comparison (one plate of IL-13–treated primary lung fibroblasts compared with one plate of untreated lung fibroblast). (*D*) Comparison of fold ratios of average of multiple samples with fold ratios of pooled samples. Pools were generated as described in Figure 1. Data in *A*, *B*, and *D* was generated as previously described (33), and data in *C* was previously described (23). *Blue dots* represent genes that significantly distinguished between experimental groups (*P* < 0.05, using TNoM score), *gray dots* represent genes with *P* > 0.05.

ues (Figure 2D) than in the other comparisons. This analysis suggests (as one would intuitively presume) that basing an observation on fold ratio may easily lead to spurious results, especially when a small amount of samples is analyzed. A slightly modified version of looking at fold changes that deals with the diversity of the data is D's (n-1) score (23). In this simple approach, we set up a fold threshold for most of the samples, for example demanding that the fold ratio of n-1 of the experiments will be more than Z-fold over the mean of the controls. This score could be made stricter by setting the Z threshold higher, or by demanding that the value Z will be calculated against the maximum (for increased) or minimum (for decreased) values of the control group. The major problem with relying on fold changes for finding relevant genes is that they do not provide any means to measure the statistical relevance of the results.

Scoring Methods

Statistical considerations are essential when trying to find genes whose expression characterizes a specific experimental condition. The available methods can be divided into parametric and nonparametric methods.

Parametric Methods

These approaches model expression profiles within a parametric representation, and ask how different the parameters of the experimental groups are. A simple example is the classic t test (19). Other examples of parametric approaches are the separation score (24) and the Bayesian t test (25). The significance analysis of microarrays (SAM) method developed at Stanford University Statistics and Biochemistry Labs (26) is a method for identifying genes with statistically significant changes in gene expression. It deals with the specific issues of multiple testing (*see* STATISTICAL SIGNIFICANCE) and is easily installed as an Excel add-in. SAM is freely available from http://www-stat.stanford.edu/~tibs/SAM/ and we frequently use it.

Nonparametric Methods

No *a priori* assumptions are made about the distribution of expression profiles in the data. Instead, we attempt to directly examine the degree to which the two groups of expression measurements are distinguished. The methods that we use were introduced by Ben-Dor and coworkers (13) and applied successfully to the study of breast cancer (27), melanoma (18), and recently to idiopathic pulmonary fibrosis (28).

Threshold Number of Misclassifications

The threshold number of misclassifications (TNOM) measures how successful we are in separating the two groups of samples by a simple threshold over the expression values. That is, we search for the threshold value of the gene's expression that will distinguish the experimental conditions. A gene is scored by the number of misclassifications made by the best threshold that we can find for it. If the expression value of the gene allows us perfectly separate the groups, the gene has a TNOM score = 0. On the other hand, if the two groups are interspersed, the gene has a score that may be close to the size of the smallest group of samples.

INFO score is a refined version of TNOM that measures the misclassifications made by a simple threshold in terms of the information lost (or entropy) of the labels of samples in each side of the threshold.

There are several other scores that we often use including the Gaussian score and D's score (n-1) (23, 28). ScoreGene—a package that calculates the scores for a given dataset—will soon be available on our website at http://fgusheba.cs.huji.ac.il/software.htm.

The main weakness of parametric approaches is the assumptions that they make about the data. For example, outliers can significantly skew the t test score by changing the variance estimated in the sample. Similarly, scale transformation (e.g., working in logarithmic scale) can have drastic impact on the scores. Nonparametric approaches are more robust to this type of phenomenon. This comes at a price, as nonparametric approaches are less sensitive to the actual expression values.

Statistical Significance and Multiple Testing

A unique challenge posed by the results of microarray experiments to the statistical analysis is the asymmetry between the number parameters (genes) measured and the number of samples. No matter what statistical method that we apply to the data, it will always be possible that a certain percentage of the "significant" P values will be spurious. To handle this problem it is imperative to determine the degree of significance of the computed P values.

One approach is to find a value q, such that the probability that any of the events (i.e., the smallest one) has a P value less than q is small. This standard way of selecting q is using the Bonferroni threshold (29), defined as the allowed error probability divided by the number of parameters measured (genes in our case). For example, to ensure that the probability of a false recognition is < 0.05 (i.e., 95% significance level), we need to set the Bonferroni threshold q as 0.05 divided by the number of the genes in the analysis. This will lead to a P value threshold of $\sim 10^{-6}$. The stringency of the Bonferroni threshold ensures that each and every validated scoring event is indeed a significant event. Our aim, however, is slightly different. We want to retrieve a set of events, such that most of them are not spurious. A statistical method that addresses this kind of requirement is the False Discovery Rate (FDR) method (30). In this method, all P values are ranked and tested against different thresholds. The genes are ranked by their P values. The best P value is tested against the Bonferroni threshold; however, the next P value is tested against a more relaxed threshold, and so on. This replaces a strict validation test of single events with a more tolerable version validating a group of events. Tools such as the FDR threshold allow us relax the Bonferroni threshold and to locate promising genes for further examination with minimal spurious events.

Overabundance Analysis

Another approach is to ask "how surprising is the data set?" The approach that we take to this problem is to examine the number of genes at different P values (i.e., significance levels) and compare them with the expected num-

ber under the null-hypothesis (the assumption that the separation of the samples is random). We can visualize this difference using overabundance graphs (Figure 3). The difference between the expected and observed number of genes in each significant P value is an estimate of the overabundance of information in the analyzed dataset (Figure 3). We find this analysis extremely useful and very easy to perform using the ScoreGene package (http://fgusheba.cs.huji.ac.il/software.htm).

Loading the Results with Biologic Meaning

One of the most painful stages in the analysis of the results of microarray experiments is the postcomputational stage. The long-awaited and dreaded computational analysis has finally been completed, and now the investigator has to examine the lists of genes he or she so eagerly anticipated and to generate new hypotheses. In the easiest cases, more than half of the genes in these lists will be familiar to the investigator; however, it will still be difficult to put them in a meaningful biologic framework without focusing mainly on what was previously known. To fully realize the potential of microarray experiments as hypothesis-generating tools, it is essential to have tools that will facilitate the process of looking for biologic meaning in the data. We would like to mention three examples for directions to automatically load the results with biologic meaning.

Probabilistic Relational Models

Probabilistic relational models (PRM) allow the inclusion of multiple types of information in the computational process itself (31). For example, experiments can be annotated by their gene expression patterns, experimental or clinical data, the cell type or strain used in the experiment, the cellular phenotype triggered by each condition, etc. The genes can be annotated by the gene expression data, sequence elements present in the gene promoters, functional information, gene ontology definitions, protein motifs, and more. The analysis then creates context-specific groupings (clusters) of genes and experiments that are enriched with biologic information, thus facilitating the hy-
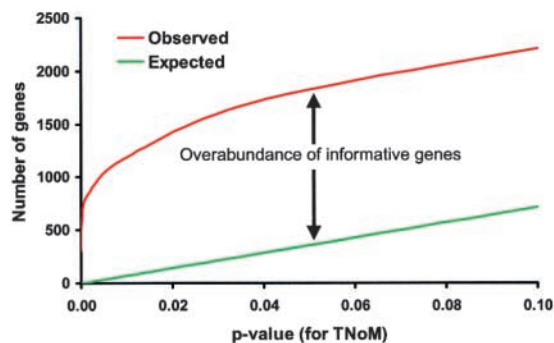
pothesis-generating process. In our view PRMs represent the first tool that allows statistically sound, unbiased integration of biologic information and microarray results.

GenMAPP

GenMAPP represents a completely different approach, which is to analyze gene expression data in the context of known biologic pathways (32). It allows the uploading of gene expression data unto known pathways and gene families using complex selection criteria. Furthermore, it allows investigators to author their own pathways using a simple and intuitive tool. We often run our data through GenMAPP to browse through processes with which we are less familiar, to see whether there is a specific signal worth following up. GenMAPP is available at http://www.genmapp.org/.

Pubgene

Available at http://www.pubgene.uio.no/cgi-bin/PubGene.cgi, this tool allows the user to upload results of microarray experiments and to search for interesting literature clusters. The literature clusters are then ranked according to the data analysis parameters and the user-defined number of the highest scoring clusters are displayed as a result. It allows the investigator to rapidly get a feeling about the relationship of the results of his microarray experiments to published literature. Naturally, this does not replace the actual work of looking for and reading the literature, but it can quickly and easily highlight a specific literature cluster that might have been otherwise overlooked.

Conclusions and Practical Suggestions

In this article, we provided practical approaches to analysis of microarray experiments. We presented our approach to data filtering, commonly used clustering methods, tools for finding informative genes, and some of the newer tools that go "beyond" clustering. As opposed to the situation two years ago, today there is a wide range of analysis tools for examining microarray experiments. There is no single "best" method for analysis, but there are multiple tools that can be applied to the data and allow querying it from different angles. Clustering methods may help in looking at gene expression patterns, scoring methods to identify genes that are statistically associated with the process studied, and the advanced tools to facilitate the generation of new hypotheses.

We would like to conclude with several practical suggestions that are based on our experience:

- When looking at the results of microarray experiments, do not get intimidated by the wealth of information. Remember the gene funnel. Define a set of "legal genes" and "active genes" and look at them. Query the data. This will give you a preliminary sense of the data.
- Do not get hooked on fold ratios.
- It is important to collaborate with computational scientists. However, do not leave the analysis completely to them. Most of the tools mentioned are relatively easy to use. Being comfortable with using the tools will allow you to become familiar with the results and to determine what you really need from your collaborators.

Figure 3. Overabundance graph of the significant genes that distinguish lung fibroblasts from human airway epithelial cells. *Red*, observed genes; *green*, expected number under the null-hypothesis (random labels). The *x*-axis denotes P value and the *y*-axis the number of genes. The expected number of genes is the P value multiplied by the number of genes in the data set. Experiments were previously described by us (23).

- Remember that the most productive groups in microarray research function as multidisciplinary teams. Make sure that your computational collaborators understand your experimental system. Involve them early on in the design of the experiments. Do not allow the analysis to become a service.
- Repeat your experiments. If you cannot afford to run repeats for every experimental point, do pools; they represent the data better than a single arbitrary sample. If you are using RNA pooled from several repeats of an experiment for your microarray experiment, make sure that you also keep aliquots of RNA from the individual experiments. This way you will still be able to verify the results using other methods on individual samples.
- Statistical considerations and analysis are important. In contrast to two years ago, there are currently many tools for statistical analysis of the results of microarray experiments. There is no justification to neglect the statistical analysis of the data.
- Use more than one analytic approach, scoring method, or clustering application on your dataset. This will help you to gain confidence in your observations.
- Do not rush to purchase commercial software solutions; there are many publicly available resources that provide visualization and analysis tools. Actually, many of the tools described here do not exist in any commercial package.
- Share your tools. If you developed an analytic approach or tool, share it. The main reason that hierarchical clustering is so popular is because it was freely available very early on. Furthermore, as many groups develop approaches that deal with specific aspects of data analysis, sharing of tools and source code would lead to improved tools.
- Share your data. One good thing about the results of microarray experiments is that there is enough for everybody. The full impact of the results can only be realized if they are freely available to the scientific community (naturally after publication). It is important to also share results of experiments that will not get published or negative results. The creation of repositories for results of large-scale gene expression experiments will of course facilitate this process.

References

1. Schena, M., D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis. 1996. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA* 93:10614–10619.
2. Tavazoie, S., J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. 1999. Systematic determination of genetic network architecture. *Nat. Genet.* 22:281–285.
3. Lockhart, D. J., and E. A. Winzeler. 2000. Genomics, gene expression and DNA arrays. *Nature* 405:827–836.
4. Miki, R., K. Kadota, H. Bono, Y. Mizuno, Y. Tomaru, P. Carninci, M. Itoh, K. Shibata, J. Kawai, H. Konno, S. Watanabe, K. Sato, Y. Tokusumi, N. Kikuchi, Y. Ishii, Y. Hamaguchi, I. Nishizuka, H. Goto, H. Nitanda, S. Satomi, A. Yoshiki, M. Kusakabe, J. L. DeRisi, M. B. Eisen, V. R. Iyer, P. O. Brown, M. Muramatsu, H. Shimada, Y. Okazaki, and Y. Hayashizaki. 2001. Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. *Proc. Natl. Acad. Sci. USA* 98:2199–2204.
5. Chu, S., J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz. 1998. The transcriptional program of sporulation in budding yeast. *Science* 282:699–705.
6. Whitfield, M. L., G. Sherlock, A. Saldanha, J. I. Murray, C. A. Ball, K. E. Alexander, J. C. Matese, C. M. Perou, M. M. Hurt, P. O. Brown, and D. Botstein. 2002. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* (In press, published online ahead of print March 21, 2002.)
7. van't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415:530–536.
8. Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, Jr., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403:503–511.
9. Boldrick, J. C., A. A. Alizadeh, M. Diehn, S. Dudoit, C. L. Liu, C. E. Belcher, D. Botstein, L. M. Staudt, P. O. Brown, and D. A. Relman. 2002. Stereotyped and specific gene expression programs in human innate immune responses to bacteria. *Proc. Natl. Acad. Sci. USA* 99:972–977.
10. Quackenbush, J. 2001. Computational analysis of microarray data. *Nat. Rev. Genet.* 2:418–427.
11. Miles, M. F. 2001. Microarrays: lost in a storm of data? *Nat. Rev. Neurosci.* 2:441–443.
12. Friedman, N., and N. Kaminski. 2002. Statistical methods for analyzing gene expression data for cancer research. Ernst Schering Research Foundation Workshop 38: Bioinformatic and Genome Analysis. *In* H. W. Mewes, H. Seidel, and B. Weiss, editors. Springer Verlag. 109–113.
13. Ben-Dor, A., L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. 2000. Tissue classification with gene expression profiles. *J. Comput. Biol.* 7:559–583.
14. Affymetrix. 2002. Statistical Algorithms Reference Guide. http://www.affymetrix.com/support/technical/technotes/statistical_reference_guide.pdf.
15. Kaminski, N., J. D. Allard, J. F. Pittet, F. Zuo, M. J. Griffiths, D. Morris, X. Huang, D. Sheppard, and R. A. Heller. 2000. Global analysis of gene expression in pulmonary fibrosis reveals distinct programs regulating lung inflammation and fibrosis. *Proc. Natl. Acad. Sci. USA* 97:1778–1783.
16. Kannan, K., N. Amariglio, G. Rechavi, J. Jakob-Hirsch, I. Kela, N. Kaminski, G. Getz, E. Domany, and D. Givol. 2001. DNA microarrays identification of primary and secondary target genes regulated by p53. *Oncogene* 20:2225–2234.
17. Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95:14863–14868.
18. Bittner, M., P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, and V. Sondak. 2000. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406:536–540.
19. Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 96:6745–6750.
20. Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96:2907–2912.
21. Ben-Dor, A., R. Shamir, and Z. Yakhini. 1999. Clustering gene expression patterns. *J. Comput. Biol.* 6:281–297.
22. Sharan, R., and R. Shamir. 2000. CLICK: A clustering algorithm with applications to gene expression analysis. Proc. 8th International Conference on Intelligent Systems for Molecular Biology (ISMB '00). AAAI Press, Menlo Park, CA. 307–316.[/conf]
23. Lee, J. H., N. Kaminski, G. Dolganov, G. Grunig, L. Koth, C. Solomon, D. J. Erle, and D. Sheppard. 2001. Interleukin-13 induces dramatically different transcriptional programs in three human airway cell types. *Am. J. Respir. Cell Mol. Biol.* 25:474–485.
24. Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531–537.
25. Baldi, P., and A. D. Long. 2001. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17:509–519.
26. Tusher, V. G., R. Tibshirani, and G. Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* 98:5116–5121.
27. Hedenfalk, I., D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon,

P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi, B. Wilfond, A. Borg, and J. Trent. 2001. Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* 344:539–548.

28. Zuo, F., N. Kaminski, E. Eugui, J. Allard, Z. Yakhini, A. Ben-Dor, L. Lollini, D. Morris, Y. Kim, B. DeLustro, D. Sheppard, A. Pardo, M. Selman, and R. A. Heller. 2002. Gene expression analysis reveals matrilysin as a key regulator of pulmonary fibrosis in mice and humans. *Proc. Natl. Acad. Sci. USA* 99:6292–6297.

29. Durrett, R. editor. 1991. Probability theory and examples. Wadsworth and Brooks, Cole, CA.

30. Benjamini, Y., and Y. Hochberg. 1995. Controling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 57:289–300.

31. Segal, E., B. Taskar, A. Gasch, N. Friedman, and D. Koller. 2001. Rich probabilistic models for gene expression. *Bioinformatics* 17:S243–S252.

32. Dahlquist, K. D., N. Salomonis, K. Vranizan, S. C. Lawlor, and B. R. Conklin. 2002. GenMAPP: a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.* 31:19–20.

33. Cojocaru, G., N. Friedman, M. Krupsky, P. Yaron, D. Simansky, A. Yellin, G. Rechavi, Y. Barash, A. Ben-Dor, Z. Yakhini, and N. Kaminski. 2002. Transcriptional profiling of non-small cell lung cancer using oligonucleotide microarrays. *Chest* 121:44S.