Analysis of expression patterns: The scope of the problem, the problem of scope

Yidong Chen[a], Zohar Yakhini[b], Amir Ben-Dor[b], Edward Dougherty[c], Jeffrey M. Trent[a] and Michael Bittner[a,*]

[a]*Cancer Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, MD 20892, USA*

[b]*Chemical and Biological Systems Department, Agilent Laboratories, Palo Alto, CA 94304, USA*

[c]*Department of Electrical Engineering, Texas A & M University, College Station, TX 77843, USA*

Studies of the expression patterns of many genes simultaneously lead to the observation that even in closely related pathologies, there are numerous genes that are differentially expressed in consistent patterns correlated to each sample type. The early uses of the enabling technology, microarrays, was focused on gathering mechanistic biological insights. The early findings now pose another clear challenge, finding ways to effectively use this kind of information to develop diagnostics.

1. Introduction

The profiling of transcription patterns is a central and long-standing tool in molecular biology. Recently, it has become possible to gather this kind of data for many genes simultaneously, allowing a wider view of the transcriptional activity of a particular cell type or tissue [1,2]. The fundamental assumption concerning the utility of such transcript abundance profiles is that transcription profiles convey information about the processes operating in a cell of a given type or state. The view obtained is limited to those parts of operations directly influencing the transcriptional activity of the cell. Still, it is obvious that even a clear and complete listing of the variances in transcription between the operations of a healthy cell and those of a diseased cell of the same type would provide useful information. At a minimum, those differences that are most extreme could provide useful markers to differentiate the diseased cells from the healthy ones in routine diagnostic testing. In some cases, this information could conceivably be much more useful. In the field of oncology, our particular area of interest, the variances might point to some difference that could be exploited to kill or terminally differentiate the cancerous cells through some form of treatment. These tantalizing possibilities have led many researchers to pursue the development of methods for gathering and analyzing expression data.

The experiences gained through these early efforts have begun to outline useful approaches to the collecting and analyzing expression profile data and to identify the most serious obstacles to realizing the desired benefits. This review will summarize some of the strengths of data viewing and analysis approaches used to date, and sketch some of the approaches and limitations to the development of more powerful forms of analysis.

2. Methods of analysis

2.1. Clustering/correlation

As might be expected, the first attempts to harness the information provided in profiles centered on the most readily detected forms of relationships in these kinds of observations, simple correlation of similarities [3]. The typical way to view similarities is to perform a clustering operation. This type of comparison is meant as a very preliminary way to look at data, a way of discovering trends. The methods do not offer any estimate of the way in which the variance arising from the system biology or the measurement procedure will affect the reproducibility of the resulting groupings. Additionally, the similarity measurement steps used frequently incorporate averaging and normalization steps that make it impossible to compare the resulting sets of similarity measurements in a quantitative fashion.

*Address for correspondence: Michael Bittner, NHGRI/NIH, Building 49, Room 4A52, 9000 Rockville Pike, Bethesda, MD 20850, USA. Tel.: +1 301 496 7980; Fax: +1 301 402 3241; E-mail: mbittner@nhgri.nih.gov.

Disease Markers 17 (2001) 59–65
ISSN 0278-0240 / $8.00

In spite of these drawbacks, the use of correlation and clustering has a rich history in mathematics and engineering, and a large and growing number of the extant approaches to clustering data are being examined in the context of expression profile analysis. They provide a very quick way to see the most evident relationships, and when supplemented with other forms of analysis or connections to prior knowledge can help identify differences in expression worth further consideration.

2.2. Gene by gene correlation

At the start of efforts to gather transcription profiles, there were two simple expectations for clustering results, both based on historical knowledge of transcriptional regulation. The first was that genes responding to a given type of signal, such as a fundamental change in metabolism, would be identifiable in data from a sample series that spanned such a transition, due to the similarity of their transcriptional response to the signal. This was quickly demonstrated in yeast undergoing a diauxic shift [4]. Many subsequent experiments in a variety of systems have demonstrated that correlating similarities of response can be a very powerful way of grouping genes involved in processes such as serum response [5], or cell cycling [6,7]. Very complex data sets containing mixtures of developmental time-course and mutant analysis series have been shown to allow clustering along functional lines [8].

A further power of this form of correlation is its power to group genes whose expression is driven by the same kinds of cis-regulatory sites. Demonstrations of the clustering of genes sharing known or new cis-regulatory sites have been presented in several studies of transcription dynamics associated with the yeast cell cycle [6,9]. While these results provide a demonstration of the competence of clustering to disclose explicit co-regulation, both these studies, and one examining transcription during meiosis [10] suggest cautious expectations for the extent of utility of clustering in disclosing cis-regulatory sharing. Even in yeast, it appears to be difficult to identify a large fraction of the motifs that must be in place. Clustering paired with promoter sequence searching is most useful for those elements that have the least ambiguity and the greatest length. This signal to noise constraint makes the cluster and search strategy likely to be most useful in identifying genes sharing already discovered cis-regulatory elements in multi-cellular organisms. This is due to their large genome size, their typically short and variable cis-regulatory sequences and their highly combinatorial use of these regulators (extensively reviewed by Davidson in [11]).

The results of a number of systematic informatics efforts can also be expected to enhance the kinds of insight that correlation can provide. Various efforts to associate the currently available knowledge of genes' functions and relationships with other genes are under development. One approach that has already reached a fairly high level of sophistication is the production of gene ontologies [12]. This is a large, collaborative effort that seeks to provide curated information about the biological role of genes in many different organisms using a unitary set of classifiers (controlled vocabulary). As the human gene ontology becomes more complete, it can be used to provide a summary view of the various biological activities represented in a particular cluster. Known genes in other species could supplement this process by suggesting that a human EST of unknown function that is related by sequence to the known gene may have a function that makes sense in the context of the cluster in which it falls. Similar aid in deciphering function may become available from a less supervised, indexing approach to evaluating gene functions. One example of this type of approach is the High-density Array Pattern Interpreter program (http://array.ucsd.edu/tools.htm). This program uses controlled terminology hierarchies, based on the National Library of Medicine's Medical Subject Headings, to delineate how genes have been described in the published literature. In general, any further characterization that can be associated with a gene is likely to improve the odds of estimating whether a gene with a biologically interesting expression pattern should be further studied as a candidate marker or target.

2.3. Sample by sample correlation

The second a priori expectation for expression patterns was that it would be possible to discern differing types of healthy and pathological cells by considering the overall profile of similarities and differences across many genes. There have been many demonstrations of the use of overall transcription similarity measures to group samples into their previously known classes [13–15]. There have also been demonstrations of the practicality of using this strategy to discover classes within samples that could not be subdivided in this fashion by conventional measurements [16–18]. Surveying the uses of this strategy, it has become apparent that there are both differing biological bases driving the separa-

tions and a wide spectrum in the number of genes that strongly contribute to the separation.

One clear component of the separation between cell types, a "tissue of origin" component arises from the particular differentiation state of the cells being studied. It is easy to imagine that vastly different cell types, such as muscle versus neuron, would show considerable differences in their expression of specialized gene products and be readily classified based on those differences. It was less easy to predict that even lymphomas arising from close relatives in the B-cell lineage have a large number of transcriptional differences that can be easily exploited for classification [16]. Some other ways in which surprising degrees of variation have been seen are evident in a study of tumor material from breast cancer [17]. A very clear finding of this study was the distinctiveness of individual tumors. Primary tumors and their metastases were found to have the highest degree of similarity in this study, which encompassed a variety of breast tumor types. That a tumor growing at a distant site and time than its primary is much more similar to its primary than to another metastasis arising from the same type of tissue implies that there is a large "space" of transcriptional settings available to a developing tumor. It also implies that the "position" the tumor occupies in that space is well separated from that chosen by other tumors of similar origin.

As the volume of expression space inhabited by cells of different types does have a significant impact on the ease, resolution and reliability of discrimination between diseases that can be achieved via expression profiling, it is worth having a look at the amount and magnitude of variance there is between samples. It is difficult to get a feel for this from the processed data usually presented in papers, since the most common representation is the use of a color scale, which tends to understate the differences. A much different intuition is conveyed by a scatter plot comparing the average intensities of the two fluorescent cDNA probes hybridized at each immobilized detector on a cDNA array. Three such plots are presented in Fig. 1. Panels A and B show that when an mRNA pool derived from either a melanoma cell line or a myeloid cell line is profiled against itself, there is very little variance in the intensity of any of the genes detected. Panel C shows the contrasting result of widespread variation, ranging from minor to major differences, when these different mRNA pools are compared to each other. This is not an exceptional result; it is the typical outcome. It appears that every cell state has unique settings for the expression levels of most of the genes expressed in the cell, including both the ubiquitously expressed genes and the genes specific to particular states. This level of idiosyncracy of expression differences provides both vast opportunities and complications to the identification of useful disease markers. The likelihood of finding a small set of accurate, discriminative markers that could be pressed into service in a traditional format such as immunohistochemistry increases with the number of differentially expressed genes available, however one must be ready to sift many candidates for consistency and tolerance to noise.

2.4. Simple gene ranking methods

When looking at the numerous correlates between genes and samples that result from a large array study one is frequently overwhelmed by the many possibly interesting genes that would make sense to study, based on the relevance of their known biological activities to the system being explored. To further reduce the number of candidates, one can choose a variety of filters based on the pattern of gene expression in the various sample types. Two simple tools have been described for carrying out such analysis. One, the "Weighted-List" method, is based on estimating the compactness within sample groups and the separation between sample groups that a given gene or genes' expression values would produce between the sample types [15,18]. This approach is conceptually related to the standard statistical F and t-tests. A geometric interpretation of the weight value produced by this analysis is presented in Fig. 2. The other method, Threshold Number of Misclassification (TNoM), is based on finding the minimal error rate of separation for ranked samples. The samples are ranked according to the expression values a gene produces, and then separated at a point that produces the least misclassifications [18–20].

Both of these methods are based on the supposition that those genes whose action is of strong consequence to the biological differences between the samples will accurately separate them. The Weighted List method emphasizes the average extent to which expression values separate the samples, making it a useful tool for finding genes with relatively large shifts in expression between sample types. The TNoM method emphasizes the integrity of the separation, making it useful for finding genes that accurately separate the samples but do not have as large ratio shifts. Genes that produce particularly clear discriminations achieve high scores with both measurements. Both methods of analysis are tested by forming permuted sample groups of the

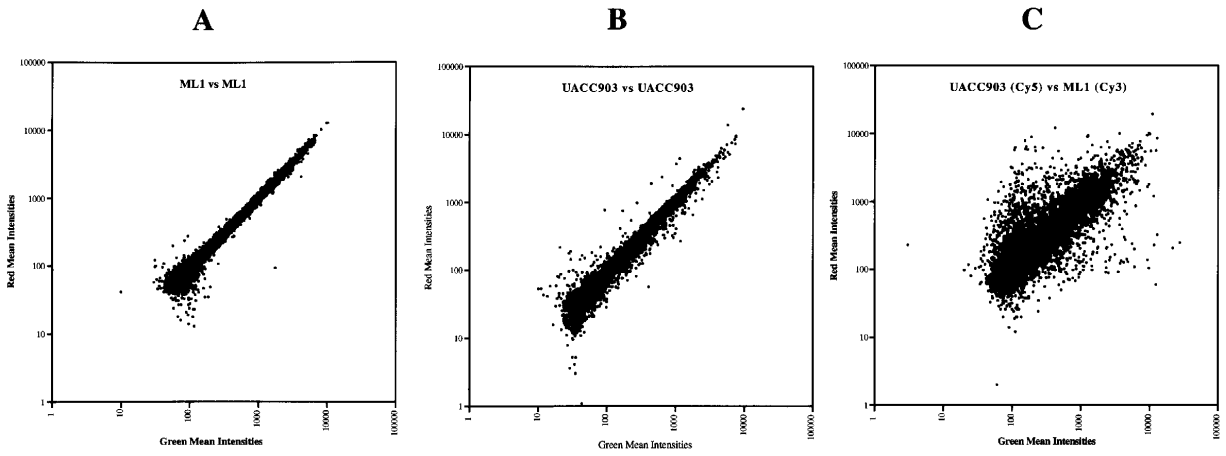

Fig. 1. Scatter-plots of average channel intensity per gene. The average red (y-axis) and green (x-axis) intensities at each immobilized gene detector element on an array of approximately 7000 genes is plotted. A) RNA from cell line ML1 used for both channels. B) RNA from cell line UACC903 used for both channels. C) RNA from UACC 903 used for red channel, RNA from cell line ML1 used for green channel. (From [22]).

same size as the authentic sample sets, but with randomized membership. Running and scoring a thousand such permuted sets provides an empirical estimate of the highest expected weight or TNoM value in a random collection of biological samples, providing a useful estimate of the lower limit on values that are significant. Figure 3 is a diagram showing this kind of analysis applied to a fairly homogeneous subset of 19 melanomas versus 12 melanomas having much greater diversity in their expression profiles [17]. The black line depicts the actual number of genes able to separate the samples with the indicated level of accuracy (on the x-axis). The gray line depicts the expected number of such separating genes when 19 samples are chosen uniformly at random and designated as a class. The error bars indicate the 95% confidence interval for these numbers, under the same stochastic model. The difference between the authentic sample curve and the permuted sample/theoretical curve shows that there are many genes whose expression pattern aligns with the sample sets in a very non-random way. A similar differential is seen with the Weighted List results. Sharp overabundance of informative or highly separating genes is also observed in other studies such as [14–16].

Other approaches to finding highly discriminative genes include ones where methods similar or identical to those used in formal statistical sample classification are explored. Examples include studies of differences in gene expression between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) [14], and between breast tumors arising in patients with or without the breast cancer predisposing mutations BRCA1 or BRCA2. The methods utilized in these studies explored the results of allowing larger numbers of genes to participate to varying extents in a decision about what class a given sample was in. Such studies provide another way of probing the robustness of the differentiation in expression patterns. An estimate of the consistency and robustness of the differentiation is achieved by serially building the decision function using all of the samples save one, for all of the N-1 sample sets, a process known as leave-one-out-cross-validation. The results may be further queried to determine whether a significant fraction of identical deciding genes are employed in all sets and whether the classification is approximately equally accurate in all cases. Forming permuted sample groups with randomized membership and reconstituting and re-scoring a classifier, as above, allows estimation of the significance of the achieved classification.

These and many other approaches to finding discriminating genes for further study in mechanistic or diagnostic settings are in the early phases of development. A more refined sense of their practical utility will emerge as experimental determination of the importance of the high scoring genes to the phenotypic differences between the sample sets is carried out.

2.5. Classification

The ability to employ microarray methodology to carry out formal diagnostic classification of tissue is a reasonable long-term goal, given the demonstrated ability of the method to discern differences in the patterns of gene expression between normal and healthy

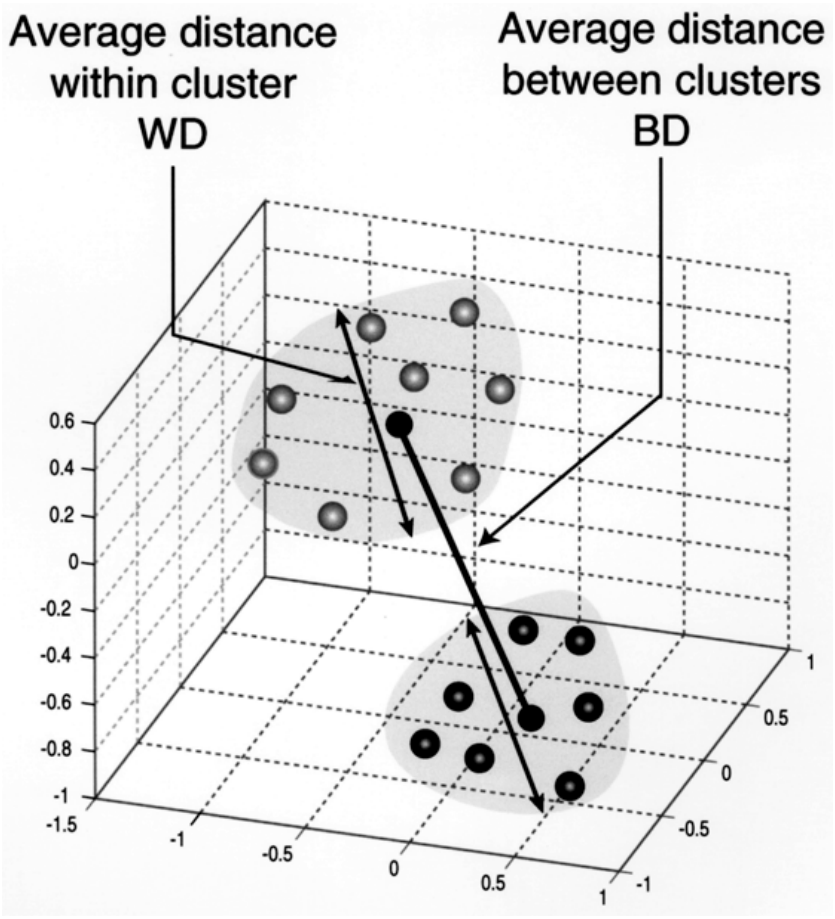

Fig. 2. Weighted Discriminator Method. Assuming K categories (or clusters) for a set of samples, a discriminative weight for each gene can be evaluated by w = Average(BD) / (Average(WD) + a) where Average(BD) is the average of the between cluster Euclidean distance for all pairs of clusters (total of (K*K-1)/2 pairs), and Average(WD) is the weighted average of the within-cluster distance (weighted by the number of samples in the cluster). The within-cluster distance is the average distance of all pairs of samples in the cluster. a is a small constant to prevent zero denominator case. (See http://www.nhgri.nih.gov/DIR/Microarray/discriminative.html).

tissue, and between differing types of diseased tissue. At the present there appear to be two main obstacles to making profiling a sufficiently practical form of diagnostic to find wide use. The first difficulty is an analytical one. How can very good candidate diagnostic panels be rapidly developed from profile data? The ideal panel would be one that used a very small number of genes, each of which provided at least some unique information (i.e. information that was not equivalent to the contribution of the other genes) and which was relatively insensitive to the levels of biological variance and measurement noise routinely encountered.

The problems associated with finding small classifier gene sets that meet robustness and uniqueness criteria, using expression-profiling data, have been concisely reviewed by Dougherty [21]. The general basis of the problem is that expression studies tend to be carried out as surveys aimed at developing insight into the biological mechanics of pathology. In studies of human tissue the goal has been to sample the broadest number of genes possible with the limited number of tissue samples and microarrays available. A consequence of this strategy has been that in most cases, there are neither sufficient numbers of samples nor sufficient numbers of replications of data sets to get good estimates of the error rates over the general population of the various genes in classification. Given small sample sets and large numbers of genes being sampled, it becomes possible to identify many small sets of genes for which the estimated error of classification is zero. In many cases, this estimate will not be markedly improved by small-sample-number validation procedures, such as leave-one-out cross-validation. As was mentioned in the Gene Ranking section above, there are

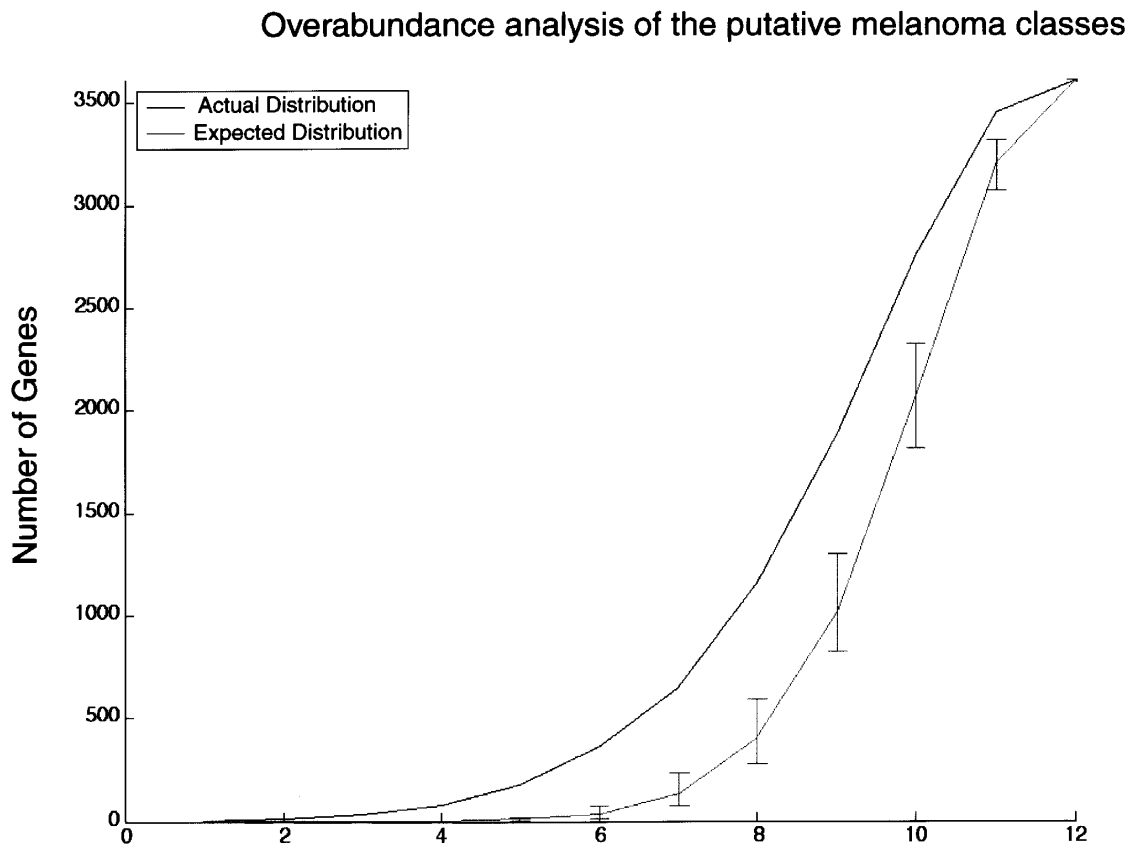

Fig. 3. Threshold Number of Misclassifications data from melanoma study [18]. The black line shows the number of genes in the original data set capable of producing the given number of misclassifications. The gray line is the result if the samples in the sets containing 19 and 12 members are permuted. Error bars show the calculated 95% confidence interval for the same size data set if gene expression behavior is independent and random relative to the samples.

many genes whose differences in expression pattern are aligned with differences in sample type in a very simple way, being more highly expressed in one sample type than the other. This produces a considerable overlap of the information content in relation to sample type in these sets of genes with the attendant problem that combining these genes in a classifier can easily lead to decreased performance via increased noise. An urgent need is therefore some readily computable analysis of the data that will help identify the most noise-resistant and least redundant classifier gene sets.

In addition to the problems of designing a classifier and choosing the genes that will provide the highest accuracy in the classifier, there are pragmatic problems that further complicate the use of expression profiling as a diagnostic. The primary analyte in the technique is mRNA, which is much less stable than DNA or protein, placing considerable constraints on sample collection. The methods of converting the mRNA into a species that can be detected and scored is very sensitive to the integrity of the mRNA and to contaminants that co-purify with the mRNA during sample preparation. The technique, as now practiced, requires significantly more cells than many diagnostics, and could be confounded by the presence or variable content of cells other than disease cells in the sample. Were the technology to be used in diagnosis, its focus would need to be shifted from breadth of examination toward precision. The starting point for practical diagnosis would be a small, specialized set of genes, not as many genes as possible, with sufficient replicates of this set to provide the required degree of measurement precision.

3. Conclusion

Expression profiles can be seen to provide a rich source of data on the differential expression of genes

between cell states. Early results have demonstrated that it is possible to find many genes that exhibit state-dependent patterns of expression, even between closely related pathologies. The expression studies carried out to date have been of sufficiently limited scope to provide the large amounts of data needed for confident design of classifiers based on expression data, however even with limited data the trends are encouraging. Technologic improvements will continue to increase the precision and reproducibility of measurement that can be achieved. Larger studies designed to support the development of disease markers will no doubt be undertaken. In the shorter term, a good analytic method for identifying robust candidate classifier gene panels based on smaller sample number could be developed. With such a tool, it may be possible to use limited information to construct immunohistochemical assays, usable within the sphere of current diagnostic practice.

References

[1] M. Schena, D. Shalon, R.W. Davis and P.O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science* **270**(5235) (1995), 467–470.

[2] D.J. Lockhart, H. Dong AND M.C. Byrne et al., Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nat Biotechnol* **14**(13) (1996), 1675–1680.

[3] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci USA* **95**(25) (1998), 14863–14868.

[4] J.L. DeRisi, V.R. Iyer and P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science* **278**(5338) (1997), 680–686.

[5] V.R. Iyer, M.B. Eisen and D.T. Ross et al., The transcriptional program in the response of human fibroblasts to serum, *Science* **283**(5398) (1999), 83–87.

[6] P.T. Spellman, G. Sherlock and M.Q. Zhang et al., Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Mol Biol Cell* **9**(12) (1998), 3273–3297.

[7] P. Tamayo, D. Slonim and J. Mesirov et al., Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, *Proc Natl Acad Sci USA* **96**(6) (1999), 2907–2912.

[8] A. Ben-Dor, R. Shamir and Z. Yakhini, Clustering gene expression patterns, *J Comput Biol* **6**(3–4) (1999), 281–297.

[9] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho and G.M. Church, Systematic determination of genetic network architecture, *Nat Genet* **22**(3) (1999), 281–285.

[10] M. Primig, R.M. Williams and E.A. Winzeler et al., The core meiotic transcriptome in budding yeasts, *Nat Genet* **26**(4) (2000), 415–423.

[11] E.H. Davidson, *Genomic Regulatory Systems in Development and Evolution,* Academic Press, London, 2001.

[12] M. Ashburner, C.A. Ball and J.A. Blake et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet* **25**(1) (2000), 25–29.

[13] J. Khan, R. Simon and M. Bittner et al., Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays, *Cancer Res* **58**(22) (1998), 5009–5013.

[14] T.R. Golub, D.K. Slonim and P. Tamayo et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286**(5439) (1999), 531–537.

[15] I. Hedenfalk, D. Duggan and Y. Chen et al., Gene-expression profiles in hereditary breast cancer, *N Engl J Med* **344**(8) (2001), 539–548.

[16] A.A. Alizadeh, M.B. Eisen and R.E. Davis et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature* **403**(6769) (2000), 503–511.

[17] C.M. Perou, T. Sorlie and M.B. Eisen et al., Molecular portraits of human breast tumours, *Nature* **406**(6797) (2000), 747–752.

[18] M. Bittner, P. Meltzer and Y. Chen et al., Molecular classification of cutaneous malignant melanoma by gene expression profiling, *Nature* **406**(6795) (2000), 536–540.

[19] A. Ben-Dor, L. Bruhn and N. Friedman et al., Tissue classification with gene expression profiles, *J Comput Biol* **7**(3–4) (2000), 559–583.

[20] A. Ben-Dor, N. Friedman and Z. Yakhini, *Scoring genes for relevance,* Palo Alto, Agilent Technologies, 1999.

[21] E.R. Dougherty, Small sample issues for microarray-based classification, *Comp Funct Genom* **2** (2001), 28–34.

[22] Y. Jiang, J. Lueders, A. Glatfelter, C. Gooden and M. Bittner, in: *Profiling human gene expression with cDNA microarrays. Current Protocols in Human Genetics,* N. Dracopoli, ed., John Wiley & Sons, New York, 2000.