From patterns to pathways: gene expression data analysis comes of age

Donna K. Slonim

doi:10.1038/ng1033

Many different biological questions are routinely studied using transcriptional profiling on microarrays. A wide range of approaches are available for gleaning insights from the data obtained from such experiments. The appropriate choice of data-analysis technique depends both on the data and on the goals of the experiment. This review summarizes some of the common themes in microarray data analysis, including detection of differential expression, clustering, and predicting sample characteristics. Several approaches to each problem, and their relative merits, are discussed and key areas for additional research highlighted.

Advances in the molecular understanding of disease have already had widespread practical applications. Pre-genomic molecular biology produced diagnostics such as prostate specific antigen screening for prostate cancer and drugs such as the protein kinase inibitor Gleevec and the monoclonal antibody Herceptin, respectively effective in well-defined subsets of leukemia and breast cancer patients. The advent of microarray technologies for large-scale transcriptional profiling has fueled hopes that similar advances might become commonplace, leading to new methods of diagnosis and treatment for any number of diseases. The response has been profound, with researchers, clinicians and companies rushing to embrace the new techniques.

As more and more researchers jump on the microarray bandwagon, however, it has become increasingly clear that simply generating the data is not enough; one must be able to extract from it meaningful information about the system being studied. Despite the combined efforts of biologists, computer scientists, statisticians and software engineers, there is no one-size-fits-all solution for the analysis and interpretation of genome-wide expression data. As transcriptional profiling has grown in popularity, statistical methods for interpreting the data have proliferated. The advantage of this growth is that a wealth of tools are now available for those hoping to sift valuable nuggets of knowledge from the widening river of data. However, there are now so many options available that choosing among them is challenging. An understanding of both the biology and the computational methods is essential for tackling the associated 'data mining' tasks without being distracted by the abundant fool's gold.

Detecting differential expression

The most basic question one can ask in a transcriptional profiling experiment is which genes' expression levels changed significantly. Answering this question involves many considerations. There may be two experimental conditions or many, the conditions may be independent or related to each other in some way (as in a time series), or there may be many different combinations of experimental variables. Replicates, if present at all, might be samples from different animals or repeated hybridizations of the same samples. Reflecting this variety, many different methods are commonly used for identifying significant changes.

Most of the earliest transcriptional profiling experiments measured differential expression by the ratio of expression levels between two samples. Genes with ratios above a fixed cut-off k (that is, those whose expression underwent a k-fold change) were said to be differentially expressed[1–4]. Costly replication of arrays was rare. As suggested elsewhere in this issue (see review by G. Churchill, pages 490–495)[5], replication is essential in experimental design because it allows an estimate of variability (see also ref. 6). The ability to assess such variability allows identification of biologically reproducible changes in gene expression levels. As researchers recognized this, experiments with replicates became more common. Yet many analyses still designated as differentially expressed those genes with expression ratios or 'fold-changes' above a fixed threshold in more than one of the replicates[7,8].

Li and Wong[9] introduced a more sophisticated fold-change approach to analyzing oligonucleotide array data. They first fit a model that accounts for random, array- and probe-specific noise, and then evaluated whether the 90% confidence interval for each gene's fold-change excludes 1.0. Unlike standard fold-change approaches, this method incorporates available information about variability in the gene-expression measurements. However, because the error model is fitted to the entire data set, it can suffer when the data set is either too small or too heterogeneous. Other model-based methods designed for two-color arrays[10,11] also incorporate data-derived estimates of variation.

More typically, researchers now rely on variants of common statistical tests. These generally involve two parts: calculating a test statistic and determining the significance of the observed statistic. A standard statistical test for detecting significant change between repeated measurements of a variable in two groups is the t-test; this can be generalized to multiple groups via the ANOVA F statistic (see, for example, ref. 12). Variations on the t-test statistic (often called 't-like tests') for microarray analysis are abundant[13–15]. The use of non-parametric rank-based statistics is also common, via both traditional statistical methods[16] and more *ad hoc* ones designed specifically for microarray data[17,18]. For most practical cases, computing a standard t or F statistic is appropriate, although referring to the t or F distributions to determine significance is often not, as discussed below. The main hazard in using such methods occurs when there are too few

Department of Genomics, Wyeth Research, 35 Cambridge Park Drive, Cambridge, Massachusetts 02140, USA (e-mail: dslonim@wyeth.com).

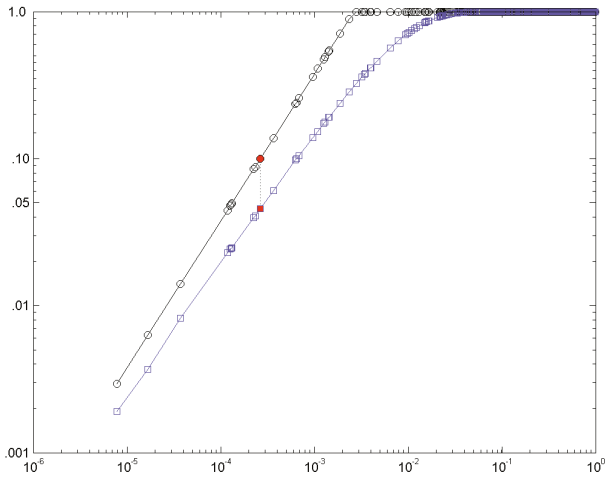

Fig. 1 The advantage of permutation-based adjustment for multiple testing. The data set (from ref. 97) contains expression data for 376 genes in 30 cancer cell lines selected for sensitivity or resistance to the compound cytochalasin D (see Supplementary information). The *x*-axis shows unadjusted *p*-values derived from independent *t* tests for each gene to detect differential expression between sensitive and resistant cell lines. The *y*-axis shows the adjusted *p*-values using Bonferroni correction (black circles) and Westfall and Young's permutation-based method[18,20] (blue squares). At the adjusted cutoff of 0.05, the permutation method finds 11 significantly changing genes (instead of 7 without permutation). For example, the gene shown in red (*COL6A2*) with unadjusted *p*-value of 0.00027 is adjusted to 0.1 by the Bonferroni method but 0.0457 using permutation, a difference of more than two-fold (dotted line).

replicates to obtain an accurate estimate of experimental variances. In such cases, modeling methods that use pooled variance estimates[9] may be helpful.

Regardless of the test statistic used, one must determine its significance. Standard interpretations of *t*-like tests assume that the data are sampled from normal populations with equal variances. Expression data may fail to satisfy either or both of these constraints. Although log transformation can improve normality (see review by J. Quackenbush, pages 496–501, this issue)[19] and help equalize variances, ultimately the best estimates of the data's distribution come from the data themselves. Permutation tests, generally carried out by repeatedly scrambling the samples' class labels and computing *t* statistics for all genes in the scrambled data, best capture the unknown structure of the data[13,14,20]. Such permutation tests are ideal when the number of arrays is sufficient to offer the desired degree of confidence.

One advantage of permutation methods is that they allow more reliable correction for multiple testing. The issue of multiple tests is crucial, as microarrays typically monitor the expression levels of thousands of genes. Standard Bonferroni correction (that is, multiplying the uncorrected *p*-value by the number of genes tested) is overly restrictive. Step-down methods designed to minimize this overcorrection[21] are little better for thousands of genes. Both methods are overly strict because they are based on the assumption that each gene represents an independent test. In fact, the correlation structure between gene-expression patterns is significant and complex. To capture this structure, Dudoit *et al.*[20] propose a permutation-based approximation of Westfall and Young's method[22], for which C code is available online (http://www.cbil.upenn.edu/tpWY). (A package of R functions for other techniques evaluated in ref. 20 is available at http://www.stat.berkeley.edu/users/terry/zarray/Software/smacode.html.) The advantage of permutation is apparent in Fig. 1.

All these approaches focus on determining the 'family-wise error rate,' the overall chance that at least one gene is incorrectly identified as differentially expressed. For microarray studies focusing on finding sets of predictive genes, it may instead be acceptable to bound the 'false discovery rate' (FDR), the probability that a given gene identified as differentially expressed is a false positive. A simple method for bounding the FDR is proposed by Benjamini and Hochberg[23]. While this, too, assumes that each gene is an independent test[20], a permutation-based approximation of this method is implemented in the SAM program by Tusher *et al.*[13], and a more permissive permutation-based approach to bounding the FDR appears in the Whitehead's GeneCluster software package[14]. Although in some data sets even the lowest FDR may be prohibitively high, this can be a valuable approach to finding some valid leads when more stringent analyses find none.

Time-series analysis. One common differential expression problem that has received relatively little specific attention until recently is time-series analysis. Entire books have been written on time-series analysis methods[24,25], and growing numbers of available data sets follow biological processes over time[1,2,26]. Yet so far, most have been analyzed by the statistical methods described above, perhaps supplemented by pattern-discovery techniques, without accounting for the known temporal relationships between samples. Exceptions include the use of time-series data for inference of regulatory pathways, as discussed below, and a small but growing number of papers that seek to exploit this knowledge more fully.

The canonical time-series data in the field come from two experiments following the yeast cell cycle[27,28]. Spellman's analysis incorporates a Fourier transform to test the periodicity of individual genes in three separate data sets, before combining these into a single significance score used to rank the genes. Later analyses of the same data sets[29,30] look at other time-warping or phase-shifting algorithms to test periodicity. Software for several of these is available online[29]. Evaluating or modifying time-series analysis methods for the microarray domain, particularly given the difficulty of taking sufficiently frequent array measurements to monitor many processes of interest, is an area ripe for additional attention. Also of interest is the suitability of such methods for analysis of samples related in other ways, such as cells exposed to different doses of a drug, or expression patterns from related bacterial strains.

Pattern discovery

Pattern discovery provides a high-level overview of a data set and may be the first analysis step in a study that ultimately involves other analytical methods. Such techniques include a variety of dimension-reduction methods such as singular value decomposition, as well as various 'clustering' techniques designed for finding groups within the data. What these methods have in common is that they simplify the data set, ideally in ways that impart additional information about its structure, and that they are considered 'unsupervised', meaning that the reduction is derived solely from the data rather than reflecting any previous knowledge or classification scheme.

Principal components analysis[31,32], singular value decomposition[33,34] and multidimensional scaling[35,36] are related dimension-reduction techniques that can be used for visualizing large data sets. Each of these approaches projects the data into a new space based on linear combinations of variables that retain a large amount of the original data's variation. These techniques rely on the idea that most of the data's variation can be explained by a smaller number of transformed variables. When this is true, a two- or three-dimensional visual representation of highly multidimensional data may provide valuable insight (Fig. 2*a*). However, much information may be lost, so the potential strength inherent in these methods is also their greatest peril. Recalling

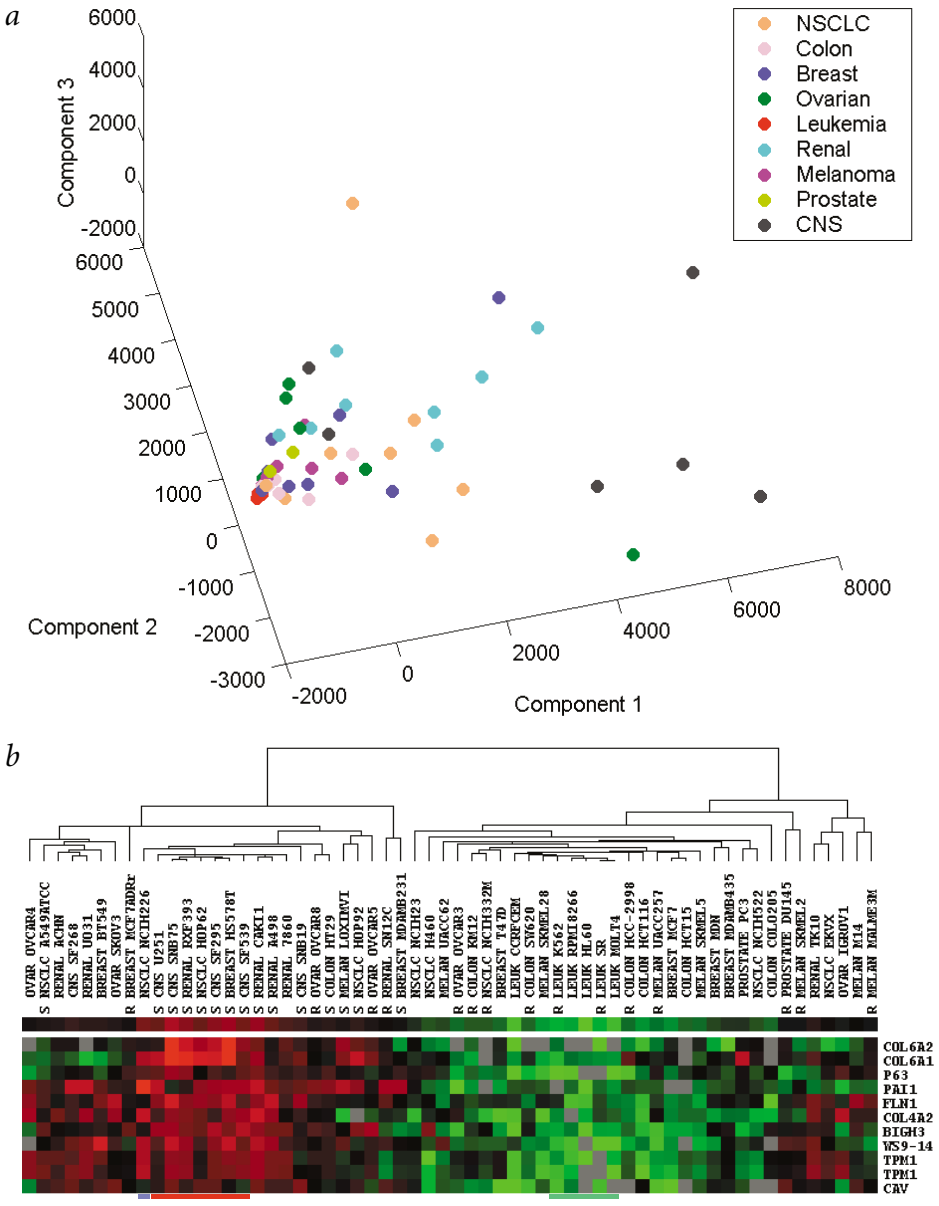

Fig. 2 Two pattern-discovery techniques. Data for both figures measure expression for 11 genes characterizing sensitivity to compound cytochalasin D in 60 cancer cell lines[97]. ***a***, The first three principal components, plotted using Matlab software (Mathworks). Apparent features include a tight cluster of leukemia samples (red dots, nearly superimposed) and the more scattered outlying cluster of CNS tumors (black dots). A single lung cancer sample (NSCLC-NCIH226) also appears as an outlier — the solitary orange dot at the top. ***b***, Hierarchical clustering of the same data, using Cluster/TreeView (http://rana.lbl.gov/EisenSoftware.htm). Names of samples extremely sensitive or resistant to cytochalasin D (see Supplementary information) are prefixed 'S' and 'R' respectively. The samples fall into two main clusters, roughly, but not perfectly, separating the sensitive and resistant samples. As in ***a***, fine structure shows a tight leukemia cluster (underlined in green) and a tight CNS cluster (underlined in red), but does not suggest that the CNS cluster or NSCLC-NCIH226 (underlined in blue) are outliers. Apparent in both ***a*** and ***b*** is the relative heterogeneity of the breast cancer cell lines.

that data-reduction and visualization tools are projecting many thousands of dimensions into two or three may prevent frustration if the reduced data fail to capture the expected aspects of a data set.

Clustering. The term 'clustering' applies to a wide variety of unsupervised methods for organizing multivariate data into groups with roughly similar patterns[37]. Clustering has many applications in expression-data analysis. Clues to unknown gene function may be inferred from clusters of genes similarly expressed across many samples[26,38]. Clustering samples over the expression levels of multiple genes has been proposed as a way of defining new disease subclasses[14,39]. Cluster analysis may be used primarily for data reduction and visualization, or it may be used to generalize or predict the categorization of new samples[40]. To solve any of these problems, researchers can choose from a vast library of techniques for grouping multivariate data.

Perhaps most familiar to biologists are the hierarchical clustering methods[26,28]. In this family of techniques, all data instances start in their own clusters, and the two clusters most closely related by some similarity metric are merged. The process of merging the two closest clusters is repeated until a single cluster remains. This arranges the data into a tree structure that can be broken into the desired number of clusters by cutting across the tree at a particular height. Tree structures are easily viewed and understood (Fig. 2*b*), and the hierarchical structure provides potentially useful information about the relationships between clusters. Trees are known to reveal close relationships very well. However, as later merges often depend on aggregated measures of clusters containing many scattered elements, the broadest clusters can sometimes be hard to interpret.

Another common family of clustering methods is that of partition or centroid algorithms. These methods generally require specification of the number, *k*, of clusters, and start with *k* data points that may be chosen either randomly or deliberately. These *k* points are used as the 'centroids' — the multidimensional center points — of an initial set of clusters. The algorithm then partitions the samples into the *k* clusters, optimizing some objective function (such as within-cluster similarity) by iteratively assigning samples to the nearest centroid's cluster and adjusting the centroids to represent the new clusters' center points. The *k*-means method[37] is a well-known centroid approach. A variation that allows samples to influence the location of neighboring clusters is known as the self-organizing map or Kohonen map[41,42]. Such maps are particularly valuable for describing the relationships between clusters. New centroid methods specifically for microarray data have also been proposed[43].

Other techniques abound. Some seek to optimize a measure of within-cluster similarity or separation between clusters, but avoid specifying the number of clusters ahead of time, instead specifying

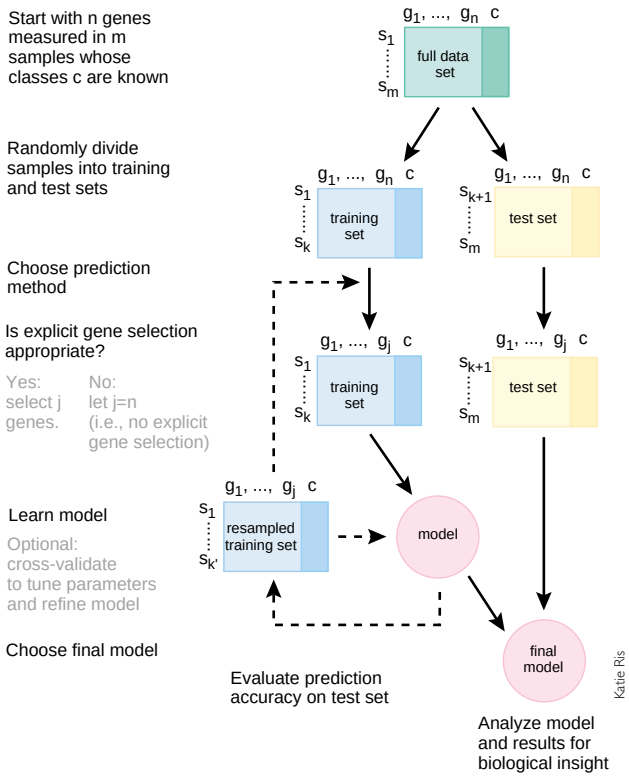

Fig. 3 An overview of the process for building a prediction model to classify samples. The partition into training and test data is ideally chosen at random across the entire set of samples. Many prediction methods require tuning some parameter (such as the number of genes, the number of nearest-neighbors to consider, or the number of decision trees built). This choice is often evaluated by cross-validation — the process of repeatedly removing smaller test sets from the training set, building new models (starting with the gene selection process) with the remaining data, and evaluating performance across all the different models built. For example, "leave-one-out cross validation" (also called "*n*-way") builds *n* models, each using *n*–1 training examples and evaluated on the remaining one; the accuracy for predicting all *n* samples is reported. Observing that predictors may succeed by chance even in cross-validation, Radmacher *et al.* suggest using permutation testing to determine the significance of the observed results[98]. Ultimately the final model, perhaps chosen during the cross-validation process, is then tested on entirely new data not used in the model generation process. The model itself, as well as the prediction results and the influential genes, may yield new biological insights.

information-theoretic bounds on cluster membership[44–46]. Model-based methods assume the data can be generated by a specified statistical model (such as a mixture of Gaussian distributions), and search for model parameters that best fit the data[47,48]. So-called 'fuzzy' clustering finds groups, but may allow elements to appear in more than one cluster or in no clusters at all[49].

Evaluating clusters. How does one choose among this panoply of techniques? Intuitively, a good clustering method should ensure that objects in the same cluster are similar to each other and different from objects in other clusters. Many cluster-evaluation metrics have been designed to formalize these intuitions[40,46]. There is, however, no single best way to evaluate a clustering method, and no single best clustering method for a data set. Different techniques often highlight different patterns in the data, so complementary methods may be helpful in analyzing a single data set (Fig. 2).

How best to compare clustering approaches depends on the purpose of clustering and on the information available. For example, if the 'true' or best clustering is known (as it might be with simulated data), a good metric might measure the fraction of co-clustered pairs from the true solution that are grouped together by the new method[40,46]. If clustering is to be used primarily for data reduction, one might evaluate it strictly from that point of view — the best clustering is the one that allows expression of the entire data set in minimal space. This generally requires making assumptions about data distributions and the data representation format. If clusters are to be used to predict classifications of other samples, one might choose to evaluate each clustering by its predictive power[50]. Another desirable property of a clustering is stability; that is, if the experiment were repeated again, one would hope to obtain similar clusters. A standard technique for testing cluster reliability involves adding a small amount of noise to the data and re-clustering. Several microarray studies have incorporated these techniques, either using simple but reasonable noise models[35,51], or by sampling the noise distribution directly from the data[52].

Choosing the right number of clusters is crucial for many hierarchical and partitioning algorithms. Although this problem has been addressed extensively by statisticians[53], it is particularly problematic for microarray data, which may be somewhat evenly distributed in gene expression space and thus may not have any solution featuring isolated and compact clusters. One approach to setting the desired number of clusters is based on the observation that good clusters will probably not change dramatically if a small randomly chosen subset of the samples is discarded. Ben-Hur *et al.* thus use repeated sampling to determine the number of clusters that provides the most stable solution[54]. Another approach, the 'gap statistic' of Tibshirani *et al.*, compares a measure of within-cluster distances from the cluster centroids to its expectation, and chooses the number of clusters maximizing the difference[55].

Overall, choosing a clustering method is still very much dependent on the data, the goals, and the investigator's personal preferences. One caveat, however, is to be sure that the primary goal is the pattern-discovery or dimension reduction that clustering offers. If the intent is to distinguish between currently defined sample classes (say, tumors and normal tissue samples), the class prediction methods described below may be more effective. While there are many exciting examples of clustering algorithms that happen to identify a desirable distinction between samples[39], there are probably many more (generally unpublished) examples in which clustering identified only some known artifact (for example, array production lot) or obvious characteristic of the patient (for example, age). Clustering can chance on the 'correct' sample classification only if the desired distinction is quite prominent among the sources of variation in the data set. So it is important to note that, while a clustering algorithm might fail to find a desired separation, the prediction methods described below may well succeed in defining the desired classes given the same data.

Class prediction

In contrast to pattern discovery, class prediction methods are techniques specifically designed to classify objects into known groups. A wealth of machine-learning literature[54] describes computational techniques for classifying multidimensional data. Most such methods include a training phase run on samples whose classes are already known, and a testing phase, in which the algorithm generalizes from the training data to predict classifications of previously unseen samples (Fig. 3). Because of this directed training phase, prediction methods are referred to as 'supervised' classification methods. For microarray data, prediction generally refers to the classification of patients' samples by characteristics such as disease subtype or response to treatment. The goal may be diagnostic, offering a new way to distinguish similar-looking diseases[14,56,57], or it may be a true effort to pre-

dict clinical outcome[39,58–60] (see also review by C.M. Perou, pages 533–540, this issue)[61].

Gene-expression data presents unusual challenges for machine-learning algorithms, which are generally designed to work with large numbers of samples over relatively few variables. In contrast, a typical microarray experiment measures thousands of genes (variables), but includes only tens or hundreds of patients' samples. Most algorithms stumble when faced with problems of these dimensions. A common problem is 'overfitting' the data, which refers to the case where an algorithm models the training samples too exactly, without sufficient generalization. In consequence, classification of the training examples may well be perfect, but subsequent attempts to classify new test data fail dismally. Compounding the challenge of learning from microarray data is the high level of noise. The variance of array measurements can be substantial, many data points may be missing entirely, and occasionally training examples may even be misclassified. All these traits make sample prediction from array data particularly challenging.

As with clustering, choosing a prediction method requires selecting from a vast range of techniques. Some of the most straightforward linear and quadratic discriminant methods are described clearly by Dudoit *et al.*[62]. Related methods include weighted voting[14], shrunken centroids[63] and compound covariates[64]. A deceptively simple but powerful approach is *k*-nearest-neighbor prediction, in which the prediction for a test sample *x* is the most common class label among the *k* training samples most similar to *x* (refs 58,62,65,66). Simple neural networks[66] may be effective at learning the complex functions often inherent in multiclass diagnostic problems[56]. New pattern-discovery algorithms such as Splash[67] have shown some success at learning non-linear functions of the input variables. Two other well-studied classes of algorithms are of growing interest for microarray prediction problems: support vector machines and decision tree classifiers.

Support vector machines (SVMs) are a family of statistical machine-learning methods that have been proposed as particularly suitable to the dimensions of microarray learning problems[68,69]. Intuitively, SVMs try to draw a hyperplane in *n*-dimensional gene-expression space between the training examples from two classes. If no separating hyperplane exists, the samples are mapped into a higher-dimensional space where such a separator does exist. The algorithms minimize potential overfitting problems by choosing the separator farthest from the training samples, thus leaving room for generalization. More complex mapping functions provide non-linear mappings into higher dimensional spaces, resulting in a non-linear classifier for the original data. While these models may be difficult to interpret, they are potentially quite powerful.

Decision tree algorithms classify samples by filtering them through a tree-like structure, testing at each branchpoint (called a 'node') some simple attribute of that sample, such as whether the expression of *p53* is greater than 100 (ref. 66). Single decision trees are particularly prone to overfitting. However, as tree models are easily built, easily understood, and able to model quite complex functions, there are many modified tree-based techniques for avoiding overfitting and improving performance. Solutions to overfitting include 'pruning' the tree; that is, restricting the number of consecutive branches so that it is forced to generalize. More powerful solutions are possible by repeatedly sampling the data to build many trees and combining these trees into a single predictive model using techniques known as 'bagging'[70] and 'boosting'[71,72]. Combined tree models may be harder to interpret than single trees, but standard approaches allow determination of which genes contributed most heavily to the models' predictive powers[73].

Choosing a prediction method. In deciding how best to approach a prediction problem, it is first important to consider the desired outcome. Are there just two classes to be distinguished, or many? Is it desirable to find the minimal number of predictive genes, in order to minimize the number of leads or to provide a simple diagnostic tool? Would it be better to have an easily interpretable model, which may help provide new medical insights, or is the only goal the greatest prediction accuracy possible? If the output will ultimately affect patients' treatment, it may be essential to have an accurate confidence estimate for each prediction. All of these issues can influence the choice of a prediction method.

There are few unbiased comparative studies of prediction methods for gene-expression analysis. Some appear in the supplementary information of published studies reporting particular biomedical results[58,74], and a thorough study by Dudoit *et al.* compares several standard statistical prediction methods on a diverse collection of public data sets[62]. Furthermore, most array data sets lack enough samples to prove a method clearly superior; generally, only a few errors separate the winners from the losers[69]. However, a few trends emerge. Simple *k*-nearest neighbors tends to perform well after a gene-filtering step[58,62]. Dudoit *et al.* also find that diagonal linear discriminant analysis does well overall, while CART (a decision tree algorithm prone to overfitting) and Fisher's linear discriminant lagged behind on most data sets evaluated. The aggregated tree models fell somewhere in between, but had the advantage of presenting relatively accurate confidence estimates. Pomeroy *et al.* found that *k*-nearest neighbors, weighted voting, and SVMs were all comparable, but a combination of several methods outperformed any single predictor[58]. Overfitting may be avoided by using simpler SVM mapping functions when complex ones are not needed[69,75], or by limiting the number of iterations of boosting algorithms[75]. There is some consensus that simpler methods outperform complex ones in the typical case where the number of genes is much larger than the number of samples[62,69,75].

The number of classes in the prediction problem may impose modest constraints on the choice of algorithm. Whereas neural networks, decision trees and *k*-nearest neighbors can, in principle, separate any number of output classes, SVMs and various linear methods are inherently binary — they only distinguish between two classes. It is, however, possible to combine binary predictors together to separate multiple classes[65,76]. A more important question is whether the prediction algorithm should consider the data for all available genes, or whether prior gene selection (often by the methods described above for detecting differential expression) should be used to reduce the dimensionality of the problem. For methods like *k*-nearest neighbors and weighted voting, which can be confused by too many irrelevant variables, preselection of genes may be appropriate[66]. However, there is evidence that SVM prediction improves given all the data[65]. Recent work confirms that pairs of genes selected for their combined ability to distinguish output classes predict better than genes selected individually[77]. If the genes used for prediction are to provide additional clues to the biological system under investigation, it may be particularly helpful to give the prediction method a chance to find these interesting gene interactions.

Inferring regulatory pathways and networks

Analysis of differential expression may provide new information about the biological pathways involved in a process. This is often done by looking for over-representation of functional classes in gene clusters derived from expression data[39]. Yet this approach relies heavily on existing functional annotation, which is notoriously incomplete for most organisms of interest. Looking specif-

ically for information on gene interactions indicated by expression data may ultimately suggest new pathways and associations. Even simple pairwise comparisons can indicate novel interactions[78]. However, the prediction results above imply that more complex gene relationships may be discovered as we learn to combine the data in more complex ways.

Although it is optimistic to assume that expression data alone will be sufficient for the inference of complete regulatory pathways, several recent studies successfully tackle parts of the problem. Studying the properties of synchronous, Boolean network models suggested new strategies for inferring regulatory networks from expression data[79].

Bayesian network models and variations are now the focus for a growing number of researchers concerned with discovering novel interactions, information dependencies and regulatory relationships from expression data. Whereas the posterior probabilities of all models are likely to be very low, repeated random resampling of the data (called 'bootstrapping') can help in identifying 'high-probability' gene relationships shared by a significant fraction of the models built from the different data subsets[79]. Methods for designing new experiments to discriminate among contradictory network models consistent with existing data have also been described[81].

More recent modifications of Bayesian network methods focus further on finding probabilistically supported gene interactions or on combining these into subnetworks[82,83], on modeling 'latent' or hidden variables representing biological information unavailable to the model[84,85], and on incorporating prior biological knowledge or annotation[82,86]. Most of these methods have been tested on previously published data, from which they rediscover some known relationships, propose revisions or contradictions of others, and suggest many novel interactions. Current pathway methods seem to do reasonably well in finding correlated sets of genes — genes that are co-expressed or are all targets of the same transcription factor. However, it has proved more difficult to infer the direction of causal relationships successfully directly from transcriptional data. In general, models that incorporate existing constraints from other data sources seem to produce hypotheses that agree better with existing biological knowledge than do models learned from the expression data alone[82].

Future directions

The basic problems described here are still areas of active research. The desire to find ever more reliable answers to harder problems demands more powerful statistical methods coupled with better understanding of the data. Future projects may focus on finding modestly sized sets of predictive genes, better characterizing the structure and predictive power of gene-expression data, or combining the knowledge gained from multiple clustering approaches. One goal of network inference methods is the simulation of cellular and systemic responses to interventions such as gene knockouts or drug treatment. While current approaches are just initial attempts to answer complex biological questions, the potential of such methods remains tantalizing.

The interpretation of microarray results remains a crucial issue. Because of the observational nature of many microarray studies, possible bias and confounding variables are substantial concerns[87]. Thus, transcriptional profiling is often used primarily to generate hypotheses for further testing by less or differently biased methods. Ultimately, the greatest contributions to understanding function will probably come from directly combining microarray data with other sources of genomic and biomedical information. Algorithms for integrating different types of data are already showing promise. Integrating clinical data from patients' records has been proposed as an approach to aiding interpretation[87,88], and preliminary approaches suggest ways of combining microarray data with other clinical or experimental variables[89,90]. Several groups have considered ways of directly combining expression data with analysis of gene regulatory motifs[91–93]. Furthermore, much of the evidence supporting or disproving hypotheses derived from microarray studies is found in the existing medical literature. Thus, systems to augment expression analysis with automated literature extraction or organization[94–96] are likely to prove extremely valuable in drawing meaningful and reproducible conclusions. By continuing efforts to boost the rigor and power of analyses and to integrate knowledge from complementary sources, we are progressing toward realizing the full potential of these powerful new technologies.

Acknowledgments

I thank Gene Brown, Lenore Cowen, Steve Haney, Andrew Hill, Steve Rozen and Timm Triplett for helpful discussions and comments.

1. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705 (1998).
2. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
3. Schena, M. *et al.* Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl Acad. Sci. USA* **93**, 10614–10619 (1996).
4. Wodicka, L., Dong, H., Mittmann, M., Ho, M.H. & Lockhart, D.J. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **15**, 1359–1367 (1997).
5. Churchill, G.A. Fundamentals of experimental design for cDNA microarrays. *Nature Genet.* **32**, 490–495 (2002).
6. Yang, Y.H. & Speed, T. Design issues for cDNA microarray experiments. *Nature Rev. Genet.* **3**, 579–588 (2002).
7. Fambrough, D., McClure, K., Kazlauskas, A. & Lander, E.S. Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather than independent, sets of genes. *Cell* **97**, 727–741 (1999).
8. Holstege, F.C. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717–728 (1998).
9. Li, C. & Hung Wong, W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* **2**, research0032 (2001).
10. Roberts, C.J. *et al.* Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873–880 (2000).
11. Ideker, T., Thorsson, V., Siegel, A.F. & Hood, L.E. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.* **7**, 805–817 (2000).
12. Zar, J.H. *Biostatistical Analysis*, 663 (Prentice-Hall, Upper Saddle River, NJ, 1999).
13. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5121 (2001).
14. Golub, T.R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
15. Model, F., Adorjan, P., Olek, A. & Piepenbrock, C. Feature selection for DNA methylation based cancer classification. *Bioinformatics* **17 Suppl 1**, S157–S164 (2001).
16. Zhan, F. *et al.* Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells. *Blood* **99**, 1745–1757 (2002).
17. Ben-Dor, A., Friedman, N. & Yakhini, Z. Scoring genes for relevance. Technical Report 2000-38 (Institute of Computer Science, Hebrew University, Jerusalem, 2000).
18. Park, P.J., Pagano, M. & Bonetti, M. A nonparametric scoring algorithm for identifying informative genes from microarray data. *Pac. Symp. Biocomput.* 52–63 (2001).
19. Quackenbush, J. Microarray data normalization and transformation. *Nature Genet.* **32**, 496–501 (2002).
20. Dudoit, S., Yang, Y.-H., Callow, M.J. & Speed, T.P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report 578 (Department of Statistics, University of California at Berkeley, Berkeley, CA, 2000).
21. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
22. Westfall, P.H. & Young, S.S. *Resampling-Based Multiple Testing*, 340 (John Wiley & Sons, New York, 1993).
23. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B* **57**, 289–300 (1995).
24. Chatfield, C. *The Analysis of Time Series: An Introduction* (5th ed.), 283 (Chapman & Hall, London, 1996).
25. Shumway, R.H. & Stoffer, D.S. *Time Series Analysis and Its Applications*, 560 (Springer Verlag, New York, 2000).
26. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
27. Cho, R.J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**, 65–73 (1998).
28. Spellman, P.T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
29. Aach, J. & Church, G.M. Aligning gene expression time series with time warping algorithms. *Bioinformatics* **17**, 495–508 (2001).

30. Filkov, V., Skiena, S. & Zhi, J. Analysis techniques for microarray time-series data. *J. Comput. Biol.* **9**, 317–330 (2002).
31. Raychaudhuri, S., Stuart, J.M. & Altman, R.B. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.* 455–466 (2000).
32. Landgrebe, J., Wurst, W. & Welzl, G. Permutation-validated principal components analysis of microarray data. *Genome Biol.* **3**, research0019 (2002).
33. Holter, N.S. *et al*. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl Acad. Sci. USA* **97**, 8409–8414 (2000).
34. Alter, O., Brown, P.O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA* **97**, 10101–10106 (2000).
35. Bittner, M. *et al*. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**, 536–540 (2000).
36. Khan, J. *et al*. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58**, 5009–5013 (1998).
37. Jain, A.K. & Dubes, R.C. *Algorithms for Clustering Data* (Prentice-Hall, Englewood Cliffs, NJ, 1988).
38. Wen, X. *et al*. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl Acad. Sci. USA* **95**, 334–339 (1998).
39. Alizadeh, A.A. *et al*. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
40. Yona, G. Methods for global organization of all known protein sequences. PhD. thesis (Institute of Computer Science, Hebrew University, Jerusalem, Israel, 1999).
41. Kohonen, T. *Self-Organizing Maps* (Springer, Berlin, 1997).
42. Tamayo, P. *et al*. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA* **96**, 2907–2912 (1999).
43. Ben-Dor, A., Shamir, R. & Yakhini, Z. Clustering gene expression patterns. *J. Comput. Biol.* **6**, 281–297 (1999).
44. De Smet, F. *et al*. Adaptive quality-based clustering of gene expression profiles. *Bioinformatics* **18**, 735–746 (2002).
45. Heyer, L.J., Kruglyak, S. & Yooseph, S. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.* **9**, 1106–1115 (1999).
46. Sharan, R. & Shamir, R. CLICK: a clustering algorithm with applications to gene expression analysis. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 307–316 (2000).
47. Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E. & Ruzzo, W.L. Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977–987 (2001).
48. Fraley, C. & Raftery, A.E. Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Stat. Assoc.* **97**, 611–631 (2002).
49. Hastie, T. *et al*. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* **1**, research0003 (2000).
50. Yeung, K.Y., Haynor, D.R. & Ruzzo, W.L. Validating clustering for gene expression data. *Bioinformatics* **17**, 309–318 (2001).
51. McShane, L.M. *et al*. Methods of assessing reproducibility of clustering patterns observed in analysis of microarray data. *Bioinformatics* **18**, 1462–1469 (2002).
52. Kerr, M.K. & Churchill, G.A. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl Acad. Sci. USA* **98**, 8961–8965 (2001).
53. Gordon, A.D. *Classification* (Chapman & Hall/CRC, Boca Raton, FL, 1999).
54. Ben-Hur, A., Elisseeff, A. & Guyon, I. A stability based method for discovering structure in clustered data. *Pac. Symp. Biocomput.* 6–17 (2002).
55. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a dataset via the gap statistic. *J. Roy. Statist. Soc. B* **63**, 411–423 (2001).
56. Khan, J. *et al*. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med.* **7**, 673–679 (2001).
57. Armstrong, S.A. *et al*. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genet.* **30**, 41–47 (2002).
58. Pomeroy, S.L. *et al*. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**, 436–442 (2002).
59. Sorlie, T. *et al*. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98**, 10869–10874 (2001).
60. van 't Veer, L.J. *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
61. Chung, C.H., Bernard, P.S. & Perou, C.M. Molecular portraits and the family tree of cancer. *Nature Genet.* **32**, 533–540 (2002).
62. Dudoit, S., Fridlyand, J. & Speed, T.P. Comparison of discrimination methods for the classification of tumors using gene expression data. Technical Report 576. (Department of Statistics, University of California at Berkeley, Berkeley, CA, 2000).
63. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA* **99**, 6567–6572 (2002).
64. Hedenfalk, I. *et al*. Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* **344**, 539–548 (2001).
65. Ramaswamy, S. *et al*. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA* **98**, 15149–15154 (2001).
66. Mitchell, T.M. *Machine Learning*, 414 (WCB McGraw-Hill, Boston, 1997).
67. Califano, A., Stolovitzky, G. & Tu, Y. Analysis of gene expression microarrays for phenotype classification. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 75–85 (2000).
68. Brown, M.P. *et al*. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA* **97**, 262–267 (2000).
69. Furey, T.S. *et al*. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**, 906–914 (2000).
70. Breiman, L. Bagging predictors. *Machine Learning* **24**, 123–140 (1996).
71. Schapire, R.E., Freund, Y., Bartlett, P. & Lee, W.S. Boosting the margin: a new explanation for the effectiveness of voting methods. *Annls Stat.* **26**, 1651–1686 (1998).
72. Schapire, R.E. The strength of weak learnability. *Machine Learning* **5**, 197–227 (1990).
73. Breiman, L. *Manual on Setting Up, Using, and Understanding Random Forests v3.1*. (University of California at Berkeley, Berkeley, CA, 2002).
74. Shipp, M.A. *et al*. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Med.* **8**, 68–74 (2002).
75. Ben-Dor, A. *et al*. Tissue classification with gene expression profiles. *J. Comput. Biol.* **7**, 559–583 (2000).
76. Su, A.I. *et al*. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res.* **61**, 7388–7393 (2001).
77. Bo, T. & Jonassen, I. New feature subset selection procedures for classification of expression profiles. *Genome Biol.* **3**, research0017 (2002).
78. Butte, A.J. & Kohane, I.S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.* 418–429 (2000).
79. Liang, S., Fuhrman, S. & Somogyi, R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.* 18–29 (1998).
80. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7**, 601–620 (2000).
81. Ideker, T.E., Thorsson, V. & Karp, R.M. Discovery of regulatory interactions through perturbation: inference and experimental design. *Pac. Symp. Biocomput.* 305–316 (2000).
82. Hartemink, A.J., Gifford, D.K., Jaakkola, T.S. & Young, R.A. Combining location and expression data for principled discovery of genetic regulatory network models. *Pac. Symp. Biocomput.* 437–449 (2002).
83. Pe'er, D., Regev, A., Elidan, G. & Friedman, N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* **17 Suppl 1**, S215–S224 (2001).
84. Segal, E., Taskar, B., Gasch, A., Friedman, N. & Koller, D. Rich probabilistic models for gene expression. *Bioinformatics* **17 Suppl 1**, S243–S252 (2001).
85. Yoo, C., Thorsson, V. & Cooper, G.F. Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational DNA microarray data. *Pac. Symp. Biocomput.* 498–509 (2002).
86. Hartemink, A.J., Gifford, D.K., Jaakkola, T.S. & Young, R.A. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.* 422–433 (2001).
87. Potter, J.D. At the interfaces of epidemiology, genetics and genomics. *Nature Rev. Genet.* **2**, 142–147 (2001).
88. Kohane, I.S. Bioinformatics and clinical informatics: the imperative to collaborate. *J. Am. Med. Inform. Assoc.* **7**, 512–516 (2000).
89. Butte, A.J., Tamayo, P., Slonim, D., Golub, T.R. & Kohane, I.S. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl Acad. Sci. USA* **97**, 12182–12186 (2000).
90. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A.F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18 Suppl 1**, S233–S240 (2002).
91. Chiang, D.Y., Brown, P.O. & Eisen, M.B. Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. *Bioinformatics* **17 Suppl 1**, S49–S55 (2001).
92. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. Systematic determination of genetic network architecture. *Nature Genet.* **22**, 281–285 (1999).
93. Holmes, I. & Bruno, W.J. Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 202–210 (2000).
94. Shatkay, H., Edwards, S., Wilbur, W.J. & Boguski, M. Genes, themes and microarrays: using information retrieval for large-scale gene analysis. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 317–328 (2000).
95. Masys, D.R. *et al*. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics* **17**, 319–326 (2001).
96. Jenssen, T.K., Laegreid, A., Komorowski, J. & Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet.* **28**, 21–28 (2001).
97. Staunton, J.E. *et al*. Chemosensitivity prediction by transcriptional profiling. *Proc. Natl Acad. Sci. USA* **98**, 10787–10792 (2001).
98. Radmacher, M.D., McShane, L.M. & Simon, R. A paradigm for class prediction using gene expression profiles. *J. Comput. Biol.* **9**, 505–511 (2002).