Analyzing array data using supervised methods

Markus Ringnér[1][†],
Carsten Peterson[2] &
Javed Khan[3]

[†]*Author for correspondence*
[1]*Cancer Genetics Branch,*
National Human Genome
Research Institute,
National Institutes of Health,
Building 50, Room 5142,
50 South Drive MSC 8000,
Bethesda, MD 20892, USA
[2]*Complex Systems Division,*
Department of Theoretical
Physics, Lund University,
Sölvegatan 14A,
SE-223 62 Lund, Sweden
[3]*Advanced Technology Center,*
National Cancer Institute,
National Institutes of Health,
Room 134E, 8717
Grovemont Circle,
Gaithersburg,
MD 20877, USA
Tel: +1 301 496 5148;
Fax: +1 301 402 3241;
E-mail: mringner@nhgri.
nih.gov.

Pharmacogenomics is the application of genomic technologies to drug discovery and development, as well as for the elucidation of the mechanisms of drug action on cells and organisms. DNA microarrays measure genome-wide gene expression patterns and are an important tool for pharmacogenomic applications, such as the identification of molecular targets for drugs, toxicological studies and molecular diagnostics. Genome-wide investigations generate vast amounts of data and there is a need for computational methods to manage and analyze this information. Recently, several supervised methods, in which other information is utilized together with gene expression data, have been used to characterize genes and samples. The choice of analysis methods will influence the results and their interpretation, therefore it is important to be familiar with each method, its scope and limitations. Here, methods with special reference to applications for pharmacogenomics are reviewed.

As the sequencing and gene annotation projects of entire genomes of many species are headed towards completion [1], massive mapping efforts are now focused on the genes' functions and interactions. Microarray (also known as DNA chip, gene chip or biochip) technology is a rapid method of analyzing large numbers of genes simultaneously. A DNA microarray system usually consists of DNA probes formatted on a microscale on a glass substrate surface, together with instruments to read reporter (fluorescent) molecules (scanner) and to analyze the data (images) generated. Gene transcripts or genomic DNA extracted from samples are labeled with reporter molecules and hybridized to the probes formatted on the slides. For two-color fluorescence systems, the relative abundance of the sequences hybridized to each DNA probe is subsequently read from the glass slides. The glass slides can be constructed using:

- double stranded complimentary DNA clones (cDNA arrays) [2]
- short oligonucleotides (~ 23 mer) synthesized *in situ* [3]
- synthesized long oligonucleotides (30–70 mer) [4]
- genomic clones [5]

The microarray technology has been reviewed elsewhere [6-8]. This review is limited to the analysis of gene expression measurements generated by DNA microarrays.

Two major designs of microarray expression experiments exist: time series and static. In time series experiments, which for many experimental systems have so far been limited to cell culture experiments (cell lines), each experiment corresponds to a discrete measured time point. Potential applications include investigations of gene expression responses, for different genotypes, to external stimuli, such as drugs, environment or hormones. In static applications each experiment typically corresponds to a different tissue, cell line or blood sample. In terms of analysis objectives, one aims at relating the measured gene expression patterns to phenotypes, such as diagnosis, outcome, treatment response or drug resistance and in this process determine the most important genes for the questions posed.

Approaches to the computational analysis of gene expression data can be separated into two groups: unsupervised and supervised. In unsupervised methods the gene expression patterns are grouped based solely on the expression data. Unsupervised methods are particularly useful to analyze the data in an exploratory fashion, for example, to enable the formulation of novel hypotheses or to discover experimental artifacts at an early stage of investigation. If one has some prior information or hypothesis about which samples or genes are expected to group together, this information can be utilized in a supervised method. The main reason for choosing a supervised method is that one desires a classifier or predictor. To use a supervised method, one has to know the 'correct' classification for at least some of the samples, which are to be used as a training set to calibrate the method. Therefore, the choice of method to analyze the data is a fundamental

Table 1. Methods used for array data analysis.

Category	Method
Unsupervised clustering	Hierarchical clustering, *K*-means clustering, MDS, self-organizing maps
Supervised discriminatory gene classifiers	F-test, t-test, Mann-Whitney U-test, Wilcoxon rank score, total number of mis-classifications score, signal-to-noise statistic, MDS weighted gene analysis, ANOVA
Supervised machine learning classifiers	Support vector machines, multi-layer perceptron artificial neural networks

ANOVA: Analysis of variance; MDS: Multi-dimensional scaling.

decision-making step in experimental design, prior to the initiation of the experiments. For example, to make sure that sufficient numbers of samples with known classifications are profiled for the training set to be used in a supervised method. In addition, once a classifier has been constructed using a supervised method it is crucial to use an independent test set or a cross-validation technique to estimate its classification error.

This paper reviews applications of supervised methods in the analysis of microarray experiments with special reference to pharmacogenomics. It is widely appreciated that there are many important applications of microarrays in pharmacogenomics, for example:

- molecular target identification and drug discovery
- toxicology
- molecular diagnostics

The massive amount of data generated by genomic methods has led to a need for computational methods to manage and analyze this data and the methods used will influence the results and their interpretation. The data mining tools employed range from various clustering techniques to supervised learning schemes [9]. The main emphasis of this review is on supervised classifications methods, a brief summary of some of the unsupervised methods used for array analysis is also given. This will provide some necessary requisites for the following discussion on the advantages of using supervised methods in the context of pharmacogenomics. The most common methods used to analyze array data are listed in **Table 1**.

Preprocessing

Prior to applying any computational analysis tools, one needs to assess, and if necessary correct for, the quality of the data. The simplest and most straightforward approach is to apply cuts based on intensities and spot areas. This can be done more elegantly by using an error model to estimate whether an expression ratio is departing from unity due to measurement errors alone [10]. However, such procedures may remove genes that only have low quality measurements for a few samples from the entire data set. Remedies for this include taking the quality into account explicitly in the analysis by weighting measurements with a quality factor [11] or using missing value algorithms. The latter can be quite elaborative and include user-defined choices and parameters [12]. More profound and sophisticated corrections for noise have been suggested [13,14]. Here the biological signals are separated from other effects (for example, noise and experimental variation) and the latter are modeled and the model is fitted to the data to allow relevant signals to be extracted. For most experiments, the number of samples is relatively small, as compared to the number of measured genes, this could lead to erroneous conclusions as one might distort the relevant biological signal. Nevertheless, as larger samples and replication of experiments become standard, this will be the basic approach to determine significant signals. When supervised machine learning approaches are used one might take a more pragmatic attitude and assume that the calibrated feature model implicitly corrects for features that are not related to the relevant biology but present in the data (as verified by the success on an independent test set).

Unsupervised analysis and dimensional reduction

Many of the algorithms used for the analysis of array data are based on pair-wise comparison of expression patterns of either genes or samples. This is addressed by mathematically defining a measure of distance (or similarity) between genes or samples in 'expression space'. Unsupervised clustering algorithms group samples or genes based on their separation in expression space, as given by the distance metric. Different choices of distance metric will place different objects in different clusters [9].

The most commonly used method for clustering in gene expression space is hierarchical clustering. It has been used both to reveal sample closeness, for example for rhabdomyosarcomas [15], B cell lymphomas [16], breast tumors [17], colon adenocarcinomas [18] and lung adenocarcinomas [19,20], as well as to cluster genes with similar behavior in time course experiments [21],

with the aim to find functionally related genes. Other clustering approaches that have been used extensively are K-means clustering [22] and self-organizing maps (SOMs) [23].

Since the number of genes measured is very large, one cannot visualize samples in expression space directly. One way to reduce the dimensionality of the samples that is aimed at qualitative displays rather than quantitative analysis, is multi-dimensional scaling (MDS). This method has been frequently used in expression analysis to display samples, for example to characterize alveolar rhabdomyosarcoma [15] and cutaneous melanoma [24]. Another standard tool to visualize samples (or genes) is principal component analysis (PCA), which is a technique that rotates expression space, such that the variance of expression is dominated by as few linear combinations of genes (or samples) as possible. Not only can this be a good visualization tool, when the two or three leading components are retained [25] but in contrast to MDS an analytical form for the transformation exists. Hence, it can be used as a preprocessing tool [9]; in particular for supervised learning [26] as will be discussed below.

Supervised classification

Supervised approaches are well suited to categorize samples into known phenotypes. Typically, two goals are on the agenda in these investigations:

- develop a robust classifier with validation procedures that can successfully handle blinded test data
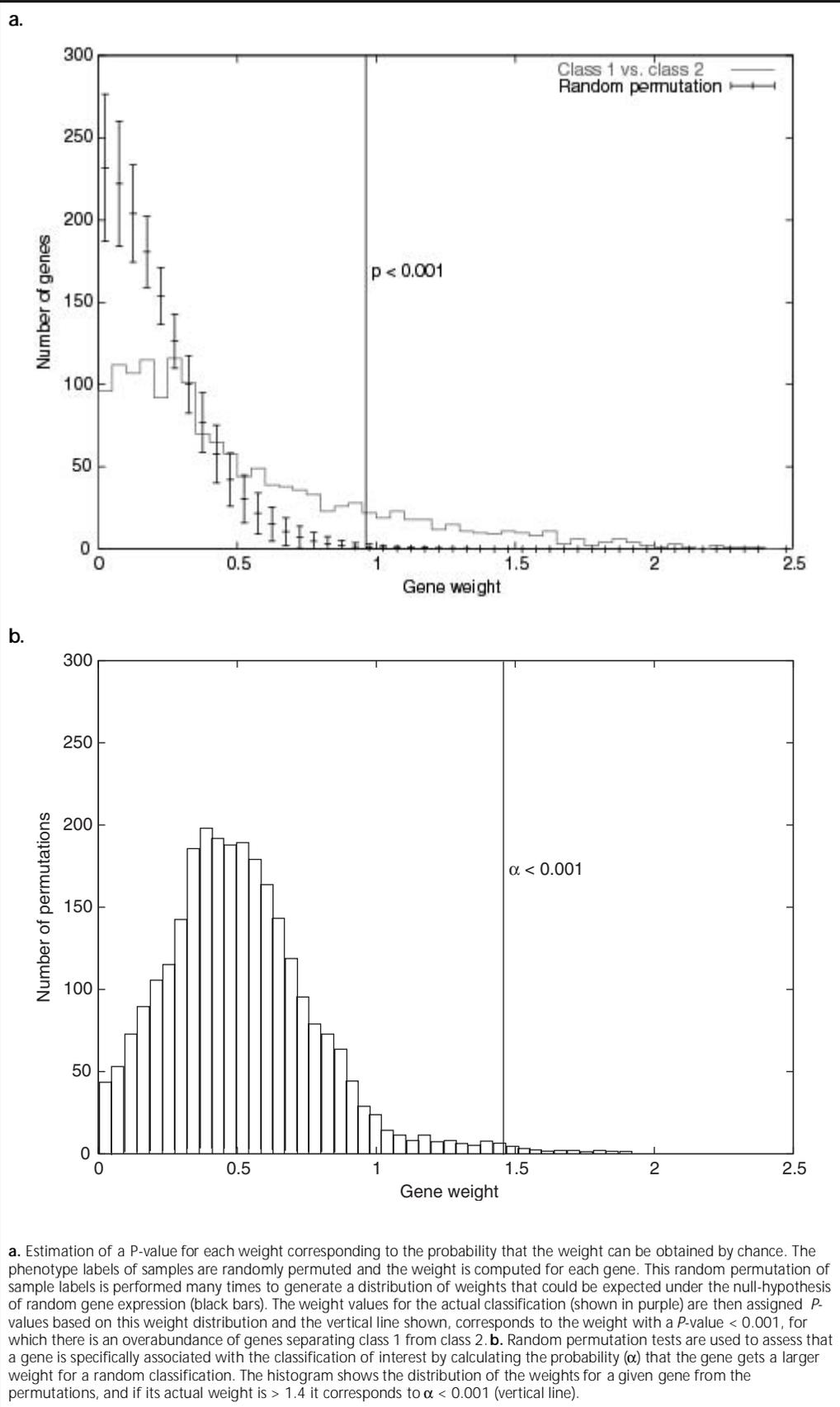- identify the genes that are most important for the classification

Hence, two objectives are achieved simultaneously. A diagnostic/prognostic tool for clinical use, that can be used to diagnose a disease or to predict the outcome or treatment response for samples is obtained, at the same time insights into the underlying molecular biology are gained. The latter can be explored to find candidate drug targets or to better understand why a treatment is not working for some patients. This section provides an overview of how supervised methods have been and can be used for analyzing array data. As more and more microarray data become publicly available it is likely that rigorous evaluations of the performance of different methods in this context will be undertaken [27].

Discriminatory gene classifiers

Disregarding collective effects amongst the genes, limiting the investigation to single gene dependencies, the second goal (above) is generally achieved using various statistical measures to, gene-by-gene, correlate expression levels of a single gene with a phenotype of interest [24,28-32]. In this way a discriminatory weight is calculated for each gene. Typically, the number of genes is much greater than the number of experiments, it is expected by random chance that some genes correlate highly with the phenotype, resulting in large weights. Therefore, it is crucial to estimate a P-value for each weight that corresponds to the probability that the weight can be obtained by chance for such a large number of genes. P-values are readily calculated using random permutation tests. In these tests, the phenotype labels of samples are randomly permuted and the weight is computed for each gene. This random permutation of sample labels is performed many times to generate a distribution of weights that could be expected under the null-hypothesis of random gene expression. The weight values for the actual classification can then be assigned P-values based on the weight distribution from the random permutations [24,28]. In this way, it can be verified that the discriminatory genes have significant weights and that there is an overabundance of genes discriminating between the phenotypes (**Figure 1a**). In addition, random permutation tests can be used to assess whether a gene is specifically associated with the classification of interest by calculating the probability (α) that the gene gets a larger weight for a random classification [29,31] (**Figure 1b**). Together, a small P and α indicate a good discriminatory gene. Once the genes are ranked according to the discriminatory weights, supervised classifiers can be constructed using the top-ranked genes. A general approach to classify additional samples is that each gene gives a flat vote or a vote weighted with the weight, the gene's expression level or the difference between the average expressions in the classes. Alternatively, the significant genes can be used in for example, a (K-)nearest-neighbor classifier [25,33] or similar methods [34]. A plethora of statistical measures have been used for discriminatory weights. Golub *et al.* used a signal-to-noise statistic, designed to find genes that on average were expressed differently in two groups, but also had a small variation of expression within each group to discriminate acute myeloid leukemia from acute lymphoblastic leukemia [28]. The signal-to-noise statistic has subsequently been extensively used, for example, to select marker genes for distinct lung adenocarcinoma subclasses [19] and to characterize the

Figure 1. Identification of discriminatory genes.

a.

b.

a. Estimation of a P-value for each weight corresponding to the probability that the weight can be obtained by chance. The phenotype labels of samples are randomly permuted and the weight is computed for each gene. This random permutation of sample labels is performed many times to generate a distribution of weights that could be expected under the null-hypothesis of random gene expression (black bars). The weight values for the actual classification (shown in purple) are then assigned *P*-values based on this weight distribution and the vertical line shown, corresponds to the weight with a *P*-value < 0.001, for which there is an overabundance of genes separating class 1 from class 2. **b.** Random permutation tests are used to assess that a gene is specifically associated with the classification of interest by calculating the probability (α) that the gene gets a larger weight for a random classification. The histogram shows the distribution of the weights for a given gene from the permutations, and if its actual weight is > 1.4 it corresponds to α < 0.001 (vertical line).

expression profiles of gastrointestinal stromal tumors with *KIT* mutations [31]. A standard *t*-test has been used to discriminate *BRCA1* from *BRCA2* breast tumors [29] and to identify therapeutic targets for metastatic medulloblastoma [19]. Other standard statistical methods such as *F*-test or analysis of variance (ANOVA) can easily be applied in a similar way. To find genes associated with survival in esophageal tumors, the rank based Mann-Whitney *U*-test has been applied [30]. Another approach is to find discriminatory genes using the total number of mis-classification (TNoM) score which, based on a threshold expression value, measures the number of mis-classified samples [24,29,35,36]. A supervised method, inspired by MDS analysis, to weight and rank discriminatory genes has been used to identify genes associated with a highly aggressive subset of melanomas [24] and genes differing in expression between human prostate cancer and benign prostatic hyperplasia [37]. Furthermore, Fisher's linear discriminant can be used as a method for expression-based tumor classifications [38].

It has been suggested that the simple idea of looking for genes that are preferentially expressed in a given tissue and are located in regions containing unidentified disease genes provides a shortcut to finding genes implicated in human diseases [39]. This is in agreement with the finding by Allander *et al.* [31] that *KIT* itself is indeed the number one ranked discriminator for gastrointestinal stromal tumors with *KIT* mutations. Systematic investigations into genes that are distinguished not only by their relative expression in the affected tissue as compared to other tissues, but also by their absolute mRNA abundance in the tissue in question, are likely to be a fruitful approach to reduce the number of candidate genes in searches for disease genes.

Machine learning methods

For supervised learning that includes collective and nonlinear effects among genes one can pursue two different paths:

- support vector machines (SVMs) [40]
- artificial neural networks (ANNs) [41]

Both methods are computer-based supervised learning algorithms that can be trained to recognize and characterize complex patterns. Pattern recognition is achieved by adjusting the parameters of the models fitting the data by a process of error (for example, mis-classification) minimization through learning from experience (using training samples). Both approaches have pros and cons. Since array data are very high dimensional, ANNs generally require some preprocessing to avoid overfitting, which is not the case for SVMs. On the other hand, the results from ANNs allow for a straightforward probability interpretation and ANNs are more easily generalized to multi-class classification problems. In addition, ANNs can be used not only to classify samples according to a dichotomous distinction (such as, cured versus fatal/refractory disease) but also according to more sample specific phenotypes such as time of survival (a continuous variable).

Support vector machines

SVMs [40], a supervised machine learning technique, are well suited to work with high dimensional data such as array-based expression data. When used for classification, SVMs separate one class from the other in a set of binary training data with a hyper-plane that is maximally distant (called the maximal margin hyper-plane) from the training examples. However, most real-world problems involve data for which no such hyper-plane exists. SVMs solve this inseparability by mapping the data from the original input space into a higher dimensional space and define a hyper-plane that separates the data there. This higher dimensional space is called feature space and the hyper-plane found in this space corresponds to a nonlinear decision boundary in the original input space. An appealing feature with SVMs is that data does not need to be explicitly represented in feature space; the hyper-plane can be located simply by defining a kernel function, which plays the role of a dot product in the feature space. This dot product can be viewed as analogous to the distance measures used in the clustering algorithms above. However, SVMs are capable of using a larger variety of such functions. The fact that SVMs only use a kernel function and do not need an explicit high dimensional representation of the data, is what makes them appealing for use in the supervised classification of multi-dimensional array (MDA) data (typically having relatively small sample numbers). One obvious pitfall with the introduction of feature space is that by artificially separating the data optimally in this way, one may risk finding trivial solutions that overfit the data. Sometimes, such as in the presence of noise, it is better to trade some training accuracy for better predictive power. SVMs address this problem by using a soft margin that tolerates training errors. Hence, a support vector machine is specified by

choosing both a kernel function and setting a parameter that controls the training error.

The initial application of SVMs to array data aimed at functional classification of genes based on their expression patterns [42]. Brown *et al.* did this by training SVMs to recognize genes belonging to five functional classes, as defined in a database based on biochemical and genetic studies. A sixth class not expected to exhibit similar expression profiles was included as a control group. The study showed that SVMs provided superior performance as compared to four non-SVM methods (not including ANNs) and to unsupervised clustering methods such as hierarchical clustering.

Furey *et al.* applied SVMs to the classification and validation of cancer tissue samples using microarray expression data [43]. The method was primarily applied to the classification of ovarian tissue samples. The aim was to separate ovarian cancer samples from normal ovarian and nonovarian tissue samples. Leave-one-out cross-validation was performed to evaluate the classification performance. The classification results were relatively good and interestingly, the one sample that was most difficult to classify turned out to be incorrectly labeled. This shows the potential of supervised learning methods to identify difficult cases. However, no method to identify the genes most important for SVM classification, which is of particular interest in pharmacogenomics, was presented. Instead, the signal-to-noise statistic [28] was used to select a subset of genes to be used in the SVM analysis.

Ramaswamy *et al.* [44] and Su *et al.* [45] used SVMs to diagnose multiple common adult malignancies. These studies demonstrated the feasibility of multi-class molecular cancer classifications. Since SVMs are not easily directly adapted to multi-class classifications, they both used an approach in which a committee of classifiers, each identifying one class of cancers from all others, was used. Su *et al.* compared various classification methods and got somewhat better classification results using methods that make no assumptions about the distribution of the data (such as SVMs or ANNs), as compared to supervised weighted correlations methods [28] and other supervised learning methods (for example, Fisher's linear discriminant). Both devised methods to rank genes according to their importance in classifying samples. Su *et al.* filtered for discriminatory genes using the Wilcoxon rank score, followed by ranking genes based on their predictive accuracy in a leave-one-out cross-validation scheme. Ramaswamy *et al.,* on the other hand, analyzed the calibrated SVMs and ranked the genes according to their contribution to defining the decision hyper-plane, that is according to their importance in classifying the samples. This latter approach in principle allows each gene to be ranked for each sample. Since it is likely that distinct clinical behaviors are explained by different molecular mechanisms in different patients, this approach hints at the potential use of machine learning methods to extract an individual's genetic profile – thereby creating the possibility of tailoring treatment for each patient.

Artificial neural networks

The first generation of ANNs, so-called perceptrons, was simple linear logistic regression methods. More elaborate ANNs in the form of a multilayer perceptron is another machine learning approach that has proven to be powerful when classifying tumor array-based expression data [26] (see [46] for a review on applications of ANNs for biological systems). A multilayered perceptron consists of a set of layers of perceptrons, modeled on the structure and behavior of neurons in the human brain. The input data, in this case the gene expression data, is fed into the so-called input layer and triggers a response in the following so-called hidden layer(s). The response in the hidden layer(s) in turn triggers a response in the output layer. In the case of classification, each perceptron in the output layer typically represents a class. When the gene expression pattern of a sample is fed into the ANN, ideally only the output perceptron representing the class that the sample belongs to should respond. For calibration, samples belonging to the classes of interest are presented to the ANNs, which are trained to recognize them in a supervised fashion by a process of error minimization. Since the number of perceptrons in the input layer depends on the dimension of the input data, a large number of perceptrons is needed for high dimensional data. Furthermore, the more perceptrons there are in the ANN, the more training samples are needed to calibrate all the perceptrons in such a way that the classifier has good predictive power. In the case of array data where the number of samples is much less than the number of measured genes this leads to a large risk of overfitting. There are two parts to the solution to this problem. Firstly, the dimension of the data can be reduced, either by using a dimensional reduction algorithm

such as PCA [26] or by selecting a smaller set of genes as input to the classifier in a supervised way by using a discriminatory score (see for example [45]). Secondly, the learning process can be carefully monitored using a cross-validation scheme to avoid overtraining [26]. Another advantage with using a cross-validation scheme is that it results in a set of models, each trained on a subset of the samples, which can be used as a committee to classify test samples in a robust way [26]. Other methods to avoid overtraining of ANNs that may reduce the need to reduce the dimension of the data include regularization (for example, weight decay), pruning and training with noise [41]. A possible way to use ANNs to classify array data is shown, together with illustrations of the calibration procedure, in **Figure 2**. The schematic illustration in **Figure 2a** of the analysis process is similar to that used for many supervised classification methods. As compared to SVMs, ANNs have been shown to represent most functions that in this case map expression space onto phenotypes, while SVMs instead map the data onto a higher-dimensional space in which the data are linearly separable into two phenotypes.

Khan *et al.* [26] used ANNs for classification and diagnostic prediction of small, round blue cell tumors, belonging to four different diagnostic categories. To determine which genes were most important for the classification, Khan *et al.* analyzed the calibrated ANNs and ranked the genes according to how sensitive the output was with respect to each gene's expression level. As an example, they found *FGFR4*, a tyrosine kinase receptor, to be highly expressed in rhabdomyosarcoma, a finding with therapeutic potential. This gene ranking method shares its philosophy with the approach used by Ramaswamy *et al.* [44] for SVMs. In particular, it can rank each gene for each patient individually. This study demonstrated the potential applications of ANN-based methods for tumor diagnoses and the identification of candidate targets for therapy.

Gruvberger *et al.* used ANNs to investigate the phenotype associated with estrogen receptor (ER) α status in human breast cancer and found that the ANNs could accurately classify the tumors [47] into ER-positive and -negative samples. Furthermore, they found that the ANNs could accurately predict the ER status even when excluding top discriminator genes. These results provided evidence that ER-positive and ER-neg-ative tumors display remarkably different gene-expression phenotypes.

Even though the optimal number of genes selected for use in SVMs or ANNs is typically chosen by optimizing the training performance, it should be noted that random permutation tests to assess the significance of each highly ranked gene are also feasible in this context. An advantage with ANNs is that they can easily be adopted to predict continuous values instead of classes. This can, for example, be used to predict protein levels of the ER-receptor instead of classifying samples into binary ER-positive or -negative classes. Such a prediction method can potentially be used to gain further insights into the relevant genes and is likely to be useful for patient outcome predictions, where survival times may be of importance.

Companies involved in array analysis

Some of the companies involved in analysis of microarray data for pharmacogenomic applications are listed in **Table 2**.

Conclusions and expert opinion

In recent years, as microarrays have begun to be used to a larger extent for investigating expression profiles of diseases, the emphasis has, for the analysis methods, shifted from unsupervised to supervised clustering methods. This is largely because they are better suited to identify genes specific to a given phenotype, such as patient outcome after treatment. One of the great hopes of microarrays has always been to use the pattern reflecting the molecular state of a sample, under some specific condition, to identify particular characteristics of an individual, such as propensity to a disorder or response to a drug. The identification of these characteristics depends crucially on the analysis methods used. Recently, supervised machine learning classification techniques have been used to extract the genes most important for classification, in ways that allow for the genes to be investigated for each sample individually [26,44]. This hints at the future use of machine learning methods not only as new diagnostic tests but also to help physicians develop and choose the drugs that work best and have least side effects for a given individual.

To illustrate differences in results obtained using unsupervised and supervised methods a comparison of the investigations by Alizadeh *et al.* [16] and Shipp *et al.* [32] serves as a good example. Both analyze gene expression patterns of diffuse large B cell lymphoma (DLBCL). Ali-

Figure 2. An ANN based classifier.

a.

Figure 2a. Schematic illustration of the analysis process. The entire data set of N samples is quality filtered **1.** and then the dimensionality is further reduced by PCA to 10 PCA projections **2.** from the original M expression values. Next the N2 test experiments are set aside and N1 training experiments are randomly partitioned into three groups **3.** One of these groups is reserved for validation and the two remaining groups are used for calibration **4.** ANN models are then calibrated using the 10 PCA values for each sample as input and the phenotype category as output **5.** For each model the calibration is optimized with a number of iterative cycles (epochs). This is repeated using each of the three groups for validation **6.** Samples are again randomly partitioned and the entire training process repeated **7.** For each selection of validation group one model is calibrated resulting in a total of 3 x K trained models. Once the models are calibrated, they are used to rank the genes according to their importance for the classification **8.** The entire process (2–7) is repeated using only the top ranked genes **9.** The N2 test experiments are subsequently classified using all the calibrated models.
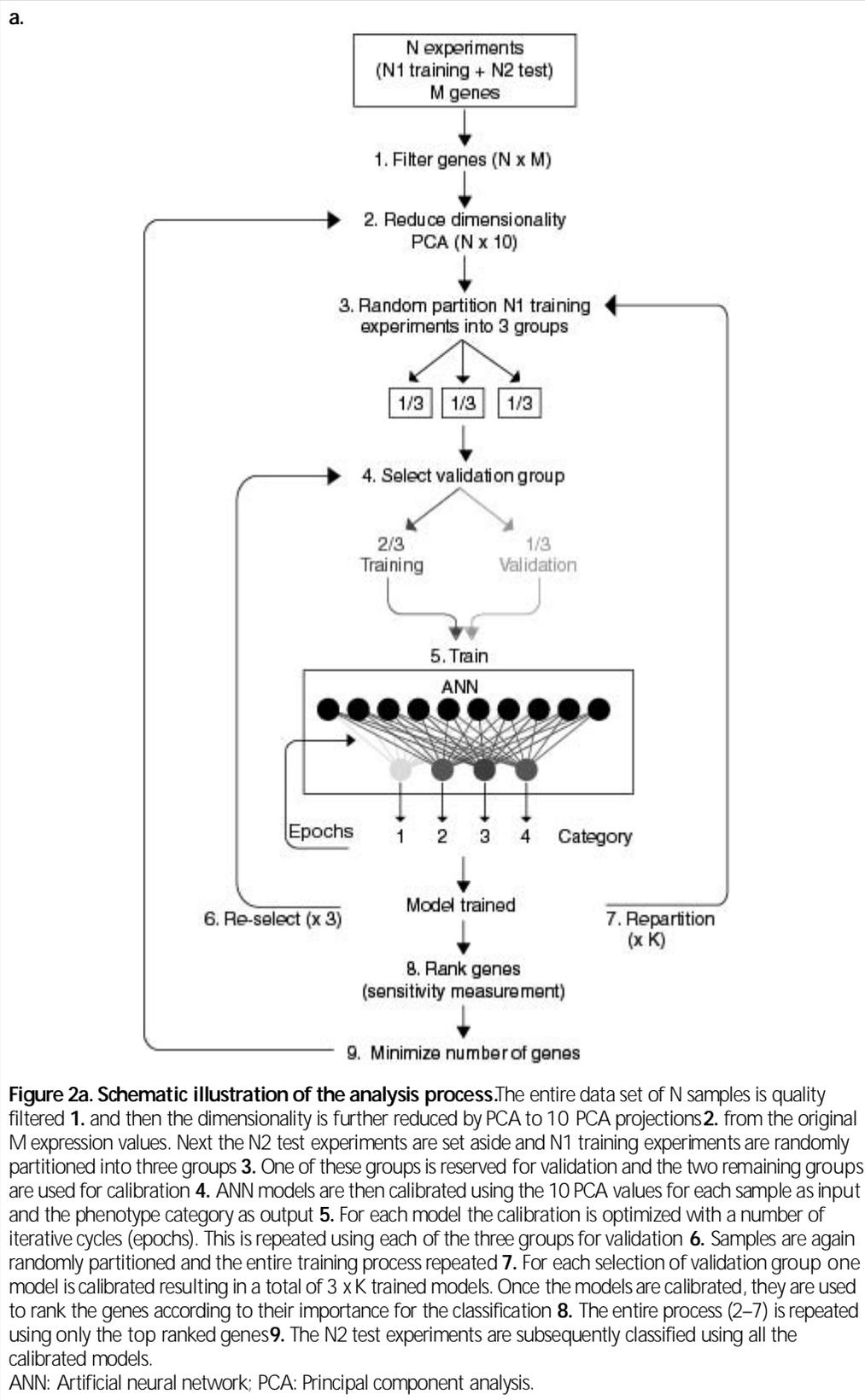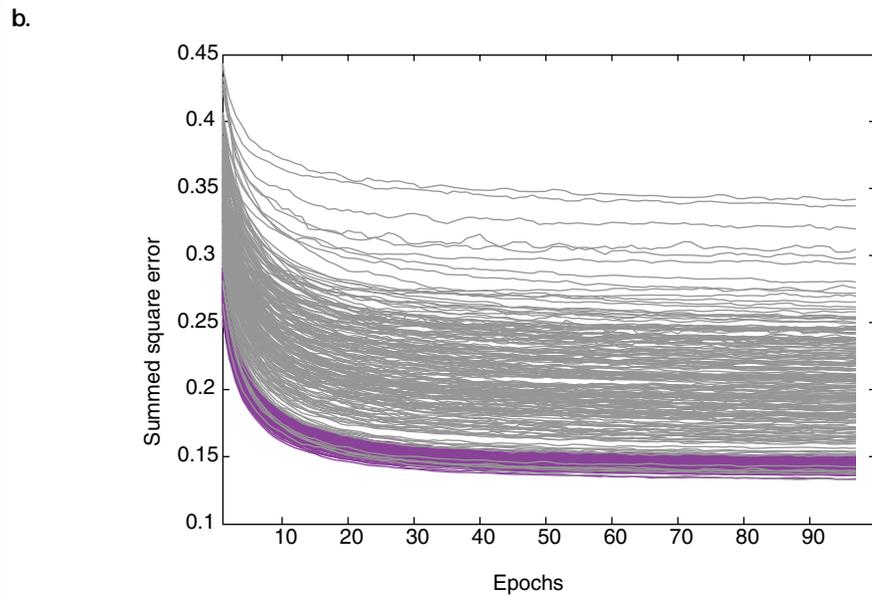ANN: Artificial neural network; PCA: Principal component analysis.

Figure 2. An ANN based classifier.

b.

Figure 2b. Monitoring the calibration of the models. The average classification error per experiment (using a summed square error measure) is plotted during the training iterations (epochs). A pair of lines, purple (training) and gray (validation) represents one model (each corresponding to a random partitioning of the data). The decrease in the errors with increasing epochs demonstrates the learning of the models to classify the experiments. All the models perform well for both training and validation. In addition, there is no sign of over-fitting, which would result in an increase in the error for the validation at the point where the models begin to learn features in the training set that are not present in the validation set.

c.
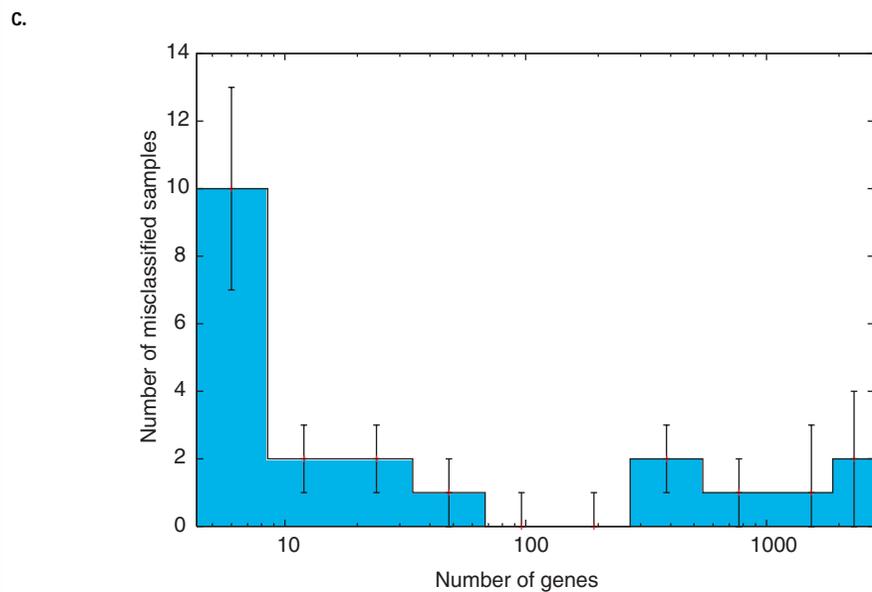
Figure 2c. Minimizing the number of genes. The average number of misclassified samples for all models is plotted against increasing number of used genes. As can be seen using the 96 highest ranked genes results in zero mis-classifications for this example.

Reprinted and adapted with permission from [26]. ©2001 Nature Publishing Group.

ANN: Artificial neural network.

Table 2. Some of the companies involved in developing microarray data analysis methods for pharmacogenomics.

Company	Software	Methods/applications
Affymetrix	Data Mining Tool	Clustering and discriminatory gene analysis.
Applied Maths	GeneMaths	Clustering tools with bootstrap methods to indicate significance. PCA, SOM and discriminant analysis. Provides an integrated platform together with Array-Pro™ from Media Cybernetics.
BioDiscovery	ArrayPack™ GeneSight™	Integrated expression management system. K-means and hierachical clustering, SOMs and PCA. Discriminatory gene analysis.
Clustan	ClustanGraphics	Numerous clustering methods. Cluster validation.
GeneData	GeneData Expressionist™	K-means and hierachical clustering and SOMs. Discriminatory gene analysis.
InforMax	Xpression NTI	Hierarchical and nonhierarchical clustering methods.
InforSense	Kensington Discovery	Methods for clustering, time-series analysis, classification (decision tree, neural network and Bayesian), predictive modeling, dimensional reduction and discriminatory gene analysis.
Iobion Informatics	GeneTraffic™	Hierarchical and K-means clustering. PCA and MDS.
Lion Biosciences	ArraySCOUT™	Connectivity to modules for analysis of molecular networks and biological pathways.
Molmine	J-express	Hierarchical and K-means clustering. SOMs and PCA. Profile similarity search.
OmniViz	OmniViz Pro™	Clustering, dimensionality reduction and projection methods. Correlating genotypes such as SNPs to therapeutic outcome or response.
Partek	Partek Pro Partek Discoverer Partek Infer Partek Predict	PCA, MDS, cluster analysis, neural network regression models and discriminant analysis. Bootstrap for model validation.
Rosetta Biosoftware	Rosetta Resolver™	Bayesian classifiers, PCA and discriminatory gene analysis.
Silicon Genetics	GeneSpring™ GeNet™ Metamine™	Machine learning tools, clustering methods and PCA. Integrated platform for gene expression research.
Spotfire	DecisionSite™	Clustering and prediction tools. Integrated platform for functional genomics.

MDS: Multidimensional scaling; PCA: Principal component analysis; SOM: Self-organizing map.

zadeh *et al.* used unsupervised hierarchical clustering [21] to show that DLBCLs fell into two groups according to the biological origin of the malignancies:

- those with expression profiles similar to normal germinal center (GC) B cells
- those with expression profiles similar to *in vitro* activated peripheral blood B cells

Although this study was not primarily aimed at predicting the outcome of disease, Alizadeh *et al.* found the GC-like DLBCLs to have a more favorable outcome. On the other hand, Shipp *et al.* used a supervised weighted voting scheme based on the signal-to-noise statistic [28] to directly relate the gene expression patterns of DLBCLs to patient outcome (after standard chemotherapeutic treatment) and the method was evaluated using a leave-one-out cross-validation method. The highest outcome prediction accuracy was achieved using predictors containing 13 genes, and the results at diagnosis indicate the presence of a gene expression signature for outcome in DLBCL. Of note is the connection between this classification and the cell-of-origin classification by Alizadeh *et al.* Shipp *et al.* found

that the discriminatory genes described by Alizadeh *et al.* also significantly separated their DLBCL samples according to the cell-of-origin distinction. However, in this data set this distinction was not correlated with patient outcome. On the other hand, those of the 13 genes used in the Ship *et al.* predictors that were present on the Alizadeh *et al.* arrays were, when evaluated as single markers, clearly correlated with outcome in the Alizadeh *et al.* expression data. Even though larger sample sets are needed to pinpoint optimal genes related to patient outcome in DLBCLs, these results illustrate important differences between supervised and unsupervised approaches. The supervised approach found genes associated with significant outcome differences in both data sets and some of the genes were related to apoptotic responses to receptor engagement and potentially to cytotoxic therapy. This suggests that the advantage with a supervised analysis method is not only that it is extendable to a clinical setting in the form of a classifier or predictor but also that the approach may clearly suggest strategies for the use and development of therapies. In conclusion, if one has a clear hypothesis about different categories of samples a supervised method is advantageous and allows the construction of a classifier/predictor. However, since there are many genes compared to the number of samples, it is crucial to validate a supervised classifier using an independent sample set. Otherwise the classifier may depend on features that are only present in the training set, thus having poor predictive power. This is particularly important in the relatively new field of microarray research, since the measurements are noisy and subject to experimental

variability and the sample sets are small and may not be as well matched as in traditional clinical investigations. A potential advantage with unsupervised methods is to generate novel hypotheses, as illustrated by the finding of the two groups of DLBCLs [16].

In addition to expression arrays, other microarrays exist or are under development, which also hold great promise as diagnostic tools and aids to biological research. One example is antibody-based arrays to measure the levels of proteins in tissues. From an analysis point of view, methods that can use and correlate information from different kinds of microarrays in a supervised fashion will be of great value. For example, comparative genomic hybridization can be performed to analyze the genomic content of tissue samples using the same cDNA microarrays used for expression analysis [48,49]. In this way, one obtains both the expression levels and copy numbers for the same large set of genes. Supervised methods can then be used to find genes whose expression levels are significantly attributable to their amplification or deletion status. Such transcripts and their encoded proteins would be ideal targets for anticancer therapies, as demonstrated by the clinical success of therapies against amplified oncogenes, such as *ERBB2* [50] and *EGFR* [51] in breast cancer and other solid tumors. Another type of array that has great potential to rapidly uncover the functions of genes is cell arrays [52]. These constructs can be queried for the consequences of expressing or potentially knocking-out genes, such that casual connections between genes are revealed. This is in contrast to expression arrays from which only correlation ('guilt by association') information regarding relations between genes is gained. Cell arrays will provide plenty of opportunities for the development of analysis methods to discover gene products that alter cellular physiology, unravel their pathways and identify small molecule targets affecting them. Furthermore, expression array data can be used together with drug activity patterns to elucidate gene-drug relationships [53-55].

Many problems, such as large costs, need for elaborate tissue preparation skills and difficulties in easily accessing patient samples, remain to be solved before microarrays will live up to their full potential for research and clinical applications. However, these problems are likely to be overcome with time. Similarly, the potential use of supervised analysis methods as diagnostic and drug target discovery tools has so far been somewhat limited by small sample sets. As larger sample sets and more ingenious array techniques

Highlights

- Unsupervised methods are useful for exploration of data sets for initial quality control, 'class discovery' and the formulation of novel hypotheses.
- Supervised methods are used for class prediction and identification of the most important genes for classification.
- Supervised methods require *a priori* knowledge of the 'correct' classification of at least some of the samples, which are used to calibrate the classifier.
- It is crucial to evaluate the performance of a supervised classifier using an independent test set or a cross-validation technique.
- The utility of multi-class molecular cancer classifications using supervised methods has been demonstrated.
- Several studies have demonstrated the feasibility of using supervised methods to predict outcome for various diseases.
- Methods to extract the genes most important for the classification from supervised methods have been developed, and are likely to identify possible novel targets for therapy.

become readily available, the promise of analyzing array data for pharmacogenomics applications is likely to finally be fulfiled.

Outlook

In the last decade, we have seen a rapid growth in the application of genomics and proteomics in biological, translational and clinical research. Array-based methods to diagnose and predict the outcome of diseases have been developed. Over the next decade, high-dimensional data generated from microarrays and proteomic-based

applications are likely to be taken to the clinic. One can predict that it will be possible to use these methods for rapid diagnosis and prediction of outcome. A large part of this will involve therapeutics and toxicology for the individualization of therapy and for optimizing personal dosage to minimize toxic side-effects of drugs – individual patient management. New genes and their products which are potential targets for therapy will be identified. We expect that supervised analysis methods will be an integral part of this translational research.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Lander ES, Linton LM, Birren B *et al.*: Initial sequencing and analysis of the human genome. *Nature* 409, 860-921 (2001).

2. Schena M, Shalon D, Davis RW, Brown PO: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467-470 (1995).

3. Lockhart DJ, Dong H, Byrne MC *et al.*: Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14, 1675-1680 (1996).

4. Hughes TR, Mao M, Jones AR *et al.*: Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* 19, 342-347 (2001).

5. Pinkel D, Segraves R, Sudar D *et al.*: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* 20, 207-211 (1998).

6. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: Expression profiling using cDNA microarrays. *Nat. Genet.* 21, 10-14 (1999).

7. Khan J, Bittner ML, Chen Y, Meltzer PS, Trent JM: DNA microarray technology: the anticipated impact on the study of human disease. *Biochim. Biophys. Acta* 1423, M17-M28 (1999).

8. Jain KK: Applications of biochip and microarray systems in pharmacogenomics. *Pharmacogenomics* 1, 289-307 (2000).

9. Quackenbush J: Computational analysis of microarray data. *Nat. Rev. Genet.* 2, 418-427 (2001).
   • **A concise review that provides knowledge about most common methods used to analyze array data.**

10. Marton MJ, DeRisi JL, Bennett HA *et al.*: Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat. Med.* 4, 1293-1301 (1998).

11. Chen Y, Kamat V, Dougherty ER, Bittner ML, Meltzer PS, Trent JM: Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics* (2002) In press.

12. Troyanskaya O, Cantor M, Sherlock G *et al.*: Missing value estimation methods for DNA microarrays. *Bioinformatics* 17, 520-525 (2001).

13. Ideker T, Thorsson V, Siegel AF, Hood LE: Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.* 7, 805-817 (2000).

14. Kerr MK, Churchill GA: Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. USA* 98, 8961-8965 (2001).

15. Khan J, Simon R, Bittner M *et al.*: Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* 58, 5009-5013 (1998).

16. Alizadeh AA, Eisen MB, Davis RE *et al.*: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503-511 (2000).
    •• **A very nice application of hierarchical clustering to characterize and subgroup human cancer.**

17. Perou CM, Sorlie T, Eisen MB *et al.*: Molecular portraits of human breast tumours. *Nature* 406, 747-752 (2000).
    • **Another nice illustration of hierarchical clustering to profile expression patterns of human cancer.**

18. Notterman DA, Alon U, Sierk AJ, Levine AJ: Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and

normal tissue examined by oligonucleotide arrays. *Cancer Res.* 61, 3124-3130 (2001).

19. MacDonald TJ, Brown KM, LaFleur B *et al.*: Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease. *Nat. Genet.* 29, 143-152 (2001).

20. Garber ME, Troyanskaya OG, Schluens K *et al.*: Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl. Acad. Sci. USA* 98, 13784-13789 (2001).

21. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95, 14863-14868 (1998).
    •• **This paper showed the potential of clustering methods to analyze genome-wide expression patterns.**

22. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281-285 (1999).

23. Tamayo P, Slonim D, Mesirov J *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).

24. Bittner M, Meltzer P, Chen Y *et al.*: Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406, 536-540 (2000).

25. Pomeroy SL, Tamayo P, Gaasenbeek M *et al.*: Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415, 436-442 (2002).

26. Khan J, Wei JS, Ringnér M *et al.*: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* 7, 673-679 (2001).
    •• **The first application of supervised artificial neural networks to classify human cancer**

and to identify the most important genes.

27. Dudoit S, Fridlyand J, Speed TP: Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.* 97, 77-87 (2002).

28. Golub TR, Slonim DK, Tamayo P *et al.*: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-537 (1999).

•• **The first application of a supervised method to predict cancers based on gene expression patterns.**

29. Hedenfalk I, Duggan D, Chen Y *et al.*: Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* 344, 539-548 (2001).

30. Kihara C, Tsunoda T, Tanaka T *et al.*: Prediction of sensitivity of esophageal tumors to adjuvant chemotherapy by cDNA microarray analysis of gene-expression profiles. *Cancer Res.* 61, 6474-6479 (2001).

31. Allander SV, Nupponen NN, Ringnér M *et al.*: Gastrointestinal stromal tumors with KIT mutations exhibit a remarkably homogeneous gene expression profile. *Cancer Res.* 61, 8624-8628 (2001).

32. Shipp MA, Ross KN, Tamayo P *et al.*: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* 8, 68-74 (2002).

•• **This paper nicely illustrates differences between unsupervised and supervised methods, as applied to the identification of genes relevant to patient outcome.**

33. Armstrong SA, Staunton JE, Silverman LB *et al.*: MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.* 30, 41-47 (2002).

34. Van't Veer LJ, Dai H, Van de Vijver MJ *et al.*: Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530-536 (2002).

35. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z: Tissue classification with gene expression profiles. *J. Comput. Biol.* 7, 559-583 (2000).

36. Ben-Dor A, Friedman N, Yakhini Z: Scoring genes for relevance. *Agilent Tech. Report* (2000) AGL-2000-13:

37. Luo J, Duggan DJ, Chen Y *et al.*: Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res.* 61, 4683-4688 (2001).

38. Xiong M, Li W, Zhao J, Jin L, Boerwinkle E: Feature (gene) selection in gene expression-based tumor classification. *Mol. Genet. Metab.* 73, 239-247 (2001).

39. Blackshaw S, Fraioli RE, Furukawa T, Cepko CL: Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes. *Cell* 107, 579-589 (2001).

40. Cristianini N, Shawe-Taylor J: An introduction to support vector machines: and other kernel-based learning methods. Cambridge University Press, Cambridge (2000).

41. Bishop CM: Neural networks for pattern recognition. Clarendon Press, Oxford (1995).

42. Brown MP, Grundy WN, Lin D *et al.*: Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* 97, 262-267 (2000).

•• **The first application of support vector machines to array data.**

43. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16, 906-914 (2000).

44. Ramaswamy S, Tamayo P, Rifkin R *et al.*: Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA* 98, 15149-15154 (2001).

• **This paper shows how the genes most relevant for the classification can be extracted from support vector machines.**

45. Su AI, Welsh JB, Sapinoso LM *et al.*: Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res.* 61, 7388-7393 (2001).

46. Almeida JS: Predictive non-linear modeling of complex data by artificial neural networks. *Curr. Opin. Biotech.* 13, 72-76 (2002).

47. Gruvberger S, Ringnér M, Chen Y *et al.*: Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.* 61, 5979-5984 (2001).

48. Pollack JR, Perou CM, Alizadeh AA *et al.*: Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* 23, 41-46 (1999).

49. Heiskanen MA, Bittner ML, Chen Y *et al.*: Detection of gene amplification by genomic hybridization to cDNA microarrays. *Cancer Res.* 60, 799-802 (2000).

50. Ross JS, Fletcher JA: The HER-2/neu oncogene: prognostic factor, predictive factor and target for therapy. *Semin. Cancer Biol.* 9, 125-138 (1999).

51. Arteaga CL: The epidermal growth factor receptor: from mutant oncogene in nonhuman cancers to therapeutic target in human neoplasia. *J. Clin. Oncol.* 19, 32S-40S (2001).

52. Ziauddin J, Sabatini DM: Microarrays of cells expressing defined cDNAs. *Nature* 411, 107-110 (2001).

53. Scherf U, Ross DT, Waltham M *et al.*: A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* 24, 236-244 (2000).

54. Staunton JE, Slonim DK, Coller HA *et al.*: Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci. USA* 98, 10787-10792 (2001).

55. Zembutsu H, Ohnishi Y, Tsunoda T *et al.*: Genome-wide cDNA microarray screening to correlate gene expression profiles with sensitivity of 85 human cancer xenografts to anticancer drugs. *Cancer Res.* 62, 518-527 (2002).