LU TP 01-33
April 4, 2002
Revised Version

Analyzing Tumor Gene Expression Profiles

Carsten Peterson[1]

Complex Systems Division, Department of Theoretical Physics
Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden

Markus Ringnér[2]

Cancer Genetics Branch, National Human Genome Research Institute
National Institutes of Health, 50 South Drive MSC 8000
Bethesda, MD 20892, USA

to appear in *Artificial Intelligence in Medicine*

Keywords: artificial neural networks, bioinformatics, diagnostic prediction, drug target identification, genes, microarray

Abstract:

A brief introduction to high throughput technologies for measuring and analyzing gene expression is given. Various supervised and unsupervised data mining methods for analyzing the produced high-dimensional data are discussed. The main emphasis is on supervised machine learning methods for classification and prediction of tumor gene expression profiles. Furthermore, methods to rank the genes according to their importance for the classification are explored. The approaches are illustrated by exploratory studies using two examples of retrospective clinical data from routine tests; diagnostic prediction of small round blue cell tumors of childhood and determining the estrogen receptor status of sporadic breast cancer. The classification performance is gauged using blind tests. These studies demonstrate the feasibility of machine learning based molecular cancer classification.

[1]Corresponding author. Tel.: +46-46-2229002; fax: +46-46-2229686.
E-mail addresses: carsten@thep.lu.se (C. Peterson), markus@thep.lu.se (M. Ringnér)
WWW addresses: http://www.thep.lu.se/complex/, http://research.nhgri.nih.gov/microarray/
[2]Present address: Complex Systems Division, Lund University, Sweden.

1 Introduction

As the sequencing and gene annotation projects of entire genomes of many species are headed towards completion (see *e.g.* [13]), massive mapping efforts in biology are now focused on how the genes interact.

The molecular interactions of genes and gene products underlie fundamental questions of biology. Genetic interactions are, for example, central to the understanding of molecular structure and function, cellular metabolism, development of cells and tissues, and response of organisms to their environments. If such interaction patterns can be measured for various kinds of tissues and the corresponding data can be interpreted, potential clinical benefits are obvious and novel tools for diagnostics, identification of candidate drug targets, and predictions of drug effectiveness for *e.g.* cancer diseases will emerge.

Measuring which genes and gene products are active can be done on two levels. Since genes are composed of a given alphabet (A,T,C,G) with fixed pairing properties A-T and C-G, probes can be constructed that attract gene transcripts extracted from cell tissues and cultures. Using such probes, one can obtain a fingerprint of the gene expression activity in a macroscopic sample. Microarrays are one such tool that allows for the study of expression of thousands of genes simultaneously. For proteins, no obvious such probe technique exists. Rather, one has here to rely upon 2D gel measurements, which in general are quite noisy.

Microarrays can be constructed using gene sequences or their complimentary DNA (cDNA arrays) [24], oligonucleotides synthesized in situ (DNA chips) [19], or genomic sequences [21]. In what follows we limit ourselves to the analysis of gene expression measurements obtained using microarrays with small spots of DNA fixed to glass slides. The cDNA spots are typically 200 microns or less in size, with each slide containing 5000-50000 spots. Levels of gene expression are measured using a preparation of fluorescently labeled tissue RNA (copies of DNA), together with reference RNA labeled with a different fluorochrome, hybridized onto the slides. Typically when data are presented, the fluorescent intensity from the tissue is pseudo-colored red and the intensity from the reference green, and the logarithmic ratio of background corrected red and green intensities for each gene (spot) is subject to analysis (see Fig. 1). Part of a cDNA microarray image is shown in Fig. 2. There exist two major designs of microarray experiments; time series and static ones. In time series experiments, which for many experimental systems are confined to laboratory cell culture experiments (cell-lines), each slide corresponds to a measured time point. In clinical applications, which is of most relevance here, each slide corresponds to a tissue or blood sample, *e.g.* a biopsy. In terms of analysis objectives, in the latter case one aims at relating the measured gene expression to phenotypes, such as diagnosis or drug resistance and in this process determine the most important genes for the questions posed. The

data mining tools employed range from various clustering techniques to supervised learning schemes.

This paper is not intended as a comprehensive review (for a mini-review see *e.g.* [22]). Rather, we very briefly go through the most common methods employed in microarray analysis and then deal with two applications in some detail. The paper is organized as follows. In Section 2 we list the toolboxes with emphasis on a combined multilayer perceptron approach with a principal component analysis preprocessor. Case studies using this perceptron approach, diagnostic prediction of small round blue cell tumors of childhood [16] and determination of estrogen receptor status [10], are presented in Section 3 and a summary and outlook can be found in Section 4.

2 Analysis Tools

2.1 Preprocessing

Prior to applying any data mining processing, one needs to assess, and if necessary correct for, the quality of the data. The simplest and most straightforward preprocessing strategy is to apply cuts on intensities and spot areas (see [5] for a thorough discussion). However, such procedures might remove genes that have low quality measurements for a few samples only, which has unwanted consequences on the entire data set. Remedies for this would involve missing value algorithms, which could be quite elaborate and include user-specific choices and parameters (for a discussion on this see [28]).

More profound and sophisticated corrections for noise have been suggested [14]. Here the signals are decomposed into biological plus other effects, where the latter are modeled and the model is fitted to the data. With relatively few data points ($O(100)$ experiments), as compared to the number of measured genes, this could be dangerous and one might distort or loose the relevant biological signal. In cases, where supervised learning is used for the analysis, one might take a more pragmatic attitude and assume that the calibrated feature models (e.g. multilayer perceptrons (MLP)) implicitly corrects for features that are not related to the relevant biology, but present in the data.

2.2 Clustering and Dimensionality Reduction

Before discussing the algorithms used in the main theme of this review (supervised ones), we briefly list some clustering and dimensional reduction schemes that have proven to be useful in the cDNA microarray analysis context.

- **Hierarchical clustering.** This method is commonly used for pairwise clustering in gene expression space. For example, to reveal sample closeness for static data [20], or to cluster genes with similar behavior in time course experiments [7], thereby zeroing in on groups of functionally related genes.
- Both **K-means clustering** [27] and **self-organizing networks** [26] have been used extensively in similar situations.
- **Reshuffling** [2] is an interesting novel clustering variant, which renumbers data according to similarity from a global standpoint. Here one does not cluster the data into groups but rather renumbers the data points such that similar features appear adjacent in the ordering. This approach is suitable for time course data.
- **Multidimensional scaling (MDS).** This method, which is more aimed at qualitative displays than quantitative analysis, has been frequently used in expression analysis to display samples (see *e.g.* [15]). Briefly, one projects the high-dimensional data points of the different samples onto 2 or 3 dimensions, while preserving the distances in gene expression space between the samples as well as possible. It has also been extended to identify the genes most important for the separation of clusters [3].
- **Principal component analysis (PCA).** With this standard tool, one rotates gene space, such that the variance is dominated by as few linear combinations as possible. Not only can this be a good visualization tool when retaining the 2 or 3 leading directions, but in contrast to MDS an analytic form exists for the transformation. Hence, it can be used as a preprocessing tool [22]; in particular, for supervised learning [16] as will be discussed below.

2.3 Supervised Learning

When categorizing samples into known phenotypes, it is convenient to use supervised approaches. Typically, one then has two goals on the agenda:

1. Develop robust classifiers with validation procedures that successfully handle "blind test" data.
2. Identify the genes most important for the classification.

Hence, when investigating tissues two objectives are simultaneously achieved. One obtains a diagnostic/prognostic tool for the clinic at the same time as insights into the underlying molecular biology are gained.

Limiting the investigation to single gene dependencies, point 2 is easily done using *e.g.* the signal-to-noise statistic [9], where for each gene a classification weight is

computed. The corresponding P value, the probability that the obtained weight can be obtained by chance, is then readily computed using random permutation tests. In these tests, sample labels are randomly permuted and the weight for each gene is computed again. This random permutation of sample labels is performed many times to generate a weight distribution that could be expected under the assumption of random gene expression. The weight values for the actual classification are then assigned P values based on the weight distribution from the random permutations.

For supervised learning that include collective effects among genes, one can pursue two different paths; kernel methods, *e.g.* support vector machines (SVM) [4][8][23][25] and multilayer perceptrons (MLP) [10][16]. They both have their pros and cons. Since cDNA data is very high-dimensional, MLPs generally require some preprocessing to avoid over-fitting, which is not the case for SVMs. On the other hand, the results from MLPs allow for a straightforward probability interpretation and MLPs are more easily generalized to multi-class instances. In what follows, to exemplify the kind of results that can be achieved, we restrict ourselves to MLP methods.

2.4 Multilayer Perceptron (MLP) Models

We next lay out an MLP scheme that has proven to be powerful when classifying tumor cDNA data. In Section 3 two case studies will be discussed using this scheme.

Reducing the dimensionality. In cDNA microarray experiments the dimension of the data (N) is typically several orders of magnitude larger than the number of samples (M). Hence, to allow for a supervised regression model with no over-training, one needs to reduce the dimensionality of the samples. This can be done using PCA. Even though the formal dimension of the problem is given by the number of genes, the effective dimension is just one less than the number of samples. Hence the eigenvalue problem underlying PCA can be solved without diagonalizing $N \times N$ covariance matrices by using singular value decomposition. Thus each sample is represented by M numbers, which are the results from projection of the gene expressions using the M PCA eigenvectors with non-zero eigenvalues. One then keeps the $O(10)$ dominant (corresponding to the largest eigenvalues) projections to represent the expression data. A potential risk when using PCA on relatively few data points is that components might be singled out due to strong noise in the data or to signals not related to the sample categories of interest. Provided the final performance of the method is gauged by an independent test set, one might as an alternative select, by trial-and-error, the most important PCA components.

Architectures and calibration. Simple single hidden layer architectures are in general used with summed square error Langevin updating. The calibrations are monitored both for the training set and a validation set. The latter is not used

for training but for optimizing architecture and training parameters. The resulting weights for a completed training defines a "model". An independent test set is subsequently used to estimate the generalization performance of the model.

Cross-validation and ensemble of models. In order to validate the results, 3-fold cross validation is used. Furthermore, the learning procedure is repeated L times, each with a different random partitioning of the samples into training and validation sets. The training is monitored by measuring the committee output ($3 \times L$ models) of each data point when it appears in the validation set (*i.e.* L models are used).

Sensitivity analysis. The sensitivity S_k of the classification upon different genes (k) is determined by the absolute value of the partial derivative of the outputs (o_i) with respect to the gene expressions (x_k) (see *e.g.* [18]), averaged over outputs, samples M and models.

$$S_k \propto \sum_i \sum_{\text{samples}} \sum_{\text{models}} \left| \frac{\partial o_i}{\partial x_k} \right| \tag{1}$$

A large sensitivity for a gene implies that changing the expression influences the output significantly. In this way the genes can be ranked. The term sensitivity here should not be confused with its usage in the context of classifier performance (see Sect. 3).

α-values. An immediate question that arises is to what extent the rankings based upon S_k are statistically significant? By randomly permuting the class labels of the training data and performing all the steps above, including the sensitivity calculation, for each permutation, one can estimate the probability that a ranked gene would have a larger S_k by chance; the α-value. In a parametric test, *e.g.* t-test, this corresponds to a P value. Similarly, one can obtain P values, corresponding to the probability that any gene has a larger S_k by chance. So far, with sample sizes around 50-100, it appears that 5000-10000 permutations suffice and that $\alpha = 0.01$ is a reasonable cut for regarding a gene as significant for the classification.

2.5 Class Discovery

Supervised learning approaches, such as SVMs and MLPs, to classify human disease states using patterns of gene expression are very promising and they can potentially have great impact on the classification of cancer. However, the advantage with supervised methods, that one can make use of previous knowledge about classes of disease states, restricts their usefulness to investigations where one has previous knowledge. They can in particular not be directly applied to finding new classes of cancer. Of note is that once new classes of cancer have been suggested by unsupervised class discovery methods they can be verified using supervised classification schemes. The field of class discovery based on gene expression patterns is still in an early stage and

great activity is directed towards developing methods for this application. That there typically is an overabundance of genes separating known classes can be exploited to discover classes in gene expression data, by seeking partitions of samples with an overabundance of differentially expressed genes [1]. A slightly simpler approach inspired by the method in [1] has been used to sub-classify familial breast cancer [12]. Briefly, for a given partition of samples into two classes (with n_1 and n_2 samples, respectively) a univariate discriminative weight is calculated for each gene using the signal-to-noise statistic [9]. Random permutation tests are then used as described above to generate a weight distribution that could be expected for two classes with n_1 and n_2 samples under the assumption of random gene expression. Candidate partitions of the data are scored with the number of statistically significant (*e.g.* $P < 0.001$) weights, *i.e.* the number of genes significantly different in expression between samples in the two classes. A simulated annealing [17] scheme is used in which partitions are updated by changing the class of a randomly selected sample, to find the partition of samples into the two classes with the highest score. Gene expression analysis can in this way be used to subset families into more homogeneous groups prior to conventional genetic analysis, and may thereby potentially help in the search for novel breast cancer predisposing genes.

3 Case studies

Below we summarize the application of some of the tools described above on cDNA microarray data for two tumor classification problems. As mentioned above, using cDNA data for such problems has two goals; obtaining a reliable classifier and gaining knowledge about the genes most important for the separation. Many of the datasets used for published microarray results are publicly available. The small round blue cell tumor dataset discussed below is available through our webpages (http://www.thep.lu.se/complex and http://www.nhgri.nih.gov/DIR/microarray).

3.1 Small Round Blue Cell Tumors of Childhood

The small round blue cell tumors (SRBCT) of childhood, including neuroblastoma (NB), rhabdomyosarcoma (RMS), Burkitt's lymphoma (BL) and the Ewing's family of tumors (EWS) exhibit similar appearance on routine histology. However, accurate diagnosis is essential since treatment options, response to therapy and prognosis depend strongly upon the diagnosis. These cancers are difficult to distinguish by light microscopy. Several conventional techniques are utilized in clinical practice to diagnose them. However, no single test exists that can precisely distinguish these cancers.

The cDNA data originate from microarrays containing 6567 genes onto which 63 (training/validation) and 25 (blind test) samples were hybridized. The composition of the 63 samples used for training/validation are shown in Table 1.

As can be seen, the training samples include both tumor biopsy material and cell lines. The latter were added to improve the sample statistics. In principle, this might be dangerous since cell lines may not be representative of the disease *in vivo*, but in this case the underlying disease signals turned out to dominate the picture.

The number of genes was reduced to 2308 after filtering for a minimal level of expression. PCA further reduced the dimensionality of the inputs to 10 projections per sample, using the 10 dominant of the original 88 PCA eigenvectors. The binary encoding for the four possible outputs is also shown in Table 1.

A 3-fold cross-validation procedure was redone 1250 times giving rise to a total of 3750 MLP models, which correctly classified the 63-sample training set in 100% of the cases. When comparing the training with the validation set, all the models performed well based upon manual inspection, and there was no sign of "over-training".

We next determined the classification error with increasing numbers of the ranked genes from Eq. 1 in order of significance. The classification error rate was minimized to 0% for 96 genes, as can be seen in Fig. 3. Using only these 96 genes the models were re-calibrated and again correctly classified all 63 samples – a consistency check. Out of the 96 genes, roughly 50% have not been previously described as associated with these diseases. To be able to reject secondary choices for the classifications of samples as well as test samples that are significantly different from the samples used in training and validation, we proceeded as follows. A squared Euclidean distance between a sample and each disease category was computed between the predicted outputs and an ideal classification, normalized such that it was unity between disease categories. Using 1250 MLP models for each validation sample we constructed for each disease category an empirical probability distribution for the distances. From these distributions we calculated the distance corresponding to the 95 percentile (see Fig. 4), outside which we did not diagnose samples.

The diagnostic classification capabilities of the models were tested on a set of 25 blinded samples. These contained 6 EWS, 7 RMS, 7 NB and 3 BL (mostly tumors with some cell lines), as well as 5 non-SRBCT (muscle tissues and three cancer cell lines). The latter samples were used to test the ability of the models to reject a diagnosis. Using the 3750 models calibrated with the 96 genes, we correctly classified 100% of the 20 SRBCT test samples (see Fig 4). All five of the non-SRBCTs were confidently excluded from any of the four diagnostic categories. The sensitivity of the MLP models for diagnostic classification was for EWS 93%, for RMS 96% and for both NB and BL 100%. The specificity was 100% for all four diagnostic categories. In addition, hierarchical clustering using Pearson correlation and average linkage using

the 96 genes identified from the models correctly classified all 20 of the test samples (see Fig. 5). This is not the case if hierarchical clustering is performed on all the genes. If the sum over outputs (i) is excluded in Eq. 1 the sensitivity can be calculated for each cancer category individually. In this way we also ranked the genes according to their importance for the classification of each category separately.

3.2 Breast Cancer Estrogen Receptor Status

Estrogens are important regulators in the development and progression of breast cancer and regulate gene expression via estrogen receptor alpha (ER), which can be encoded as "on" or "off"; ER+ and ER− respectively. Since approximately two-thirds of all breast cancers are ER+ at the time of diagnosis, the expression of the receptor has important implications for their biology. The ER-status of cells can be measured by standard biochemical methods.

cDNA images of lymph node-negative sporadic breast tumors were studied with respect to ER status using 6768 cDNA clones in [10] with the sample composition shown in Table 2.

The main objective is to see how well the ER status can be predicted from gene expression profiles and to identify the genes involved in this classification. In the latter context it is also interesting to remove top genes from the ranking list to investigate how "deep" the ER pathways are, *i.e.* how many other genes that are affected by the ER status.

To this end, we proceeded very much like in the SRBCT case above. After filtering 3389 genes remained. The 8 largest PCA components were kept for further processing. MLP models with 8 inputs, 4 hidden units and 1 output unit were calibrated with 3-fold cross validations performed 200 times yielding a total of 600 models. Thereafter, sensitivities (Eq. 1) were computed and the genes ranked accordingly.

With the extracted top 100 genes forming the input for another and final calibration, all 47 samples were correctly classified in the validation phase. The output values from the MLP committee is shown for all (training/validation) 47 samples in Fig. 6. The majority of the samples, in both groups, obtained output values close to either 0 or 1, with little variance between the output results from the different models – a clear separation between ER+ and ER− tumors. An interesting issue is to what extent ER+ and ER− tumors can be separated when excluding top genes on the ranking list. To test this, a series of classifications using different sets of 100 genes was done, starting from the top of the discriminator list by excluding the top 50 genes and following this by the stepwise exclusion of 50 additional genes for every classification (i.e. excluding the top 50, 100, 150, 250, 300 genes, respectively). The number of correctly classified samples and the Receiver Operator Characteristic (ROC) [11]

area for the predictions of both the 47 tumors in the validation set as well as the 11 blind test tumors are shown in Table 3. The ROC area is the probability that for a randomly chosen pair of samples, one belonging to and one not belonging to a category, the one belonging to the category is the one with the closest distance to the ideal for that particular category.

Although the success of the predictions declined when using genes lower down on the discriminator list, the network performance was still fairly good (see Table 3 and Fig. 7). Hence, the ER has a "deep" pathway and its status more or less defines a very distinct and broad genotype. Using a classifier together with a sensitivity measure to establish such a feature is an appealing and useful approach.

4 Summary and Outlook

A brief introduction to analysis of cDNA microarray data has been given, with emphasis on tumor classification methods.

Using PCA as a preprocessor to committees of MLP models turns out to provide powerful classifiers. When applying a sensitivity measure to the calibrated models, valuable insights into the underlying biology can be obtained.

In two case studies, small round blue cell tumors (SRBCT) of childhood and the estrogen receptor (ER) status of breast cancer, impressive performance on "blind test" sets were obtained. In addition, novel genes in this context were obtained for SRBCT and remarkable pathway depths were discovered for ER.

The studies above should be considered as "first generation" ones with limited sample sizes. Yet, the results are convincing. With $O(1000)$ samples, it should be feasible to further disentangle the diseases and, very importantly, to predict clinical outcomes for various therapies. Using microarray technology together with advanced data mining tools is likely to be routine clinical practice in a not too distant future.

The focus has been on clinical situations, which are limited to static data. Hence, from a biology standpoint only correlation type conclusions can be drawn. Understanding causal dependencies among the genes requires time course measurements, which is feasible with cell lines. For model systems like yeast such investigations are now quite mature (see *e.g.* [6]).

Acknowledgments

This work was in part supported by the Swedish Foundation for Strategic Research, the Swedish Research Council and the Knut and Alice Wallenberg Foundation through the SWEGENE consortium.

References

[1] A. Ben-Dor, N. Friedman, Z. Yakhini. Class discovery in gene expression data. In: T. Lengauer, D. Sankoff, S. Istrail, P. Pevzner and M. Waterman, eds., *Proc 5th Annual International Conference on Computational Molecular Biology (RECOMB)*. ACM Press, New York, 2001.

[2] S. Bilke. Shuffling yeast gene expression data, LU TP 00-18 2000, http://www.thep.lu.se/complex/publications.html.

[3] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, N. Hayward, J. Trent. Molecular classification of cutaneous malignant melanoma by gene expression profiling, *Nature* 2000;406:536-540.

[4] M.P. Brown, W.N. Grundy, D. Lin, N. Cristianini, C. Walsh Sugnet, T.S. Furey, M. Ares, Jr., D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proc. Natl. Acad. Sci. USA* 2000;97:262-267.

[5] Y. Chen, V. Kamat, E.R. Dougherty, M.L. Bittner, P.S. Meltzer, J.M. Trent. Ratio statistics of gene expression levels and applications to microarray data analysis, *Bioinformatics* 2002;18:1207-1215.

[6] F. Devaux, P. Marc, C. Jacq. Transcriptome, transcription activators and microarrays, *FEBS Letters* 2001:498:140-144.

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

[8] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics* 2000;16:906-914.

[9] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.D. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* 1999;286:531-537.

[10] S. Gruvberger, M. Ringnér, Y. Chen, S. Panavally, L.H. Saal, Å. Borg, M. Fernö, C. Peterson, P.S. Meltzer. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns, *Cancer Res.* 2001;61:5979-5984.

[11] J.A. Hanley, B.J. McNeil. The meaning and use of the area under the receiver operating characteristic (ROC) curve, *Radiology* 1982;143:29-36.

[12] I. Hedenfalk, M. Ringnér, A. Ben-Dor, Z. Yakhini, Y. Chen, G. Chebil, R. Ach, H. Olsson, N. Loman, P.S. Meltzer, Å. Borg, J. Trent. Molecular classification of familial non-*BRCA1/2* Breast Cancer, submitted to *Proc. Natl. Acad. Sci. USA*.

[13] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome, *Nature* 2001;409:860-921.

[14] M.K. Kerr, G.A. Churchill. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments, *Proc. Natl. Acad. Sci. USA* 2001;98:8961-8965.

[15] J. Khan, R. Simon, M. Bittner, Y. Chen, S.B. Leighton, T. Pohida, P.D. Smith, Y. Jiang, G.C. Gooden, J.M. Trent, P.S. Meltzer. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays, *Cancer Res.* 1998;58:5009-5013.

[16] J. Khan, J.S. Wei, M. Ringnér, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Atonescu, C. Peterson, P.S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nat. Med.* 2001;7:673-679.

[17] S. Kirkpatrick, C. Gelatt, M. Vecchi. Optimization by simulated annealing, *Science* 1983;220:671-680.

[18] P.J.G. Lisboa, A.R. Mehridehnavi. Sensitivity methods for variable selection using the MLP. In: *International Workshop for Neural Networks for Identification, Control, Robotics and Signal Processing, Venice*. IEEE-Computer Society Press, 1996.

[19] D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, E.L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nat. Biotechnol.* 1996;14:1675-1680.

[20] C.M. Perou, T. Sørlie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, H. Johnsen, L.A. Akslen, Ø. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P.E. Lønning, A-L. Børresen-Dale, P.O. Brown, D. Botstein. Molecular portraits of human breast tumours, *Nature* 2000;406,747-752.

[21] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W-L. Kuo, C. Chen, Y. Zhai, S.H. Dairkee, B-M. Ljung, J.W. Gray, D.G. Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays, *Nat. Genet.* 1998;20:207-211.

[22] J. Quackenbush. Computational analysis of microarray data, *Nat. Rev. Genet.* 2001;2:418-427.

[23] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, T.R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures, *Proc. Natl. Acad. Sci. USA* 2001;98:15149-15154.

[24] M. Schena, D. Shalon, R.W. Davis, P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science* 1995;270:467-470.

[25] A.I. Su, J.B. Welsh, L.M. Sapinoso, S.G. Kern, P. Dimitrov, H. Lapp, P.G. Schultz, S.M. Powell, C.A. Moskaluk, H.F. Frierson, Jr., G.M. Hampton. Molecular classification of human carcinomas by use of gene expression signatures, *Cancer Res.* 2001;61,7388-7393.

[26] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, T.R. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and applications to hematopoietic differentiation, *Proc. Natl. Acad. Sci. USA* 1999;96:2907-2912.

[27] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho and G.M. Church. Systematic determination of genetic network architecture, *Nat. Genet.* 1999;22:281-285.

[28] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman. Missing value estimation methods for DNA microarrays, *Bioinformatics* 2001;17:520-525.

	Category		Samples	MLP Encoding
BL	Burkitt's Lymphoma		8	1000
EWS	Ewing's Sarcoma	tumor	13	0100
		cell-line	10	
NB	Neuroblastoma		12	0010
RMS	Rhabdomyosarcoma	tumor	10	0001
		cell-line	10	
	Total:		63	
	Blind Tests[a]**:**		25	

[a]The 25 blind tests were provided after completed calibrations.

Table 1: Composition of the SRBCT data set.

Category	Samples
ER+	23
ER−	24
Total:	47
Blind Tests:	11

Table 2: Composition of the estrogen receptor status data set.

Genes	Correct[a]	ROC Area
1-100	11	100.0%
51-150	9	100.0%
101-200	11	100.0%
151-250	9	100.0%
201-300	11	100.0%
251-350	9	93.3%
301-400	8	96.7%
Random[b]	5.5	53.0%

[a]Out of a total of 11.

[b]100 random genes from top 401-3389.

Table 3: Number of correctly classified blind test samples.

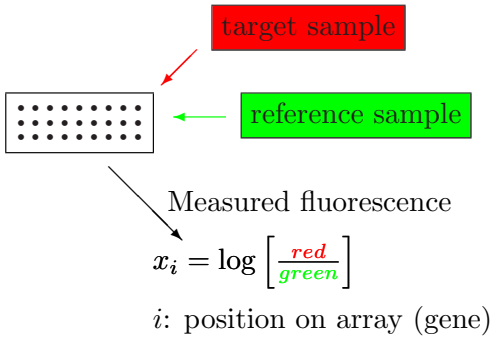

Figure 1: The cDNA microarray preparation process.

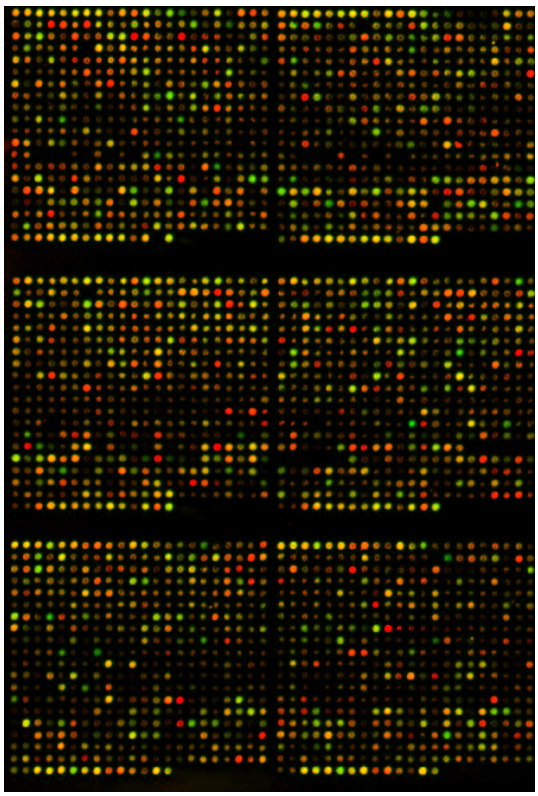

Figure 2: Example of (part of) a cDNA microarray slide. Red, yellow and green indicate over-, similar and under-expression of genes as compared to a reference sample.

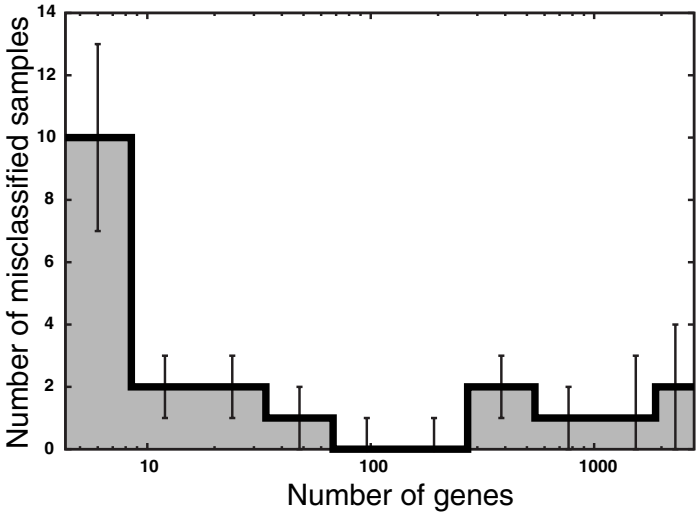

Figure 3: Number of mis-classified samples as a function of the number of top ranked genes used in the analysis. Reproduced with permission from [16] ©2001 Nature Publishing Group.

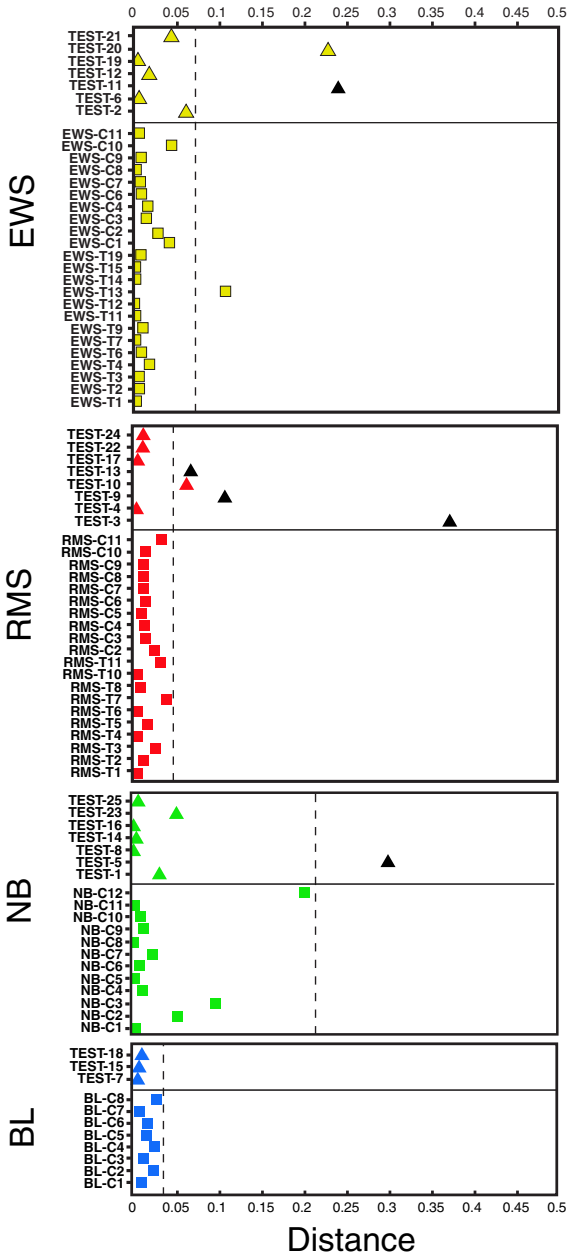

Figure 4: Distances from outputs for validation (squares) and test set (triangles) samples to ideal classifications. Each category is shown separately and the validation and test samples are separated by horisontal lines. Vertical lines denote the 95% percentile outside which samples are not diagnosed and black triangles represent the 5 non-SRBCT samples. Reproduced with permission from [16] ©2001 Nature Publishing Group.

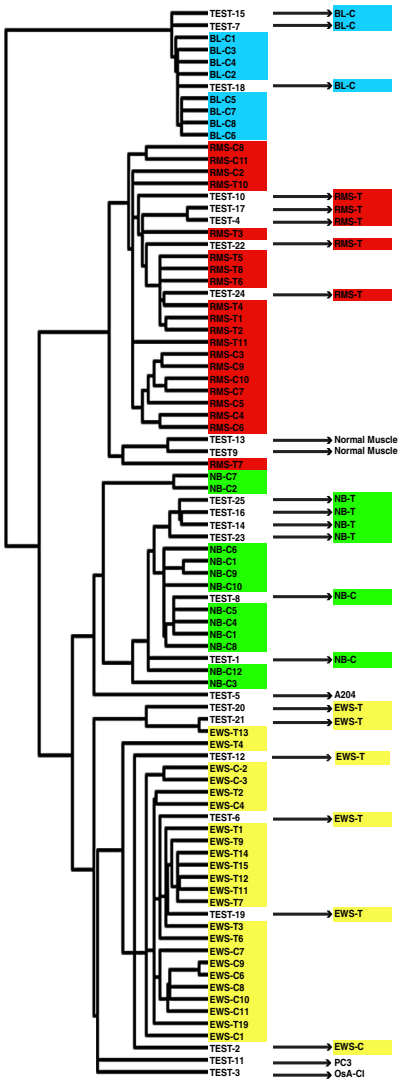

Figure 5: Hierarchical clustering of 63 SRBCT and 25 "blind test" samples using the top 96 genes. Reproduced with permission from [16] ©2001 Nature Publishing Group.

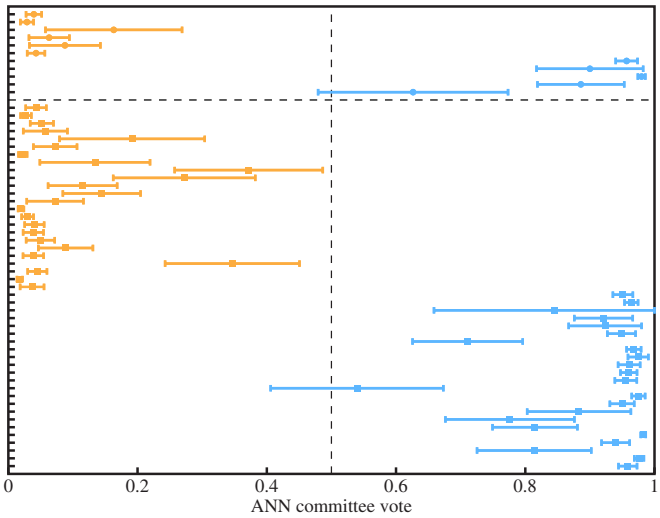

Figure 6: Committee output values for ER classification of 47 validation and 11 test samples (separated by the horisontal line), using genes 1-100 on the ranking list. ER− samples are colored yellow and ER+ blue. Reproduced with permission from [10] ©2001 AACR.

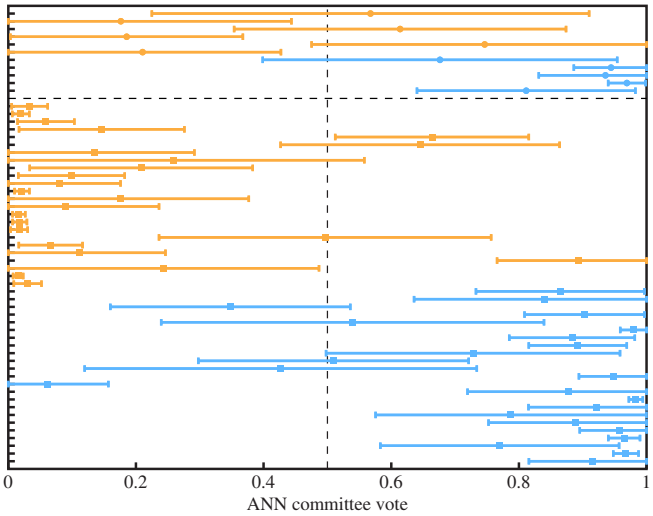

Figure 7: Committee output values for ER classification of 47 validation and 11 test samples (separated by the horisontal line), using genes 301-400 on the ranking list. ER− samples are colored yellow and ER+ blue. Reproduced with permission from [10] ©2001 AACR.