Microarray-Based Cancer Diagnosis with Artificial Neural Networks

Markus Ringnér and Carsten Peterson
Complex Systems Division, Department of Theoretical Physics, Lund University, Sweden

BioTechniques 34:S30-S35 (March 2003)

ABSTRACT

In recent years, the advent of experimental methods to probe gene expression profiles of cancer on a genome-wide scale has led to widespread use of supervised machine learning algorithms to characterize these profiles. The main applications of these analysis methods range from assigning functional classes of previously uncharacterized genes to classification and prediction of different cancer tissues. This article surveys the application of machine learning algorithms to classification and diagnosis of cancer based on expression profiles. To exemplify the important issues of the classification procedure, the emphasis of this article is on one such method, namely artificial neural networks. In addition, methods to extract genes that are important for the performance of a classifier, as well as the influence of sample selection on prediction results are discussed.

INTRODUCTION

The possibility to simultaneously record gene expression levels of thousands of genes using DNA arrays (1,2) has led to new ways of looking at organisms on a genome-wide scale. In particular, it is now possible to monitor gene expression of thousands of genes in cancer (3). DNA array experiments involve measuring tens of thousands of gene expression levels under different conditions. In addition to allowing for a quick scan for interesting single genes, such measurements provide insights into collective effects among the gene expressions.

Clustering methods have been used extensively to investigate array data along the two dimensions, conditions, and genes, with two aims: (*i*) either to cluster the conditions (tissue type, phenotype, treatment response, patient outcome, etc.) based on expression levels regarded as their molecular portraits (4); or (*ii*) to classify genes with expression patterns that are correlated across various conditions (5). The clustering of conditions has been applied to molecular classification of cancer in several studies (6–10). If there is some prior knowledge about the classes to be analyzed, supervised methods can be used and may be advantageous to unsupervised clustering methods (for a discussion, see Reference 11). In a supervised method, features of classes are extracted in a training process in order to learn how to identify and utilize them for classification. Our goal in this article is to describe how supervised machine learning algorithms have been used to learn to classify cancers based on features of their gene expression profiles.

MACHINE LEARNING APPROACHES FOR ARRAY ANALYSIS

Supervised methods that have been applied to molecular classification of cancer tissues include correlation-based classification methods (8), artificial neural networks (ANNs) in the form of supervised layered perceptrons (12), and support vector machines (SVMs) (13).

In the correlation-based methods, various statistical measures are used to correlate, gene-by-gene, expression levels with a condition of interest (tissue type, phenotype, treatment response, patient outcome, etc.). In this way, a discriminatory weight is calculated for each gene. Once the genes are ranked according to the discriminatory weights, supervised classifiers can be constructed using only the top-ranked genes. To classify samples, each gene typically gives a flat vote or a vote weighted by some function of the gene's expression levels. For example, Golub et al. used a signal-to-noise statistic to discriminate acute myeloid leukemia from acute lymphoblastic leukemia (8), and van't Veer et al. used Pearson's correlation to predict clinical outcome of breast cancers (14).

ANNs and SVMs have, in contrast to correlation-based methods, the potential to include collective and nonlinear effects among the genes. Both methods are computer-based algorithms that are capable of learning to identify complex patterns, such as gene expression data, in a training process. Once trained, the parameters of the ANN or SVM can give relevant information on the relative importance of each gene in the learning of the classes (12,15). Both SVMs (16) and ANNs (17) have been applied to assigning functional annotations to uncharacterized genes based on expression signatures of genes belonging to existing functional classes. Furey et al. used SVMs primarily to classify ovarian tissue samples based on their gene expression profiles (13). SVMs have also been applied to diagnose multiple common adult malignancies (15,18). ANNs have been applied to molecular classification of small round blue cell tumors (12), breast cancer (19), esophageal cancer and premalignancy (20), and colorectal lesions (21). To exemplify how machine learning methods can be used to analyze array data, we will next outline how ANNs can be used for molecular classification of cancers based on their gene expression profiles.

CLASSIFICATION OF TUMORS USING ANNs

In its simplest form, so-called perceptrons (22), ANNs are just simple linear logistic regression methods. More general ANNs in the form of a multilayer perceptron (MLP) have proven to be powerful when classifying tumor array-based expression data (19) and also for applications in other biological systems (for a review see Reference 23). An MLP consists of a set of layers of units inspired by the structure and behavior of neurons in the mammal brain (Figure 1). 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 node in the output layer represents a class. The simple perceptron corresponds to the case with no hidden layer. When a gene expression pattern of a sample is fed into the MLP, ideally only the output node representing the class that the sample belongs to should respond. For calibration, samples belonging to the classes of interest are presented to the MLPs, which are trained to recognize them in a supervised fashion by a process of error minimization; the parameters of the MLPs are adjusted such that error of the output units are minimized for the calibration samples. The number of nodes in the input layer is equal to the dimension of the input data. Calibrating an MLP follows the golden rule of any data fitting process; one needs more samples than the number of tunable parameters in the model. Since the number of inputs typically far exceeds the number of samples in cancer profiling studies, one has a potential problem, and there is a risk of overfitting. There are two possible solutions to this problem. First, the dimension of the data can be reduced, either by using a dimensional reduction algorithm, such as principal component analysis (PCA) (24), or by selecting a smaller set of genes as input to the classifier in a supervised way by using a discriminatory score (see e.g., Reference 8). Second, the learning process can be carefully monitored using a cross-validation scheme to avoid overtraining (12). 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, that can be used as a committee to classify test samples in a robust way (12). Typically, both measures are taken; dimensional reduction and cross-validation. After completed calibration of the ANN models, genes are ranked according to their importance for the classification. One can then remove genes from the bottom of the list to identify the optimal number of genes with respect to calibration set performance. With this procedure, two goals are achieved; one obtains an optimized classifier for diagnostic prediction with few genes, and the other singles out genes important for discovering novel biology. We next illustrate the method with two cases: (*i*) the diagnostic prediction of small round blue cell tumors (12); and (*ii*) the determination of estrogen receptor (ER) status in sporadic breast cancer (19).

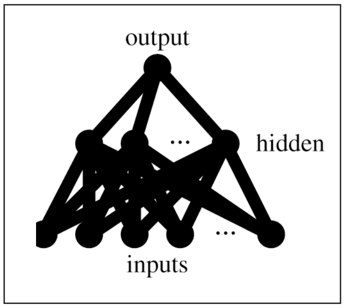

Figure 1. An MLP. Input data, in this case from gene expression measurements, is fed into a so-called input layer, triggering a response in the hidden units that subsequently results in a signal in the output layer. Each link corresponds to a parameter that is tuned such that the output signal matches labeled classes describing the measured samples (classification).

In Reference 12, ANNs are used for classification and diagnostic prediction of small round blue cell tumors, which belong to four different diagnostic categories and are based on data from cDNA microarrays containing 6567 genes. 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. A calibration material of 63 tumors and cell lines was used for 3-fold cross-validation training of 3750 ANNs, each of which were trained on a different subset of the samples to provide for a robust and diverse committee of ANNs, while gauging the performance of the committee vote on the validation sets. The procedure was repeated for different sets of ranked genes, where these had been successively selected from the top of the ranking lists which resulted from each committee calibration. The resulting performance, in terms of error in classification, is shown in Figure 2. As can be seen, optimal performance is obtained using 96 genes. The ANNs classified all of the 63 samples correctly. An interesting question is, to what extent could such classification results have been obtained by chance. To probe this issue, the four tumor classes are randomly relabeled, and the procedure above is repeated 1000 times for different random permutations of the sample labels. The resulting performance distribution is shown in Figure 3. As can be seen, the tail from the randomly relabeled samples is far from the predictions for the original sample labels (the vertical line indicating all 63 tumors correctly identified). The committee of models are then used on 25 blinded test samples that are provided after the calibration is frozen. It turned out that 20 of these belonged to the four categories, whereas the other 5 represented other tissues. Not only were the trained ANNs able to correctly classify the 20 tumors, but very importantly, they also rejected tissues not belonging to any of the four classes. This is possible since the signals from the ANN output nodes need not be binary, but should, for a correct classification, be close to its encoding binary value. Hence, a distance measure from the encoding value can be used to reject samples that do not confidently belong to any of the encoded groups. In terms of interpreting the underlying biology from the ranked genes, one should mention that, as an example in Reference 12, FGFR4, which is a tyrosine kinase receptor, was found to be highly expressed in rhabdomyosarcoma. This finding was confirmed at the protein level and is of therapeutic interest because of the potential role of FGFR4 in tumor growth and in prevention of terminal muscle differentiation (12). Regarding the gene rankings, it is of course possible to rank the genes individually for each patient, thereby opening up the possibility for customized treatment.

Gruvberger et al. (19) used ANNs to investigate the phenotype associated with ER status in human breast cancer and found that the ANNs could accurately classify the 47 tumors into ER-positive and ER-negative samples. The data analyzed was from cDNA microarrays containing 6728 clones. Again, PCA preprocessing was employed to reduce the number of input variables followed by training a committee of 1250 models using 3-fold cross-validation. Also, the genes were ranked. It was found that the ANNs could accurately predict ER status even when excluding top discriminator genes, including ER and *GATA3*, for the validation, as well as a blinded test set of 11 tumors. In fact, by systematically removing genes from the top of the list, predictive power was retained down to removing more than 10% of the top genes. In Table 1, the performance of the ANNs is shown for different stages of removing genes from the top. The results are presented in terms of correctly classified test set samples and the corresponding receiver operating characteristics (ROC) areas (25). These results provided evidence that ER-positive and ER-negative tumors display remarkably different gene expression phenotypes. This procedure of establishing different distinct gene expression programs could be useful in studying cancer diseases, in which many phenotypes are involved prior to dividing up the material for diagnostic prediction studies.

Table 1. Prediction of ER Status

Genes	Correct[a]	ROC area[b]
Top-100	11	100%
51–150	9	100%
101–200	11	100%
151–250	9	100%
201–300	11	100%
251–350	9	93%
301–400	8	97%
Random[c]	5.5	53%

[a] Number of correct classifications of 11 test samples.

[b] The ROC area is identical to another more intuitive measure: the probability that the ER-positive sample in a randomly chosen pair of samples, one being ER-positive and one ER-negative, is classified with the output value closest to the ER-positive category. Thus, if the ROC area is 100%, it is possible to define a cutoff for the ANN output, such that all samples are correctly classified.

[c] Results for 100 randomly picked genes among top 401–3400.

OUTCOME PREDICTION AND THE INFLUENCE OF SAMPLE SELECTION

An application of array analysis of cancer that attracts a lot of interest is to predict which tumors will recur. In general, tumors detected early have good clinical outcomes, but there are no markers that will correctly predict if or when a tumor will recur. A consequence of this is that most cancer patients receive intensive treatments, often associated with severe side-effects, with the hope of achieving a positive outcome.

MacDonald et al. identified a gene expression profile that could classify if a medulloblastoma would metastasize or not (26). Shipp et al. developed a supervised classifier that could separate cured vs. refractory disease based on expression profiles of diffuse large B cell lymphoma (27). For breast cancer, there have been studies to investigate lymph node status (28) and whether a tumor would develop metastases (14). The determination of lymph node involvement is the most important factor in disease outcome for breast cancer. West et al. managed to identify a gene expression profile that could classify their samples according to whether they, at the time of diagnosis, had metastasized to the lymph nodes or had not spread beyond the breast (28). van't Veer et al. developed a classifier that could separate their samples into those that developed metastases in less than 5 years and those that did not (14). The prediction accuracy of their supervised classifier was 80%–90% in a cross-validation scheme and for a blinded test set.

These studies imply that a gene expression profile present in tumors at the time of diagnosis can predict the following course

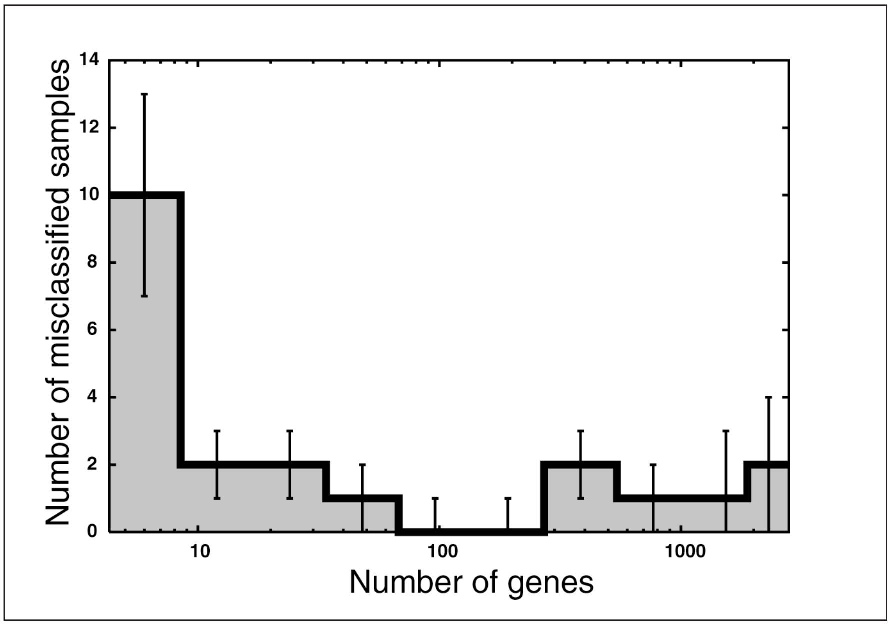

Figure 2. Minimizing the number of genes. The average number of misclassified samples for all models is plotted against the number of genes used in the classifier. As can be seen, using the 96 highest ranked genes results in zero misclassifications for this example. Reprinted and adapted with permission from Reference 12. ©2001 Nature Publishing Group.

of the cancer. Once these profiles are sufficiently characterized, they can be measured for primary tumors with the aim of using them in clinical decisions. Needless to say for this latter aim, one needs, as for any clinical procedure, to demonstrate sufficient sensitivity, specificity, reproducibility, robustness, and reliability. A key point here is a thorough characterization of these profiles. The studies have so far been carried out on relatively small sample sets. In the sample cohort studied by van't Veer et al., the predictor based on a set of genes with expression correlated with outcome performed better, but not significantly better, than predictors based on standard histopathological and clinical prognostic factors (14). Gruvberger et al. applied the part of the gene set used to predict outcome by van't Veer available on their own arrays on a different cohort of samples and failed to get significant outcome prediction results (29). There were important differences in the two sample cohorts. In the sample set studied by Gruvberger et al., standard prognostic markers were matched between samples in the two groups having good and bad outcome. This meant that a classifier based on standard prognostic factors could not predict outcome for these samples, in contrast to the results for the sample cohort studied by van't Veer et al. In addition, there were differences in the treatments received by the patients in the two studies. Thus, it seems that care should be taken in sample selection when defining a gene expression profile associated with outcome for breast cancers.

It should be noted that consequences of features characteristic of a specific sample set and not of a larger cohort cannot be compensated for by using a more sophisticated analysis method. In applications of any supervised method to predict disease phenotypes, the influence of the sample selection and the importance of an independent test set should be emphasized. As larger cohorts of samples are studied, we will see if expression profiles present in tumors at the time of diagnosis will become better characterized and allow for expression signatures to add to standard factors in decisions to direct treatment. In particular, age, disease stage, and other patient characteristics should be representative of the spectrum of patients. Furthermore, many of the microarray-based studies have so far analyzed patients that have undergone different therapies, which may confound outcome expression signatures.

OUTLOOK

Array-based technologies hold great promise for the diagnosis and prognosis of disease outcomes. As with any microarray application, care has to be exercised in all steps of the experiment and analysis, including: (*i*) sample selection; (*ii*) experimental design; (*iii*) proper preprocessing and normalization of the raw data (30); and (*iv*) good practice when it comes to employing appropriate data mining methods, including validation procedures. Recently, supervised machine learning classification techniques have been used successfully in studies in which there are data to guide the analysis. In the case of cancer diagnoses, one often has such data. The aim is to use available data such as patient survival, treatment response, etc., to train an algorithm to recognize patterns in the expression profiles that can be utilized to diagnose and make predictions for unknown cases. For these applications, supervised methods, such as ANNs, are proving to be more effective than unsupervised clustering methods. Nevertheless, it seems, as has often been the case in clinical research, that issues of sample selection are at least as critical as the choice of analysis method for array-based approaches to fulfill their promise for clinical applications.

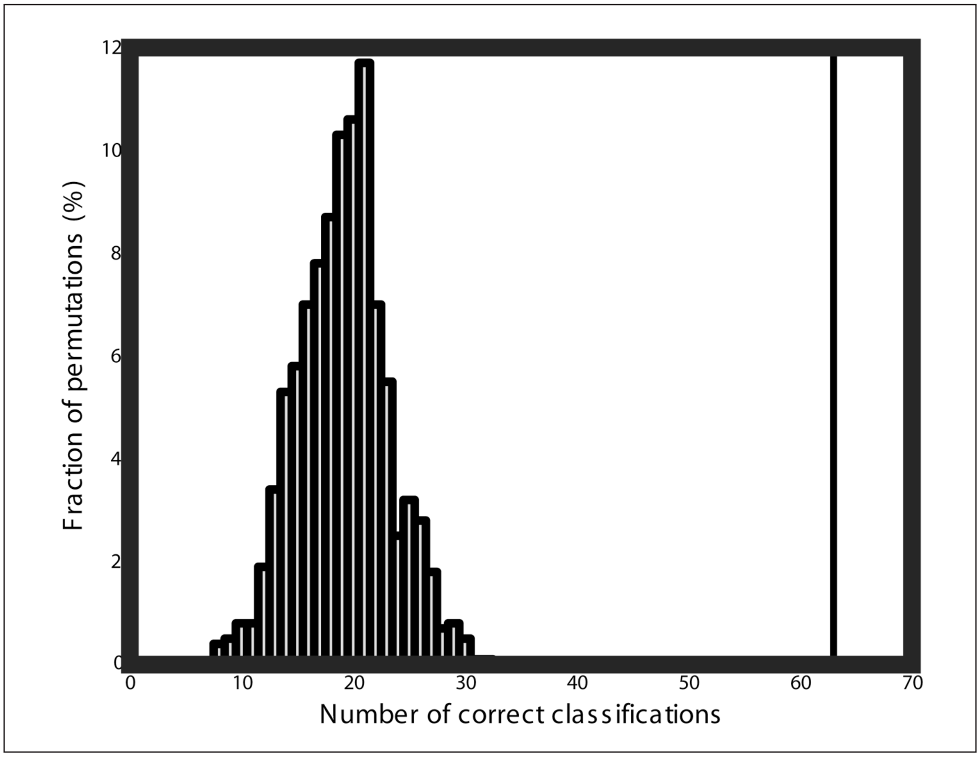

Figure 3. Random permutation test. Validation results for randomly permuted sample labels (disease category) using a committee of ANNs from a 3-fold cross-validation scheme. The number of correctly classified samples is histogrammed for the random permutations. Typically, 20 samples are correctly classified for a random permutation, whereas all 63 samples are correct for the diagnostic categories (vertical line). Reprinted with permission from Reference 31. ©2002 Kluwer Academic Publishers.

ACKNOWLEDGMENTS

We would like to thank P. Edén, S. Gruvberger, P. Johansson, J. Khan, and P. Meltzer for valuable discussions. This work was in part supported by the Knut and Alice Wallenberg Foundation through the SWEGENE consortium.

REFERENCES

1. **Lockhart, D.J., H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, et al.** 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol. *14*:1675-1680.
2. **Schena, M., D. Shalon, R.W. Davis, and P.O. Brown.** 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*:467-470.
3. **DeRisi, J., L. Penland, P.O. Brown, M.L. Bittner, P.S. Meltzer, M. Ray, Y. Chen, Y.A. Su, et al.** 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat. Genet. *14*:457-460.
4. **Khan, J., R. Simon, M. Bittner, Y. Chen, S.B. Leighton, T. Pohida, P.D. Smith, Y. Jiang, et al.** 1998. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. Cancer Res. *58*:5009-5013..
5. **Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein.** 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *95*:14863-14868.
6. **Alizadeh, A.A., M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, et al.** 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*:503-511.
7. **Bittner, M., P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, et al.** 2000. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature *406*:536-540.
8. **Golub, T.R., D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, et al.** 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*:531-537.
9. **Hedenfalk, I., D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, et al.** 2001. Gene-expression profiles in hereditary breast cancer. N. Engl. J. Med. *344*:539-548.
10. **Perou, C.M., T. Sorlie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, et al.** 2000. Molecular portraits of human breast tumours. Nature *406*:747-752.
11. **Ringnér, M., C. Peterson, and J. Khan.** 2002. Analyzing array data using supervised methods. Pharmacogenomics *3*:403-415.
12. **Khan, J., J.S. Wei, M. Ringnér, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, et al.** 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. *7*:673-679.
13. **Furey, T.S., N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler.** 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics *16*:906-914.
14. **van ‘t Veer, L.J., H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, et al.** 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*:530-536.
15. **Ramaswamy, S., P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, et al.** 2001. Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA. *98*:15149-15154.
16. **Brown, M.P., W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares, Jr., and D. Haussler.** 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. USA *97*:262-267.
17. **Mateos, A., J. Dopazo, R. Jansen, Y. Tu, M. Gerstein, and G. Stolovitzky.** 2002. Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. Genome Res. *12*:1703-1715.
18. **Su, A.I., J.B. Welsh, L.M. Sapinoso, S.G. Kern, P. Dimitrov, H. Lapp, P.G. Schultz, S.M. Powell, et al.** 2001. Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res. *61*:7388-7393.
19. **Gruvberger, S., M. Ringnér, Y. Chen, S. Panavally, L.H. Saal, Å. Borg, M. Fernö, C. Peterson, et al.** 2001. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. Cancer Res. *61*:5979-5984.
20. **Xu, Y., F.M. Selaru, J. Yin, T.T. Zou, V. Shustova, Y. Mori, F. Sato, T.C. Liu, et al.** 2002. Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett’s esophagus and esophageal cancer. Cancer Res. *62*:3493-3497.
21. **Selaru, F.M., Y. Xu, J. Yin, T. Zou, T.C. Liu, Y. Mori, J.M. Abraham, F. Sato, et al.** 2002. Artificial neural networks distinguish among subtypes of neoplastic colorectal lesions. Gastroenterology *122*:606-613.
22. **Rosenblatt, F.** 1958. The perceptron: a probabilistic model for information storage in the brain. Psychol. Rev. *65*:386-407.
23. **Almeida, J.S.** 2002. Predictive non-linear modeling of complex data by artificial neural networks. Curr. Opin. Biotechnol. *13*:72-76.
24. **Joliffe, I.T.** 1986. Principal Component Analysis. Springer, New York.
25. **Hanley, J.A. and B.J. McNeil.** 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology *143*:29-36.
26. **MacDonald, T.J., K.M. Brown, B. LaFleur, K. Peterson, C. Lawlor, Y. Chen, R.J. Packer, P. Cogen, et al.** 2001. Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease. Nat. Genet. *29*:143-152.
27. **Shipp, M.A., K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C. Aguiar, M. Gaasenbeek, M. Angelo, et al.** 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat. Med. *8*:68-74.
28. **West, M., C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J.A. Olson, Jr., et al.** 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc. Natl. Acad. Sci. USA *98*:11462-11467.
29. **Gruvberger, S.K., M. Ringnér, P. Edén, Å. Borg, M. Fernö, C. Peterson, and P.S. Meltzer.** 2003. Expression profiling to predict outcome in breast cancer: the influence of sample selection. Breast Cancer Res. *5*:23-26.
30. **Quackenbush, J.** 2002. Microarray data normalization and transformation. Nat. Genet. *32(Suppl 2)*:496-501.
31. **Ringnér, M., P. Edén, and P. Johansson.** 2002. Classification of expression patterns using artificial neural networks, p. 201-215. *In* D.P. Berrar, W. Dubitzky, and M. Granzow (Eds.), A Practical Approach to Microarray Data Analysis. Kluwer Academic Publishers, Boston.

Address correspondence to:

Dr. Carsten Peterson
Complex Systems Division
Department of Theoretical Physics
Sölvegatan 14A
SE-223 62 Lund, Sweden
e-mail: carsten@thep.lu.se