

# Supervised classification for gene network reconstruction

L.A. Soinov<sup>1</sup>

Microarray Informatics Group, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, U.K.

## Abstract

One of the central problems of functional genomics is revealing gene expression networks – the relationships between genes that reflect observations of how the expression level of each gene affects those of others. Microarray data are currently a major source of information about the interplay of biochemical network participants in living cells. Various mathematical techniques, such as differential equations, Bayesian and Boolean models and several statistical methods, have been applied to expression data in attempts to extract the underlying knowledge. Unsupervised clustering methods are often considered as the necessary first step in visualization and analysis of the expression data. As for supervised classification, the problem mainly addressed so far has been how to find discriminative genes separating various samples or experimental conditions. Numerous methods have been applied to identify genes that help to predict treatment outcome or to confirm a diagnosis, as well as to identify primary elements of gene regulatory circuits. However, less attention has been devoted to using supervised learning to uncover relationships between genes and/or their products. To start filling this gap a machine-learning approach for gene networks reconstruction is described here. This approach is based on building classifiers – functions, which determine the state of a gene's transcription machinery through expression levels of other genes. The method can be applied to various cases where relationships between gene expression levels could be expected.

## Introduction

So far, great effort has been put into various applications of supervised learning to the problems of discrimination between different sample classes or experimental conditions based on corresponding expression profiles. Multi-class cancer recognition [1,2], disease diagnosis [3,4] and prediction of treatment/survival outcome [5,6] are among examples of successful applications of supervised techniques. Learning functional classification of uncharacterized genes or proteins represents another area for use of supervised methodologies [7–10].

All the studies above aim to discover genes and their characteristic expression parameters, which help to discriminate between object classes. Therefore, all of them fall into the category of data mining methods for determining whether or not a previously unobserved example belongs to a known predefined class. One of the earliest and most distinctive examples of such applications is the study by Golub et al. [2], in which the authors generally considered two problems: the problem of identification of new cancer classes (class discovery) and the problem of assignment of new samples to known classes (class prediction). For the first problem they used self-organizing maps, a clustering method [11,12], while a supervised classification procedure resembling weighted voting was used for the second.

Although in many cases, and the above examples demonstrate this quite well, classification models do not have to be interpretable to be useful [13], there is a different point of view that it would be productive to construct the learning process as step-by-step generation of essential descriptive hypotheses, which can be verified by and compared with conventional knowledge, e.g. with the facts reported in literature.

## Learning, definitions and terms

In this review the applications of machine learning methods are considered only in the context of gene expression networks reconstruction from microarray data. However, in many respects, this study is relevant also for understanding protein–protein interaction networks as well as metabolic networks that represent the major components of the biochemical network.

We start with definitions and concepts in machine learning and some practical examples of their use in the analysis of expression data. All learning techniques fall into two broad classes: *supervised* and *unsupervised*. The main difference between them is that the first effectively make use of prior knowledge in order to achieve accurate results, while the latter are constructed in such a way that they do not need additional, previously derived, information about the data to be used. In practice, this is revealed in the initial requirements needed to run the algorithms. In the case of supervised methods the researcher has to introduce a set of *training examples* of known classification to make an *induction*

**Key words:** gene network, microarray, supervised learning.

<sup>1</sup>e-mail lev@ebi.ac.uk

*algorithm* learn from these examples and produce a result in the form of a *classifier* – a function that, when applied to previously unseen examples, assigns them to one of the classes. Consequently, any object considered for classification in supervised learning has a set of *features* and a *label* assigned to the feature's set according to the classifier. The purpose of a classification is to assign a label to a given set of features *correctly* [14].

As for the unsupervised techniques, they do not need examples of correct classification to be used. Normally, one should only introduce a relative measure of how distant or how close objects are. As soon as the matrix of pairwise distances is constructed, the researcher can start to unite the objects into homogeneous groups, which are, in turn, supposed to be maximally separated by means of some criteria.

At this point textbooks on machine learning usually contain a notion that unsupervised learning is for discovering *new* patterns in data and supervised methods are mainly for mining data according to prior knowledge. Although being formally correct, this notion should be used carefully depending on what exactly we mean by 'pattern' or 'knowledge' (discussed below).

Two more concepts, *hypothesis* and *space of hypotheses*, are necessary for our study. Once we are given a training set, in order to choose the appropriate learning algorithm and to tune its parameters, we should make some assumptions (hypotheses) about what would be a rough model of the system under consideration. To get an idea of how to restrict the range of possible hypotheses the following questions need to be answered: what are the processes that underlie the modelled relationships and what might their properties (linear/non-linear, number of parameters, method of representation, etc.) be? The ultimate goal of every classification procedure is to organize optimally the search through the space of possible hypotheses to choose those that are *accurate* and *consistent* with the data.

## Expression data

The main form of representation of expression data is via a *gene expression data matrix*  $X$ , in which rows represent genes and columns represent samples or experimental conditions [15]. Each element of  $X$  specifies the expression level of a gene in a particular sample and is called a *gene expression value*. Expression values may be represented in various ways, but the key point is that only consistent measurements can be used within a single dataset.

One of the major problems of comparative studies of microarrays is that data often come from different platforms, laboratories, etc. It is often difficult or even impossible to compare results of experiments done by different research groups. Nevertheless, there exist cases when the data obtained in different experiments were combined, renormalized and reanalysed successfully [16]. For instance, Spellman et al. [17] used the Cho et al. [18] dataset in their search for cell-cycle-regulated genes of *Saccharomyces cerevisiae*. However, often it is impossible to unite different experimental datasets [19].

The method of expression data analysis discussed in the next section of this review is promising, for it allows comparing hypotheses generated from different datasets, removing the restriction of compulsory consistency of measurement units across individual experiments.

## Reconstruction by learning

### Expression data in learning terms

For the purposes of gene network reconstruction it is usually assumed that the transcription machinery of a gene can be in a finite number of *states*. The exact definition and biological interpretation of 'state' depends on the type of hypotheses that one wishes to generate and explore. For example, for Boolean networks only two states, on and off, are needed. The flexibility of supervised classification is that one can incorporate various types of prior information, along with specific reasoning, when constructing a training set.

Assuming that genes are connected via the expression network, the goal is to find the state of a particular gene  $g$  from the expression measurements of other genes. The gene  $g$  is called the *predicted gene*, while the genes with which we make the prediction are *explaining genes*. If the expression data matrix is rearranged so that the predicted gene  $g$  is in the bottom row with the expression values transformed into states, it is called a *prediction matrix* for gene  $g$  (Table 1). In terms of machine learning, each column of the prediction matrix is an example of correct classification, where the last element of a column is the label and the other elements, expression values of other genes, are the features. Note that the features and the label do not necessarily have to be from the same sample; for example, we may look for the dependencies between gene expression levels at different time points.

Our goal here is not to 'predict' but to find the exact rules that define the states of given genes through the expression values of other genes. That is, we aim to uncover unknown relationships between genes from the data. Even though supervised classification methods are being used to solve

**Table 1 | Example of a prediction matrix for gene 4 (log scale)**

The expression data matrix is rearranged so that predicted gene 4 is represented by the last row in which expression values are transformed into states.

	Sample 1	Sample 2	Sample 3	Sample 4	...
Gene 1	-1.20	1.30	0.85	-1.57	...
Gene 2	2.20	-0.15	-0.95	1.47	...
Gene 3	-1.25	-1.90	0.25	-1.57	...
Gene 5	1.92	1.62	-1.32	-1.52	...
Gene 6	-1.32	-1.12	-2.32	-0.12	...
...	...	...	...	...	...
Gene 4	'1'	'0'	'1'	'1'	...

this problem, this is a problem of knowledge discovery, the discovery of unknown gene interrelations (Table 1).

One could argue here that the results of classification may be highly dependent on the particular definition of state of a gene. On the other hand, the way of defining the state of a gene depends on the meaning of a gene network and the constraints imposed on the space of hypotheses. For example, one can consider a two-state static model as a graph where nodes represent genes and a connection between two nodes represents the fact that the product of the first gene binds to a binding site in the promoter region of the other [20]. In contrast, if it is *known* that the product of a gene interacts with *three* different proteins at different expression levels then at least three states should be considered for that gene if the connection between genes means the existence of interaction between their products.

Let us define for each gene the *normal expression level range* – the range of mRNA levels observed in a steady unperturbed state of a cell. Generally, the following related questions can be considered for a particular gene: how does its expression profile compare with some reference level, chosen from the normal expression level range (e.g. the average expression level under unperturbed conditions), under different conditions and in different cell types? Producing hypotheses answering these questions requires information about the expression levels of genes of treated and untreated cells, which normally are available in most types of comparative microarray experiments. Furthermore, we should use a common, unified language to be able to construct such hypotheses for different microarray experiments separately. The technical details of how to achieve an accurate hypothesis description are not discussed here. However, the general idea is to switch from quantitative (exact) to qualitative (relative) descriptions, e.g. by indicating for a gene that its expression is more/less than the normal expression level. In other words, we aim to describe expression profiles in terms of significant/non-significant changes. The chosen description language would then determine how finely we can detect these changes and distinguish between real changes and noise [21].

### Some properties of gene networks

Now we briefly consider some general properties of gene regulatory networks, as related to the question of hypothesis selection. In the classification approach the relationships between genes are expressed in the form of classifiers. Mathematically, classifiers are functions, and thus, to start the analysis one has to determine the type of functions that most accurately represent the genes' interrelations. Whatever type is chosen (Boolean, linear, non-linear, etc.), one needs to know the number of arguments and parameters of the model. To get an insight into this issue we need to explore the space of the network's states. Though a lot of work has been devoted to this, even the question of what proportion of theoretically possible states may be observed in practice remains open.

The common view in modern biology is that the whole gene regulatory system may be divided into fairly small

subpopulations of genes carrying out some vital functions and connected to each other via the global signalling network [22,23]. Thus the states of genes in the network are often considered in relation to some cellular processes, modules, etc. [24,25]. This makes possible selecting subsets of genes for further detailed analysis. Such a selection along with ensuring that the selected genes are expressed differentially from sample to sample in the considered experiments is crucial due to the restricted amount of data. It may seem that the power of high-throughput methodologies is not fully used by such an approach. Indeed, what would be the point in measuring expression levels of all genes in the genome, if you are interested only in some subset? The answer is that apart from interactions within the set of selected genes it is essential to know the interactions of the subset with the other parts of the system under various conditions. This is one of the key reasons for using supervised algorithms, when we do not need to pretend that nothing is known about the interactions within the biochemical networks and we do not have to consider them as 'black boxes'.

Given a state for each particular gene, we can determine the state for the whole network as a superposition of given genes' states. Even in theory, only a small fraction of all 'theoretically possible' states are expected to be occupied by a real biological system [26,27]. Therefore, it is likely that there are many states, which we never see in our experiments. The living cell, in practice, cannot be forced into all kinds of conditions in order to observe the relationships between input and output signals, the common procedure in *reverse engineering*, just because the biological systems may not be functional (dead) at that point. Due to this uncertainty, we cannot learn genes' precise interconnections by observing all the biologically realizable states. Thus, with the expression data and additional biological information available now, the best strategy would be to construct essential hypotheses for further experimental verification [25,28]. The important question is how confident we are that these hypotheses are not produced just by chance, e.g. due to the noise and the limited amount of available data. This is a subject of a separate study and will not be discussed in detail here, although some related aspects will be considered further in the text.

To summarize the preceding discussion, I shall list the steps of analysis involved:

- (i) pre-select a set of 'interesting' genes, e.g. ones known to participate in a certain cellular process;
- (ii) make sure that the selected genes are expressed differentially;
- (iii) for each gene define its states;
- (iv) pre-define the space of hypotheses as narrowly as possible;
- (v) for each pre-selected gene construct a training set using relevant biological information.

Now we can start to discuss the questions of verification of obtained relations between genes in the network.

**Table 2 | Errors in the prediction matrix**

The last column exemplifies the relevance of explaining genes for the prediction of the states of gene 4. Sample 3 is wrongly labelled (see Table 1). Thus the data shown in italics here are either irrelevant or misclassified.

	Sample 1	Sample 2	Sample 3	Sample 4	...	Relevance
Gene 1	-1.20	1.30	<i>0.85</i>	-1.57	...	Relevant
Gene 2	<i>2.20</i>	<i>-0.15</i>	<i>-0.95</i>	<i>1.47</i>	...	Irrelevant
Gene 3	-1.25	-1.90	<i>0.25</i>	-1.57	...	Relevant
Gene 5	<i>1.92</i>	<i>1.62</i>	<i>-1.32</i>	<i>-1.52</i>	...	Irrelevant
Gene 6	<i>-1.32</i>	<i>-1.12</i>	<i>-2.32</i>	<i>-0.12</i>	...	Irrelevant
...	...	...	...	...	...	...
Gene 4	'1'	'0'	'0'	'1'	...	...

### Errors in expression data

When inferring dependencies between predicted and explaining genes we may generally consider two types of error. The first is related to the columns of the expression data matrix and the second is related to its rows. Due to experimental noise/variations as well as due to the lack of information there may be examples (samples) with wrong labels; and, there may be features (genes) that are irrelevant for the classification of a given predicted gene (Table 2).

There exist various approaches, used in addition to normalization, to deal with noise in microarray data: data randomization [29,30], receiver operating characteristic (ROC) curve analysis [31], estimation of the statistical significance of the results [32–34], etc. For machine learning, there also are several techniques to handle noisy data and this section describes some of them.

The most common way to reduce the effects from misclassified instances (columns of expression data matrix) is to perform *cross-validation* [35]. When cross-validating the results, the training set is split into  $n$  folds, and then the learning algorithm is *trained*  $n$  times, each time on all but one of the  $n$  folds and is *tested* on the remaining one, leaving out a different fold each time. Two parameters are typically traced during this procedure: the accuracy of the tests and the stability of the classifier. If the classifier is stable (unchanged) under different  $n$ -fold splits, we can conclude that the level of noise associated with the misclassified samples is reasonably low. Furthermore, stability implies that test accuracy, averaged across all the  $n$ -fold splits, is a reliable estimate of the prediction accuracy. Therefore, one should prefer those classifiers that have the highest average accuracy and are most stable under multiple  $n$ -fold splits. Although the cross-validation procedure might seem to be rather simple, multiple studies show that, without the possibility of making reliable assumptions about the data in advance (which is practically always the case for microarray data), it is one of the most natural and effective techniques [36].

It is generally assumed that a reasonably small number of explaining genes are sufficient for accurate classification [20,37]. The search for features (rows of expression data matrix) relevant for classification of a particular gene is known

as a *feature subset selection* problem. For feature selection either *filter* or *wrapper* methods can be used [39]. To find the most predictive subset of features by filter approach, some objective function is used in a filtering procedure before running the induction algorithm. Since the specific properties of a particular algorithm are completely ignored in this method, it is regarded as its weakness. The wrapper approach uses the induction algorithm itself to evaluate an objective function and to find relevant features. In the search for the best features this method 'wraps' the induction algorithm into a searching procedure feeding it with several heuristically selected feature subsets and getting feedback as an output of some objective function (normally the accuracy of classification). Wrappers have been reported as performing better than the filter approach for various learning problems [14].

At this point the list of steps involved in the data analysis should be continued:

- (vi) for each selected gene create a classifier using feature subset selection and cross-validating the results obtained;
- (vii) select stable classifiers with the high cross-validation accuracy estimates.

### Hypothesis selection

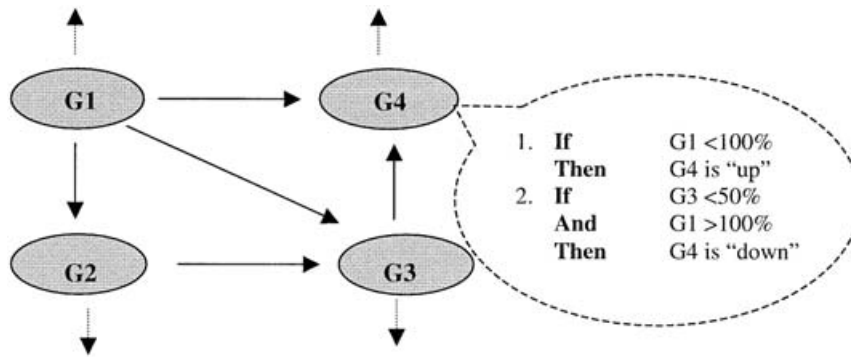
The preceding sections describe how to find relationships between genes from expression data. It should be understood that the structure of such relationships, given in the form of classifiers, can be quite complex (graphs, functions of various types, probability distributions, etc.). To be able to compare or verify them some additional steps should be taken. Thus, when the classifiers are produced, the next problem is to transform them into the hypotheses presented in a simple form suitable for further comparisons and verification.

Classifiers in the form of decision trees or even more complex functions can be decomposed into decision rules each of which may be considered to be the desired simple hypothesis [35,38,39]. Once produced, rules are not bound to the classifiers anymore and can be verified and evaluated separately. This is an important property of the presented approach, for it gives an idea of how to solve the problem of incompatibility of different microarray datasets, mentioned in the Introduction. The possibility to evaluate rules independently is also advantageous because classifiers constructed for the whole dataset may perform differently on different parts of the data; for example, branches of a decision tree may have unequal accuracy rates on the training/test examples covered by them.

Each extracted rule is an indication of some recurring pattern in the data. To be relatively significant this pattern necessarily has to be more than a reflection of some local properties of the dataset and should hold true for several separate datasets. Also, the accuracy of prediction of the predicted gene state must be high. And, finally, to be trustworthy, the rule has to be stable, i.e. to remain unchanged under moderate perturbations of the data [14]. Given the

**Figure 1 | Example of a gene sub-network**

A list of possible decision rules is provided for gene G4.  $G1 > 100\%$  means that expression of G1 is more than 100% of its *normal* level. Two states are presented here for G4: 'up-regulated' and 'down-regulated'. See also Table 2.



rules in a unified representation (see Expression data in learning terms for clarification), we may select and compare them based on the following criteria: significance/coverage, accuracy and stability.

Since we are looking for patterns confirmed by different distinct studies, the most promising hypotheses are those that are preserved between different experiments. Therefore, in the list of the analysis steps formulated earlier we should include:

- (viii) transform classifiers into relevant hypotheses for the following verification;
- (ix) generate hypotheses for each dataset separately and in unified form;
- (x) select hypotheses according to their significance, accuracy and stability;
- (xi) find those that are common for the majority of given datasets;
- (xii) connect genes according to the produced hypotheses (Figure 1).

The structure of decision rules allows connecting genes in an interaction network in a straightforward manner. Any decision rule is directed, i.e. genes on the left-hand side of the rule explain the gene on the right-hand side, thus defining the direction of the connection [39,40]. The list of rules for a certain predicted gene represents a control function transforming the expression levels of explaining genes (input) into the states of the predicted gene (output). Once the rules are verified independently it may turn out that only a part of them can be selected as high-quality hypotheses, because some may be questionable and some incorrect [40]. Though the list of high-quality rules may be incomplete, covering only some of the possible inputs, this should be considered to be an advantage rather than a weakness of the method, for it allows the flexibility to add new rules as new knowledge comes to light and to define relatively complex dependencies between gene expression levels.

## Conclusions

Supervised classification techniques have proven to be an effective and powerful tool for uncovering significant recurring patterns in various types of data. It looks promising to expand their applications to the area of gene network reconstruction. In this study one of the possible methodologies for such applications is described. The method was successfully tested [40] and hopefully this will encourage experts in the field to participate in its further progress.

This work was financially supported by AstraZeneca. I am particularly thankful to Misha Kapushesky, EBI Microarray Group, for valuable discussions and comments. I also would like to thank Alvis Brazma, EBI Microarray Group leader, and Helen Parkinson, ArrayExpress, for critical assessments of this review.

## References

- 1 Nutt, C.L., Mani, D.R., Betensky, R.A., Tamayo, P., Cairncross, J.G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M.E., Batchelor, T.T. et al. (2003) *Cancer Res.* **63**, 1602–1607
- 2 Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A. et al. (1999) *Science* **286**, 531–537
- 3 Dyrskjot, L., Thykjaer, T., Kruhoffer, M., Jensen, J.L., Marcussen, N., Hamilton-Dutoit, S., Wolf, H. and Orntoft, T.F. (2003) *Nat. Genet.* **33**, 90–96
- 4 Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P. et al. (2001) *Proc. Natl. Acad. Sci. U.S.A.* **98**, 15149–15154
- 5 Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S. et al. (2002) *Nat. Med.* **8**, 68–74
- 6 van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T. et al. (2002) *Nature (London)* **415**, 530–536
- 7 Pavlidis, P., Weston, J., Cai, J. and Noble, W.S. (2002) *J. Comput. Biol.* **9**, 401–411
- 8 Mateos, A., Dopazo, J., Jansen, R., Tu, Y., Gerstein, M. and Stolovitzky, G. (2002) *Genome Res.* **12**, 1703–1715
- 9 Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, Jr, M. and Haussler, D. (2000) *Proc. Natl. Acad. Sci. U.S.A.* **97**, 262–267

- 10 Hvidsten, T.R., Komorowski, J., Sandvik, A.K. and Laegreid, A. (2001) *Pac. Symp. Biocomput.* 2001, 299–310
- 11 Kohonen, T. (1992) *Biol. Cybernetics* **43**, 59–69
- 12 Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dittrovsky, E., Lander, E.S. and Golub, J.R. (1999) *Proc. Natl. Acad. Sci. U.S.A.* **96**, 2907–2912
- 13 Mjolsness, E. and DeCoste, D. (2001) *Science* **293**, 2051–2055
- 14 Kohavi, R. (1995) PhD Thesis, Stanford University
- 15 Brazma, A. and Vilo, J. (2000) *FEBS Lett.* **480**, 17–24
- 16 Stolovitzky, G. (2003) *Curr. Opin. Struct. Biol.* **13**, 370–376
- 17 Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) *Mol Biol Cell.* **9**, 3273–3297
- 18 Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J. and Davis, R.W. (1998) *Mol. Cell* **2**, 65–73
- 19 Brazma, A. (2001) *Bioinformatics* **17**, 113–114
- 20 Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I. et al. (2002) *Science* **298**, 799–804
- 21 Quackenbush, J. (2002) *Nat. Genet.* **32** (suppl.), 496–501
- 22 Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. and Barkai, N. (2002) *Nat. Genet.* **31**, 370–377
- 23 Maslov, S. and Sneppen, K. (2002) *Science* **296**, 910–913
- 24 van Someren, E.P., Wessels, L.F., Backer, E. and Reinders, M.J. (2002) *Pharmacogenomics* **3**, 507–525
- 25 Alm, E. and Arkin, A.P. (2003) *Curr. Opin. Struct. Biol.* **13**, 193–202
- 26 Wuensche, A. (1998) *Pac. Symp. Biocomput.* 1998, 89–102
- 27 Kauffman, S.A. (1993) *The Origins of Order: Self Organization and Selection in Evolution*, Oxford University Press, Oxford
- 28 Cross, F.R., Archambault, V., Miller, M. and Klovdstad, M. (2002) *Mol. Biol. Cell* **13**, 52–70
- 29 Kerr, M.K. and Churchill, G.A. (2001) *Proc. Natl. Acad. Sci. U.S.A.* **98**, 8961–8965
- 30 Nguyen, D.V. and Rocke, D.M. (2002) *Bioinformatics* **18**, 1216–1226
- 31 Bilban, M., Buehler, L., Head, S., Desoye, G. and Quaranta, V. (2002) *BMC Genomics* **3**, 19
- 32 Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T.R. and Mesirov, J.P. (2003) *J. Comput. Biol.* **10**, 119–142
- 33 Lyons-Weiler, J., Patel, S. and Bhattacharya, S. (2003) *Genome Res.* **13**, 503–512
- 34 Pan, W. (2002) *Bioinformatics* **18**, 546–554
- 35 Mitchell, T.M. (1997) *Machine Learning*, McGraw-Hill, Boston
- 36 Kohavi, R. (1995) in *Proc. IJCAI-95*, pp. 1137–1143, Montreal, 20–25 August
- 37 Ciuelziu, N., Bottani, S., Bowgine, P. and Kepes, F. (2002) *Nat. Genet.* **31**, 60–63
- 38 Quinlan, J.R. (1993) *C4.5 Programs for Machine Learning*, Morgan Kaufmann Publishers, San Francisco
- 39 Witten, E.H. and Frank, E. (1999) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, San Francisco
- 40 Soinov, L.A., Krestyaninova, M.A. and Brazma, A. (2003) *Genome Biol.* **4**, R6

---

Received 11 August 2003