BIOINFORMATICS

Vol. 19 no. 18 2003, pages 2465–2472
DOI: 10.1093/bioinformatics/btg361

A CART-based approach to discover emerging patterns in microarray data

Anne-Laure Boulesteix[1,], Gerhard Tutz[1] and Korbinian Strimmer[2]*

[1]Seminar for Applied Stochastics, Department of Statistics, University of Munich, Akademiestrasse 1, D-80799 Munich, Germany and [2]Laboratory of Statistical Genetics and Bioinformatics, Department of Statistics, University of Munich, Ludwigstrasse 33, D-80539 Munich, Germany

Received on April 19, 2003; revised on June 25, 2003; accepted on July 7, 2003

ABSTRACT

Motivation: Cancer diagnosis using gene expression profiles requires supervised learning and gene selection methods. Of the many suggested approaches, the method of emerging patterns (EPs) has the particular advantage of explicitly modeling interactions among genes, which improves classification accuracy. However, finding useful (i.e. short and statistically significant) EP is typically very hard.

Methods: Here we introduce a CART-based approach to discover EPs in microarray data. The method is based on growing decision trees from which the EPs are extracted. This approach combines pattern search with a statistical procedure based on Fisher's exact test to assess the significance of each EP. Subsequently, sample classification based on the inferred EPs is performed using maximum-likelihood linear discriminant analysis.

Results: Using simulated data as well as gene expression data from colon and leukemia cancer experiments we assessed the performance of our pattern search algorithm and classification procedure. In the simulations, our method recovers a large proportion of known EPs while for real data it is comparable in classification accuracy with three top-performing alternative classification algorithms. In addition, it assigns statistical significance to the inferred EPs and allows to rank the patterns while simultaneously avoiding overfit of the data. The new approach therefore provides a versatile and computationally fast tool for elucidating local gene interactions as well as for classification.

Availability: A computer program written in the statistical language R implementing the new approach is freely available from the web page http://www.stat.uni-muenchen.de/~socher/

Contact: boulesteix@stat.uni-muenchen.de

*To whom correspondence should be addressed.

INTRODUCTION

In cancer research microarray technology is now routinely used as a diagnostic tool to classify tumor samples. Because many genes are expressed differentially according to tumor type and therefore a large variety of different genetic markers are available, microarrays are believed to allow finer and more reliable identification of tumor classes than the usual clinical methods (Dudoit *et al*., 2002). On the statistical side, analysis of gene expression profiles involves the application of particular supervised learning schemes. These must be suited for the typical data situation with a small number of patients n (=observations) and a large number of genes p (=variables), the so-called 'small n large p' paradigm in gene expression analysis (West *et al*., 2000).

While a large body of literature deals with classification methods in general and their application to microarray gene expression data in particular, see Hastie *et al*. (2001) and Dudoit *et al*. (2002) for a first overview, only few approaches are designed explicitly to consider interaction among the investigated genes. It is well understood that the (co-)expression of genes in a cell is governed by a complicated network of regulatory controls. Hence, to achieve optimal classification accuracy these interdependencies among the genes clearly need to be taken into account.

Emerging patterns (EPs) are among the simplest examples for the use of interaction structures in classification. They were first introduced by Dong and Li (1999) in the context of a general data mining framework that was subsequently applied to microarray data (Li and Wong, 2001, 2002). EPs are expression patterns of the form $expr(X_1) > a_1 \wedge expr(X_2) < a_2$ that have differing frequencies in the considered classes, where $expr(X_i)$ is the measured expression level of gene X_i and the a_i are boundary constants (that are eventually inferred from the data). For illustration of this concept consider Figure 1 where two genes A and B are employed as markers for two cancer classes $Y = 1, 2$. Figure 1a shows an idealized case where class 2 tissue is fully separated from class 1 tissue according to the (emerging)

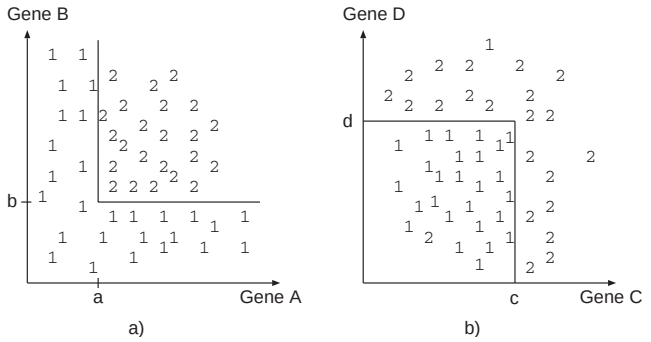

Fig. 1. Examples for possible configurations for two genes with '2' denoting cancer tissue and '1' normal.

pattern $expr(A) \geq a \wedge expr(B) \geq b$. Note that both genes A and B are necessary to discriminate class 2 from class 1 samples. Similarly, in Figure 1b almost all class 1 tissue samples can be distinguished from those of class 2 using the condition $expr(C) < c \wedge expr(D) < d$. Genes A and B and genes C and D interact in Figure 1 and this provides an essential piece of information for classification. Unfortunately, the inference of interaction patterns among genes is difficult, and standard approaches for gene selection and classification frequently miss genes involved even in interactions as simple as in the examples in Figure 1.

Here, we present a simple and versatile statistical approach for inferring EPs and their use in classification. We first introduce the notion of a statistically relevant EP and subsequently address the problem of determining suitable (i.e. short and statistically significant) patterns by suggesting a decision tree (CART)-based method to discover relevant EPs as well as a classification scheme to use these EPs for supervised learning.

The rest of the paper is organized as follows. In the next section, we present the underlying mathematical principles and algorithms of our approach. Subsequently, we demonstrate the power of the method by applying it to a number of simulated data sets as well as to two publicly available 'benchmark' microarray data sets. Finally, we discuss the merits and limitations of our method relative to other supervised learning methods used in the analysis of microarray data.

METHODS

Statistical definition of EPs

Emerging patterns are 'item sets whose support increase significantly from one data set D_1 to another, D_2' (Dong and Li, 1999). Let $n_1 = |D_1|, n_2 = |D_2|$ denote the sample size of two data sets D_1 and D_2 and $n_{P,1}$ and $n_{P,2}$ the counts for a specific pattern P (e.g. $expr(A) \geq 1.023 \wedge expr(B) \geq 0.789$) within D_1 and D_2. The support of pattern P in data set D_i is simply the frequency of occurrence of the pattern, i.e. $supp_{D_i}(P) = n_{P,i}/n_i$, and the growth rate from D_1 to D_2 is defined as

$$r_{D_1D_2}(P) = \frac{n_{P,2}/n_2}{n_{P,1}/n_1} = \frac{supp_{D_2}(P)}{supp_{D_1}(P)}. \qquad (1)$$

Table 1. Examples of EPs

Pattern P	$supp_{D_1}(P)$ (%)	$supp_{D_2}(P)$ (%)	$r_{D_1D_2}(P)$	Type
$expr(A) \geq b \cap$ $expr(B) \geq b$	0	100	$+\infty$	II
$expr(C) < c \cap$ $expr(D) < d$	96.3	4.55	0.047	I

For abbreviations see main text. See also Figure 1.

EPs with a growth rate smaller than one are EPs of type I, otherwise they are of type II. The order of an EP is the number of genes k considered in the EP. For the two EPs of order 2 displayed in Figure 1 these properties are summarized in Table 1.

From the biological point of view, the most interesting EPs in microarray data are those whose support differs significantly between two investigated data samples. Unfortunately, it is not straightforward to determine a general cut-off value for the growth rate that would define a statistically relevant EP. In Li and Wong (2001, 2002) this problem is circumvented by focusing on EPs with infinite growth rate only. In our view this is unsatisfactory for two reasons. First, it seems too restrictive to require infinite growth rate. Second, microarray data are inherently noisy and thus statistical rather than simple deterministic modeling is warranted.

We therefore suggest the following alternative definition of a statistical EP as a pattern of the form

$$P = expr(g_1) \diamond a_1 \wedge \cdots \wedge expr(g_k) \diamond a_k, \qquad (2)$$

where the diamond $\diamond$ stands for either $\leq$ or $>$, for which the hypothesis of equal support for P in the two data sets D_1 and D_2 is tested and thus can be rejected to a certain confidence level. This definition requires an associated test statistic. In our approach we use the deviance and Fisher's exact test (see below).

Discovering EPs with CART trees

In order to find EPs with high statistical significance we have to conduct a search through the space of all possible patterns. For this difficult task Dong and Li (1999) suggested an enumeration-based algorithm. Instead, we use an approach based on decision trees.

A decision tree is a statistical model that recursively partitions the measurement space (i.e. the gene expression measurements for all genes in the data set) into subsets by successive application of a splitting criterion (Breiman *et al.*,

1984). In each step, the subset A is divided into two parts,

$$A_1(j,\mu) = \{x \in A | x_j \leq \mu\} \quad \text{and}$$
$$A_2(j,\mu) = \{x \in A | x_j > \mu\}, \tag{3}$$

so that A is split by use of one variable, x_j, with the split simply depending on a threshold μ from the range of x_j. As a result from d splittings one obtains subsets of the form

$$\{x|x_{i_1} \leq \mu_1\} \cap \{x|x_{i_2} > \mu_2\} \cap \cdots \cap \{x|x_{i_d} \leq \mu_d\}. \tag{4}$$

The relationship between decision trees and EPs is thus simple: a pattern of the form of Equation (2) is equivalent to the agglomerated splitting rules for a leaf in the tree.

For growing the decision tree we employ the variant of the CART algorithm implemented in the R package `tree`. As splitting criterion for the recursive partitioning algorithm we chose the deviance, a statistic that measures how far the fitted model $p(P|D_1) = p(P|D_2)$ deviates from the data. After the tree is grown, we first use Fisher's exact test to determine the maximum order of a pattern, i.e. we test the null-hypothesis that the growth rate is larger in shorter patterns. Subsequently, Fisher's test is also used to evaluate the significance of the pattern, i.e. whether the null-hypothesis $p(P|D_1) = p(P|D_2)$ of equal support given two data sets D_1 and D_2 can be rejected. Fisher's test has the advantage that it is exact and thus can be applied to small leaves (short EPs) whereas other conceivable deviance-based tests are only valid asymptotically. As a last step, we eliminate the gene involved in the first splitting and build another CART tree with the remaining variables. The whole procedure is repeated until no variables (genes) are left.

It turns out that due to the small number of available observations in present microarray data almost all patterns of order greater than two are typically not statistically significant (see also Results section). Longer EPs also may not be observed due to data structure. Hence, the search for statistically significant EPs can be sped up by restricting to short patterns. However, provided the number of observations is large enough to allow longer patterns to be statistically relevant, larger decision trees will need to be grown.

Classification with EPs

The EPs inferred by our tree-based approach can subsequently be used for classification as follows. We define binary covariates based on the m inferred EPs and apply linear discriminant analysis (LDA), a classical supervised learning method (Hastie *et al*., 2001), on these new covariates.

Suppose that we have a learning set $\mathcal{L}$ and a test set $\mathcal{T}$. Let n_L denote the number of observations in $\mathcal{L}$ and n_T denote the number of observations in $\mathcal{T}$. Then we can define two new data matrices $\mathcal{L}'$ of dimensions $(n_L \times m)$ and $\mathcal{T}'$ of dimensions $(n_T \times m)$ as follows (m is the number of inferred EPs):

$$\mathcal{L}'(i,j) = \begin{cases} 1, & \text{if the } i\text{-th training observation is in the } j\text{-th EP} \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

$$\mathcal{T}'(i,j) = \begin{cases} 1, & \text{if the } i\text{-th test observation is in the } j\text{-th EP} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Maximum-likelihood LDA is then employed to predict the class of observations from $\mathcal{T}$ using the matrix $\mathcal{L}'$ as learning data set and $\mathcal{T}'$ as test data set.

For discriminant analysis, we make the following distributional assumptions for $\mathbf{X}^{\mathrm{T}} = (X_1, \ldots, X_m)$, where the $X_j, j = 1, \ldots, m$, stand for the new variables:

$$\mathbf{X}|Y = 1 \sim \mathcal{N}_2(\mu_1, \lambda\mathbf{I})$$
$$\text{where } \mu_1 \text{ is the mean of } \mathbf{X} \text{ in class 1, and} \tag{7}$$
$$\mathbf{X}|Y = 2 \sim \mathcal{N}_2(\mu_2, \lambda\mathbf{I})$$
$$\text{where } \mu_2 \text{ is the mean of } \mathbf{X} \text{ in class 2,} \tag{8}$$

where $\mathbf{I}$ is the identity matrix of dimension $(m \times m)$ and λ is a constant. In particular, the variables are considered to be independent and to have the same variance λ. This simplified discriminant analysis method is also known as nearest centroid approach. Its underlying assumptions are quite strong, nevertheless it offers good performance as it avoids estimating too many parameters from the sparse number of observations.

Prescreening with empirical distribution function

While comparatively fast, our method for discovering EPs with trees is still computationally intensive if applied to all genes in a data set simultaneously. Thus, a prescreening or gene selection step is needed.

As can be seen from the example in Figure 1, an EP typically involves genes that do not necessarily discriminate well when used on their own. Thus selection methods which score the genes separately can be too restrictive so that some interesting genes may be left out. On the other hand, in order to be part of an EP a variable nevertheless has to have some discriminatory power.

This provides the rationale for a simple heuristic for prescreening genes on the basis of the empirical distribution functions F_1 and F_2 of the observations from data sets D_1 and D_2 for each gene. Our selection criterion is whether there exists a point where the empirical distribution function is less than α for one class and more than β for the other class, or more than $1-\alpha$ for one class and less than $1-\beta$ for the other class, where α is a 'small parameter' (say between 0 and 0.1) and β is a 'large parameter' (say between 0.5 and 0.7).

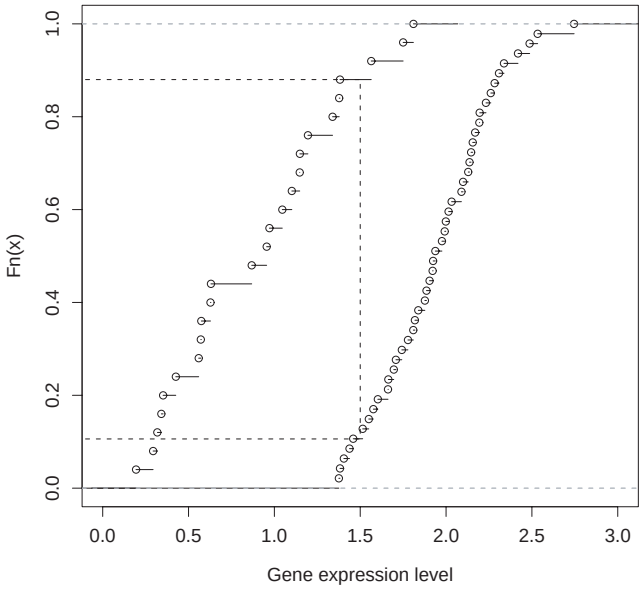

Fig. 2. Example of the empirical distribution of a gene for class 1 and class 2.

For an example consider Figure 2, which shows the empirical distributions of a gene from the leukemia data set (Golub *et al*., 1999). Setting, e.g. $\alpha = 0.1$ and $\beta = 0.7$ it can be seen that this gene will pass the prescreening process as there exists an interval where the empirical distribution function for class 1 is smaller than α and the empirical distribution function for class 2 is larger than β. For illustration one of the points contained in this interval is marked in the panel. Provided α is large enough and β is small enough, this procedure will select most interesting genes. Thus, in Figure 1 genes A, B, C and D would be selected whereas they may be missed by usual gene selection approaches if they do not discriminate well individually.

Recipe for analysis

In summary, our approach to infer EPs consists of the following simple steps:

(1) Employ the prescreening algorithm to determine a set S of candidate genes.

(2) Grow a classification tree with the variables from S with maximal depth two (or any other depth if the number of observable data points is large).

(3) For each inferred pattern P use Fisher's exact test (with significance level p_S) to determine whether the corresponding second splitting in the tree is relevant, i.e. whether the pattern has the maximum order 2.

(4) For each pattern P of order 2 employ Fisher's exact test (with significance level p_G) to assess the null-hypothesis of equal support of the pattern for D_1 and D_2.

(5) Select and store the significant pattern(s) and their dominant class (in the case of duplicate patterns predicting the same class keep only the most significant pattern).

(6) Remove the gene involved in the first splitting of the tree from the set of variables S. Repeat construction of decision trees and evaluation of the resulting EP (steps 2–5) until all genes have been eliminated or the desired number of significant EPs has been retrieved.

As an alternative for removing a single gene in step 6 one could also eliminate all the genes involved in the discovered EPs. This makes the algorithm slightly faster, but has the pitfall that one may miss some potentially interesting EPs.

These steps, along with classification using LDA, have been implemented in a computer program written in the statistical language R. The code, complete with examples and explanation of the normalization procedures, can be freely downloaded from the web page http://www.stat.uni-muenchen.de/~socher/

RESULTS

We used a number of simulated data sets to investigate the power of our approach to infer EPs. Subsequently, we also tested classification accuracy based on EPs using biological data sets.

Simulating data

In order to simulate data for p genes and n observations (n_1 in class 1, n_2 in class 2) we proceeded as follows. First, we generated n_{EP1} EPs of type I and n_{EP2} EPs of type II. Each pattern involved two genes, for which the boundary thresholds were randomly drawn from the uniform distribution between 0.25 and 0.75. The type of ordering ('>' or '<') was also randomly chosen. Second, we simulated expression values in the range $[0, 1]$ according to the following scheme. For genes not involved in an EP, the expression level was drawn randomly from the uniform distribution between zero and one. For genes assigned to an EP, we generated expression values with probability q within the correct EP boundaries, and with probability $1 - q$ outside those boundaries. In our simulations, the values for the free parameters were fixed at $p = 1000, n = 100, n_1 = n_2 = 50, n_{EP1} = n_{EP2} = 50$ and $q = 0.95$. These settings correspond roughly to typical values found in real data sets.

Performance of EP discovery method

To evaluate the performance of our CART-based method to discover EPs from simulated data we introduce two binary variables for each pair of genes: e, which equals 0 if the pair does not form an EP and 1 if the pair forms an EP, and d, which equals 0 if the pair is not detected as an EP by our method and

1 if it is detected as an EP. For each simulated data matrix, this allows to compute four summary statistics:

- n_{ed}, number of detected EPs;
- $n_{\bar{e}d}$, number of non-EP gene pairs which were detected as EPs;
- $n_{e\bar{d}}$, number of EPs which were not detected; and
- $n_{\bar{e}\bar{d}}$, number of non-EP gene pairs which were not detected as EPs.

If an EP of type I or II is diagnosed as EP of type II or I, respectively, the EP is not considered as detected; rather it will be counted in $n_{e\bar{d}}$. The hit rate (HR) is then defined as the median proportion of discovered EPs among the n_{EP} real EPs, i.e.

$$HR = \frac{med(n_{ed})}{n_{EP}}. \tag{9}$$

Similarly, the false alarm rate (FAR) is defined as the median proportion of gene pairs which were discovered as EPs among the non-EP pairs, i.e.

$$FAR = \frac{med(n_{\bar{e}d})}{p(p-1)/2 - n_{EP}}. \tag{10}$$

Subsequently, we tested our method for prescreening genes and discovering EPs for 12 different combinations of the corresponding parameters ($\alpha = 0.1, \beta = 0.3, 0.4, p_G = 10^{-16}, 10^{-18}, 10^{-20}, p_S = 10^{-4}, 10^{-8}$). The parameters α and β control the prescreening, and p_G and p_S are the significance levels of the two tests used for inferring EPs (see Methods section). For each setting, we generated 100 data sets and estimated the HR and the FAR. The results are summarized in Table 2 and the corresponding boxplots are shown in Figures 3 and 4.

Three important features are revealed in the simulation study. First, the prescreening parameter β does not seem to have much impact on both the HR and the FAR. This indicates that a large β can select most important genes. Second, both the FAR and the HR decrease when p_S decreases. Third, a small p_G parameter leads to a distinct decrease in the FAR but not of the HR. Therefore, in analysis of a real data set a small value of the significance level p_G should be advantageous, whereas the other parameters do not seem a particular influence on the performance of the discovering method. However, we expect that for real data set the choice of the prescreening parameter β will be more difficult, in particular for very noisy data where the the distinction between informative and uninformative genes is not straightforward.

EP classification accuracy

To test our classification method we randomly divided two labeled real microarray data sets (colon and leukemia cancer data, see below) into a learning set $\mathcal{L}$ and a test set $\mathcal{T}$, following the procedure described in Dudoit *et al*. (2002). We fixed the size of the test set at 10 observations and repeated the entire procedure of generating $\mathcal{L}$ and $\mathcal{T}$, prescreening genes and learning 50 times to estimate the expected classification error. In our study, we fixed the first prescreening parameter at $\alpha = 0.1$ and set the confidence level $p_S = 10^{-4}$. The second prescreening parameter and the confidence level p_G were varied ($\beta = 0.3, \ldots, 0.7$ and $p_G = 10^{-8}, \ldots, 10^{-15}$). If for at least one partition no EP was found (thus making the discrimination impossible) we indicate this in the Tables 3 and 4 by marking the respective entry with an asterisk. In this case, the mean classification error is calculated using only the partitions that yielded at least one EP.

Table 2. HR and FAR for various parameter combinations and simulated data

p_S	p_G	β	HR	FAR
10^{-4}	10^{-16}	0.3	0.55	$4.1 \cdot 10^{-5}$
10^{-4}	10^{-16}	0.4	0.55	$4.2 \cdot 10^{-5}$
10^{-4}	10^{-18}	0.3	0.55	$2.8 \cdot 10^{-5}$
10^{-4}	10^{-18}	0.4	0.50	$2.8 \cdot 10^{-5}$
10^{-4}	10^{-20}	0.3	0.45	$2.0 \cdot 10^{-5}$
10^{-4}	10^{-20}	0.4	0.45	$2.0 \cdot 10^{-5}$
10^{-8}	10^{-16}	0.3	0.45	$2.9 \cdot 10^{-5}$
10^{-8}	10^{-16}	0.4	0.50	$2.9 \cdot 10^{-5}$
10^{-8}	10^{-18}	0.3	0.45	$1.8 \cdot 10^{-5}$
10^{-8}	10^{-18}	0.4	0.45	$2.0 \cdot 10^{-5}$
10^{-8}	10^{-20}	0.3	0.43	$1.2 \cdot 10^{-5}$
10^{-8}	10^{-20}	0.4	0.40	$1.2 \cdot 10^{-5}$

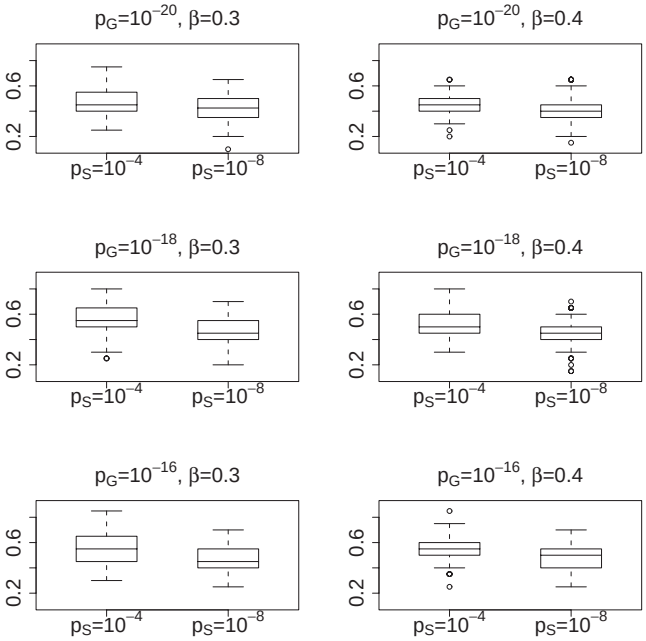

Fig. 3. Boxplots of the HR for the simulated data sets and for different parameter combinations.

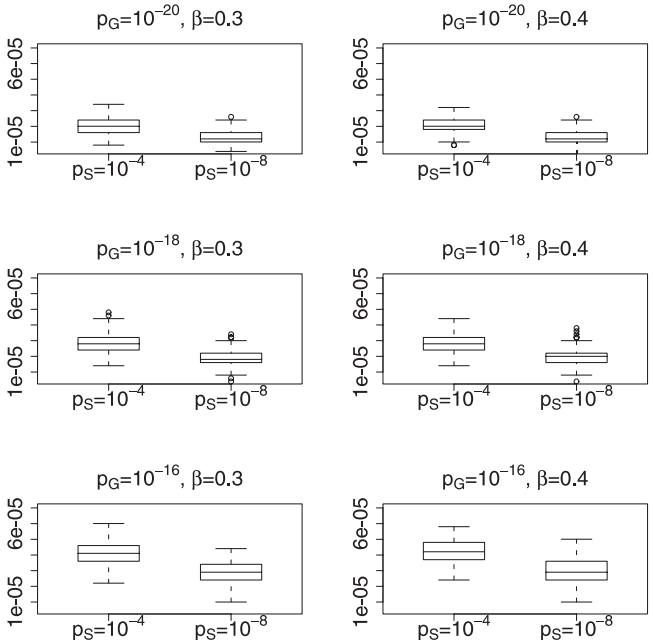

Fig. 4. Boxplots of the FAR for the simulated data sets and for different parameter combinations.

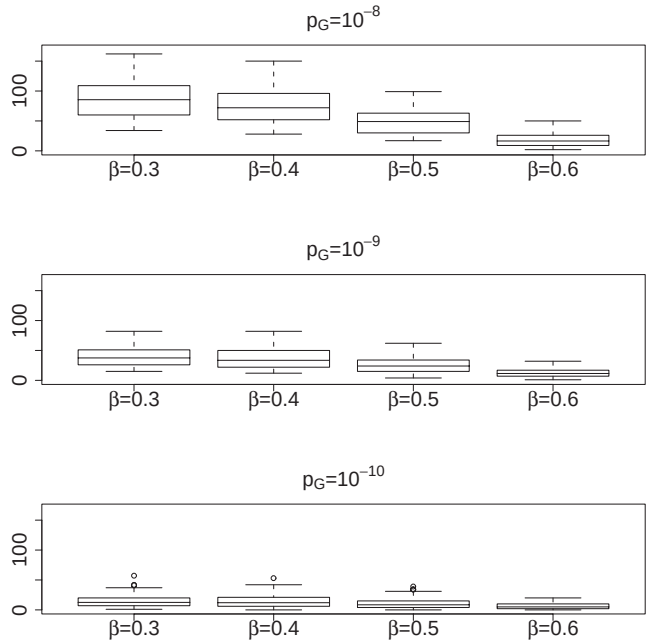

Fig. 5. Boxplots of the number of identified EPs for the colon data set.

Figure 5 and Table 3 show the results for the colon cancer microarray data first investigated in Alon *et al*. (1999). This data set contains 2000 genes for 22 normal and 40 cancer samples. Before applying our EP-based approach for classification we normalized the data and removed duplicate genes with identical expression levels across all 62 samples. Table 3 (top) contains the estimated mean error rate for this data set for different values of p_G and β. After prescreening about 700 (for $\beta = 0.3$) to 60 (for $\beta = 0.6$) genes remained. The overall classification accuracy increases when either β decreases or p_G increases. The fact that a low β increases the accuracy is not surprising, since in this case more potentially informative genes are selected. On the other hand, it is more difficult to explain the correlation between p_G and the mean error rate. Theoretically, a stronger selection criterion for the EPs should prevent the selection of irrelevant EPs. However, classifiers based on a larger number of EPs are also more robust and hence exhibit a lower error rate. Indeed, the boxplot in Figure 5 shows that for small values of p_G the number identified EPs is very low and very variable, with the consequence of a decreased classification accuracy.

Table 3. Mean classification error for the colon cancer data set

	$p_G = 10^{-8}$	$p_G = 10^{-9}$	$p_G = 10^{-10}$
$\beta = 0.3$	0.134	0.128	0.158
$\beta = 0.4$	0.138	0.146	0.165*
$\beta = 0.5$	0.146	0.158	0.206*
$\beta = 0.6$	0.192	0.206	0.234*
	LDA	3-NN	SVM
10 genes	0.120	0.124	0.118
20 genes	0.122	0.152	0.122
50 genes	0.122	0.164	0.114
100 genes	0.126	0.150	0.122
200 genes	0.128	0.160	0.122

Results for (top) EP-based classification and (bottom) three standard classification methods (see text).

For further assessment, we investigated the performance of EP-based classification using a microarray data from leukemia cancer studies (Golub *et al*., 1999). After preprocessing as described by Dudoit *et al*. (2002), we obtained a data matrix with 3571 genes and 72 samples [47 of tissue type acute lymphoblastic leukemia (ALL) and 25 of type acute myeloid leukemia (AML)]. Table 4 (top) contains the estimated mean error rate based on 50 random partitions into learning and test sets $\mathcal{L}$ and $\mathcal{T}$ for various values of p_G and β. After prescreening about 700 (for $\beta = 0.4$) to 100 (for $\beta = 0.7$) genes remained. The overall picture is similar to that of the colon data set and classification accuracy increases with p_G. However, the parameter β has less of an influence both on accuracy and the number of identified EPs (Fig. 6).

Comparison with other supervised learning methods

Using the same study design, we tested three standard classification methods (Hastie *et al*., 2001), diagonal LDA (DLDA), nearest-neighbors with $k = 3$ (3-NN) and the support vector machine (SVM) approach. We chose these methods because

Table 4. Mean classification error for the leukemia cancer data set

	$p_G = 10^{-13}$	$p_G = 10^{-14}$	$p_G = 10^{-15}$
$\beta = 0.4$	0.028	0.028	0.044
$\beta = 0.5$	0.026	0.030	0.046
$\beta = 0.6$	0.024	0.030	0.038
$\beta = 0.7$	0.026	0.032	0.049*
	LDA	3-NN	SVM
10 genes	0.040	0.044	0.048
20 genes	0.030	0.036	0.040
50 genes	0.028	0.040	0.052
100 genes	0.034	0.044	0.042
200 genes	0.032	0.044	0.036

Results for (top) EP-based classification and (bottom) three standard classification methods (see text).

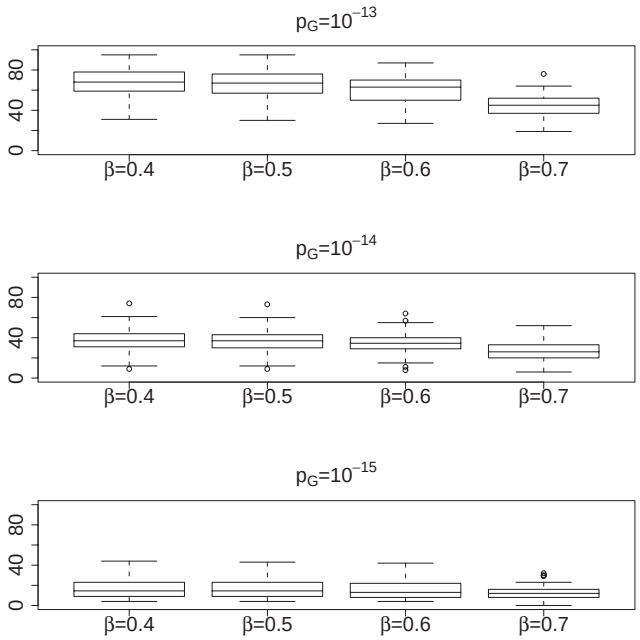

Fig. 6. Boxplots of the number of identified EPs for the leukemia data set.

DLDA and *k*-NN were top-performers in a recent comparative study (Dudoit *et al*., 2002) and SVM is also believed to be well suited for microarray data (Furey *et al*., 2000). For the 3-NN method, we used the R function `knn` from the library `class` and chose the Euclidean distance as distance metric. For SVMs, we used the function `svm` from the R package `e1071`. Since these methods work much better with few genes, we performed a preliminary gene selection using the robust Wilcoxon statistic as described by Dettling and Bühlmann (2003).

The classification accuracy for the three investigated methods are shown in Table 3 (bottom) for the colon data set and in Table 4 (bottom) for the leukemia data set. For the colon data, the results are a bit better than those obtained using our EP-based approach, while for the leukemia data set our results are better. Thus, the classification accuracy using EPs are comparable with those of the best available methods. However, our method has the added benefit of additionally identifying dependency structures in the data. This is a distinct advantage over the usual approaches that only filter highly differentially expressed genes and classify the samples.

EPs identified in the colon data set

It is instructive to analyze the EPs identified by our CART-based approach in real microarray data sets. For this purpose, we ran the prescreening and EP discovery algorithm on the whole colon cancer data set (Alon *et al*., 1999), with $\beta = 0.3$. In Table 5 we list all identified EPs with a p-value lower than 10^{-11}, i.e. those that are most significant.

Three different things are worth pointing out. First, the numbers of EPs of type I (Table 5, top) and of type II (bottom) are approximately equal. Note that this balanced situation is a desirable property for classification. Second, another interesting observation is that not all the genes involved in the EPs listed in Table 5 are good classifiers individually. For instance, gene L38810 (involved in the first EP of type II) and gene T41207 (involved in the sixth EP of type II) are ranked 868 and 533, respectively, according to the Wilcoxon statistic. This shows it may be too restrictive to select genes for classification based on an univariate criterion, such as the t-statistic or the Wilcoxon statistic. Third, some genes such as R55310 and T62947 take part in more than one pattern, indicating that there is higher-level interaction in the data.

Our EPs inferred for the colon data set are very different from the EPs given by Li and Wong (2001, 2002). Several reasons can be put forward to explain this discrepancy. First, Li and Wong looked only for what could be referred as 'perfect EPs', i.e. EPs with infinite growth rate. From the statistical point of view, it makes sense to consider non-perfect EPs as well, especially for noisy data like microarray data. Second, our EPs are shorter (order two), because we focus on statistically significant and reproducible EPs. In contrast, the EPs found by Li and Wong are typically very long and thus are likely to contain highly correlated genes. From a statistical perspective long EPs are also indicative of overfitting, i.e. some of the additional genes will eliminate one or two observations for the training data set but not generally be useful for additional independent data. Finally, Li and Wong used a very restrictive prescreening procedure that left only 35 genes. Since EPs are based on genes interacting with each other, it is questionable whether such dramatic data elimination is helpful.

DISCUSSION

In this paper, we have introduced a new approach to supervised learning and exploring interactions between genes based on

Table 5. Significant EPs identified in the colon data set

Gene 1	*Gene*2	Freq. in D_1	Freq. in D_2
$H06524 \in [-0.54, +\infty)$	$Z50753 \in [0.16, +\infty)$	1	0.075
$H11084 \in [-\infty, 0.33)$	$Z50753 \in [-0.55, +\infty)$	0.91	0.025
$U04953 \in [-\infty, 0.07)$	$M63391 \in [1.17, +\infty)$	0.86	0
$R81330 \in [-0.45, +\infty)$	$R36977 \in [-\infty, -0.08)$	0.91	0.05
$M82919 \in [-1.05, +\infty)$	$X12369 \in [0.49, +\infty)$	0.82	0
$T51493 \in (-\infty, -0.77]$	$R64115 \in (-\infty, 0.58]$	0.91	0.05
$T64467 \in [0.67, +\infty)$	$H72234 \in (-\infty, -0.10]$	0.82	0
$U04953 \in (-\infty, 0.07]$	$R60883 \in [-0.38, +\infty)$	0.91	0.025
$T64467 \in [0.67, +\infty)$	$T51493 \in (-\infty, -0.72]$	0.82	0
$R55310 \in [0.32, +\infty)$	$U09564 \in (-\infty, -0.15]$	0.82	0
$R55310 \in [0.32, +\infty)$	$H72965 \in (-\infty, -0.51]$	0.86	0
$L38810 \in (-\infty, 1.48]$	$M76378 \in (\infty, 1.19]$	0	0.9
$X87159 \in (-\infty, 0.68]$	$X63629 \in [-0.90, +\infty)$	0	0.875
$D14812 \in [0.20, +\infty)$	$U25138 \in (-\infty, -0.44]$	0.14	0.975
$T62947 \in [-1.06, +\infty)$	$M76378 \in (-\infty, 1.18]$	0	0.875
$T62947 \in [-1.12, +\infty)$	$H20709 \in (-\infty, 2.80]$	0.05	0.925
$T41207 \in (-\infty, -0.11]$	$T92451 \in (-\infty, 1.94]$	0.05	0.925
$T71025 \in (-\infty, 2.19]$	$H08393 \in [-1.19, +\infty)$	0	0.875
$M91463 \in (-\infty, -0.59]$	$R44418 \in (-\infty, 0.54]$	0.14	0.975

Emerging patterns of (top) type I and of (bottom) type II.

the concept of EPs. This tree-based method is computationally fast and intuitive and also assigns statistical relevance to the identified patterns. In contrast to previous algorithms, it allows to rank EPs by statistical criteria and avoids overfitting the observed data.

We have compared our method using simulated and real microarray gene expression data with other widely used approaches for classification. Our approach has classification accuracy comparable with those of the best methods available but at the same time additionally infers local interactions among two or more genes in the form of EPs. Furthermore, we demonstrated that it is not generally necessary to conduct strong prescreening and data reduction before classification. Our investigations also emphasize that there are genes that are poorly suited for classification on their own, but are critical in association with other genes because of reciprocal interactions.

A potential drawback of the present approach is that we have considered only data sets with two tissue classes. However, generalization to the multi-class case is possible, and future version of our algorithm (and program) is planned to be applicable to this case. Furthermore, we have chosen suitable values for the confidence level parameters p_S and p_G for a given data set on a heuristic basis. These parameters are linked with the number of analyzed genes and therefore should be better controlled by a suitable multiple testing procedure (Benjamini and Hochberg, 1995). Finally, the prescreening step could potentially be improved by employing a multivariate rather than univariate selection criterion. However, this would increase substantially the computational costs of the algorithm.

To summarize, our CART-based approach for searching for EPs in gene expression data offers a versatile new tool both for elucidating local gene interaction and for classifying tissue samples. We believe that future development of classification approaches will even show a tighter integration of higher-order interaction, such as in the form of genetic networks. Ultimately, it is the knowledge of these gene interactions that will help biologist to understand better the function of genes and the mechanisms of cancer.

ACKNOWLEDGEMENTS

We are grateful for financial support from the Deutsche Forschungsgemeinschaft (DFG) within the Emmy Noether program (K.S.) and within the SFB 386 (A.-L.B. and G.T.).

REFERENCES

Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D. and Levine,A. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Cell Biol.*, **96**, 6745–6750.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JR Statist. Soc. B*, **57**, 289–300.

Breiman,L., Friedman,J.H., Olshen,J.A. and Stone,C.J. (1984) *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Dettling,M. and Bühlmann,P. (2003) Boosting for tumor classification with gene interaction data. *Bioinformatics*, **19**, 1061–1069.

Dong,G. and Li,J. (1999) Efficient mining of emerging patterns: discovering trends and differences. *Proceedings of the SIGKDD (5th ACM International Conference on Knowledge Discovery and Data Mining)*, **5**, ACM, NY, pp. 43–52.

Dudoit,S., Fridlyand,J. and Speed,T.P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *JASA*, **97**, 77–87.

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

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

Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learning*. Springer, New York.

Li,J. and Wong,L. (2001) Emerging patterns and gene expression data. *Genome Inform.*, **12**, 3–13.

Li,J. and Wong,L. (2002) Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. *Bioinformatics*, **18**, 725–734.

West,M., Nevins,J.R., Marks,J.R., Spang,R. and Zuzan,H. (2000) Bayesian regression analysis in the 'large p, small n' paradigm with application in DNA microarray studies. *Technical Report* 15, Institute of Statistics and Decision Sciences, Duke University, USA.