

Software

Open Access

## Multiclass discovery in array data

Yingchun Liu and Markus Ringnér\*

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

Email: Yingchun Liu - [spring@thep.lu.se](mailto:spring@thep.lu.se); Markus Ringnér\* - [markus@thep.lu.se](mailto:markus@thep.lu.se)

\* Corresponding author

Published: 04 June 2004

Received: 21 January 2004

BMC Bioinformatics 2004, 5:70

Accepted: 04 June 2004

This article is available from: <http://www.biomedcentral.com/1471-2105/5/70>

© 2004 Liu and Ringnér; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

### Abstract

**Background:** A routine goal in the analysis of microarray data is to identify genes with expression levels that correlate with known classes of experiments. In a growing number of array data sets, it has been shown that there is an over-abundance of genes that discriminate between known classes as compared to expectations for random classes. Therefore, one can search for novel classes in array data by looking for partitions of experiments for which there are an over-abundance of discriminatory genes. We have previously used such an approach in a breast cancer study.

**Results:** We describe the implementation of an unsupervised classification method for class discovery in microarray data. The method allows for discovery of more than two classes. We applied our method on two published microarray data sets: small round blue cell tumors and breast tumors. The method predicts relevant classes in the data sets with high success rates.

**Conclusions:** We conclude that the proposed method is accurate and efficient in finding biologically relevant classes in microarray data. Additionally, the method is useful for quality control of microarray experiments. We have made the method available as a computer program.

### Background

A common application in microarray data analysis is to identify genes that, based on their expression levels, discriminate between known classes of experiments. This identification is often achieved by using various statistical measures to, gene-by-gene, correlate the expression levels with the classes of interest. In this way a discriminatory weight is calculated for each gene. For example, Golub *et al.* used a signal-to-noise statistic to find genes with expression patterns that discriminate between samples obtained from patients with acute myeloid leukemia and patients with acute lymphoblastic leukemia [1]. Other examples include using a standard *t*-test to discriminate between breast tumors from carriers of *BRCA1* mutations and carriers of *BRCA2* mutations [2]. For an overview of applications see [3]. In most studies, the number of genes is much larger than the number of experiments. For such

a large number of genes, it is crucial to estimate how many genes would correlate with the classes of interest by chance. Often, a *P* value corresponding to the probability of obtaining a given weight by chance is calculated for each weight. One can then investigate if there is an over-abundance of discriminatory genes for classes of interest as compared to randomly selected classes. Indeed, such an over-abundance has been found for many microarray-based classification applications (see *e.g.* [1,2,4]).

Often clustering methods, such as hierarchical clustering [5], *k*-means clustering [6], or self-organizing maps (SOM) [7] are used for unsupervised classification of array data (see [8] for an overview). For example, hierarchical clustering has been used to discover two subtypes of diffuse B-cell lymphoma [9], three subtypes of breast tumors [10], and two subtypes of cutaneous melanoma [4].

Dugas *et al.* have developed an iterative *k*-means clustering method for class discovery in array data [11]. Examples of SOM-based methods for discovery of cancer subtypes include applying SOMs to automatically discover the distinction between acute myeloid leukemia and acute lymphoblastic leukemia [1], and to separate 14 different tumor subtypes [12]. SOMs and *k*-means clustering require the user to predefine the number of clusters to be found. Hsu *et al.* proposed an unsupervised hierarchical self-organizing map approach that automatically identifies a suitable number of clusters, and applied it to a couple of publicly available array data sets [13].

In these clustering methods, experiments are clustered based on the distance between them in gene expression space. An alternative unsupervised classification approach to discover classes in gene expression data, which exploits the fact that there typically is an over-abundance of genes separating known classes was proposed by Ben-Dor *et al.* [14]. In their method, classes are discovered by seeking partitions of experiments with an over-abundance of discriminatory genes. In contrast to many clustering methods, no metric to define distances between experiments is required. Furthermore, classes are discovered based only on the subset of genes that are differentially expressed between the classes, whereas in unsupervised clustering the distances are often based on all the genes. A similar classification method, which also searches for binary class distinctions in a set of samples that show separation in the expression of subsets of genes, has been developed by von Heydebreck *et al.* [15]. These classification methods are well-suited to discover several significant partitions of experiments, each based on a different subset of genes.

Inspired by the method by Ben-Dor *et al.*, we have previously used a similar approach to sub-classify familial breast cancer into two classes [16]. Briefly, the approach was as follows. For a given partition of the experiments into two classes (with  $n_1$  and  $n_2$  experiments, respectively), a discriminative weight was calculated for each gene using the signal-to-noise statistic [1]. To assign *P* values to the weights one has to perform random permutation tests. Such a test was used to generate a weight distribution that could be expected for two classes with  $n_1$  and  $n_2$  experiments under the assumption of random gene expression. Using this weight distribution, each weight was assigned a *P* value corresponding to the probability to obtain the weight or larger for a random partitioning into  $n_1$  and  $n_2$  experiments. Candidate partitions of the data were scored with the number of statistically significant weights, that is the number of genes that were significantly different in expression between the classes. A simulated annealing [17] scheme was used, in which partitions were updated by changing the class of a randomly selected experiment, to find the partition of experiments into the classes with

the highest score. Our approach is different from the Ben-Dor *et al.* method in two respects. First, they use the total number of misclassification (TNoM) score to find discriminatory genes [18]. Second, we use a fixed *P* value cut-off to find the number of discriminatory genes, whereas they instead use surprise scores [14].

In this work, we have extended our unsupervised classification method for discovery of more than two classes and to allow for missing values in gene expression data. Furthermore, we have made the method publicly available as a computer program. For the breast cancer study [16], we performed random permutation tests for all possible  $n_1$  and  $n_2$  that add up to the total number of experiments. Extending the method to find a preset but arbitrary number of classes and to allow for missing values would result in performing random permutation tests for many more combinations of class sizes. For nonparametric rank-based statistics, analytically calculated *P* values correspond to what would be obtained by random permutation tests. Therefore, we decided to use such statistics instead of a parametric test. Moreover, nonparametric methods have been shown to be robust conservative (low numbers of false positives) in its application to the identification of discriminatory genes in gene expression data [19].

Here, we describe the unsupervised classification method used in our class discovery program in detail, and results from applying it on two publicly available data sets.

**Implementation**  
**Identification of differentially expressed genes**

The Wilcoxon rank sum test (WT) is used to identify genes differentially expressed in two classes [21]. The nonparametric WT tests for equality of medians of two samples of data, but unlike the *t*-test it makes no assumption that the data is normally distributed. It operates on rank-transformed data rather than the raw values. In our method, the expression values of each gene are ranked across experiments from low to high, disregarding to which class each experiment belongs. For a given partition of experiments into two classes, a discriminatory weight  $u_g$  is calculated for each gene (*g*),

$$\text{class 1: } n_1 \text{ samples} \quad \text{class 2: } n_2 \text{ samples} \quad n_2 > n_1$$

$$w_g = \sum_{e \in \text{class 1}} \text{rank}_g(e)$$

$$u_g = w_g - n_1(n_1 + 1)/2$$

where *e* denotes an experiment with  $\text{rank}_g(e)$  for *g*. Next, we want to calculate a *P* value of  $u_g$  for the null hypothesis that expression values for all experiments are drawn from the same probability distribution. If the *P* value is near

zero, it casts doubt on the null hypothesis and suggests that the medians of expression values are significantly different in the two classes. For  $n_1 > 8$  (and thus  $n_2 > 8$  also), the  $P$  value can be calculated by using a normal approximation [22],

$$\begin{aligned} \text{mean}_{u_g} &= n_1 n_2 / 2 \\ \text{var}_{u_g} &= n_1 n_2 (n_1 + n_2 + 1) / 12 \\ z &= \left( u_g - \text{mean}_{u_g} \right) / \sqrt{\text{var}_{u_g}} \\ z &\in N(0, 1). \end{aligned}$$

For partitions with  $n_1$  being at least 9, we use this normal approximation. For partitions with  $n_1$  smaller than 9, we assign  $P$  values to  $u_g$  using a random permutation test. For each  $n_1$  and  $n_2$ , the test is based on 50,000 random permutations of class labels. For the WT, we use two-sided tests.

For partitions of experiments into three or more classes, the Kruskal-Wallis test (KWT) is used to identify discriminatory genes [23]. The KWT is a nonparametric version of the one-way analysis of variance (ANOVA), uses the ranks of the data, and is an extension of the WT to more than two groups. It tests for equality of medians of  $k$  samples of data. For a given partition of experiments into  $k$  classes, a discriminatory weight  $H_g$  is calculated for each gene,

class  $i$  :  $n_i$  samples

$$\begin{aligned} n &= \sum_{i=1}^k n_i \\ R_g(i) &= \sum_{e \in \text{class } i} \text{rank}_g(e) \\ H_g &= \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_g(i)^2}{n_i} - 3(n+1). \end{aligned}$$

For partitions where all classes have at least 5 samples, the sampling distribution of  $H_g$  can be approximated very well by a chi-square distribution with  $k - 1$  degrees of freedom [22]. Hence, the  $P$  value for the null hypothesis that expression values for all experiments are drawn from the same probability distribution can be calculated from  $\chi^2(k - 1)$ . If the  $P$  value is close to zero it suggests that the median of expression levels for at least one class is significantly different from the others. This does not necessarily mean that every group differs from every other group. For partitions with all classes having at least 5 experiments, we use the chi-square approximation.  $P$  values are not calculated for partitions into more than 2 classes for which a class have less than 5 experiments.

Missing values is handled by using, for each gene, only those experiments for which there exists measurements in the calculation of the statistical test (WT or KWT).

### Scoring partitions of experiments

Each partition of experiments is assigned a score corresponding to the number of genes with  $P$  values smaller than or equal to a user specified cut-off. Thus, the score is the number of genes with significantly different expression in the classes. Because the total number of genes for a data set is identical for all partitions, we do not correct  $P$  values for multiple testing. For random gene expressions, we would, for a typical partition, expect a score of  $P$  multiplied by the total number of genes. We denote this expected score by  $E$ .

### Finding the partition with the highest score

We use simulated annealing (a global optimization method) to find the partition of the experiments into a given number of classes with the highest score [17,24]. This procedure is described in Fig. 1. In addition to the best partition, we extract all partitions investigated in the search that have a score larger than a user specified cut-off. Table 1 contains the values of the parameters used in the analysis of the data sets.

### Peeling discriminatory genes

For a data set there may be several biologically relevant partitions. However, one of them is often supported by many more discriminatory genes than the others. In such a scenario, the discovered partitions with the highest scores will mostly be similar to the best one, because shifting one or two experiments will still result in a higher score as compared to other relevant partitions. Hence, it may be difficult to discover important but not dominating partitions. One way to address this issue is to remove genes that contribute to the score for the best partition and run the class discovery program again using this smaller set of genes [14]. Using such a procedure, one can iteratively peel discriminatory genes from the data set to systematically investigate the presence of further partitions with significant scores.

### Evaluation

To further evaluate discovered classes we used hierarchical clustering. Clustering was performed using EPCLUST <http://ep.ebi.ac.uk/EP/EPCLUST/> with the distance measure parameter set to linear correlation based distance (Pearson).

To investigate discovery of two classes, we used the non-*BRCA1/2* familial breast cancer (termed *BRCAX*) data set by Hedenfalk *et al.* [16]. This data set consists of 16 *BRCAX* samples, for which intensity ratios of 4,795 clones considered to be well-measured are provided. Following

## 1. Initialization

- (a) Initialize the labeling of the experiments by randomly assigning each experiment to one of  $k$  classes. For  $k > 2$  require each class to have at least 5 experiments.
- (b) Initialize the 'temperature'  $T = T_{\text{start}}$ .

## 2. Procedure to optimize the partition score.

- (a) Calculate the partition score ( $S$ ) for the labeling.
- (b) Randomly pick an experiment and change its label to a different label.
  - if  $k = 2$ , the experiment is randomly selected from all experiments.
  - if  $k > 2$ , the experiment is randomly selected from classes having more than 5 experiments.
- (c) Calculate the score for the changed labeling ( $S_{\text{new}}$ ).
  - Accept the changed labeling, if  $S_{\text{new}} > S$ ; otherwise, accept it with probability  $e^{(S_{\text{new}} - S)/T}$ .
  - If the changed labeling is accepted set  $S = S_{\text{new}}$ . If not, keep the original labeling and  $S$ .
- (d) Decrease  $T$  by a factor  $\eta$  ( $T \leftarrow \eta T$ ), if  $N_{\text{success}}$  changed labelings have been accepted, or  $N_{\text{total}}$  changed labelings have been proposed at the current  $T$ .
- (e) Repeat steps 2(b)-2(d) until  $T$  in step 2(d) becomes smaller than  $T_{\text{end}}$ .

## 3. Extract all partitions with scores larger than a threshold ( $S_c$ ) and the discriminatory genes associated with each of them.

### Figure 1

The essential algorithmic steps in the class discovery procedure. For actual values of the parameters used in the analysis see Table 1.

Hedenfalk *et al.*, we performed pre-processing of the data such that the log intensity ratios were mean-centered and these values were used as a measure of the expression levels.

To investigate discovery of more than two classes, we used the data for small round blue cell tumors (SRBCTs) of childhood by Khan *et al.* [20]. This data set consists of 88 samples, separated into a training and a test set, for which

**Table 1: Parameters in the class discovery procedure and the values used for the SRBCT and the BRCAx data.**

Parameter	Value
$T_{start}$	3.0
$T_{end}$	0.1
$\eta$	0.9
$N_{success}$	10
$N_{total}$	150 (SRBCT) or 50 (BRCAx)

**Table 2: Batches of production for the 88 SRBCT microarrays.**

Batch <sup>a</sup>	Category	Experiments <sup>b</sup>
104	EWS	T1, T2, T3, T4, C1, C2, C3, C4
	BL	C1, C2, C3, C4
	NB	C1, C2, C3
	RMS	T1, T2, T3, T4, C8, C11
	TEST	5, 24
118	EWS	T6, T7, T9, T11, T12, T13, T14, T15, T19
	RMS	T5, T6, T7, T8, C3, C4
	TEST	6, 9, 11, 20, 21
119	EWS	C6, C7, C8, C9, C10, C11
	BL	C5, C6, C7, C8
	NB	C4, C5, C6, C7, C8, C9, C10, C11, C12
	RMS	C2, C5, C6, C7, C9, C10
	TEST	3
143	RMS	T11
	TEST	1, 2, 4, 7, 12, 17
163	RMS	T10
	TEST	8, 10, 13, 14, 15, 16, 18, 19, 22, 23, 25

<sup>a</sup>Identifier of batch of production <sup>b</sup>T: tumor samples; C: cell lines

relative intensities of 2,308 filtered genes are provided. In the training set, there are 63 samples belonging to four different SRBCT types, neuroblastoma (NB), rhabdomyosarcoma (RMS), Burkitt's lymphoma (BL) and Ewing's sarcoma (EWS). The test set consists of 20 SRBCTs (belonging to the four types) and 5 non-SRBCTs. Following Khan *et al.*, we used the logarithm of the relative intensities as a measure of the expression levels. The 88 experiments were performed on microarrays from 5 different batches of production, and included samples from both tumor biopsy material and cell lines (Table 2).

## Results and Discussion

### Discovery of two classes in the breast cancer data set

To validate our class discovery program, we first applied it to the Hedenfalk *et al.* BRCAx data set and looked for two classes, as in our original analysis [16]. In that work, a signal-to-noise statistic [1] and a  $P$  value cut-off of 0.001 was used to discover two classes supported by 60 discriminatory genes. For a fixed  $P$  value cut-off, we now expect less

discriminatory genes, because of the change to a rank-based statistic. Therefore, we decided to use a somewhat larger cut-off to facilitate a comparison of our set of discriminatory genes with the original set. Importantly, we found the classes discovered to be rather insensitive to changes in  $P$  value cut-off. Our best scoring partition was supported by 133 discriminatory genes ( $P \leq 0.005$ ), whereas we would expect 24 discriminatory genes for a random partition ( $E = 24$ ). Our partition was similar to the one in the original study, except that three samples had shifted class. Moreover, the Hedenfalk *et al.* partition was also highly significant (score 117), and the top scoring partitions were dominated by partitions very similar to it. Of our 117 discriminatory genes 57 overlapped with the 60 found in ref. [16]. We conclude that the partition found in the original study is robust to changes in the details of the class discovery algorithm.

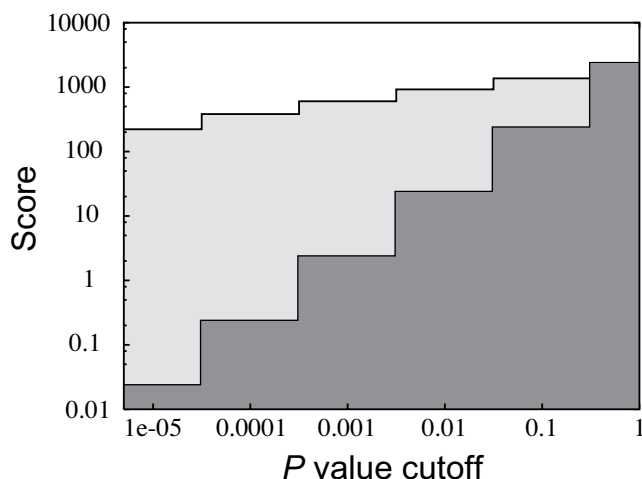
Setting the values of the parameters in the annealing schedule requires experimentation. For our data sets, we selected parameter values for which the program runs relatively fast (5–10 minutes on a standard personal computer), and for which running the program again, with a new random initialization of the classes, often resulted in again finding the best partition found previously. For more conservative parameter values ( $T_{start} = 25$ ,  $N_{success} = 100$ ,  $N_{total} = 500$ ) for which the program takes 10–15 times longer to run, we found that the best partition was essentially found every time the program was run (9 out of 10 times for the breast cancer data).

### Discovery of two classes in the SRBCT data set

We applied our class discovery method to the 63 experiments in the SRBCT data training set, leaving the 25 experiments in the test set for verification. For two classes, our best scoring partition was supported by 602 genes ( $P \leq 0.001$ ;  $E = 2.3$  genes). We investigated the score (number of discriminatory genes) for our best scoring partition as a function of  $P$  value cut-off. The significance of the partition is relatively insensitive to the  $P$  value cut-off (Fig. 2). The two classes essentially separated cell lines from tumors, with 4 cell lines (EWS-C4, RMS-C3, C8 and C11) in the tumor class. Thus, we found, in agreement with principal component analysis of the expression data [25], that the dominant separation of the experiments is into cell lines and tumors (94% correct for our two classes). The score for perfect partitioning into cell lines and tumors was 513.

### Discovery of three classes in the SRBCT data set

Next, we investigated discovery of more than two classes. For three classes, our best scoring partition was supported by 934 discriminatory genes ( $P \leq 0.001$ ;  $E = 2.3$ ). The three classes separated experiments according to batches of microarray production (Table 2). Experiments on print



**Figure 2**  
The number of discriminatory genes (score) as a function of the cut-off in *P* value. The data shown is for discovery of two classes in the SRBCT data set. The two curves are for the best partition found (light gray) and for random partitions (dark gray). For the *P* value cut-off 0.001, the best partition is supported by 602 genes, whereas the expectation for a random partition is 2.3 genes.

batches 104, 118 and 119 were in one class each. There was only one error: RMS-T4 (print batch 104) was in the class with experiments on print batch 118, as were the two experiments from batches 143 and 163. When only using the 61 training experiments belonging to batches 104, 118 and 119, the score for a perfect partitioning into these three batches was 923.

**Discovery of four classes in the SRBCT data set**

For four classes, our best scoring partition was supported by 1051 discriminatory genes ( $P \leq 0.001$ ;  $E = 2.3$ ). The four classes were identical to our result for three classes (according to print batches), except that the BL experiments (on batches 104 and 119) were in a class of their own. Thus, apart from finding the BL class, the dominant separation of the experiments into four classes is according to print batches. In agreement with our result, BL was in the original supervised analysis found to be the category with the most distinct expression profile [20]. The score for perfect partitioning into the four SRBCT categories was 544, which though not the best score is highly significant compared to random expectations.

**Peeling genes discriminatory for batches of array production**

To reveal partitions into four biologically relevant classes, we proceeded by removing genes discriminatory for print

**Table 3: The four classes of experiments identified by the class discovery program (score = 470;  $P < 0.001$ ;  $E = 1.4$ ) after removing 923 genes discriminatory for batches of array production.**

Category <sup>a</sup>	Class 1	Class 2	Class 3	Class 4
BL	8	0	0	0
EWS-C	0	8	1	1
EWS-T & RMS-T	0	0	22	1
NB-C & RMS-C	0	2	2	18

<sup>a</sup>T: tumor samples; C: cell lines

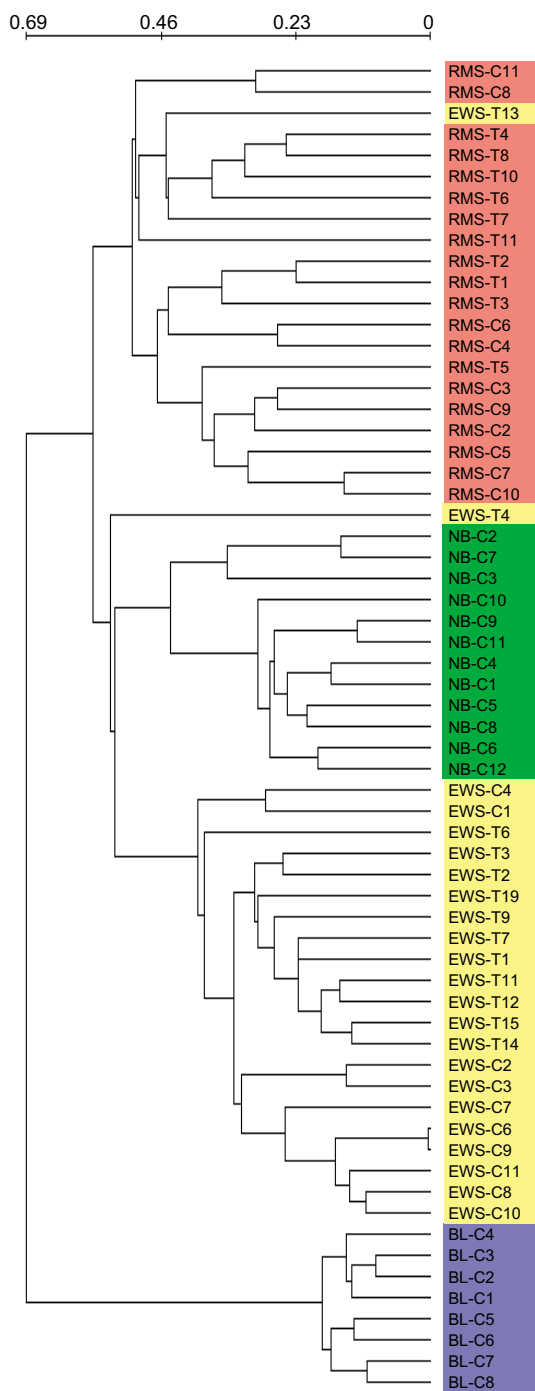
**Table 4: The four classes of experiments identified by the class discovery program (score = 353;  $P < 0.001$ ;  $E = 1.2$ ) after removing 1076 genes discriminatory for batches of array production or cell lines versus tumors.**

Category	Class 1	Class 2	Class 3	Class 4
BL	8	0	0	0
EWS	0	12	10	1
RMS	0	4	13	3
NB	0	0	0	12

batches from the dataset. We removed the 923 genes found to be discriminatory between the 3 major print batches. The best partition found was supported by 470 discriminatory genes ( $P \leq 0.001$ ;  $E = 1.4$ ). The four classes corresponded to BL, EWS cell lines, EWS and RMS tumors, and NB and RMS cell lines, respectively (see Table 3), with 89% of the experiments correctly assigned to these categories. Using the four SRBCT categories instead, the four classes corresponded to correct assignment of 48% of the experiments. For this reduced data set, the score for perfect partitioning into the four SRBCT categories was 390. Hence, the removal of genes discriminatory for print batches had the desired effect: the dominant partition no longer reflected print batches, but instead biologically relevant categories. However, there was a separation of tumors from cell lines.

**Peeling genes discriminatory between cell lines and tumors**

To further reveal relevant partitions, we also removed the 513 genes discriminatory between cell lines and tumors. There was an overlap between these genes and the 923 previously removed print batch discriminatory genes, resulting in the removal of a total of 1076 genes. Using neural networks, Khan *et al.* identified a set of 96 genes with which they were capable of classifying the four categories. Of these 96 genes, 62 remained in our dataset after peeling. For the peeled data set, the best partition found was supported by 353 discriminatory genes ( $P \leq 0.001$ ;  $E = 1.2$ ), including 46 of the 62 genes identified by Khan *et*



**Figure 3**  
 Hierarchical clustering of the 63 SRBCT training experiments. The clustering was performed using the 353 genes discriminatory for the best partition found in the data set reduced for genes discriminating cell lines versus tumors or between print batches. Using the discriminatory genes found by our unsupervised method results in clusters that correspond to the disease categories. The scale shows the linear correlation based distance used to construct the dendrogram.

al. The separation of the SRBCT categories was improved, and with 71% of the experiments correctly assigned, the four classes corresponded to the SRBCT categories (see Table 4). Most of the mis-classifications were mistakes between RMS and EWS experiments, and is due to differences between cell lines and tumors that still remain in the reduced data set. The score for perfect partitioning into the four SRBCT categories was 319, including all 62 of the genes identified by Khan *et al.* Furthermore, hierarchical clustering using the 353 genes significant for the best partition, clearly clustered the SRBCT categories into distinct clusters (Fig. 3). Only two samples were misplaced: EWS-T13 were in the RMS cluster and EWS-T4 was an outlier. We conclude that by removing genes discriminatory between cell lines and tumors or between print batches, we can discover the four SRBCT categories with a high success rate.

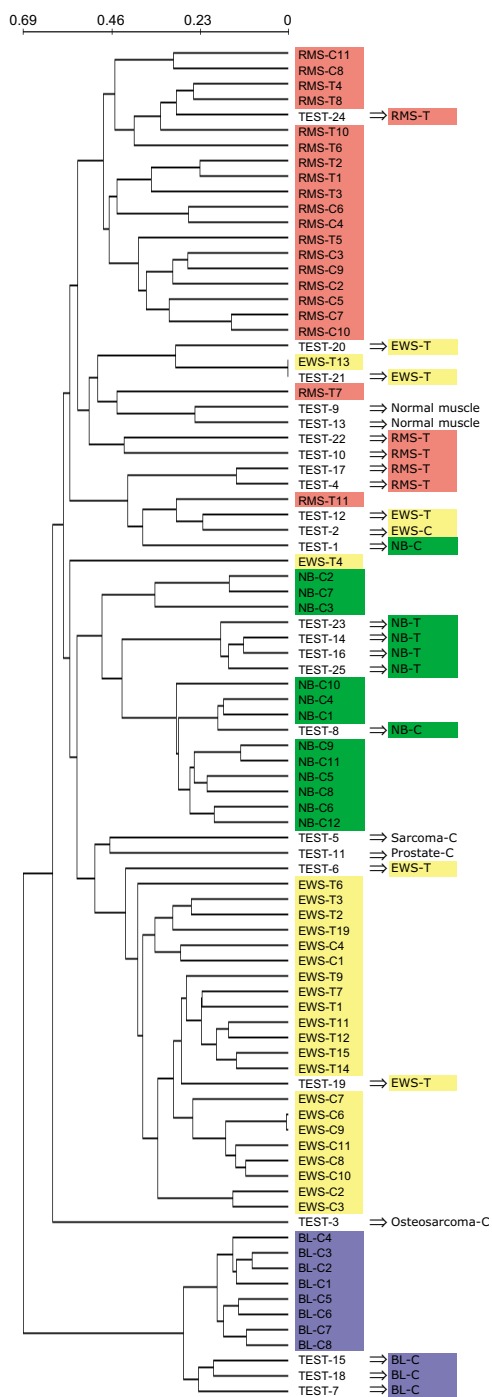
**Results for additional SRBCT test samples**

Finally, we wanted to investigate the robustness of the discovered classes using our unsupervised method on independent test data. Therefore, we included the 25 test experiments and performed hierarchical clustering of all the 88 experiments in the SRBCT data set using the 353 genes (Fig. 4). Again the SRBCT categories clustered into fairly distinct clusters. 15 of the 20 test experiments belonging to one of the four SRBCT categories clustered into clusters dominated by their category. The 5 misplaced experiments (4 EWS and one NB) were in clusters dominated by RMS. Some of these mistakes reflect that many of the test experiments were performed on print batches not corrected for in our class discovery analysis because they were rarely used for training experiments. We conclude that the discriminatory genes found in an unsupervised class discovery analysis can be used to successfully cluster additional experiments. Nonetheless, one should keep in mind that by incorporating which SRBCT category each experiment belongs to in a supervised analysis, one can separate the experiments with 100% success rate [20].

**Conclusions**

We have developed an unsupervised classification method for the discovery of two or more classes of experiments in microarray data. The method has been implemented as a publicly available computer program. We have tested the method on two published gene expression data sets and conclude that the proposed method is effective in finding relevant classes. We are planning to make the program available as a plugin for BASE [26], the open-source database for array data maintained by our group.

When applying our method on these data sets, we have found that the best partitions discovered are relatively insensitive to the cut-off in *P* value. Moreover, even



**Figure 4**  
 Hierarchical clustering of all 88 SRBCT experiments. The clustering was performed using the 353 genes discriminatory for the best partition found using our unsupervised method applied to the training data set reduced for genes discriminating cell lines versus tumors or between print batches. Using these genes, the test samples cluster in clusters dominated by the correct disease category. The scale shows the linear correlation based distance used to construct the dendrogram.

though there is no guarantee that the simulated annealing algorithm finds the optimal score, we found that high-quality local minimas are discovered. Nevertheless, any user of the program will benefit from experimenting with the values of the parameters in the algorithm to explore the particular details of each data set.

The method was designed for unsupervised classification, but it can also be very useful for quality control when classes of experiments are known. It is common to look for an over-abundance of discriminatory genes separating known classes. In such a scenario, it may be useful to seek the partitions having the largest overabundance of discriminatory genes. Thereby, one can rule out potential problems, as highlighted by our example of the print batches for the SRBCT experiments. The SRBCT experiments were performed on arrays produced by the first generation of cDNA microarray printers. Using our class discovery program on more recent data sets our experience is that differences due to print batches are much smaller (data not shown). Nevertheless, we think that our results using the SRBCT data set illustrates how experimental artifacts can be found and corrected for when using our class discovery program. Here, one should keep in mind the crucial importance of random experimental design: if each SRBCT category had been investigated using a unique print batch, there would be no way to disentangle the relevant biology from artifacts. Moreover, these findings illustrate that it is important to know the procedural steps underlying an experiment to be able to interpret discovered classes.

**Availability and requirements**

Project name: MCD – Multiclass discoverer

Project homepage: <http://bioinfo.thep.lu.se/classdiscoverer>

Operating systems: Linux, Windows, Mac OS X

Programming language: Perl

Other requirements: The Perl modules: Algorithm::Numerical::Shuffle, POSIX, Statistics::Distributions, Storable, and Tie::RefHash

License: GNU general public license

Any restrictions to use by non-academics: none

**Authors' contributions**

YL developed and tested this software under the supervision of MR. Both authors wrote the manuscript.



## Acknowledgments

We thank Amir Ben-Dor and Zohar Yakhini for valuable discussions. This work was in part supported by the National Research School in Genomics and Bioinformatics and the Knut and Alice Wallenberg Foundation through the Swegene consortium.

## References

- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ED: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**:531-537.
- Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg Å, Trent J: **Gene-expression profiles in hereditary breast cancer.** *N Engl J Med* 2001, **344**:539-548.
- Ringnér M, Peterson C, Khan J: **Analyzing array data using supervised methods.** *Pharmacogenomics* 2002, **3**:403-415.
- Bittner M, Meltzer P, Chen Y, Jiang Y, Sefror E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampsas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V, Hayward N, Trent J: **Molecular classification of cutaneous malignant melanoma by gene expression profiling.** *Nature* 2000, **406**:536-540.
- Sokal RR, Michener CD: **A statistical method for evaluating systematic relationships.** *Univ Kans Sci Bull* 1958, **38**:1409-1438.
- Hartigan JA, Wong MA: **A K-means clustering algorithm.** *Applied Statistics* 1979, **28**:100-108.
- Kohonen T: *Self-Organizing Maps* 3rd edition. Berlin: Springer; 2001.
- Quackenbush J: **Computational analysis of microarray data.** *Nat Rev Genet* 2001, **2**:418-427.
- Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JJ, Yang L, Marti GE, Moore T, Hudson Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403**:503-511.
- Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406**:747-752.
- Dugas M, Merk S, Breit S, Dirschedl P: **Mdclust – Exploratory microarray analysis by multidimensional clustering.** *Bioinformatics* 2004, **20**:931-936.
- Covell DG, Wallqvist A, Rabow AA, Thanki N: **Molecular classification of cancer: unsupervised self-organizing map analysis of gene expression microarray data.** *Mol Cancer Ther* 2003, **2**:317-332.
- Hsu AL, Tang SL, Halgamuge SK: **An unsupervised hierarchical dynamic self-organizing approach to cancer class discovery and marker gene identification in microarray data.** *Bioinformatics* 2003, **19**:2131-2140.
- Ben-Dor A, Friedman N, Yakhini Z: **Class discovery in gene expression data.** In *Proceedings of the Fifth Annual Conference on Computational Biology (RECOMB): 2001; Montreal* Edited by: Lengauer T, Sankoff D, Istrail S, Pevzner P, Waterman M. ACM Press; 2001:31-38.
- von Heydebreck A, Huber W, Poustka A, Vingron M: **Identifying splits with clear separation: a new class discovery method for gene expression data.** *Bioinformatics* 2001, **17(Suppl 1)**:S107-114.
- Hedenfalk I, Ringnér M, Ben-Dor A, Yakhini Z, Chen Y, Chebil G, Ach R, Loman N, Olsson H, Meltzer P, Borg Å, Trent J: **Molecular classification of familial non-BRCA1/BRCA2 breast cancer.** *Proc Natl Acad Sci USA* 2003, **100**:2532-2537.
- Kirkpatrick S, Gelatt C, Vecchi M: **Optimization by simulated annealing.** *Science* 1983, **220**:671-680.
- Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z: **Tissue classification with gene expression profiles.** *J Comput Biol* 2000, **7**:559-583.
- Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altmann RB: **Nonparametric methods for identifying differentially expressed genes in microarray data.** *Bioinformatics* 2002, **18**:1454-1461.
- Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Atonescu CR, Peterson C, Meltzer PS: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.** *Nat Med* 2001, **7**:673-679.
- Wilcoxon F: **Individual comparisons by ranking methods.** *Biometrics* 1945, **1**:80-83.
- Walpole RE, Myers RH: *Probability and Statistics for Engineers and Scientists* 3rd edition. New York: Macmillan; 1985.
- Kruskal WH, Wallis WA: **Use of ranks in one-criterion variance analysis.** *J Amer Statist Assoc* 1952, **47**:583-621.
- Press WH, Teukolsky SA, Vetterling WT, Flannery BP: *Numerical Recipes in C* 2nd edition. Cambridge, UK: Cambridge University Press; 1992.
- Ringnér M, Edén J, Johansson P: **Classification of expression patterns using artificial neural networks.** In *A Practical Approach to Microarray Data Analysis* Edited by: Berrar DP, Dubitzky W, Granzow M. Boston: Kluwer Academic Publishers; 2002:201-215.
- Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg Å, Peterson C: **Bioarray software environment: a platform for comprehensive management and analysis of microarray data.** *Genome Biol* 2002, **3**:software0003.1-0003.6.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
http://www.biomedcentral.com/info/publishing\_adv.asp

