BMC Bioinformatics

Methodology article

Open Access

Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data

Junbai Wang*[1,4], Trond Hellem Bø[2], Inge Jonassen[2,3], Ola Myklebost[1,4] and Eivind Hovig[1]

Address: [1]Departments of Tumor Biology, The Norwegian Radium Hospital, N0310 Oslo, Norway, [2]Departments of Informatics, University of Bergen, HIB, N5020 Bergen, Norway, [3]Computational Biology Unit, Bergen Center for Computational Sciences, University of Bergen, Norway and [4]Department for Molecular Bioscience, University of OSLO, Norway

Email: Junbai Wang* - junbaiw@radium.uio.no; Trond Hellem Bø - trondb@ii.uib.no; Inge Jonassen - Inge.Jonassen@ii.uib.no; Ola Myklebost - ola.myklebost@biokjemi.uio.no; Eivind Hovig - j.e.hovig@labmed.uio.no

* Corresponding author

Published: 02 December 2003

BMC Bioinformatics 2003, **4**:60

Received: 11 August 2003
Accepted: 02 December 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/60

Abstract

Background: Using DNA microarrays, we have developed two novel models for tumor classification and target gene prediction. First, gene expression profiles are summarized by optimally selected Self-Organizing Maps (SOMs), followed by tumor sample classification by Fuzzy C-means clustering. Then, the prediction of marker genes is accomplished by either manual feature selection (visualizing the weighted/mean SOM component plane) or automatic feature selection (by pair-wise Fisher's linear discriminant).

Results: The proposed models were tested on four published datasets: (1) Leukemia (2) Colon cancer (3) Brain tumors and (4) NCI cancer cell lines. The models gave class prediction with markedly reduced error rates compared to other class prediction approaches, and the importance of feature selection on microarray data analysis was also emphasized.

Conclusions: Our models identify marker genes with predictive potential, often better than other available methods in the literature. The models are potentially useful for medical diagnostics and may reveal some insights into cancer classification. Additionally, we illustrated two limitations in tumor classification from microarray data related to the biology underlying the data, in terms of (1) the class size of data, and (2) the internal structure of classes. These limitations are not specific for the classification models used.

Background

Generally, cancer classification has been based primarily on the morphological appearance of the tumor, but tumors with similar histopathological appearance can follow significantly different clinical courses and show different responses to therapy. Current microarray technology (such as high density oligonucleotide arrays and cDNA arrays) enables researchers to partially overcome this limitation, by enabling tumor subclass identification through global gene expression analysis. Research in this direction has gained wide attention, as illustrated by molecular classification of various clinical samples, such as in acute leukemia, human cancer cell lines and brain tumors [9,12,16], and in tumor subclass prediction, e.g. in diffuse large B-cell lymphoma and breast cancer [1,18]. Several analytical approaches have been applied for this task, such as k-nearest neighbours, weighted voting [9], support vector machines [23], partial least squares [14], hierarchical clustering, artificial neural networks [12], and supervised clustering [5]. Even if these approaches show promising results, classification of clinical samples remains a challenging task due to the complexity and high dimensionality of microarray gene expression data [6].

In this paper, we propose two novel classification models: A combination of optimally selected Self-Organizing Maps (SOMs), followed by Fuzzy C-means clustering (FCC) and the use of pair-wise Fisher's linear discriminant (PFLD). The SOM approach has previously been successfully applied in microarray data analysis [19]. Here, we introduce a new statistical procedure (a stress function) to automatically estimate the boundaries of SOM reference vectors to generate optimally selected SOM. The aim of applying this SOM procedure in the current model is to find map units that can represent the configuration of the input dataset, and at the same time to achieve a continuous mapping from the input gene space to a lattice. The Fuzzy C-means clustering (FCC) algorithm is the fuzzy equivalent of the "hard" k-means clustering, where the assignment of fuzzy membership values can serve as a confidence measure in tumor classification. The Fisher's linear discriminant is a general method in discrimination analysis, which searches for good separation between groups by finding the maximal ratio of the between-group-sum of squares to the within-group-sum of squares. The cross validation of the selected feature is accomplished by a newly developed pair-wise version of Fisher's linear discriminant [10]. The performance of the proposed models was illustrated on four publicly available microarray datasets: leukemia (2 classes) [9], colon cancer (2 classes) [2], brain tumors (5 classes) [16] and NCI cancer cell lines (8 classes) [17], which all have been studied by a number of authors. The last three data sets are well known for their high misclassification rates [6]. Finally, a systematic learning of the internal structure of different tumor classes from microarray expression data has been carried out in this paper.

Results

In the following sections, we demonstrate the performance of the two suggested models using four microarray data sets: (1) leukaemia http://www-genome.wi.mit.edu/cgi-bin/cancer/publications/pub_menu.cgi; (2) colon cancer http://microarray.princeton.edu/oncology/affydata/index.html; (3) brain tumors http://www-genome.wi.mit.edu/mpr/CNS/; and (4) cancer cell lines from the NCI60 data set http://genome-www.stanford.edu/nci60/. All data sets are publicly available. In this work, the search of optimal number of SOM reference vectors was increased from 2 to 1120 and is demonstrated in figure (1). The feature map units selected by model one (manual feature selection) marked by light green square as shown in figure (2), and the empirical cumulative distribution of the significant score d_E of feature genes (clustered in feature map units) shown in figure (3).

Leukemia data

The data set used here is an acute leukemia data set published by Golub et al. The original training data set consisted of 38 bone marrow samples, containing 27 acute lymphoblastic leukemias (ALL) and 11 acute myeloid leukemias (AML). The independent (test) data set contained 20 ALL and 14 AML cases. The gene expression intensities were obtained from Affymetrix high-density oligonucleotide microarrays, containing probes for 6817 genes. A variation filtering procedure [9] was applied to the raw gene expression values before log transformation of the ratios. The data were further standardized to have a mean ratio of zero and variance of one across samples.

In figure (1a), we illustrate the results of using a forward search algorithm to estimate the boundaries of the SOM component plane with 48 training samples. The upper panel of figure (1a) shows a plot of stress versus map size, and the corresponding chi-square test is displayed in the lower panel. This figure clearly demonstrates that the decrease of stress becomes unnoticeable when the number of map size figure (1a) reaches 30, and that the probability P of the chi-square test exceeds the 95% significance level at this point (marked by red vertical lines in figure (1a)). In the subsequent calculations, the two proposed classification models were applied to SOMs with map size 10 × 3. The final mean test error $E(T_{error})$ of model one is 2.4% and model two is 4%, as shown in Table (1). The low misclassification rate obtained from the leukemia data was not surprising, as the expression structure of ALL and AML was rather distinguishable in figure (2a), where 60% of the SOM reference vectors (18 feature map units) were differentially expressed between the two types of tumors. The performance of other machine learning approaches on the leukemia data give test errors ranging from 2.62% to 5.88% Table (1). We also compared our predicted marker genes with the 50 marker genes from the original publication [9], and of these, 28 were detected by our model two.

Colon cancer

Using Affymetrix oligonucleotide arrays, expression levels of 40 tumor and 22 normal colon tissues were measured for 6500 human genes. A dataset containing intensities of 2000 genes in 22 normal and 40 tumor colon tissues was available from [2], where the genes were chosen to give the highest minimal intensity across all samples. The data were pre-processed by transforming the raw intensities to

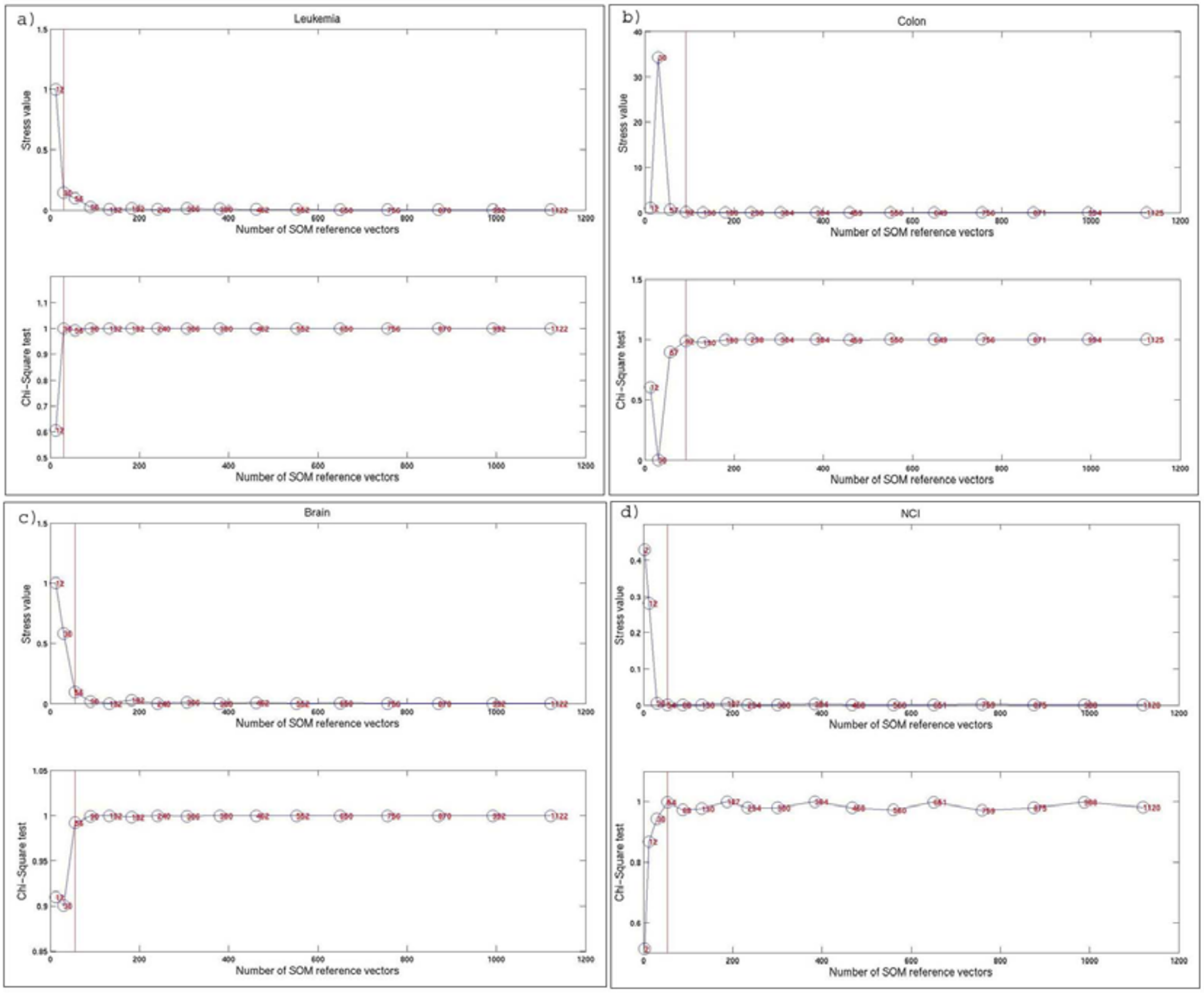

Figure 1

Stress as a function of SOM reference vectors in model one. a) Leukemia data set. **b)** Colon data set. **c)** Brain tumor data set. **d)** NCI60 cancer cell line data set. In each plot, the optimal number of SOM reference vectors was marked by red vertical line and the number of SOM reference vectors was indicated by red text.

base 10 logarithmic values and standardizing each sample to zero mean and variance one.

For the colon data set, we found that 92 (23 × 4) SOM reference vectors may well explain the input gene space of all training samples, as shown in figure (1b), where the P value of the chi-square test also increases beyond the 95% threshold. Thus, the classification of colon tissues into normal and tumor tissues was based on the data distribution of 92 SOM reference vectors. Because the classification results obtained from FCC were poor, in the application of model one, we chose to use the mean component planes of each tissue type, shown in figure (2b), for the visualization of gene expression structure and for manual feature selection. From figure (2b), we found that the expression patterns of normal and tumor colon tissue were extremely similar, where only around 8.7% of SOM reference vectors (8 feature map units) had distinct expression levels between two types of colon tissues (marked by light green squares). This may explain the poor classification results obtained on the colon data set using other methods, see Table (1). The mean test error

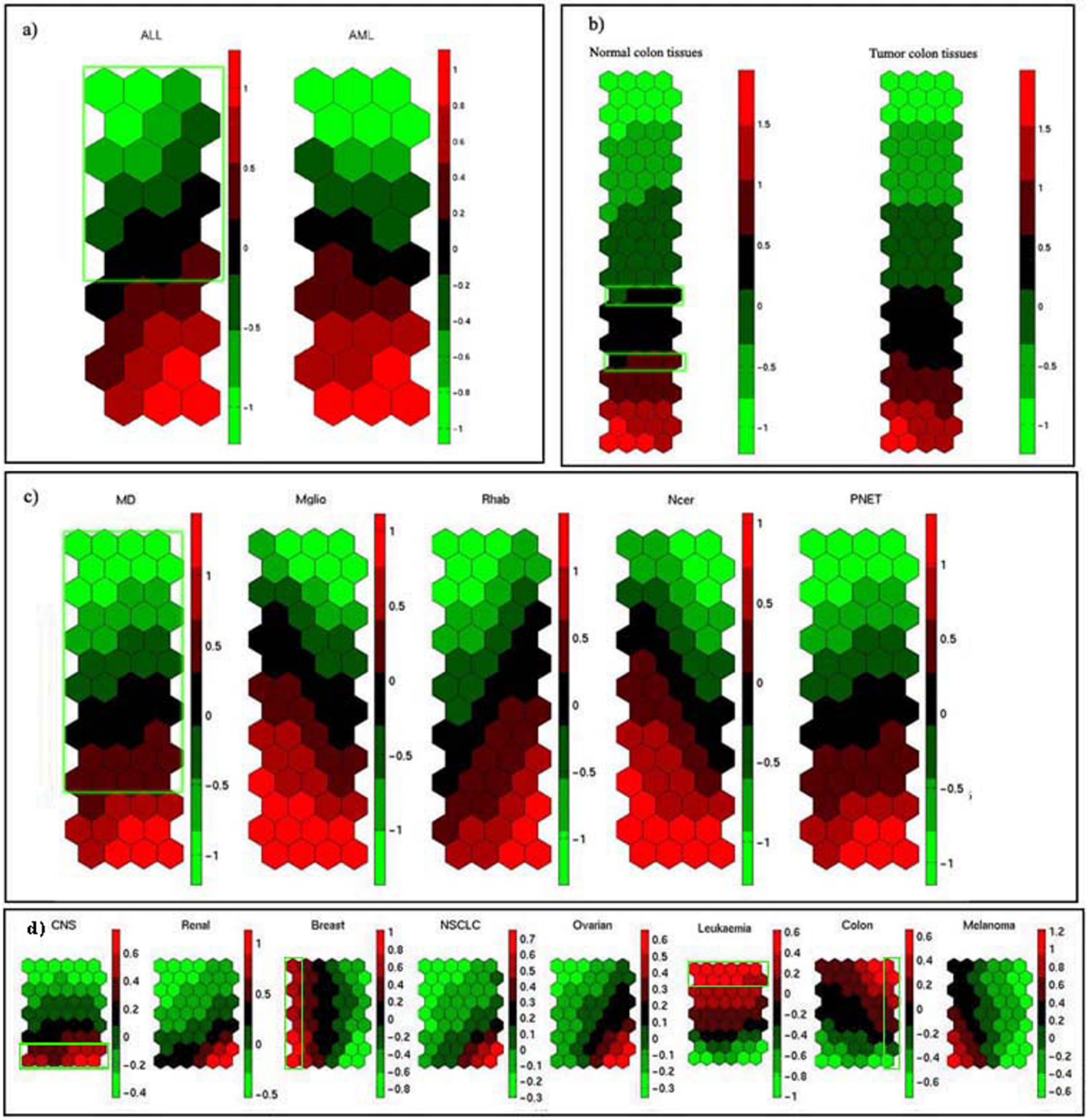

Figure 2

Weighted/mean SOM component plane. a) Weighted component planes of ALL and AML type of tumors in leukemia data set. **d)** Mean component planes of Normal and Tumor colon tissues in colon data set. **c)** Weighted component planes of MD, Mglio, Rhab, Ncer and PNET type of tumors in brain tumor data set. **d)** Weighted component planes of CNS, Renal, Breast, NSCLC, Ovarian, Leukemia, Colon and Melanoma type of cancer cell lines in NCI60 data set. In each plot, feature map units that identified by the manual feature selection of model one were marked by light green squares and detailed information of selected SOM map units can be found in our web supplement [22]. The color scale of weighted/mean component plane represented the expression level of SOM reference vectors, where red indicates high expression and green indicates low expression

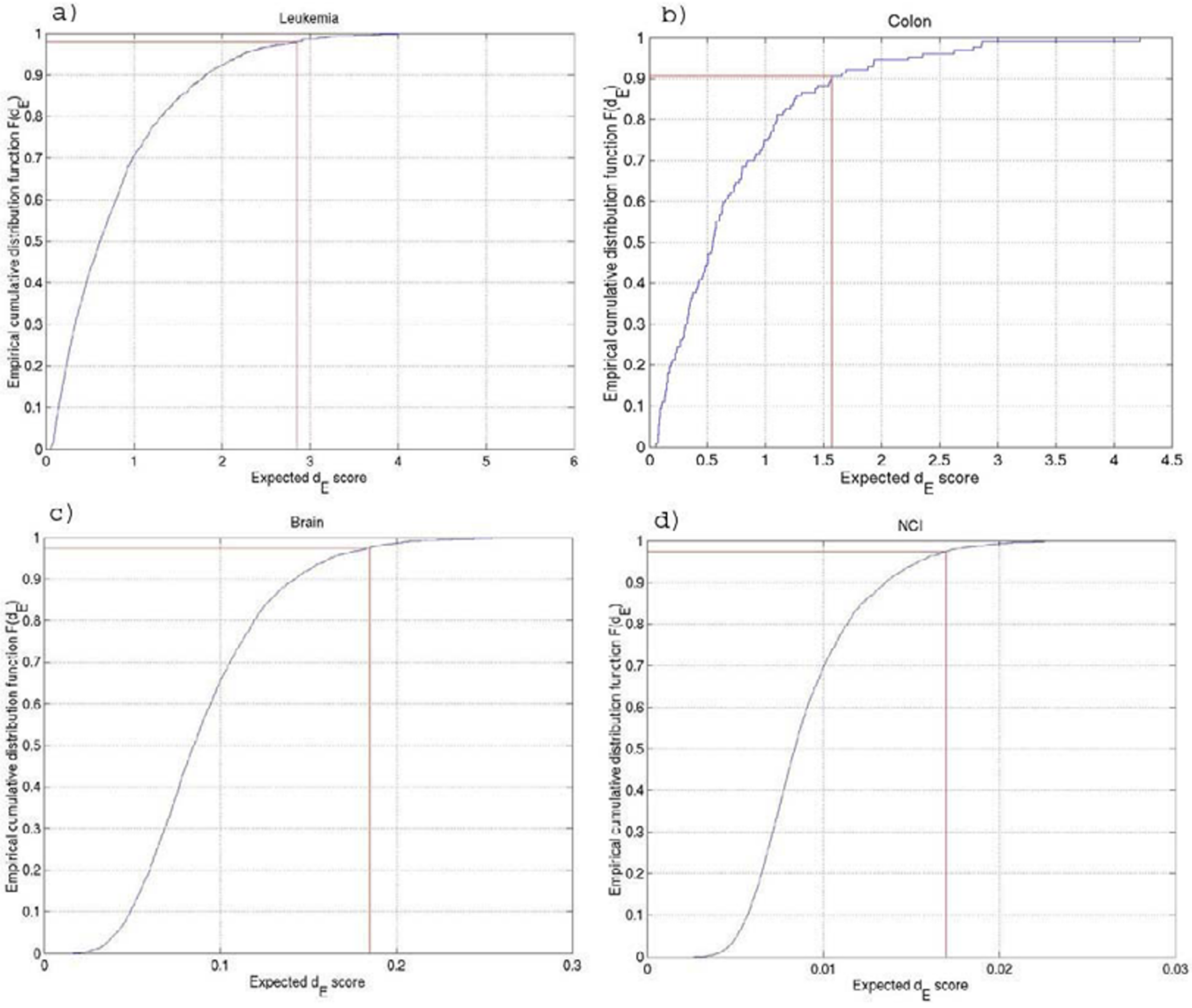

Figure 3
Empirical cumulative distribution of the significant scores d_E. a) Leukemia data set. **b)** Colon data set. **c)** Brain tumor data set. **d)** NCI60 cancer cell line data set. In each plot, the percentage of $F(d_E)$ that maximizes the classification performance was marked by red smooth line.

$E(T_{error})$ in model one is 12.27%, and in model two it is 11.36%. The test error of both models was superior to supervised clustering, and approaches the lowest misclassification rate (9.68%) we have found in the literature (Table 1). This low value was achieved by the use of support vector machines. For the colon data set, no genuine marker genes were available. Therefore, we verified biological functions of our predicted marker genes (50 genes) from the model two, and found that 7 genes were ribosomal protein genes and smooth muscle related genes. This finding was in agreement with previous studies [2] in that expression levels of ribosomal protein genes are relatively low in normal colon tissues and high in colon tumor tissues; and conversely that smooth muscle related genes had high intensities in normal tissues and low intensities in tumors.

Brain tumors

Having obtained good performance on datasets with two classes, we next test the proposed models on a more complicated data set, consisting of 42 brain tumor samples containing 10 medulloblastomas (MD), 10 maglignant

Table 1: Comparison of test error against literature and an independent test. a) The test error of supervised clustering from [5]. b) The test error of weighted voting on leukemia data from [9], on brain tumor data from [16]. c) The test error of support vector machines from [7]. d) The test error of the boosting method on leukemia and NCI data from [6], on colon data from [4]. e) The test error of nearest neighbors on leukemia and NCI data from [6], on colon data from [4]. f) The test error of an independent test, by using the same data set that had been tested on our proposed models with the t-test and Fisher's linear discriminant. Here, NA means that the test error is not available, because we either did not find classification results in the literature (i.e. weighted voting, support vector machines, boosting and nearest neighbors) or the model was not able to perform multiple class classification.

	Leukemia (2 class)	Colon (2 class)	Brain (5 class)	NCI (8 class)
Model one (manual feature selection): mean test error	2.4%	12.27%	10%	24%
Model one (manual feature selection): median test error	4%	13.64%	8.82%	22.73%
Model two (automatic feature selection): mean test error	4%	11.36%	13.53%	22.27%
Model two (automatic feature selection): median test error	4%	11.36%	14.71%	22.73%
a) Supervised clustering	2.62%	15.95%	16.86%	26.5%
b) Weighted voting	4.17%	NA	16.67%	NA
c) Support vector machines	5.88%	9.68%	NA	NA
d) Boosting	2.94%	17.7%	NA	42.86%
e) Nearest neighbors	2.94%	19.4%	NA	42.86%
f) T-test plus Fisher's linear discriminant	4%	18%	NA	NA

gliomas (Mglio), 10 atypical teratoid/rhabdoid tumors (Rhab), 8 primitive neuroectodermal (PNET) and 4 normal cerebella tumors (Ncer). The gene expression profiles of 42 patient samples were obtained from oligonucleotide microarrays containing probes for 6817 genes. The raw expression data were subjected to a variation filter that excluded genes showing minimal variation across all samples [16]. The expression rates were log transformed and normalized by standardizing each sample to a mean of 0 and a variance of 1.

With this data set, 56 SOM reference vectors (14 × 4) were considered as a reasonable subspace of the original high dimensional expression data, with a P = 99.23% for a chi-square test (figure (1c)). Based on the configuration of 56 SOM reference vectors, we applied two classifier models to predict the marker genes of each type of tumor class. From model one, the difficulties of multi-class classifications (the number of classes greater than 3) were easily visualized by the expression patterns of five brain tumor classes (weighted component planes). In figure (2c), a large number of overlapping structures among five types of brain tumors were readily visible, indicating that some tumor classes may share some of the same activated genes with other classes, *i.e.* some map units were highly expressed across all five classes (the last two rows of the weighted component planes in figure (2c)). Mglia and Ncer type tumor shared a number of map units that had the same trend in up regulation or down regulation, and the expression pattern of PNET type tumors had strong correlation with MD. Therefore, the manual selection of feature map units in model one was based on the internal structures of each class (marked by light green squares in figure (2c)). In model two, the selection of feature map units was only considered by a statistical significance test (PFLD) that has commonly been used in other types of classification models [5]. Both our models produced classification errors lower than those previously reported. The mean test error of model one was 10%, and model two, which is unsupervised, resulted in a 13.53% error. This result also indicated that model one was more robust on noisy data. Additionally, we found that most of the misclassified samples appeared in the PNET type of tumors. The tumors were often falsely labelled as Mglia, MD or Ncer. This problem was also mentioned in the original paper [16], where weighted voting was applied to the same 42 samples. They found 7 misclassifications, and 4 of them were primitive neuroectodermal tumors. This demonstrated that the tumor classes shared common expression patterns (figure 2c, MD with PNET and Mglia with Ncer), which may dramatically reduce the performance of a machine learning algorithm, resulting in an increased error rate. Thus, we concluded that the internal structure of the catalogue of tumor classes has a potential effect in tumor classification and marker gene prediction. Additionally, we compared our predicted marker genes (around 110 genes) of the model two with 50 marker genes that had been manually selected in the original paper [16], and found 20 of them were identical in both studies.

NCI60 data

The NCI60 data set contained 61 cell lines derived from human cancers from a variety of tissues and organs; 5 central nervous system (CNS), 9 renal, 9 breast, 9 non-small-lung (NSCLC), 6 ovarian, 8 leukemia, 7 colon and 8 melanoma, and the data set included approximately 8000 distinct genes in each cDNA array [17]. Here, we tested our two models on a data subset with 6665 genes and 61 samples, where all genes had less than 20% missing values across the 8 classes. The missing data were imputed by the k-nearest neighbour algorithm [15], raw ratios were log2 transformed and standardized by using a mean of 0 and a variance of 1 across samples.

We first summarized the input gene space into an optimally selected subspace, SOM, where the chi-square test of the efficiency of dimensional reduction figure (1d) suggested that 54 SOM reference vectors would be a good approximation of the features of the original number of genes (P = 99.97%). According to the "configuration" of this optimally selected SOM (map size 9 × 6), the two proposed models were used to predict marker genes and to classify test samples into 8 classes. An overview of 8 weighted component planes (8 tumor classes), as shown in figure (2d), displayed a high degree of interconnection among the 8 tumor classes. For instance, the renal type of cancer cell lines had an expression structure almost identical to the NSCLC cell lines, where highly expressed genes only gathered at the lower right part of the component planes. Also, both renal and NSCLC cell lines showed a strong similarity with ovarian cancer. Activated genes in breast cancer cell lines were often shared by CNS, leukemia and melanoma type of cancers; and same overlapping structures also existed between leukemia and colon cancer cell lines. Moreover, the SOM also indicated that the internal structure of NCI60 data was much more complicated than the brain tumor data, *i.e.* 3 or 4 cancer classes shared the same gene cluster, while 8 classes were to be classified. Because of this gene cluster sharing, it was not unexpected that our mean test error rate on NCI60 data was dramatically higher than for the other data sets, for model one $E(T_{error})$ = 24% and model two $E(T_{error})$ = 22.73%. Although the misclassification rate was high compared to that achieved on the other three data sets, our proposed models performed better on the NCI60 data than any other classification models available in the literature (Table (1)). As can be seen, the misclassification rates reported from other approaches varied between 26.5% and 42.86%. Additionally, we identified two groups of cancer classes, (NSCLC and ovarian cancer cell lines) and (breast cancer cell lines, CNS and melanoma cancer cell lines), where incorrectly assigned class labels often came from the same group. There were around 83% misclassified samples belonging to the above two groups, emphasizing that the internal structure of the classes and class size had a strong influence on the performance of classification models. To test the robustness of our predicted marker genes (around 140 genes) by model two, we collected a list of genes (around 400) known to be related to tissue characteristics in the cell lines [17] and found that 34 of our predicted marker genes belonged to this list.

Discussion

Microarray data analysis has some similarity with information theory, where one of the central tasks is compression. In order to obtain optimal compression, an optimal machine learning approach that discovers and exploits subtle patterns in the data is required. For that reason, given the ability to ignore the noise inherent in expression data, and given the ability to find expression pattern features among various tumor classes, then it would be possible to identify the real marker genes of each type of tumor class. Our proposed models both meet the above requirements usefully, where the noisy gene expression profiles are first summarized into SOM with optimally selected map units (estimated by stress function), then the feature selection is performed on the weighted/mean component plane, by either manual feature selection (model one), or automatic feature selection (model two). The test error rates obtained from our models were generally better than those reported for other classification methods, *i.e.* supervised clustering, weighted voting and nearest neighbours etc. In particular, model one provided the best misclassification rate on brain tumor data (5 classes, around 6% improvement) and NCI60 data (8 classes, around 4% improvement) when compared with available results from the literature (see Table (1)). Given the improvement from the proposed models, the models are potentially very attractive for multi-class tumor classification using gene expression data.

We have also compared the performance differences among various classification methods according to the class size of data. Normally, simple discrimination methods and well designed classification models (i.e. our models considering the expression features) have similar performance on binary classes, i.e. the test error of the proposed models and supervised clustering on leukemia is between 2.4% and 4%, on colon data between 11.36% and 15.95%; and other methods had 2.94% to 5.88% on leukemia, 9.68% to 19.4% on colon data Table (1). However, clear differences were found in multi-class problems where the designed classification models gave an almost 50% reduction in the misclassification rate compared to others (Table 1). More detailed discussions of comparisons may be found in [5]. We further investigated the possible effect of different feature selection procedures on tumor classification and marker gene prediction. In model one, the prediction of marker genes is determined by the combination of internal structures of classes and statistical significance tests of expression levels. In model two and in supervised clustering, only the statistical significance tests are considered. In other words, model one may avoid predicting genes that have statistical significance, but no real biological significance across all tumor samples [6]. This is a likely explanation for why model two had a similar performance as supervised clustering,

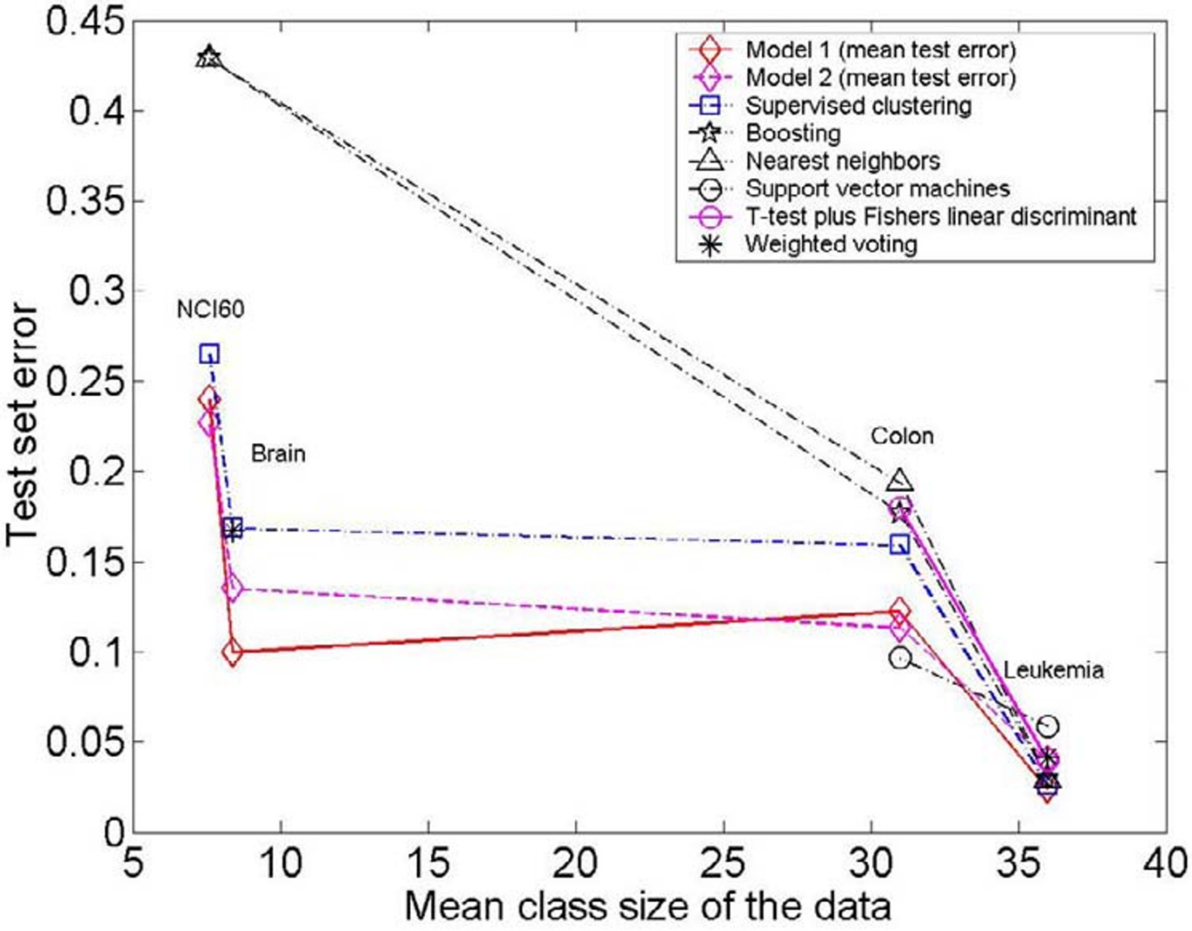

Figure 4
Test set error as a function of mean class size of the data set.

whereas our model one had better performance than supervised clustering on some data sets, as seen in figure (4). It may also be inferred that the link to other information, *i.e.* internal structure of the classes, clinical data of the samples, gene sequence information or gene functional information [13], with the statistical significance test of gene expression data in the prediction of markers, may lower the test error rate of tumor classification. Thus, feature selection plays an essential role in tumor classification and marker gene prediction using microarray data.

Finally, the rich visualization features (weighted/mean component plane figure (2)) of the proposed models provide an opportunity to make a systematic examination of other possible effects that may influence the tumor classification using expression profiles. As mentioned previously, we identified two reasons for the poor classification performance that were related to the biology underlying the data, rather than to the technical aspects of the classification models. They are (1) class size and (2) the internal structure of the classes. Figure (4) shows the test set error rate as a function of the mean class size of the data. There is a clear trend for the test error rate to decrease with increasing class size. The reason is that classes of large size are more likely to be learned by classification models. The second important factor that determines the test error rate of a data set is the internal structure of the classes. figure (2) shows the gene expression structures of various tumor classes in each data set. As can be seen for binary classes, the leukemia data set had a more clear and distinguishable structure than did the colon data set. Their relative test error rate in figure (4) showed a strong correlation with their class structure, while data sets with more overlapping structures among classes gave higher test error rates. The same phenomenon was also found with multiple classes, *i.e.* brain tumors and NCI60 data as shown in figure (4). Such overlapping structures (the same genes shared by a number of tumor classes) are biologically understandable, as genes tend to work in a complex and highly interacting manner [13]. Current microarray data only represent a snapshot of the dynamic gene interactions in the real world, and the static state of experimental data lack the information to describe the interaction of different biochemical processes inherent in biology. Therefore, the first factor (class size) may sometimes be easily overcome by an increase in the sample size of each class, i.e. collecting tumor tissue from more patients. The second factor is quite difficult to overcome. Either new techniques in experimental design or extra information (i.e. internal structure of classes or gene functional information etc.) is needed to guide the classification models. However, the second factor will always set a limitation for tumor classification using microarray data. That is to say, to a certain degree a number of tumor samples will not be correctly classified, and the misclassification will occur for every classification model if there are intersections among multiple tumor classes.

Conclusions

We have proposed two novel models for classification of tumors using microarray data. Our model one gave the best test error rate on four published data sets, when compared to other results in the literature. Particularly for multi-class problems, our models represent approximately a 4% improvement (NCI60 dataset) in error rate compared to other classification models. Additionally, we explored the importance of feature selection on tumor classification and marker gene prediction. The main limitations in tumor classification from microarray data are related to the biology underlying the data in terms of (1) the class size of data and (2) the internal structure of classes. These limitations are not aspects of the classification models used. A future development of our approach may be to design a numerical score to assess the complexity of overlapping structures among multiple tumor classes.

Methods

Estimation of boundaries in the SOM component plane using a stress function

The SOM component plane

The SOM approach is one of the most popular machine learning approaches, and is based on unsupervised competitive learning [11]. In our models, the SOM acts as a dimensionality reduction tool, which reassembles the data distribution of raw expression profiles by a two-dimensional SOM component plane with an optimally selected map size. Each component plane describes a gene expression structure of a tumor sample or a class, and the component plane is displayed by taking from each map the value of the component, and depicting this as a color on the grid. As presented in our previous study [21], a SOM component plane may reveal the essential biological difference among various tumor classes found through microarray data. But it remains a challenge to systematically estimate the boundaries of SOM reference vectors.

Stress function

To estimate the number of SOM reference vectors that best fit the data distribution of a high dimensional input space, we used a forward searching algorithm with a stress function to detect the boundaries of SOM reference vectors. The general form of the stress function is as follows:

$$[\Sigma\Sigma\ (F_{ij} - D_{ij})^2/\ \Sigma\Sigma D_{ij}^2]^{1/2}$$

In this equation, F_{ij} and D_{ij} is a dissimilarity measure (1 - the correlation coefficient of SOM reference vectors) in m's and m+1's iteration, m = 1,2, ... maximum number of iterations. The forward searching algorithm starts with 2 map units, during each iteration another 2 map units are added in both the row and column of the SOM. Then, a stress value is calculated. We expect that the true dimensionality of data will be revealed by the rate of decline of stress as the map units increase. A chi-square test is used to estimate the quality of the fit of newly increased SOM reference vectors, and we assume that the stress value has an asymptotic chi-square distribution with the degrees of freedom given as: (the difference of SOM row number units between adjacent iterations - 1) times (the difference of SOM column number units between adjacent iterations - 1). If the probability of the chi-square test is greater than 0.95, then the forward searching algorithm stops. In this case, we increase the number of map units until there is no significant change in the "configuration" of SOM reference vectors. At this point, we consider this the optimal number of SOM reference. We used the SOM toolbox built into Matlab [20] to perform the SOM calculations and to produce the SOM visualizations.

Manual feature selection: fuzzy c-means clustering and weighted/mean SOM component planes

Fuzzy c-means clustering

Fuzzy c-means clustering (FCC) has previously been used by Gasch etc. to identify overlap clusters of yeast genes based on microarray gene expression data [8]. In their study, the use of FCC resulted in a good performance when extracting biological insights from gene expression data. Below is a brief description of the FCC algorithm [3]: Given an input data space $X = \{x_{1m}, x_{2m},...,x_{nm}\}$, where n is the number of tumor samples and m is the dimension of the gene space, we assume the existence of C clusters (tumor class), whose centers are unknown and are given the initial values $C=\{y_{10}, y_{20}, \ldots y_{C0}\}$. The degree of membership of x_{im} in class C_k is denoted by a C by n matrix U, the elements of U must satisfy the following constraints:

$\sum_{i=1,\ldots C} U_{ik=1}$; $U_{ik} \in [0,1]$;

we are interested in minimizing the following cost:

$J(U) = \sum_{i=1,\ldots n} \sum_{j=1,\ldots C} U^{\alpha}_{ji} || X_i - C_j ||^2, \alpha > 1$;

The parameter α controls the degree of fuzziness in the process. The following algorithm finds a solution that converges to a local minimum of J(U). (1) Initialize C and U randomly. (2) set $\alpha > 1$. (3) For $1 \leq i \leq n$ and $1 \leq j \leq C$ calculate membership values $U_{ji} = 1/[\sum_{k=1,\ldots C} (||X_i - C_j ||^{2/(\alpha-1)}/||X_i - C_k||^{2/(\alpha-1)})]$. (4) For $1 \leq j \leq C$ update the cluster centers by $C_j = (\sum_{i=1,\ldots n} U^{\alpha}_{ji} X_i)/(\sum_{i=1,\ldots n} U^{\alpha}_{ji})$. (5) The process stops when the difference in the U_{ji}'s between two consecutive iterations is smaller than a given tolerance ε; otherwise go to step 3.

It may be especially advantageous to introduce fuzzy sets in tumor classification, where frequently unlabeled tumor samples may not necessarily be clear members of one class or another. Using crisp techniques, an ambiguous sample will be assigned to one class only, resulting in an aura of precision and definiteness to the assignment that is not warranted. On the other hand, fuzzy techniques will specify to what degree the object belongs to each class, which is information that will frequently be useful [3]. For instance, if we apply the FCC on the optimally selected SOM, we then use its fuzzy membership values to construct a weighted SOM component plane of each type of tumor class figure (2). By visual inspection of the component plane, we may identify some important expression features of the tumor class.

Weighted/Mean component plane

By introducing the fuzzy membership value U_{ks} into the SOM component plane, we can generate the weighted component plane $W_{Ck} = [\sum_{s=1,\ldots p} (U_{ks} W_{rs})]/(\sum_{s=1,\ldots p} U_{ks})$ for each type of tumor class C_k, where W_{rs} is the SOM reference vectors, r is the map size and p is the number of tumor samples that is labelled as class C_k. The mean component plane $_k = \sum_{s=1,\ldots p} W_{rs}/p$ simply represents the data distribution of the mean SOM component plane of p tumor samples, where the class label is given by prior knowledge. The exploration of clustering structures, and the manual selection of SOM feature map units can be easily achieved by parallel visualization of the weighted/mean component plane of all tumor classes. By examination of the weighted/mean component plane, we obtain an improved understanding of the gene expression structure of each class, and thus in the prediction of marker genes.

Automatic feature selection: pair-wise Fisher's linear discriminant

Our goal might be to automatically identify a set of genes or feature SOM reference vectors that have significant expression difference across all classes. The difference score d(i) to determine the significance of these changes can be defined in terms of the Fisher's linear discriminant [10]. For example, a set of n tumor samples that consists of k non-overlapping subclasses, such that the tumor subtype $y_j \in \{1,2,\ldots,k\}$. Define $C_k = \{j: y_j = k\}$. Let n_k = number of tumor samples in C_k. The average gene expression in each subclass is $x_k(i) = \sum_{j \in C(k)} x_j(i)/n_k$ and the average gene expression for all n samples is $x(i) = \sum_j x_j(i)/n$. Then define: $r(i) = \{(\sum_k n_k/\prod_k n_k) \sum_k n_k [x_k(i) - x(i)]^2\}^{1/2}$ is the between-group-sum of squares, $s(i) = \{[\sum_k (1/n_k)/\sum_k (n_k - 1)] \sum_k \sum_{j \in C(k)} [x_j(i) - x_k(i)]^2\}^{1/2}$ is the within-group-sum of squares and $d(i) = r(i)/(s(i) + s_0)$; $i = 1,\ldots,m$; m is the number of available genes; the value of s_0 was chosen as the median value of s(i).

An important issue of the prediction of significant features is cross-validation, which tries to minimize potentially confounding effects from the differences in various tumor samples. Cross validation of the selected feature can be accomplished by leaving out a portion of the data, building a prediction rule on the remaining data. For that reason, we developed a pair-wise Fisher's linear discriminant (PFLD) by randomly deleting part of (i.e., 5%) tumor samples from each class C_k at a time, followed by pair-wise comparison of all the classes and computing the difference score $d_p(i)$. The whole process is repeated P times and the final expected difference is $d_E(i) = \sum_p d_p(i)/P$. We set P equal to 100 to ensure that the pair-wise Fisher's linear discriminant analysis provides a more realistic estimate of the significant feature than one can expect when applying the predictor to independently collected tumor samples. For the selection of significant genes that can maximize the classification performance, we fit the expected significance score d_E to an empirical cumulative distribution function $F(d_E)$ that is defined as $F(d_E)$ = (Number of significant scores $\leq d_E$) / (Total number of significant scores) for all values in d_E. Thus, the significant genes ($F(d_E) \geq 90\%$) may be automatically identified.

Classifier design: the selection of marker genes

In microarray data analysis, a more ambitious, difficult, and potentially useful computational problem than clustering, *i.e.* classifier design, refers to the identification of a few typical genes from all available gene expression profiles. Once they are defined, a classifier is capable of labeling every tumor sample in the entire sample collection. Sometimes this is termed as supervised learning (in this

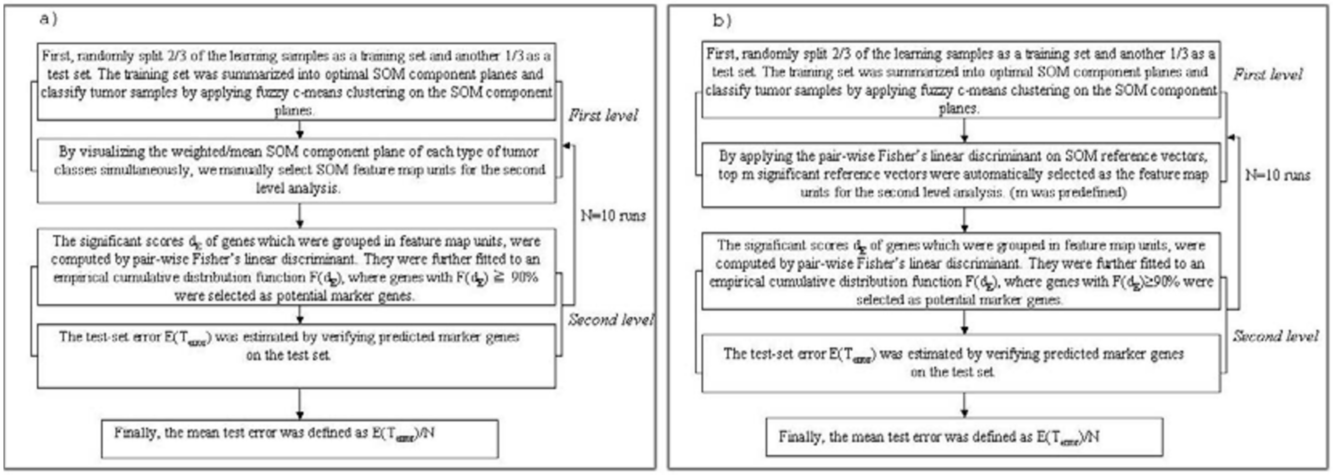
a)
First, randomly split 2/3 of the learning samples as a training set and another 1/3 as a test set. The training set was summarized into optimal SOM component planes and classify tumor samples by applying fuzzy c-means clustering on the SOM component planes.
By visualizing the weighted/mean SOM component plane of each type of tumor classes simultaneously, we manually select SOM feature map units for the second level analysis.
The significant scores d_Σ of genes which were grouped in feature map units, were computed by pair-wise Fisher's linear discriminant. They were further fitted to an empirical cumulative distribution function $F(d_\Sigma)$, where genes with $F(d_\Sigma) \geqq 90\%$ were selected as potential marker genes.
The test-set error $E(T_{error})$ was estimated by verifying predicted marker genes on the test set.
Finally, the mean test error was defined as $E(T_{error})/N$
First level
N=10 runs
Second level
b)
First, randomly split 2/3 of the learning samples as a training set and another 1/3 as a test set. The training set was summarized into optimal SOM component planes and classify tumor samples by applying fuzzy c-means clustering on the SOM component planes.
By applying the pair-wise Fisher's linear discriminant on SOM reference vectors, top m significant reference vectors were automatically selected as the feature map units for the second level analysis. (m was predefined)
The significant scores d_Σ of genes which were grouped in feature map units, were computed by pair-wise Fisher's linear discriminant. They were further fitted to an empirical cumulative distribution function $F(d_\Sigma)$, where genes with $F(d_\Sigma) \geq 90\%$ were selected as potential marker genes.
The test-set error $E(T_{error})$ was estimated by verifying predicted marker genes on the test set.
Finally, the mean test error was defined as $E(T_{error})/N$
First level
N=10 runs
Second level

Figure 5
Diagrams of proposed two classifier models. a) The model one with the manual feature selection. **b)** The model two with the automatic feature selection.

context we are learning the genes' biological contribution in each type of tumor). By the combination of above three techniques (optimally selected SOM, FCC and PFLD), we have created two types of classifier models. Model one is implemented with manual feature selection and model two is applied with automatic feature selection to predict the marker gene of each type of tumor class. The detailed illustration of these models is shown in figure (5). Some features of the proposed models will be explained here: First, the preprocessing of microarray data was essential in that different choices may affect the outcome of comparison. Thus, we followed exactly the preprocessing protocol in [5], i.e. thresholding, filtering, a logarithmic transformation, and a standardization of each dataset that enables us to have a fair comparison with other methods. After the preprocessing, each dataset was subjected to model one and model two (see figure (5) for the further details), where no preprocessing steps were involved in the cross validation. Secondly, for both models, the marker genes obtained from each run will subsequently be used to predict class labels of the test dataset (randomly selecting 1/3 of all learning samples) and to calculate the test-set error T_{error}. Finally, for a possible comparison between two proposed models, the number of feature map units (m) used by the automatic feature selection (model two) is defined as m = number of tumor classes times β, where β is a parameter that leads m has the closest value to the size of feature map units that were identified by manual feature selection (model one).

Authors' Contributions

JBW designed and carried out the study and drafted the manuscript. THB implemented JAVA program for T-test and Fisher's linear discriminant. IJ and OM participated in validation of the study. EH supervised the study. All authors read and approved the final manuscript.

Acknowledgements

We thank Bjarte Dysvik for very helpful discussion. JBW was supported by the Norwegian Cancer Society http://www.kreft.no. IJ was supported by grants from the Research Council of Norway and by the TEMBLOR grant from the European Commission.

References

1. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Staudt LM: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2002, **403:**503-511.
2. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.** *Proc Natl Acad Sci USA* 1999, **96:**6745-6750.
3. Bezdek JC, Pal SK: **Fuzzy models for pattern recognition method that search for structures in data.** *IEEE press New York* 1992.
4. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Washington U: **Tissue classification with gene expression profiles.** *RECOMB Tokyo Japan* 2000.
5. Dettling M, Buhlmann P: **Supervised clustering of genes.** *Genome Biol* 2002, **3:**12.
6. Dudoit S, Fridlyand J, Speed Tp: **Comparison of discrimination methods for the classification of tumors using gene expression data.** *J Am Stat Assoc* 2002, **97:**77-87.
7. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: **Support vector machine classification and validation of cancer tissue samples using microarray expression data.** *Bioinformatics* 2000, **16:**906-914.

8. Gasch AP, Eisen MB: **Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering.** *Genome Biol* 2002, **3:**11.
9. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.
10. Johnson RA, Wichern DW: **Applied multivariate statistical analysis.** *Prentice-Hall New Jersey* 1998.
11. Kohonen T: **Self-organizing maps.** *Berlin Springer* 1997.
12. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu 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.
13. Lægreid A, Hvidsten TR, Midelfart H, Komorowski J: **Predicting gene ontology biological process from temporal gene expression patterns.** *Genome Rese* 2003, **13:**965-979.
14. Nguyen DV, Rocke DM: **Tumor classification by partial least squares using microarray gene expression data.** *Bioinformatics* 2002, **18:**39-50.
15. Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, Russ B Altman: **Missing value estimation methods for DNA microarrays.** *Bioinformatics* 2001, **17:**520-525.
16. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, Mclaughlin ME, Kim JYH, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR: **Prediction of central nervous system embryonal tumor outcome based on gene expression.** *Nature* 2002, **415:**436-442.
17. Ross TD, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Rijn MVD, Waltham M, Pergamenschikov A, Lee JCF, Lashkari D, Shalon D, Myers TG, Weinstein JN, Bostein D, Brown PO: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet* 2000, **24:**227-235.
18. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, Rijn MVD, Jeffrey S, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lønning PE, Børresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proc Natl Acad Sci USA* 2001, **98:**10869-10874.
19. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci USA* 1999, **96:**2907-2912.
20. Vesanto J: **SOM-Based data visualization methods.** *Intelligent Data Analysis* 1999, **3(2):**111-126.
21. Wang JB, Delabie J, Aasheim HC, Smeland E, Myklebost O: **Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study.** *BMC Bioinformatics* 2002, **3:**36.
22. Wang JB, Bø TH, Jonassen I, Myklebost O, Hovig E: **Supplementary information for "Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data".** 2003 [http://www.uio.no/~junbaiw/mfuzzy/index.html].
23. Yeang CH, Ramaswamy S, Tamayo P, Mukherjee S, Rifkin RM, Angelo M, Reich M, Lander E, Mesirov , Golub T: **Molecular classification of multiple tumor types.** *Bioinformatics* 2001, **17:**S316-S322.