Gene selection from microarray data for cancer classification—a machine learning approach

Yu Wang [a,*], Igor V. Tetko [a], Mark A. Hall [b], Eibe Frank [b], Axel Facius [a], Klaus F.X. Mayer [a], Hans W. Mewes [a,c]

[a] *Institute for Bioinformatics,German Research Center for Environment and Health,Ingolstädter Landstraβe 1, D-85764 Neuherberg, Germany*

[b] *Department of Computer Science, University of Waikato, Private Bag 3105, Hamilton, New Zealand*

[c] *Department of Genome-Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, Alte Akademie 10, D-85354 Freising-Weihenstephan, Germany*

Abstract

A DNA microarray can track the expression levels of thousands of genes simultaneously. Previous research has demonstrated that this technology can be useful in the classification of cancers. Cancer microarray data normally contains a small number of samples which have a large number of gene expression levels as features. To select relevant genes involved in different types of cancer remains a challenge. In order to extract useful gene information from cancer microarray data and reduce dimensionality, feature selection algorithms were systematically investigated in this study. Using a correlation-based feature selector combined with machine learning algorithms such as decision trees, naive Bayes and support vector machines, we show that classification performance at least as good as published results can be obtained on acute leukemia and diffuse large B-cell lymphoma microarray data sets. We also demonstrate that a combined use of different classification and feature selection approaches makes it possible to select relevant genes with high confidence. This is also the first paper which discusses both computational and biological evidence for the involvement of zyxin in leukaemogenesis.

Key words:
Microarray; Gene selection; Machine learning; Cancer classification; Feature Selection

* Corresponding author.Tel: (+49) 89/3187-2627; Fax: (+49) 89/3187-3585.
Email address: yu.wang@gsf.de (Yu Wang).

1 INTRODUCTION

Accurate cancer diagnosis is vital for the successful application of specific therapies. Although cancer classification has improved over the last decade, there is still a need for a fully automated and less subjective method for cancer diagnosis. Recent studies demonstrated that DNA microarrays could provide useful information for cancer classification at the gene expression level due to their ability to measure the abundance of messenger ribonucleic acid (mRNA) transcripts for thousands of genes simultaneously.

Several machine learning algorithms have already been applied to classifying tumors using microarray data. Voting Machines and Self-organising Maps (SOM) were used to analyse acute leukemia (Golub *et al.*, 1999). Support Vector Machines (SVMs) were applied to multi-class cancer diagnosis by (Ramaswamy *et al.*, 2001). Hierarchical Clustering was used to analyse colon tumor (Alon *et al.*, 1999). The best classification results are reported by Li *et al.* (2003) and Antonov *et al.* (2004). Li *et al.* employed a rule discovery method and Antonov *et al.* Maximal Margin Linear Programming (MAMA).

Given the nature of cancer microarray data, which usually consists of a few hundred samples with thousands of genes as features, the analysis has to be carried out carefully. Work in such a high dimensional space is extremely difficult if not impossible. One straightforward approach to select relevant genes is the application of standard parametric tests such as the t-test (Thomas *et al.*, 2001; Tsai *et al.*, 2003) and a non-parametric test such as the Wilcoxon score test (Thomas *et al.*, 2001; Antoniadis *et al.*, 2003). Wilks's lambda score was proposed by (Hwang *et al.*, 2002) to access the discriminatory power of individual genes. A new procedure (Antonov *et al.*, 2004) was designed to detect groups of genes that are strongly associated with a particular cancer type.

In this paper we consider two general approaches to feature subset selection, more specifically, wrapper and filter approaches, for gene selection. Wrappers and filters differ in how they evaluate feature subsets. Filter approaches remove irrelevant features according to general characteristics of the data. Wrapper approaches, by contrast, apply machine learning algorithms to feature subsets and use cross-validation to evaluate the score of feature subsets. Most methods of gene selection for microarray data analysis focus on filter approaches, although there are a few publications on applying wrapper approaches (Inza *et al.*, 2004; Xiong *et al.*, 2001; Xing *et al.*, 2001). Nevertheless, in theory, wrappers should provide more accurate classification results than filters (Langley, 1994). Wrappers use classifiers to estimate the usefulness of feature subsets. The use of "tailor-made" feature subsets should provide a better classification accuracy for the corresponding classifiers, since the features are selected according to their contribution to the classification accuracy of the classifiers. The disadvantage of the wrapper approach is its computational requirement when combined with sophisticated algorithms such as support vector machines.

As a filter approach, correlation-based feature selection (CFS) was proposed by Hall (1999). The rationale behind this algorithm is "a good feature subset is one that contains features highly correlated with the class, yet uncorrelated with each other." It has been shown in Hall (1999) that CFS gave comparable results to the wrapper and executes many times faster. It will be shown later in this paper that combining CFS with decision trees, the naive

Bayes algorithm and SVM, provides classification accuracy on cancer microarray data that is similar or better than published results.

The rest of this paper is organised as follows. We begin with a brief introduction to feature subset selection, followed by a description of feature wrappers, filters and CFS, which is essentially a filter algorithm. We discuss the advantages and disadvantages of using wrappers and filters to select feature subsets. Thereafter, we present the experimental results on acute leukemia and lymphoma microarray data. The last section discusses the results and concludes this paper.

2 METHODS

2.1 Feature Subset Selection

We now define the basic notions used in the paper. Given a microarray cancer data set $\mathcal{D}$, which contains n samples from different cancer types or subtypes, we have to build a mathematical model which can map the samples to their classes. Each sample has m genes as its features. The assumption here is that not all genes measured by a microarray are related to cancer classification. Some genes are irrelevant and some are redundant from the machine learning point of view. It is well-known that the inclusion of irrelevant and redundant information may harm performance of some machine learning algorithms.

Feature subset selection can be seen as a search through the space of feature subsets. Four questions have to be answered in terms of the search process (Langley, 1994):

(1) *Where to start the search in the feature space?* The starting point will decide the direction of the search. The search can start with an empty set and successively add useful features to this set. This is called *forward selection*. An alternative would be starting with a full set and successively removing useless features. This is called *backward elimination*. Starting the search from somewhere in the middle of the feature set is also possible. The search could be performed by either adding useful or removing useless features.

(2) *How to evaluate subsets or features?* There exist two general strategies, namely *filters* and *wrappers*. Most *filter* approaches evaluate features by giving them a score according to general characteristics of the training set. By setting a threshold, they then remove irrelevant features. If the score of a gene is above the threshold, the gene will be selected. There are also some filter approaches, such as CFS, that assign a score to subsets of features. *Wrapper* approaches, by contrast, take biases of machine learning algorithms into account when selecting features. They apply a machine learning algorithm to feature subsets and use cross-validation to compute a score for them.

(3) *How to search?* An exhaustive search of the entire feature subspace is impractical even with the current standard of computational power. A typical microarray cancer data set contains a few thousands genes as features. With m genes there exist 2^m possible feature subsets. Heuristic search strategies such as greedy hill climbing and best first are usually applied. Greedy hill climbing search considers only local changes to a feature subset. It evaluates all the possible local changes to the current feature set, such as adding one feature to the set or removing one.

It chooses the best or simply the first change that improves the score of the feature subset. Once a change is made for a feature subset, it is never reconsidered. Best first search is similar to greedy hill climbing but with the difference that it can backtrack to a more promising previous subset if it finds the current subset is not worthy to be explored.

(4) *When to stop searching?* The addition or removal of features should be stopped when none of the alternatives improves the score of a current feature subset. Another criterion would be to revise the feature subset continuously as long as the score does not degrade or to continue generating feature subsets until reaching the other end of the feature space and then select the best.

One major problem of *filters* that score individual features is the selection of a threshold by which to discard features. Although all the features will be given a score by the filter algorithm, it is not clear how to determine the optimal threshold for the data. One heuristic approach (the so called *n-1* rule) in microarray cancer analysis chooses the top *n-1* genes to start the analysis (Li and Yang, 2002). Golub *et al.* (1999) chose 50 genes most closely correlated with leukemia subtypes. Nevertheless, ranking genes by filters does present an overall picture of the microarray data. It is therefore a nice starting point for the data analysis.

In general, *filters* are much faster than *wrappers*. However, as far as the final classification accuracy is concerned, *wrappers* normally provide better results. The general argument is that the classifier that will be built from the feature subset should provide a better estimate of accuracy than a separate measure that may have an entirely different classification bias. The main disadvantage of *wrapper* approaches is that during the feature selection process, the classifier must be repeatedly called to evaluate a subset. For some computationally expensive algorithms such as SVMs or artificial neural networks, wrappers can be impractical. This will be demonstrated in our experiments.

2.2 The Choice of Feature Filter Algorithms and Classifiers

2.2.1 Feature Filter Algorithms

Apart from CFS, we consider four other filter methods in this paper. They are described as follows:

(1) χ^2 **Statistic** This criterion measures the worth of a feature by computing the value of the χ^2 statistic with respect to the class.

(2) **Information Gain** This criterion measures the worth of a feature by measuring the information gain with respect to the class. Information gain is given by

$$InfoGain = H(Y) - H(Y|X),$$

where X and Y are features, and

$$H(Y) = -\sum_{y \in Y} p(y) log_2(p(y)).$$
$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) log_2(p(y|x)).$$

Both, the information gain and the χ^2 statistic, are biased in favour of features with higher dispersion.

(3) **Symmetrical Uncertainty** This criterion measures the worth of a feature by measuring the symmetrical uncertainty with respect to the class, and compensates for information gain's

bias (Press *et al.*, 1988).

$$SU = 2.0 \times \frac{InfoGain}{H(Y) + H(X)}.$$

(4) **ReliefF** This is a feature weighting algorithm that is sensitive to feature interactions. The key idea of ReliefF is to rate features according to how well their values distinguish among instances of different classes and according to how well they cluster instances of the same class (Kira and L.A.Rendell, 1992; Kononenko, 1994). To this end, ReliefF repeatedly chooses a single instance at random from the data, and then locates the nearest instances of the same class and the nearests instances pertaining to different classes. The feature values of these instances are used to update the scores for each feature.

2.2.2 Correlation-based Feature Selection

CFS evaluates a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them (Hall, 1999).

$$CFS_S = \frac{k\bar{r_{cf}}}{\sqrt{k + k(k-1)\bar{r_{ff}}}}$$

where CFS_S is the score of a feature subset S containing k features, $\bar{r_{cf}}$ is the average feature to class correlation ($f \in S$), and $\bar{r_{ff}}$ is the average feature to feature correlation. The distinction between normal filter algorithms and CFS is that while normal filters provide scores for each feature independently, CFS presents a heuristic "merit" of a feature subset and reports the best subset it finds.

2.2.3 Classification algorithms

In this study we use three well-known classifiers, namely the decision tree learner C4.5, the simple Bayesian classifier naive Bayes, and a Support Vector Machine (SVM) (Vapnik, 1998) to demonstrate the advantages and disadvantages of feature selection algorithms. For a more thorough discussion of the first two algorithms and the corresponding feature selection methods, we refer to (Witten and Frank, 1999; Hall, 1999).

Decision trees have been popular in practice due to their simplicity, fast evaluation speed, and interpretability. The training of decision trees directly on high dimensional microarray cancer data can sometimes overfit the data, generating an overly large tree. Removing irrelevant and redundant information results in smaller, more predictive trees.

Naive Bayes assumes that features are independent given the class. Its performance on data sets with redundant features can be improved by removing such features. A forward search strategy is normally used with naive Bayes as it should immediately detect dependencies when harmful redundant features are added.

SVMs use a kernel function to implicitly map data to a high dimensional space. Then, they construct the maximum-margin hyperplane by solving an optimization problem on the training data. Sequential Minimal Optimization (SMO) (Platt, 1998) is used in this paper to train an SVM. SVMs have been shown to work well for high dimensional microarray data sets (Furey *et al.*, 2000). Howerver, due to the high computational cost it is not very practical to use the wrapper method to select genes for SVMs, as will be shown in our experimental results section.

2.3 Experimental Procedure

The experiments were performed with the Weka machine learning package (Witten and Frank, 1999). We used the following three general strategies to identify predictive features.

2.3.1 Selecting genes using feature-ranking filters

(1) Use a filter to rank all the genes in the data.
(2) Choose the first *n-1* genes as the best feature subset.

Note that the data has to be discretized before χ^2, Information Gain and Symmetrical Uncertainty filters can be applied. Weka's implementation uses an MDL-based discretization method for this purpose (Fayyad and Irani, 1993).

2.3.2 Selecting genes using CFS

(1) Choose a search algorithm.
(2) Perform the search, keeping track of the best subset encountered according to CFS_S.
(3) Output the best subset encountered.

2.3.3 Selecting genes using a wrapper method

(1) Choose a machine learning algorithm to evaluate the score of a feature subset.
(2) Choose a search algorithm.
(3) Perform the search, keeping track of the best subset encountered.
(4) Ouput the best subset encountered.

The search algorithm we used was best-first with forward selection, which starts with the empty set of genes. In this paper we report accuracy estimates for classifiers built from the best subset found during the search. The search for the best subset is based on the training data only. Once the best subset has been determined, and a classifier has been built from the training data (reduced to the best features found), the performance of that classifier is evaluated on the test data.

3 RESULTS

3.1 Analysis of Acute Leukemia Data

The acute leukemia data of Golub *et al.* (1999) consists of samples from two different types of acute leukemia, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The training data set has 38 bone marrow samples (27 ALL and 11 AML). Each sample has expression patterns of 7129 genes measured by the Affymetrix oligonucleotide microarray. The test data set consists of 24 bone marrow and 10 peripheral blood samples (20 ALL and 14 AML).

Feature-ranking filters provide a natural way to rank genes according to their ability to distinguish AML and ALL according to different criteria. The first 10 genes selected by χ^2, InfoGain, ReliefF and Symmetrical Uncertainty are listed in Table 1.

The genes in Table 1 are listed in the order according to their χ^2 score. Nevertheless, we find that the order of genes according to information gain, symmetrical uncertainty, and χ^2 does not differ much, while the ReliefF measure produces a substantially different ranking. This is due to the fact that ReliefF takes gene interactions into account while the other

three measures do not. However, we notice that the score of zyxin is high in each case. It is ranked first by χ^2, InfoGain and symmetrical uncertainty. The ReliefF filter ranks zyxin ninth.

We used the wrapper method and CFS in conjunction with a best-first search to select genes from the training set. With two classifiers, the decision tree learner J48 (Weka's implementation of C4.5) and naive Bayes, and the wrapper, only one gene is selected. This gene is zyxin, which is also the only gene selected by CFS. The SMO wrapper selected two genes, zyxin and hum_alu_at. A leave-one-out cross validation procedure was performed to investigate the robustness of the feature selection procedures. In 38 runs, zyxin was selected 34 times (92%) by CFS, 34 times (92%) by the J48 wrapper and 28 times (74%) by the naive Bayes wrapper.

It is interesting to note that zyxin is repeatedly selected by CFS, and different wrapper algorithms. Moreover, it is scored highly by the filter algorithms. This is the same gene identified by the Emerging Patterns algorithm (Li and Wong, 2002). A box plot of zyxin expression levels in the training set is presented in Figure 1. This figure clearly indicates that the expression levels of zyxin can be used to distinguish ALL from AML in the training set. The median and mean of ALL are 360.0 and 349.9, respectively. For AML, the median is 2947 and the mean is 3064.

The training result for J48 is shown in Table 2. The following rule can be created from the decision tree: if the expression level of zyxin of the sample is less than or equal to 938, it is classified as ALL. If the expression level of zyxin is larger than 938, it is classified as AML. Thirty one test samples are correctly classified by this simple rule . There are only three mistakes, one for AML, two for ALL. In Figure 2, the expression levels of zyxin from the test set are plotted individually for each sample in the test set.

Figure 2 shows three errors in the test set, two for ALL, one for AML. The x axis represents the samples and the y axis represents the expression levels of zyxin. The black line across the lower part of the figure is the threshold line y=938. The three misclassified samples have expression levels of zyxin which are far from the threshold. The median and mean of ALL in the test set are 215.00 and 416.30, respectively. Those of AML are 3029 and 3492.

The previous result reported by Golub *et al.* (1999), using a voting machine with 50 genes, can correctly predict 29 samples on the test set. In Table 3, our results are shown along with some previously published results, obtained by the Emerging Pattern algorithm (Li and Wong, 2002), the Voting Machine method (Golub *et al.*, 1999), SVMs (Furey *et al.*, 2000), and MAMA (Antonov *et al.*, 2004), on the same test set.

The results obtained by us and others suggest that the expression level of zyxin plays

Fig. 1. The expression levels of zyxin in the training set

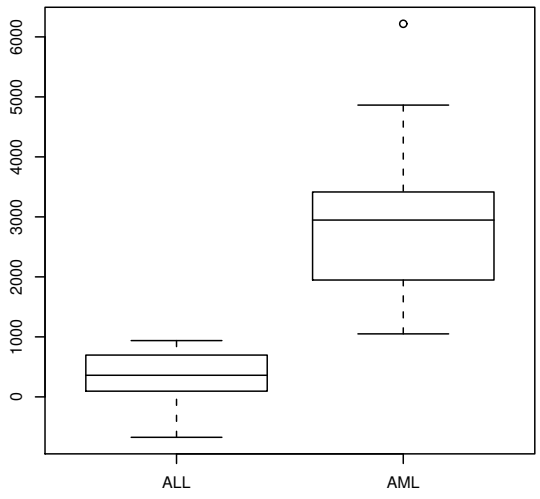
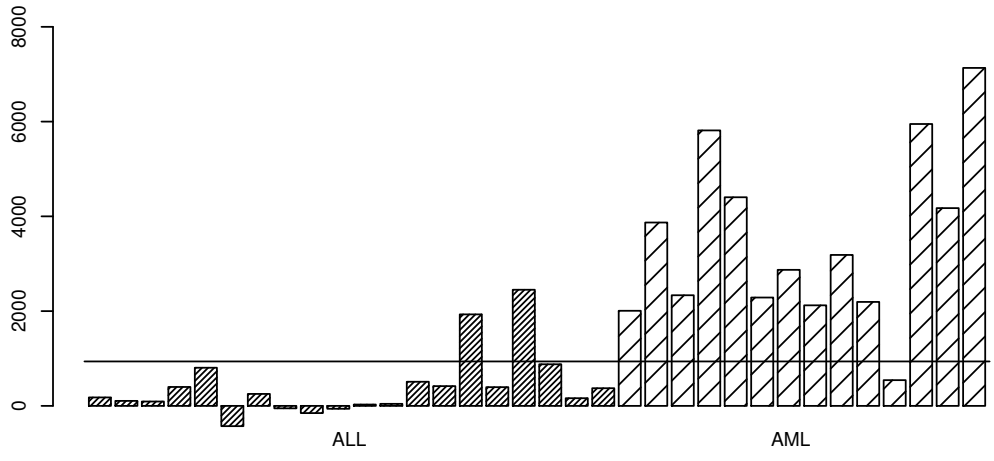


Table 1
Genes ranked by feature filters to classify subtypes of acute leukemia. The gene selected by the wrappers is marked with *.

Probe ID	Gene Annotation	χ^2		InfoGain		ReliefF		Symmetrical Uncertainty	
		score	rank	score	rank	score	rank	score	rank
X95735*	Zyxin	38.00	1	0.87	1	0.27	9	1.00	1
M55150	FAH Fumarylacetoacetate	33.54	2	0.74	2	0.26	14	0.83	2
M27891	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	33.31	3	0.70	3	0.28	7	0.83	3
M31166	PTX3 Pentaxin-related gene, rapidly induced by IL-1 beta	33.31	3	0.70	3	0.12	151	0.83	3
X70297	CHRNA7 Cholinergic receptor, nicotinic, alpha polypeptide 7	29.77	5	0.66	5	0.12	148	0.73	5
U46499	GLUTATHIONE S-TRANSFERASE, MICROSOMAL	29.77	5	0.66	5	0.22	21	0.73	5
L09209_s	APLP2 Amyloid beta (A4) precursor-like protein 2	29.77	5	0.66	5	0.20	31	0.73	5
M77142	NUCLEOLYSIN TIA-1	29.77	5	0.66	5	0.06	991	0.73	5
J03930	ALKALINE PHOSPHATASE, INTESTINAL PRECURSOR	29.02	9	0.60	9	0.11	267	0.56	45
M23197	CD33 CD33 antigen (differentiation antigen)	28.95	10	0.59	10	0.30	5	0.71	9

Fig. 2. The expression levels of zyxin in the test set



an important role in distinguishing acute lymphoblastic leukemia and acute myeloid leukemia. However, no one has yet reported direct involvement of zyxin in hematopoiesis. Zyxin has been shown to encode a LIM domain protein important in cell adhesion in

Table 2
J48 classifier for Leukemia data set

J48 pruned tree

The expression level of zyxin $\leq$ 938: ALL (27.0)

The expression level of zyxin $>$ 938: AML (11.0)

Number of Leaves : 2

Size of the tree : 3

Table 3
The comparison of classification results for AML/ALL classification. The result column shows the number of correctly classified samples in the test set (total 34).

Method	Number of features	Result
J48	1	31
Naive Bayes	1	31
SMO-CFS	1	31
SMO-Wrapper	2	30
Emerging Patterns [a]	1	31
SVM [b]	25-1000	30-32
Voting Machine [c]	50	29
MAMA [d]	132-549	**34**

[a] (Li and Wong, 2002)
[b] (Furey *et al.*, 2000)
[c] (Golub *et al.*, 1999)
[d] (Antonov *et al.*, 2004)

fibroblast (Crawford and Beckerle, 1991). Recent research has demonstrated that zyxin may enter the nucleus by association with other proteins, but is exported from the nucleus by means of intrinsic leucine-rich nuclear export sequences. Zyxin proteins may regulate gene transcription by interaction with transcription factors. In some cases, misregulation of nuclear functions of zyxin proteins appear to be associated with pathogenic effects (Wang and Gilmore, 2003).

Among the proteins which are interaction partners of zyxin (Wang and Gilmore, 2003), H-warts/LATS1, p130^{CAS} and CasL are of interest since we are looking for involvement of zyxin in acute leukemia. Zyxin is phosphorylated specifically during mitosis (Hirota *et al.*, 2000), most likely by Cdc2 kinase, and the phosphorylation regulates association with h-warts/LATS1. These findings suggest that h-warts/LATS1 and zyxin play a crucial role in controlling mitosis progression by forming a regulatory complex on the mitotic apparatus.

It was reported that zyxin LIM(1-2) are necessary and sufficient for CasL/HEF1 interaction (Yi *et al.*, 2002). CasL/HEF1 interacts with a Crk family adaptor protein called Crkl. p130^{CAS} is found to be trosine phosphorylated and associated with Crkl in BCR/ABL expressing cell lines and in samples obtained from chronic myeloid leukemia (CML) and a type of ALL(Ph$^+$ ALL) (Salgia *et al.*, 1996). BCR/ABL is an oncogene which is sufficient to produce CML. A study by (Yagi *et al.*, 2003) identified zyxin as one of 35 genes which were associated with the prognosis of pediatric AML. Given all these facts, it is tempting to speculate that zyxin plays a role in leukemogenesis.

Recently zyxin has been shown to be upregulated by RASSF1A in non-small cell lung cancer and neuroblastoma. RASSF1A is a 3p21.3 tumor suppressor gene (Agathanggelou *et al.*, 2003). Harada *et al.* (2002) investigated aberrant promoter methylation and silencing of the RASSF1A gene in pediatric tumours and cell lines. They found that 17% of ALL are methylated, but methylation is absent in AML. This might be one of the reasons why the expression levels of zyxin are

high in AML samples and low in ALL samples. This hypothesis needs to be confirmed by experiments.

Could zyxin be one of the molecular targets in acute leukemia? Research (van der Gaag *et al.*, 2002) on the role of zyxin in differential cell spreading and proliferation of melanoma cells and melanocytes showed that zyxin is significantly up-regulated in melanoma cells compared to melanocytes. Treatment of melanoma cells with 12-O-tetradecanoylphorbol-13-acetate down-regulates zyxin expression, inhibits cell spreading and proliferation, and promotes differentiation. We believe more experiments are needed to verify zyxin's role in leukemia.

Most of the other genes at the top of the list in Table 1 are also selected and discussed by Golub *et al.* (1999), except for one. It is called PTX3, which is a Pentaxin-related gene. This gene has been shown to be up-regulated by C/EBPα in BCR/ABL cell lines (Tavor *et al.*, 2003). Recently, mutations that abrogated transcriptional activation of C/EBPα have been detected in AML patient samples. Moreover, the progression of CML to blast crisis in patients is correlated with down-modulation of C/EBPα. This evidence suggests that the expression level of PTX3 should not be neglected when expression data of acute leukemia is analyzed.

3.2 Analysis of Diffuse Large B-cell Lymphoma Data

Diffuse large B-cell lymphoma (DLBCL) is the most common subtype of non-Hodgkin's lymphoma. There are no reliable morphological or immunohistochemical indicators that can be used to recognise subtypes of DLBCL. Alizadeh *et al.* (2000) identified two molecularly distinct forms of DLBCL which had gene expression patterns indicative of different stages of B-cell differentiation. One type expressed genes characteristic of germinal centre B cells—germinal centre B-like DLBCL (GC-DLBCL); the second type expressed genes normally induced during *in vitro* activation of peripheral blood B cells—activated B-like DLBCL (ABC-DLBCL). Patients with GC-DLBCL had a significantly better survival rate than those with ABC-DLBCL. Alizadeh *et al.* (2000) designed a specialised cDNA microarray, the 'Lymphochip' to analyse 45 samples. We have divided this data into a training set of 36 samples and a test set of 9 samples. Each sample has expression values of 4026 genes.

The first 10 genes selected by the filtering algorithms from the training set are listed in Table 4. At the end of this table additional genes selected by the naive Bayes and SMO wrappers are also shown (see Table 5).

We can draw the same conclusion from the results of the filters in Table 4 as in the acute leukemia case. χ^2, InfoGain, and Symmetrical Uncertainty filters give more or less the same ranking for genes while the ReliefF filter ranks genes quite differently. From the biological application point of view it is not clear which filter to choose. The first 25 genes selected by χ^2, InfoGain, and Symmetrical Uncertainty filters are different from Alizadeh *et al.* (2000) except for JAW1 and FMR2. There are several reasons:

(1) We have divided the data set into a training set and a test set. Our feature selection is performed on the training set. Alizadeh *et al.* (2000) used the whole data set to select the most informative genes.
(2) We have used different selection criteria.

The wrappers chose the genes shown in Ta-

Table 4
Genes ranked by feature filters to distinguish subtypes of DLBCL. The genes selected by the wrappers are marked with *.

Gene ID	Gene Annotation	χ^2		InfoGain		ReliefF		Symmetrical Uncertainty	
		score	rank	score	rank	score	rank	score	rank
GENE3330X*	Unknown	25.08	1	0.59	2	0.15	6	0.59	2
GENE3328X*	Unknown UG Hs.136345 ESTs	25.08	1	0.59	2	0.16	4	0.59	2
GENE3967X	Deoxycytidylate deaminase	25.02	3	0.59	4	0.09	43	0.59	4
GENE3261X	Unknown	24.91	4	0.63	1	0.17	3	0.64	1
GENE3259X	Unknown UG Hs.124922 ESTs	23.22	5	0.54	8	0.13	13	0.54	8
GENE3258X	JAW1, lymphoid-restricted membrane protein	22.74	6	0.57	5	0.14	10	0.59	5
GENE3256X	JAW1, lymphoid-restricted membrane protein	22.74	6	0.57	5	0.11	19	0.59	5
GENE3939X	Unknown UG Hs.169081 ets variant gene 6 (TEL oncogene)	22.53	8	0.54	7	0.16	5	0.54	7
GENE3512X	zinc finger protein 42 MZF-1	22.09	9	0.51	9	0.13	12	0.51	10
GENE3966X	Deoxycytidylate deaminase	21.94	10	0.51	11	0.07	105	0.51	11
GENE3165X*	Unknown	0.00	331	0.00	331	-0.00	2367	0.00	331
GENE1063X*	PMS6, DNA mismatch repair protein	11.42	187	0.24	248	0.06	144	0.25	262

ble 5 to build the classifiers. Table 5 also lists the numbers of samples that are correctly classified by the classifiers both on the training set and the test set.

The J48 wrapper chose only GENE3328X to build the decision tree. GENE3328X also scored quite high with each of the filters. It is ranked first by χ^2, second by InfoGain and Symmetrical Uncertainty, and fourth by ReliefF. The decision tree built on GENE3328X achieves 89% accuracy in a leave-one-out cross-validation on the training set and 89% on the test set. GENE3328X is a cDNA clone from germinal centre B cells.

The naive Bayes wrapper chose GENE3165X, GENE3330X, and GENE1063X. The combination of these three genes gives a good performance for naive Bayes. On the training set it gets 97% accuracy. On the test set, it is 100%. Among the three, only GENE3330X is ranked high by the filters. It is ranked first by χ^2, second by InfoGain and Symmetrical

Table 5
Experimental results of the wrappers for classification of DLBCL. The numbers count correctly classified samples in each data set. The numbers inside parentheses are the total numbers of samples in each data set.

Method	Genes selected	Training	Test
J48	GENE3328X	33(36)	8(9)
Naive Bayes	GENE3165X GENE3330X GENE1063X	35(36)	9(9)
SMO	GENE3330X GENE1063X	36(36)	9(9)

Uncertainty, and sixth by ReliefF. Unfortunately, the function of GENE3330X is not known, neither is its origin. GENE1063X is PMS6, also called PMS2L4. It encodes a DNA mismatch repair protein. At the RNA level, it is found at spleen, prostate and lymphoid. The filter algorithms do not rank GENE1063X highly (see Table 4). The scores of GENE3165X from the filters are all zero. This gene would be ignored if we only relied on the filter results. Without this gene, naive Bayes only achieves 94% accuracy on the training set.

The SMO wrapper chose GENE3330X and GENE1063X. On both the training set and the test set, the SVM gets the best classification result among the three machine learning algorithms. All the samples are classified correctly.

CFS chose the genes shown in Table 6 as the best subset. About half of the genes have low ranks according to the other filters. These genes would certainly escape the notice of an investigator if a heuristic threshold like the *n-1* rule (see Section 2.1) were applied.

Table 7 shows the classification results of the three learning algorithms with genes selected by CFS. Both naive Bayes and the SVM perform the same as in the wrapper case on the test data, only J48 is slightly worse.

4 DISCUSSION

We have shown in this paper that feature subset selection algorithms, namely wrappers, filters and CFS, can be very useful in extracting relevant information in microarray data analysis. Wrapper approaches can choose the best genes for building classifiers while filters can provide a nice overview by ranking the genes for the particular problem at the hand. CFS can choose genes which are highly correlated to cancers yet uncorrelated to each other.

When the methods agree and select the same genes, we can have more confidence in the result. In our study we demonstrated that several different methods used in the wrapper approach as well as several different filters indicated an involvement of zyxin in distinguishing AML and ALL. This result is in agreement with previous work (Li and Wong, 2002). However, contrary to previous studies, we collected in this study important biological evidence that suggests at least indirect involvement of zyxin in acute leukemia. To our knowledge, this is the first study that combines both computational and biological evidence and generates a clear hypothesis about zyxin that can be tested experimentally.

We have applied wrappers, filters and CFS to acute leukemia data and diffuse large B-cell lymphoma microarray data. Although CFS and wrappers based on decision trees, naive Bayes, and SVMs, do no select as many genes as previous research suggests (Golub *et al.*, 1999; Alizadeh *et al.*, 2000), the final classifiers built with these few genes yield surprisingly good performance. However, given the nature of microarray cancer data, which on the one hand has low signal to noise ratio, and on the other hand has a limited number of samples, we are very cautious to suggest that these genes are sufficient to build good classifiers for the diagnosis of the analysed cancers.

Filter algorithms provide a natural way to present an overview of microarray cancer data. Four feature-ranking filters, namely χ^2, Information Gain, Symmetrical Uncertainty and Relief, have been investigated in this paper, each of which has been quite popular in the machine learning community. The first three filters give more or less the same

Table 6
Genes selected by CFS with their corresponding filter scores

Gene ID	Gene Annotation	χ^2		InfoGain		ReliefF		Symmetrical Uncertainty	
		score	rank	score	rank	score	rank	score	rank
GENE3941X	Unknown UG Hs.143722 ESTs	17.70	36	0.45	26	0.07	106	0.47	20
GENE3499X	Unknown UG Hs.123387 ESTs	9.71	255	0.25	208	0.01	1020	0.29	154
GENE3718X	47-kD autosomal chronic granulomatous disease protein	8.32	298	0.21	287	0.02	934	0.26	242
GENE2322X	Unknown UG Hs.140489 ESTs	10.43	233	0.27	178	0.03	577	0.31	122
GENE3132X	NERF, ets family transcription factor	10.06	242	0.26	188	0.04	271	0.30	137
GENE3325X	Unknown UG Hs.120245	16.60	44	0.37	51	0.12	16	0.39	56
GENE3258X	JAW1,lymphoid-restricted membrane protein	22.74	6	0.57	5	0.14	10	0.59	5
GENE3259X	Unknown UG Hs.124922 ESTs	23.22	5	0.54	8	0.13	13	0.54	8
GENE3256X	JAW1,lymphoid-restricted membrane protein	22.74	6	0.57	5	0.11	19	0.59	5
GENE3261X	Unknown	24.91	4	0.63	1	0.17	3	0.64	1
GENE2739X	Unknown UG Hs.136952 ESTs	9.71	255	0.25	208	0.07	100	0.29	154
GENE1940X	Low-affinity IgG Fc receptor II-B and C isoforms	18.84	23	0.45	26	0.07	109	0.46	21
GENE1354X	Casein kinase I delta	9.71	255	0.25	208	0.05	212	0.29	154
GENE3967X	Deoxycytidylate deaminase	25.02	3	0.59	4	0.09	43	0.59	4
GENE3932X	core binding factor alpha1b subunit	15.48	63	0.39	41	0.15	8	0.42	32
GENE236X	metallothionein-II	9.71	255	0.25	208	0.05	253	0.29	154
GENE547X	GCF-2,GC-rich sequence DNA binding factor	9.71	255	0.25	208	0.04	373	0.29	154
GENE763X	Eukaryotic translation initiation factor 4E	11.50	180	0.29	117	0.03	529	0.33	95
GENE427X	p18-INK6, Cyclin-dependent kinase 6 inhibitor	12.88	123	0.33	86	0.03	410	0.36	66
GENE404X	Unknown UG Hs.140559 EST	18.44	30	0.46	20	0.10	32	0.49	15
GENE958X	DNA alkylation repair protein	13.93	94	0.35	65	0.03	589	0.39	54
GENE1798X	Unknown	9.71	255	0.25	208	0.04	344	0.29	154
GENE3821X	Unknown	13.41	112	0.34	76	0.07	110	0.38	60
GENE1720X	cysteine rich protein with LIM motif	11.50	180	0.29	117	0.08	77	0.33	95
GENE1567X	CXC chemokine	11.50	180	0.29	117	0.05	211	0.33	95

ranking for the genes, but the ranking obtained from ReliefF is quite different, since ReliefF is sensitive to feature interactions. It is up to the practitioner to decide which filter to use. Since there are many filter algorithms available—see, for example, Hero (2003) and Su *et al.* (2003)—one idea is to combine scores from different filters to produce an overall score. Further research needs to clarify exactly how to achieve this.

Another important decision for filter algorithms a practitioner might face is the number of genes to be selected. In other words, a practitioner must choose a threshold for the filters. It is not clear how to determine

an optimal value for the threshold. CFS and the wrapper do not have this problem. By testing combinations of genes from the data, they will automatically select an appropriate subset of genes. In our lymphoma example, we show that Gene3165X selected by the wrapper with naive Bayes and SMO, scores around zero for all four filters we have used. Considering only the filters, this gene would certainly have been ignored. A major drawback of some feature-ranking filter algorithms is that they evaluate each gene individually, but in reality the combination of expression levels of several genes might be responsible for cancer. Filters might miss these genes if their individual expression levels are not informative enough for the cancer classification.

Due to their high computational costs, it is not easy to combine wrappers with some machine learning algorithms such as SMO. Table 8 shows that the SMO wrapper needs more time to run, on average a magnitude more than the time required by CFS and the other wrappers.

Table 7
Experimental results of CFS for classification of DLBCL. The numbers count correctly classified samples in each data set. The numbers inside parentheses are the total numbers of samples in each data set.

Method	Training set	test set
J48	36(36)	7(9)
Naive Bayes	36(36)	9(9)
SMO	36(36)	9(9)

Table 8
CPU time (in seconds) spent on the data sets by CFS and the wrapper.

Data set	CFS	Wrapper Methods	
Leukemia	671.74	J48	3838.74
		Naive Bayes	4866.97
		SMO	60228.97
Lymphoma	2246.42	J48	2354.25
		Naive Bayes	4001.75
		SMO	49101.68

Our experimental results show that the classifiers built from the genes selected by filters, CFS, and wrappers, demonstrated a similar performance on the analyzed datasets. Thus, the filters and CFS could be recommended for fast analysis of data. However, in order to better validate the results and to select a few genes that could be further investigated for cancer treatment, the wrapper approaches can be recommended. Indeed, wrappers selected in general just a few genes in our experiments. A consensus analysis of results given by filter and wrapper approaches will provide selection of relevant genes with high confidence.

We believe that our results will motivate more microarray practitioners to use wrappers, filters and CFS as their analysis tools. These machine learning algorithms are implemented in WEKA, a publicly available open-source software package. This software can be used both by experienced and novice users. WEKA has been already applied in a number of bioinformatics studies as reviewed elsewhere (Frank *et al.*, 2004).

5 ACKNOWLEDGEMENT

We would like to thank Dr. Marco Zaffalon for proofreading the manuscript and validating our results with his algorithm, Dr. Franceso Bertoni for his advice on lymphoma data analysis, and Annina Neumann for proofreading the manuscript. We are also grateful for the comments given by reviewers, which have significantly improved this paper.

References

Agathanggelou, A., Bieche, I., Ahmed-Choudhury, J., Nicke, B., Dammann, R., Baksh, S., Gao, B., Minna, J., Downward, J., Maher, E. and Latif, F. (2003) Identification of novel gene expression targets for the ras association domain family 1 (rassf1a) tumor suppressor gene in non-small cell lung cancer and neuroblastoma. *Cancer Research*, **63(17)**, 5344–5351.

Alizadeh, A. A., M.B. Eisen, R. D., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Powell, J., Yang, L., Marti, G., Moore, T., Hudson, J. J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, J., Warnke, R., Levy, R., Wilson, W., Grever, M., Byrd, J., Botstein, D., Brown, P. and Staudt, L. M. (2000) Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, **403(6769)**, 503–11.

Alon, U., Barkai, N., Notterman, D., 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. *Proc. Natl. Acad. Sci*, **96(12)**, 6745–6750.

Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003) Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, **19**, 563–570.

Antonov, A. V., Tetko, I. V., Mader, M. T., Budczies, J. and Mewes, H. W. (2004) Optimization models for cancer classification: extracting gene interaction information from microarray expression data. *Bioinformatics*, **20**, 644–652.

Crawford, A. and Beckerle, M. (1991) Purification and characterization of zyxin, an 82,000-dalton component of adherens junctions. *The Journal of Biological Chemistry*, **266(9)**, 5847–53.

Fayyad, U. and Irani, K. (1993) Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of IJCAI-93, 13th International Joint Conference on Artificial Intelligence*, pp. 1022–1027.

Frank, E., Hall, M., Trigg, L., Holmes, G. and Witten, I. H. (2004) Data mining in bioinformatics using Weka. *Bioinformatics*, **20**, 2479 – 2481.

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

Hall, M. A. (1999) Correlation-based feature selection for machine learning. *PhD thesis, Department of Computer Science, University of Waikato.*

Harada, K., Toyooka, S., Maitra, A., Maruyama, R., Toyooka, K., Timmons, C., Tomlinson, G., Mastrangelo, D., Hay, R., Minna, J. and Gazdar, A. (2002) Aberrant promoter methylation and silencing of the rassf1a gene in pediatric tumors and cell lines. *Oncogene*, **21(27)**, 4345–4349.

Hero, A. (2003) Gene selection and ranking with microarray data. *Proc. of Intl Conf on Signal Processing and Applications, Paris.*

Hirota, T., Morisaki, T., Nishiyama, Y., Marumoto, T., Tada, K., Hara, T., Masuko, N., Inagaki, M., Hatakeyama, K. and Saya, H. (2000) Zyxin, a regulator of actin filament assembly, targets the mitotic apparatus by interacting with h-warts/lats1 tumor suppressor. *The Journal of Cell Biology,*,

149, 1073–1086.

Hwang, D., Schmitt, W. A., Stephanopoulos, G. and Stephanopoulos, G. (2002) Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics*, **18**, 1184–1193.

Inza, I., Larranaga, P., Blanco, R. and Cerrolaza, A. (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine*, **31(2)**, 91–103.

Kira, K. and L.A.Rendell (1992) A practical approach for feature selection. *Proceedings of the Ninth International Conference on Machine Learning*, pp. 249–256.

Kononenko, I. (1994) Estimating attributes: Analysis and extensions of RELIEF. *European Conference on Machine Learning*, pp. 171–182.

Langley, P. (1994) Selection of relevant features in machine learning. *Proceedings of AAAI Fall Symposium on Relevance*, pp. 140–144.

Li, J., Liu, H., Ng, S.-K. and Wong, L. (2003) Discovery of significant rules for classifying cancer diagnosis data. *Bioinformatics*, **19**, 93ii–102ii.

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.

Li, W. and Yang, Y. (2002) How many genes are needed for a discriminant microarray data analysis. *Methods of Microarray Data Analysis, Kluwer Academic*, pp. 137–150.

Platt, J. (1998) Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods - Support Vector Learning, MIT Press.*

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1988) *Numerical Recipes in C.*

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., Poggio, T., Gerald, W., Loda, M., Lander, E. and Golub, T. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci*, **98(26)**, 15149–15154.

Salgia, R., Pisick, E., Sattler, M., Li, J., Uemura, N., Wong, W., Burky, S., Hirai, H., Chen, L. and Griffin, J. (1996) p130^{CAS} forms a signaling complex with the adapter protein crkl in hematopoietic cells transformed by the bcr/abl oncogene. *The Journal of Biological Chemistry*, **271(41)**, 25198–25203.

Su, Y., Murali, T. M., Pavlovic, V., Schaffer, M. and Kasif, S. (2003) Rankgene: identification of diagnostic genes based on expression data. *Bioinformatics*, **19**, 1578–1579.

Tavor, S., Park, D., Gery, S., Vuong, P., Gombart, A. and Koeffler, H. (2003) Restoration of c/ebpalpha expression in a bcr-abl+ cell line induces terminal granulocytic differentiation. *The Journal of Biological Chemistry*, **278(52)**, 52651–9.

Thomas, J. G., Olson, J. M., Tapscott, S. J. and Zhao, L. P. (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.*, **11**, 1227–1236.

Tsai, C.-A., Chen, Y.-J. and Chen, J. J. (2003) Testing for differentially expressed genes with microarray data. *Nucl. Acids. Res.*, **31**, e52.

van der Gaag, E., Leccia, M., Dekker, S., Jalbert, N., Amodeo, D. and Byers, H. (2002) Role of zyxin in differential cell spreading and proliferation of melanoma cells and melanocytes. *Journal of Investigative Dermatology*, **118(2)**, 246–54.

Vapnik, V. N. (1998) *Statistical Learning Theory, Wiley.*

Wang, Y. and Gilmore, T. (2003) Zyxin and paxillin proteins: focal adhesion plaque lim domain proteins go nuclear. *Biochim Biophys Acta.*, **1593(2-3)**, 115–120.

Witten, I. H. and Frank, E. (1999) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann.*

Xing, E., Jordan, M. and Karp, R. (2001) Feature selection for high-dimensional genomic microarray data. *Proceedings of the Eighteenth International Conference on Machine Learning.*

Xiong, M., Fang, X. and Zhao, J. (2001) Biomarker identification by feature wrappers. *Genome Research*, **11(11)**, 1878–1887.

Yagi, T., Morimoto, A., Eguchi, M., Hibi, S., Sako, M., Ishii, E., Mizutani, S., Imashuku, S., Ohki, M. and Ichikawa, H. (2003) Identification of a gene expression signature associated with pediatric aml prognosis. *Blood*, **102(5)**, 1849–1856.

Yi, J., Kloeker, S., Jensen, C., Bockholt, S., Honda, H., Hirai, H. and Beckerle, M. (2002) Members of the zyxin family of lim proteins interact with members of the p130cas family of signal transducers. *The Journal of Biological Chemistry*, **277(11)**, 9580–9.