A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and Microarray Data Classification

Yukyee Leung and Yeungsam Hung

Abstract—Filters and wrappers are two prevailing approaches for gene selection in microarray data analysis. Filters make use of statistical properties of each gene to represent its discriminating power between different classes. The computation is fast but the predictions are inaccurate. Wrappers make use of a chosen classifier to select genes by maximizing classification accuracy, but the computation burden is formidable. Filters and wrappers have been combined in previous studies to maximize the classification accuracy for a chosen classifier with respect to a filtered set of genes. The drawback of this single-filter-single-wrapper (SFSW) approach is that the classification accuracy is dependent on the choice of specific filter and wrapper. In this paper, a multiple-filter-multiple-wrapper (MFMW) approach is proposed that makes use of multiple filters and multiple wrappers to improve the accuracy and robustness of the classification, and to identify potential biomarker genes. Experiments based on six benchmark data sets show that the MFMW approach outperforms SFSW models (generated by all combinations of filters and wrappers used in the corresponding MFMW model) in all cases and for all six data sets. Some of MFMW-selected genes have been confirmed to be biomarkers or contribute to the development of particular cancers by other studies.

Index Terms—Filters, gene selection, hybrid classification models, microarray data classification, wrappers.

1 Introduction

The rapid advances of gene expression microarray technology enable the simultaneous monitoring of thousands of genes in a single experiment (referred to as a sample). With a certain number of samples, investigations can be made into whether there are patterns or dissimilarities across samples of different types, such as cancerous versus normal, or even within subtypes of diseases. The problem is referred to as sample classification.

Microarray sample classification has been studied extensively using classification techniques in machine learning and pattern recognition. Classification tools such as weighted voting (WV) [1], k-nearest neighbor (k-NN) [2], support vector machine (SVM) [3], and Fisher's linear discriminant analysis (LDA) [4] have been used for microarray data classification. However, these tools have not been effective for identifying biomarkers, which are substances (genes in the present context) used for detecting whether a patient has got a particular disease or not [5], [6]. In a microarray chip, the number of genes available is far greater than that of samples, a well-known problem called the curse of dimensionality [7]. However, most genes in a microarray give little benefits to the sample classification problem. Therefore, prior to sample classification, it is important to perform gene selection whereby more interpretable genes are identified as biomarkers, so that a more efficient, accurate, and reliable performance in classification can be expected. These biomarkers may also be useful for assessing disease risk [6] and understanding the basic biology of a disorder [8].

There are, in general, two approaches to gene selection, namely filters and wrappers [9], [10]. The filter approach selects genes according to their discriminative powers with regard to the class labels of samples [11]. Methods such as Signal-to-Noise Ratio (SNR) [1], t-statistics (TS) [11], threshold number of misclassifications (TNoM) score [12], and F-test [13] have been shown to be effective scores for measuring the discriminative power of genes. A comparison of these methods can be found in [14]. In all cases, genes are ranked according to their statistical scores, and a certain number of the highest ranking genes are selected for the purpose of classification [15]. Although gene selection using filters are simple and fast, the method has several shortcomings:

1. The criterion used for gene selection in filters does not necessarily associate with the classifiers to be applied [13], [16].
2. The filter approach does not take into account correlation between genes, which reduces the usefulness of the selected genes for sample classification.
3. Despite reports that classifiers with few genes (less than 15-20) are able to achieve good performances [17], [18], there is little theoretical support for determining how many genes should be chosen for classification, and the number used is somewhat arbitrary [19].

• *The authors are with the Department of Electrical and Electronic Engineering, CYC512, Chow Yei Ching Building, University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: {yyleung, yshung}@eee.hku.hk.*

Manuscript received 30 Oct. 2007; revised 20 Mar. 2008; accepted 4 May 2008; published online 7 May 2008.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2007-10-0145. Digital Object Identifier no. 10.1109/TCBB.2008.46.

1545-5963/10/$26.00 © 2010 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

The wrapper approach for gene selection goes some way to address all of the above problems of the filter approach. In the wrapper approach, genes are selected sequentially one by one so as to optimize the training accuracy of a particular classifier [10]. That is, the classifier is first trained using one single gene, and this training is performed for the entire original gene set. The gene that gives the highest training accuracy is selected. Then, a second gene is added to the selected gene and the gene that gives the highest training accuracy for the two-gene classifier is chosen. This process is continued until a sufficiently high accuracy is achieved with a certain gene subset. Contrary to the filter approach which selects genes with no consideration for the classifier, the wrapper approach selects genes that are "tailor-made" for a particular classifier. Furthermore, the process ensures that newly added genes are complementary rather than correlated with genes already selected. The method also provides a stopping criterion when a certain number of genes are sufficient for attaining the required accuracy. However, a major disadvantage of the wrapper approach is that its computation requirement is formidable [5], particularly if the original gene set is large. Because of this, wrappers are not frequently used in microarray data analysis [17], [20].

In view of the drawbacks of the filter and wrapper approaches, hybrid filter-wrapper models have been proposed [21], [22], [23] that take advantage of the simplicity of the filter approach for initial gene screening and then make use of the wrapper approach to optimize classification accuracy in final gene selection. In the hybrid model, a filter is first used to screen out a majority of (irrelevant) genes from the original set to give a filtered subset of a relatively small size (say, a few hundred from an original set of several thousands). Then, the wrapper is applied to select genes from the filtered subset to optimize the training accuracy. As the filter efficiently reduces the size of the gene set by an order of magnitude or more, the computations of the subsequent wrapper become acceptable. We will refer to this hybrid model as a single-filter-single-wrapper (SFSW) approach, as it is necessary to select a particular filter and a specific classifier in the process. The SFSW approach however has its own difficulties:

1. Different filters yield different filtered subsets that may leave out some relevant biomarkers which consequently do not have a chance to be considered in the wrapper evaluation.
2. Different wrappers will select different genes from the filtered set despite achieving the same training accuracy.
3. Some SFSW models are better than the others in terms of attaining the required training accuracy.

The last point can be addressed by trying different filter-wrapper combinations in the SFSW model to suit the data set. However, this is not a satisfactory approach, as the need to tune the choice of filter and classifier in the SFSW model suggests that the model is not robust to variations among data sets. Furthermore, the fact that different filters and wrappers would select different genes makes it doubtful whether the genes selected by one particular model are true biomarkers.

To address the above problems, we propose in this paper a multiple-filter-multiple-wrapper (MFMW) approach for the hybrid model. In the MFMW hybrid model, we consider using multiple filters to select genes and then combining them to provide a merged filtered subset of genes. The use of multiple filters with different filter metrics ensures that useful biomarkers are unlikely to be screened out in the initial filter stage. The use of multiple wrappers is intended to enhance the reliability of the classification by establishing consensus among several classifiers. As a result, there is some kind of consensus among the different classifiers in the wrapper step as to which genes should be selected. Hence, the final genes selected can be considered to be more robust with a mixture of characteristics that fit several wrappers, and are therefore better qualified as biomarkers. Furthermore, since the MFMW model already incorporates the characteristics of multiple filters and wrappers, it is no longer necessary to try different filter-wrapper combinations in order to search for a suitable combination that yields the highest classification accuracy. We will show that the proposed MFMW model provides predictive accuracies that are either comparable or better than the best existing results obtained using all available SFSW methods.

This paper is organized as follows: In Section 2, some preliminaries on filter metrics and classifiers are given. Our proposed algorithm MFMW is given in Section 3. In Section 4, the MFMW approach is evaluated by means of six benchmark data sets available in the public domain. Comparisons are made between the results obtained using the MFMW and different SFSW models. In Section 5, discussions are given on how many genes should be selected by each filters, how many filters and classifiers should be chosen in an MFMW model, and biological significance of MFMW-selected genes. Some concluding remarks are given in Section 6.

2 Preliminaries

In a microarray data set Z, each sample is an n-vector storing the gene expression values, and may be considered as a data point in an n-dimensional space, where n may take a value of several thousands. Each sample carries a specific label which is utilized for supervised learning. This means that $Z = [X|Y]$ can be partitioned into two groups X and Y which represent, for example, cancerous and normal samples, respectively. Hence, the $g\text{th}$ row $Z_g = [X_g|Y_g]$ of Z is the profile of gene g across all samples.

2.1 Filter Metrics

Feature selection is a crucial and critical step in microarray data analysis for removing irrelevant data and reducing computational complexity. One way to do so is by filtering, whereby the "goodness" of a gene is evaluated by measuring the relationship between gene expression and the class label using statistical techniques [14]. Three of the most commonly used metrics are 1) SNR, 2) TS, and 3) Pearson Correlation Coefficient (PC).

2.2 Classifiers

The aim of supervised classification is to develop a decision rule to discriminate between samples of different

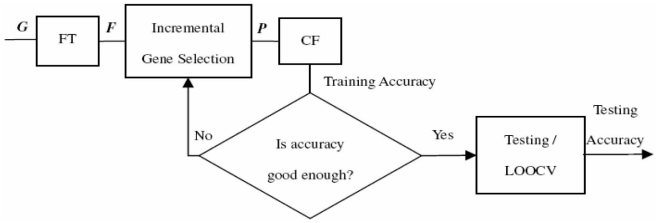


Fig. 1. SFSW model for gene selection.

classes based on the gene expression profile. In this study, three widely used classifiers are considered, namely 1) WV, 2) k-NN, and 3) SVM.

WV define the vote of gene g as

$$v_g = w_g\Big(x_g - \frac{\mu_1 + \mu_2}{2}\Big), \tag{1}$$

where the weighting factor w_g is the SNR score of gene g, x_g is the expression level of gene g in a sample S, and μ_1 and μ_2 are the mean expression levels of gene g for samples of class 1 and class 2, respectively [1]. Assuming $\mu_1 > \mu_2$, $v_g > 0$ (respectively < 0) indicates a vote for class 1 (respectively 2). Let V_{win} and V_{lose} be the sums of votes for the winning and losing class, respectively. The prediction strength (PS) provides a measure of the margin of victory:

$$PS = \frac{V_{win} - V_{lose}}{V_{win} + V_{lose}}. \tag{2}$$

If PS exceeds a certain threshold, the sample S is assigned to winning class, otherwise, sample S is unclassified.

k-NN classifies samples based on closest training examples in the gene space. All training samples are mapped into a multidimensional gene space which is partitioned into regions by class labels of the training samples. A point in the space is assigned to the class c if it is the most frequent class label among the k nearest training samples [24]. The choice of k depends upon the data—larger values of k reduce the effect of noise, but make boundaries between classes less distinct. We choose $k = 3$ in our study.

SVM—Vapnik [25] proposed the SVM for performing classification by locating the hyperplane that maximally separates two sets of points in an n-dimensional feature (e.g., genes) space R^n. The SVM classifier then assigns a class label to a new sample according as which side of the hyperplane the sample lies.

3 Multiple-Filter-Multiple-Wrapper (MFMW) Model

Since the MFMW hybrid model is a generalization of the SFSW model, we will begin by introducing the structure of the SFSW model.

In the SFSW model shown in Fig. 1, the full set of genes G is first filtered by a filter (FT) using a particular filter metric to extract a subset of genes F that is much reduced in size compared to G. The genes in F are then selected in an incremental manner using a chosen classifier (CF) with an aim to optimize the training accuracy. To start with, each gene in F is considered as a candidate for a single-gene classifier and the gene(s) giving the highest accuracy is/are identified and retained as the first gene in the set P. Then, the remaining genes in F are combined with the first gene in P to give a two-gene classifier, and the second gene that together with the first selected gene yield the highest accuracy is identified. This process of adding genes to P is repeated until P contains a sufficient number of genes to achieve the required training accuracy. The final gene set P is then evaluated using either an independent data set or by performing Leave-One-Out Cross-Validation (LOOCV). In LOOCV, one sample from the data set is excluded, and the rest of the samples are used to train the classifier. The trained classifier is then used to predict the class label of the left-out sample, and this is performed for each sample in the data set. The LOOCV estimate of classification accuracy is defined as the overall number of correct classifications divided by the total number of samples in the data set. If independent testing and training data sets are available, the classifier can first be trained using the training data set and then applied to the testing data set. In this case, the testing accuracy represents the unbiased predictive power of the gene subset P. Note that the above LOOCV strategy means that only the classification (as opposed to the entire process of gene filtering) is cross validated [26], [27].

In order to overcome the weaknesses of using SFSW as discussed in Section 1, we now propose the MFMW model shown in Fig. 2.

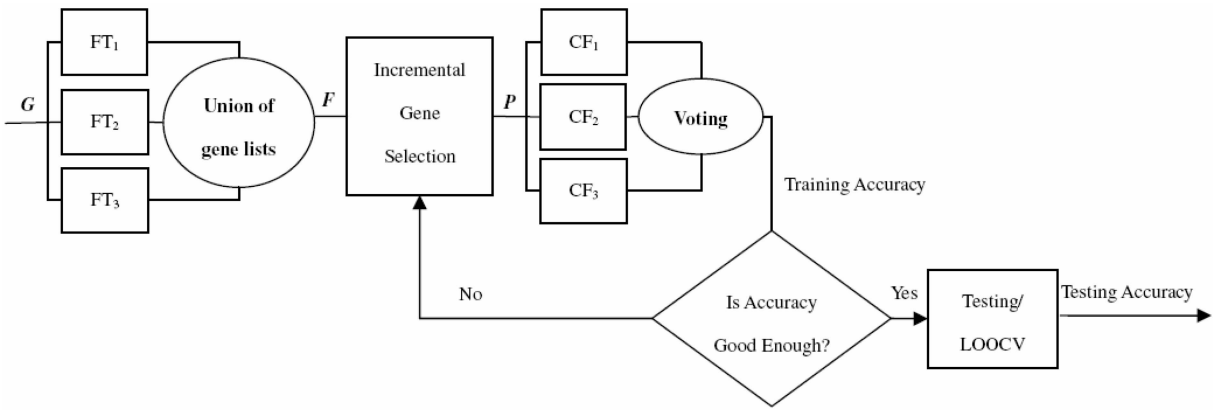


Fig. 2. MFMW model for gene selection.

TABLE 1
All Possible Combinations of Outputs of Three Classifiers in a Two-Class Classification Problem

	C1	C2	C3	C4	C5	C6	C7	C8
True Class Label	A	A	A	A	A	A	A	A
CF 1	A	A	A	A	B	B	B	B
CF 2	A	A	B	B	A	A	B	B
CF 3	A	B	A	B	A	B	A	B
Classification Output	A	X	X	X	X	X	X	B
Prediction Status	R	I	I	I	I	I	I	W

By unanimous voting, only cases C1 and C8 will receive a classification output A or B, giving a prediction status of R and W, respectively. All remaining samples have indecisive classification output "X" and consequently indecisive prediction status I.

In the MFMW model, several (e.g., three in Fig. 2) filters $FT_i(i = 1, 2, 3)$ are employed, each for selecting a predefined number of (e.g., a few hundred) genes. Different filters have their own characteristics and there is typically a fair proportion of both overlapping and nonoverlapping genes among the lists of genes selected by two different filters. The filtered gene subset F is formed by taking the union of the lists of genes obtained by all the filters. After that, the genes in F are selected by means of a wrapper consisting of multiple (e.g., three in Fig. 2) classifiers. Since different classifiers may give different classification labels for the same sample, there is a need to resolve this conflict when it arises. It is natural to resort to some kind of voting scheme among the classifiers. Two possibilities are majority voting and unanimous voting. The advantage of majority voting is that there will always be a decision if the number of classifiers is odd. However, the "winning" class label may win by only a small margin, in which case the classification output may not be reliable. For this reason, we have chosen to use unanimous voting to decide on the overall classification output based on the outputs of the classifiers. In the case where a unanimous vote cannot be reached, the classification output is regarded as indecisive (denoted by "X"). Table 1 illustrates the voting result for all possible combinations of the outputs of three classifiers in a two-class (with labels "A" and "B") classification problem where the sample has a true class label "A." Out of these combinations of classifier outputs, only two will produce a unanimous vote of class A or B, one of which is right and the other is wrong (with prediction status *R* and *W*, respectively), and the other six produce an indecisive outcome (denoted as *I*).

By the unanimous voting scheme, each sample can be categorized as either *R*, *I*, or *W*. This information will be used for gene selection in the MFMW model. The number of *I* and *W* prediction statuses across all samples will be used to determine the usefulness of the set of genes. A prediction status *W* implies that all classifiers misclassify the sample under consideration, and is therefore particularly undesirable. Our first objective in gene selection is therefore to minimize the number of *W*, preferably to zero. Beyond that, the next goal will be to reduce the number of *I* (i.e., the indecisive cases). The procedure is summarized in the following algorithm:

TABLE 2
Microarray Data Sets Utilized for Experiments

Dataset	*No. of Classes (Samples): Distribution among classes*	
LEU 38 [1]	2 classes (38 samples)	27 ALL & 11 AML
COL 62 [28]	2 classes (62 samples)	40 tumor & 22 normal
BR-ER 49 [29]	2 classes (49 samples)	25 ER+ & 24 ER-
LYM 77 [30]	2 classes (77 samples)	58 DLBCL & 19 FL
PROS 102 [31]	2 classes (102 samples)	50 normal & 52 tumor
LUNG 181 [32]	2 classes (181 samples)	31 MPM & 150 ADCA

MFMW Algorithm:

FILTER

1. Choose m = the number of genes to be selected by each filter.
2. For each filter $FT_i(i = 1, 2, 3)$,
   a) Calculate the statistical scores for all genes and rank the scores from the highest to the lowest.
   b) Select m genes with top ranking scores in each list.
3. Take the union of the list of genes obtained by $FT_i(i = 1, 2, 3)$ to give F.

WRAPPER

1. Initialize $P = \Phi$ (the set of selected genes) and $n = 0$ (number of genes in P).
2. Put $n = n + 1$; Repeat for all genes $g_n \in F \backslash P$,
   a) Using $P \cup \{g_n\}$ as the gene set, classify all samples using $CF_i(i = 1, 2, 3)$ individually.
   b) Based on unanimous voting, determine the number of samples that receive *W* or *I* prediction statuses.
   Select the gene subset with the smallest number of *W*. If more than one gene subset has the same number of *W*, the one(s) with the smallest number of *I* is/are chosen. Update P by adding the selected gene to the set.
3. Repeat step 2 until the number of *W* and number of *I* are below a prescribed threshold, or cannot be reduced further.
4. Output the selected genes set(s) P.

Ideally, both the number of *W* and the number of *I* should be zero at the end of the MFMW algorithm. In practice, this may not be achievable as the data set may contain outliers or corrupted samples, in which case a sufficiently small number of *W* and number of *I* will be regarded as acceptable. In any case, the algorithm will be terminated when first the number of *W* and then the number of *I* cannot be reduced further by adding more genes to the set P. Hence, the number of iterations in the algorithm (i.e., the number of loopings between steps 2 and 3) is given by the number of genes selected in the final MFMW model.

4 Experimental Results

The proposed MFMW model was evaluated by means of six DNA microarray data sets that have been used in the diagnosis of different cancers, namely LEU (LEU38 and LEU72) [1], COL62 [28], BR-ER49 [29], LYM77 [30], PROS102 [31], and LUNG181 [32]. A brief description of these data sets is given below with a summary in Table 2:

1. *LEU (LEU38 and LEU72) data set*—Seventy-two samples were analyzed with Affymetrix oligonucleotide arrays. The training set (LEU38) contains 38 samples (27 Acute Lymphoblastic Leukemia (ALL), 11 Acute Myeloid Leukemia (AML)), with 7,129 probes from 6,817 genes. The testing set contains an independent group of 34 samples: 20 ALL and 14 AML. LEU72 data set was formed by combining the training set and testing set [1].
2. *COL62 data set*—Gene expression in 40 tumor and 22 normal colon tissue samples were analyzed with Affymetrix oligonucleotide arrays [28]. Two thousand out of around 6,500 genes were selected based on the confidence in the measured expression levels.
3. *BR-ER49 data set*—Breast tumors are either positive or negative for the estrogen receptors (ER). The collection of tumors consists of 25 ER+ tumors and 24 ER- tumors. All samples were tested using HuGeneFL Affymetrix microarray, with 6,817 genes included in each chip [29].
4. *LYM77*—Samples were obtained from 77 patients with 58 of them belonging to diffuse large B-cell lymphoma (DLBCL) group, and the remaining 19 belong to follicular lymphoma (FL). These biopsies were subjected to transcriptional profiling using oligonucleotide microarrays containing probes from 6,817 genes [30].
5. *PROS102*—Samples were divided into two groups: with 50 normal and 52 tumor prostate specimens included. A total of 12,600 are present on the Affymetrix chips used for these experiments [31]. The preprocessing steps are specified in the website [33].
6. *LUNG181*—Samples came from two different types of lung cancer. Thirty-one of them belong to malignant pleural mesothelioma (MPM) and the remaining 150 belong to adenocarcinoma (ADCA). These experiments were performed on Affymetrix human U95A oligonucleotide probe arrays. There are altogether 12,600 genes for each sample [32].

In the MFMW model, the three filters we used are SNR, TS, and PC (see Section 2). Each filter is used to generate a list of $m\ (=300)$ genes from the full set G. Taking the union of the three lists consolidates overlapping genes and reduces the size of the combined list of filtered genes for the respective data sets to: 484 for LEU38, 467 for LEU72, 451 for COL62, 463 for BR-ER49, 473 for LYM77, 455 for PROS102, and 473 for LUNG181. Then, three classifiers WV, k-NN $(k=3)$, and SVM (see Section 2) are used to determine the gene subset P. The training accuracy is defined as

$$\frac{T-N_W-N_I}{T}\times 100\ \%, \tag{3}$$

where T is the total number of samples in the data set and $N_W(N_I)$ is the number of samples with prediction status $W(I)$.

Since it is not appropriate to use the training accuracy as defined in (3) for evaluating the classification performance of MFMW, to get a more realistic estimate of the classification accuracy, the MFMW model is evaluated using LOOCV (only R outputs contribute toward the accuracy). Table 3 shows the comparison of MFMW against the best reported existing SFSW models on the six data sets [21], [22], [23], [34], [35]. Although LEU is a benchmark data set, we are not aware of any studies applying the SFSW model to the LEU38 data set. For ease of comparison, LEU72 data set was used. Results given in Table 3 show that MFMW achieves the same accuracy, namely 100 percent, as in [23] for the LEU72 data set and 98.34 percent, as in [34] for the LUNG181 data set. MFMW performs better in all existing SFSW models for the remaining four data sets.

TABLE 3
Comparison of MFMW against Existing SFSW Models on Six Data Sets

	Best Existing SFSW Model			*MFMW*	
Dataset	*No. of genes*	*Method*	*Performance*	*No. of genes*	*Performance*
			LOOCV		*LOOCV*
LEU 72 [23]	4	Information theoretic gene selection + C4.5 Classifier	100 %	4	100 %
COL 62 [34]	4	Redundancy-based Feature Selection + C4.5 Classifier	93.55 %	6	95.16 %
BR-ER 49 [22]	12	Mutual Information + k-NN Classifier	89.80 %	6	100 %
LYM 77 [21]	5-6	Multi-Objective Evolutionary Algorithm + Weighted Voting	93.51 %	6	100 %
PROS 102 [35]	2-4	T-statistics + Support Vector Machine	97.06 %	6	98.04 %
LUNG 181 [34]	6	Redundancy Based Filter (RBF) algorithm + C4.5 Classifier	98.34 %	6	98.34 %

Details on the best set of selected genes P using MFMW for the six different data sets are shown in Table 4. Accession numbers of the Affymetrix chip, as well as the symbols and names of our selected genes, are given.

Next, we examine the results for LYM77 in more details. Table 5 shows the ranking of the MFMW-selected genes in the original lists of genes generated by SNR, TS, and PC. The first two genes selected by MFMW are either present in both the SNR and TS lists, or in the PC list only. This shows that using any one of these filters alone cannot generate this best subset of genes that yields 100 percent classification accuracy.

To further compare the MFMW with SFSW approaches, nine SFSW models are built by considering all possible combinations of the three filters and three classifiers used in our MFMW model. Of the six data sets we used, only LEU contains an independent testing data set. We first perform gene selection on LEU38 (using SFSW and MFMW, respectively). Out of the nine SFSW models, eight of them achieve perfect training classification accuracies, with different number of selected genes (see Table 6). When the chosen gene lists of the nine SFSW models are compared, few overlaps can be found (results not shown). This is undesirable as the genes selected depend on which filter and wrapper are used, and hence the selected genes can hardly be justified as

TABLE 4
Lists of Genes (Identified by Accession Numbers of the Affymetrix, Symbols, and Names) Selected by MFMW for Different Data Sets

Dataset	Accession Number	Symbol	Name
LEU 72	J03779_at	MME	Membrane metallo-endopeptidase
	M23197_at	CD33	CD33 molecule
	M55150_at	FAH	Fumarylacetoacetate hydrolase
	D13988_at	GDI2	GDP dissociation inhibitor 2
COL 62	Hsa.9218	t51858	EST
	Hsa.1832	MYL9	Myosin, light chain 9, regulatory
	Hsa.31933	CNNM4	Cyclin M4
	Hsa.1410	r54097	EST
	Hsa.2291	GSN	Gelsolin (amyloidosis, Finnish type)
	Hsa.2842	LAMC2	Laminin, gamma 2
BR-ER 49	X93499_at	x93499_at	EST
	X89109_s_at	CORO1A	Coronin, actin binding protein, 1A
	J04027_at	ATP2B1	ATPase, Ca transporting, plasma membrane 1
	U12424_s_at	GPD2	Glycerol-3-phosphate dehydrogenase 2
	D61391_at	PRPSAP1	Phosphoribosyl pyrophosphate synthetase-associated protein 1
	S54005_s_at	TMSB10	Thymosin, beta 10
LYM 77	X59812_at	CYP27A1	Cytochrome P450, family 27, subfamily A, polypeptide 1
	U00238_rna1_at	PPAT	Phosphoribosyl pyrophosphate amidotransferase
	L04510_at	TRIM23	Tripartite motif-containing 23
	D87119_at	TRIB2	Tribbles homolog 2 (Drosophila)
	U73514_at	HSD17B10	Hydroxysteroid (17-beta) dehydrogenase 10
	D00596_at	d00596_at	EST
PROS 102	36151_at	PLD3	Phospholipase D family, member 3
	40597_g_at	TCOF1	Treacher Collins-Franceschetti syndrome 1
	39527_at	PAIP2B	Poly(A) binding protein interacting protein 2B
	364_s_at	PLCB3	Phospholipase C, beta 3
	37550_at	F8	Coagulation factor VIII, procoagulant component (hemophilia A)
	32129_at	ZNF364	Zinc finger protein 364
LUNG 181	32186_at	SLC7A5	Solute carrier family 7, member 5
	38614_s_at	OGT	O-linked N-acetylglucosamine transferase
	41283_at	HNRPH3	Heterogeneous nuclear ribonucleoprotein H3
	35855_s_at	GRIK1	Glutamate receptor, ionotropic, kainate 1
	33541_s_at	LAIR2	Leukocyte-associated immunoglobulin-like receptor 2
	35081_at	FGF9	Fibroblast growth factor 9 (glia-activating factor)

biomarkers. Prediction accuracies were obtained on the 34 testing samples (see Table 6), and it can be observed that the testing accuracies vary significantly from 67.65 percent to 91.18 percent even though the training accuracies are mostly 100 percent (except one at 94.73 percent). In contrast (see Table 7), a total of 21 pairs of genes were selected using MFMW, each giving a perfect training accuracy (defined as in (3)). These 21 MFMW models are more robust in their testing accuracies, ranging from 91.18 percent to 100 percent.

Similar experiments using all nine possible combinations of SFSW models were performed for the remaining five data sets (detailed results of individual SFSW models not shown), and LOOCV was used for evaluation. For all these data sets, the estimated classification accuracies of the proposed MFMW model (given in Table 3) consistently outperform those obtained from every one of the nine individual combinations. The results are summarized in Table 8.

TABLE 5
Ranking of MFMW-Selected Genes in the Original SNR, TS, and PC Lists in LYM77 Data Set

Gene Accession #	Rank in SNR list	Rank in TS list	Rank in PC list
U94832_at	NA	NA	162
D78134_at	1	1	NA
D55716_at	101	101	3

NA means that the corresponding gene is "Not Available" in the list.

TABLE 6
Nine SFSW Models for LEU

	WV	3-NN	SVM
	Training / Testing Acc	Training / Testing Acc	Training / Testing Acc
	# of genes	# of genes	# of genes
SNR	100 % / 76.47 %	100 % / 79.41 %	100 % / 91.18 %
	2 genes	3 genes	3 genes
TS	100 % / 88.24 %	100 % / 91.18 %	100 % / 79.41 %
	2 genes	3 genes	3 genes
PC	100 % / 67.65 %	94.73 % / 82.35 %	100 % / 82.35 %
	3 genes	15 genes	8 genes

TABLE 7
Twenty-One Pairs of Genes (Shown by Accession #) Selected by MFMW on Leukemia Training Set, with the Training and Testing Accuracies Obtained Using MFMW

No. of Pairs	Accession # (First gene)	Accession # (Second gene)	Training (38 samples)	Testing (34 samples)
2	X95715_at	J04615_at, X62535_at	100 %	91.18 %
13	X95715_at	HG2855-HT2995_at, M11722_at, M92287_at, U32944_at, U33822_at, X59417_at, X66401_cds1_at, X74801_at, X86691_at, Z15115_at, D26156_s_at, U22376_cds2_s_at, M63838_s_at	100 %	94.12 %
4	X95715_at	M19507_at, M33680_at, U37055_rna1_s_at, X07730_at	100 %	97.06 %
2	X95715_at	HG1612-HT1612_at, X51521_at	100 %	100 %

5 Discussion

5.1 Study on Varying the Number of Genes Selected by Each Filter

To study the effect of varying the number of genes selected by each filter, we repeat all the experiments of Section 4 for each data set with $m = 200$ and 350, to ascertain whether $m = 300$ is a reasonable choice. The LOOCV performance results for different values of m are summarized in Table 9. Note that those in bold are the best LOOCV performance for each data set.

TABLE 8
Estimated Classification Accuracies of the MFMW Model versus the Best of SFSW Models, for the COL62, BR-ER49, LYM77, PROS102, and LUNG181 Data Sets

Dataset	Best SFSW Model				MFMW Model	
	Filter	Wrapper	# of genes	LOOCV	# of genes	LOOCV
COL 62	SNR	SVM	10	90.32 %	6	95.16 %
BR-ER 49	PC	WV	12	73.47 %	6	100 %
LYM 77	PC	SVM	8	94.81 %	6	100 %
PROS 102	TS	SVM	4	97.06 %	6	98.04 %
LUNG 181	SNR	3NN	5	97.79 %	6	98.34 %

TABLE 9
LOOCV Performance Accuracies Variation on Different Number of Genes Selected by Each Filter

m / Dataset		LEU 38	LEU 72	COL 62	BR-ER 49	LYM 77	PROS 102	LUNG 181
200	No. of genes	2	4	6	6	4	6	9
	Performance	100%	100%	96.77%	95.92%	98.70%	96.08%	95.03%
300	No. of genes	2	4	6	6	6	5	6
	Performance	100%	100%	95.16%	100%	100%	98.04%	98.34%
350	No. of genes	2	4	6	6	6	5	6
	Performance	100%	100%	91.94%	95.92%	100%	98.04%	98.34%

For all the data sets, the best accuracy can be achieved with $m = 300$, except for COL62, where the accuracy has decreased from the case of $m = 200$ to that of $m = 300$ due to the increase of one misclassification. Theoretically, the use of a larger gene subset should always increase the accuracy. However, the MFMW algorithm is not guaranteed to converge to the global optimal solution, and the results of Table 9 suggest that the use of an unnecessarily large gene set may cause the algorithm to be trapped at a local minimum, as are the cases when m is increased from 300 to 350 for the COL62 and BR-ER49 data sets. Hence, for the data sets under consideration, $m = 300$ would be the best choice of number of genes to be retained by each filter.

5.2 Study on Varying the Number of Filters and Classifiers in the MFMW Model

The idea of using multiple filters is to include as many relevant genes of different characteristics as possible using different filters, so as not to leave out any potential biomarkers at the filtering stage. In this regard, it would be beneficial to include more filters. Apart from the three statistical filters used in our study, most of the other filter metrics are based on information theory, which do not quite match the classifiers in our study. Hence, by using SNR, TS, and PC, we have effectively included all filters relevant to our classification methods. The remaining question is whether a smaller number of filters can serve the same purpose as the three selected filters. Our study shows that for "good" data sets such as LEU38, two filters may generate sufficient genes for achieving the same accuracy as three filters. However, for more difficult data sets such as LYM77, some two-filter combinations fail to include some useful genes and as a result cannot achieve the same accuracy as the three-filter MFMW model.

To study the effect of changing the number of classifiers on performance, we select seven classifiers, namely, LDA, Quadratic Discriminant Analysis (QDA), WV, two k-NN classifiers (with $k = 3$ or 5), and two SVM classifiers (with kernels of polynomial order 1 or 2), and use different combinations of two and four classifiers in the MFMW model. For the LYM77 data set, 3 out of the 21 $(= {}_7C_2)$ cases of two-classifier MFMW model cannot achieve the same accuracy as the three-classifier MFMW model, whereas the use of four classifiers do not improve the accuracy any further. Hence, the use of three classifiers is optimal for this data set.

Clearly, it is not possible to conclude that the use of three filters and three classifiers is optimal in general—even testing this hypothesis just for all the data sets considered in this paper is computationally prohibitive. However, our experiences based on a wide range of experiments (many of which are not reported here) suggest that the choice of three filters and three classifiers in an MFMW model incorporating the kind of popular filtering and classification methods considered in this paper should be close to an optimal compromise between accuracy and computational efficiency.

TABLE 10
Five of the 21 Genes Selected by MFMW as Complementary to ZYX Are Confirmed by Other Studies as Significantly over- or under-Expressed (with Very Low p-Values)

Accession # (Gene Symbol)	Over/Under-expressed (class)	Dataset: p-value	Function
M92287_at (CCND3)	Over-expressed (ALL)	Golub [1]: $1.3E^{-12}$	Essential for the control of the cell cycle at the G1/S transition.
		Andersson [39]: $2.2E^{-8}$	
Z15115_at (TOP2B)	Over-expressed (ALL)	Golub [1]: $1.2E^{-13}$	Control of topological states of DNA; widely used in the treatment of AML and ALL
		Armstrong [38]: $7.8E^{-10}$	
		Andersson [39]: $2.5E^{-7}$	
U22376_cds2_s_at (MYB)	Over-expressed (ALL)	Armstrong [38]: $1.3E^{-6}$	Involved in hematopoiesis of both AML and ALL
M63838_s_at (IFI16)	Over-expressed (ALL)	Armstrong [38]: $1.2E^{-14}$	Controls cellular proliferation by modulating the functions of cell cycle regulatory factors, e.g. p53
		Andersson [39]: $1.2E^{-11}$	
		Golub [1]: $1.8E^{-6}$	
M19507_at (MPO)	Under-expressed (ALL)	Golub [1]: $7.6E^{-8}$	Involved in phagocytosis
		Andersson [39]: $2.7E^{-9}$	

Functions of these genes as given in Oncomine also suggest that they are related to AML/ALL.

5.3 Biological Interpretation of the Selected Genes

We would like to explore the biological significance of the MFMW-selected genes on all data sets. The contributions to each cancer types of these genes have been confirmed by other existing studies. With the help of the Oncomine database [36], where p-values obtained by performing student t-tests between two classes in a data set are given for each gene, we can determine how significantly over- or under-expressed the MFMW-selected genes are.

1. *LEU data set*—First, genes chosen from LEU38 data set are discussed (see Table 7). The first selected gene in our gene subsets is X95715_at (with symbol ZYX), which is also selected in [1] as under-expressed in the ALL class (with p-value: $4.1E^{-16}$). The gene may be a component of a signal transduction pathway that mediates adhesion-stimulated changes in gene expression. Note that MFMW is capable of consistently identifying ZYX as the first gene in all 21 gene sets. Furthermore, of the 21 genes given in Table 7 which, when added to ZYX give a perfect training accuracy, functions of five of them are proven to be related to leukemia with very low p-values. Studies which report these five genes as statistically over- or under-expressed using p-values generated by student t-test are summarized in Table 10.

   Next, we consider the four genes in Table 4 as selected by MFMW using the LEU72 data set. Note

that ZYX is not among one of the four selected genes. This may be due to the inconsistencies present in the training and testing data sets. Of the four genes, two of them (namely MME and CD33) are biologically useful. MME is an important cell surface marker in the diagnostic of human ALL [37]. It is over-expressed in ALL of three data sets: p-value: 7.5E^{-20} [38], p-value: 4.9E^{-15} [39], and p-value: 5.1E^{-5} [1]. CD33 is under-expressed in ALL with a p-value of 1.2E^{-20}, 9.9E^{-15}, and 2.4E^{-5} in data sets of Andersson et al. [39], Golub et al. [1], and Armstrong et al. [38], respectively. It induces apoptosis in AML (in vitro) [40].

2. *COL62 data set*—Since no genes were identified to distinguish between cancerous and normal samples in [28], no comparison can be made. Our selected genes include CD36, GSN, and MORF4L2. CD36 was reported to increase the risk of predisposing colon cancer [41] by functioning as a cell adhesion molecule. GSN is likely an effector of morphologic change during apoptosis, and is found to be over/under-expressed in cancerous samples by Alon et al. [28] (p-value: 1.1E^{-8}) and Notterman et al. [42] (p-value: 5.2E^{-8}). MORF4L2 is responsible for the activation of transcriptional programs associated with oncogene mediated growth induction.
3. *BR-ER49 data set*—None of our six MFMW-selected genes are in the list of genes claimed to be differentially expressed between ER+ and ER- samples in the original paper [29]. However, three potential biomarkers are found: PIM1, PGM5, and E2F2. PIM1 contributes to both cell proliferation and survival and thus facilitates tumorigenesis. It is over-expressed in ER+ group: p-value of 4.5E^{-14} [43] and p-value of 3.2E^{-11} [44]. E2F2 controls genes regulating S phase in the cell cycle regulation. It is found to be under-expressed in ER+: p-value of 4.2E^{-8} [43] and p-value of 2.9E^{-6} [45]. Though no function is assigned for PGM5, it was found to be over-expressed in ER+: p-value of 3.8E^{-6} [43].
4. *LYM77 data set*—All three MFMW-selected genes, namely KHSRP, CIRBP, and MCM7, are good biomarkers for lymphoma. The first selected gene is KHSRP, a type of KH-type splicing regulatory protein, which plays a role in mRNA trafficking. This gene is also found to be over-expressed (p-value: 6.8E^{-8}) in DLBCL group of Alizadeh's lymphoma cDNA data set (IMAGE: 1340925) [46]. The second selected gene, CIRBP, is significantly over-expressed (p-value: 9.4E^{-9}) for the FL group [30]. This cold inducible RNA binding protein plays an essential role in cold-induced suppression of cell proliferation. The last one, MCM7, was selected by Shipp et al. [30]. It was under-expressed (p-value: 8.1E^{-11}) for the FL group. It is required for DNA replication and cell proliferation.
5. *PROS102 data set*—Genes selected by our MFMW model have little correlation with existing knowledge on the differentiation between prostate cancer and normal tissues. Only gene PAIP2B is significantly over-expressed in the cancerous tissue in [47] (p-value: 7.6E^{-7}).
6. *LUNG181 data set*—Only two of our selected genes are good at differentially between the two types of lung cancers (MPM and ADCA). The first gene, SLC7A5, is significantly over-expressed in ADCA of three data sets: p-value: 8.6E^{-8} [48], p-value: 3.3E^{-7} [49], and p-value: 3.3E^{-5} [32]. The second gene is OGT, which is also significantly over-expressed in ADCA in [50] (p-value: 2.3E^{-6}).

6 Conclusion

Although combining filters and wrappers for gene selection appears to be a sensible approach, SFSW models yield results that are dependent on the specific choice of filter and wrapper. In this paper, an MFMW approach has been proposed to overcome the difficulties of SFSW models, to improve the classification accuracy, and possibly to identify biomarker genes. By combining genes selected by different filters, an expanded feature subset is created to ensure that relevant genes are not left out. Improvement in classification accuracy is achieved through the use of multiple filters with a unanimous voting scheme.

The proposed MFMW model was tested on six microarray data sets containing samples of cancerous versus normal, or subtypes of diseases. The experimental results show that the LOOCV estimated classification accuracies obtained by the MFMW model outperforms that of SFSW models generated from all possible combinations of filters and wrappers used in the MFMW model. Many of the MFMW-selected genes have been confirmed to be biomarkers or contribute to the development of particular cancers by other studies.

We have not compared the results of the MFMW with classification results using popular methods based on PCA because of the different nature of the two classes of methods. PCA methods perform gene selection by feature extraction, whereas MFMW is designed to optimize the classification accuracy. Hence, MFMW is expected to outperform the feature extraction methods, as can be readily shown by comparing the results presented in this paper with existing ones based on feature extraction methods.

Acknowledgments

The authors would like to thank the reviewers for their valuable comments which helped improve the manuscript in many ways. This work was supported by a HKU CRCG research grant.

References

[1] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science,* vol. 286, no. 5439, pp. 531-537, 1999.

[2] L. Li, C.R. Weinberg, T.A. Darden, and L.G. Pedersen, "Gene Selection for Sample Classification Based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the GA/KNN Method," *Bioinformatics,* vol. 17, no. 12, pp. 1131-1142, 2001.

[3] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler, "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data," *Bioinformatics,* vol. 16, no. 10, pp. 906-914, 2000.

[4] M.M. Xiong, L. Jin, W. Li, and E. Boerwinkle, "Tumor Classification Using Gene Expression Profiles," *Biotechniques,* vol. 29, pp. 1264-1270, 2000.

[5] Y. Wang, I.V. Tetko, M.A. Hall, E. Frank, A. Facius, K.F.X. Mayer, and H.W. Mewes, "Gene Selection from Microarray Data for Cancer Classification—A Machine Learning Approach," *Computational Biology and Chemistry,* vol. 29, no. 1, pp. 37-46, 2005.

[6] M. Xiong, X. Fang, and J. Zhao, "Biomarker Identification by Feature Wrappers," *Genome Research,* vol. 11, pp. 1878-1887, 2001.

[7] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification.* Wiley, 2000.

[8] M.L. Chow, E.J. Moler, and I.S. Mian, "Identifying Marker Genes in Transcription Profiling Data Using a Mixture of Feature Relevance Experts," *Physiological Genomics,* vol. 5, pp. 99-111, 2001.

[9] P. Langley, "Selection of Relevant Features in Machine Learning," *Proc. AAAI Fall Symp. Relevance,* pp. 1-5, 1994.

[10] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," *Artificial Intelligence,* vol. 97, nos. 1/2, pp. 273-324, 1997.

[11] T.P. Speed, *Statistical Analysis of Gene Expression Microarray Data.* Chapman & Hall/CRC, 2003.

[12] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, "Tissue Classification with Gene Expression Profiles," *Proc. Fourth Ann. Int'l Conf. Computational Molecular Biology (RECOMB '00),* pp. 54-64, 2000.

[13] C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," *Proc. IEEE Computer Soc. Bioinformatics Conf. (CSB '03),* pp. 523-528, 2003.

[14] C. Lai, M.J.T. Reinders, L.J. van't Veer, and L.F.A. Wessels, "A Comparison of Univariate and Multivariate Gene Selection Techniques for Classification of Cancer Datasets," *BMC Bioinformatics,* vol. 7, no. 235, 2006.

[15] R. Diaz-Uriarte and S. Alvarez de Andres, "Gene Selection and Classification of Microarray Data Using Random Forest," *BMC Bioinformatics,* vol. 7, no. 3, 2006.

[16] T.K. Paul and H. Iba, "Extraction of Informative Genes from Microarray Data," *Proc. Genetic and Evolutionary Computation Conf. (GECCO '05),* pp. 453-460, 2005.

[17] I. Inza, P. Larranaga, R. Blanco, and A.J. Cerrolaza, "Filter versus Wrapper Gene Selection Approaches in DNA Microarray Domains," *Artificial Intelligence in Medicine,* vol. 31, no. 2, pp. 91-103, 2004.

[18] W. Li and Y. Yang, "How Many Genes are Needed for a Discriminant Microarray Data Analysis?" *Proc. Critical Assessment of Microarray Data Analysis Workshop (CAMDA '00),* pp. 137-150, 2000.

[19] T. Li, C. Zhang, and M. Ogihara, "A Comparative Study of Feature Selection and Multiclass Classification Methods for Tissue Classification Based on Gene Expression," *Bioinformatics,* vol. 20, no. 15, pp. 2429-2437, 2004.

[20] E.P. Xing, M.I. Jordan, and R.M. Karp, "Feature Selection for High-Dimensional Genomic Microarray Data," *Proc. Int'l Conf. Machine Learning (ICML '01),* pp. 601-608, 2001.

[21] J. Liu and H.B. Zhou, "Tumor Classification Based on Gene Microarray Data and Hybrid Learning Method," *Proc. Int'l Conf. Machine Learning and Cybernetics,* pp. 2275-2280, 2003.

[22] X. Liu, A. Krishnan, and A. Mondry, "An Entropy-Based Gene Selection Method for Cancer Classification Using Microarray Data," *BMC Bioinformatics,* vol. 6, no. 76, 2005.

[23] M. Ng and L. Chan, "Informative Gene Discovery for Cancer Classification from Microarray Expression Data," *Proc. IEEE Workshop Machine Learning for Signal Processing (MLSP '05),* pp. 393-398, 2005.

[24] J.W. Lee, J.B. Lee, M. Park, and S.H. Song, "An Extensive Comparison of Recent Classification Tools Applied to Microarray Data," *Computational Statistics and Data Analysis,* vol. 48, no. 4, pp. 869-885, 2005.

[25] V.N. Vapnik, *Statistical Learning Theory.* Wiley, 1998.

[26] C. Ambroise and G.J. McLachlan, "Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data," *Proc. Nat'l Academy of Sciences USA,* vol. 99, no. 10, pp. 6562-6566, 2002.

[27] R. Simon, M.D. Radmacher, K. Dobbin, and L.M. McShane, "Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification," *J. Nat'l Cancer Inst.,* vol. 95, no. 1, pp. 14-18, 2003.

[28] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proc. Nat'l Academy of Sciences USA,* vol. 96, no. 12, pp. 6745-6750, 1999.

[29] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J.A. Olson Jr., J.R. Marks, and J.R. Nevins, "Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles," *Proc. Nat'l Academy of Sciences USA,* vol. 98, no. 20, pp. 11462-11467, 2001.

[30] M.A. Shipp, K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C.T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus, T.S. Ray, M.A. Koval, K.W. Last, A. Norton, T.A. Lister, J. Mesirov, D.S. Neuberg, E.S. Lander, J.C. Aster, and T.R. Golub, "Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling and Supervised Machine Learning," *Nature Medicine,* vol. 8, pp. 68-74, 2002.

[31] D. Singh, P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D'Amico, and J. Richie, "Gene Expression Correlates of Clinical Prostate Cancer Behavior," *Cancer Cell,* vol. 1, no. 2, pp. 203-209, 2002.

[32] G.J. Gordon, R.V. Jensen, L.L. Hsiao, S.R. Gullans, J.E. Blumenstock, S. Ramaswamy, W.G. Richards, D.J. Sugarbaker, and R. Bueno, "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma," *Cancer Research,* vol. 62, no. 17, pp. 4963-4967, 2002.

[33] http://download.cancercell.org/supplementarydata/ccell/1/2/203/DC1/index.htm, 2008.

[34] L. Yu and H. Liu, "Redundancy Based Feature Selection for Microarray Data," *Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '04),* pp. 737-742, 2004.

[35] F. Tan, X. Fu, H. Wang, Y. Zhang, and A.G. Bourgeois, "A Hybrid Feature Selection Approach for Microarray Gene Expression Data," *Proc. Int'l Conf. Computational Science (ICCS '06),* pp. 678-685, 2006.

[36] D.R. Rhodes, S. Kalyana-Sundaram, V. Mahavisno, R. Varambally, J. Yu, B.B. Briggs, T.R. Barrette, M.J. Anstet, C. Kincead-Beal, P. Kulkarni, S. Varambally, D. Ghosh, and A.M. Chinnaiyan, "Oncomine 3.0: Genes, Pathways, and Networks in a Collection of 18,000 Cancer Gene Expression Profiles," *Neoplasia,* vol. 9, no. 2, pp. 166-180, 2007.

[37] *Swissprot,* http://www.ebi.ac.uk/swissprot/, 2008.

[38] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub, and S.J. Korsmeyer, "MLL Translocations Specify a Distinct Gene Expression Profile that Distinguishes a Unique Leukemia," *Nature Genetics,* vol. 30, no. 1, pp. 41-47, 2001.

[39] A. Andersson, C. Ritz, D. Lindgren, P. Edén, C. Lassen, J. Heldrup, T. Olofsson, J. Råde, M. Fontes, A. Porwit-MacDonald, M. Behrendtz, M. Höglund, B. Johansson, and T. Fioretos, "Microarray-Based Classification of a Consecutive Series of 121 Childhood Acute Leukemias: Prediction of Leukemic and Genetic Subtype as Well as of Minimal Residual Disease Status," *Leukemia,* vol. 21, no. 6, pp. 1198-1203, 2007.

[40] C. Vitale, C. Romagnani, A. Puccetti, D. Olive, R. Costello, L. Chiossone, A. Pitto, A. Bacigalupo, L. Moretta, and M.C. Mingari, "Surface Expression and Function of p75/AIRM-1 or CD33 in Acute Myeloid Leukemias: Engagement of CD33 Induces Apoptosis of Leukemic Cells," *Proc. Nat'l Academy of Sciences USA,* vol. 98, no. 10, pp. 5764-5769, 2001.

[41] K. Kuriki, N. Hamajima, H. Chiba, Y. Kanemitsu, T. Hirai, T. Kato, T. Saito, K. Matsuo, K. Koike, S. Tokudome, and K. Tajima, "Relation of the CD36 Gene A52C Polymorphism to the Risk of Colorectal Cancer among Japanese, with Reference to with the Aldehyde Dehydrogenase 2 Gene Glu487Lys Polymorphism and Drinking Habit," *Asian Pacific J. Cancer Prevention,* vol. 6, no. 1, pp. 62-68, 2005.

[42] D.A. Notterman, U. Alon, A.J. Sierk, and A.J. Levine, "Transcriptional Gene Expression Profiles of Colorectal Adenoma, Adenocarcinoma, and Normal Tissue Examined by Oligonucleotide Arrays," *Cancer Research,* vol. 61, pp. 3124-3130, 2001.

[43] M.J. van de Vijver, Y.D. He, L.J. van't Veer, H. Dai, A.A. Hart, D.W. Voskuil, G.J. Schreiber, J.L. Peterse, C. Roberts, M.J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E.T. Rutgers, S.H. Friend, and R. Bernards, "A Gene-Expression Signature as a Predictor of Survival in Breast Cancer," *New England J. Medicine,* vol. 347, no. 25, pp. 1999-2009, 2002.

[44] Y. Wang, J.G. Klijn, Y. Zhang, A.M. Sieuwerts, M.P. Look, F. Yang, D. Talantov, M. Timmermans, M.E. Meijer-van Gelder, J. Yu, T. Jatkoe, E.M. Berns, D. Atkins, and J.A. Foekens, "Gene-Expression Profiles to Predict Distant Metastasis of Lymph-Node-Negative Primary Breast Cancer," *Lancet,* vol. 365, no. 9460, pp. 671-679, 2005.

[45] L.J. van 't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, and S.H. Friend, "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer," *Nature,* vol. 415, no. 6871, pp. 530-536, 2002.

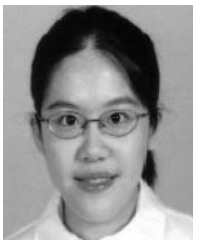
[46] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson Jr., L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, and L.M. Staudt, "Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling," *Nature,* vol. 403, no. 6769, pp. 503-511, 2000.

[47] J. Lapointe, C. Li, J.P. Higgins, M. van de Rijn, E. Bair, K. Montgomery, M. Ferrari, L. Egevad, W. Rayford, U. Bergerheim, P. Ekman, A.M. DeMarzo, R. Tibshirani, D. Botstein, P.O. Brown, J.D. Brooks, and J.R. Pollack, "Gene Expression Profiling Identifies Clinically Relevant Subtypes of Prostate Cancer," *Proc. Nat'l Academy of Sciences USA,* vol. 101, no. 3, pp. 811-816, 2004.

[48] A. Bhattacharjee, W.G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E.J. Mark, E.S. Lander, W. Wong, B.E. Johnson, T.R. Golub, D.J. Sugarbaker, and M. Meyerson, "Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct Adenocarcinoma Subclasses," *Proc. Nat'l Academy of Sciences USA,* vol. 98, no. 24, pp. 13790-13795, 2001.

[49] S. Tomida, K. Koshikawa, Y. Yatabe, T. Harano, N. Ogura, T. Mitsudomi, M. Some, K. Yanagisawa, T. Takahashi, H. Osada, and T. Takahashi, "Gene Expression-Based, Individualized Outcome Prediction for Surgically Treated Lung Cancer Patients," *Oncogene,* vol. 23, pp. 5360-5370, 2004.

[50] A.H. Bild, G. Yao, J.T. Chang, Q. Wang, A. Potti, D. Chasse, M.B. Joshi, D. Harpole, J.M. Lancaster, A. Berchuck, J.A. Olson, J.R. Marks, H.K. Dressman, M. West, and J.R. Nevins, "Oncogenic Pathway Signatures in Human Cancers as a Guide to Targeted Therapies," *Nature,* vol. 439, no. 7074, pp. 353-357, 2006.



Yukyee Leung received the BEng degree in medical engineering from the University of Hong Kong. She is currently a PhD student in the Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong. Her research interests include machine learning, pattern recognition, bioinformatics, molecular biology, and specifically microarray data mining. She is a student member of the IEEE Engineering in Medicine and Biology Society (EMBS).

Yeungsam Hung received the BSc (Eng) degree in electrical engineering and the BSc degree in mathematics from the University of Hong Kong and the MPhil and PhD degrees from the University of Cambridge. He has worked as a research associate at the University of Cambridge and as a lecturer at the University of Surrey before he joined the University of Hong Kong, where he is currently a professor and the head of the Department of Electrical and Electronic Engineering. His research interests include robust control systems theory, robotics, computer vision and, more recently, biomedical engineering. He is a chartered engineer (UK) and a fellow of the Institution of Engineering and Technology (IET) and the Hong Kong Institution of Engineers (HKIE).

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.