Physiol Genomics
5: 99–111, 2001.

Identifying marker genes in transcription profiling data using a mixture of feature relevance experts

M. L. CHOW,[1,3] E. J. MOLER,[1,2] AND I. S. MIAN[1]
[1]*Radiation Biology and Environmental Toxicology Group, Department of Cell and Molecular Biology, Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720;* [2]*Chiron Corporation, Emeryville, California 94608; and* [3]*Gene Logic Incorporated, Berkeley, California 94704*

Received 25 September 2000; accepted in final form 30 January 2001

Chow, M. L., E. J. Moler, and I. S. Mian. Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. *Physiol Genomics* 5: 99–111, 2001.—Transcription profiling experiments permit the expression levels of many genes to be measured simultaneously. Given profiling data from two types of samples, genes that most distinguish the samples (marker genes) are good candidates for subsequent in-depth experimental studies and developing decision support systems for diagnosis, prognosis, and monitoring. This work proposes a mixture of feature relevance experts as a method for identifying marker genes and illustrates the idea using published data from samples labeled as acute lymphoblastic and myeloid leukemia (ALL, AML). A feature relevance expert implements an algorithm that calculates how well a gene distinguishes samples, reorders genes according to this relevance measure, and uses a supervised learning method [here, support vector machines (SVMs)] to determine the generalization performances of different nested gene subsets. The mixture of three feature relevance experts examined implement two existing and one novel feature relevance measures. For each expert, a gene subset consisting of the top 50 genes distinguished ALL from AML samples as completely as all 7,070 genes. The 125 genes at the union of the top 50s are plausible markers for a prototype decision support system. Chromosomal aberration and other data support the prediction that the three genes at the intersection of the top 50s, cystatin C, azurocidin, and adipsin, are good targets for investigating the basic biology of ALL/AML. The same data were employed to identify markers that distinguish samples based on their labels of T cell/B cell, peripheral blood/bone marrow, and male/female. Selenoprotein W may discriminate T cells from B cells. Results from analysis of transcription profiling data from tumor/nontumor colon adenocarcinoma samples support the general utility of the aforementioned approach. Theoretical issues such as choosing SVM kernels and their parameters, training and evaluating feature relevance experts, and the impact of potentially mislabeled samples on marker identification (feature selection) are discussed.

marker genes; mixture of experts; support vector machines; adipsin; cystatin C; azurocidin

DNA MICROARRAY TECHNOLOGY generates a panoramic survey of genes expressed in a sample of cells. Comparing the transcription profiles of different types of samples permits identification of marker genes, genes that best distinguish samples. When the samples correspond to different pathological states of the same tissue or subtypes of the same malignancy, transcription profiling holds promise as a method for classifying and analyzing cancers from a molecular rather than morphological perspective (1, 2, 11). Despite difficulties in obtaining sufficient, high quality, homogeneous tissue samples from an in situ environment rather than, for example, cell lines, transcription profiling affords an opportunity to identify novel and/or uncharacterized genes that are potential candidates for developing faster and more reliable systems for clinical diagnosis, prognosis, and monitoring. Furthermore, these marker genes represent putative targets for therapeutic agents and understanding the basic biology of the disorder. A typical profiling study measures the expression levels of thousands of genes (features) L across tens of samples N, with each sample labeled as being of one type or another. The problem considered here is that of identifying marker genes given N labeled L-feature sample profile vectors.

A variety of techniques have been employed to address three statistical tasks associated with analysis of profile data (3, 9, 11, 13, 16–18, 20–22). The first, unsupervised learning, involves discovering and characterizing the classes present in unlabeled profile vectors. This clustering procedure can suggest previously unrecognized cancer (sub)types. The second task, supervised learning, involves discriminating between profile vectors with different labels and assigning the label of a new profile vector. Given profiling data for a sample of unknown origin, this classification and prediction procedure can indicate the origin of the sample, for example, whether it is from tumor or nontumor tissue. The third task and subject of this work is feature relevance, ranking, and selection. This involves defining a feature relevance expert which *1*) implements an algorithm that quantitates the degree to which a gene distinguishes samples, *2*) reorders genes according to this relevance value, *3*) selects nested subsets of ranked genes and uses them to train a supervised learning system, and *4*) identifies highly

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: I. S. Mian, Dept. of Cell and Mol. Biol., MS 74-197, Life Sciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd., Berkeley, CA 94720 (E-mail: SMian@lbl.gov).

informative or marker genes based on the ability of subsets to assign accurately the label for samples not used for training, i.e., the generalization performance of the subset. Thus, gene subsets corresponding to marker genes can be identified by varying a single parameter, the number of ranked features used to train and evaluate the supervised learning system. For a given data set, different feature relevance experts can be compared via their generalization performance on the same number of ranked genes.

Recently, two independent studies employed different techniques to address the three aforementioned tasks. The first study applied naive Bayes models, support vector machines (SVMs) and naive Bayes global relevance (NBGR) (16) to sixty-two 1,988-feature experiment profile vectors derived from colon adenocarcinoma samples labeled as tumor or nontumor (2). The NBGR requires unlabeled profile vectors as input, since it is computed from the probability parameters of profile vector classes discovered by a naive Bayes model. The second study applied self-organizing maps (SOMs), neighborhood analysis, and weighted voting and gene/class correlation to seventy-two 7,070-feature experiment profile vectors derived from bone marrow (BM) and peripheral blood (PB) samples labeled as acute lymphoblastic or myeloid leukemia (ALL, AML) (11). The relevance measure, referred to here as the mean aggregate relevance (MAR), requires labeled profile vectors, since it is computed from the mean and standard deviation of the expression levels of genes in samples labeled ALL and AML. For the {Tumor, Nontumor} and {ALL, AML} binary supervised learning problems, each study identified 50 markers that had the same generalization performance as the full repertoire of, respectively, 1,988 or 7,070 genes (11, 16).

This work considers three distinct but interrelated feature relevance-, ranking-, and selection-related problems. Currently, the number of training examples, N sample profile vectors, is considerably smaller than their dimensionality, L measured gene expression levels ($N \ll L$). The first problem is identifying P marker genes for development of a robust decision support system to assign the cancer (sub)type for a new sample as accurately as or better than the original L genes ($P \ll L$). The second problem involves reducing the dimensionality even further by defining the Q marker genes best-suited for subsequent experimental investigations ($Q < P \ll L$). The third problem concerns multiply-labeled profile vectors and increasing the utility of profiling studies beyond their original purpose. Apart from the primary ALL and AML labels, each leukemia sample had 1–3 additional labels: {PB, BM}, {T cell, B cell}, and {Male, Female} (11). Since it is unlikely that all 7,070 genes are involved in differentiating ALL from AML, it is possible that some (or all) could provide a readout on other aspects of the samples. The question becomes whether the L genes analyzed to address a primary supervised learning problem can be employed to identify markers for secondary problems defined by additional sample labels. Here, a mixture of feature relevance experts is used to address the first and second problems. The validity of the premise underlying the third problem is demonstrated using data from the leukemia samples. Since submission of this work, a variety approaches for identifying marker genes have been proposed (see, for example, Refs. 7, 8, 10, 12, and 23).

METHODS AND APPROACH

Gene expression data. The transcription profiling studies reconsidered here both employed Affymetrix technology to monitor gene expression levels. The adenocarcinoma study provides measurements for 1,988 probes in 62 human colon adenocarcinoma tissue samples, 40 labeled *tumor* and 22 nontumor (2). The leukemia study provides measurements for 7,070 probes in 72 human leukemia samples, 47 *ALL* and 25 AML (total 72), 10 *PB* and 62 BM (total 72), 9 *T cell* and 38 B cell (total 47), and 26 *male* and 23 female (total 49) (11). For these five aforementioned binary supervised learning problems, samples having the label shown in italics are, without loss of generality, defined to be positive training examples.

Although not examined in this work, a variety of other supervised learning problems can be derived from the leukemia data. For example, a binary problem might involve distinguishing leukemia subtypes on the basis of tissue of origin ({ALL+PB, ALL+BM} or {AML+PB, AML+BM}). A multiclass problem could include discriminating samples on the basis of tissue origin and subtype ({ALL+PB, ALL+BM, AML+PB, AML+BM}). For convenience, each functionally defined nucleic acid sequence probe whose expression level is monitored will be termed a "gene," irrespective of whether it is actually a gene, an expressed sequence tag, or DNA from another source.

Feature relevance experts. The three feature relevance experts evaluated here implement relevance measures that are based upon labeled (MAR, MVR) or unlabeled (NBGR) sample profile vector training examples. These measures are designed to be illustrative rather than comprehensive, because, for example, all treat genes as independent of one another, whereas the transcription levels of some genes are likely to be correlated. In general, each measure generates a ranking of features and defines nested gene subsets $Top1 \subset Top2 \subset \ldots \subset TopL$, where L is the number of genes monitored in the profiling study (here $L = 1988, 7070$). $Top1$ denotes the top-ranked or most distinctive gene according to the relevance measure, $Top2$ denotes the top 2, and so on. Evaluating all possible gene subsets in terms of how well they perform on a particular classification and prediction problem using a supervised learning method (here SVMs) is a computationally demanding task. Hence, the focus is on a small number of selected gene subsets, for example, $Top4 \subset Top5 \subset Top11 \subset Top25 \subset Top50 \subset Top100 \subset TopL$, as well as the bottom and middle 50 ranked genes.

It remains to be determined whether the degrees of difficulty of the supervised learning and feature selection problems posed by the leukemia and adenocarcinoma data sets are typical of cancer profiling studies. The strategy deployed here is sufficiently general that other feature relevance measures, ranking and selection techniques, supervised learning methods, training and evaluation procedures, and methods for combining predictions from experts could be utilized.

Median vote relevance. For gene F_l, let x_l^n be its expression level in sample n. Let $\nu_i(F_l)$ and $\nu_j(F_l)$ be the median values for samples belonging to classes i (positive training examples) and j (negative examples). Each sample casts a vote

$V(n, l)$ according to whether the expression level is closer to the median value of class i or j. The median vote relevance (MVR) is the sum over all N samples

$$\mathrm{MVR}(F_l) = \sum_{n=1}^{N} V(n, l) = \begin{cases} 1 & \text{if } |\nu_i(F_l) - x_l^n| < |\nu_j(F_l) - x_l^n| \\ 0 & \text{Otherwise} \end{cases}$$

The larger the score, the better the gene distinguishes classes (two or more genes can have the same value). Although both require labeled training examples, the MVR is less sensitive to outliers than the mean-based MAR, because the median is a more robust estimate of the center of a population sample. MVR values were computed using a spreadsheet program.

Naive Bayes global relevance. Given the K classes identified and characterized by a naive Bayes model estimated from N unlabeled L-feature profile vectors, the NBGR (16) is the sum of the relevance over pairwise combinations of classes

$$\mathrm{NBGR}(F_l) = \frac{1}{K} \sum_{i=1}^{K} \sum_{j=i+1}^{K} \log \left[\frac{4}{N} \sum_{n=1}^{N} \frac{P(x_l^n|c_{i,l})P(x_l^n|c_{j,l})}{[P(x_l^n|c_{i,l}) + P(x_l^n|c_{j,l})]^2} \right]$$

$P(x_l^n|c_{k,l})$ is the probability of the expression level given class k. The greater the absolute magnitude, the better the gene distinguishes all K classes. A naive Bayes model was estimated using AutoClass C version 3.3 (5) and the 72 unlabeled 7,070-feature leukemia sample profile vectors (the reported expression values were not shifted or scaled in any way). An expectation maximization algorithm finds a mixture of Gaussian probability distributions, and a Bayesian approach finds the maximum posterior probability classification and optimum number of classes K. Thus, $P(x_l^n|c_{k,l}) = [2\pi\sigma_{k,l}^2]^{-1/2}\exp[-1/2\{(x_l^n - \mu_{k,l})/\sigma_{k,l}\}^2]$ where $[\mu_{k,l}, \sigma_{k,l}]$ is the [mean, standard deviation] of the Gaussian modeling class k. For each feature, gene l, a lower bound for $\sigma_{k,l}$ was set to 1/10 of the standard deviation of all N expression levels, $\{x_l^1, \ldots, x_l^N\}$.

A naive Bayes model of the adenocarcinoma experiment profile vectors identified four underlying classes (16) rather than the two indicated by the tumor and nontumor labels (2). NBGR values were calculated using Gaussian parameters determined directly from the values of gene F_l in the tumor and nontumor samples, i.e., $K = 2$. The generalization performance of the top 50 genes from this "supervised" NBGR expert was considerably worse than that of the top 50 from an "unsupervised" NBGR expert that employed the $K = 4$ classes estimated from data.

Mean aggregate relevance. This is the correlation between a gene and the ALL/AML classes (11). Unlike the MVR, the MAR utilizes both the location and spread of samples in classes i and j

$$\mathrm{MAR}(F_l) = \frac{\mu_{i,l} - \mu_{j,l}}{\sigma_{i,l} + \sigma_{j,l}}$$

where $[\mu_{i,l}, \sigma_{i,l}]$ and $[\mu_{j,l}, \sigma_{j,l}]$ are the mean and standard deviation of the log of the expression level of gene F_l in classes i and j. A large absolute magnitude signifies a strong correlation. A positive (negative) sign indicates that the gene is more highly expressed in class i (j). MAR(F_l) is related to the Fisher criterion score $|(\mu_{i,l} - \mu_{j,l})/(\sigma_{i,l}^2 + \sigma_{j,l}^2)|$.

Leukemia and adenocarcinoma genes: feature ranking and selection. The 7,070 genes in the leukemia data were ranked separately according to their NBGR value and MVR value for the labels {ALL, AML}, {PB, BM}, {T cell, B cell}, and {Male, Female} (a total of five different rankings). The 1,988 genes in the adenocarcinoma data were ranked separately according to their NBGR value and MVR value for the label {Tumor, Nontumor} (two different rankings). For each of these seven rankings, nine representative gene subsets were created by selecting different numbers of top-, middle-, and bottom-ranked genes. Two additional gene subsets based on the {ALL, AML} labels were defined. The first, taken from figure 3*A* of Ref. 11 and referred to as the MAR 50, represents the 25 genes with the highest positive values and the 25 genes with the highest negative values. The second subset consists of genes common to the MAR 50, the NBGR top 50, and the MVR top 50. For the multiply-labeled leukemia data, the NBGR ranking reflects the importance of genes in distinguishing ALL from AML, so it may be uninformative in terms of the other labels.

SVMs: training and evaluation. Because of the limited number of training examples, a leave-one-out cross validation strategy was utilized. A pool of N known positive and negative training examples was partitioned into two disjoint sets (here $N = 62, 72$). The estimation set, $N - 1$ examples, was used to determine the parameters of an SVM, and the test set, 1 example, was used to assess its generalization performance. The label assigned by a trained SVM to a test example can be a true positive (known positive test example, assigned positive label), true negative (negative example, negative label), false positive (negative example, positive label), or false negative (positive example, negative label). This procedure was repeated for each training example in turn. The generalization performance of these leave-one-out studies is the total number of SVMs that make true positive or true negative assignments (the maximum possible generalization performance is N). Elsewhere (11), the 72 leukemia training examples were partitioned into estimation and test sets containing 38 and 34 examples, respectively. The generalization performance of this "38 estimation, 34 test" partitioning is how many of the 34 test examples were assigned to be true positives or true negatives. The roles of the two sets were then reversed, and the generalization performance of a "34 estimation, 38 test" partitioning was determined in a similar manner.

In addition to training examples, estimating an SVM requires specifying an inner-product kernel function, a measure of similarity between two profile vectors $\mathbf{X}_L^i = \{x_1^i, \ldots, x_L^i\}$ and $\mathbf{X}_L^j = \{x_1^j, \ldots, x_L^j\}$. Since there is no general theory for determining the most appropriate kernel for a particular learning problem, two kernels were employed. The first was the dot product kernel $K(\mathbf{X}_L^i, \mathbf{X}_L^j) = \sum_{l=1}^{L} x_l^i x_l^j$. The second was a radial basis kernel function $K(\mathbf{X}_L^i, \mathbf{X}_L^j) = \exp(-\|\mathbf{X}_L^i - \mathbf{X}_L^j\|^2/2\sigma^2)$, where $\gamma = 1/2\,\sigma^2$ is a user-defined width parameter. Two different width parameters were used: $\gamma_\mathrm{f} = 0.01$, a data-independent value employed in earlier work (16); and 2) γ_d, a data-dependent value in which σ is set equal to the median of the Euclidean distances from each positive training example to the nearest negative training example (3).

SVMs were trained and evaluated using SVM$^{\text{light}}$ version 3.02 (15). Each gene subset was employed to create training examples in which the input profile vectors contained only the selected genes. Rather than working directly with the reported expression levels, x_l^n, each value was normalized using $x_l^n/[\sum_{l \in S}(x_l^n)^2]^{1/2}$ where S is the subset of interest. For simplicity and to illustrate the basic approach, genes were ranked once using all N training examples and not reranked for each estimation set. To account for unequal numbers of positive and negative examples, each estimation set was balanced by duplicating as many randomly chosen examples as necessary from the smaller set to yield the same number of examples as the larger set. Elsewhere (3), imbalanced data sets were handled by adding a diagonal to the kernel matrix (different values for positive and negative examples).

RESULTS

Leukemia sample profile vector classes. Each of the $N = 72$ samples could be assigned uniquely to one of three naive Bayes model classes because the probability of the profile vector for that class was 1.0 (Table 1). Although *class 3* contains only ALL samples, none of the other labels exhibit any clear association with specific classes. The unsupervised learning method utilized here determines the number of classes from the data, whereas the published SOM approach (11) requires this number be specified a priori (only a four class SOM was reported). For the adenocarcinoma data set also (16), the number of classes estimated by a naive Bayes model is greater than the two that might be expected given the {Tumor, Nontumor} labels. Further research is required to ascertain whether these discrepancies are the result of deficiencies in the modeling method or reflect the fine structure and complexity of the data that is masked by the original (known) labels. Interestingly, ranking genes on the basis of these estimated (pure and mixed) classes does not diminish the ability of top-ranked gene subsets to address the {ALL, AML} and {Tumor, Nontumor} supervised learning problems.

Markers for decision support systems. Table 2 shows that the maximum generalization performance achieved is less than the maximum possible for both the {ALL, AML} (71 vs. 72) and {Tumor, Nontumor} (55 vs. 62) problems. This may be because both data sets contain outliers and potentially mislabeled samples. For the five {ALL, AML} experiments with a performance of 71, the single error is a false positive. Previously, 6/62 adenocarcinoma samples were assigned as false positive or false negative across the 17 gene subsets examined (16). Subsets with the same performance may differ in their false positive and false negative assignments. Decreasing the number of ranked genes below the top 11 degrades performance. Overall, the NBGR and MVR rankings are effective because the top 50 perform better than the middle 50 and significantly better than the bottom 50. Some subsets generalize as well as or better than the full repertoire of 1,988 or 7,070 genes. Thus the top 25–100 genes of each expert are potential markers for use in developing decision support systems aimed at distinguishing tumor from nontumor colon adenocarcinoma samples and ALL from AML samples.

SVM kernel function and kernel parameters. No kernel function or parameter setting is optimal in terms of generalization performance. For example, a data-dependent width parameter γ_d gives superior results compared to the data-independent parameter γ_d for the {ALL, AML} problem. The reverse is true for the {Tumor, Nontumor} problem. The poorer performance of a data-dependent width parameter for the {Tumor, Nontumor} problem may be due to the larger number of potentially misclassified examples in the adenocarcinoma versus the leukemia data set. In previous analysis of the adenocarcinoma data (16), training exam-

Table 1. *Naive Bayes model class assigned to leukemia samples by a model trained using 72 unlabeled 7,070-feature sample profile vectors*

Naive Bayes Class	Sample Number	Primary Label {ALL, AML}	Secondary Label		
			{PB, BM}	{T cell, B cell}	{Male, Female}
1	3	ALL	BM	T cell	M
1	1	ALL	BM	B cell	M
1	17	ALL	BM	B cell	M
1	49	ALL	BM	B cell	M
1	7	ALL	BM	B cell	F
1	8	ALL	BM	B cell	F
1	27	ALL	BM	B cell	F
1	39	ALL	BM	B cell	F
1	40	ALL	BM	B cell	F
1	56	ALL	BM	B cell	F
1	4	ALL	BM	B cell	
1	62	AML	PB		
1	63	AML	PB		
1	64	AML	PB		
1	54	AML	BM		
1	28	AML	BM		
1	30	AML	BM		
1	31	AML	BM		
1	32	AML	BM		
1	33	AML	BM		
1	34	AML	BM		
1	35	AML	BM		
1	36	AML	BM		
1	37	AML	BM		
1	38	AML	BM		
1	50	AML	BM		
1	51	AML	BM		
1	53	AML	BM		
1	58	AML	BM		
1	61	AML	BM		
2	67	ALL	PB	T cell	M
2	70	ALL	PB	B cell	F
2	71	ALL	PB	B cell	
2	6	ALL	BM	T cell	M
2	10	ALL	BM	T cell	M
2	23	ALL	BM	T cell	M
2	22	ALL	BM	B cell	M
2	25	ALL	BM	B cell	M
2	45	ALL	BM	B cell	M
2	47	ALL	BM	B cell	M
2	12	ALL	BM	B cell	F
2	18	ALL	BM	B cell	F
2	26	ALL	BM	B cell	F
2	41	ALL	BM	B cell	F
2	43	ALL	BM	B cell	F
2	44	ALL	BM	B cell	F
2	46	ALL	BM	B cell	F
2	55	ALL	BM	B cell	F
2	59	ALL	BM	B cell	F
2	19	ALL	BM	B cell	
2	52	AML	PB		
2	60	AML	BM		M
2	57	AML	BM		F
2	29	AML	BM		
2	65	AML	BM		
2	66	AML	BM		
3	68	ALL	PB	B cell	M
3	69	ALL	PB	B cell	M
3	72	ALL	PB	B cell	
3	2	ALL	BM	T cell	M
3	9	ALL	BM	T cell	M
3	11	ALL	BM	T cell	M
3	14	ALL	BM	T cell	M
3	16	ALL	BM	B cell	M
3	21	ALL	BM	B cell	M
3	24	ALL	BM	B cell	M
3	13	ALL	BM	B cell	F
3	15	ALL	BM	B cell	F
3	42	ALL	BM	B cell	F
3	48	ALL	BM	B cell	F
3	5	ALL	BM	B cell	
3	20	ALL	BM	B cell	

ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; PB, peripheral blood; BM, bone marrow.

Table 2. *Identifying marker genes using two different feature relevance experts*

	Leukemia {ALL, AML}					
	NBGR			MVR		
		Radial Basis Function			Radial Basis Function	
Gene Subset	Dot Product	$\gamma_d = 50.8$, 9537.0	$\gamma_f = 0.01$	Dot Product	$\gamma_d = 51.4$, 25641.0	$\gamma_f = 0.01$
All 7,070	70 (2,0)	$\boxed{70}$ (1,1)	45 (25,2)	70 (2,0)	70 (1,1)	45 (25,2)
Top 100	$\boxed{71}$ (1,0)	$\boxed{70}$ (2,0)	68 (4,0)	$\boxed{71}$ (1,0)	70 (2,0)	$\boxed{68}$ (4,0)
Top 50	70 (2,0)	68 (4,0)	$\boxed{70}$ (2,0)	70 (2,0)	$\boxed{71}$ (1,0)	66 (6,0)
Top 25	$\boxed{71}$ (1,0)	$\boxed{70}$ (1,1)	68 (4,0)	70 (2,0)	$\boxed{71}$ (1,0)	65 (7,0)
Top 11	70 (2,0)	69 (2,1)	$\boxed{70}$ (2,0)	70 (2,0)	69 (1,2)	64 (7,1)
Top 5	60 (8,4)	63 (6,3)	53 (19,0)	69 (2,1)	63 (4,5)	58/69* (11,0)
Top 4	67 (3,2)	64 (5,3)	51/70* (19,0)	49 (2,21)	39 (19,14)	41 (7,24)
Middle 50	54 (14,4)	48 (18,16)	49 (22,1)	59 (10,3)	57 (8,7)	46 (25,1)
Bottom 50	47 (25,0)	39 (19,14)	45 (21,6)	43 (7,22)	36 (19,17)	40 (23,9)

	Colon Adenocarcinoma {Tumor, Nontumor}					
	NBGR			MVR		
		Radial Basis Function			Radial Basis Function	
Gene Subset	Dot Product	$\gamma_d = 3.0$, 56.1	$\gamma_f = 0.01$	Dot Product	$\gamma_d = 3.0$, 116.4	$\gamma_f = 0.01$
All 1,998	$\boxed{53}$ (5,4)	$\boxed{54}$ (4,4)	$\boxed{55}$ (3,4)	$\boxed{53}$ (5,4)	$\boxed{54}$ (4,4)	$\boxed{55}$ (3,4)
Top 100	46 (7,9)	$\boxed{54}$ (4,4)	53 (6,3)	51 (4,7)	52 (6,4)	$\boxed{55}$ (3,4)
Top 50	49 (8,5)	53 (6,3)	$\boxed{55}$ (3,4)	48 (6,8)	52 (6,4)	$\boxed{55}$ (3,4)
Top 25	47 (6,9)	53 (6,3)	53 (6,3)	46 (8,8)	50 (6,6)	52 (3,7)
Top 11	50 (5,7)	52 (6,4)	54 (3,5)	49 (5,8)	50 (7,5)	53 (3,6)
Top 5	45 (9,8)	51 (8,3)	42 (9,11)	$\boxed{53}$ (3,6)	49 (7,6)	52 (4,6)
Top 4	38 (12,12)	42 (13,7)	40 (12,10)	52 (4,6)	50 (5,7)	54 (3,5)
Middle 50	39 (10,13)	43 (11,8)	45 (7,10)	51 (4,7)	48 (9,5)	50 (5,7)
Bottom 50	37 (12,13)	34 (12,16)	39 (11,12)	40 (11,11)	33 (16,13)	33 (16,13)

The full repertoire of genes assayed in the leukemia (7,070) and colon adenocarcinoma (1,988) profiling studies were ranked separately using the NBGR and MVR measures. Each entry gives the generalization performance of leave-one-out SVMs trained using a specific gene subset (All 7,070, Top 100,. . .), kernel function (dot product, radial basis function), and radial basis function width parameter γ ($\gamma_d =$ *minimum, maximum,* minimum and maximum values across the subsets; γ_f, fixed value across all subsets). The triplet of numbers indicates "true positive + true negative (false positive, false negative)" assignments. The maximum generalization performance possible for the leukemia profiling study was 72; the maximum possible score for colon adenocarcinoma study was 62. The maximum generalization performance achieved in a column is boxed. *Experiments in which only some of the leave-one-out partitioning of the training examples resulted in estimation sets capable of yielding models (the maximum generalization performance possible decreases from 72 to 69 and 70). NBGR, naive Bayes global relevance; MVR, median vote relevance; SVM, support vector machine.

ples that constituted support vectors in each of the 62 leave-one-out SVMs were used to pinpoint potentially mislabeled samples (support vectors are training examples that define the location of the decision surface). Similarly, it may be instructive to examine how the nature and number of such invariant support vector training examples vary according to feature subset, kernel function, and kernel parameters.

SVM training and evaluation. Table 3 indicates that performance is influenced by how the training examples are partitioned (compare the false positives and false negatives in the "38 estimation, 34 test" and "34

Table 3. *The generalization performance of different partitionings of the {ALL, AML} training examples*

	NBGR Top 50			MVR Top 50			MAR 50		
		Radial Basis Function			Radial Basis Function			Radial Basis Function	
Partitioning of Training Set	Dot Product	$\gamma_d = 39.8$, 102.3	$\gamma_f = 0.01$	Dot Product	$\gamma_d = 41.5$, 100.0	$\gamma_f = 0.01$	Dot Product	$\gamma_d = 40.2$, 80.5	$\gamma_f = 0.01$
72 leave-one-out	70 (2,0)	68 (4,0)	66 (6,0)	70 (2,0)	71 (1,0)	70 (2,0)	68 (3,1)	68 (2,2)	70 (2,0)
38 estimation, 34 test	32 (2,0)	32 (2,0)	30 (4,0)	33 (1,0)	33 (1,0)	33 (1,0)	33 (1,0)	33 (1,0)	32 (1,1)
34 estimation, 38 test	37 (1,0)	37 (1,0)	34 (4,0)	36 (2,0)	37 (1,0)	36 (2,0)	37 (1,0)	37 (1,0)	35 (3,0)

Each triplet of numbers refers to "true positive + true negative (false positive, false negative)" assignments. For the "72 leave-one-out" partitioning, they are assignments made by 72 leave-one-out SVMs for their single test example. For the "38 estimation, 34 test" partitioning, they are assignments made by a single SVM for 34 test examples. For the "34 estimation, 38 test" partitioning, they are assignments made by a single SVM for 38 test examples.

Table 4. *The leukemia NBGR top 50 genes*

	Gene ID	Gene Annotation
1	U14394_at	METALLOPROTEINASE INHIBITOR 3 PRECURSOR
2	L04947_at	KDR Kinase insert domain receptor (a type III receptor tyrosine kinase)
3	M11353_at	EEF1G Translation elongation factor 1 gamma
4	X56468_at	14-3-3 PROTEIN TAU
5	**M27891_at**	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)
6	**M84526_at**	DF D component of complement (adipsin)
7	M29581_at	ZNF8 Zinc finger protein 8 (clone HF.18)
8	U75679_at	Histone stem-loop binding protein (SLBP) mRNA
9	X95190_at	Branched chain Acyl-CoA Oxidase
10	Y08612_at	RABAPTIN-5 protein
11	M28130_rna1_s_at	Interleukin 8 (IL8) gene
12	L07540_at	ACTIVATOR 1 36 KD SUBUNIT
13	M21624_at	TCRD T-cell receptor, delta
14	M26683_at	SCYA2 Small inducible cytokine A2 (monocyte chemotactic protein 1, homologous to mouse Sig-je)
15	D13666_s_at	Osteoblast specific factor 2 (OSF-2os)
16	**M31166_at**	PTX3 Pentaxin-related gene, rapidly induced by IL-1 beta
17	**D88422_at**	CYSTATIN A
18	M57731_s_at	GRO2 GRO2 oncogene
19	M20203_s_at	GB DEF = Neutrophil elastase gene, exon 5
20	HG4316-HT4586_at	Transketolase-Like Protein
21	J05412_at	REG1A Regenerating islet-derived 1 alpha (pancreatic stone protein, pancreatic thread protein)
22	M54995_at	PPBP Connective tissue activation peptide III
23	M27783_s_at	ELA2 Elastatse 2, neutrophil
24	U27831_at	GB DEF = Striatum-enriched phosphatase (STEP)
25	X82103_at	Beta-COP
26	X55668_at	PRTN3 Proteinase 3 (serine proteinase, neutrophil, Wegener granulomatosis autoantigen)
27	HG2887-HT3031_at	Sry-Related Hmg-Box 12 Protein (Gb:X73039)
28	U39576_at	BTN Butyrophilin
29	HG2981-HT3127_s_at	MAP kinase kinase 6 (MKK6) mRNA
30	X54667_s_at	CST4 Cystatin S
31	**J04990_at**	CATHEPSIN G PRECURSOR
32	**M96326_rna1_at**	Azurocidin gene
33	X65977_at	DEFA4 Defensin, alpha 4, corticostatin
34	U80987_s_at	GB DEF = Transcription factor TBX5 mRNA
35	**M63379_at**	CLU Clusterin (complement lysis inhibitor; testosterone-repressed prostate message 2; apolipoprotein J)
36	X04602_s_at	IL6 Interleukin 6 (B cell stimulatory factor 2)
37	X06182_s_at	KIT V-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
38	X97198_at	Receptor protein tyrosine phosphatase hPTP-J precursor
39	**M23178_s_at**	MACROPHAGE INFLAMMATORY PROTEIN 1-ALPHA PRECURSOR
40	X79981_at	CDH5 Cadherin 5, VE-cadherin (vascular epithelium)
41	U52112_rna5_at	RbP gene (renin-binding protein) extracted from Human Xq28 genomic DNA in the region of the L1CAM locus
42	X52882_at	T-COMPLEX PROTEIN 1, ALPHA SUBUNIT
43	L12392_at	HD Huntingtin (Huntington disease)
44	Y09616_at	Carboxylesterase (hCE-2) mRNA
45	X13238_at	COX6C Cytochrome *c* oxidase subunit VIc
46	M30703_s_at	Amphiregulin (AR) gene
47	U60521_at	Cysteine protease ICE-LAP6 mRNA
48	HG3454-HT3647_at	Zinc Finger Protein 20
49	X65962_s_at	CYP2C17 Cytochrome P450, subfamily IIC (mephenytoin 4-hydroxylase), polypeptide 17
50	S72008_at	CDC10 Cell division cycle 10 (homologous to CDC10 of *S. cerevisiae*)

The "Gene Annotations" in Tables 4–6 and 10–12 are text strings reproduced verbatim from the source, to facilitate searches and ease of use for the readership when extracting information from the array data. Genes in boldface are present in the {ALL, AML} MVR top 50 shown in Table 5.

estimation, 38 test" experiments). The MAR 50 subset and "38 estimation, 34 test" partitioning allow a direct comparison between the performance of SVMs and the published weighted vote predictor (11). In the latter, the estimation set was used to compute the MAR for each feature in the subset. This 50-feature predictor assigned the label for each of the 34 test examples as follows. Each gene F_l casts a weighted vote according to whether the expression level x_l is closer to the value of the gene in class $i \equiv$ ALL or $j \equiv$ AML of the estimation set, $v(F_l) = \mathrm{MAR}(F_l)(x_l - [\mu_{i,l} + \mu_{j,l}]/2)$. If the sum of the absolute values of the positive votes in the 50 genes is greater than the sum of the absolute values of the negative votes, then the test example is assigned to the positive class i. The weighted vote predictor made strong predictions for 29 of the 34 test examples, and in all instances, the assignments were true positives or true negatives. In contrast, an SVM makes true positive or true negative assignments for 33 of the 34 test examples.

ALL and AML markers for experimental studies. The original leukemia study provided biological explanations as to why members of the MAR 50 might be involved in this disorder and could distinguish AML from ALL (11). The results here indicate that the NBGR top 50, MVR top 50, and MAR 50 generalize as well as all 7,070 genes (compare the "72 leave-one-out" entry in Table 3 with the "All 7,070" entry in Table 2). Tables 4–6 list the top 50 genes of each expert. Although the precise composition of the top 50s differ, each set of 50 genes is effective in terms of discriminating between AML and ALL. The small overlap in terms of the specific genes suggests the presence of many gene subsets of a given cardinality that can generalize equally well.

Only adipsin, azurocidin, and cystatin C are common to the NBGR top 50, MVR top 50, and MAR 50. Given the large number of genes assayed (7,070) and the extensive literature on leukemia, it should be possible to provide biologically based rationales as to why three particular genes might be involved in AML/ALL, even those chosen at random and having no actual role in the disease. Although such an explanation cannot be ruled out for adipsin, azurocidin, and cystatin C, circumstantial evidence suggests that they may, indeed, be robust and reliable markers and thus good candidates for additional experimental investigation. These genes are ranked highly by three independent experts

Table 5. *The {ALL, AML} MVR top 50 genes*

Gene ID	Gene Annotation
1 X95735_at	Zyxin
2 X62320_at	GRN Granulin
3 D88422_at	CYSTATIN A
4 M23197_at	CD33 CD33 antigen (differentiation antigen)
5 M83652_s_at	PFC Properdin P factor, complement
6 M84526_at	DF D component of complement (adipsin)
7 U46499_at	GLUTATHIONE S-TRANSFERASE, MICROSOMAL
8 L09209_s_at	APLP2 Amyloid beta (A4) precursor-like protein 2
9 M63138_at	CTSD Cathepsin D (lysosomal aspartyl protease)
10 M92287_at	CCND3 Cyclin D3
11 M31523_at	TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)
12 M83667_rna1_s_at	NF-IL6-beta protein mRNA
13 M16038_at	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog
14 M31211_s_at	MYL1 Myosin light chain (alkali)
15 X62654_rna1_at	ME491 gene extracted from *H. sapiens* gene for Me491/CD63 antigen
16 X85116_rna1_s_at	Epb72 gene exon 1
17 M19507_at	MPO Myeloperoxidase
18 M63379_at	CLU Clusterin (complement lysis inhibitor; testosterone-repressed prostate message 2; apolipoprotein J)
19 M96326_rna1_at	Azurocidin gene
20 U50136_rna1_at	Leukotriene C4 synthase (LTC4S) gene
21 M32304_s_at	TIMP2 Tissue inhibitor of metalloproteinase 2
22 M55150_at	FAH Fumarylacetoacetate
23 D14664_at	KIAA0022 gene
24 J05243_at	SPTAN1 Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)
25 M31303_rna1_at	Oncoprotein 18 (Op18) gene
26 M11722_at	Terminal transferase mRNA
27 L47738_at	Inducible protein mRNA
28 M98399_s_at	CD36 CD36 antigen (collagen type I receptor, thrombospondin receptor)
29 X17042_at	PRG1 Proteoglycan 1, secretory granule
30 X90858_at	Uridine phosphorylase
31 M27891_at	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)
32 M23178_s_at	MACROPHAGE INFLAMMATORY PROTEIN 1-ALPHA PRECURSOR
33 U05572_s_at	MANB Mannosidase alpha-B (lysosomal)
34 HG3494-HT3688_at	Nuclear Factor Nf-Il6
35 M31166_at	PTX3 Pentaxin-related gene, rapidly induced by IL-1 beta
36 M19508_xpt3_s_at	MPO from Human myeloperoxidase gene, exons 1–4./ntype = DNA/annot = exon
37 M93056_at	LEUKOCYTE ELASTASE INHIBITOR
38 X59417_at	PROTEASOME IOTA CHAIN
39 Z15115_at	TOP2B Topoisomerase (DNA) II beta (180kD)
40 X98411_at	GB DEF = Myosin-IE
41 J04990_at	CATHEPSIN G PRECURSOR
42 HG1612-HT1612_at	Macmarcks
43 U05259_rna1_at	MB-1 gene
44 M22960_at	PPGB Protective protein for beta-galactosidase (galactosialidosis)
45 X16546_at	RNS2 Ribonuclease 2 (eosinophil-derived neurotoxin; EDN)
46 X07743_at	PLECKSTRIN
47 X70297_at	CHRNA7 Cholinergic receptor, nicotinic, alpha polypeptide 7
48 U70063_at	Acid ceramidase mRNA
49 M62762_at	ATP6C Vacuolar H+ ATPase proton channel subunit
50 U02020_at	Pre-B cell enhancing factor (PBEF) mRNA

The "Gene Annotations" in Tables 4–6 and 10–12 are text strings reproduced verbatim from the source, to facilitate searches and ease of use for the readership when extracting information from the array data. Genes in boldface are present in the NBGR top 50 shown in Table 4.

Table 6. *The {ALL, AML} MAR 50 genes from figure 3A of Ref. 11*

Gene ID	Gene Annotation
U22376	c-myb
X59417*	Proteasome iota
U05259*	MB-1
M92287*	Cyclin D3
M31211*	Myosin light chain
X74262	RbAp48
D26156	SNF2
S50223	HkrT-1
M31523	E2A
L47738*	Inducible protein
U32944	Dynein light chain
Z15115*	Topoisomerase II β
X15949	IRF2
X63469	TFIIEβ
M91432	Acyl-Coenzyme A dehydrogenase
U29175	SNF2
Z69881	Ca^{2+}-ATPase
U20998	SRP9
D38073	MCM3
U26266	Deoxyhypasine synthase
M31303*	Op 18
Y08612†	Rabaptin-5
U35451	Heterochromatin protein p25
M29696	IL-7 receptor
M13792	Adenosine deaminase
M55150*	Fumarylacetoacetate
X95735*	Zyxin
U50136*	LTC4 synthase
M16038*	LYN
U82759	HoxA9
M23197*	CD33
M84526*†	Adipsin
Y12670	Leptin receptor
M27891*†	Cystatin C
X17042*	Proteoglycan I
Y00787	IL-8 precursor
M96326*†	Azurocidin
U46751	p62
M80254	CyP3
L08246	MCL1
M62762*	ATPase
M28130†	IL-8
M63138*	Cathepsin D
M57710	Lectin
M69043	MAD-3
M81695	CDC11c
X85116*	Ebp72
M19045	Lysozyme
M83652*	Properdin
X04085	Catalase

MAR, mean aggregate relevance. *Genes common to MVR top 50 (Table 5). †Genes common to NBGR top 50 (Table 4). Genes in boldface are common to the MAR 50, NBGR top 50, and MVR top 50. The "Gene Annotations" in Tables 4–6 and 10–12 are text strings reproduced verbatim from the source, to facilitate searches and ease of use for the readership when extracting information from the array data.

and are located in chromosomal regions known to be sites of recurrent abnormalities in ALL and AML (Table 7). Chromatin reorganization of the 19p13.3 locus, which contains azurocidin, proteinase-3, neutrophil elastase, and adipsin, is associated with myeloid cell differentiation (24). The generalization performance achieved by these subsets of 3 (66) and 4 (64) genes is comparable to the NBGR top 4 (67) and higher than the MVR top 4 (49) (Table 8).

Cystatins C, A (GenBank accession no. D88422), and S (X54667) and cathepsins G (J04990) and D (M63138) are common to two out of the three top 50s. Cystatins are endogenous protein inhibitors of cathepsins, so these specific protease-inhibitor pains might be important in the etiology of ALL and AML. Human neutrophil-derived cathepsin G and azurocidin have been identified as chemoattractants for mononuclear cells and neutrophils (6). Experimental investigation of highly ranked genes may be warranted.

T cell/B cell, PB/BM, and male/female markers for experimental studies. The MVR expert defines 25 markers for each of the additional leukemia problems that generalize as well as all 7,070 genes (Table 9). Comparing the maximum performance achieved and the maximum possible performance indicates that the data contain sufficient information for the {PB, BM} (68 vs. 72) and {T cell, B cell} (46 vs. 47) problems, but not for the {Male, Female} (31 vs. 49) problem. Furthermore, there is little difference in performance between the {Male, Female} top 50, middle 50, and bottom 50 gene sets. This suggests little association between these sample labels and transcription profiling data. Possible explanations for the poorer {Male, Female} results include *1*) transcription profile data are poor indicators of sex, *2*) the 7,070 probe set did not include those that can distinguish males from females, and *3*) the patients (mostly children) had not achieved

Table 7. *Abnormalities associated with two chromosomal regions containing genes at the intersection of the NBGR top 50, MVR top 50, and MAR 50: azurocidin (Gene ID M96326), adipsin (M84526), and cystatin C (M27891)*

Region	Abnormality	Neoplasm	Total Cases
19p13.3	t(1;19)(q21;p13)	ALL	5
	t(1;19)(q22;p13)	ALL	2
	t(1;19)(q23;p13)	ALL	108
	t(11;19)(q23;p13)	ALL	67
	t(11;19)(q23;p13)	AML	76
	t(17;19)(q21;p13)	ALL	3
	t(17;19)(q22;p13)	ALL	5
	del(19)(p13)	AML	2
	der(19)t(1;19)(q11;p13)	AML	2
	der(19)t(1;19)(q21;p13)	ALL	8
	der(19)t(1;19)(q23;p13)	ALL	172
20p11.2	t(14;20)(q11;p11)	ALL	2
	del(20)(p11)	ALL	3
	del(20)(p11)	AML	2

The data are derived from the Breakpoint Map of Recurrent Chromosome Aberrations (**http://www.ncbi.nlm.nih.gov/CCAP**). The 19p13.3 region contains the four closely linked genes 5′ azurocidin-proteinase 3-neutrophil elastase-adipsin 3′ (24). The 20p11.2 region contains cystatin C (M27891) and cystatin S (X54667). Proteinase 3 (X55668), neutrophil elastase (M27783), and cystatin S are in the NGBR top 50. Other NBGR and MVR top-ranked genes located in sites of recurrent abnormalities include tissue inhibitor of metalloproteinase 3 (U14394; 22q12.3) and zyxin (X95735, 7q32), respectively.

Table 8. *The generalization performance of gene subsets that are good candidates for further experimental studies of ALL and AML*

Gene Subset	Dot Product	Radial Basis Function	
		γ_d	$\gamma_f = 0.01$
Intersection of NBGR, MVR, and MAR 50s M27891 Cystatin C (CST3) M84526 Adipsin M96326 Azurocidin	60/70* (10, 0)	66 (5, 1)	28/69* (18, 23)
Linked genes in 19p13.3 locus M96326 Azurocidin X55668 Proteinase 3 (PRTN3) M20203 Neutrophil elastase M84526 Adipsin	60/71* (11, 0)	64 (6, 2)	46/58* (10, 2)
NBGR Top 4 U14394 Metalloproteinase inhibitor 3 L04947 KDR Kinase insert domain receptor M11353 EEF1G Translation elongation factor 1 γ X56468 14-3-3 protein tau	67 (3, 2)	64 (5, 3)	51/70* (19, 0)
MVR Top 4 X95735 Zyxin X62320 Granulin (GRN) D88422 Cystatin A M23197 CD33 antigen	49 (2, 21)	39 (19, 14)	41 (7, 24)

"Intersection of NBGR, MVR, and MAR 50s" refers to genes common to the NBGR top 50, MVR top 50, and MAR 50. "Linked genes in 19p13.3 locus" denotes the four closely linked genes found in the NBGR top 50. "NBGR Top 4" and "MVR Top 4" are the top four genes listed in Tables 4 and 5, respectively. *Experiments in which only some of the leave-one-out partitioning of the training examples resulted in estimation sets capable of yielding models.

sexual maturity and thus not manifested any differences.

Of the 72 training examples, 47 have {T cell, B cell} labels, and there is only one false positive assignment when either all 7,070 or the top 25 genes are used. It is interesting to note that a dot product SVM trained using these 47 labeled 7,070-feature experiment profile vectors assigned a B cell label to each of the 72 − 47 = 25 test examples. These test examples are the AML samples listed in Table 1.

The three sets of MVR rankings appear to be biologically interesting (Tables 10–12). It should be noted, however, that they are valid only within the context of tissue samples derived from patients with ALL/AML. Bearing this in mind, the {T cell, B cell} top 50 contains many known T cell related genes. Genes that have no obvious annotation linking them to this cell type, such as protein disulfide isomerase, selenoprotein W, and Ras-related protein Rab-32, may be novel markers that can discriminate between T cells and B cells. Selenoprotein W is an intracellular protein that may be involved in protection against oxidative damage and muscle metabolism (4, 14). Overexpression of *Lrp*, the top ranked {PB, BM} gene, often predicts a poor response to chemotherapy in leukemia because it is one of the mechanisms by which cancer cells develop resistance to cytotoxic agents (reviewed in Ref. 19).

DISCUSSION

The principal requirement for identifying marker genes for use in developing a clinically relevant decision support system for cancer diagnosis, prognosis, and monitoring is that the resultant system generate accurate predictions. The generalization capacity of the system is of paramount importance since the number and diversity of samples available for its development are likely to be far smaller than samples for which predictions will need to be made. Undoubtedly, a variety of the extracellular and intracellular pathways that regulate and maintain interactions between cells and their microenvironment are perturbed during carcinogenesis. Hence, feature relevance experts should be designed that implement as fundamentally different notions of relevance as possible in order that each relevance measure captures a different, physiologically relevant pathway or mechanism leading to the biological end point. Given a mixture of experts, selecting gene subsets that are ranked highly by each expert and which generalize as well as or better than the full repertoire should help to pinpoint robust marker genes. Based on the results here, a prototype system for discriminating between ALL and AML samples could contain the 125 features that are the union of NBGR top 50, MVR top 50, and MAR 50.

Table 9. *Marker genes that distinguish leukemia samples according to their {PB, BM}, {T cell, B cell}, and {Male, Female} labels and identified using the MVR expert*

	{PB, BM} (72)		{T cell, B cell} (47)		{Male, Female} (49)	
		Radial Basis Function		Radial Basis Function		Radial Basis Function
Gene Subset	Dot product	$\gamma_d = 459.8, 23696.7$	Dot product	$\gamma_d = 18.5, 20833.3$	Dot product	$\gamma_d = 8.4, 12.9$
All 7,070	63 (6, 3)	64 (8, 0)	45 (1, 1)	$\boxed{46}$ (1, 0)	$\boxed{30}$ (15, 4)	$\boxed{31}$ (9, 9)
Top 100	$\boxed{67}$ (2, 3)	$\boxed{68}$ (3, 1)	$\boxed{46}$ (1, 0)	$\boxed{46}$ (1, 0)	26 (11, 12)	29 (10, 10)
Top 50	65 (3, 4)	67 (4, 1)	$\boxed{46}$ (1, 0)	$\boxed{46}$ (1, 0)	27 (12, 10)	29 (9, 11)
Top 25	64 (3, 5)	64 (5, 3)	$\boxed{46}$ (1, 0)	$\boxed{46}$ (1, 0)	29 (11, 9)	30 (11, 8)
Middle 50	56 (6, 10)	55 (10, 7)	31 (4, 12)	37 (7, 3)	28 (11, 10)	27 (12, 10)
Bottom 50	20 (5, 47)	57 (10, 5)	11 (3, 33)	31 (9, 7)	28 (12, 9)	29 (10, 10)

The maximum possible generalization performance is given in parenthesis. See legend to Table 2 for complete description and definitions.

Table 10. *The {PB, BM} MVR top 50 genes*

	Gene ID	Gene Annotation
1	X79882_at	Lrp mRNA
2	X57206_at	ITPKB Inositol 1,4,5-trisphosphate 3-kinase B
3	M37766_at	CD48 CD48 antigen (B-cell membrane protein)
4	L36818_at	INPPL1 Inositol polyphosphate phosphatase-like protein 1 (51C protein)
5	U76764_s_at	CD97 CD97 antigen (leucocyte antigen)
6	M60922_at	Surface antigen mRNA
7	D86976_at	KIAA0223 gene, partial cds
8	AF006084_at	Arp2/3 protein complex subunit p41-Arc (ARC41) mRNA
9	L32976_at	Protein kinase (MLK-3) mRNA
10	S73591_at	Brain-expressed HHCPA78 homolog [human, HL-60 acute promyelocytic leukemia cells]
11	D00591_at	CHC1 Chromosome condensation 1
12	X07767_at	PRKACA Protein kinase, cAMP-dependent, catalytic, alpha
13	D50923_at	KIAA0133 gene
14	D38305_at	Tob
15	U49187_at	Placenta (Diff48) mRNA
16	X90780_rna1_at	Cardiac troponin I gene, exons 1 to 5
17	D83735_at	Adult heart mRNA for neutral calponin
18	M72885_rna1_s_at	G0S2 gene extracted from Human GOS2 gene, 5′ flank and cds
19	D14657_at	KIAA0101 gene
20	X01703_at	Alpha-tubulin mRNA
21	M60830_at	EVI2B PROTEIN PRECURSOR TROPIC VIRAL INTEGRATION SITE 2B PROTEIN
22	U52101_at	YMP mRNA
23	U15085_at	HLA-DMB Major histocompatibility complex, class II, DM beta
24	X75962_at	OX40L RECEPTOR PRECURSOR
25	M87339_at	RFC4 Replication factor C, 37-kD subunit
26	J03600_at	ALOX5 Arachidonate 5-lipoxygenase
27	U93049_at	GB DEF = SLP-76 associated protein mRNA
28	U03851_at	Capping protein alpha mRNA, partial cds
29	HG4557-HT4962_r_at	Small nuclear ribonucleoprotein U1, 1snrp
30	U66464_at	Hematopoietic progenitor kinase (HPK1) mRNA
31	L36983_at	Dynamin (DNM) mRNA
32	U01038_at	PLK mRNA
33	J00220_cds5_at	IGHA1 gene extracted from Human Ig germline H-chain G-E-A region A: gamma-3 5′ flank
34	D25538_at	KIAA0037 gene
35	U20158_at	76 kDa tyrosine phosphoprotein SLP-76 mRNA
36	D63482_at	KIAA0148 gene
37	X59405_at	MCP Membrane cofactor protein (CD46, trophoblast-lymphocyte cross-reactive antigen)
38	U00921_at	LST1 mRNA, cLST1/E splice variant
39	X04106_at	CAPN4 Calpain, small polypeptide
40	U56418_at	Lysophosphatidic acid acyltransferase-beta mRNA
41	X78121_at	CHM Choroideremia
42	X61587_at	ARHG Ras homolog gene family, member G (rho G)
43	U80073_at	GB DEF = Tip associating protein (TAP) mRNA
44	U37022_rna1_at	Cyclin-dependent kinase 4 (CDK4) gene
45	U49278_at	Putative DNA-binding protein mRNA, partial cds
46	X63131_s_at	PML Probable transcription factor PML alternative products
47	X62048_at	WEE1-LIKE PROTEIN KINASE
48	U46751_at	Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA
49	U53204_at	Plectin (PLEC1) mRNA
50	U03105_at	B4-2 protein mRNA

The “Gene Annotations” in Tables 4–6 and 10–12 are text strings reproduced verbatim from the source, to facilitate searches and ease of use for the readership when extracting information from the array data.

Although reducing the original 7,070 leukemia genes to 125 is appropriate in terms of a decision support system, this is still too many for in-depth experimental studies. Hence, the most informative experimental markers may be genes at the intersection of the top ranked genes: adipsin, azurocidin, and cystatin C. However, they are unlikely to be the sole determinants of the difference between ALL and AML because the generalization performance of these three genes is poorer than some of the larger gene subsets. The same is true for the four closely linked genes on chromosome 19p13.3 (azurocidin-proteinase 3-neutrophil elastase-adipsin). Nonetheless, the strategy proposed here provides a protocol for pinpointing experimentally informative marker genes and thus prioritizing subsequent investigations.

In transcription profiling studies, more genes are monitored than are probably required to understand the main problem. This “overdetermined” property suggests that broader questions could be answered if additional information were available for each sample. For the leukemia {T cell, B cell} and {PB, BM} secondary problems, the 7,070 genes are sufficiently informative that 25 markers can be defined that generalize as well as all 7,070 genes. It remains to be determined whether these makers are universal or

Table 11. *The {T cell, B cell} MVR top 50 genes*

Gene ID	Gene Annotation
1 X03934_at	GB DEF = T-cell antigen receptor gene T3-delta
2 D00749_s_at	T-CELL ANTIGEN CD7 PRECURSOR
3 X00274_at	HLA CLASS II HISTOCOMPATIBILITY ANTIGEN, DR ALPHA CHAIN PRECURSOR
4 X04145_at	CD3G CD3G antigen, gamma polypeptide (TiT3 complex)
5 U23852_s_at	GB DEF = T-lymphocyte specific protein tyrosine kinase p56lck (lck) abberant mRNA
6 M23323_s_at	T-CELL SURFACE GLYCOPROTEIN CD3 EPSILON CHAIN PRECURSOR
7 X76223_s_at	GB DEF = MAL gene exon 4
8 M13560_s_at	PROBABLE PROTEIN DISULFIDE ISOMERASE ER-60 PRECURSOR
9 X00437_s_at	TCRB T-cell receptor, beta cluster
10 X59871_at	TCF7 Transcription factor 7 (T-cell specific)
11 X69398_at	CD47 CD47 antigen (Rh-related antigen, integrin-associated signal transducer)
12 M37271_s_at	T-CELL ANTIGEN CD7 PRECURSOR
13 U59878_at	Low-Mr GTP-binding protein (RAB32) mRNA, partial cds
14 L40386_s_at	DP2 (Humdp2) mRNA
15 U67171_at	GB DEF = Selenoprotein W (selW) mRNA
16 M26692_s_at	GB DEF = Lymphocyte-specific protein tyrosine kinase (LCK) gene, exon 1, and downstream promoter region
17 HG4128-HT4398_at	Anion Exchanger 3, Cardiac Isoform
18 M37815_cds1_at	CD28 gene (glycoprotein CD28) extracted from Human T-cell membrane glycoprotein CD28 mRNA
19 U14603_at	Protein tyrosine phosphatase PTPCAAX2 (hPTPCAAX2) mRNA
20 U18009_at	Chromosome 17q21 mRNA clone LF113
21 D11327_s_at	PTPN7 Protein tyrosine phosphatase, non-receptor type 7
22 X87241_at	HFat protein
23 U50743_at	Na,K-ATPase gamma subunit mRNA
24 D87292_at	Rhodanese
25 L05148_at	Protein tyrosine kinase related mRNA sequence
26 U18422_at	DP2 (Humdp2) mRNA
27 U49835_s_at	CHIT1 Chitinase 1
28 M28826_at	CD1B CD1b antigen (thymocyte antigen)
29 X14975_at	GB DEF = CD1 R2 gene for MHC-related antigen
30 U50327_s_at	Protein kinase C substrate 80K-H gene (PRKCSH)
31 X98172_at	MACH-alpha-2 protein
32 X67235_s_at	PRHX Proline-rich homeodomain-containing transcription factor (symbol provisional)
33 U16954_at	(AF1q) mRNA
34 M12886_at	TCRB T-cell receptor, beta cluster
35 S78187_at	M-PHASE INDUCER PHOSPHATASE 2
36 M16336_s_at	CD2 CD2 antigen (p50), sheep red blood cell receptor
37 X60992_at	T-CELL DIFFERENTIATION ANTIGEN CD6 PRECURSOR
38 J03077_s_at	PSAP Sulfated glycoprotein 1
39 S65738_at	Actin depolymerizing factor [human, fetal brain, mRNA, 1452 nt]
40 D38549_at	KIAA0068 gene, partial cds
41 X69433_at	IDH2 Isocitrate dehydrogenase 2 (NADP+), mitochondrial
42 X58072_at	GATA3 GATA-binding protein 3
43 D83920_at	FCN1 Ficolin (collagen/fibrinogen domain-containing) 1
44 X68742_at	GB DEF = Integrin, alpha subunit
45 HG3576-HT3779_f_at	Major Histocompatibility Complex, Class Ii Beta W52
46 U64675_at	SRI Sorcin
47 D30758_at	KIAA0050 gene
48 L08895_at	MEF2C MADS box transcription enhancer factor 2, polypeptide C (myocyte enhancer factor 2C)
49 X99584_at	SMT3A protein
50 D82345_at	NB thymosin beta

The "Gene Annotations" in Tables 4–6 and 10–12 are text strings reproduced verbatim from the source, to facilitate searches and ease of use for the readership when extracting information from the array data.

are restricted to samples originating from ALL and AML patients.

Both the leukemia and adenocarcinoma data sets contain potentially misclassified samples, samples for which the original label (the "gold standard") may be incorrect (1/72 and 6/62 respectively). In a previous study of the latter data set (16), the subset of training examples that constituted support vectors across the entire series of leave-one-out SVMs was suggested to be indicative of samples most likely to have been misclassified (the set of support vectors does appear to depend upon which training example is withheld when estimating an SVM). Misclassification may be due to simple human error during sample handling, RNA preparation, data acquisition, data analysis, and so on. Standardized protocols stipulating rigorous procedures at each step of the process should reduce this type of problem and improve the chances of creating a coherent data set. The possibility of misclassification cannot be eliminated entirely because although a sample might appear to be visually and/or histologically of one type, it might be a member of the other class in reality. By training SVMs with hard margins, assuming no a priori labeling errors, potentially mislabeled samples can be pinpointed and subjected to additional investigation to verify their label. Given the nature of the

Table 12. *The {Male, Female} MVR top 50 genes*

	Gene ID	Gene Annotation
1	L08246_at	INDUCED MYELOID LEUKEMIA CELL DIFFERENTIATION PROTEIN MCL1
2	D31887_at	KIAA0062 gene, partial cds
3	M21121_at	SCYA5 Small inducible cytokine A5 (RANTES)
4	U29656_at	NME1 Non-metastatic cells 1, protein (NM23A) expressed in
5	D64159_at	GB DEF = 3–7 gene product, partial cds
6	D14657_at	KIAA0101 gene
7	U37022_rna1_at	Cyclin-dependent kinase 4 (CDK4) gene
8	HG2788-HT2896_at	Calcyclin
9	HG417-HT417_s_at	Cathepsin B
10	M28213_s_at	RAB2 RAB2, member RAS oncogene family
11	J03925_at	ITGAM Integrin, alpha M (complement component receptor 3, alpha)
12	L38593_s_at	GB DEF = Integral membrane protein (NRAMP1) gene, exon 5
13	X01703_at	Alpha-tubulin mRNA
14	X03663_at	CSF1R Colony stimulating factor 1 receptor, formerly McDonough feline sarcoma viral (v-fms) oncogene homolog
15	J04182_at	LAMP1 Lysosome-associated membrane protein 1
16	Z50022_at	Surface glycoprotein
17	X98534_s_at	VASP gene, exons 4 to 13
18	U56402_s_at	Chromatin structural protein homolog (SUPT5H) mRNA
19	X79780_at	YPT3 mRNA
20	J04132_at	CD3Z CD3Z antigen, zeta polypeptide (TiT3 complex)
21	U90905_at	Clone 23574 mRNA sequence
22	D79990_at	KIAA0168 gene
23	U43077_at	CDC37 homolog mRNA
24	D87434_at	KIAA0247 gene
25	L13977_at	LYSOSOMAL PRO-X CARBOXYPEPTIDASE PRECURSOR
26	M33552_at	GB DEF = Lymphocyte-specific protein 1 (LSP1) mRNA
27	U61167_at	SH3 domain-containing protein SH3P18 mRNA
28	U26173_s_at	BZIP protein NF-IL3A (IL3BP1) mRNA
29	D90097_at	ALPHA-AMYLASE 2B PRECURSOR
30	U02680_at	Protein tyrosine kinase mRNA
31	M20543_at	ACTA1 Actin, alpha 1, skeletal muscle
32	Z11697_at	CD83 ANTIGEN PRECURSOR
33	U20816_s_at	GB DEF = Nuclear factor kappa-B2 (NF-KB2) gene, partial cds
34	U04810_at	DbpB-like protein mRNA
35	D14811_at	KIAA0110 gene
36	U79273_at	Clone 23933 mRNA sequence
37	M85276_at	NKG5 PROTEIN PRECURSOR
38	HG620-HT620_at	Tyrosine Phosphatase, Epsilon
39	X62534_s_at	HMG2 High-mobility group (nonhistone chromosomal) protein 2
40	X75756_at	PRKCM Protein kinase C, mu
41	X56841_at	HLA-E MHC class I antigen HLA-E
42	M22995_at	RAP1A RAP1A, member of RAS oncogene family
43	X78121_at	CHM Choroideremia
44	S81914_at	IEX-1
45	U51990_at	HPrp18 mRNA
46	M84371_rna1_s_at	CD19 gene
47	X65550_at	MKI67 Antigen identified by monoclonal antibody Ki-67
48	M29474_at	Recombination activating protein (RAG-1) gene
49	Z83741_at	GB DEF = HH2A/m gene
50	L13329_at	IDS Iduronate 2-sulfatase (Hunter syndrome)

The "Gene Annotations" in Tables 4–6 and 10–12 are text strings reproduced verbatim from the source, to facilitate searches and ease of use for the readership when extracting information from the array data.

underlying biology and technical issues surrounding generation of transcription profiling data, it is conceivable that many, it not all, cancer profiling experiments will contain noisy data and misclassified samples. Soft margin SVMs do take into consideration misclassified training examples but it is difficult to estimate the underlying error rate at the present time. To improve the reliability of downstream analyses, it may be preferable to incorporate a preprocessing step that identifies, and subsequently corrects if necessary, any misclassified samples. Once achieved, the distance of a sample to the optimal hyperplane can be used to assess confidence in an assignment.

The results from this and previous (16) work highlight a need for theoretical research in several areas. As illustrated here, the generalization performance of SVMs depends not only on the precise learning problem, but also on the training and testing procedure employed. Although leave-one-out cross-validation is costly and time-consuming, it provides a reasonable estimate of the expected generalization error. In view of uncertainties in the labels assigned to samples and the small, imbalanced sample set, a relatively simple assessment of the overall performance of SVMs was utilized: the cost function used to judge accuracy was the total number of true positive and true negative assignments. Principled, sophisticated methods need to be developed for areas such as *1*) selecting features in the presence of an unknown number of misclassified training examples, *2*) choosing the appropriate class of kernel function and determining (near) optimal kernel parameters automatically, *3*) training and evaluating a learning system that is both computationally efficient and yields biologically meaningful results, and *4*) generating an integrated prediction from a set of feature relevance experts that vary in how well they perform on the classification and prediction task at hand (boosting and bagging).

Despite the aforementioned limitations, utilizing a mixture of feature relevance experts that incorporate SVMs for supervised learning problems appears to be a promising method for identifying marker genes in cancer profiling studies. This approach can be applied directly to identifying markers in transcription profiling studies addressing other discrimination problems such as those encountered in aging and responses to different doses and dose rates of xenobiotic agents such as radiation. Similarly, the technique could be used to identify marker experiments as opposed to marker genes. These ideas can be extended to molecular profiling studies in which the features monitored are not genes, but are molecules such as proteins, metabolites, and so on.

This work was supported by the Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division, under US Department of Energy Contract No. DE-AC03-76SF0098.

REFERENCES

1. **Alizadeh AA, Eisen MB, David RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, and Staudt LM.** Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403: 503–511, 2000.
2. **Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, and Levine AJ.** Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA* 96: 6745–6750, 1999. [The data are available at http://microarray.princeton.edu/oncology/]
3. **Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Jr, and Haussler D.** Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA* 97: 262–267, 2000.
4. **Burk RF and Hill KE.** Orphan selenoproteins. *Bioessays* 21: 231–237, 1999.
5. **Cheeseman P and Stutz J.** Bayesian classification (AutoClass): theory and results. In: *Advances in Knowledge Discovery and Data Mining,* edited by Fayyad UM, Piatetsky-Shapiro, G, Smyth P, and Uthurusamy R. AAAI Press/MIT Press, 1996. [The software is available at http://ic-www.arc.nasa.gov/ic/projects/bayes-group/autoclass/index.html].
6. **Chertov O, Ueda H, Xu LL, Tani K, Murphy JM, Wang WJ, Howard OM, Sayers TJ, and Oppenheim JJ.** Identification of human neutrophil-derived cathepsin G and azurocidin/CAP37 as chemoattractants for mononuclear cells and neutrophils. *J Exp Med* 186: 739–747, 1997.
7. **Dudoit S, Yang YH, Callow MJ, and Speed TJ.** Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments [Online]. Dept. of Statistics, Univ. of California at Berkeley. http://www.stat.berkeley.edu/users/terry/zarray/Html/papersindex.html [3 Sept. 2000].
8. **Efron R, Tibshirani B, Goss V, and Chu G.** *Microarrays and Their Use in a Comparative Experiment* (Technical Report). Palo Alto, CA: Department of Statistics, Stanford University, 2000.
9. **Eisen MB, Spellman PT, Brown PO, and Botstein D.** Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95: 14863–14868, 1998.
10. **Furey T, Cristianini N, Duffy N, Bednarski D, Schummer M, and Haussler D.** Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16: 906–914, 2000.
11. **Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfeld CD, and Lander ES.** Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531–537, 1999. [The data are available at http://waldo.wi.mit.edu/MPR/data_sets.html]
12. **Guyon I, Weston J, Barnhill S, and Vapnik V.** Gene selection for cancer classification using support vector machines. *Machine Learning,* In press.
13. **Hastie T, Tibshirani R, Eisen M, Brown P, Ross D, Scherf U, Weinstein J, Alizadeh A, Staudt L, and Botstein D.** Gene shaving: a new class of clustering methods for expression arrays [Online]. Standford University. http://www-stat.stanford.edu/˜hastie/Papers/ [Jan. 2000].
14. **Holben DH and Smith AM.** The diverse role of selenium within selenoproteins: a review. *J Am Dietetic Assoc* 99: 836–843, 1999.
15. **Joachims T. Making large-scale SVM learning practical.** In: *Advances in Kernel Methods: Support Vector Learning,* edited by Schölkopf B, Burges C, and Smola A. MIT Press, 1999. [The software is available at http://ais.gmd.de/˜thorsten/svm_light]
16. **Moler EJ, Chow ML, and Mian IS.** Analysis of molecular profile data using generative and discriminative methods. *Physiol Genomics* 4: 109–126, 2000.
17. **Moler EJ, Radisky DC, and Mian IS.** Integrating naive Bayes models and external knowledge to examine copper and iron homeostasis in *S. cerevisiae. Physiol Genomics* 4: 127–135, 2000.
18. **Raychaudhuri R, Stuart JM, and Altman RB.** Principal components analysis to summarize microarray experiments: application to sporulation time series. In: *Pacific Symposium on Biocomputing,* 2000, vol. 5, p. 452–463.
19. **Ross DD.** Novel mechanisms of drug resistance in leukemia. *Leukemia* 14: 467–473, 2000.
20. **Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, and Golub TR.** Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 96: 2907–2912, 1999.
21. **Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, and Church GM.** Systematic determination of genetic network architecture. *Nat Genet* 22: 281–285, 1999.
22. **Toronen P, Kolehmainen M, Wong G, and Castren E.** Analysis of gene expression data using self-organizing maps. *FEBS Lett* 451: 142–146, 1999.
23. **Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, and Vapnik V.** Feature Selection for SVMs. *Adv Neural Inform Process Syst* 13: 2000.
24. **Wong ET, Jenne DE, Zimmer M, Porter SD, and Gilks CB.** Changes in chromatin organization at the neutrophil elastase locus associated with myeloid cell differentiation. *Blood* 94: 3730–3736, 1999.