Knowledge-based analysis of microarray gene expression data by using support vector machines

Michael P. S. Brown*, William Noble Grundy†‡, David Lin*, Nello Cristianini§, Charles Walsh Sugnet¶, Terrence S. Furey*

[redacted]

varying condition of interest, whereas the denominator is the expression level of the gene in some reference condition. The data from a series of m such experiments may be represented as a gene expression matrix, in which each of the n rows consists of an m-element expression vector for a single gene. Following Eisen *et al.* (1), we do not work directly with the ratio as discussed above but rather with its normalized logarithm. We define X_i to be the logarithm of the ratio of expression level E_i for gene X in experiment i to the expression level R_i of gene X in the reference state, normalized so that the expression vector $\vec{X} = (X_1, \ldots, X_{79})$ has Euclidean length 1:

For some data sets, the SVM may not be able to find a separating hyperplane in feature space, either because the kernel function is inappropriate for the training data or because the data contains mislabeled examples. The latter problem can be addressed by using a soft margin that allows some training examples to fall on the wrong side of the separating hyperplane. Completely specifying a support vector machine therefore requires specifying two parameters: the kernel function and the magnitude of the penalty for violating the soft margin. The settings of these parameters depend on the specific data at hand.

Given an expression vector $\vec{X}$ for each gene X, the simplest kernel $K(\vec{X}, \vec{Y}) = \vec{X} \cdot \vec{Y}$ that we can use to measure the similarity between genes

any classifier. The first five classes were selected because they represent categories of genes that are expected, on biological grounds, to exhibit similar expression profiles. Furthermore, Eisen *et al.* (1) suggested that the mRNA expression vectors for these classes cluster well using hierarchical clustering. The sixth class, the helix-turn-helix proteins, is included as a control group. Because there is no reason to believe that the members of this class are similarly regulated, we did not expect any classifier to learn to recognize members of this class based on mRNA expression measurements.

The performance of the SVM classifiers was compared with that of four standard machine learning algorithms: Parzen windows, Fisher's linear discriminant, and two decision tree learners (C4.5 and MOC1). Descriptions of these algorithms can be found at http://www.cse.ucsc.edu/research/compbio/genex. Performance was tested by using a three-way cross-validated experiment. The gene expression vectors were randomly divided into three groups. Classifiers were trained by using two-thirds of the data and were tested on the remaining third. This procedure was then repeated two more times, each time using a different third of the genes as test genes.

The performance of each classifier was measured by examining how well the classifier identified the positive and negative examples in the test sets. Each gene in the test set can be categorized in one of four ways: true positives are class members according to both the classifier and MYGD; true negatives are non-members according to both; false positives are genes that the classifier places within the given class, but MYGD classifies as non-members; false negatives are genes that the classifier places outside the class, but MYGD classifies as members. We report the number of genes in each of these four categories for each of the learning methods we tested.

To judge overall performance, we define the cost of using the method M as $C(M) = fp(M) + 2 \cdot fn(M)$, where $fp(M)$ is the number of false positives for method M, and $fn(M)$ is the number of false

Table 1. Comparison of error rates for various classification methods

Class	Method	FP	FN	TP	TN	S(M)
TCA	D-p 1 SVM	18	5	12	2,432	6
	D-p 2 SVM	7	9	8	2,443	9
	D-p 3 SVM	4	9	8	2,446	12
	Radial SVM	5	9	8	2,445	11
	Parzen	4	12	5	2,446	6
	FLD	9	10	7	2,441	5
	C4.5	7	17	0	2,443	−7
	MOC1	3	16	1	2,446	−1
Resp	D-p 1 SVM	15	7	23	2,422	31
	D-p 2 SVM	7	7	23	2,430	39
	D-p 3 SVM	6	8	22	2,431	38
	Radial SVM	5	11	19	2,432	33
	Parzen	22	10	20	2,415	18
	FLD	10	10	20	2,427	30
	C4.5	18	17	13	2,419	8
	MOC1	12	26	4	2,425	−4
Ribo	D-p 1 SVM	14	2	119	2,332	224
	D-p 2 SVM	9	2	119	2,337	229
	D-p 3 SVM	7	3	118	2,339	229
	Radial SVM	6	5	116	2,340	226
	Parzen	6	8	113	2,340	220
	FLD	15	5	116	2,331	217
	C4.5	31	21	100	2,315	169
	MOC1	26	26	95	2,320	164
Prot	D-p 1 SVM	21	7	28	2,411	35
	D-p 2 SVM	6	8	27	2,426	48
	D-p 3 SVM	3	8	27	2,429	51
	Radial SVM	2	8	27	2,430	52
	Parzen	21	5	30	2,411	39
	FLD	7	12	23	2,425	39

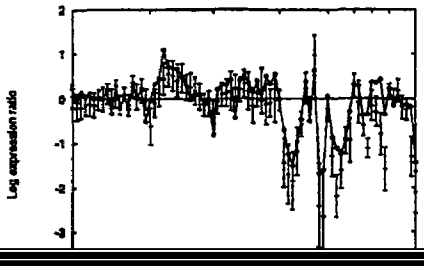
Table 2. Consistently misclassified genes

Class	Gene	Locus	Error	Description
TCA	YPR001W	CIT3	FN	Mitochondrial citrate synthase
	YOR142W	LSC1	FN	α subunit of succinyl-CoA ligase
	YLR174W	IDP2	FN	Isocitrate dehydrogenase
	YIL125W	KGD1	FN	α-ketoglutarate dehydrogenase
	YDR148C	KGD2	FN	Component of α-ketoglutarate dehydrog. complex (mito)
	YBL015W	ACH1	FP	Acetyl CoA hydrolase
Resp	YPR191W	QCR2	FN	Ubiquinol cytochrome-c reductase core protein 2
	YPL271W	ATP15	FN	ATP synthase ε subunit
	YPL262W	FUM1	FP	Fumarase
	YML120C	NDI1	FP	Mitochondrial NADH ubiquinone 6 oxidoreductase
	YKL085W	MDH1	FP	Mitochondrial malate dehydrogenase
	YGR207C		FN	Electron-transferring flavoprotein, β chain
	YDL067C	COX9	FN	Subunit VIIa of cytochrome c oxidase
Ribo	YPL037C	EGD1	FP	β subunit of the nascent-polypeptide-associated complex
	YLR406C	RPL31B	FN	Ribosomal protein L31B (L34B) (YL28)
	YLR075W	RPL10	FP	Ribosomal protein L10
	YDL184C	RPL41A	FN	Ribosomal protein L41A (YL41) (L47A)
	YAL003W	EFB1	FP	Translation elongation factor EF-1β
Prot	YHR027C	RPN1	FN	Subunit of 26S proteasome (PA700 subunit)
	YGR270W	YTA7	FN	Member of CDC48/PAS1/SEC18 family of ATPases
	YGR048W	UFD1	FP	Ubiquitin fusion degradation protein
	YDR069C	DOA4	FN	Ubiquitin isopeptidase
	YDL020C	RPN4	FN	Involved in ubiquitin degradation pathway
Hist	YOL012C	HTA3	FN	Histone-related protein
	YKL049C	CSE4	FN	Required for proper kinetochore function

The table lists all 25 genes that are most consistently misclassified by the SVMs. Two types of errors are included: a false positive (FP) occurs when the SVM includes the gene in the given class but the MYGD classification does not; a false negative (FN) occurs when the SVM does not include the gene in the given class but the MYGD classification does.

clustering, the histone cluster only identified 8 of the 11 histones, and the ribosome cluster only found 112 of the 121 genes and included 14 others that were not ribosomal genes (1).

We repeated the experiment with all four SVMs four more times with different random splits of the data. The results show that the variance introduced by the random splitting of the data is small, relative to the mean. The easiest-to-learn functional classes are those with the smallest ratio of standard deviation to mean cost savings. For example, for the radial basis SVM, the mean and standard deviations of the cost savings for the two easiest classes—ribosomal proteins and histones—are 225.8 $\pm$ 2.9 and 18.0 $\pm$ 0.0, respectively. The most difficult class, TCA cycle, had a mean and standard deviation of 10.4 $\pm$ 3.0. Results for the other classes and other kernel functions are similar (http://www.cse.ucsc.edu/research/compbio/genex).



Significance of Consistently Misclassified Annotated Genes. The five different three-fold cross-validation experiments, each performed with four different kernels, yield a total of 20 experiments per functional class. Across all five functional classes (excluding helix-turn-helix) and all 20 experiments, 25 genes are misclassified in at least 19 of the 20 experiments (Table 2). In general, these disagreements with MYGD reflect the different perspective provided by the expression data, which represents the genetic response of the cell, and the MYGD definitions, which have been arrived at through experiments or protein structure predictions. For example, in MYGD, the members of a complex are defined by biochemical co-purification whereas the expression data may identify proteins that are not physically part of the complex but contribute to proper functioning of the complex. This will lead to disagreements in the form of false positives. Disagreements between the SVM and MYGD in the form of false negatives may occur for a number of reasons. First, genes that are classified in MYGD primarily by structure (e.g., protein kinases) may have very different expression patterns. Second, genes that are regulated at the translational level or protein level, rather than at the transcriptional level as measured by the microarray experiments, cannot be correctly classified by expression data alone. Third, genes for which the microarray data

Table 3. Predicted functional classifications for previously unannotated genes

Class	Gene	Locus	Comments
TCA	YHR188C		Conserved in worm, *Schizosaccharomyces pombe*, human
	YKL039W	PTM1	Major transport facilitator family; likely integral membrane protein; similar YHL017w not co-regulated.
Resp	YKR016W		Not highly conserved, possible homolog in *S. pombe*
	YKR046C		No convincing homologs
	YPR020W	ATP20	Subsequently annotated: subunit of mitochondrial ATP synthase complex
	YLR248W	CLK1/RCK2	Cytoplasmic protein kinase of unknown function
Ribo	YKL056C		Homolog of translationally controlled tumor protein, abundant, conserved and ubiquitous protein of unknown function
	YNL119W		Possible remote homologs in several divergent species
	YNL255C	GIS2	Cellular nucleic acid binding protein homolog, seven CCHC (retroviral) type zinc fingers
	YNL053W	MSG5	Protein–tyrosine phosphatase, overexpression bypasses growth arrest by mating factor
	YNL217W		Similar to bis (5′ nucleotidyl)-tetraphosphatases
Prot	YDR330W		Ubiquitin regulatory domain protein, *S. pombe* homolog
	YJL036W		Member of sorting nexin family
	YLR387C		Three C2H2 zinc fingers, similar YBR267W not coregulated

The table lists the names for unannotated genes that were classified as members of a particular functional class by at least three of the four SVM methods. No unannotated histones were predicted.

ribosomal proteins; however, both are important for proper functioning of the ribosome. YAL003W encodes a translation elonga-

classified as a histone protein by MYGD based on its 61% amino acid similarity with histone protein H3. YKL049C is thought to act

Functional Class Predictions for Genes of Unknown Function. In addition to validating the classification accuracy of SVM methods using genes of known function, we used SVMs to classify previously unannotated yeast genes. A common trivial outcome of this experiment predicts a function for ORFs that overlap or are adjacent to

ments and have made predictions aimed at identifying the functions of unannotated yeast genes. Among the techniques examined, SVMs that use a higher-dimensional kernel function provide the best performance—better than Parzen windows, Fisher's linear discriminant, two decision tree classifiers, and