Support Vector Machine Classification of Microarray Gene Expression Data

William Noble Grundy

Department of Computer Science
University of California, Santa Cruz
http://www.cse.ucsc.edu/∼bgrundy

Outline

- Introduction
- DNA microarray expression data
- Support vector machines
- Results
- Future work

Genomics

- Genomics is the convergence of two great innovations of the latter half of the twentieth century: computers and biotechnology.
- The post-genomic era will give us our first comprehensive view of the cell, of the tree of life, and of human disease at the molecular level.

Genome-wide data sets

A variety of genome-wide data sets are available, including

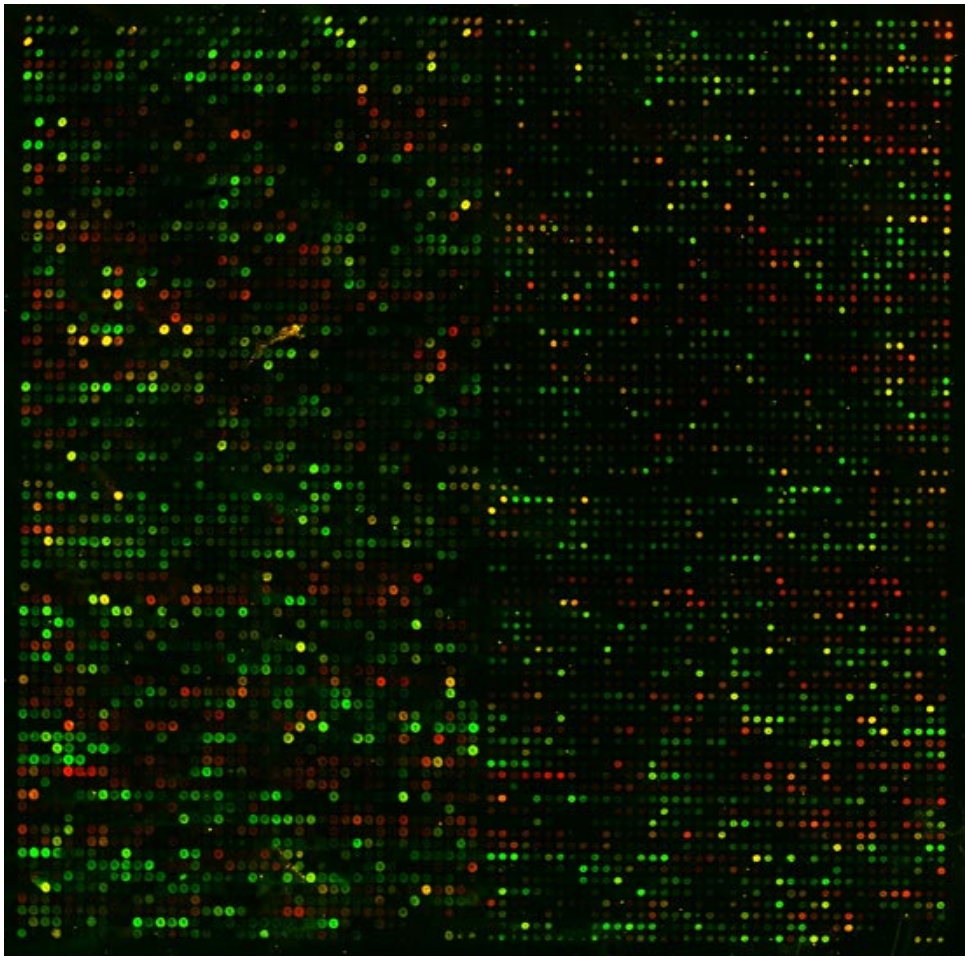
- Expressed Sequence Tag (EST) libraries;
- full genomic sequences, including all genes that are
  1. paralogs within the same organism,
  2. orthologs from closely related organisms, and
  3. remote functional and structural homologs from distantly related organisms;
- Single Nucleotide Polymorphism (SNP) databases, representing the natural variation within a population;
- mRNA expression levels obtained from microarray technology;
- protein expression levels from mass spectrometry technology;
- measurements of protein-protein and protein-DNA interactions, and
- a host of new methods to alter the expression of specific genes.

Two tasks

Given megabases of newly sequenced DNA per day, along with EST "hits", protein database "hits", and differential mRNA expression data,

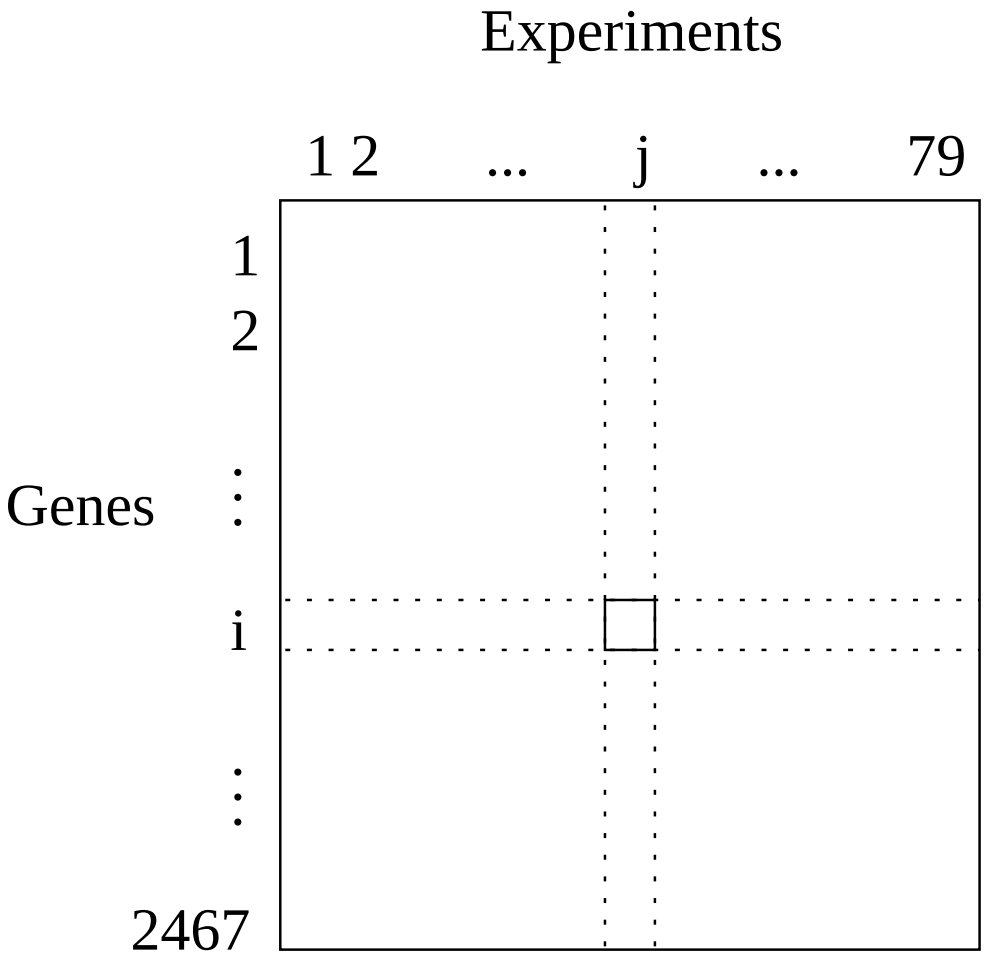
- find the protein coding genes in this DNA, identifying exons, promoter, regulatory binding sites, and other key features, and
- make an initial classification and analysis of the protein sequences derived from this DNA.

DNA microarray hybridization experiments



- Each location in the array contains a DNA sample fixed to the glass slide.
- A single experiment consists of hybridizing the microarray with two differently dyed cDNA samples from different experimental conditions.

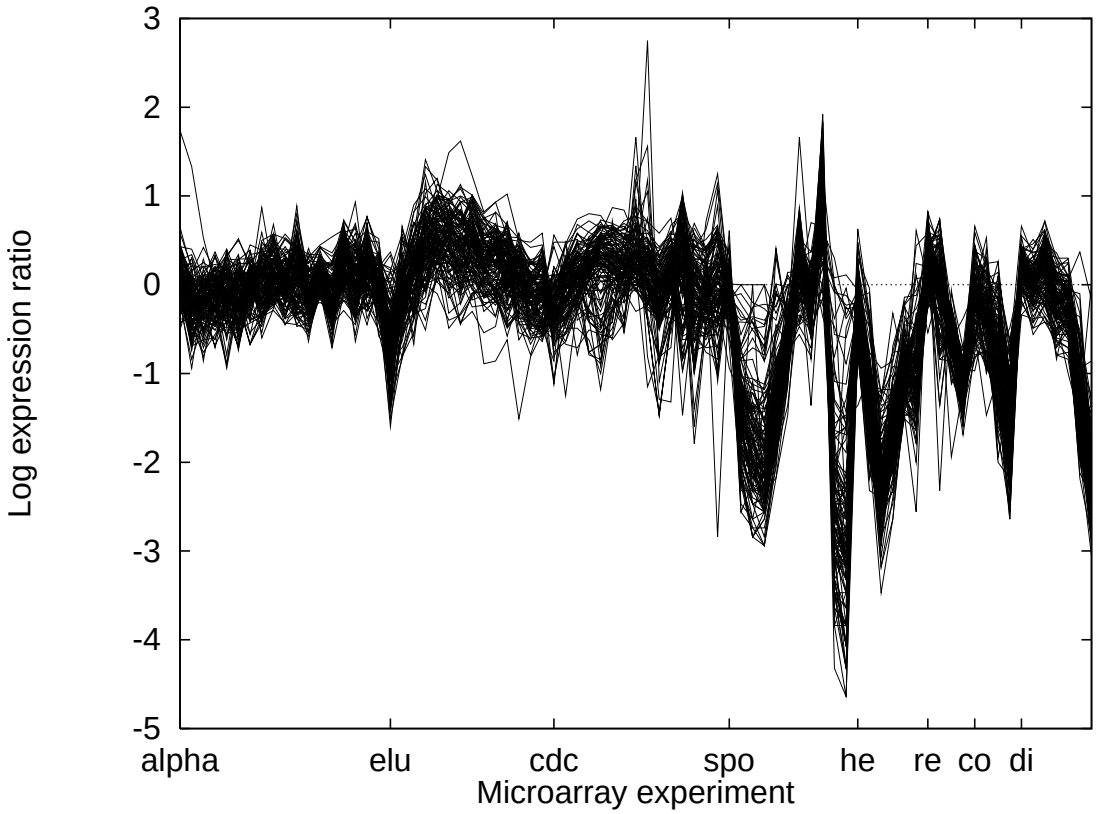
Gene expression matrix



The matrix entry at row i, column j is the logarithm of the ratio of gene i's expression level in experiment j to gene i's expression level in the reference state.

We use data from 2467 S. cerevisiae genes in 79 experimental conditions, including the cell division cycle, sporulation, heat shock, reducing shock, cold shock, and diauxic shift (http://rana.stanford.edu/clustering).

Functional similarities



Cytoplasmic ribosomal proteins

Expression profiles of functionally related genes are similar to one another.

Unsupervised learning

Previous work has employed clustering techniques to group functionally related genes.

- Each technique begins with a definition of similarity.
- Eisen *et al.* use hierarchical clustering.
- Tamayo *et al.* use self-organizing maps.

Supervised learning

Support vector machines (SVMs) are a supervised learning technique developed by Vapnik and others.

Training an SVM requires labeled training data.

A trained SVM can answer two questions:

- What other genes are related to my set?
- Does my set contain genes that do not belong?

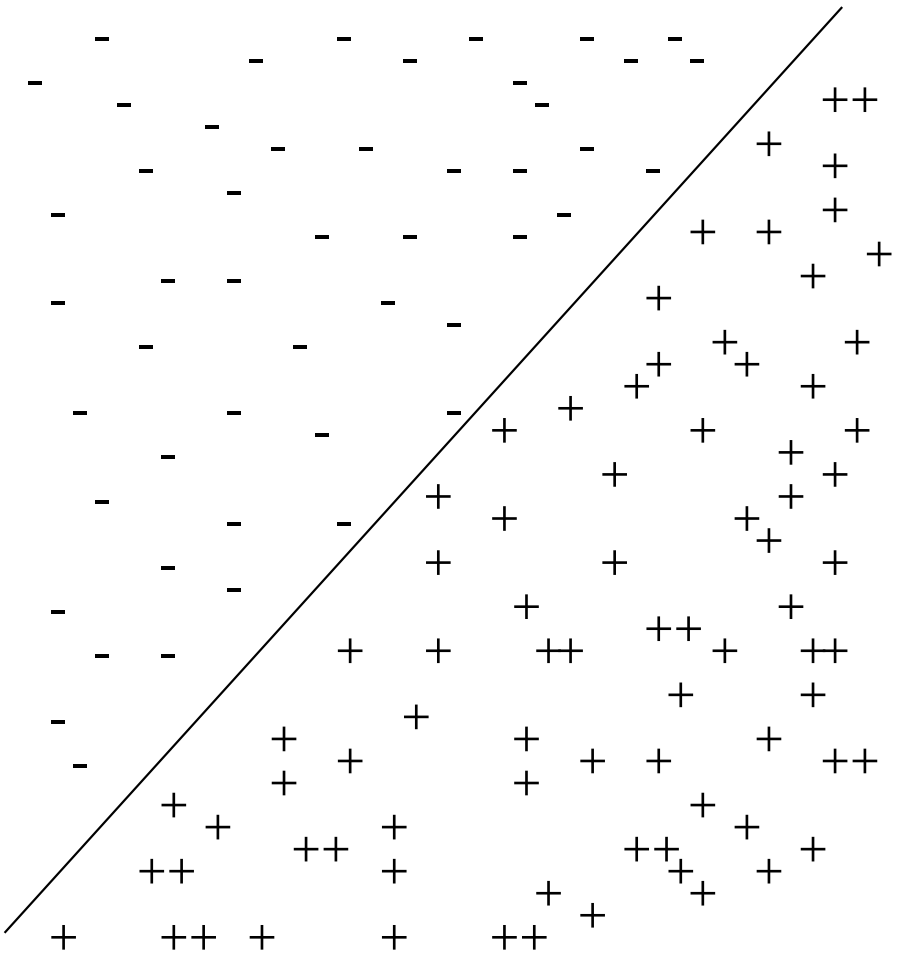
Functional classes

In this work, each gene is classified according to five different MIPS Yeast Genome Database functional classes:

- tricarboxylic-acid pathway
- respiration chain complexes
- cytoplasmic ribosomes
- proteasome
- histones

These classes have been pointed out by Eisen *et al.* in their discussion.

Separating hyperplane



- Each vector in the gene expression matrix may be thought of as a point in a 79-dimensional space.
- A simple way to build a binary classifier is to construct a hyperplane separating class members from non-members in this space.
- This is the approach taken by perceptrons, also known as single-layer neural networks.

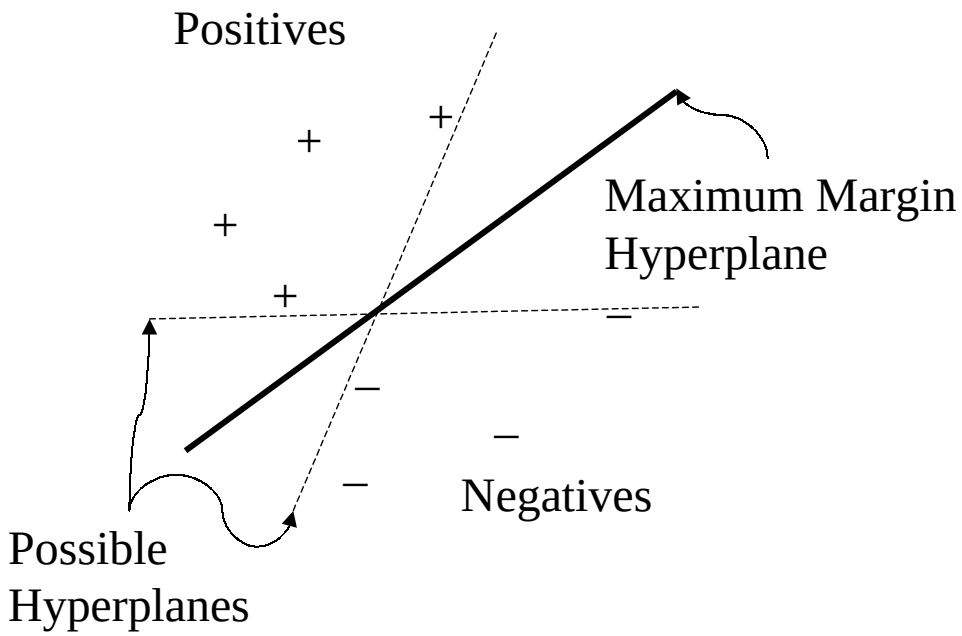
Support vector machines

Support vectors machines (SVMs) first map the data from the input space to a higher-dimensional feature space.

A separating hyperplane is found in the feature space.

The feature space is not represented explicitly, in order to avoid computational overhead.

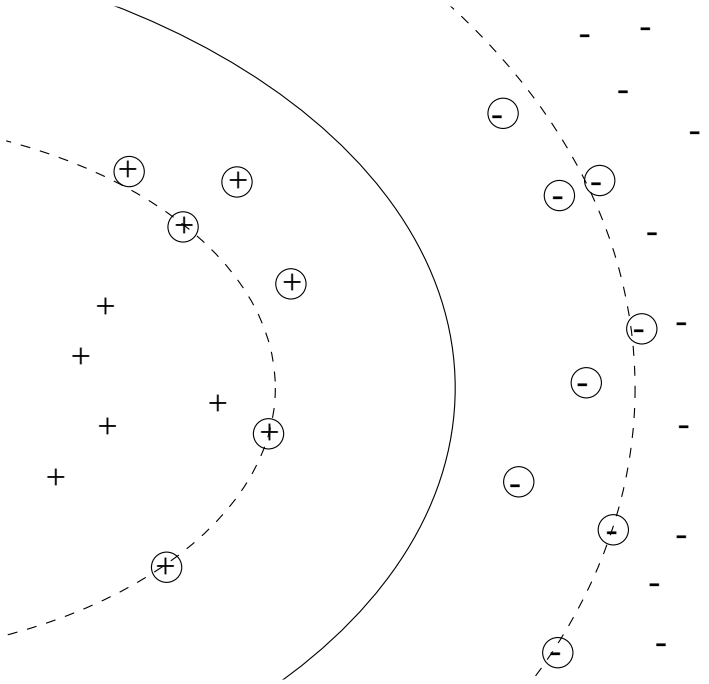
Maximum margin hyperplane



When multiple separating hyperplanes exist, the SVM uses the hyperplane that maximally separates positive from negative examples.

The use of the maximum margin hyperplane helps the SVM avoid overfitting.

Soft margin

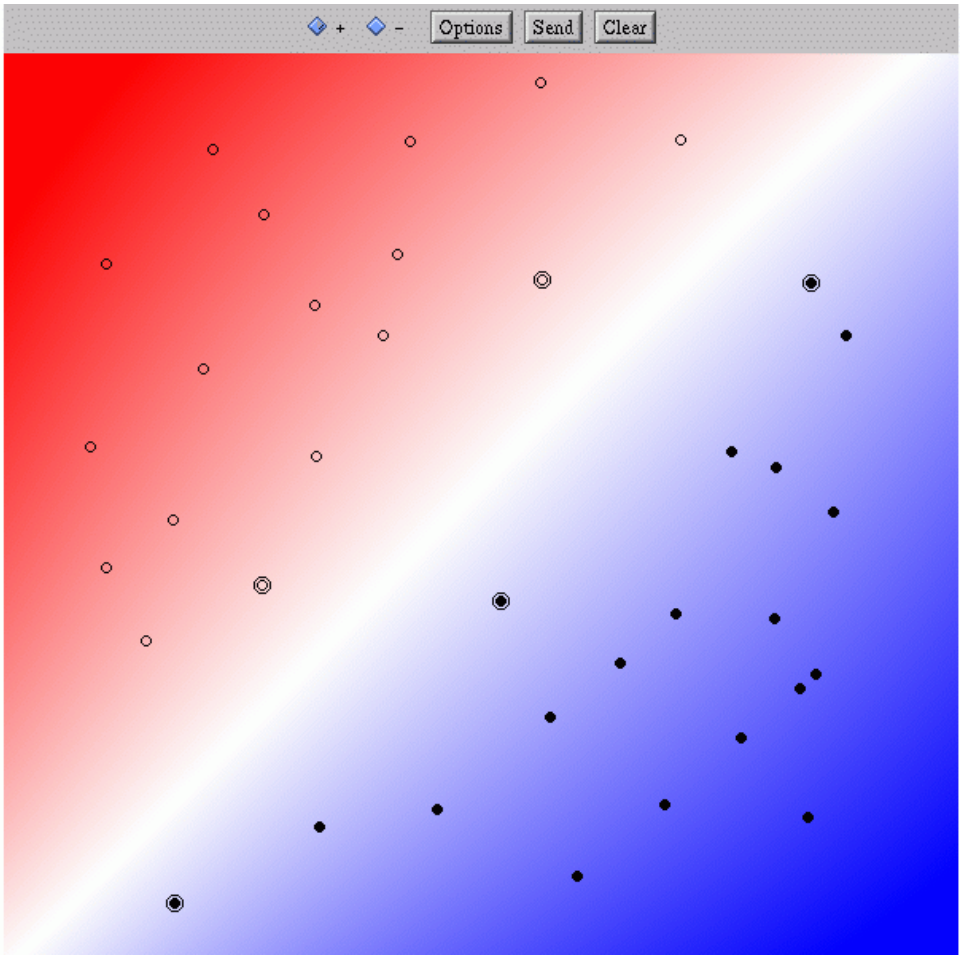


When no separating hyperplane exists, the SVM uses a soft margin hyperplane with minimal cost.

The kernel function

- Training an SVM requires finding the separating hyperplane for a given labeled training set.
- Classifying an unlabeled test example requires locating the given data point in the feature space.
- Both training and classification can be stated entirely in terms of vectors in the input space and dot products in the feature space.
- The kernel function plays the role of the dot product in the feature space.
- Any continuous, positive semi-definite function can act as a kernel function.
- The use of a kernel function allows us to avoid representing the feature space explicitly.

Linear decision boundary

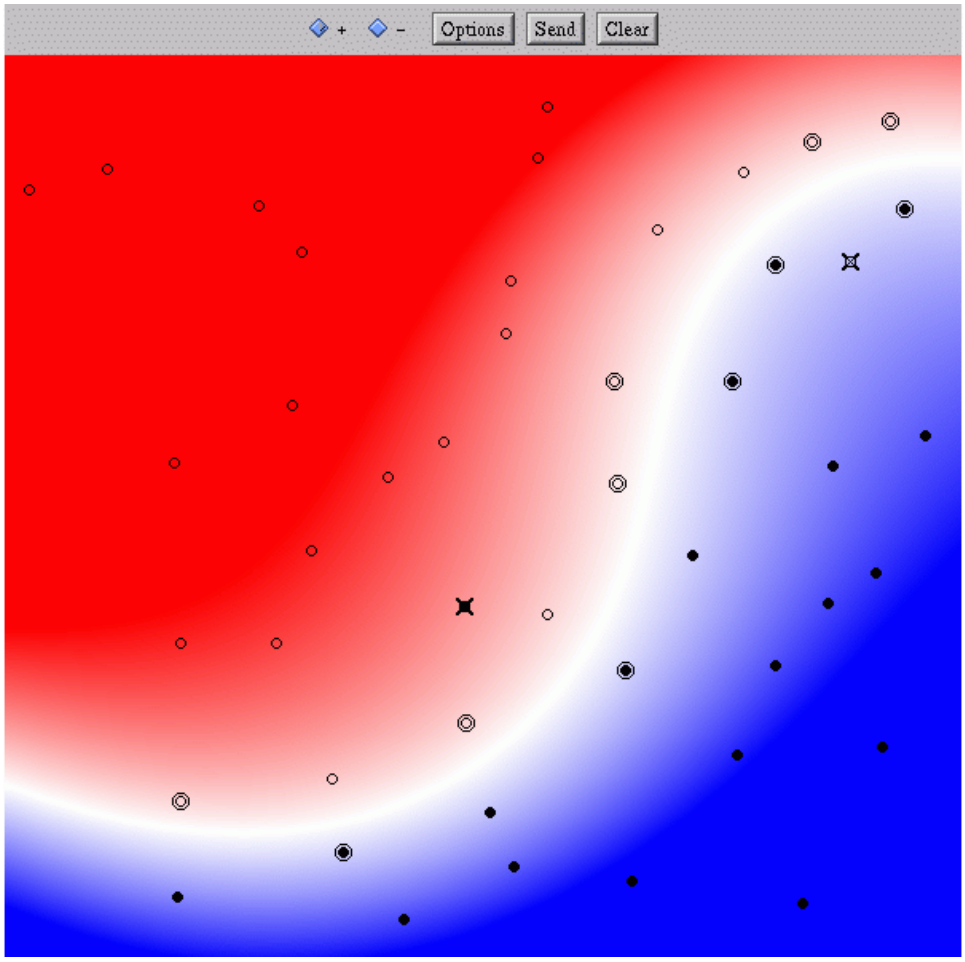


$$K(X, Y) = X \cdot Y$$

Using a simple dot product as the kernel creates a linear decision boundary in the input space.

This kernel, with normalization, corresponds to the distance metric used by Eisen *et al.* to perform hierarchical clustering.

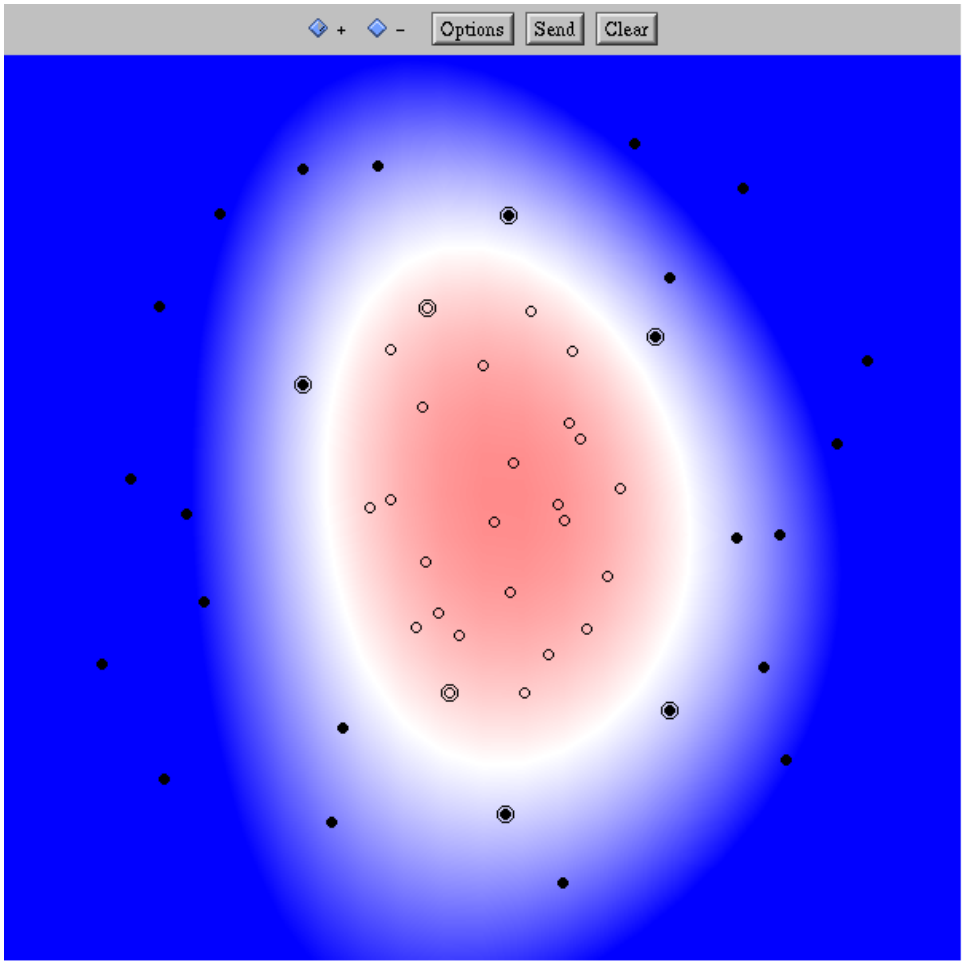
Polynomial decision boundary



$$K(X, Y) = ((X \cdot Y) + 1)^3$$

Raising the dot product to a power yields a polynomial decision boundary in the input space.

Gaussian decision boundary



$$K(X,Y) = exp\left(\frac{-||X-\bar{Y}||^2}{2\sigma^2}\right)$$

A radial basis kernel function yields a Gaussian decision boundary in the input space.

SVM summary

- A support vector machine finds a nonlinear decision function in the input space.
- This boundary corresponds to a hyperplane in a higher-dimensional feature space.
- The computational complexity of the classification operation does not depend on the dimensionality of the feature space.
- Overfitting is avoided by controlling the margin.
- The hyperplane is represented sparsely as a linear combination of points.
- The SVM automatically identifies a subset of informative points and uses them to represent the solution.
- The training algorithm solves a simple convex optimization problem.

Other supervised learning techniques

Decision trees

- Each tree is binary, with simple classifiers at internal nodes and a classification at each leaf.
- The standard algorithm of this kind is C4.5 (Quinlan 1997).
- MOC1 (Wu *et al.* 1999) is a variant motivated by VC learning theory.

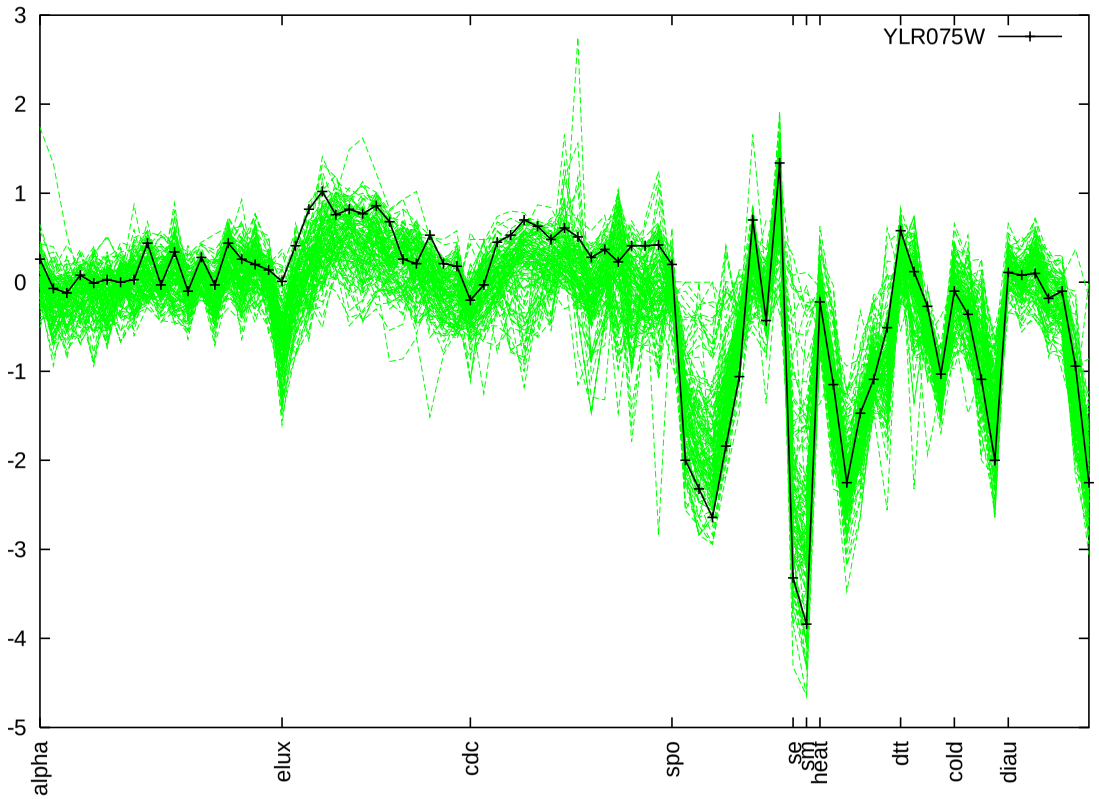
Parzen windows is a generalization of k-nearest neighbor techniques that employs a radial basis function as the distance metric.

Fisher's linear discriminant projects high-dimensional data onto a line and performs classification in this one-dimensional space.

Consistently misclassified genes

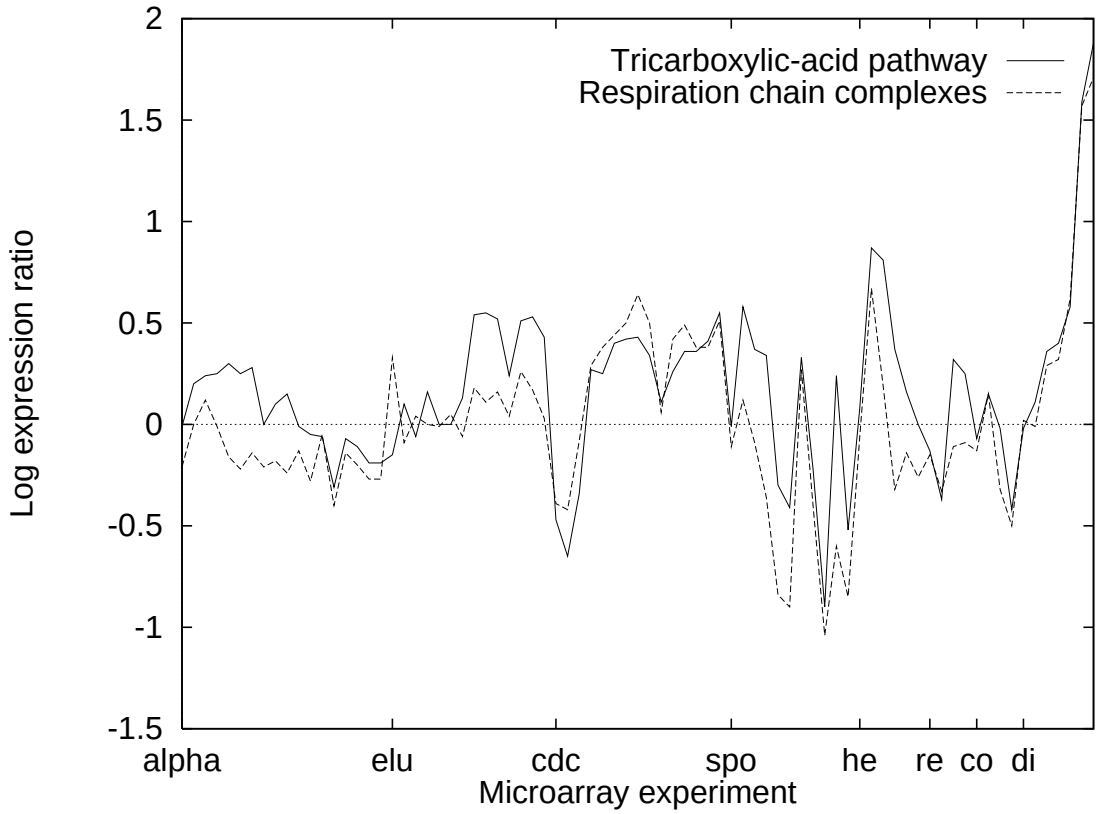
Family	Gene	Locus	Error	Description
TCA	YPR001W	CIT3	FN	mitochondrial citrate synthase
	YOR142W	LSC1	FN	α subunit of succinyl-CoA ligase
	YNR001C	CIT1	FN	mitochondrial citrate synthase
	YLR174W	IDP2	FN	isocitrate dehydrogenase
	YIL125W	KGD1	FN	α-ketoglutarate dehydrogenase
	YDR148C	KGD2	FN	component of α-ketoglutarate dehydrogenase complex in mitochondria
	YDL066W	IDP1	FN	mitochondrial form of isocitrate dehydrogenase
	YBL015W	ACH1	FP	acetyl CoA hydrolase
Resp	YPR191W	QCR2	FN	ubiquinol cytochrome-c reductase core protein 2
	YPL271W	ATP15	FN	ATP synthase epsilon subunit
	YPL262W	FUM1	FP	fumarase
	YML120C	NDI1	FP	mitochondrial NADH ubiquinone 6 oxidoreductase
	YKL085W	MDH1	FP	mitochondrial malate dehydrogenase
	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
	YLR075W	RPL10	FP	ribosomal protein L10
	YAL003W	EFB1	FP	translation elongation factor EF-1β
Prot	YHR027C	RPN1	FN	subunit of 26S proteasome
	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 pthwy
Hist	YOL012C	HTA3	FN	histone-related protein
	YKL049C	CSE4	FN	required for proper kinetochore function

An error in the training labels



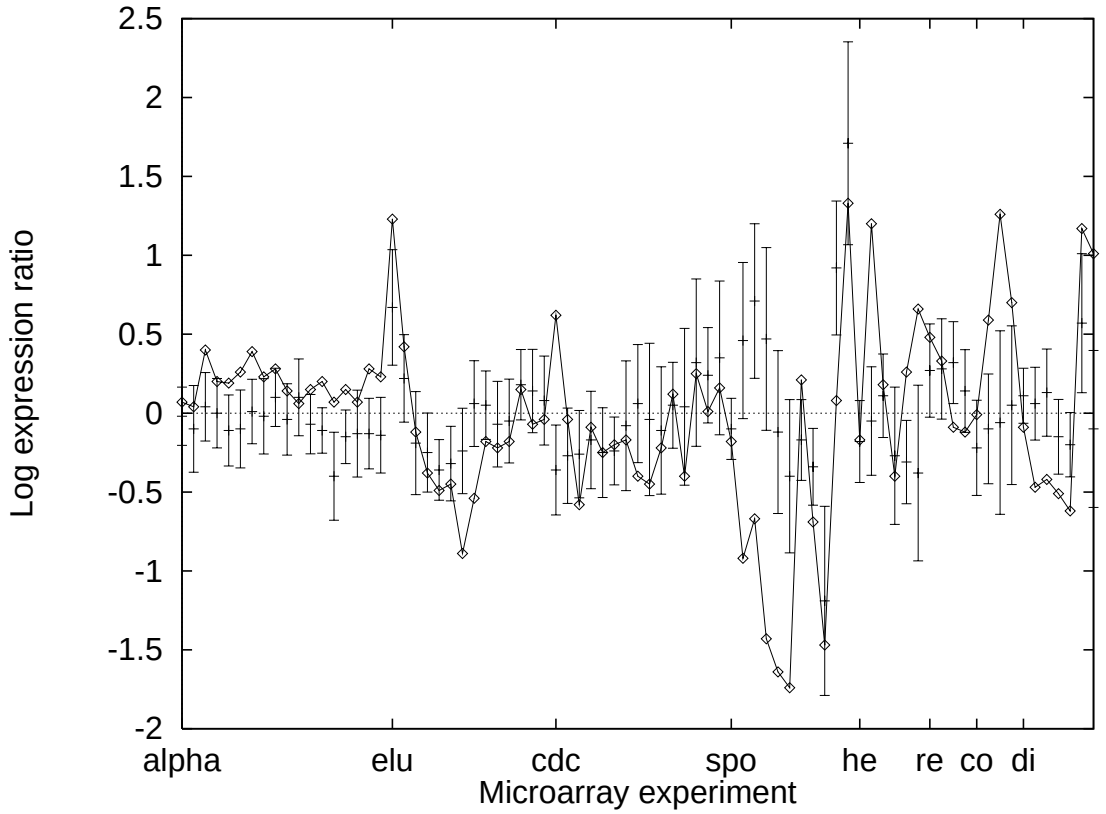
YLR075W is associated with the ribosome both by homology and purification.

Fuzzy boundaries between classes



The tricarboxylic-acid pathway and the respiration chain complexes are tightly coupled in the production of ATP.

Variation in regulation



YDL020C

YDL020C and YDR069C are auxiliary factors of the proteasome that may be regulated specifically during sporulation.

Training set outliers

Gene	Weight	Errors
YLR075W	2.093	5
YOR276W	1.016	4
YNL209W	0.977	4
YAL003W	0.930	5
YPL037C	0.833	5
YKR059W	0.815	2
YML106W	0.791	1
YDR385W	0.771	2
YPR187W	0.767	1
YJL138C	0.757	3

The magnitude of the training set weights predicts outliers.

- The "Weight" column lists the average learned weight of the gene over five different three-fold splits of the data.
- The "Errors" column lists the total number of times the gene was incorrectly classified.
- The table lists the negative genes with the largest weights.

Conclusions

SVMs provide an accurate means of functionally classifying genes using only expression data.

The radial basis SVM appears to provide better performance than SVMs using the dot product kernel.

If the complete data set were available, our trained SVMs could be used to predict functional classes of unannotated genes.

The method is scalable, which is essential, given the amount of data becoming available.

Future work

Incorporate data from other sources, such as

- protein features and
- promoter region features.

Design better kernel functions.

- The kernel function incorporates prior knowledge of the learning domain.
- The kernel function could model dependencies among experiments.

Employ SVMs in an iterative framework.

- SVMs are members of the larger class of kernel methods.
- Use an unsupervised method as the first step.
- Use the SVM to clean a given classification.

Collaborators

- Michael Brown
- David Lin
- Nello Cristianini
- Charles Sugnet
- Manuel Ares, Jr.
- David Haussler