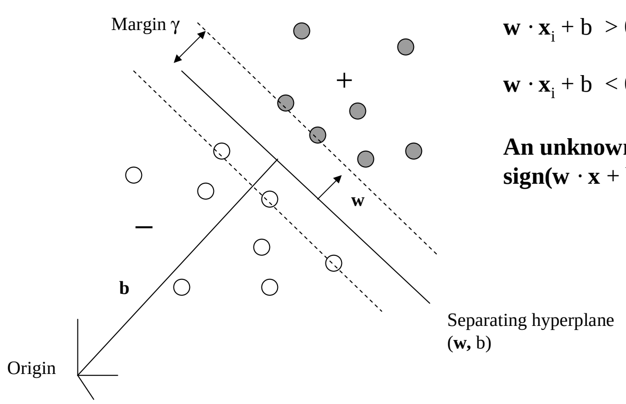
CISC 889 Bioinformatics (Spring 2004)

Support Vector Machines (II)

applications in bioinformatics

Linear SVMs: find a hyperplane (specified by normal vector **w** and perpendicular distance **b** to the origin) that separates the positive and negative examples with the largest margin.



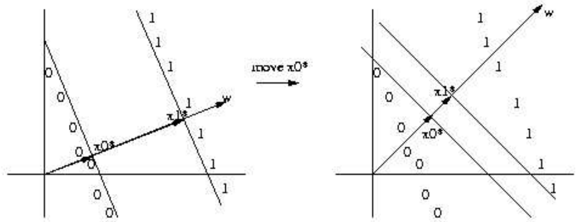
$\mathbf{w} \cdot \mathbf{x}_i + b > 0$ if $y_i = +1$

$\mathbf{w} \cdot \mathbf{x}_i + b < 0$ if $y_i = -1$

An unknown x is classified as sign(w · x + b)

Larger margin is preferred:

- converge more quickly
- generalize better



$\mathbf{w} \cdot \mathbf{x}_+ + b = +1$

$\mathbf{w} \cdot \mathbf{x}_- + b = -1$

$2 = [(\mathbf{x}_+ \cdot \mathbf{w}) - (\mathbf{x}_- \cdot \mathbf{w})] = (\mathbf{x}_+ - \mathbf{x}_-) \cdot \mathbf{w} = \| \mathbf{x}_+ - \mathbf{x}_- \| \, \|\mathbf{w}\|$

Therefore, maximizing the geometric margin $\| \mathbf{x}_+ - \mathbf{x}_- \|$ is equivalent to minimizing $\frac{1}{2} \|w\|^2$, under linear constraints: $y_i (\mathbf{w} \cdot \mathbf{x}_i) + b \geq 1$ for $i = 1, \ldots, n$.

This optimization can be solved by introducing Lagrangian multiplier α_i for each constraint

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|w\|^2 - \sum \alpha_i (y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1),$$

and then calculating

$$\frac{\partial L}{\partial \mathbf{w}} = 0, \qquad \frac{\partial L}{\partial b} = 0, \qquad \frac{\partial L}{\partial \boldsymbol{\alpha}} = 0,$$

The optimal w* and b* can be found by solving the dual problem for α to maximize:

$$L(\boldsymbol{\alpha}) = \sum \alpha_i - ½ \sum \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$$

under the constraints: $\alpha_i \geq 0$, and $\sum \alpha_i y_i = 0$.

Once $\boldsymbol{\alpha}$ is solved,

$\mathbf{w}^* = \sum \alpha_i y_i \mathbf{x}_i$

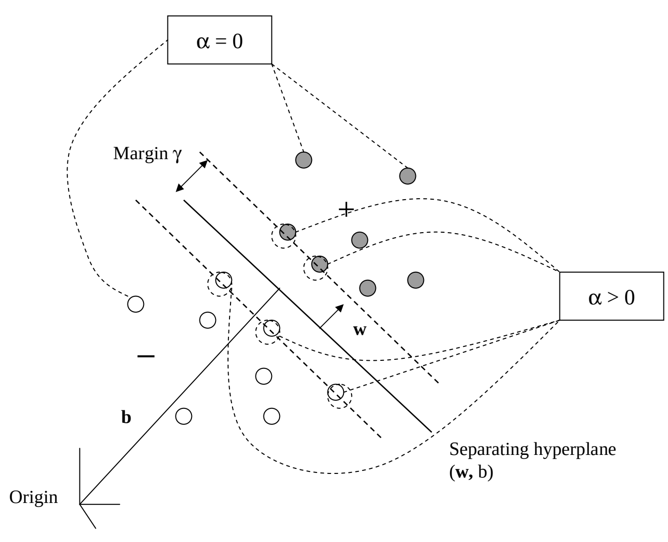
$b^* = ½ (\max_{y=-1} \mathbf{w}^* \cdot \mathbf{x}_i + \min_{y=+1} \mathbf{w}^* \cdot \mathbf{x}_i)$

And an unknown $\mathbf{x}$ is classified as

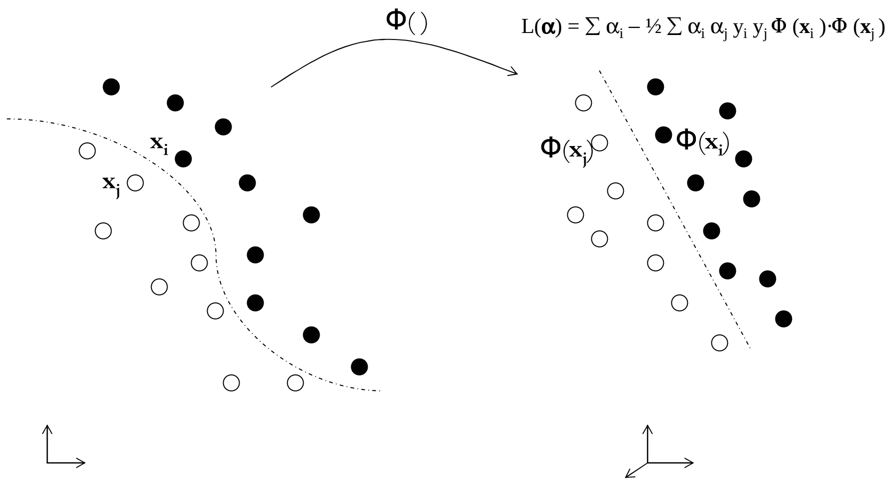
$$\text{sign}(\mathbf{w}^* \cdot \mathbf{x} + b^*) = \text{sign}(\sum \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b^*)$$

Notes:

1. Only the dot product for vectors is needed.
2. Many α_i are equal to zero, and those that are not zero correspond to $\mathbf{x}_i$ on the boundaries – support vectors!
3. In practice, instead of sign function, the actual value of $\mathbf{w}^* \cdot \mathbf{x} + b^*$ is used when its absolute value is less than or equal to one. Such a value is called a discriminant.



Non-linear mapping to a feature space



Kernels

Given a mapping $\Phi( )$ from the space of input vectors to some higher dimensional feature space, the kernel K of two vectors $\mathbf{x}_i$, $\mathbf{x}_j$ is the inner product of their images in the feature space, namely, $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$.

Since we just need the inner product of vectors in the feature space to find the maximal margin separating hyperplane, we use the kernel in place of the mapping $\Phi( )$.

Because inner product of two vectors is a measure of the distance between the vectors, a kernel function actually defines the geometry of the feature space (lengths and angles), and implicitly provides a similarity measure for objects to be classified.

Applications to Gene Expression

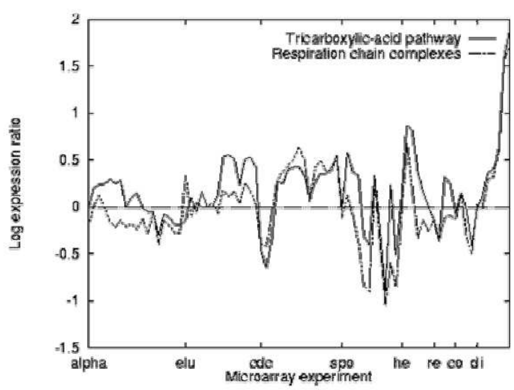
http://www.soe.ucsc.edu/research/compbio/genex

Each gene (of 2467 genes from S. cerevisiae) is represented as a vector

$X = (x_1, x_2, \ldots, x_n)$

where $x_i = \log(E_i / R_i) / (\sqrt{\sum_j \log^2 (E_j / R_j)})$, E_i/R_i is ratio of expression levels for a gene between test and control at the i-th experiment (out of n = 79 total).

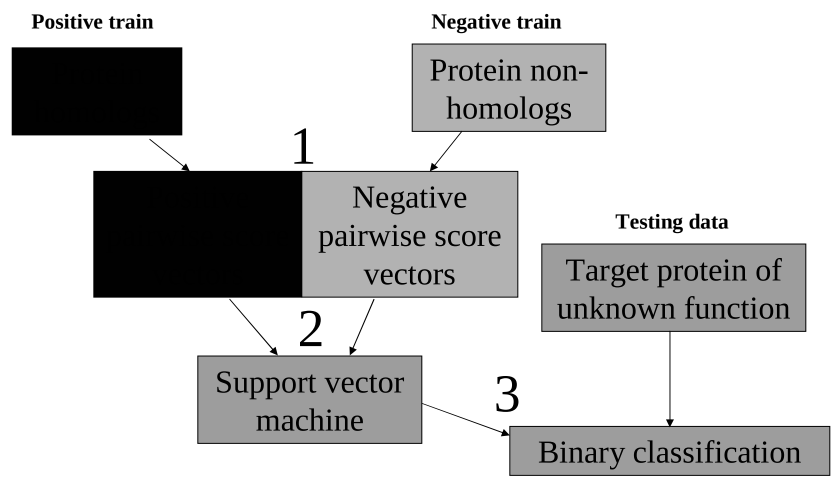
Similarity between the average expression profiles of the tricarboxylic-acid pathway and respiration chain complexes.



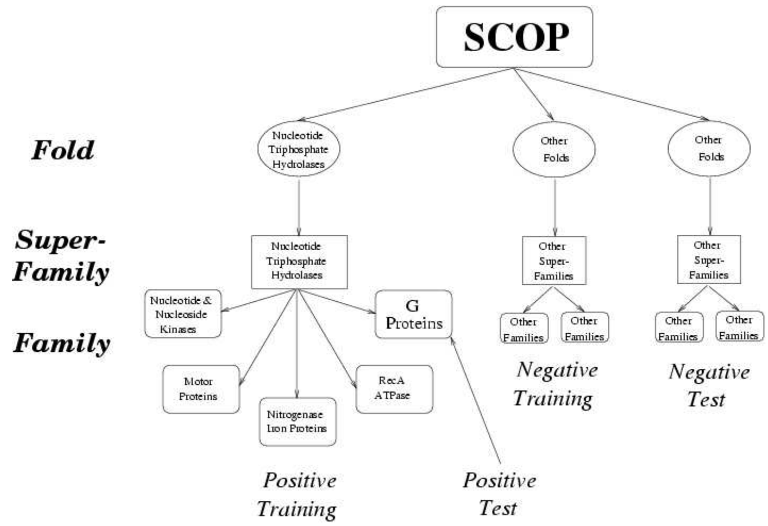
Other applications in bioinformatics

- Protein localization
  - Hua and Sun, Bioinformatics, 17(2001)721-728.
- Protein secondary structure prediction
  - Hua and Sun, JMB, 308(2001)397-407
  - Ding and Dubchak, Bioinformatics, 17(2001)349-358
- Protein superfamily
  - Jaakkola, Diekhans and haussler, JCB 7(2000)95-114
  - Liao and Noble, RECOMB 2002.
- Tissue classification
  - Guyon, Weston, Barnhill and Vapnik, Machine Learning 46(2002)389-422

Combining pairwise similarity with SVMs for protein homology detection

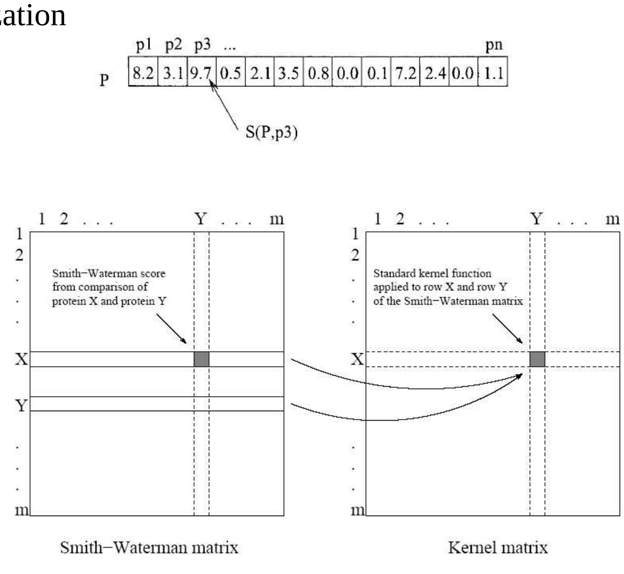


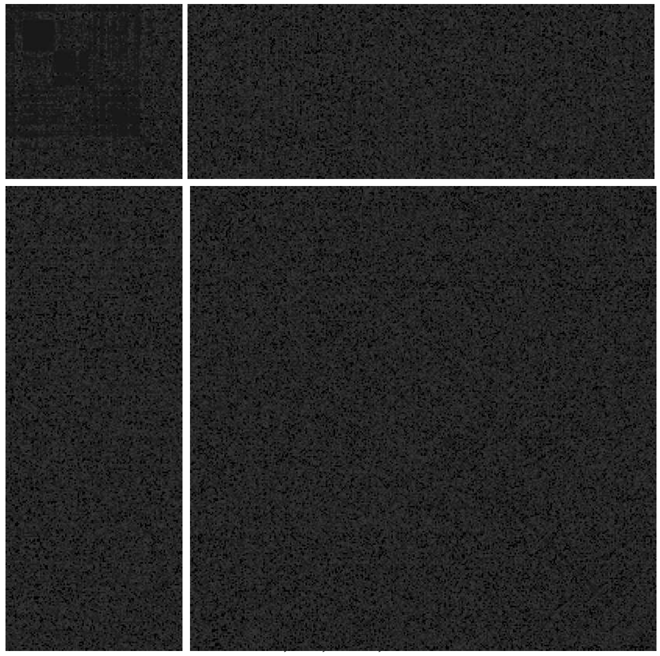
Experiment: known protein families



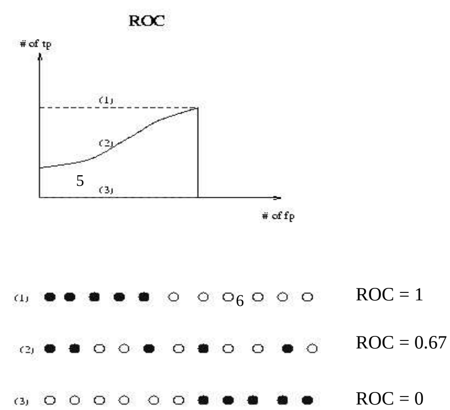
Jaakkola, Diekhans and Haussler 1999

Vectorization



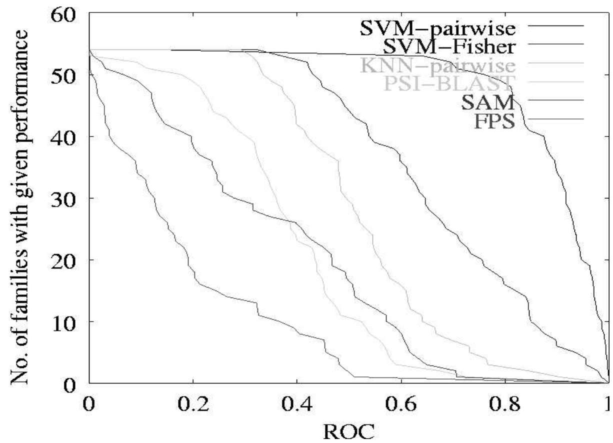


A measure of sensitivity and specificity



ROC: receiver operating characteristic score is the normalized area under a curve the plots true positives as a function of false positives

Performance Comparison (1)



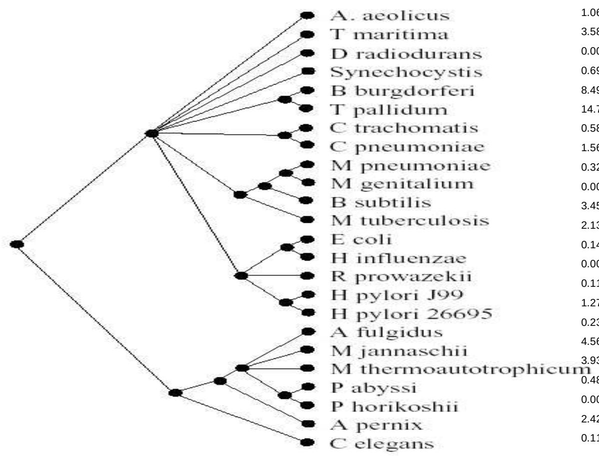
Kernel engineering

- Kernels provide a similarity measure
- Kernels should reflect special features about the data (e.g., prior knowledge)
- Special kernels can be designed for that purpose.
  - Kernels for sequences based on common subsequences
  - Kernels built from Bayesian tree models
  - Convolution kernels
  - Diffusion kernels on graphs

- Kernels based on common subsequences
  - Two strings are more similar in the feature space if they share rare common substrings
  - $K(X,Y) = p(X)p(Y) \sum_{s \in \text{common substrings}} (1/p(s))$ where p is the probability distribution on the set of sequences

- Convolution Kernels
  - X and Y are closer to each other if they share common “hidden state sequences”
  - $K(X,Y) = \sum_s p(s)p(X|s)p(Y|s)$
  - D. Haussler, Convolution kernels on discrete structures, technical report UC Santa Cruz, 1999

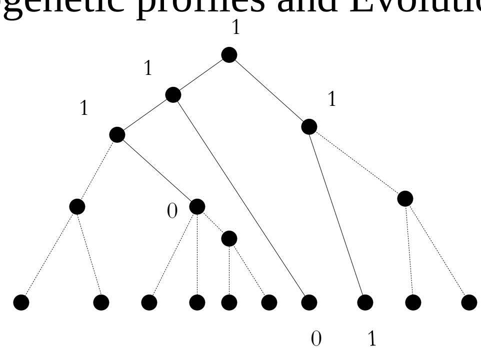
Using Phylogenetic Profiles & SVMs



YAL001C

E-value	Phylogenetic profile
0.122	1
1.064	0
3.589	0
0.008	1
0.692	1
8.49	0
14.79	0
0.584	1
1.567	0
0.324	1
0.002	1
3.456	0
2.135	0
0.142	1
0.001	1
0.112	1
1.274	0
0.234	1
4.562	0
3.934	0
0.489	1
0.002	1
2.421	0
0.112	1

phylogenetic profiles and Evolution Patterns



x 1 1 0 1 0 0 0 1 1 0

Impossible to know for sure if the gene followed exactly this evolution pattern

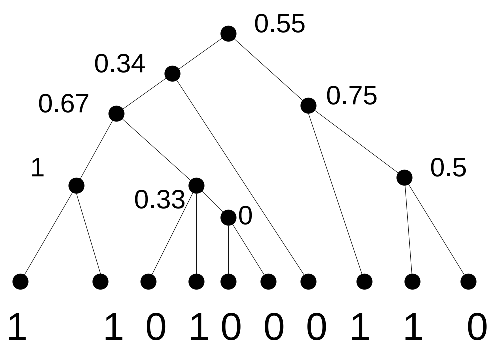
Tree Kernel (Vert, 2002)

- For a phylogenetic profile x and an evolution pattern e:
  - P(e) quantifies how "natural" the pattern is
  - P(x|e) quantifies how likely the pattern e is the "true history" of the profile x
- Tree Kernel :

$$K_{tree}(x,y) = \Sigma_e \, p(e)p(x|e)p(y|e)$$

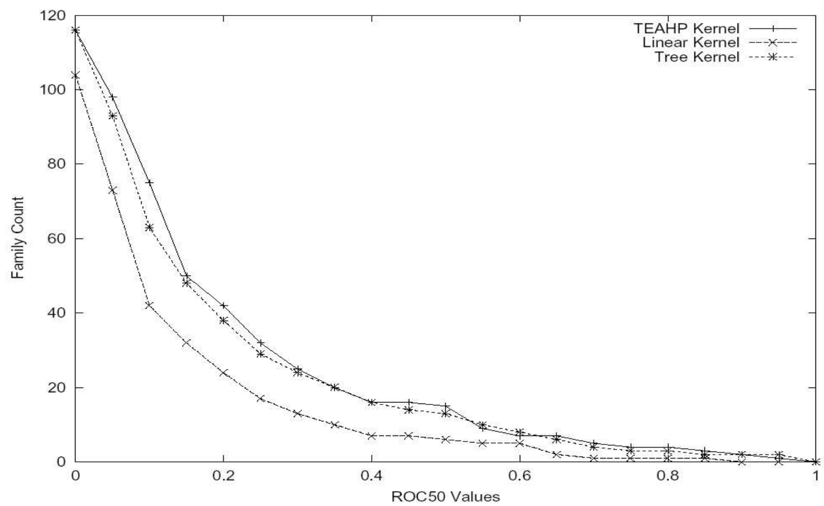
  - Can be proved to be a kernel
  - Intuition: two profiles get closer in the feature space when they have shared common evolution patterns with high probability.

Tree-Encoded Profile (Narra & Liao, 2004)



0 1 0.33 0.5 0.67 0.75 0.34 0.55

Tree Encoded Ad-hoc Hamming
Polynomial kernel



References and resources

- Cristianini & Shawe-Tayor, "*An introduction to Support Vector Machines*", Cambridge University Press, 2000.
- www.kernel-machines.org
  - SVMLight
  - Chris Burges, A tutorial
- J.-P Vert, A 3-day tutorial
- W. Noble, "Support vector machine applications in computational biology", *Kernel Methods in Computational Biology*. B. Schoelkopf, K. Tsuda and J.-P. Vert, ed. MIT Press, 2004.