Proteins are collections of “modular” domains. For example,

Coagulation Factor XII

PLAT
Modular Nature of Protein Structures

Example: Diphtheria Toxin

- Transmembrane domain
- Myoglobin
- Cellulose-binding domain
- Receptor-binding domain
- Catalytic domain
- Exotoxin

7/21/10
Domain Architecture Tools

- **CDART**
  - Protein [AAH24495](#); Domain Architecture;
  - It’s domain relatives;
  - Multiple alignment for 2\textsuperscript{nd} domain

- **SMART**
Active sites in proteins are usually hydrophobic pockets/crevices/troughs that involve sidechain atoms.
Left PDB 3RTD (streptavidin) and the first site located by the MOE Site Finder. Middle 3RTD with complexed ligand (biotin). Right Biotin ligand overlaid with calculated alpha spheres of the first site.
Secondary Structure Prediction Software

Figure 11.3 Comparison of secondary structure predictions by various methods. The sequence of flavodoxin, an αβ protein, was used as the query and is shown on the first line of the alignment. For each prediction, H denotes an α helix, E a β strand, T a β turn; all other positions are assumed to be random coil. Correctly assigned residues are shown in inverse type. The methods used are listed along the left side of the alignment and are described in the text. At the bottom of the figure is the secondary structure assignment given in the PDB title for flavodoxin (1OFV, Smith et al., 1983).
PDB: Protein Data Bank

- Database of protein tertiary and quaternary structures and protein complexes. [http://www.rcsb.org/pdb/](http://www.rcsb.org/pdb/)

- Over 29,000 structures as of Feb 1, 2005.

- Structures determined by
  - NMR Spectroscopy
  - X-ray crystallography
  - Computational prediction methods

- Sample PDB file: Click here [▪️]
Protein Folding

- Unfolded
  - Rapid (< 1s)
- Molten Globule State
  - Slow (1 - 1000 s)
- Folded Native State

How to find minimum energy configuration?
Protein Structures

- Most proteins have a **hydrophobic core**.
- Within the core, specific **interactions** take place between amino acid side chains.
- Can an amino acid be replaced by some other amino acid?
  - Limited by space and available contacts with nearby amino acids
- Outside the core, proteins are composed of loops and structural elements in contact with water, solvent, other proteins and other structures.
Viewing Protein Structures

- SPDBV
- RASMOL
- CHIME
Secondary Structure Prediction Software

Recent Ones:
GOR V
PREDATOR
Zpred
PROF
NNSSP
PHD
PSIPRED
Jnet
### Chou & Fasman Propensities

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>helix</th>
<th>Designation</th>
<th>P</th>
<th>strand</th>
<th>Designation</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>F</td>
<td>1.42</td>
<td></td>
<td></td>
<td>b</td>
<td>0.83</td>
</tr>
<tr>
<td>Cys</td>
<td>l</td>
<td>0.70</td>
<td></td>
<td></td>
<td>f</td>
<td>1.19</td>
</tr>
<tr>
<td>Asp</td>
<td>l</td>
<td>1.01</td>
<td></td>
<td></td>
<td>B</td>
<td>0.54</td>
</tr>
<tr>
<td>Glu</td>
<td>F</td>
<td>1.51</td>
<td></td>
<td></td>
<td>B</td>
<td>0.37</td>
</tr>
<tr>
<td>Phe</td>
<td>f</td>
<td>1.13</td>
<td></td>
<td></td>
<td>f</td>
<td>1.38</td>
</tr>
<tr>
<td>Gly</td>
<td>B</td>
<td>0.61</td>
<td></td>
<td></td>
<td>b</td>
<td>0.75</td>
</tr>
<tr>
<td>His</td>
<td>f</td>
<td>1.00</td>
<td></td>
<td></td>
<td>f</td>
<td>0.87</td>
</tr>
<tr>
<td>Ile</td>
<td>f</td>
<td>1.08</td>
<td></td>
<td></td>
<td>F</td>
<td>1.60</td>
</tr>
<tr>
<td>Lys</td>
<td>f</td>
<td>1.16</td>
<td></td>
<td></td>
<td>b</td>
<td>0.74</td>
</tr>
<tr>
<td>Leu</td>
<td>F</td>
<td>1.21</td>
<td></td>
<td></td>
<td>f</td>
<td>1.30</td>
</tr>
<tr>
<td>Met</td>
<td>F</td>
<td>1.45</td>
<td></td>
<td></td>
<td>f</td>
<td>1.05</td>
</tr>
<tr>
<td>Asn</td>
<td>b</td>
<td>0.67</td>
<td></td>
<td></td>
<td>b</td>
<td>0.89</td>
</tr>
<tr>
<td>Pro</td>
<td>B</td>
<td>0.57</td>
<td></td>
<td></td>
<td>B</td>
<td>0.55</td>
</tr>
<tr>
<td>Gln</td>
<td>f</td>
<td>1.11</td>
<td></td>
<td></td>
<td>h</td>
<td>1.10</td>
</tr>
<tr>
<td>Arg</td>
<td>l</td>
<td>0.98</td>
<td></td>
<td></td>
<td>l</td>
<td>0.93</td>
</tr>
<tr>
<td>Ser</td>
<td>l</td>
<td>0.77</td>
<td></td>
<td></td>
<td>b</td>
<td>0.75</td>
</tr>
<tr>
<td>Thr</td>
<td>l</td>
<td>0.83</td>
<td></td>
<td></td>
<td>f</td>
<td>1.19</td>
</tr>
<tr>
<td>Val</td>
<td>f</td>
<td>1.06</td>
<td></td>
<td></td>
<td>F</td>
<td>1.70</td>
</tr>
<tr>
<td>Trp</td>
<td>f</td>
<td>1.08</td>
<td></td>
<td></td>
<td>f</td>
<td>1.37</td>
</tr>
<tr>
<td>Tyr</td>
<td>b</td>
<td>0.69</td>
<td></td>
<td></td>
<td>F</td>
<td>1.4</td>
</tr>
</tbody>
</table>
GOR IV prediction for 1bbc

sequence length: 108

GOR IV:
alpha helix (Hh) : 50 is 46.30%
beta sheet (Ee) : 18 is 16.67%
random coil (Cc) : 40 is 37.04%

Over 29,000 structures as of Feb 1, 2005.

Structures determined by
- NMR Spectroscopy
- X-ray crystallography
- Computational prediction methods

Sample PDB file: Click here [here]
PDB Search Results

1. X62
   - Solution structure of the LIM domain of carboxyl terminal LIM domain protein 1
   - Authors: Qin, X.R., Nagashima, T., Hayashi, F., Yokoyama, S.

2. X4K
   - Solution structure of LIM domain in LIM-protein 3
   - Authors: He, F., Muto, Y., Inoue, M., Kigawa, T., Shirouzu, M., Terada, T., Yokoyama,

3. X4L
   - Solution structure of LIM domain in Four and a half LIM domains protein 2
   - Authors: He, F., Muto, Y., Inoue, M., Kigawa, T., Shirouzu, M., Terada, T., Yokoyama,
Protein Folding

- Unfolded
  - Rapid (< 1s)
- Molten Globule State
  - Slow (1 - 1000 s)
- Folded Native State

- How to find minimum energy configuration?
Protein Folding

amino acid side chains

unfolded protein

FOLDING

binding site

folded protein
Most proteins have a **hydrophobic core**.

Within the core, specific **interactions** take place between amino acid side chains.

Can an amino acid be replaced by some other amino acid?

- Limited by space and available contacts with nearby amino acids

Outside the core, proteins are composed of loops and structural elements in contact with water, solvent, other proteins and other structures.
Viewing Protein Structures

- SPDBV
- RASMOL
- CHIME
What is structural alignment of proteins?

- 3-d superimposition of the atoms as “best as possible”, i.e., to minimize RMSD (root mean square deviation).

Can be done using VAST and SARF

Structural similarity is common, even among proteins that do not share sequence similarity or evolutionary relationship.
Other databases & tools

- **MMDB** contains groups of structurally related proteins
- **SARF** structurally similar proteins using secondary structure elements
- **VAST** Structure Neighbors
- **SSAP** uses double dynamic programming to structurally align proteins
Protein Structure Prediction

- **Holy Grail** of bioinformatics

- **Protein Structure Initiative** to determine a set of protein structures that span protein structure space sufficiently well. **WHY?**
  - Number of folds in natural proteins is limited. Thus a newly discovered proteins should be within modeling distance of some protein in set.

- **CASP**: **Critical Assessment of techniques for structure prediction**
  - To stimulate work in this difficult field
PSP Methods

- **homology**-based modeling
- **methods based on fold recognition**
  - **Threading** methods
- **ab initio** methods
  - From first principles
  - With the help of databases
Best method for PSP

As proteins fold, a large number of partially folded, low-energy conformations are formed, and that local structures combine to form more global structures with minimum energy.

Build a database of known structures (I-sites) of short sequences (3-15 residues).

Monte Carlo simulation assembling possible substructures and computing energy
Modeling Servers

- SwissMODEL
- 3DJigsaw
- CPHModel
- ESyPred3D
- Geno3D
- SDSC1
- Rosetta
- MolIDE
- SCWRL
- PSIPred
- MODELLER
- LOOPY
Gel Electrophoresis for Protein

- Protein is also charged
- Has to be denatured - WHY
- Gel: SDS-Polyacrylamide gels
- Add sample to well
- Apply voltage
- Size determines speed
- Add dye to assess the speed
- Stain to see the protein bands
Protein Gel

1. Protein
2. Add SDS
   - SDS binds to amino acid residues and gives uniform negative charge to protein with heat the protein is linearized

3. Negative Electrode
   - Add Protein Sample onto SDS-PAGE
   - Gel Lane #2
   - (Protein Ladder is in Lane #1)

4. Electric Current
   - Protein Bands are separated by size

Copyright 2008 Molecular Station.com
2D-Gels

(a)
Separation in first dimension (by charge)

Protein mixture
Isoelectric focusing (IEF)

Apply first gel to top of second

pH 4.0
pH 10.0

pH 4.0
pH 10.0

Separation in second dimension (by size)

SDS electrophoresis

7/21/10
2D Gel Electrophoresis

Two-D Gels

First dimension
Isoelectric focusing
Decreasing pH

Isoelectric focusing gel is placed on SDS polyacrylamide gel.

Second dimension
SDS polyacrylamide gel electrophoresis
Decreasing pH
Decreasing Mr

(b)
Mass Spectrometry

- **Mass measurements By Time-of-Flight**: Pulses of light from laser ionizes protein that is absorbed on metal target. Electric field accelerates molecules in sample towards detector. The time to the detector is inversely proportional to the mass of the molecule. Simple conversion to mass gives the molecular weights of proteins and peptides.

- **Using Peptide Masses to Identify Proteins**: One powerful use of mass spectrometers is to identify a protein from its peptide mass fingerprint. A peptide mass fingerprint is a compilation of the molecular weights of peptides generated by a specific protease. The molecular weights of the parent protein prior to protease treatment and the subsequent proteolytic fragments are used to search genome databases for any similarly sized protein with identical or similar peptide mass maps. The increasing availability of genome sequences combined with this approach has almost eliminated the need to chemically sequence a protein to determine its amino acid sequence.
Mass Spectrometry
20 amino acids

How is it ordered?

Basis: Edman Degradation (Pehr Edman)
  - Limited ~30 residues
  - React with Phenylisothiocyanate
  - Cleave and chromatography

First separate the proteins – Use 2D gels
Then digest to get pieces
Then sequence the smaller pieces

Tedious

Mass spectrometry
Machine Learning

- Human Endeavor
  - Data → Information → Knowledge

- Machine Learning
  - Automatically extracting information from data

- Types of Machine Learning
  - Unsupervised
    - Clustering
    - Pattern Discovery
  - Supervised
    - Learning
    - Classification
Support Vector Machines

- Supervised Statistical Learning Method for:
  - Classification
  - Regression

- Simplest Version:
  - Training: Present series of labeled examples (e.g., gene expressions of tumor vs. normal cells)
  - Prediction: Predict labels of new examples.
Learning Problems
Learning Problems

- Binary Classification
- Multi-class classification
- Regression
SVM – Binary Classification

- Partition feature space with a surface.
- Surface is implied by a subset of the training points (vectors) near it. These vectors are referred to as **Support Vectors**.
- Efficient with high-dimensional data.
- Solid statistical theory
- Subsume several other methods.
Classification of 2-D (Separable) data
Classification of (Separable) 2-D data
Classification of (Separable) 2-D data

- Margin of a point
- Margin of a point set
Classification using the Separator

$\mathbf{w} \cdot \mathbf{x}_i + b > 0$

$\mathbf{w} \cdot \mathbf{x}_j + b < 0$
Perceptron Algorithm (Primal)

Given separable training set $S$ and learning rate $\eta > 0$

$\mathbf{w}_0 = 0$; // Weight
$\mathbf{b}_0 = 0$; // Bias

$k = 0; R = \max |x_i|

repeat
  for $i = 1$ to $N$
    if $y_i (\mathbf{w}_k \cdot x_i + b_k) \leq 0$ then
      $\mathbf{w}_{k+1} = \mathbf{w}_k + \eta y_i \mathbf{x}_i$
      $b_{k+1} = b_k + \eta y_i R^2$
      $k = k + 1$
  Until no mistakes made within loop

Return $k$, and $(\mathbf{w}_k, b_k)$ where $k = \#$ of mistakes

Rosenblatt, 1956
Theorem:

If margin $m$ of $S$ is positive, then

$$k \leq (2R/m)^2$$

i.e., the algorithm will always converge, and will converge quickly.
Non-linear Separators
Main idea: Map into feature space

Figure 2. The idea of SV machines: map the training data nonlinearly into a higher-dimensional feature space via \( \Phi \), and construct a separating hyperplane with maximum margin there. This yields a nonlinear decision boundary in input space. By the use of a kernel function, it is possible to compute the separating hyperplane without explicitly carrying out the map into the feature space.
Non-linear Separators
Useful URLs

- http://www.support-vector.net
Perceptron Algorithm (Primal)

Given separable training set $S$ and learning rate $\eta > 0$

$w_0 = 0; \quad // \text{Weight}$

$b_0 = 0; \quad // \text{Bias}$

$k = 0; \quad R = \max |x_i|$

repeat

for $i = 1$ to $N$

if $y_i (w_k \cdot x_i + b_k) \leq 0$ then

$w_{k+1} = w_k + \eta y_i x_i$

$b_{k+1} = b_k + \eta y_i R^2$

$k = k + 1$

Until no mistakes made within loop

Return $k$, and $(w_k, b_k)$ where $k =$ # of mistakes

Rosenblatt, 1956

$w = \Sigma a_i y_i x_i$
Perceptron Algorithm (Dual)

Given a separable training set $S$

$a = 0; b_0 = 0;$

$R = \max |x_i|$

repeat

for $i = 1$ to $N$

if $y_i (\sum a_j y_j x_i \cdot x_j + b) \leq 0$ then

$a_i = a_i + 1$

$b = b + y_i R^2$

endif

Until no mistakes made within loop

Return $(a, b)$
Perceptron Algorithm (Dual)

Given a separable training set $S$

$$a = 0; b_0 = 0;$$

$$R = \max \mid x_i \mid$$

repeat

for $i = 1$ to $N$

if $y_i \left( \sum a_j y_j \cdot (x_i, x_j) + b \right) \leq 0$ then

$$a_i = a_i + 1$$

$$b = b + y_i R^2$$

Until no mistakes made within loop

Return $(a, b)$

$\cdot (x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$
Different Kernel Functions

- Polynomial kernel
  \[ \kappa(X, Y) = (X \cdot Y)^d \]

- Radial Basis Kernel
  \[ \kappa(X, Y) = \exp\left(\frac{-\|X - Y\|^2}{2\sigma^2}\right) \]

- Sigmoid Kernel
  \[ \kappa(X, Y) = \tanh(\omega(X \cdot Y) + \theta) \]
SVM Ingredients

- Support Vectors
- Mapping from Input Space to Feature Space
- Dot Product - Kernel function
- Weights
Generalizations

- How to deal with **more than 2 classes**?
  
  **Idea:** Associate weight and bias for each class.

- How to deal with **non-linear separator**?
  
  **Idea:** Support Vector Machines.

- How to deal with **linear regression**?

- How to deal with **non-separable data**?
Applications

- **Text Categorization & Information Filtering**
  - 12,902 Reuters Stories, 118 categories (91% !!)

- **Image Recognition**
  - Face Detection, tumor anomalies, defective parts in assembly line, etc.

- **Gene Expression Analysis**

- **Protein Homology Detection**
<table>
<thead>
<tr>
<th>Class</th>
<th>Method</th>
<th>Learned threshold</th>
<th>Optimized threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FP</td>
<td>FN</td>
<td>TP</td>
</tr>
<tr>
<td>Tetracarboxylic acid</td>
<td>Radial SVM</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Dot-product-1 SVM</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Dot-product-2 SVM</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Dot-product-3 SVM</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Parzen</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>FLD</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>C4.5</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>MOCl</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>Respiration</td>
<td>Radial SVM</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Dot-product-1 SVM</td>
<td>21</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Dot-product-2 SVM</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Dot-product-3 SVM</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Parzen</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>FLD</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>C4.5</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>MOCl</td>
<td>12</td>
<td>26</td>
</tr>
<tr>
<td>Ribosome</td>
<td>Radial SVM</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Dot-product-1 SVM</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Dot-product-2 SVM</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Dot-product-3 SVM</td>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Parzen</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>FLD</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>C4.5</td>
<td>31</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>MOCl</td>
<td>26</td>
<td>26</td>
</tr>
</tbody>
</table>

Table 2: Comparison of error rates for various classification methods. Classes are as described in Table 1. The methods are the radial basis function SVM, the SVMs using the scaled dot product kernel raised to the first, second and third power, Parzen windows, Fisher's linear discriminant, and the two decision tree learners, C4.5 and MOCl. The next five columns are the false positive, false negative, true positive and true negative rates summed over three cross-validation splits, followed by the cost, which is the number of false positives plus twice the number of false negatives. These five columns are repeated twice, first using the threshold learned from the training set, and then using the threshold that minimizes the cost on the test set. The threshold optimization is not possible for the decision tree methods, since they do not produce ranked results.
<table>
<thead>
<tr>
<th>Class</th>
<th>Kernel</th>
<th>Cost for each split</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tricarboxylic acid</td>
<td>Radial</td>
<td>18 21 15 22 21 97</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>Dot-product-1</td>
<td>15 22 18 23 22 100</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-2</td>
<td>16 22 17 22 22 99</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-3</td>
<td>16 22 17 23 22 100</td>
<td></td>
</tr>
<tr>
<td>Respiration</td>
<td>Radial</td>
<td>16 18 23 20 16 93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-1</td>
<td>24 24 29 27 23 127</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-2</td>
<td>19 19 26 24 23 111</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-3</td>
<td>19 19 26 22 21 107</td>
<td></td>
</tr>
<tr>
<td>Ribosome</td>
<td>Radial</td>
<td>8 12 15 11 13 59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-1</td>
<td>13 18 14 16 16 77</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-2</td>
<td>11 16 14 16 15 72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-3</td>
<td>9 15 11 15 15 65</td>
<td></td>
</tr>
<tr>
<td>Proteasome</td>
<td>Radial</td>
<td>14 10 9 11 11 55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-1</td>
<td>16 12 12 17 19 76</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-2</td>
<td>16 13 15 17 17 78</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-3</td>
<td>16 13 16 16 17 79</td>
<td></td>
</tr>
<tr>
<td>Histone</td>
<td>Radial</td>
<td>4 4 4 4 4 20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-1</td>
<td>4 4 4 4 4 20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-2</td>
<td>4 4 4 4 4 20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-3</td>
<td>4 4 4 4 4 20</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Comparison of SVM performance using various kernels. For each of the MYGIE classifications, SVMs were trained using four different kernel functions on five different random three-fold splits of the data, training on two-thirds and testing on the remaining third. The first column contains the class, as described in Table 1. The second column contains the kernel function, as described in Table 2. The next five columns contain the threshold-optimized cost (i.e., the number of false positives plus twice the number of false negatives) for each of the five random three-fold splits. The final column is the total cost across all five splits.

<table>
<thead>
<tr>
<th>Family</th>
<th>Gene</th>
<th>Locus</th>
<th>Error</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCA</td>
<td>YPR001W</td>
<td>CIT3</td>
<td>FN</td>
<td>mitochondrial citrate synthase</td>
</tr>
<tr>
<td>YOR142W</td>
<td>LSC1</td>
<td>FN</td>
<td>α subunit of succinyl-CoA ligase</td>
<td></td>
</tr>
<tr>
<td>YNR001C</td>
<td>CIT1</td>
<td>FN</td>
<td>mitochondrial citrate synthase</td>
<td></td>
</tr>
<tr>
<td>YLR174W</td>
<td>IDP2</td>
<td>FN</td>
<td>isocitrate dehydrogenase</td>
<td></td>
</tr>
<tr>
<td>YIL125W</td>
<td>KGD1</td>
<td>FN</td>
<td>α-ketoglutarate dehydrogenase</td>
<td></td>
</tr>
<tr>
<td>YDR148C</td>
<td>KGD2</td>
<td>FN</td>
<td>component of α-ketoglutarate dehydrogenase complex in mitochondria</td>
<td></td>
</tr>
<tr>
<td>YDL066W</td>
<td>IDP1</td>
<td>FN</td>
<td>mitochondrial form of isocitrate dehydrogenase</td>
<td></td>
</tr>
<tr>
<td>YBL015W</td>
<td>ACH1</td>
<td>FP</td>
<td>acetyl CoA hydrolase</td>
<td></td>
</tr>
<tr>
<td>Resp</td>
<td>YPR191W</td>
<td>QCR2</td>
<td>FN</td>
<td>ubiquinol cytochrome-c reductase core protein 2</td>
</tr>
<tr>
<td>YPL271W</td>
<td>ATP15</td>
<td>FN</td>
<td>ATP synthase epsilon subunit</td>
<td></td>
</tr>
<tr>
<td>YPL262W</td>
<td>FUM1</td>
<td>FP</td>
<td>fumarase</td>
<td></td>
</tr>
<tr>
<td>YML120C</td>
<td>ND1</td>
<td>FP</td>
<td>mitochondrial NADH ubiquinone 6 oxidoreductase</td>
<td></td>
</tr>
<tr>
<td>YKL085W</td>
<td>MDH1</td>
<td>FP</td>
<td>mitochondrial malate dehydrogenase</td>
<td></td>
</tr>
<tr>
<td>YDL057C</td>
<td>COX9</td>
<td>FN</td>
<td>subunit Vila of cytochrome c oxidase</td>
<td></td>
</tr>
<tr>
<td>Ribo</td>
<td>YPL037C</td>
<td>EGD1</td>
<td>FP</td>
<td>β subunit of the nascent-polypeptide-associated complex (NAC)</td>
</tr>
<tr>
<td>YLR040C</td>
<td>RPL31B</td>
<td>FN</td>
<td>ribosomal protein L31B (L34B) (YL28)</td>
<td></td>
</tr>
<tr>
<td>YLR075W</td>
<td>RPL10</td>
<td>FP</td>
<td>ribosomal protein L10</td>
<td></td>
</tr>
<tr>
<td>YAL003W</td>
<td>EFB1</td>
<td>FP</td>
<td>translation elongation factor EF-1β</td>
<td></td>
</tr>
<tr>
<td>Prot</td>
<td>YHR027C</td>
<td>RPN1</td>
<td>FN</td>
<td>subunit of 26S proteasome (PA700 subunit)</td>
</tr>
<tr>
<td>YGR270W</td>
<td>VTA7</td>
<td>FN</td>
<td>member of CDC48/PAS1/SEC18 family of ATPases</td>
<td></td>
</tr>
<tr>
<td>YGR048W</td>
<td>UFD1</td>
<td>FP</td>
<td>ubiquitin fusion degradation protein</td>
<td></td>
</tr>
<tr>
<td>YDR099C</td>
<td>DOA4</td>
<td>FN</td>
<td>ubiquitin isopeptidase</td>
<td></td>
</tr>
<tr>
<td>YDL020C</td>
<td>RPN4</td>
<td>FN</td>
<td>involved in ubiquitin degradation pathway</td>
<td></td>
</tr>
<tr>
<td>Hat</td>
<td>YOL012C</td>
<td>HTA3</td>
<td>FN</td>
<td>histone-related protein</td>
</tr>
<tr>
<td>YKL049C</td>
<td>CSN4</td>
<td>FN</td>
<td>required for proper kinetochore function</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Consistently misclassified genes. The table lists all 25 genes that are consistently misclassified by SVMs trained using the MYGIE classifications listed in Table 1. Two types of errors are included: a false positive (FP) occurs when the SVM includes the gene in the given class but the MYGIE classification does not; a false negative (FN) occurs when the SVM does not include the gene in the given class but the MYGIE classification does.
Table 1: Error rates for ovarian cancer tissue experiments. For each setting of the SVM consisting of a kernel and diagonal factor (DF), each tissue was classified. Column 2 is the number of features (genes) used. Reported are the number of normal tissues misclassified (FP), tumor tissues misclassified (FN), tumor tissues classified correctly (TP), and normal tissues classified correctly (TN).

Table 5: Results for the perceptron on all data sets. The results are averaged over 5 shufflings of the data as this algorithm is sensitive to the order in which it receives the data points. The first column is the dataset used and the second is number of features in the dataset. For the ovarian and colon datasets, the number of normal tissues misclassified (FP) and the number of tumor tissues misclassified (FN) is reported. For the AML/ALL training dataset, the number of AML samples misclassified (FP) and the number of ALL patients misclassified (FN) is reported. For the AML treatment dataset, the number of unsuccessfully treated patients misclassified (FP) and the number of successfully treated patients misclassified (FN) is reported. The last two columns report the best score obtained by the SVM on that dataset.

Figure 1: SVM classification margins for ovarian tissues. When classifying, the SVM calculates a margin which is the distance of an example from the decision boundary it has learned. In this graph, the margin for each tissue sample calculated using (10) is shown. A positive value indicates a correct classification, and a negative value indicates an incorrect classification. The most negative point corresponds to tissue 1023. The second most negative point corresponds to tissue HWBC3.