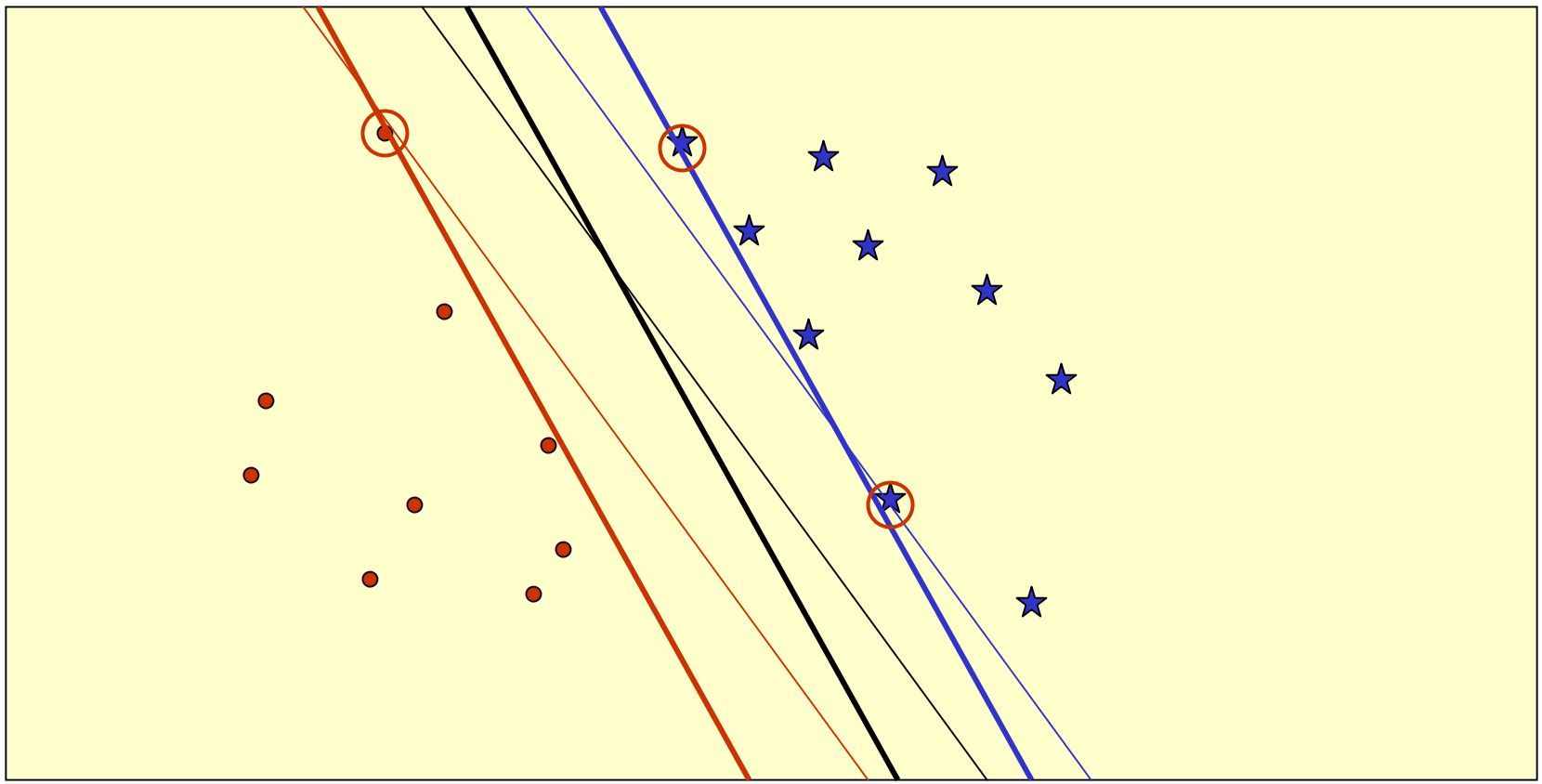
# Support Vector Machines

- Supervised Statistical Learning Method for:
  - Classification (Binary/Multi-class)
  - Regression
- Simplest Version:
  - Training: Present series of labeled examples (e.g., gene expressions of tumor vs. normal cells)
  - Prediction: Predict labels of new examples.

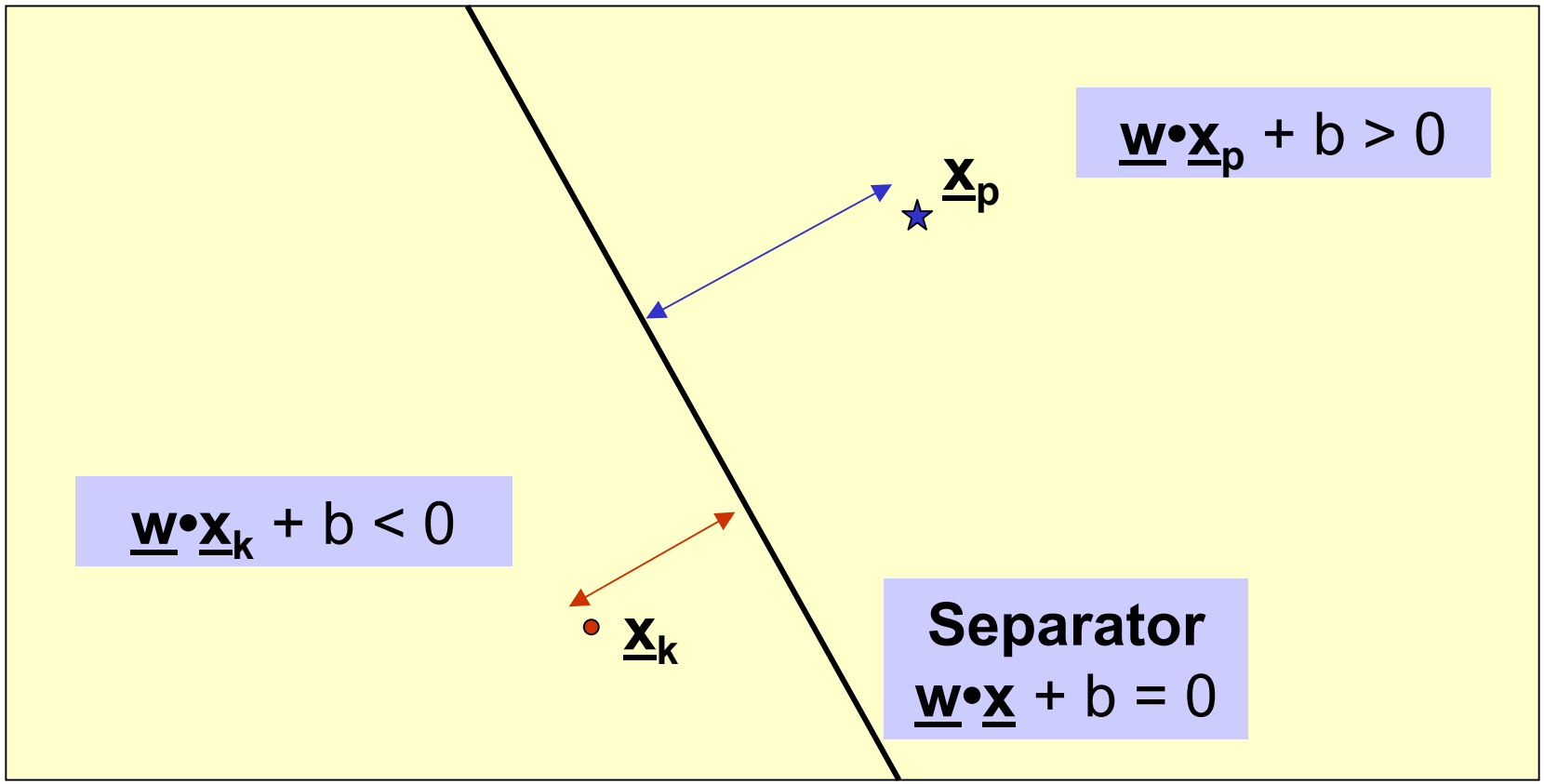
# SVM – Binary Classification

- 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 using the Separator



# Perceptron Algorithm (Primal)

Rosenblatt, 1956

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

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

$b_0 = 0$; // Bias

$k = 0$; $R = \max \|\underline{x}_i\|$

$\underline{\mathbf{w}} = \Sigma\, a_i y_i \underline{\mathbf{x}}_i$

**repeat**

  **for** i = 1 to N

    **if** $y_i (\underline{\mathbf{w}}_k \cdot \underline{x}_i + b_k) \leq 0$ **then**

      $\underline{\mathbf{w}}_{k+1} = \underline{\mathbf{w}}_k + \eta y_i \underline{\mathbf{x}}_i$

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

      $k = k + 1$

**Until** no mistakes made within loop

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

# Performance for Separable Data

**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.

# Perceptron Algorithm (Dual)

Given a separable training set S

$\underline{\mathbf{a}} = \underline{0}$; $b_0 = 0$;

$R = \max \| \underline{x}_i \|$

**repeat**

  **for** i = 1 to N

  **if** $y_i \left(\sum a_j y_j \, \underline{x}_i \bullet \underline{x}_j + b\right) \leq 0$ **then**

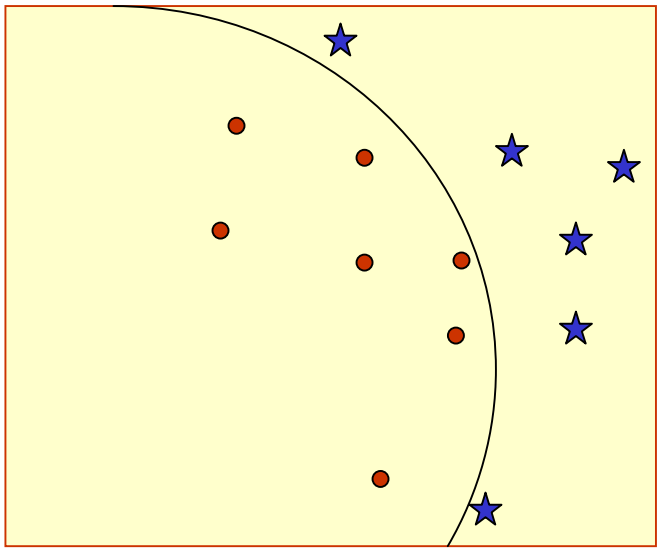
    $a_i = a_i + 1$

    $b = b + y_i R^2$

**Until** no mistakes made within loop

**Return** ($\underline{\mathbf{a}}$, b)

# Non-linear Separators



# Main idea: Map into feature space

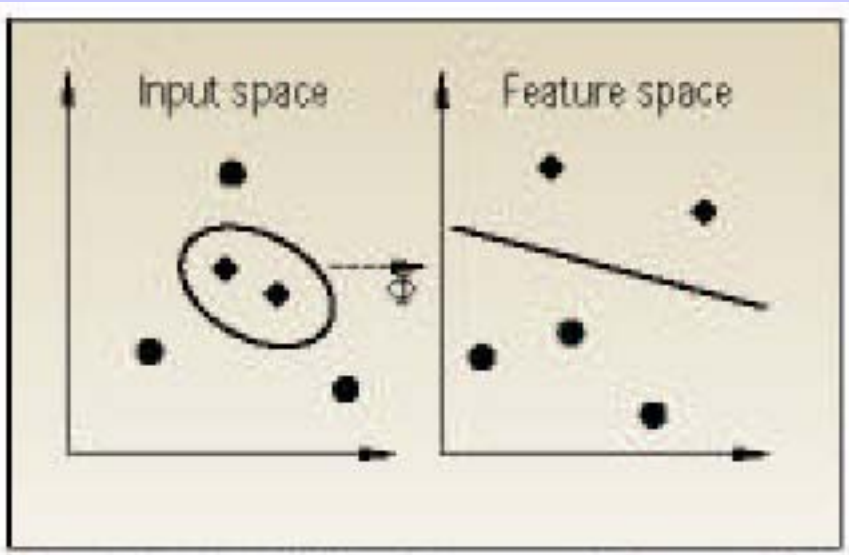
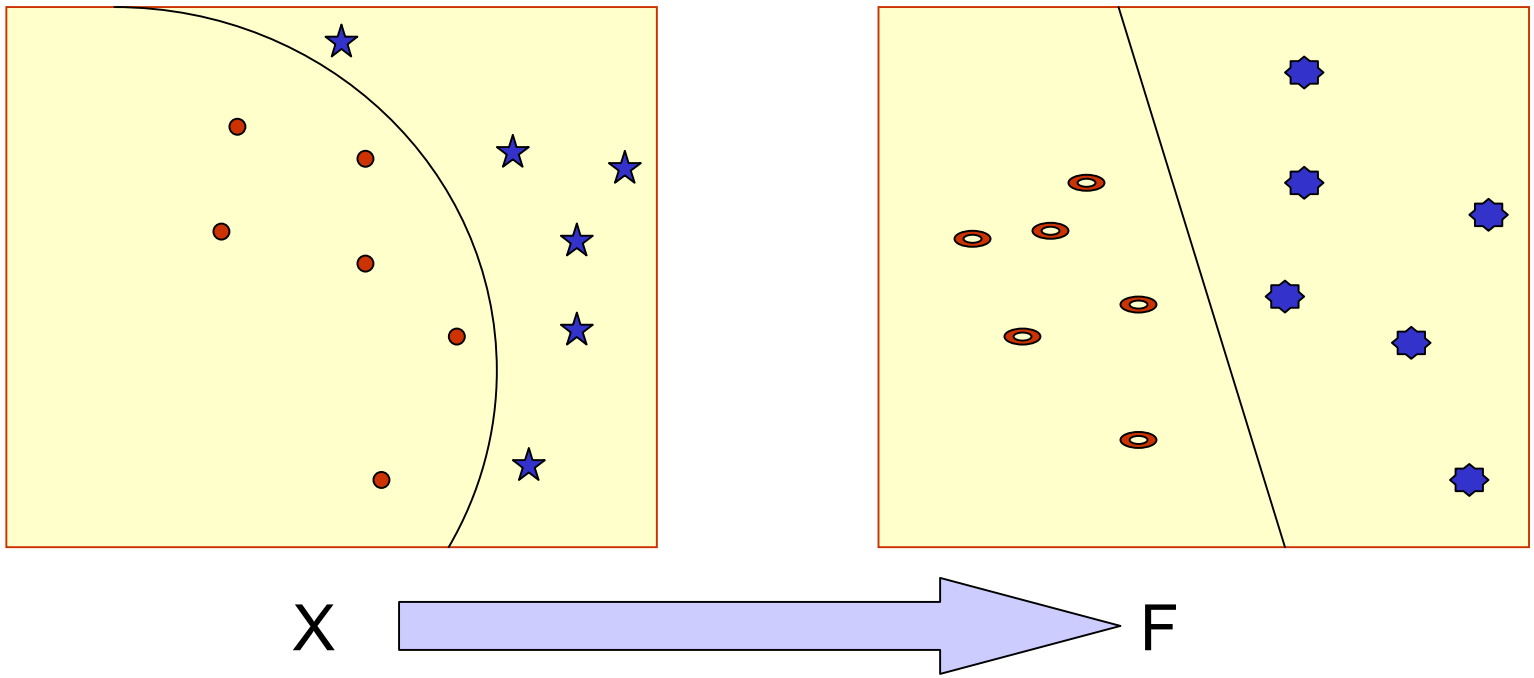



Figure 2. The idea of SV machines: map the training data nonlinearly into a higher-dimensional feature space via Φ, and construct a separating hyperplane with maximum margin there. This yields a nonlinear decision boundary in input space. By the use of a kernel function, it is possible to compute the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

# Non-linear Separators



# Perceptron Algorithm (Dual)

Given a separable training set S

$\underline{\mathbf{a}} = \underline{0}$; $b_0 = 0$;

$R = \max \|\underline{x}_i\|$

**repeat**

  **for** i = 1 to N

    **if** $y_i \left(\sum a_j y_j \, \kappa(\underline{x}_i, \underline{x}_j) + b\right) \leq 0$ **then**

      $a_i = a_i + 1$

      $b = b + y_i R^2$

**Until** no mistakes made within loop

**Return** ($\underline{\mathbf{a}}$, b)

$\kappa(\underline{x}_i, \underline{x}_j) = \Phi(\underline{x}_i) \bullet \Phi(\underline{x}_j)$

# Different Kernel Functions

- Polynomial kernel $\kappa(X,Y) = (X \bullet Y)^d$
- Radial Basis Kernel $\kappa(X,Y) = \exp\left(\frac{-\|X-Y\|^2}{2\sigma^2}\right)$
- Sigmoid Kernel $\kappa(X,Y) = \tanh(\omega(X \bullet Y) + \theta)$

# SVM Ingredients

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

# 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

| Class | Method | Learned threshold FP | FN | TP | TN | Cost | Optimized threshold FP | FN | TP | TN | Cost |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Tricarboxylic acid | Radial SVM | 8 | 8 | 9 | 2442 | 24 | 4 | 7 | 10 | 2446 | 18 |
| | Dot-product-1 SVM | 11 | 9 | 8 | 2439 | 29 | 3 | 6 | 11 | 2447 | 15 |
| | Dot-product-2 SVM | 5 | 10 | 7 | 2445 | 25 | 4 | 6 | 11 | 2446 | 16 |
| | Dot-product-3 SVM | 4 | 12 | 5 | 2446 | 28 | 4 | 6 | 11 | 2446 | 16 |
| | Parzen | 4 | 12 | 5 | 2446 | 28 | 0 | 12 | 5 | 2450 | 24 |
| | FLD | 9 | 10 | 7 | 2441 | 29 | 7 | 8 | 9 | 2443 | 23 |
| | C4.5 | 7 | 17 | 0 | 2443 | 41 | – | – | – | – | – |
| | MOC1 | 3 | 16 | 1 | 2446 | 35 | – | – | – | – | – |
| Respiration | Radial SVM | 9 | 6 | 24 | 2428 | 21 | 8 | 4 | 26 | 2429 | 16 |
| | Dot-product-1 SVM | 21 | 10 | 20 | 2416 | 41 | 6 | 9 | 21 | 2431 | 24 |
| | Dot-product-2 SVM | 7 | 14 | 16 | 2430 | 35 | 7 | 6 | 24 | 2430 | 19 |
| | Dot-product-3 SVM | 3 | 15 | 15 | 2434 | 33 | 7 | 6 | 24 | 2430 | 19 |
| | Parzen | 22 | 10 | 20 | 2415 | 42 | 7 | 12 | 18 | 2430 | 31 |
| | FLD | 10 | 10 | 20 | 2427 | 30 | 14 | 4 | 26 | 2423 | 22 |
| | C4.5 | 18 | 17 | 13 | 2419 | 52 | – | – | – | – | – |
| | MOC1 | 12 | 26 | 4 | 2425 | 64 | – | – | – | – | – |
| Ribosome | Radial SVM | 9 | 4 | 117 | 2337 | 17 | 6 | 1 | 120 | 2340 | 8 |
| | Dot-product-1 SVM | 13 | 6 | 115 | 2333 | 25 | 11 | 1 | 120 | 2335 | 13 |
| | Dot-product-2 SVM | 7 | 10 | 111 | 2339 | 27 | 9 | 1 | 120 | 2337 | 11 |
| | Dot-product-3 SVM | 3 | 18 | 103 | 2343 | 39 | 7 | 1 | 120 | 2339 | 9 |
| | Parzen | 6 | 8 | 113 | 2340 | 22 | 5 | 8 | 113 | 2341 | 21 |
| | FLD | 15 | 5 | 116 | 2331 | 25 | 8 | 3 | 118 | 2338 | 14 |
| | C4.5 | 31 | 21 | 100 | 2315 | 73 | – | – | – | – | – |
| | MOC1 | 26 | 26 | 95 | 2320 | 78 | – | – | – | – | – |

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 MOC1. 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.



| Class | Method | Learned threshold FP | FN | TP | TN | Cost | Optimized threshold FP | FN | TP | TN | Cost |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Proteasome | Radial SVM | 3 | 7 | 28 | 2429 | 17 | 4 | 5 | 30 | 2428 | 14 |
| | Dot-product-1 SVM | 14 | 11 | 24 | 2418 | 36 | 2 | 7 | 28 | 2430 | 16 |
| | Dot-product-2 SVM | 4 | 13 | 22 | 2428 | 30 | 4 | 6 | 29 | 2428 | 16 |
| | Dot-product-3 SVM | 3 | 18 | 17 | 2429 | 39 | 2 | 7 | 28 | 2430 | 16 |
| | Parzen | 21 | 5 | 30 | 2411 | 31 | 3 | 9 | 26 | 2429 | 21 |
| | FLD | 7 | 12 | 23 | 2425 | 31 | 12 | 7 | 28 | 2420 | 26 |
| | C4.5 | 17 | 10 | 25 | 2415 | 37 | – | – | – | – | – |
| | MOC1 | 10 | 17 | 18 | 2422 | 44 | – | – | – | – | – |
| Histone | Radial SVM | 0 | 2 | 9 | 2456 | 4 | 0 | 2 | 9 | 2456 | 4 |
| | Dot-product-1 SVM | 0 | 4 | 7 | 2456 | 8 | 0 | 2 | 9 | 2456 | 4 |
| | Dot-product-2 SVM | 0 | 5 | 6 | 2456 | 10 | 0 | 2 | 9 | 2456 | 4 |
| | Dot-product-3 SVM | 0 | 8 | 3 | 2456 | 16 | 0 | 2 | 9 | 2456 | 4 |
| | Parzen | 2 | 3 | 8 | 2454 | 8 | 1 | 3 | 8 | 2455 | 7 |
| | FLD | 0 | 3 | 8 | 2456 | 6 | 2 | 1 | 10 | 2454 | 4 |
| | C4.5 | 2 | 2 | 9 | 2454 | 6 | – | – | – | – | – |
| | MOC1 | 2 | 5 | 6 | 2454 | 12 | – | – | – | – | – |
| Helix-turn-helix | Radial SVM | 1 | 16 | 0 | 2450 | 33 | 0 | 16 | 0 | 2451 | 32 |
| | Dot-product-1 SVM | 20 | 16 | 0 | 2431 | 52 | 0 | 16 | 0 | 2451 | 32 |
| | Dot-product-2 SVM | 4 | 16 | 0 | 2447 | 36 | 0 | 16 | 0 | 2451 | 32 |
| | Dot-product-3 SVM | 1 | 16 | 0 | 2450 | 33 | 0 | 16 | 0 | 2451 | 32 |
| | Parzen | 14 | 16 | 0 | 2437 | 46 | 0 | 16 | 0 | 2451 | 32 |
| | FLD | 14 | 16 | 0 | 2437 | 46 | 0 | 16 | 0 | 2451 | 32 |
| | C4.5 | 2 | 16 | 0 | 2449 | 34 | – | – | – | – | – |
| | MOC1 | 6 | 16 | 0 | 2445 | 38 | – | – | – | – | – |

Table 3: **Comparison of error rates for various classification methods (continued).** See caption for Table 2.

| Class | Kernel | Cost for each split | | | | | Total |
|---|---|---|---|---|---|---|---|
| Tricarboxylic acid | Radial | 18 | 21 | 15 | 22 | 21 | 97 |
| | Dot-product-1 | 15 | 22 | 18 | 23 | 22 | 100 |
| | Dot-product-2 | 16 | 22 | 17 | 22 | 22 | 99 |
| | Dot-product-3 | 16 | 22 | 17 | 23 | 22 | 100 |
| Respiration | Radial | 16 | 18 | 23 | 20 | 16 | 93 |
| | Dot-product-1 | 24 | 24 | 29 | 27 | 23 | 127 |
| | Dot-product-2 | 19 | 19 | 26 | 24 | 23 | 111 |
| | Dot-product-3 | 19 | 19 | 26 | 22 | 21 | 107 |
| Ribosome | Radial | 8 | 12 | 15 | 11 | 13 | 59 |
| | Dot-product-1 | 13 | 18 | 14 | 16 | 16 | 77 |
| | Dot-product-2 | 11 | 16 | 14 | 16 | 15 | 72 |
| | Dot-product-3 | 9 | 15 | 11 | 15 | 15 | 65 |
| Proteasome | Radial | 14 | 10 | 9 | 11 | 11 | 55 |
| | Dot-product-1 | 16 | 12 | 12 | 17 | 19 | 76 |
| | Dot-product-2 | 16 | 13 | 15 | 17 | 17 | 78 |
| | Dot-product-3 | 16 | 13 | 16 | 16 | 17 | 79 |
| Histone | Radial | 4 | 4 | 4 | 4 | 4 | 20 |
| | Dot-product-1 | 4 | 4 | 4 | 4 | 4 | 20 |
| | Dot-product-2 | 4 | 4 | 4 | 4 | 4 | 20 |
| | Dot-product-3 | 4 | 4 | 4 | 4 | 4 | 20 |

Table 4: **Comparison of SVM performance using various kernels.** For each of the MYGD 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.



| Family | Gene | Locus | Error | Description |
|---|---|---|---|---|
| TCA | YPR001W | CIT3 | FN | mitochondrial citrate synthase |
| | YOR142W | LSC1 | FN | $\alpha$ subunit of succinyl-CoA ligase |
| | YNR001C | CIT1 | FN | mitochondrial citrate synthase |
| | YLR174W | IDP2 | FN | isocitrate dehydrogenase |
| | YIL125W | KGD1 | FN | $\alpha$-ketoglutarate dehydrogenase |
| | YDR148C | KGD2 | FN | component of $\alpha$-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 | $\beta$ subunit of the nascent-polypeptide-associated complex (NAC) |
| | YLR406C | RPL31B | FN | ribosomal protein L31B (L34B) (YL28) |
| | YLR075W | RPL10 | FP | ribosomal protein L10 |
| | YAL003W | EFB1 | FP | translation elongation factor EF-1$\beta$ |
| 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 |

Table 6: **Consistently misclassified genes.** The table lists all 25 genes that are consistently misclassified by SVMs trained using the MYGD 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 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.

| Kernel DF | Feature | FP | FN | TP | TN |
|---|---|---|---|---|---|
| dot-product 0 | 25 | 5 | 4 | 10 | 12 |
| dot-product 2 | 25 | 5 | 2 | 12 | 12 |
| dot-product 5 | 25 | 4 | 2 | 12 | 13 |
| dot-product 10 | 25 | 4 | 2 | 12 | 13 |
| dot-product 0 | 50 | 4 | 2 | 12 | 13 |
| dot-product 2 | 50 | 3 | 2 | 12 | 14 |
| dot-product 5 | 50 | 3 | 2 | 12 | 14 |
| dot-product 10 | 50 | 3 | 2 | 12 | 14 |
| dot-product 0 | 100 | 4 | 3 | 11 | 13 |
| dot-product 2 | 100 | 5 | 3 | 11 | 12 |
| dot-product 5 | 100 | 5 | 3 | 11 | 12 |
| dot-product 10 | 100 | 5 | 3 | 11 | 12 |
| dot-product 0 | 500 | 5 | 3 | 11 | 12 |
| dot-product 2 | 500 | 4 | 3 | 11 | 13 |
| dot-product 5 | 500 | 4 | 3 | 11 | 13 |
| dot-product 10 | 500 | 4 | 3 | 11 | 13 |
| dot-product 0 | 1000 | 7 | 3 | 11 | 10 |
| dot-product 2 | 1000 | 5 | 3 | 11 | 12 |
| dot-product 5 | 1000 | 5 | 3 | 11 | 12 |
| dot-product 10 | 1000 | 5 | 3 | 11 | 12 |
| dot-product 0 | 97802 | 17 | 0 | 14 | 0 |
| dot-product 2 | 97802 | 9 | 2 | 12 | 8 |
| dot-product 5 | 97802 | 7 | 3 | 11 | 10 |
| dot-product 10 | 97802 | 5 | 3 | 11 | 12 |

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 (clones) 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).

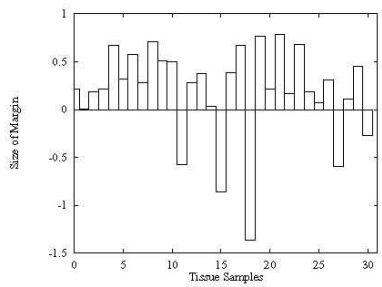
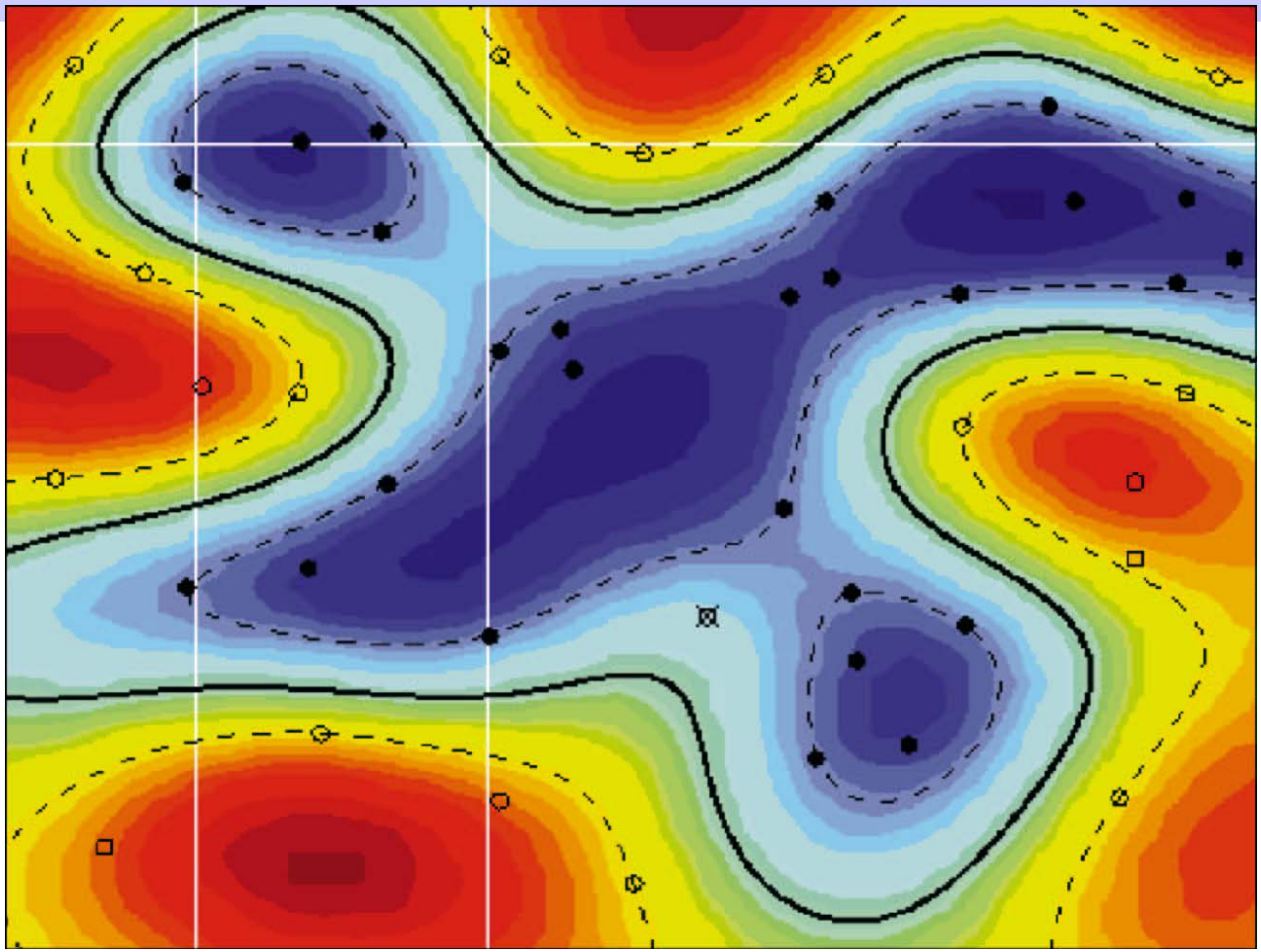



Figure 1: **SVM classification margins for ovarian tissues.** When classifying, the SVM calculates a margin which is the distance of an example from the decision boundary it has learned. In this graph, the margin for each tissue sample calculated using (10) is shown. A positive value indicates a correct classification, and a negative value indicates an incorrect classification. The most negative point corresponds to tissue N039. The second most negative point corresponds to tissue HWBC3.

| Dataset | Features | FP | FN | SVM FP | SVM FN |
|---|---|---|---|---|---|
| Ovarian(original) | 97802 | 4.6 | 4.8 | 5 | 3 |
| Ovarian(modified) | 97802 | 4.4 | 3.4 | 0 | 0 |
| AML/ALL train | 7129 | 0.6 | 2.8 | 0 | 0 |
| AML treatment | 7129 | 4.8 | 3.5 | 3 | 2 |
| Colon | 2000 | 3.8 | 3.7 | 3 | 3 |

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.

# SVM Example (Radial Basis Function)



# Useful URLs

- http://www.support-vector.net

# Sources of Variations & Errors in Microarray Data

- Variations in cells/individuals.
- Variations in mRNA extraction, isolation, introduction of dye, variation in dye incorporation, dye interference.
- Variations in probe concentration, probe amounts, substrate surface characteristics
- Variations in hybridization conditions and kinetics
- Variations in optical measurements, spot misalignments, discretization effects, noise due to scanner lens and laser irregularities
- Cross-hybridization of sequences with high sequence identity.
- Limit of factor 2 in precision of results.

Need to Normalize data

# Significance Analysis of Microarrays (SAM) [Tusher, Tibshirani, Chu, PNAS'01]

- Fold change is a typical measure to decide genes of interest.
- However, variations in gene expression are also gene dependent. If repeats are available, then such variations can be measured for each gene. This helps to give a better analysis of significant genes of interest.

# Genomics

- Study of all genes in a genome, or comparison of whole genomes.
  - Whole genome sequencing
  - Whole genome annotation & Functional genomics
  - Whole genome comparison
    - PipMaker: uses BLASTZ to compare very long sequences (> 2Mb); http://www.cse.psu.edu/pipmaker/
    - Mummer: used for comparing long microbial sequences (uses Suffix trees!)

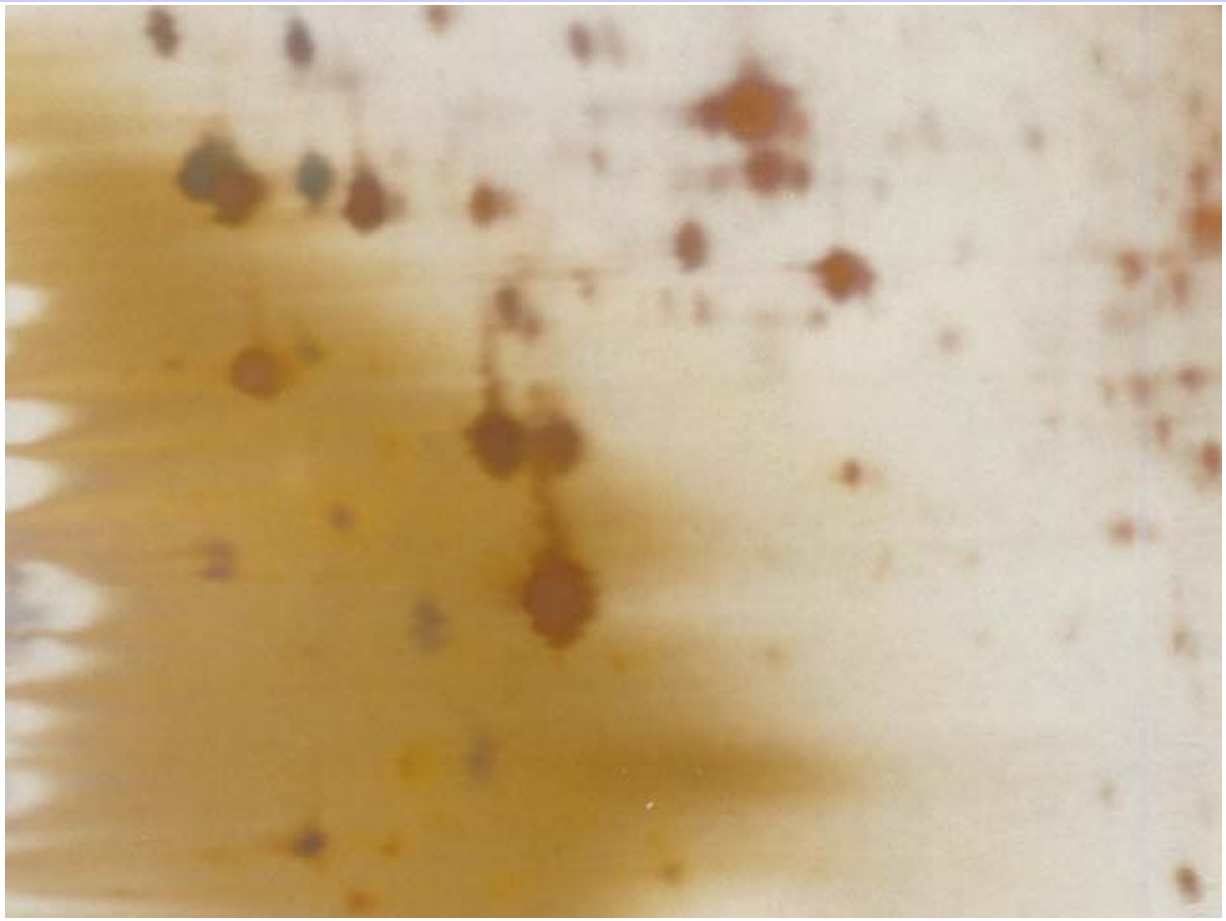
# Genomics (Cont'd)

- Gene Expression
  - Microarray experiments & analysis
    - Probe design (CODEHOP)
    - Array image analysis (CrazyQuant)
    - Identifying genes with significant changes (SAM)
    - Clustering

# Proteomics

- Study of all proteins in a genome, or comparison of whole genomes.
  - Whole genome annotation & Functional proteomics
  - Whole genome comparison
  - Protein Expression: 2D Gel Electrophoresis

# 2D Gel Electrophoresis



# Other Proteomics Tools

**From ExPASy/SWISS-PROT:**

- **AACompIdent** identify proteins from aa composition

[Input: aa composition, isoelectric point, mol wt., etc. Output: proteins from DB]

- **AACompSim** compares proteins aa composition with other proteins
- **MultIdent** uses mol wt., mass fingerprints, etc. to identify proteins
- **PeptIdent** compares experimentally determined mass fingerprints with theoretically determined ones for all proteins
- **FindMod** predicts post-translational modifications based on mass difference between experimental and theoretical mass fingerprints.
- **PeptideMass** theoretical mass fingerprint for a given protein.
- **GlycoMod** predicts oligosaccharide modifications from mass difference
- **TGREASE** calculates hydrophobicity of protein along its length

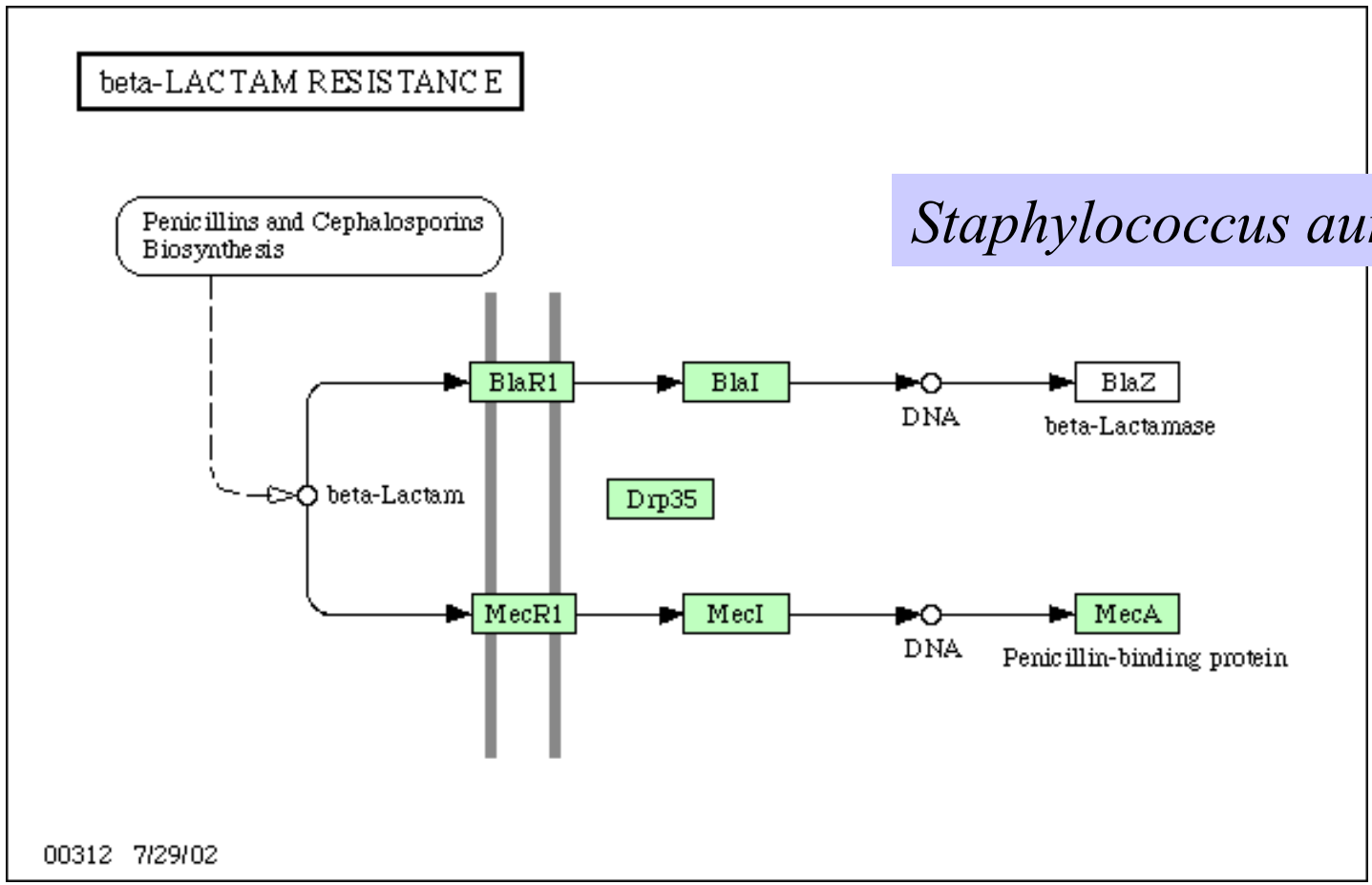
# Databases for Comparative Genomics

- PEDANT useful resource for standard questions in comparative genomics. For e.g., *how many known proteins in XXX have known 3-d structures, how many proteins from family YYY are in ZZZ, etc.*
- COGs Clusters of orthologous groups of proteins.
- MBGD Microbial genome database searches for homologs in all microbial genomes

# Gene Networks & Pathways

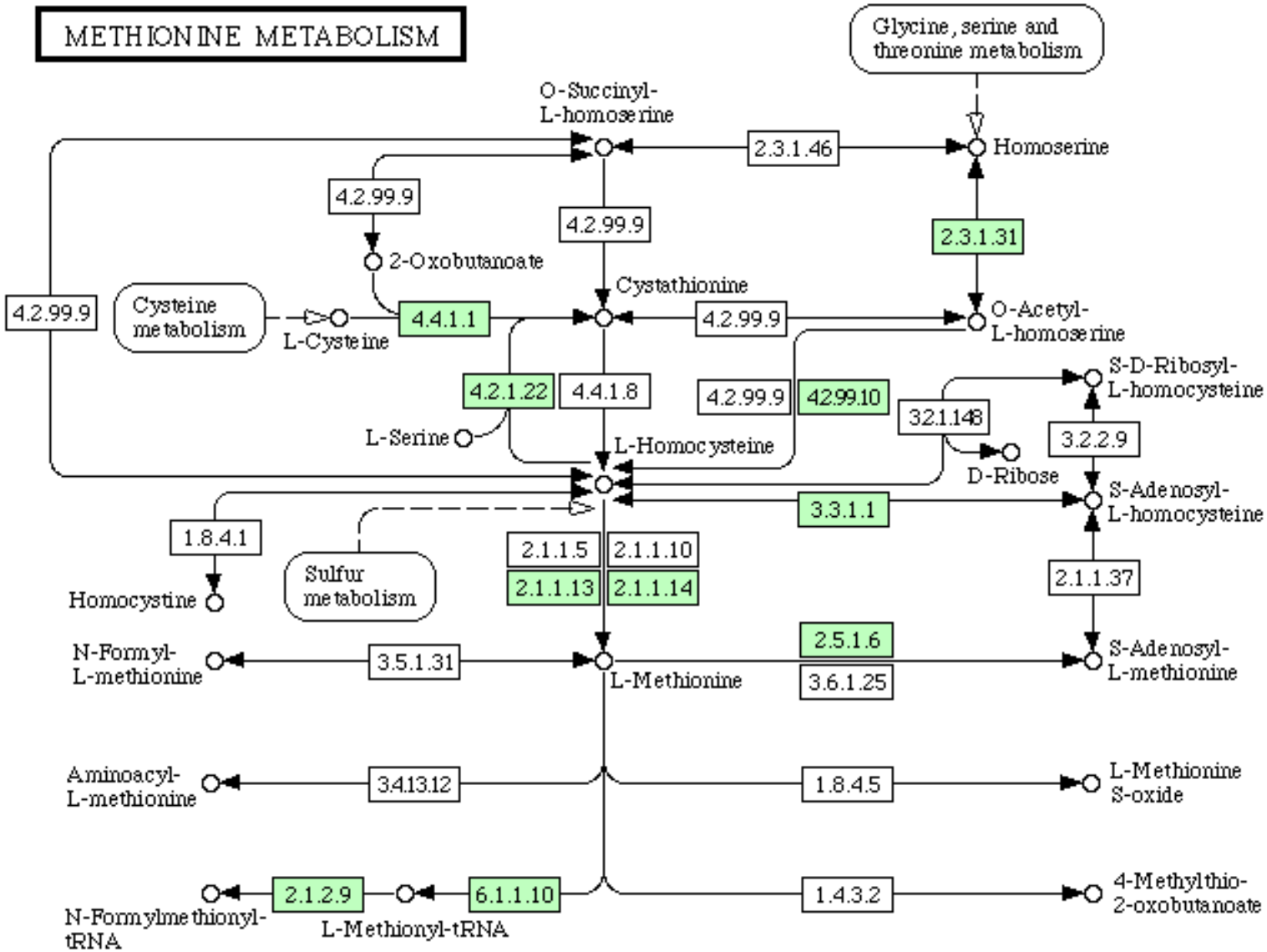
- Genes & Proteins act in concert and therefore form a complex network of dependencies.

# Pathway Example from KEGG




*Staphylococcus aureus*

# *Pseudomonas aeruginosa*



# STSs and ESTs

- Sequence-Tagged Site: short, unique sequence
- Expressed Sequence Tag: short, unique sequence from a coding region
  - 1991: 609 ESTs [Adams et al.]
  - June 2000: 4.6 million in dbEST
  - Genome sequencing center at St. Louis produce 20,000 ESTs per week.

# What Are ESTs and How Are They Made?

- Small pieces of DNA sequence (usually 200 - 500 nucleotides) of low quality.
- Extract mRNA from cells, tissues, or organs and sequence either end. Reverse transcribe to get cDNA (5' EST and 3'EST) and deposit in EST library.
- Used as "**tags**" or markers for that gene.
- Can be used to identify similar genes from other organisms (Complications: variations among organisms, variations in genome size, presence or absence of **introns**).
- 5' ESTs tend to be more useful (cross-species conservation), 3' EST often in UTR.

# DNA Markers

- Uniquely identifiable DNA segments.
- Short, <500 nucleotides.
- Layout of these markers give a map of genome.
- Markers may be polymorphic (variations among individuals). Polymorphism gives rise to alleles.
- Found by PCR assays.

# Polymorphisms

- Length polymorphisms
  - Variable # of tandem repeats (VNTR)
  - Microsatellites or short tandem repeats
  - Restriction fragment length polymorphism (RFLP) caused by changes in restriction sites.
- Single nucleotide polymorphism (SNP)
  - Average once every ~100 bases in humans
  - Usually biallelic
  - dbSNP database of SNPs (over 100,000 SNPs)
  - ESTs are a good source of SNPs

# SNPs

- SNPs often act as “disease markers”, and provide “genetic predisposition”.
- SNPs may explain differences in drug response of individuals.
- Association study: study SNP patterns in diseased individuals and compare against SNP patterns in normal individuals.
- Many diseases associated with SNP profile.