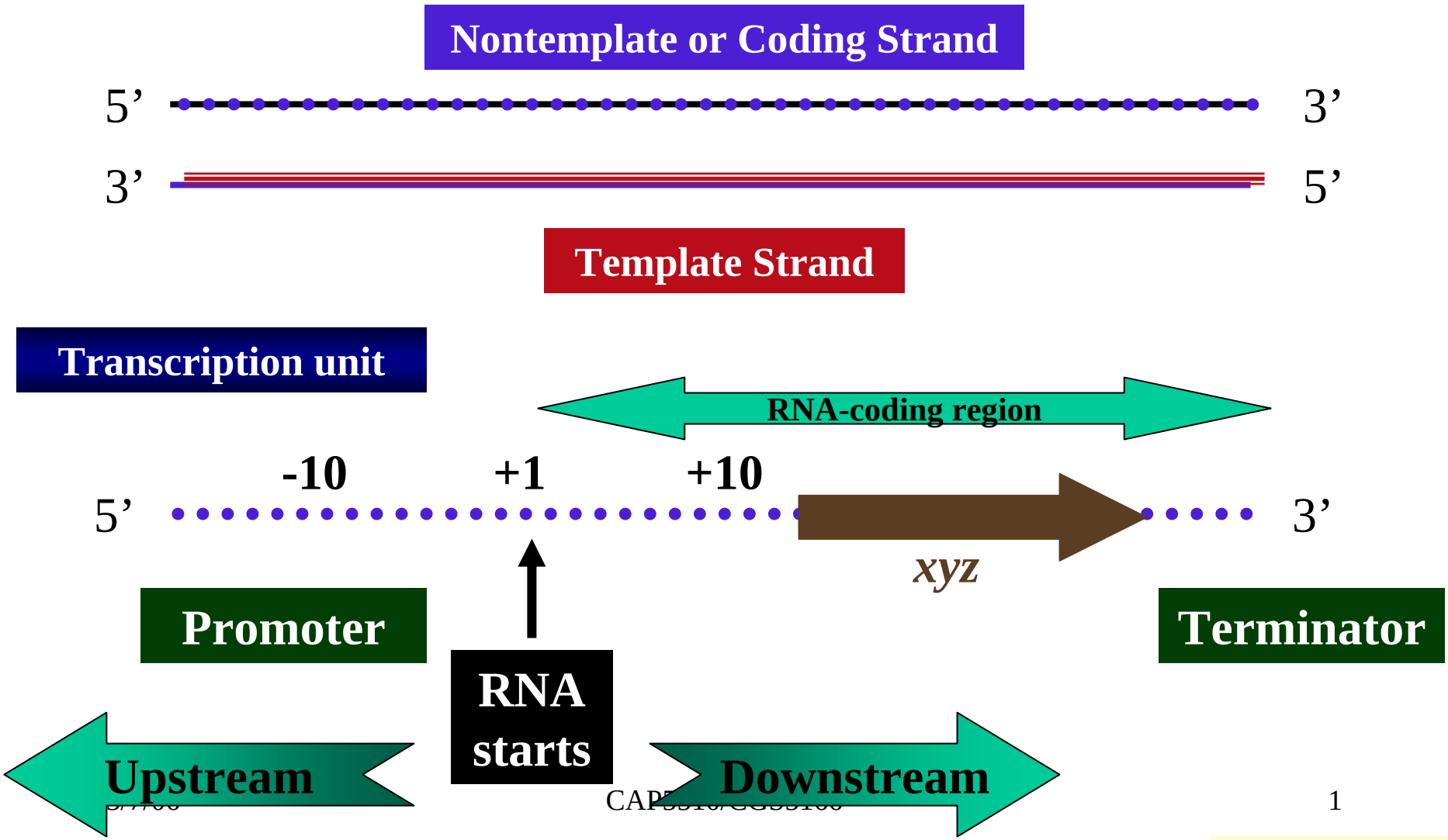
Nomenclature

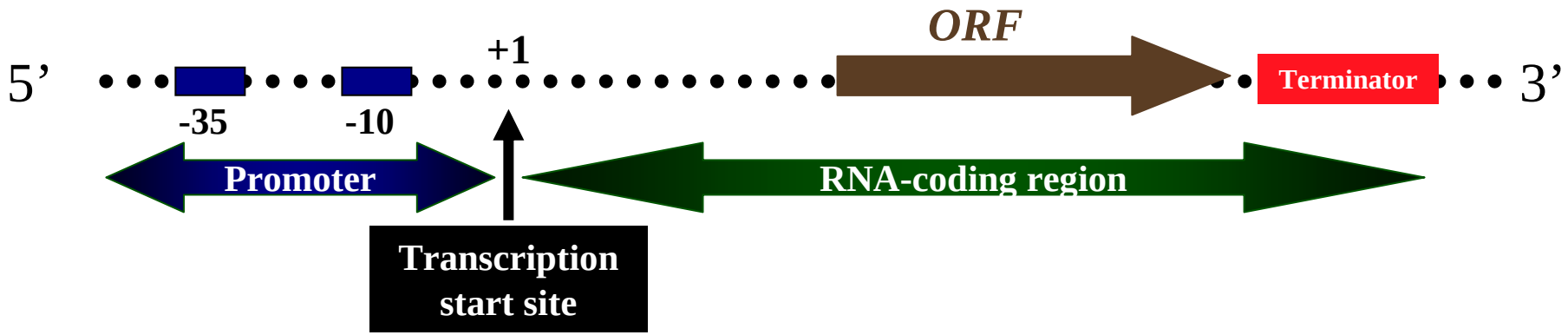
RNA Polymerization occurs 5’ to 3’



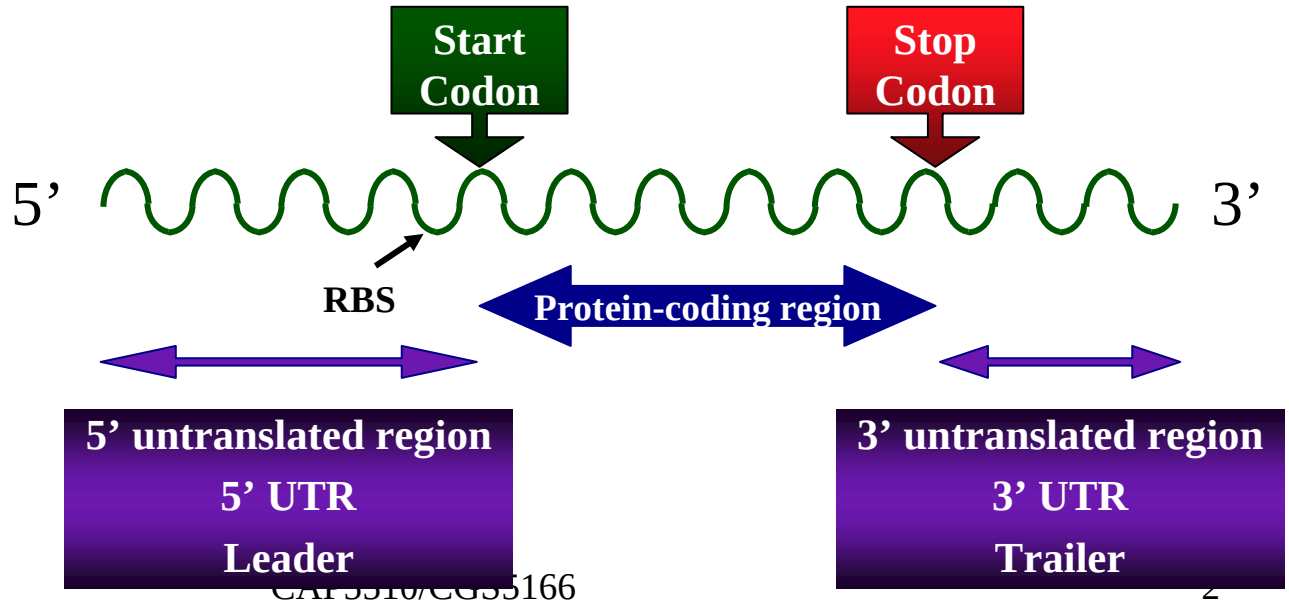
Slide courtesy Prof. Mathee

Transcriptional unit and single gene mature mRNA

Transcriptional unit



mRNA



RBS

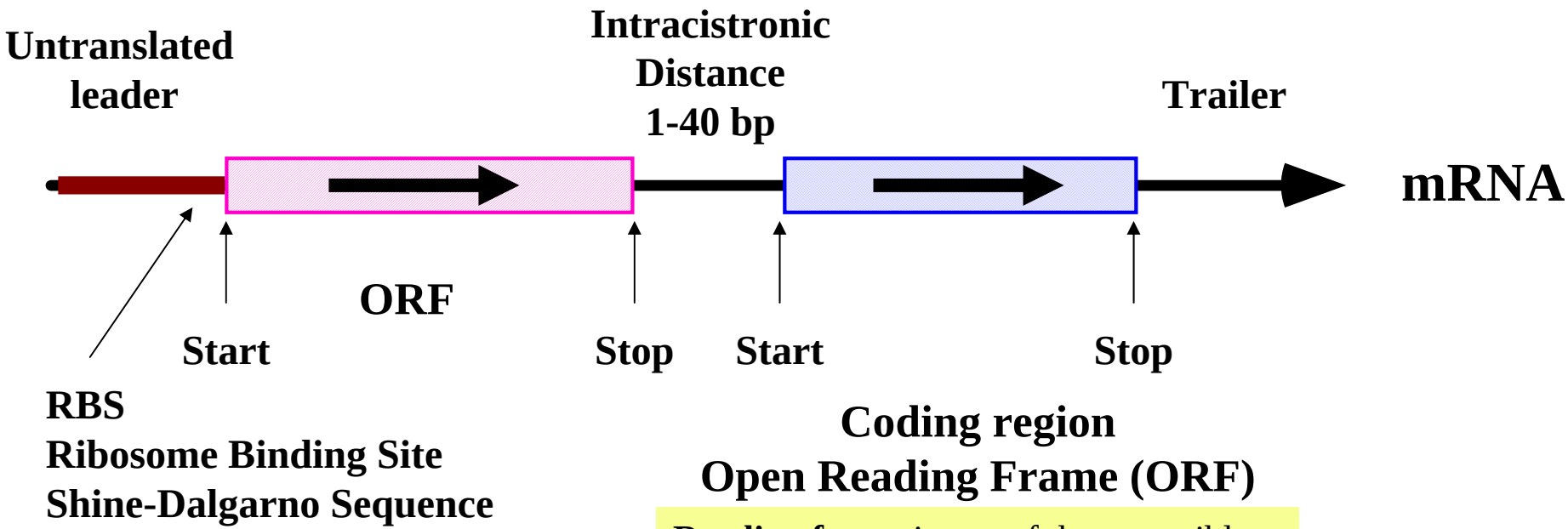
Ribosome binding site

Slide courtesy Prof. Mathee

Messenger RNA or mRNA

Initation Codon		
	AUG	Methionine
Others:	GUG	Valine
	UUG	Leucine
	AUU	Isoleucine

Termination Codons	
UAA	Ochre
UAG	Amber
UGA	Opal

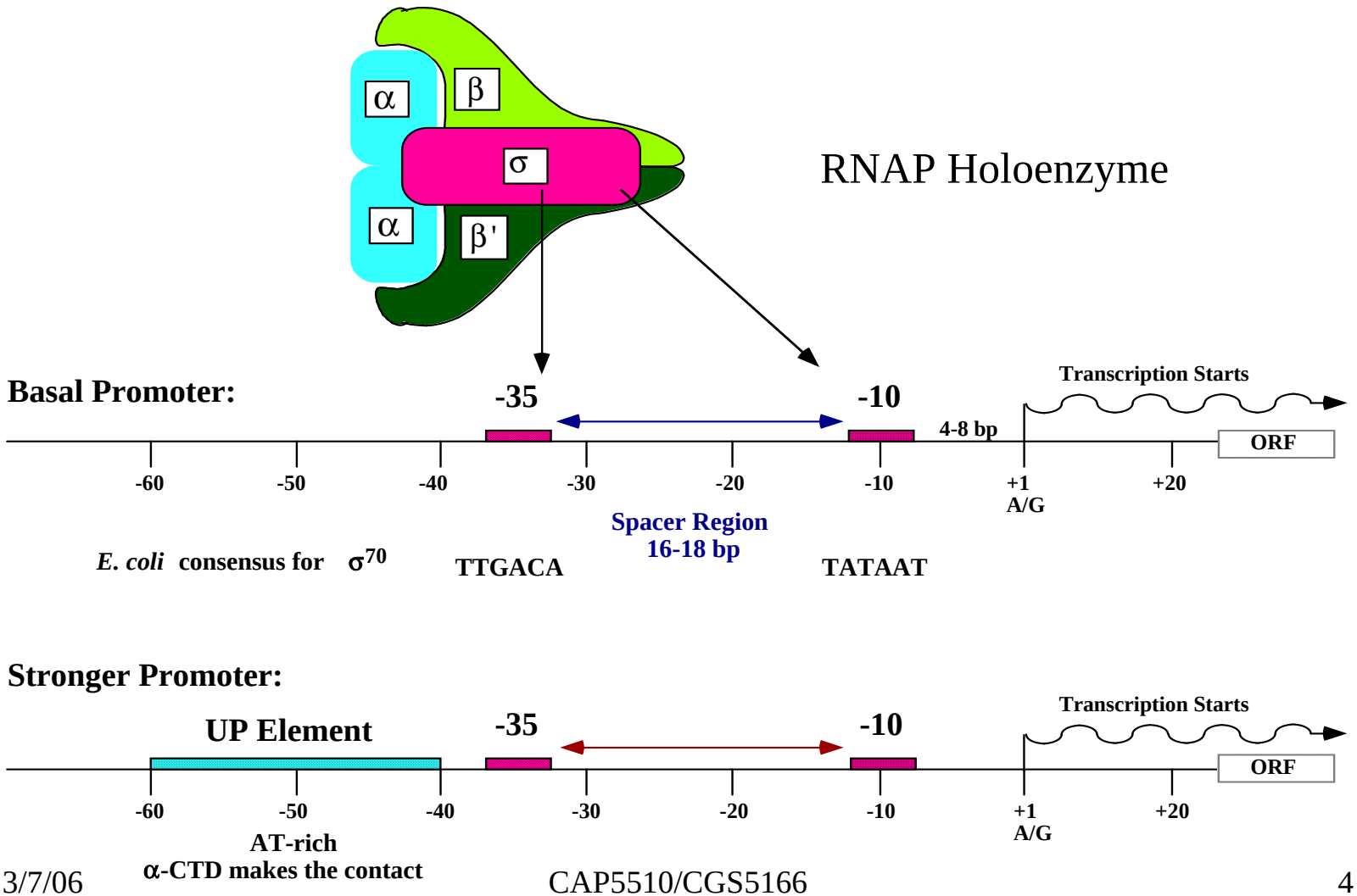


7 bp upstream of start codon
5'--AGGAGG--3'

Reading frame is one of three possible ways of reading a nucleotide sequence as a series of triplets.

Slide courtesy Prof. Mathee

Transcriptional machinery: RNA Polymerase and DNA



Slide courtesy Prof. Mathee

Prokaryotic Gene Characteristics

DNA PATTERNS IN THE *E. coli lexA* GENE

```
                GENE SEQUENCE                          PATTERN

  1 GAATTCGATAAATCTCTGGTTTATTGTGCAGTTTATGGTT      CTGNNNNNNNNNNCAG
                                          TT      TTGACA
 41 CCAAAATCGCCTTTTGCTGTATATACTCACAGCATAACTG      CTGNNNNNNNNNNCAG
    CCAA -35          -10  TATACT        >        TATAAT, > mRNA start
 81 TATATACACCCAGGGGGCGGAATGAAAGCGTTAACGGCCA      CTGNNNNNNNNNNCAG
          +10  GGGGG Ribosomal binding site        GGAGG
121 GGCAACAAGAGGTGTTTGATCTCATCCGTGATCACATCAG
161 CCAGACAGGTATGCCGCCGACGCGTGCGGAAATCGCGCAG      ATG
201 CGTTTGGGGTTCCGTTCCCCAAACGCGGCTGAAGAACATC
241 TGAAGGCGCTGGCACGCAAAGGCGTTATTGAAATTGTTTC
281 CGGCGCATCACGCGGGATTCGTCTGTTGCAGGAAGAGGAA
321 GAAGGGTTGCCGCTGGTAGGTCGTGTGGCTGCCGGTGAAC
361 CACTTCTGGCGCAACAGCATATTGAAGGTCATTATCAGGT      OPEN READING FRAME
401 CGATCCTTCCTTATTCAAGCCGAATGCTGATTTCCTGCTG
441 CGCGTCAGCGGGATGTCGATGAAAGATATCGGCATTATGG
481 ATGGTGACTTGCTGGCAGTGCATAAAACTCAGGATGTACG
521 TAACGGTCAGGTCGTTGTCGCACGTATTGATGACGAAGTT
561 [illegible]
601 TGTTGCCAGAAAATAGCGAGTTTAAACCAATTGTCGTTGA
641 CCTTCGTCAGCAGAGCTTCACCATTGAAGGGCTGGCGGTT
681 GGGGTTATTCGCAACGGCGACTGGCTGTAACATATCTCTG      TAA
721 AGACCGCGATGCCGCCTGGCGTCGCGGTTTGTTTTTCATC
761 TCTCTTCATCAGGCTTGTCTGCATGGCATTCCTCACTTCA
801 TCTGATAAAGCACTCTGGCATCTCGCCTTACCCATGATTT
841 TCTCCAATATCACCGTTCCGTTGCTGGGACTGGTCGATAC
881 GGCGGTAATTGGTCATCTTGATAGCCCGGTTTATTTGGGC
921 GGCGTGGCGGTTGGCGCAACGGCGGACCAGCT
```

Shown are matches to approximate consensus binding sites for LexA repressor (CTGNNNNNNNNNNCAG), the -10 amd -35 promoter regions relative to the start of the mRNA (TTGACA and TATAAT), the ribosomal binding site on the mRNA (GGAGG), and the open reading frame (ATG...TAA). Only the second two of the predicted LexA binding sites actually bind the repressor.

FIGURE 9.6. The promoter and open reading frame of the *E. coli lexA* gene.

Gene Expression

- Process of transcription and/or translation of a gene is called gene expression.
- Every cell of an organism has the same genetic material, but different genes are expressed at different times.
- Patterns of gene expression in a cell is indicative of its state.

Hybridization

- If two complementary strands of DNA or mRNA are brought together under the right experimental conditions they will hybridize.
- A hybridizes to B $\Rightarrow$
  - A is reverse complementary to B, or
  - A is reverse complementary to a subsequence of B.
- It is possible to experimentally verify whether A hybridizes to B, by labeling A or B with a radioactive or fluorescent tag, followed by excitation by laser.

Measuring gene expression

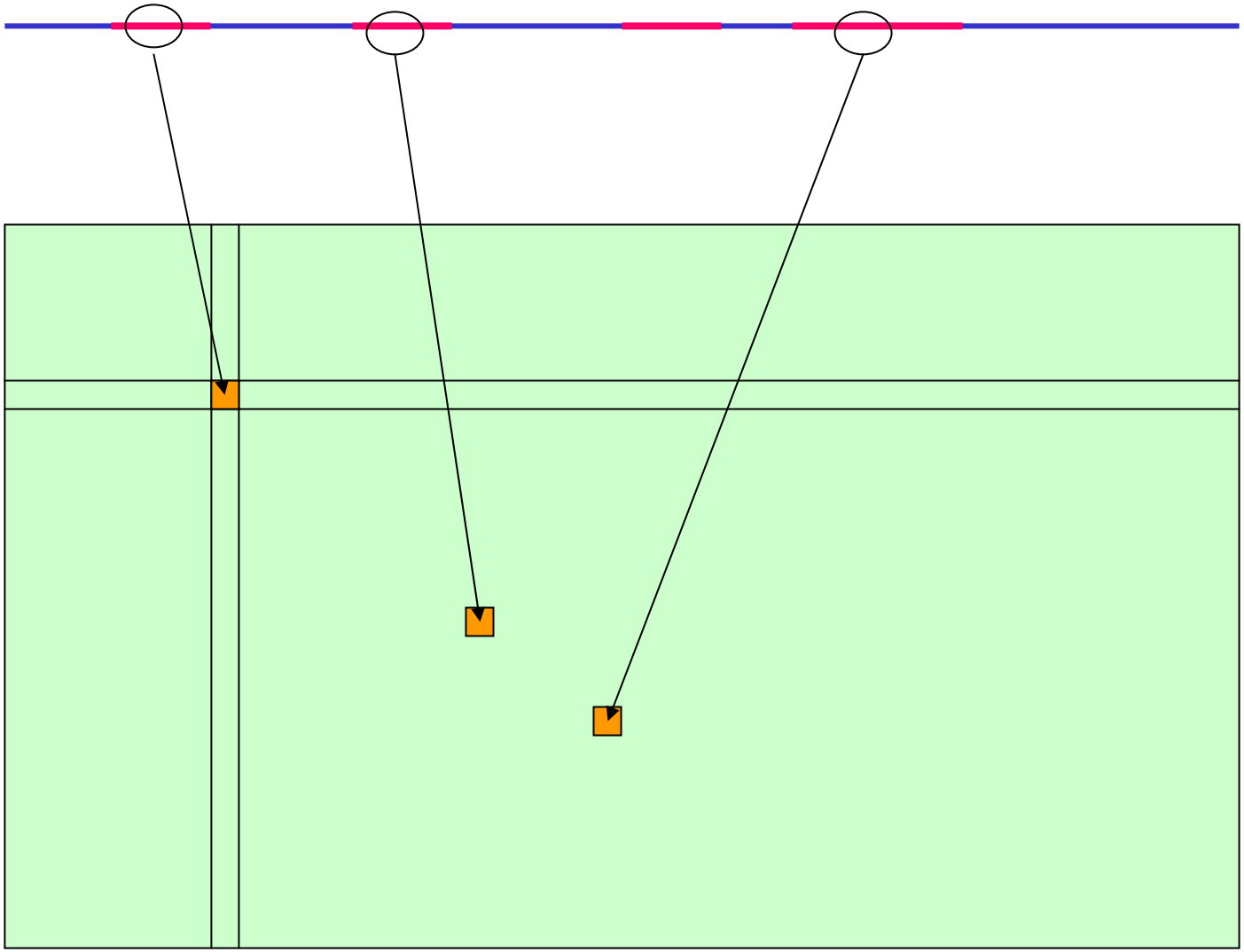
- Gene expression for a single gene can be measured by extracting mRNA from the cell and doing a simple hybridization experiment.
- Given a sample of cells, gene expression for every gene can be measured using a single microarray experiment.

Microarray/DNA chip technology

- High-throughput method to study gene expression of thousands of genes simultaneously.
- Many applications:
  - Genetic disorders & Mutation/polymorphism detection
  - Study of disease subtypes
  - Drug discovery & toxicology studies
  - Pathogen analysis
  - Differing expressions over time, between tissues, between drugs, across disease states

Microarray Data

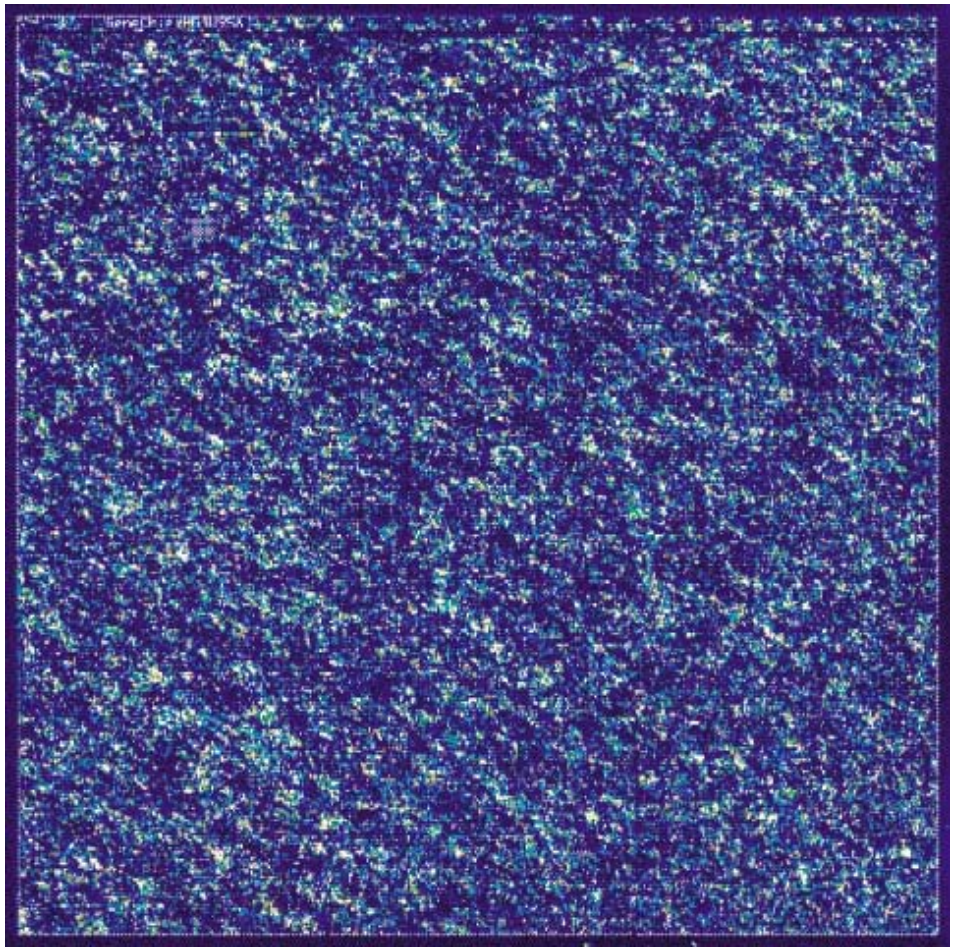
Gene	Expression Level
Gene1	
Gene2	
Gene3	
…	



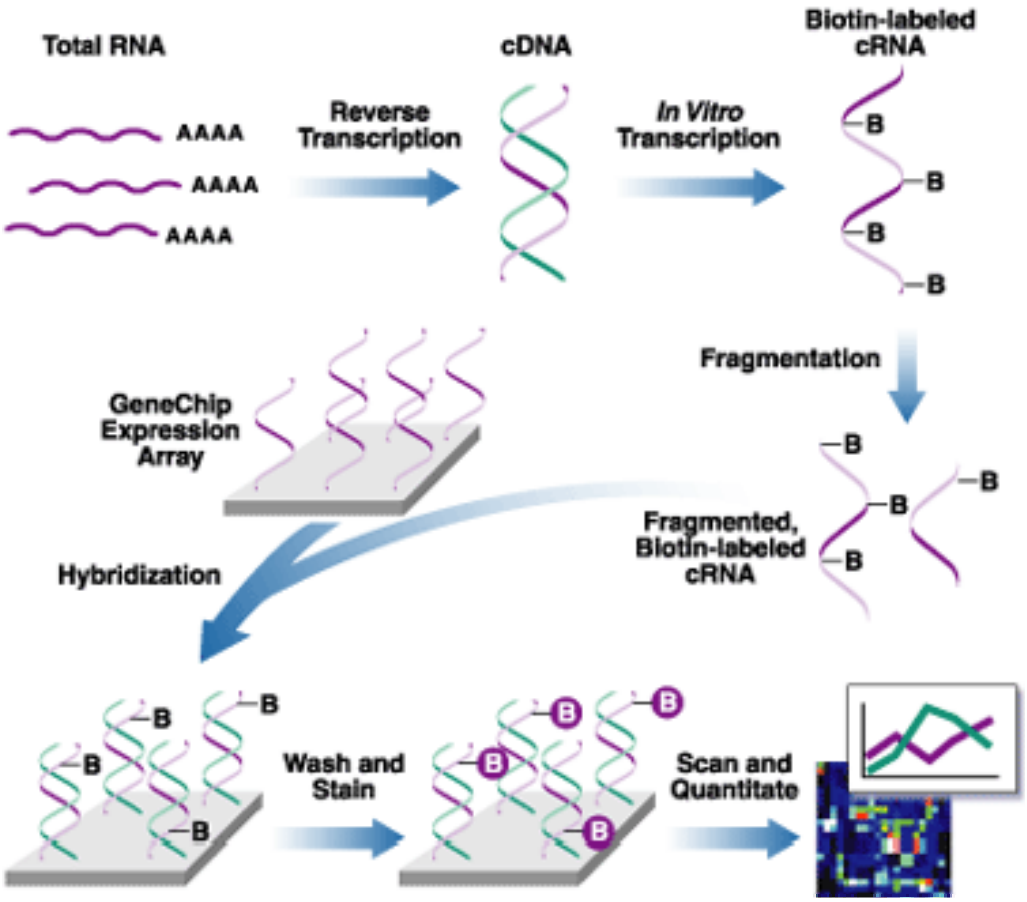
Microarray/DNA chips (Simplified)

- Construct probes corresponding to reverse complements of genes of interest.
- Microscopic quantities of probes placed on solid surfaces at defined spots on the chip.
- Extract mRNA from sample cells and label them.
- Apply labeled sample (mRNA extracted from cells) to every spot, and allow hybridization.
- Wash off unhybridized material.
- Use optical detector to measure amount of fluorescence from each spot.

Gene Chips

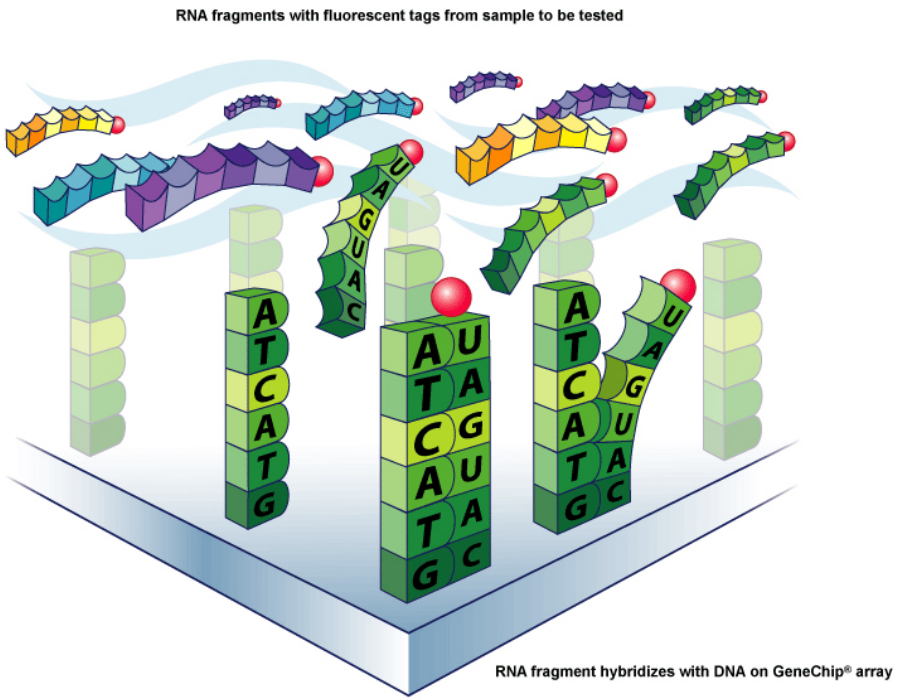


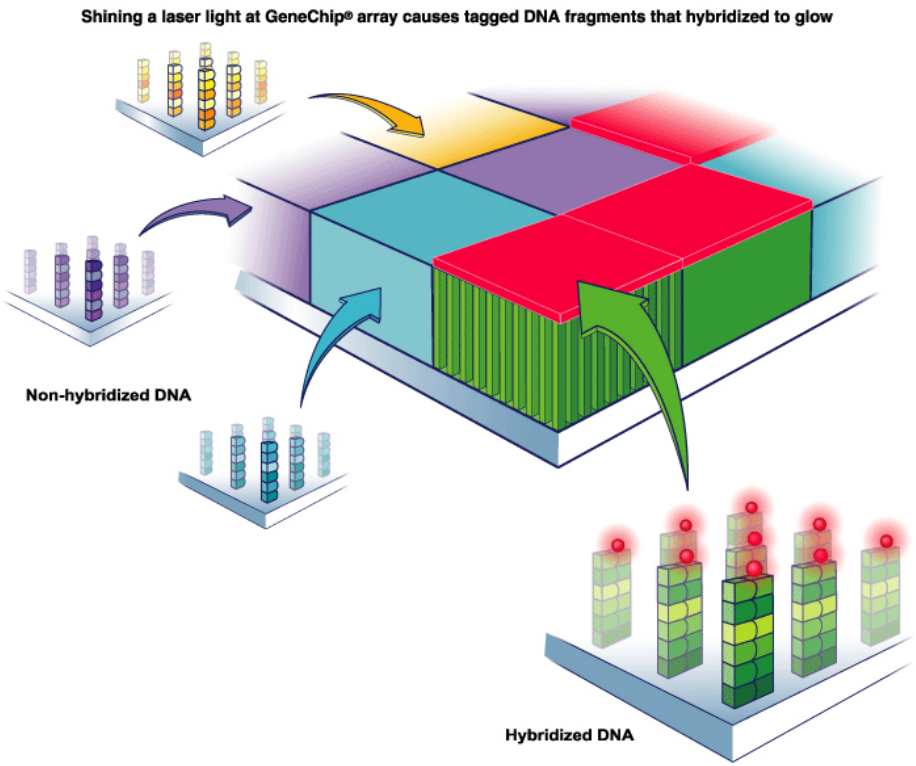
Affymetrix DNA chip schematic



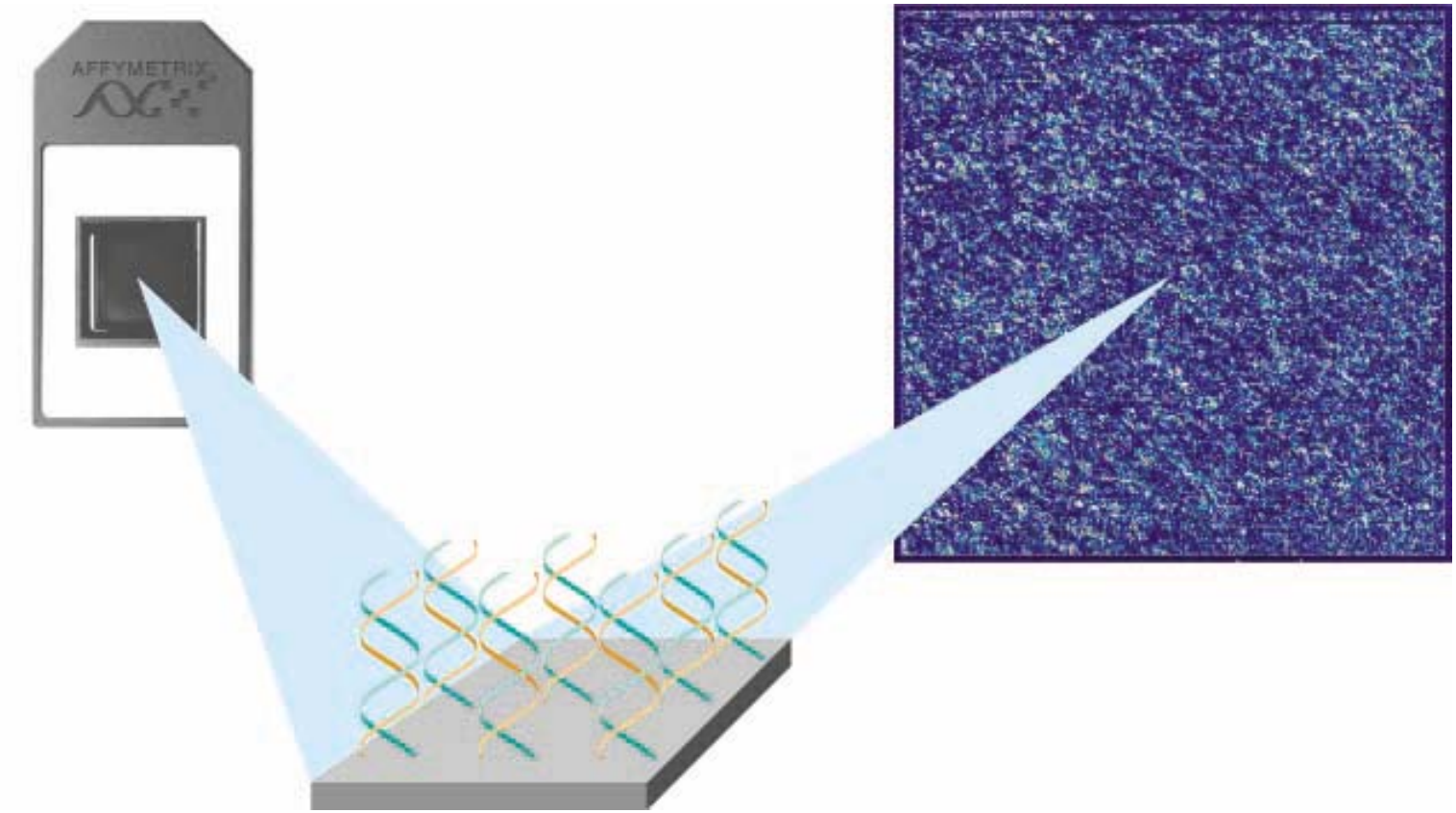
www.affymetrix.com

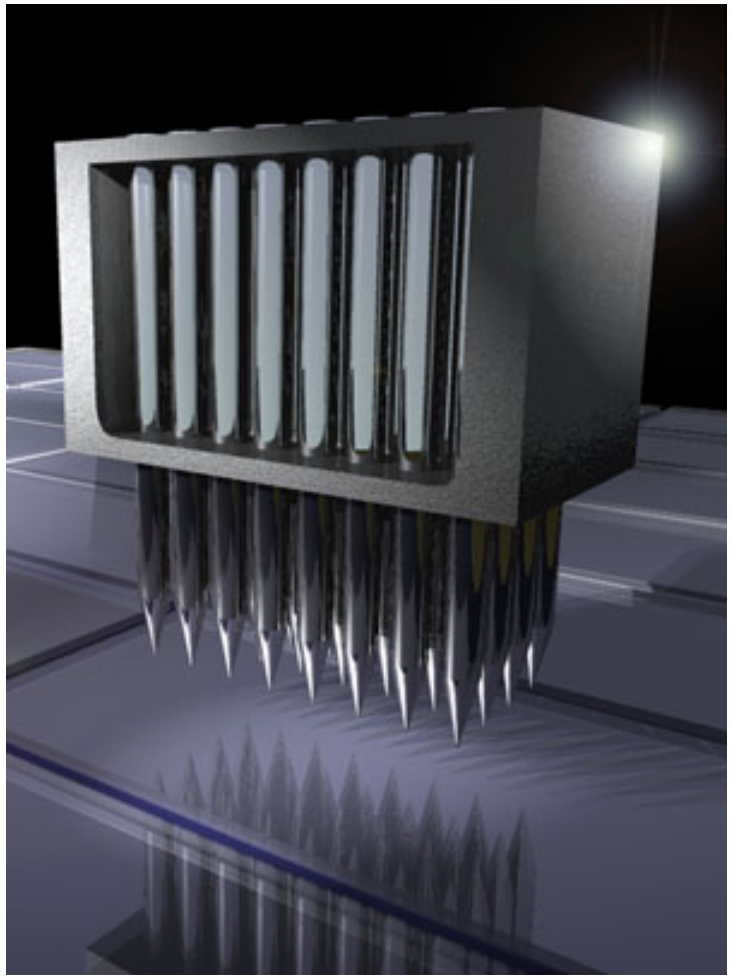
What's on the slide?





DNA Chips & Images





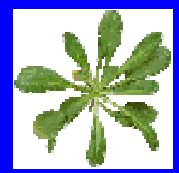
Microarrays: competing technologies

- Affymetrix & Synteni/Stanford
- Differ in:
  - method to place DNA: Spotting vs. photolithography
  - Length of probe
  - Complete sequence vs. series of fragments

How to compare 2 cell samples with Two-Color Microarrays?

- mRNA from sample 1 is extracted and labeled with a red fluorescent dye.
- mRNA from sample 2 is extracted and labeled with a green fluorescent dye.
- Mix the samples and apply it to every spot on the microarray. Hybridize sample mixture to probes.
- Use optical detector to measure the amount of green and red fluorescence at each spot.

2-color DNA microarray



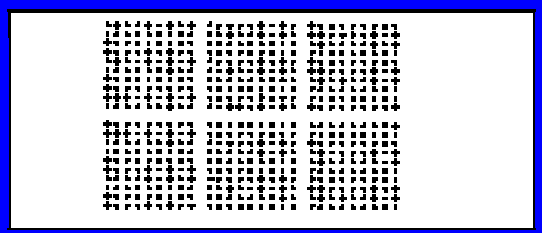
Treated

Control

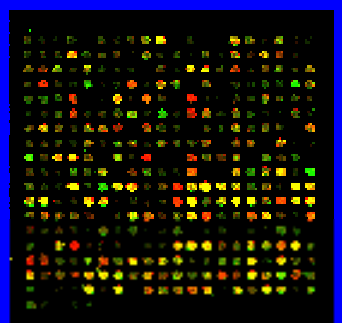
mRNA

mRNA

Cy5 Probe Cy3 Probe



Simultaneous hybridization



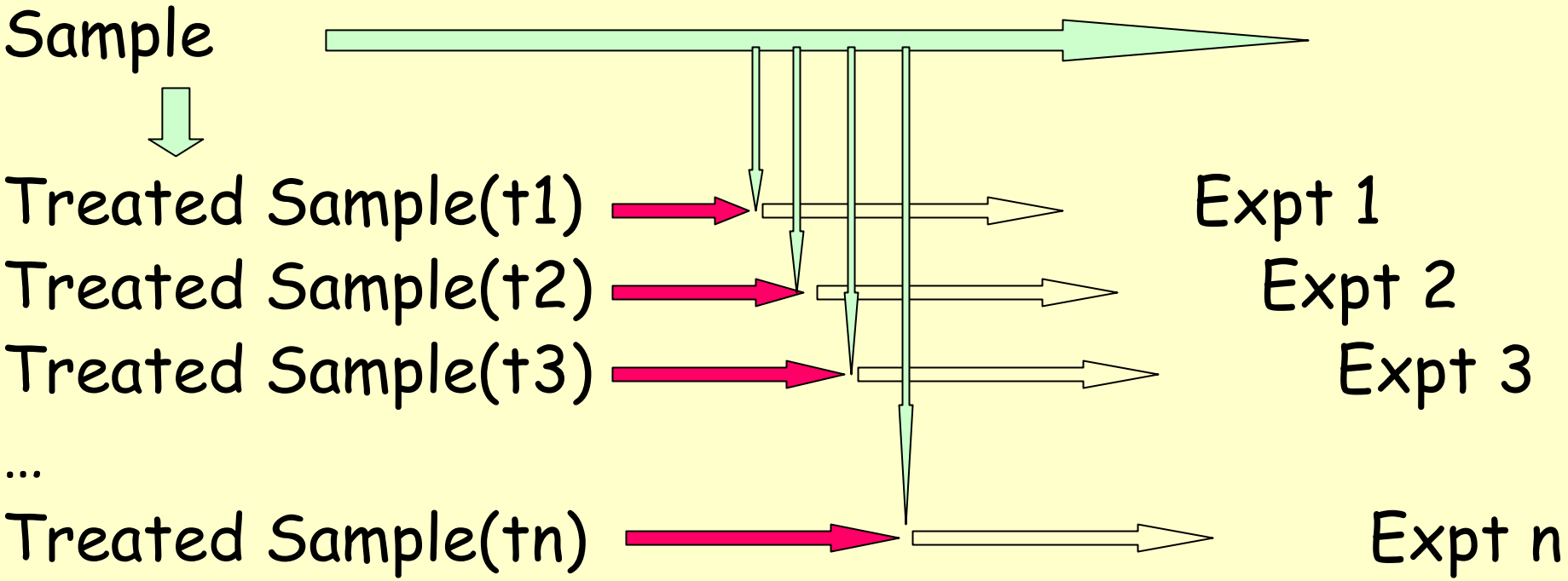
Scanning

Data extraction

Normalization

http://www.arabidopsis.org/info/2010_projects/comp_proj/AFGC/RevisedAFGC/Friday/

Study effect of treatment over time



Sources of Variations & Errors

- 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

Types of bias/variation

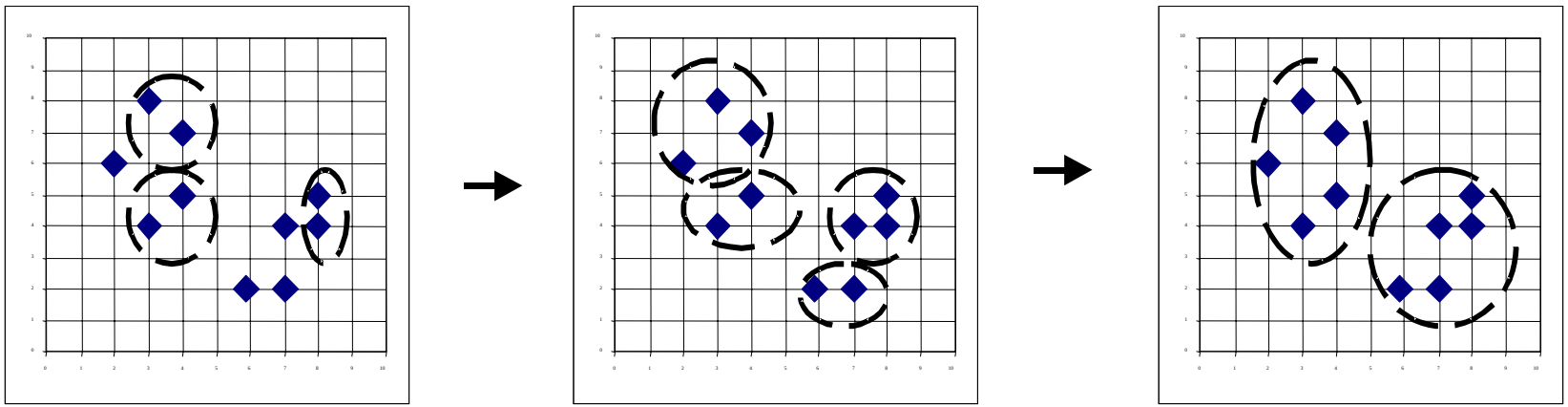
- Intensity & Range
  - Variation changes with intensity. Larger variation at lower end.
- Spatial
  - Spot location changes expression
- Plate
  - Printing plate changes expression

http://www.arabidopsis.org/info/2010_projects/comp_proj/AFGC/RevisedAFGC/Friday/index.htm

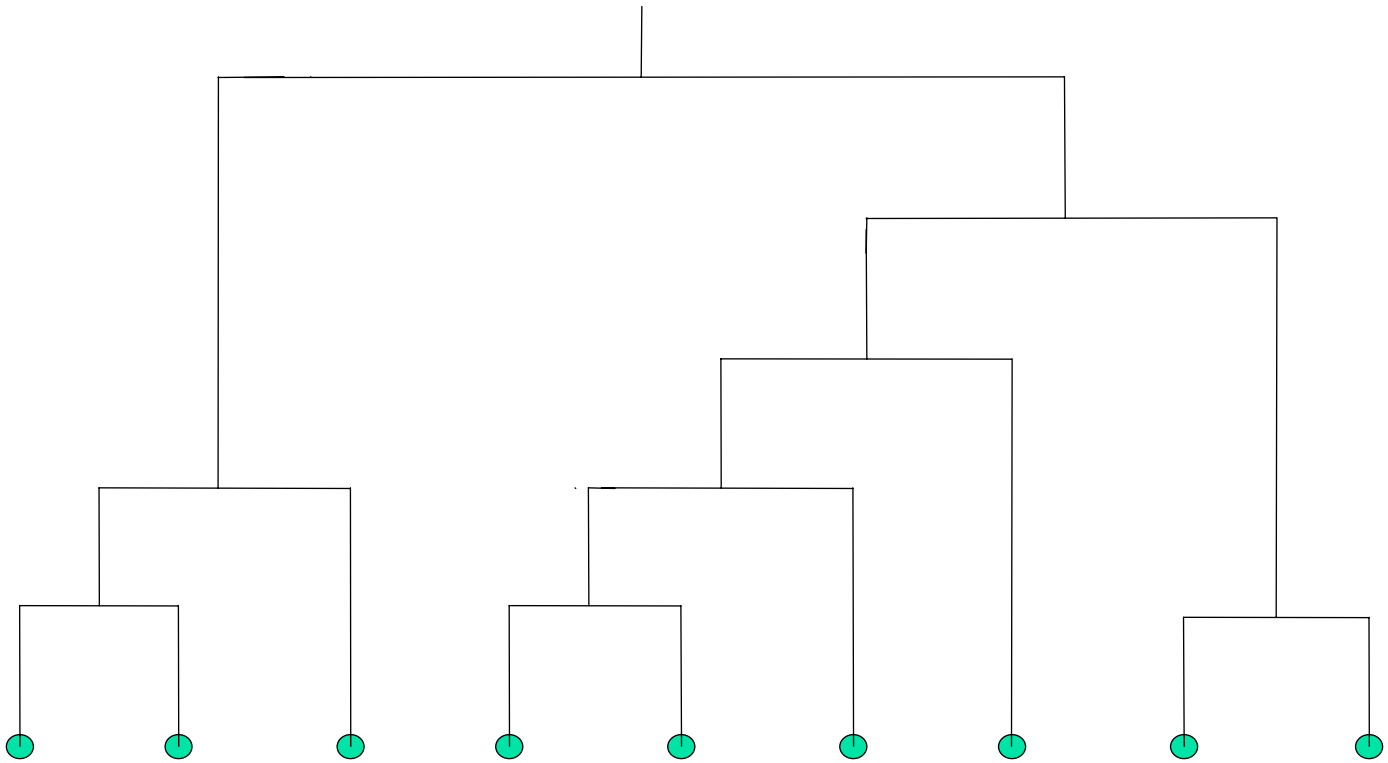
Clustering

- Clustering is a general method to study patterns in gene expressions.
- Several known methods:
  - Hierarchical Clustering (Bottom-Up Approach)
  - K-means Clustering (Top-Down Approach)
  - Self-Organizing Maps (SOM)

Hierarchical Clustering: Example



A Dendrogram



Hierarchical Clustering [Johnson, SC, 1967]

- Given n points in R^d, compute the distance between every pair of points
- While (not done)
  - Pick closest pair of points s_i and s_j and make them part of the same cluster.
  - Replace the pair by an average of the two s_{ij}

Try the applet at:
http://www.cs.mcgill.ca/~papou/#applet

Distance Metrics

- For clustering, define a distance function:
  - Euclidean distance metrics

$$D_k(X,Y) = \left[\sum_{i=1}^{d} (X_i - Y_i)^k \right]^{1/k}$$

k=2: Euclidean Distance

  - Pearson correlation coefficient

$$\rho_{xy} = \frac{1}{d} \sum_{i=1}^{d} \left(\frac{X_i - \overline{X}}{\sigma_x} \right) \left(\frac{Y_i - \overline{Y}}{\sigma_y} \right)$$

$-1 \leq \rho_{xy} \geq 1$

EXHIBIT 3.4 *Joint Probability Model for the Ratings of Two People*

(a) $\rho_{XY} = 0$

	y			
x	1	2	3	Total
3	1/9	1/9	1/9	1/3
2	1/9	1/9	1/9	1/3
1	1/9	1/9	1/9	1/3
Total	1/3	1/3	1/3	1

(b) $\rho_{XY} = \frac{1}{2}$

	y			
x	1	2	3	Total
3	1/18	1/18	4/18	1/3
2	1/18	4/18	1/18	1/3
1	4/18	1/18	1/18	1/3
Total	1/3	1/3	1/3	1

(c) $\rho_{XY} = -\frac{1}{2}$

	y			
x	1	2	3	Total
3	4/18	1/18	1/18	1/3
2	1/18	4/18	1/18	1/3
1	1/18	1/18	4/18	1/3
Total	1/3	1/3	1/3	1

(d) $\rho_{XY} = \frac{4}{9}$

	y			
x	1	2	3	Total
3	1/27	2/27	6/27	1/3
2	2/27	5/27	2/27	1/3
1	6/27	2/27	1/27	1/3
Total	1/3	1/3	1/3	1

(e) $\rho_{XY} = -\frac{5}{9}$

	y			
x	1	2	3	Total
3	6/27	2/27	1/27	1/3
2	2/27	5/27	2/27	1/3
1	1/27	2/27	6/27	1/3
Total	1/3	1/3	1/3	1

(f) $\rho_{XY} = \frac{2}{3}$

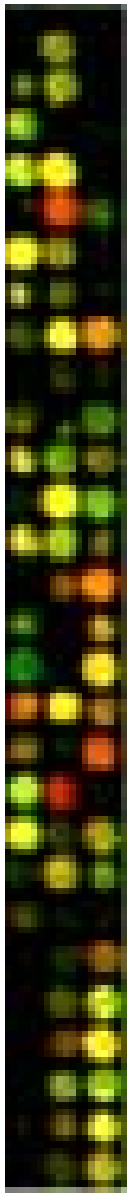
	y			
x	1	2	3	Total
3	1/36	2/36	9/36	1/3
2	2/36	8/36	2/36	1/3
1	9/36	2/36	1/36	1/3
Total	1/3	1/3	1/3	1

(g) $\rho_{XY} = -\frac{1}{3}$

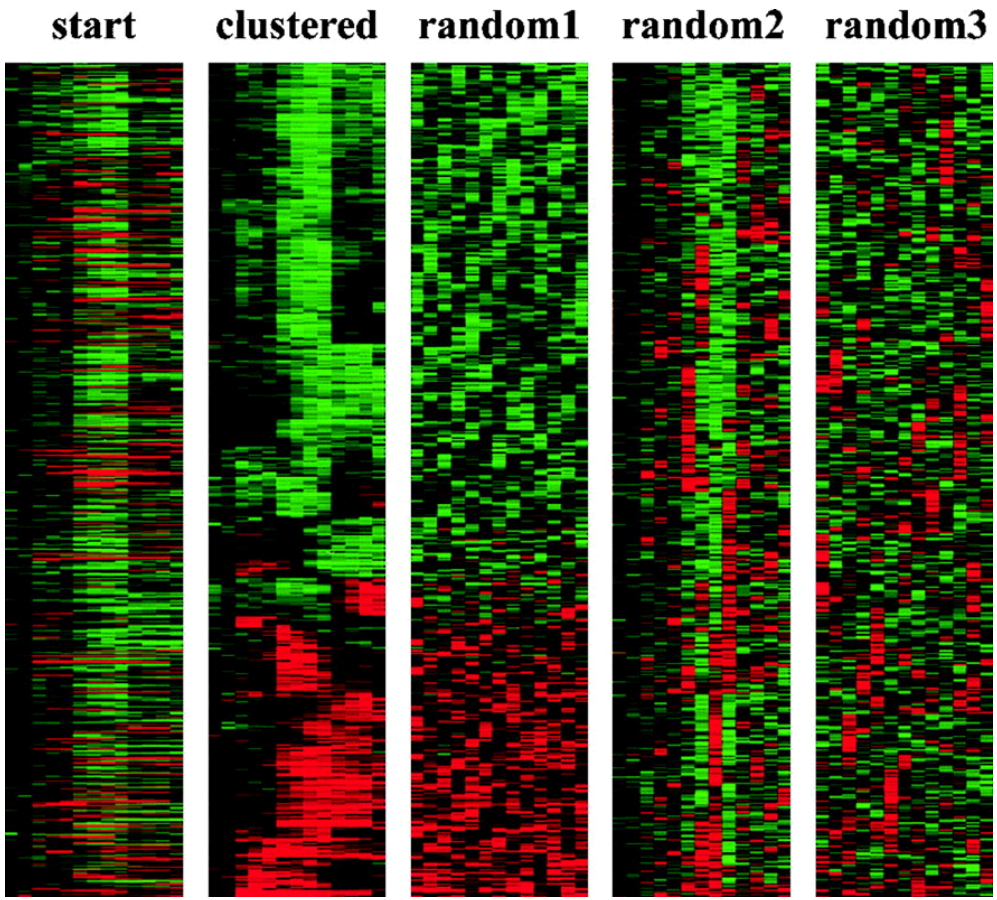
	y			
x	1	2	3	Total
3	9/36	2/36	1/36	1/3
2	2/36	8/18	2/18	1/3
1	1/36	2/36	9/36	1/3
Total	1/3	1/3	1/3	1

Clustering of gene expressions

- Represent each gene as a vector or a point in d-space where d is the number of arrays or experiments being analyzed.



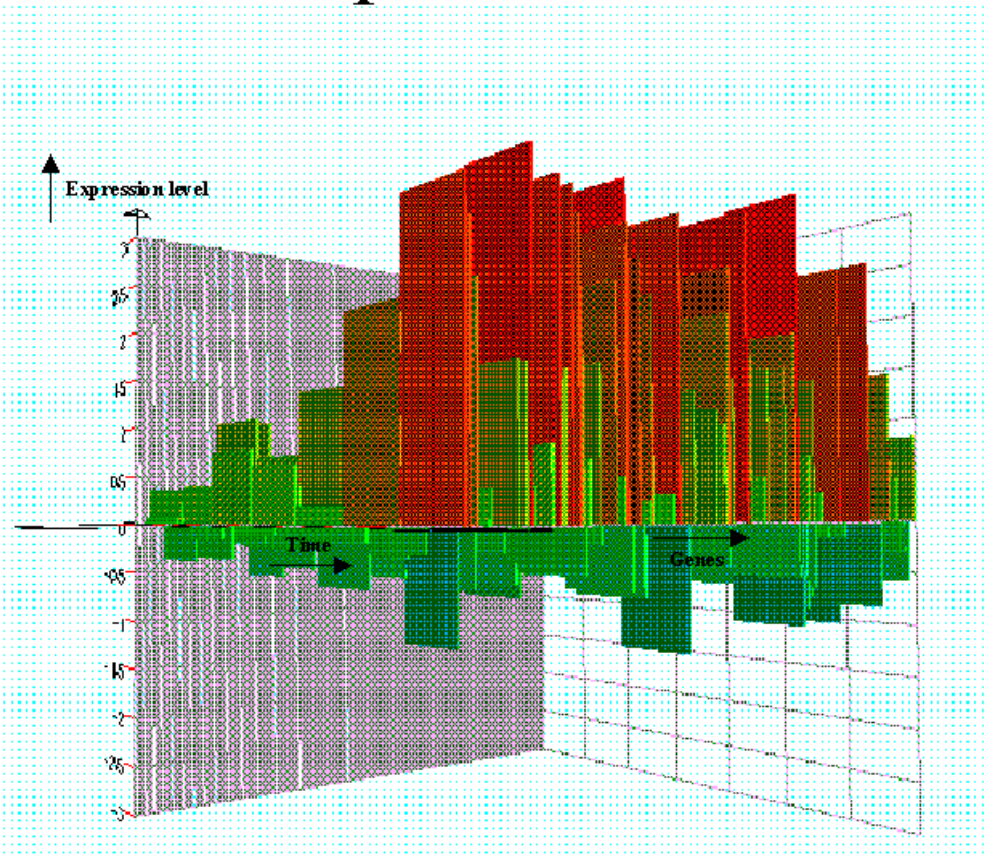
Clustering Random vs. Biological Data



From *Eisen MB, et al, PNAS 1998 95(25):14863-8*

European Bioinformatics Institute

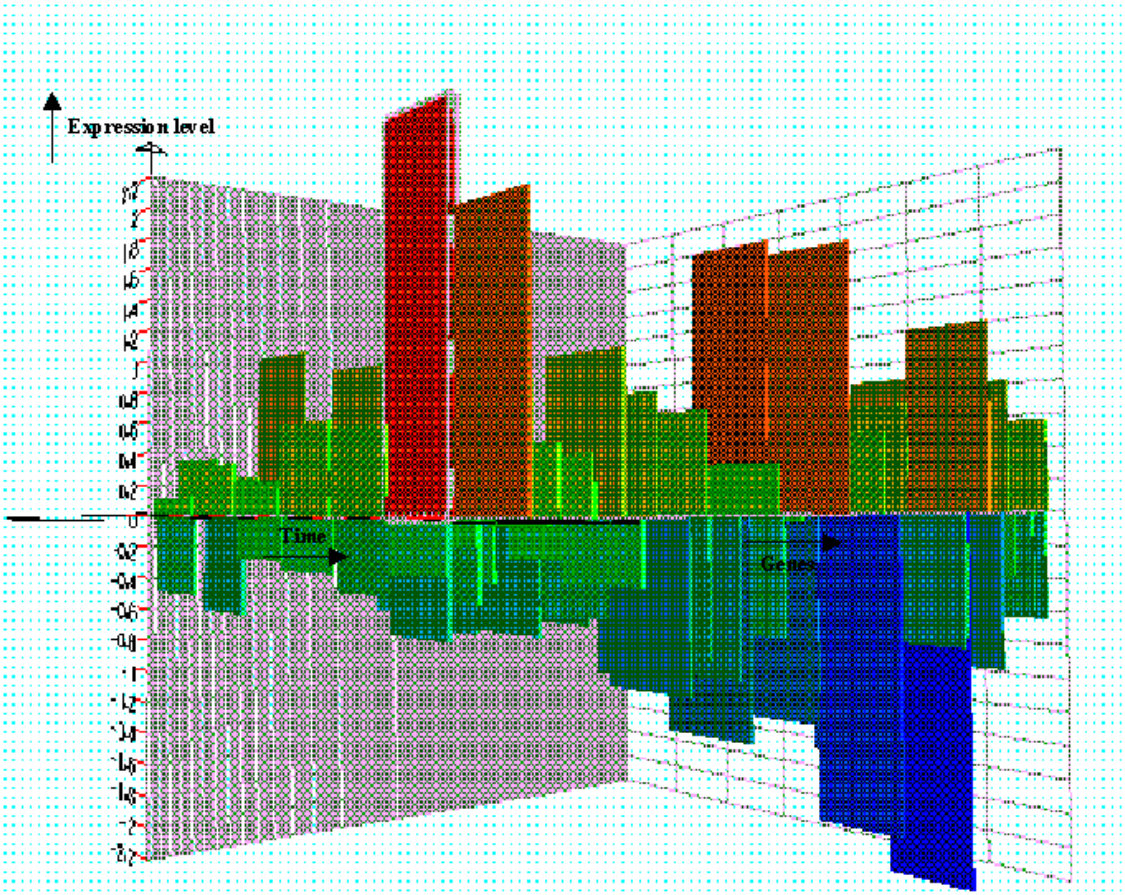
Expression Profiles for Respiration Genes



Genes show general increase in expression level during diauxic shift

European Bioinformatics Institute

Expression Profiles for Fermentation Genes



Bar two exceptions, genes show general decrease in expression level during diauxic shift

European Bioinformatics Institute

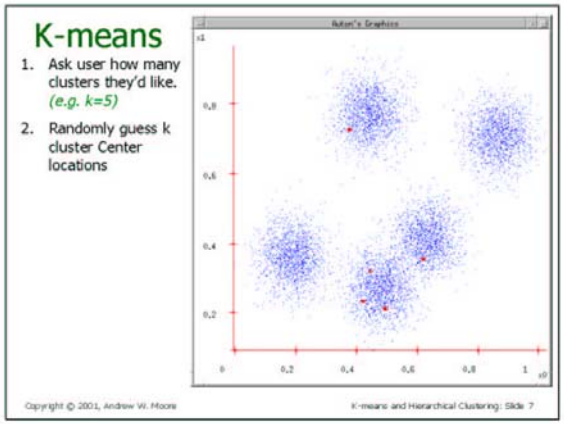
Observations

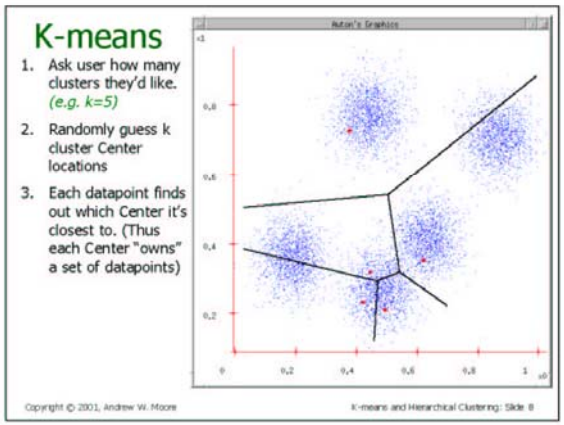
- As glucose was depleted - Marked change in the global pattern of gene expression
- ~50% of differentially expressed genes have unknown function
- Genes with similar expression profiles had common promoters
- Expression patterns observed match those observed in other types of experiments

K-Means Clustering: Example

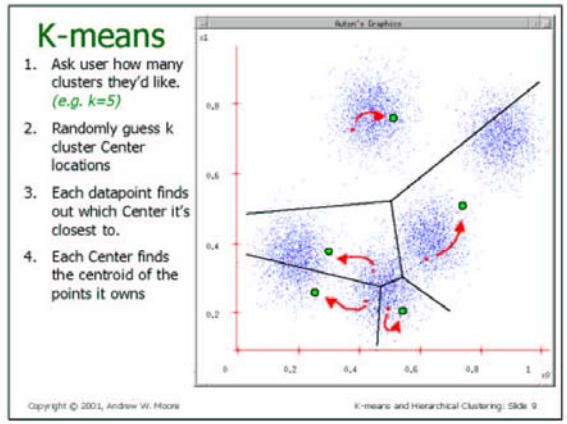
Example from Andrew Moore's tutorial on Clustering.

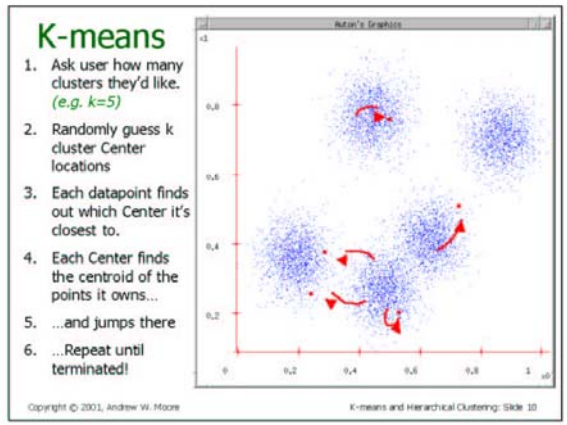
Start


K-means
1. Ask user how many clusters they'd like. (e.g. k=5)
2. Randomly guess k cluster Center locations
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 7


K-means
1. Ask user how many clusters they'd like. (e.g. k=5)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 8

4


K-means
1. Ask user how many clusters they'd like. (e.g. k=5)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 9


K-means
1. Ask user how many clusters they'd like. (e.g. k=5)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns...
5. ...and jumps there
6. ...Repeat until terminated!
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 10

5

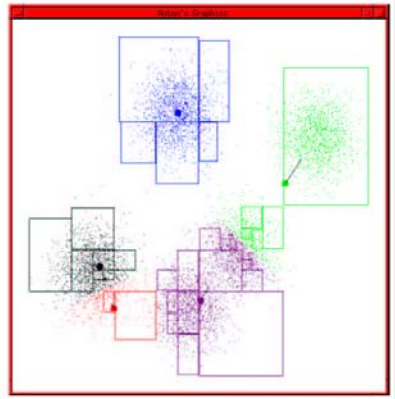
K-means Start

Advance apologies: in Black and White this example will deteriorate

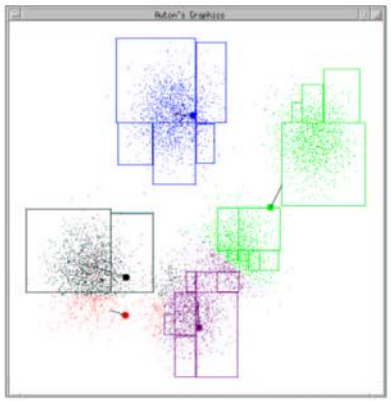
Example generated by Dan Pelleg's super-duper fast K-means system:

Dan Pelleg and Andrew Moore. Accelerating Exact k-means Algorithms with Geometric Reasoning. Proc. Conference on Knowledge Discovery in Databases 1999, (KDD99) (available on www.autonlab.org/pap.html)

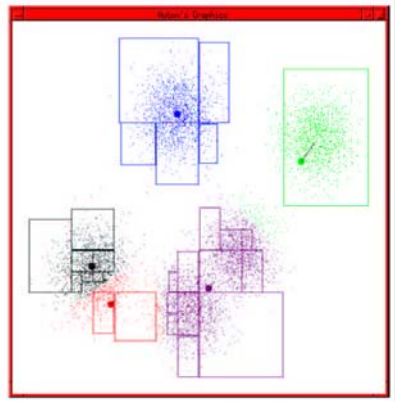
K-means continues …



K-means continues …

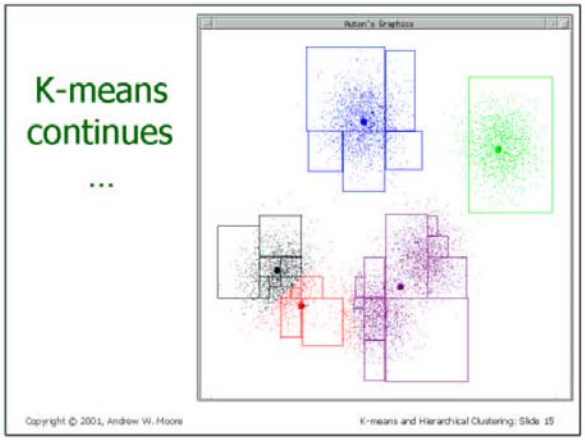


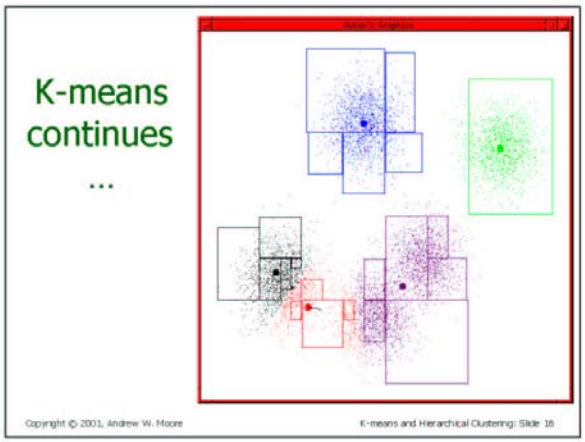
K-means continues …



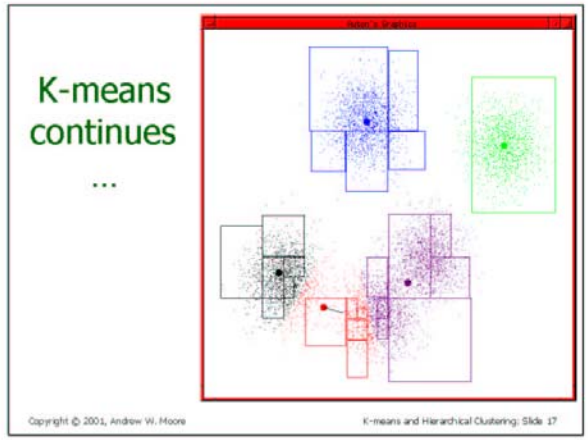
6

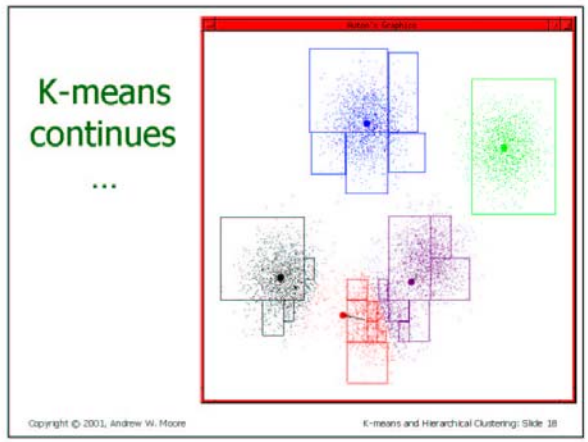
7


K-means
continues
...
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 15


K-means
continues
...
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 16

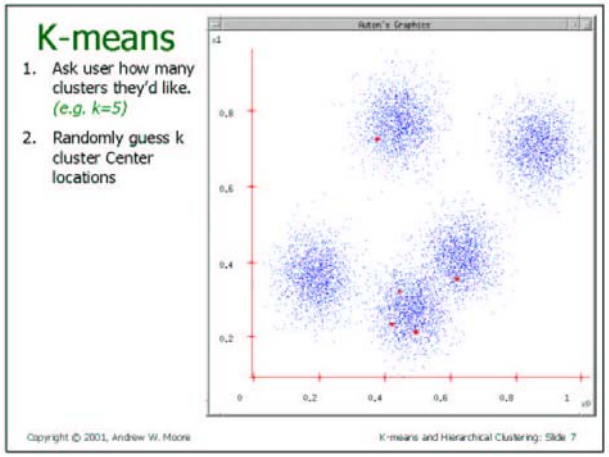
8

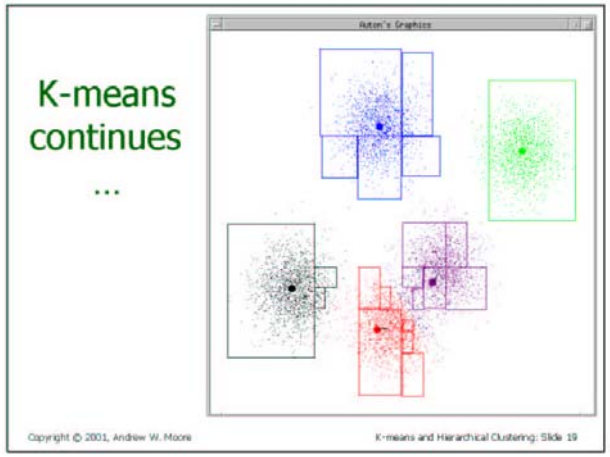

K-means
continues
...
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 17

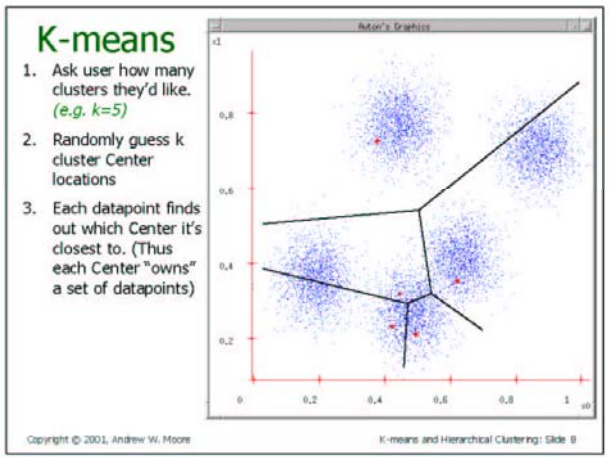

K-means
continues
...
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 18

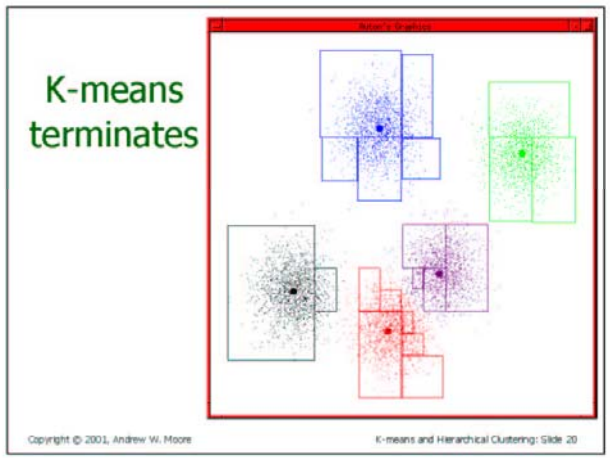
9

Start









End

4

10

K-Means Clustering [McQueen ’67]

Repeat

- Start with randomly chosen cluster centers
- Assign points to give greatest increase in score
- Recompute cluster centers
- Reassign points

until (no changes)

Try the applet at: http://www.cs.mcgill.ca/~bonnef/project.html

Comparisons

- Hierarchical clustering
  - Number of clusters not preset.
  - Complete hierarchy of clusters
  - Not very robust, not very efficient.
- K-Means
  - Need definition of a mean. Categorical data?
  - More efficient and often finds optimum clustering.

Functionally related genes behave similarly across experiments

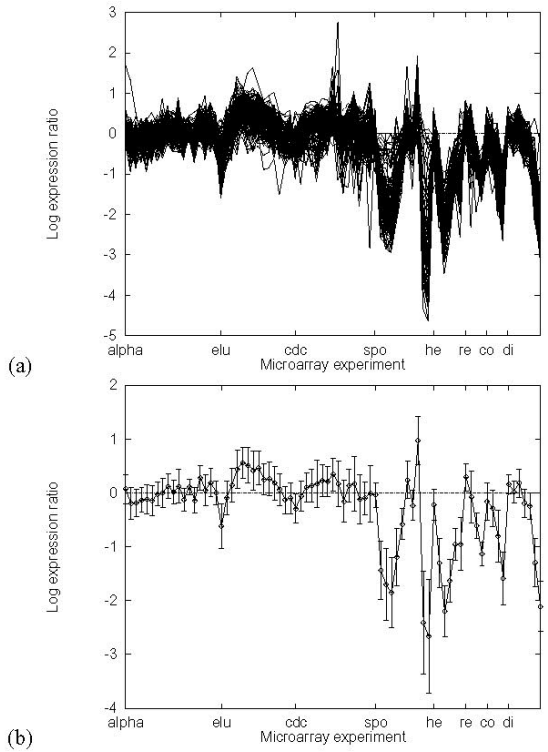


Figure 1: **Expression profiles of the cytoplasmic ribosomal proteins.** Figure (a) shows the expression profiles from the data in [Eisen et al., 1998] of 121 cytoplasmic ribosomal proteins, as classified by MYGD [MYGD, 1999]. The logarithm of the expression ratio is plotted as a function of DNA microarray experiment. Ticks along the X-axis represent the beginnings of experimental series. They are, from left to right, cell division cycle after synchronization with α factor arrest (alpha), cell division cycle after synchronization by centrifugal elutriation (elu), cell division cycle measured using a temperature sensitive *cdc15* mutant (cdc), sporulation (spo), heat shock (he), reducing shock (re), cold shock (co), and diauxic shift (di). Sporulation is the generation of a yeast spore by meiosis. Diauxic shift is the shift from anaerobic (fermentation) to aerobic (respiration) metabolism. The medium starts rich in glucose, and yeast cells ferment, producing ethanol. When the glucose is used up, they switch to ethanol as a source for carbon. Heat, cold, and reducing shock are various ways to stress the yeast cell. Figure (b) shows the average, plus or minus one standard deviation, of the data in Figure (a).

Self-Organizing Maps [Kohonen]

- Kind of neural network.
- Clusters data and find complex relationships between clusters.
- Helps reduce the dimensionality of the data.
- Map of 1 or 2 dimensions produced.
- Unsupervised Clustering
- Like K-Means, except for visualization

SOM Architectures

- 2-D Grid
- 3-D Grid
- Hexagonal Grid

SOM Algorithm

- Select SOM architecture, and initialize weight vectors and other parameters.
- **While** (stopping condition not satisfied) **do** for each input point x
  - winning node q has weight vector closest to x.
  - Update weight vector of q and its neighbors.
  - Reduce neighborhood size and learning rate.

SOM Algorithm Details

- Distance between x and weight vector: $\|x - w_i\|$
- Winning node: $q(x) = \min_i \|x - w_i\|$
- Weight update function (for neighbors):

$$w_i(k+1) = w_i(k) + \mu(k, x, i)[x(k) - w_i(k)]$$

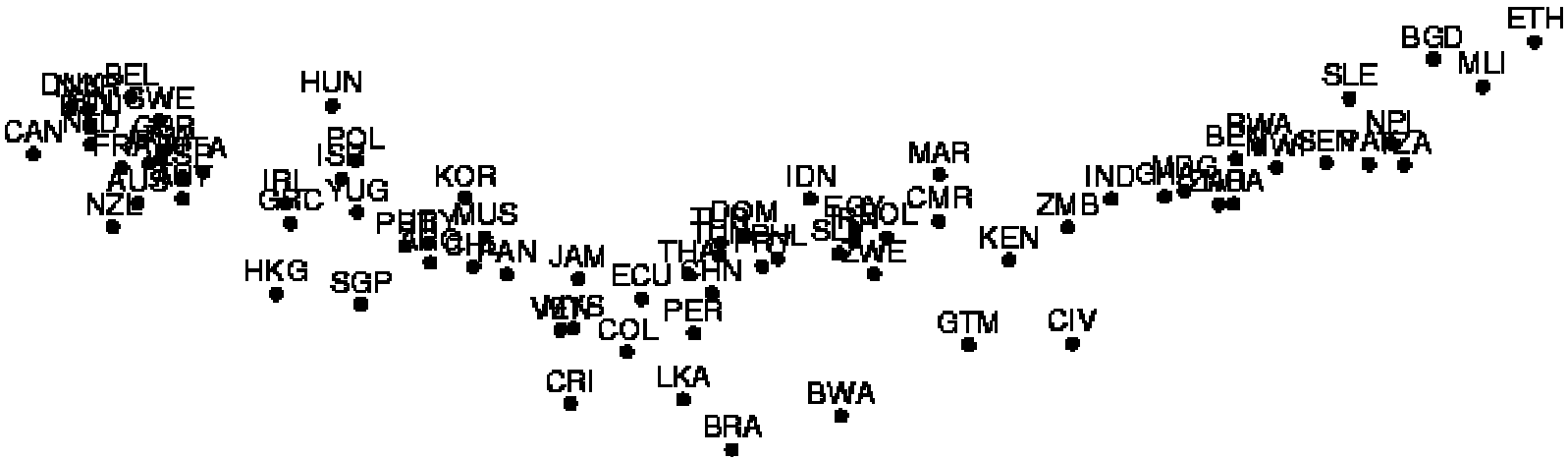
- Learning rate:

$$\mu(k, x, i) = \eta_0(k) \exp\left(\frac{-\|r_i - r_{q(x)}\|^2}{\sigma^2} \right)$$

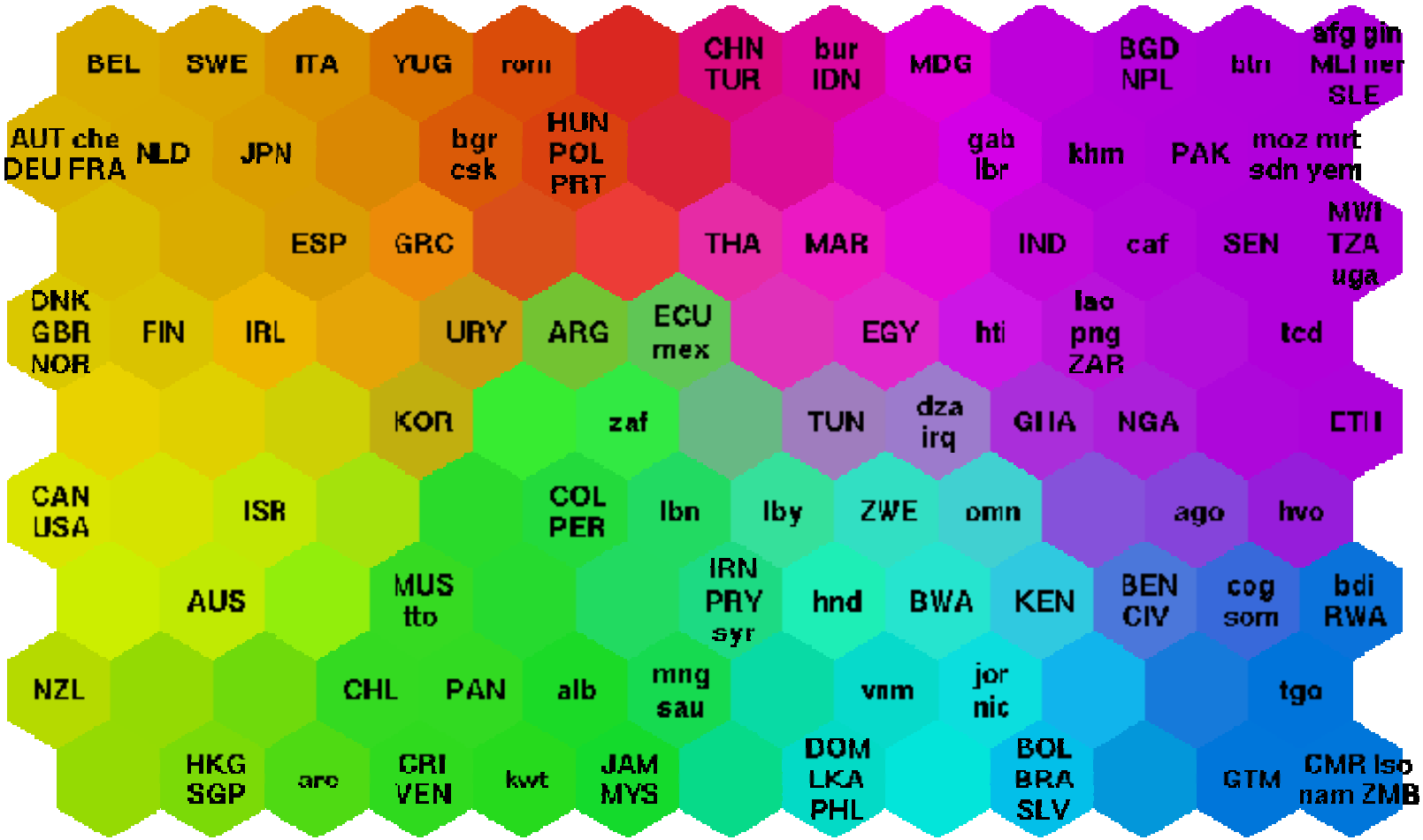
World Bank Statistics

- Data: World Bank statistics of countries in 1992.
- 39 indicators considered e.g., health, nutrition, educational services, etc.
- The complex joint effect of these factors can can be visualized by organizing the countries using the self-organizing map.

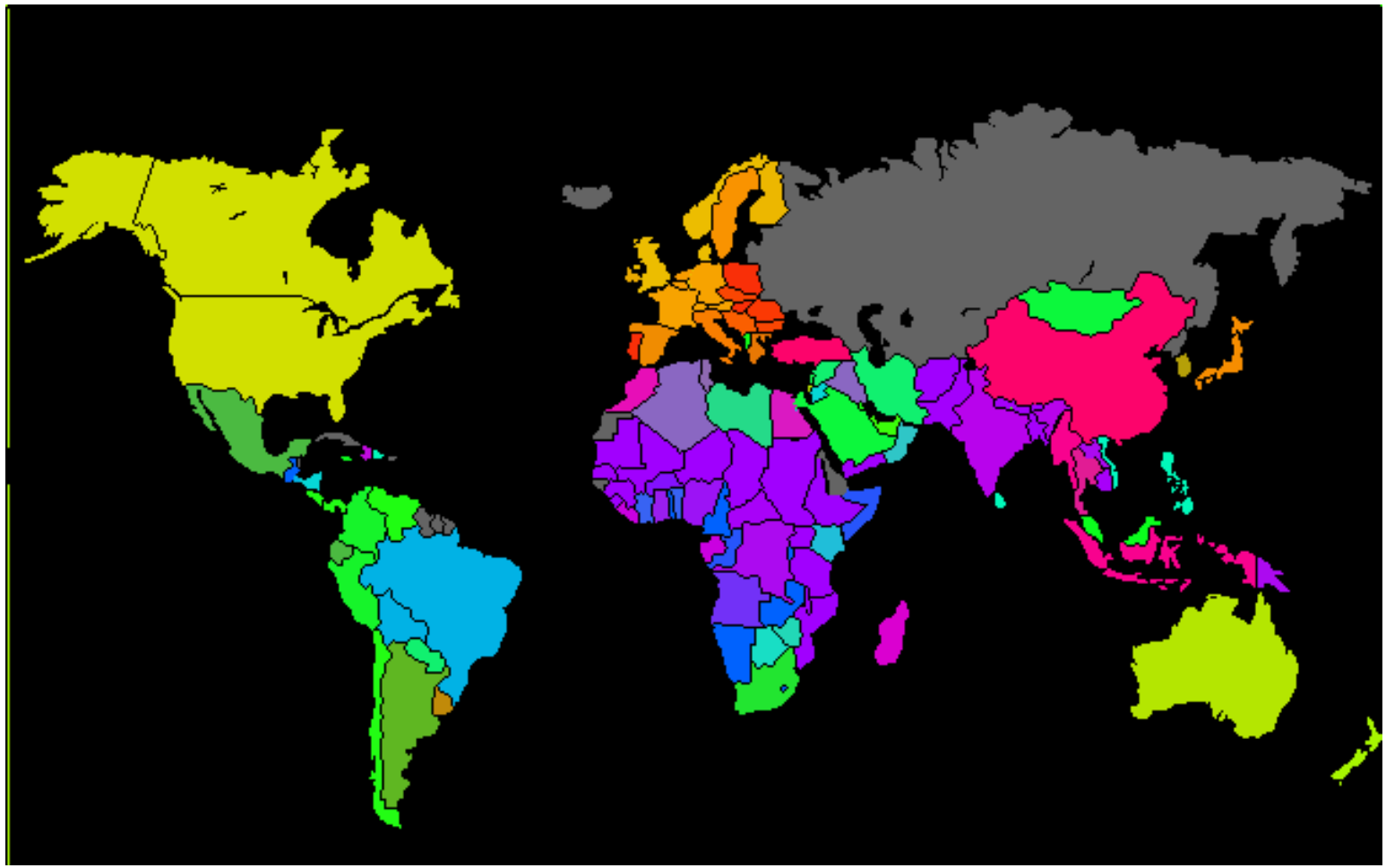
World Poverty PCA

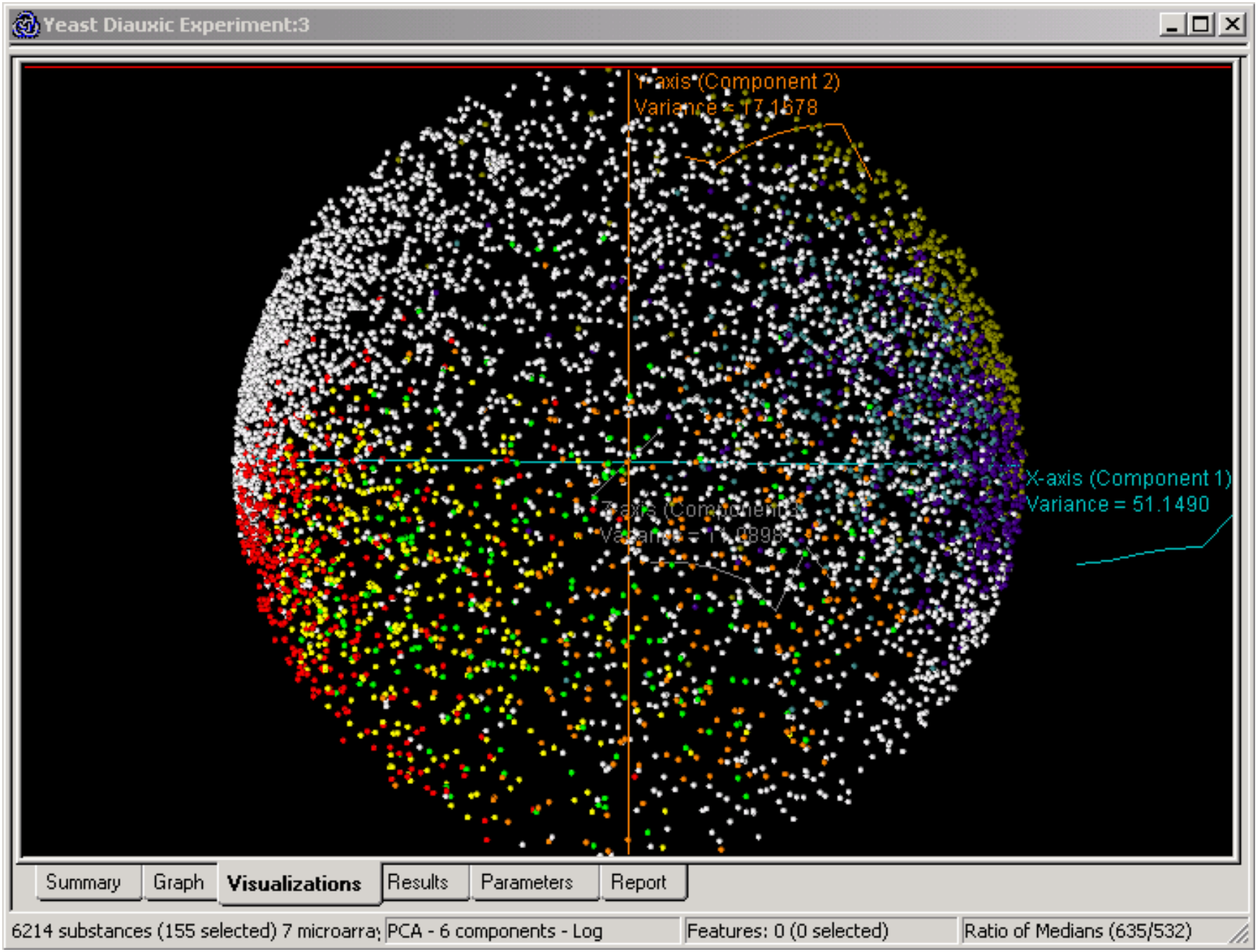


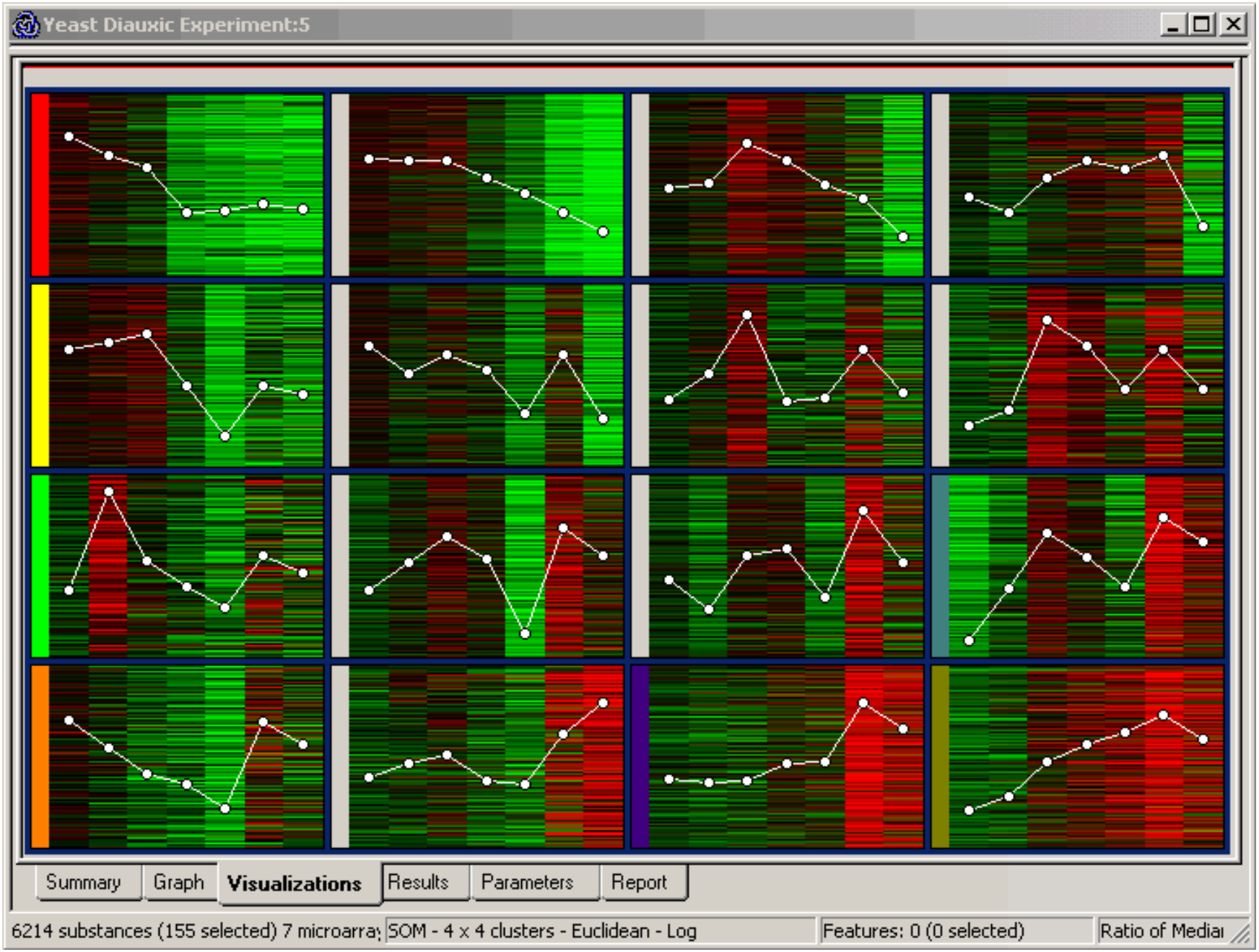
World Poverty SOM



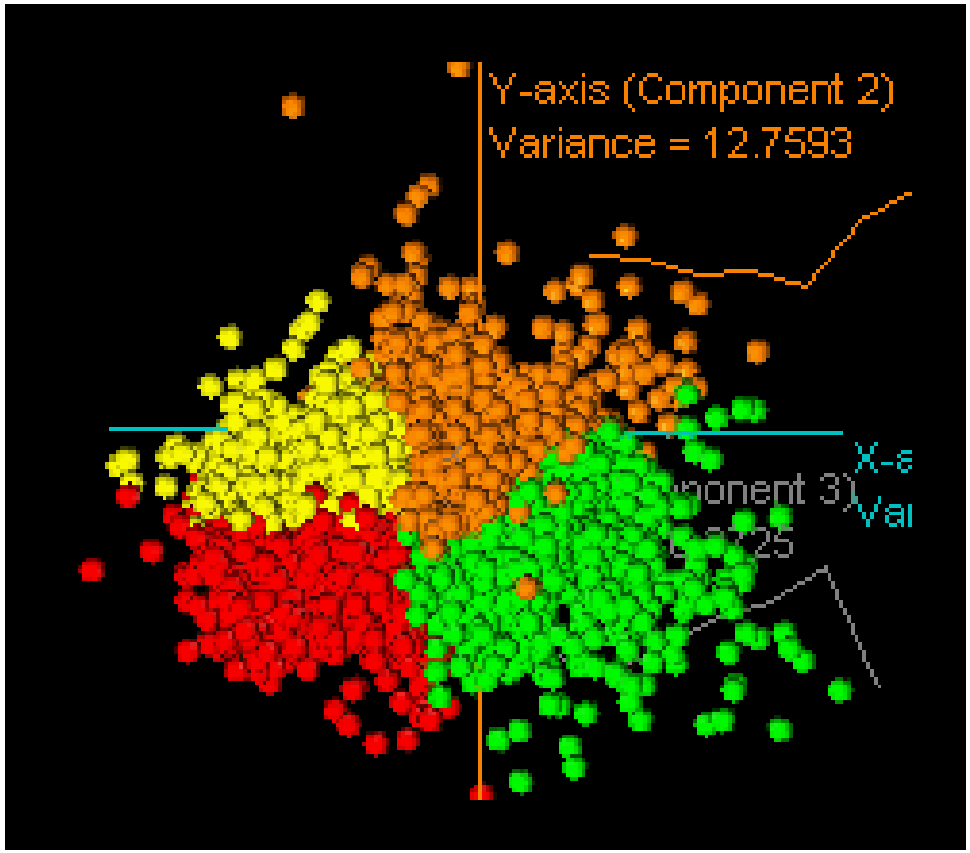
World Poverty Map



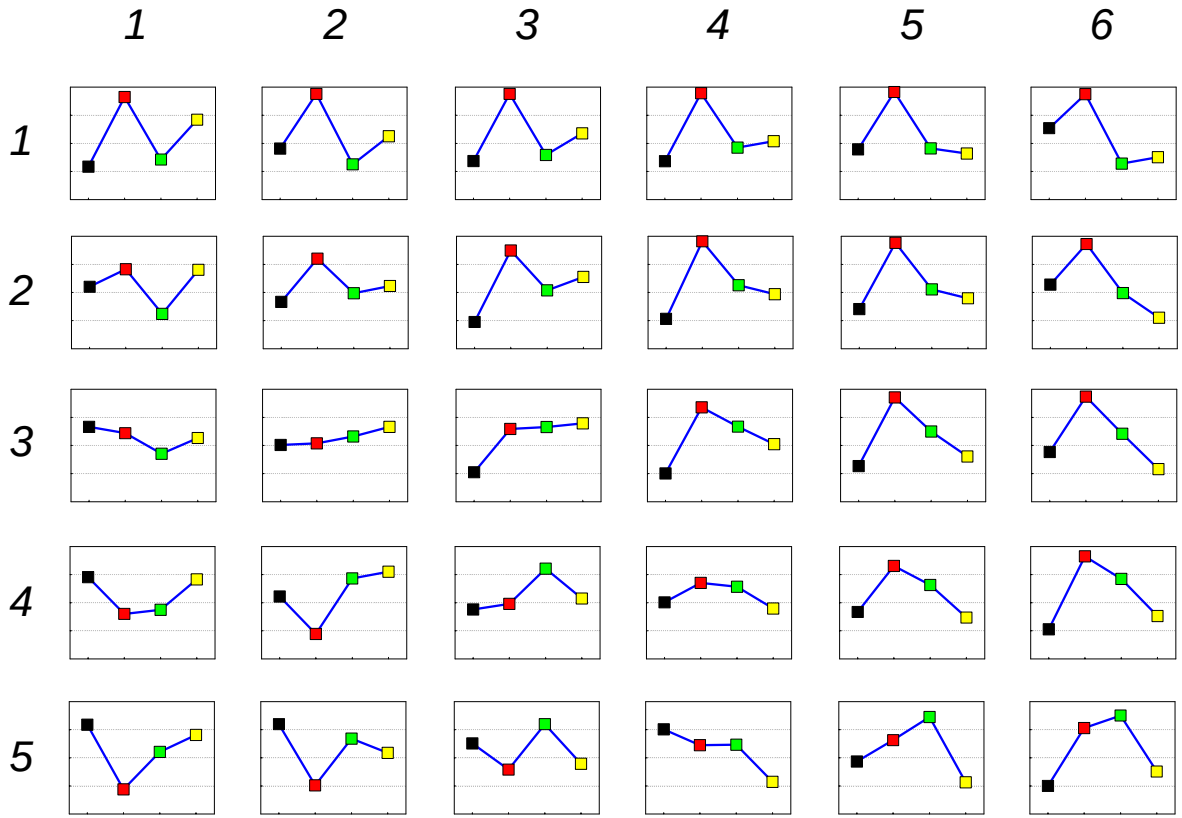

Yeast Diauxic Experiment:3
Y-axis (Component 2)
Variance = 17.1678
X-axis (Component 1)
Variance = 51.1490
Summary
Graph
Visualizations
Results
Parameters
Report
6214 substances (155 selected) 7 microarra
PCA - 6 components - Log
Features: 0 (0 selected)
Ratio of Medians (635/532)


Yeast Diauxic Experiment:5
Summary
Graph
Visualizations
Results
Parameters
Report
6214 substances (155 selected) 7 microarray
SOM - 4 x 4 clusters - Euclidean - Log
Features: 0 (0 selected)
Ratio of Median

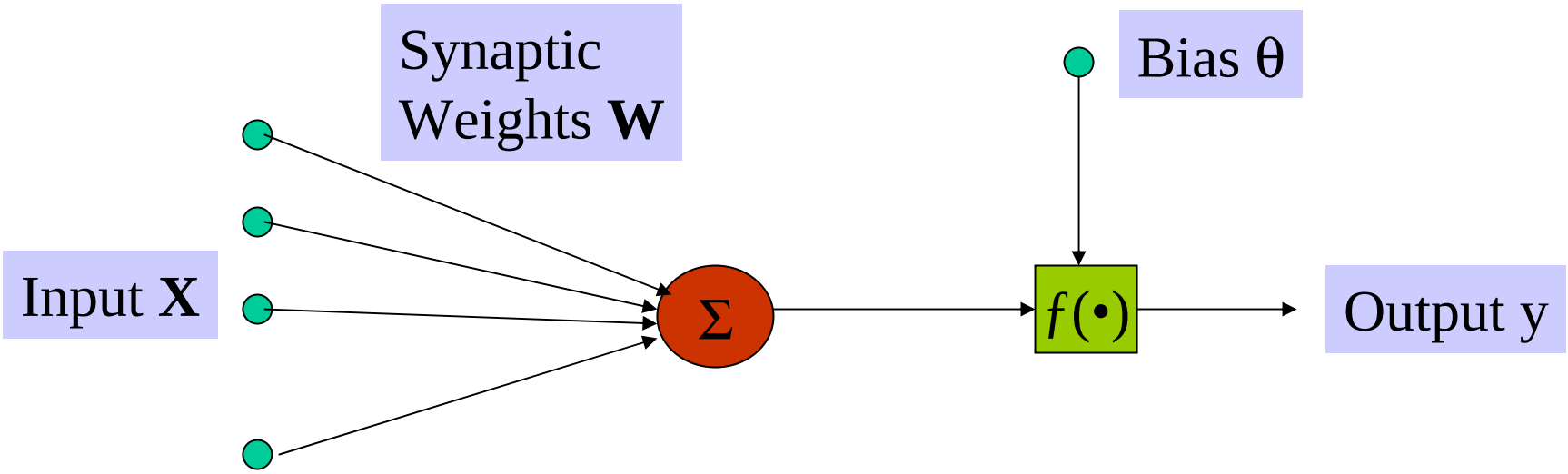
Viewing SOM Clusters on PCA axes



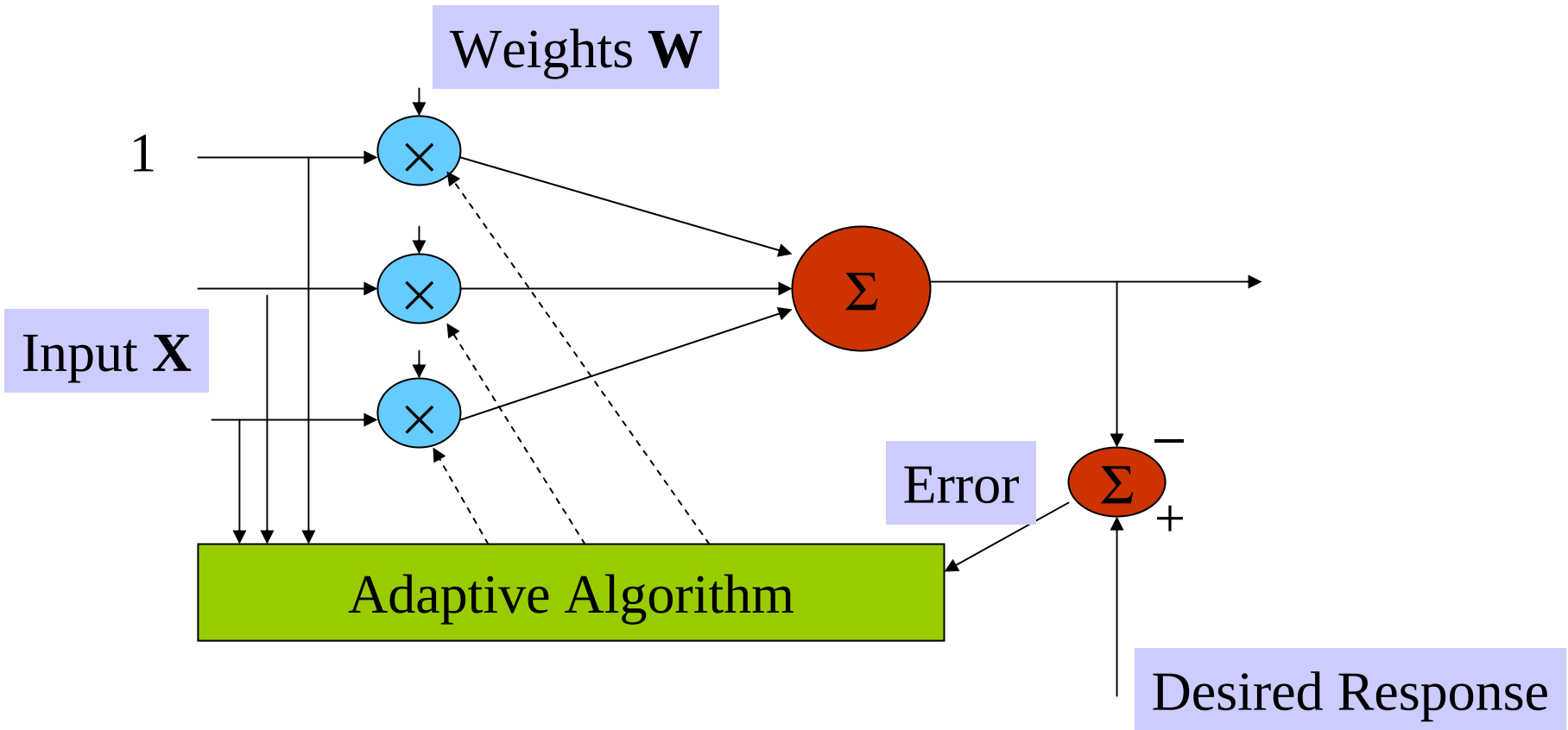
SOM Example [Xiao-rui He]



Neural Networks



Learning NN


Weights **W**
1
Input **X**
Σ
Error
Σ
−
+
Adaptive Algorithm
Desired Response

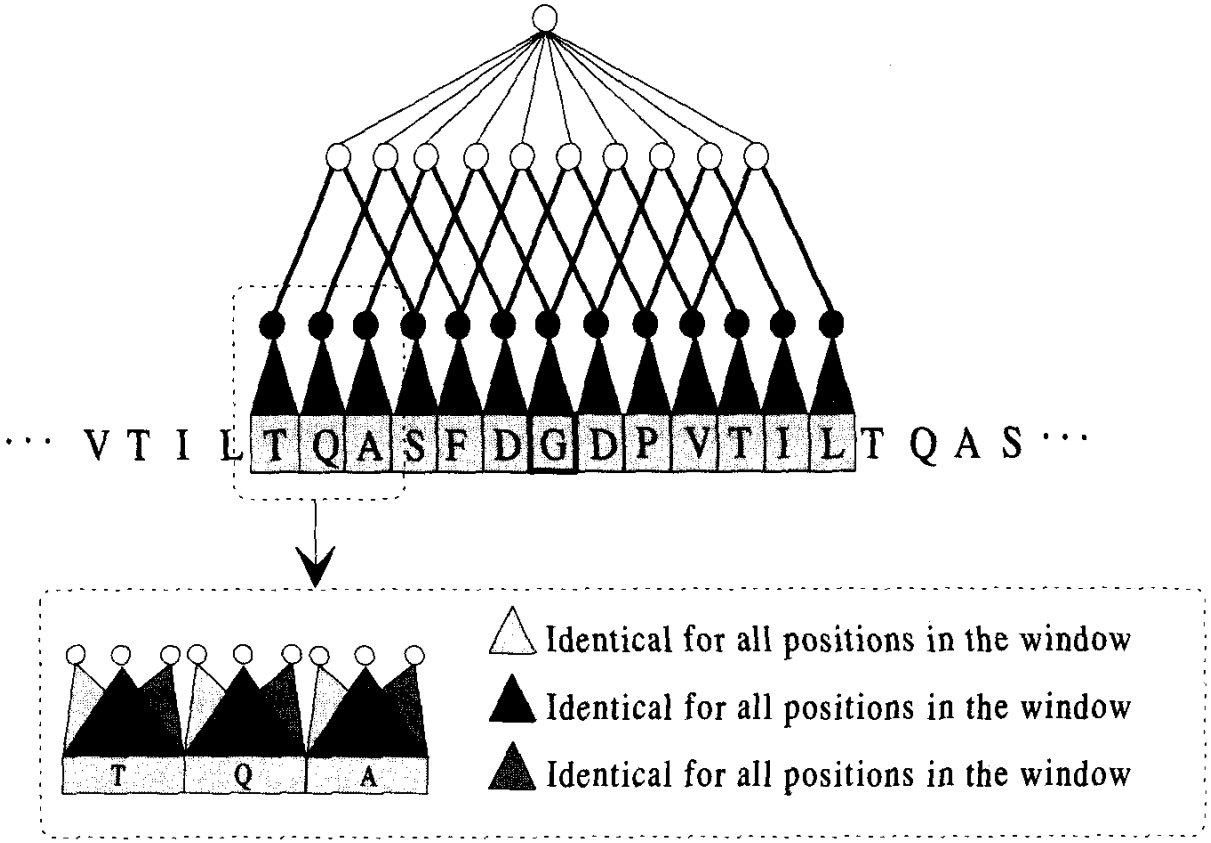
Types of NNs

- Recurrent NN
- Feed-forward NN
- Layered

Other issues

- Hidden layers possible
- Different activation functions possible

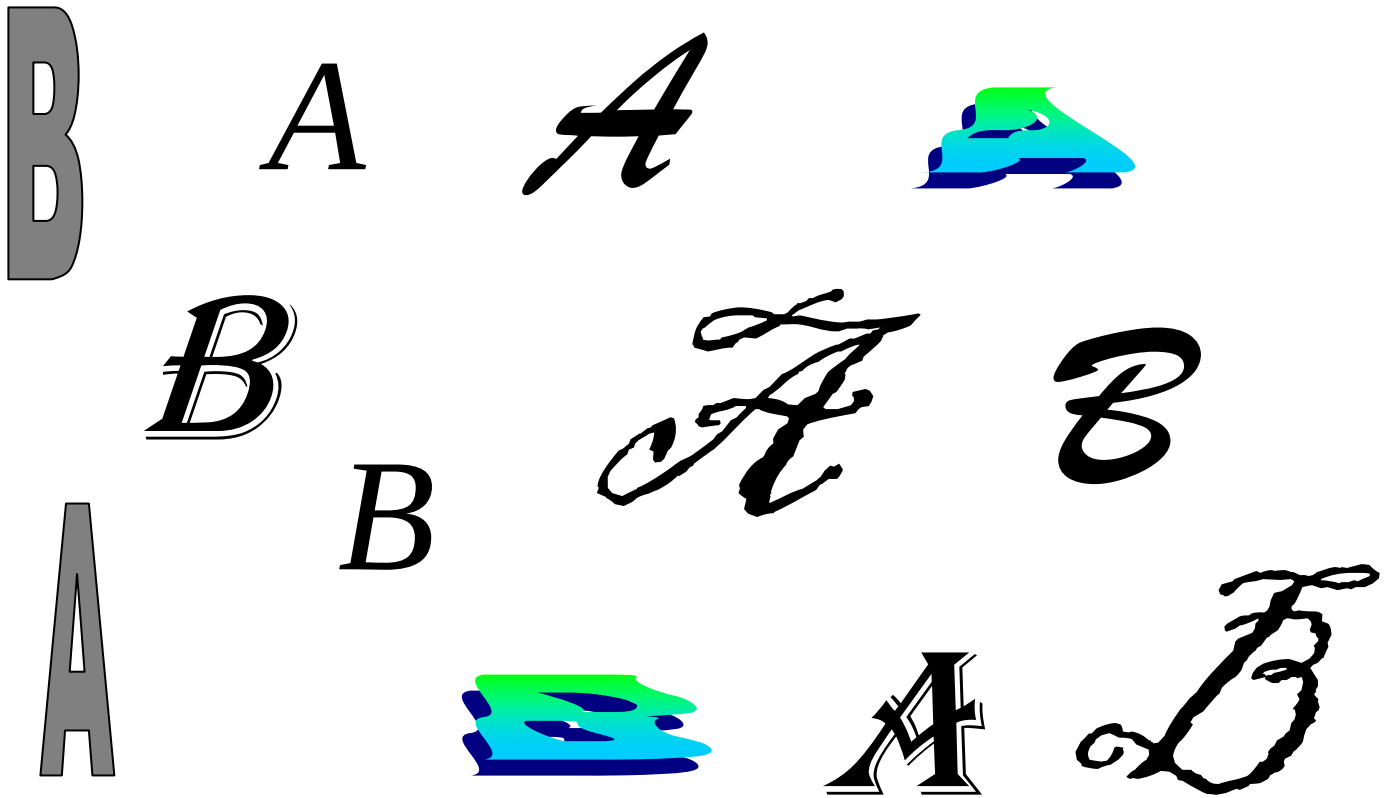
Application: Secondary Structure Prediction



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



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.

Learning Problems

- Binary Classification
- Multi-class classification
- Regression

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.

15

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.

16

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.

18

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 (NAC)
	YLR406C	RPL31B	FN	ribosomal protein L31B (L34B) (YL28)
	YLR075W	RPL10	FP	ribosomal protein L10
	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

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.

20

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).

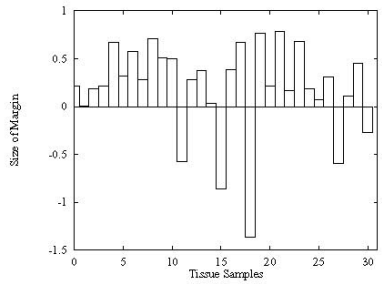


Figure 1: **SVM classification margins for ovarian tissues.** When classifying, the SVM calculates a margin which is the distance of an example from the decision boundary it has learned. In this graph, the margin for each tissue sample calculated using (10) is shown. A positive value indicates a correct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

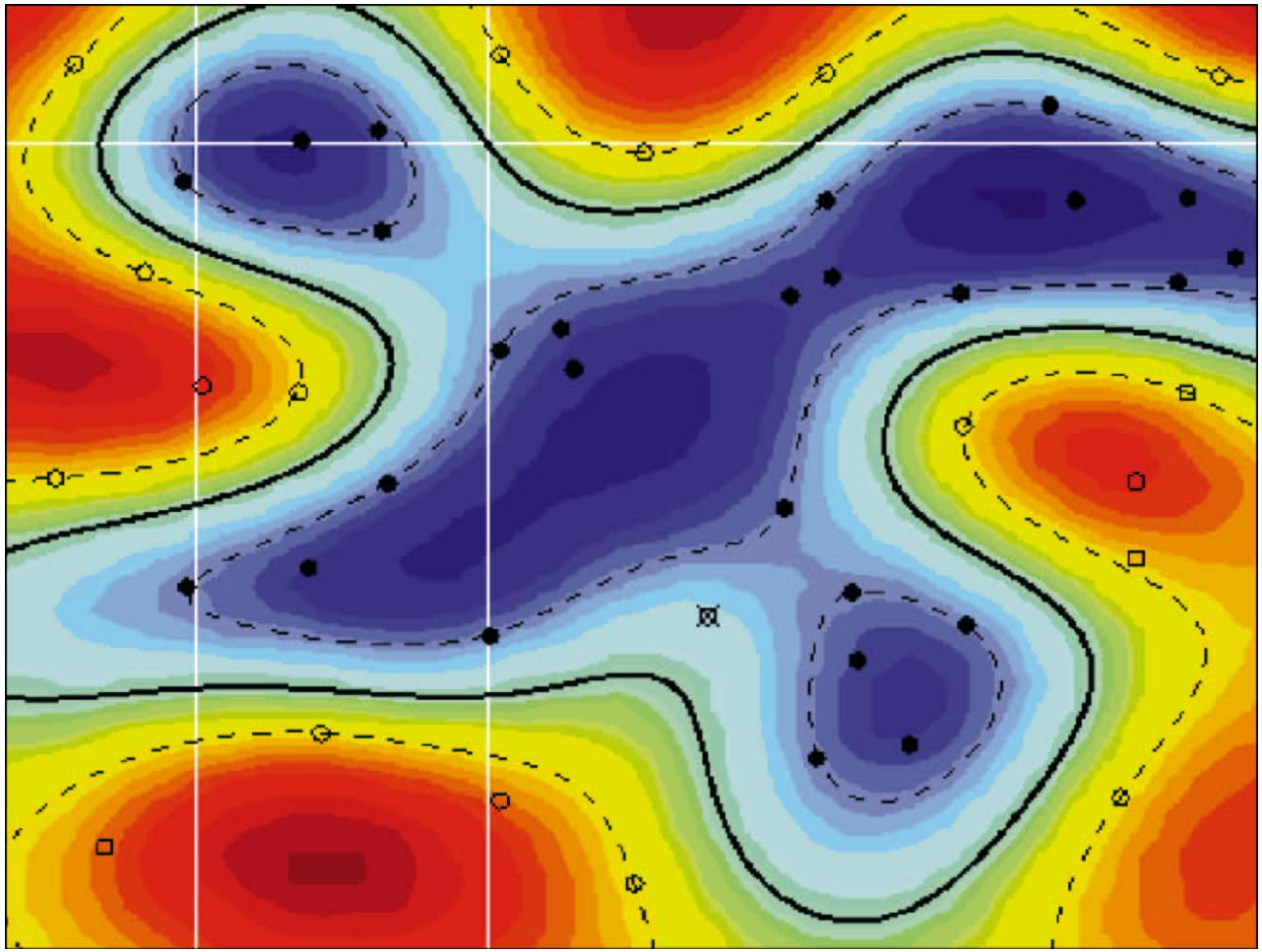
Dataset	Features	FP	FN	SVM FP	SVM FN
Ovarian(original)	97802	4.6	4.8	5	3
Ovarian(modified)	97802	4.4	3.4	0	0
AML/ALL train	7129	0.6	2.8	0	0
AML treatment	7129	4.8	3.5	3	2
Colon	2000	3.8	3.7	3	3

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 – General Principles

- SVMs perform binary classification by partitioning the feature space with a surface implied by a subset of the training points (vectors) near the separating surface. These vectors are referred to as Support Vectors.
- Efficient with high-dimensional data.
- Solid statistical theory
- Subsume several other methods.

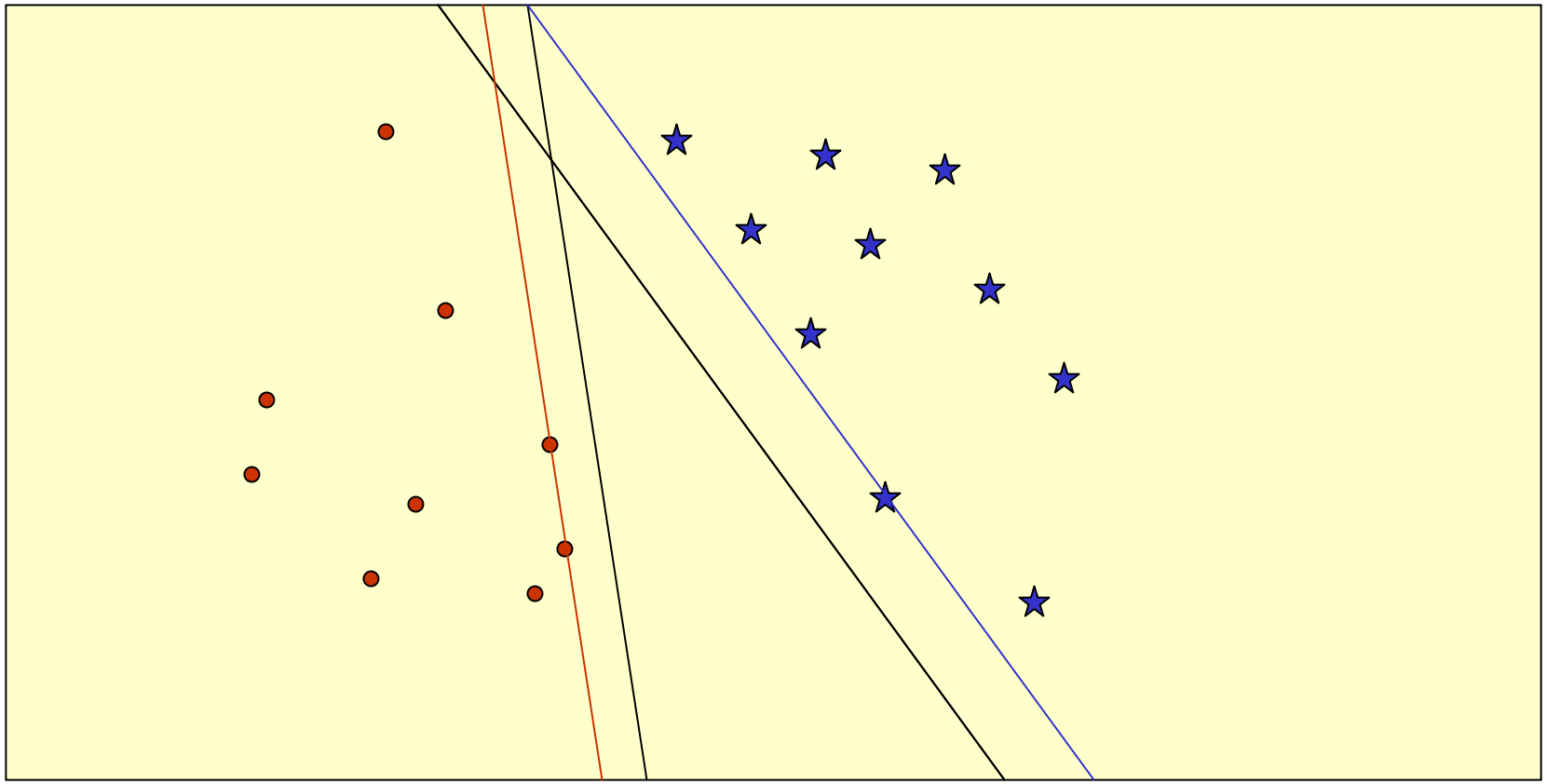
SVM Example (Radial Basis Function)



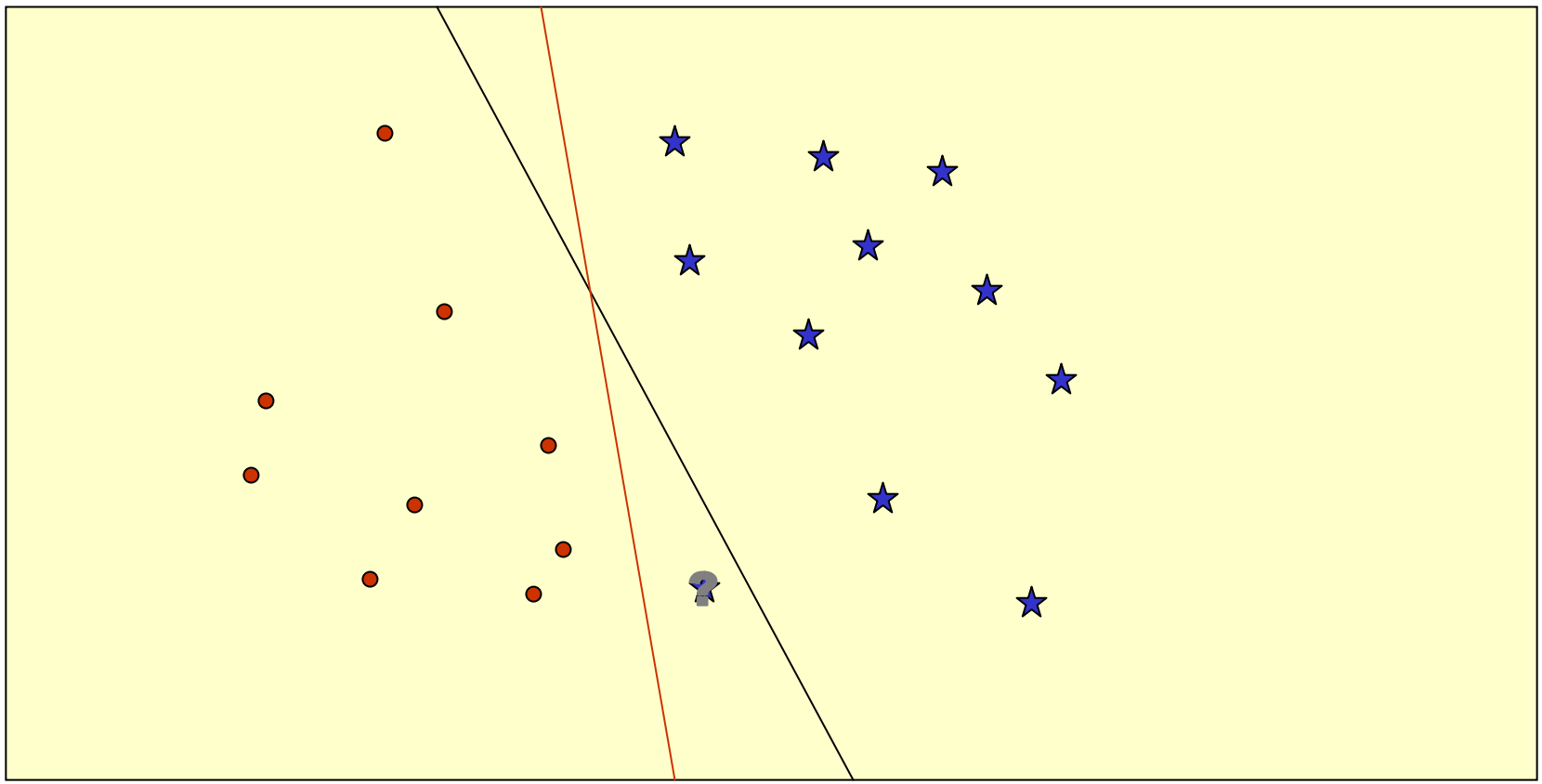
SVM Ingredients

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

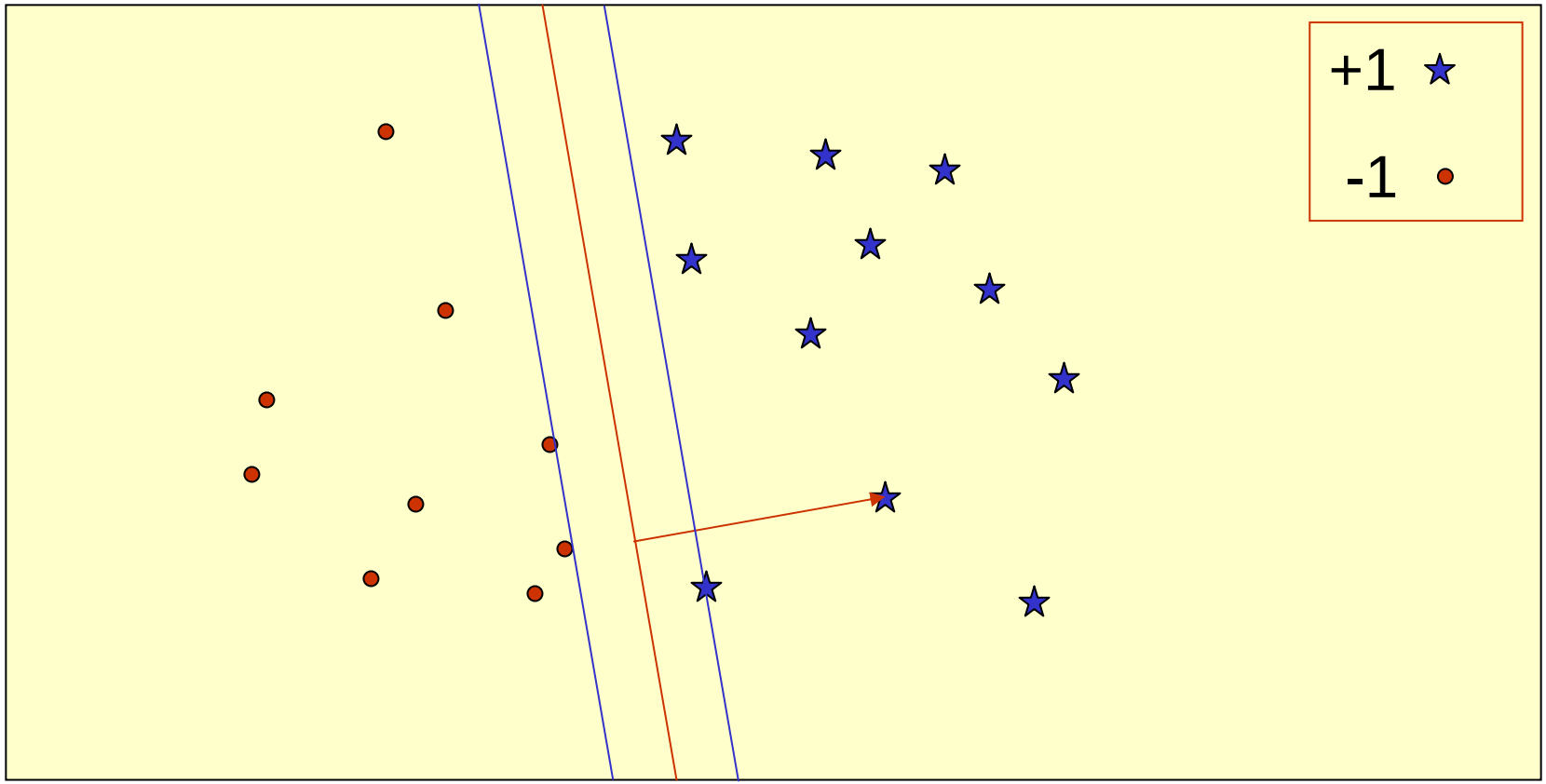
Classification of 2-D (Separable) data



Classification of (Separable) 2-D data

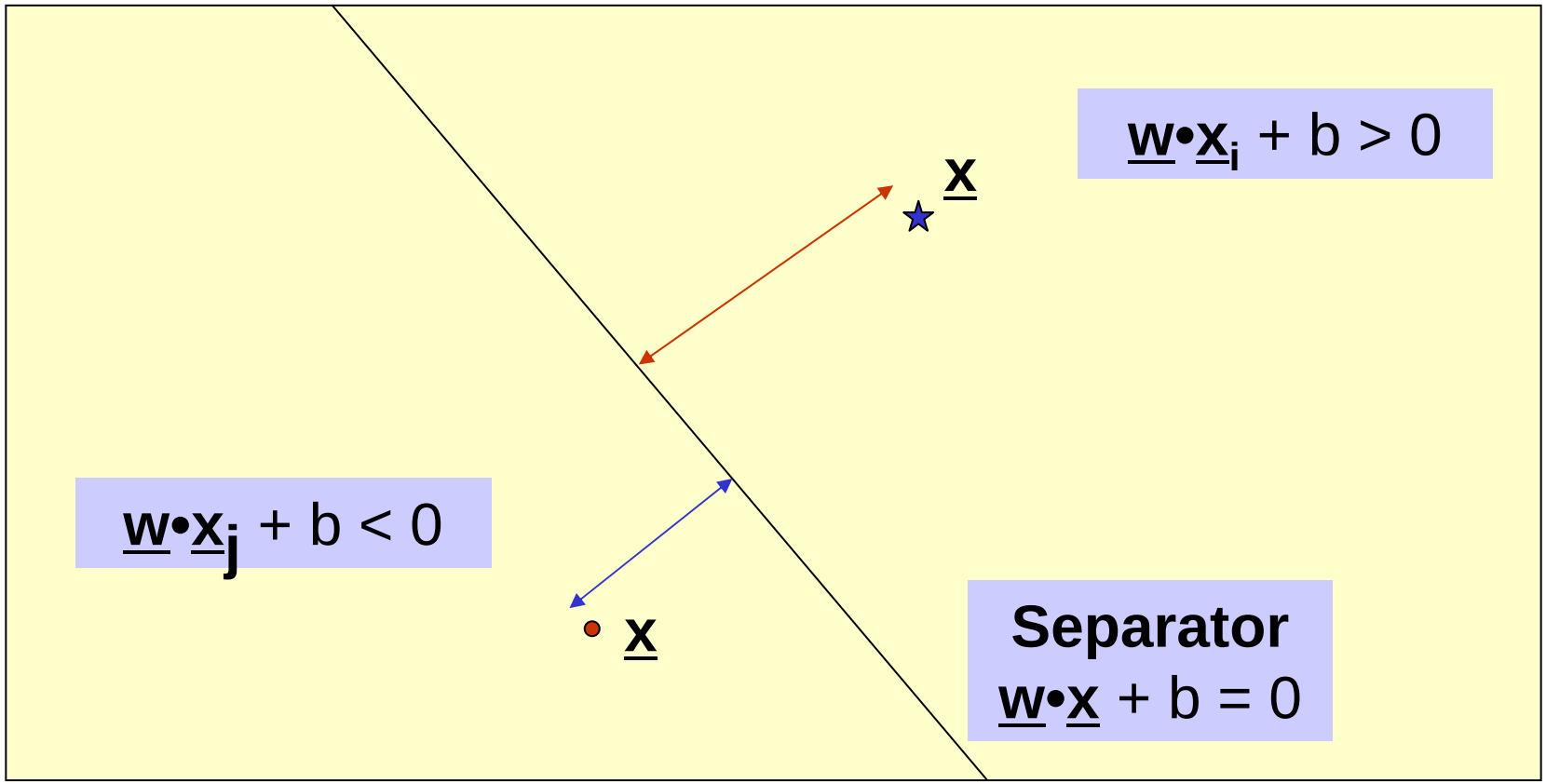


Classification of (Separable) 2-D data



- Margin of a point
- Margin of a point set

Classification using the Separator



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 $y_i\,(\underline{\mathbf{w}}_k \bullet \underline{x}_i + b_k) \leq 0$ **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**

$a_i = a_i + 1$

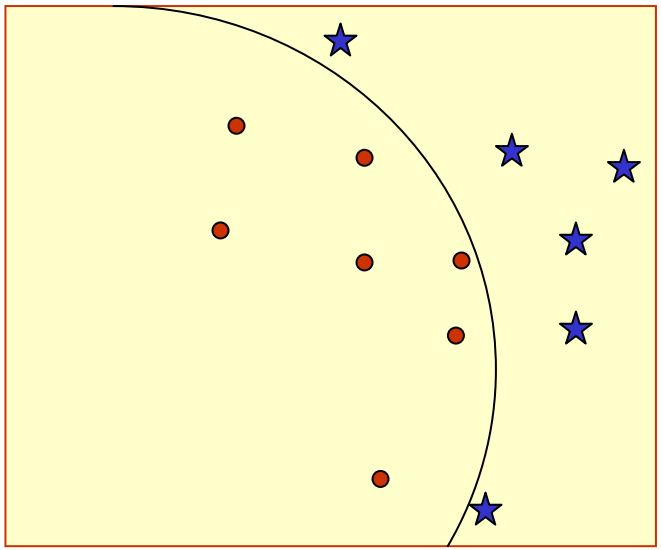
$b = b + y_i R^2$

endif

Until no mistakes made within loop

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

Non-linear Separators



Main idea: Map into feature space

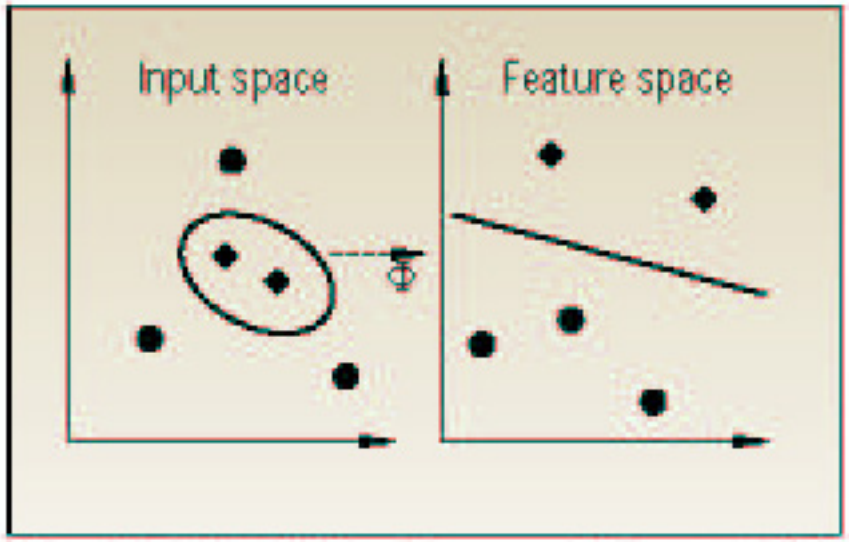
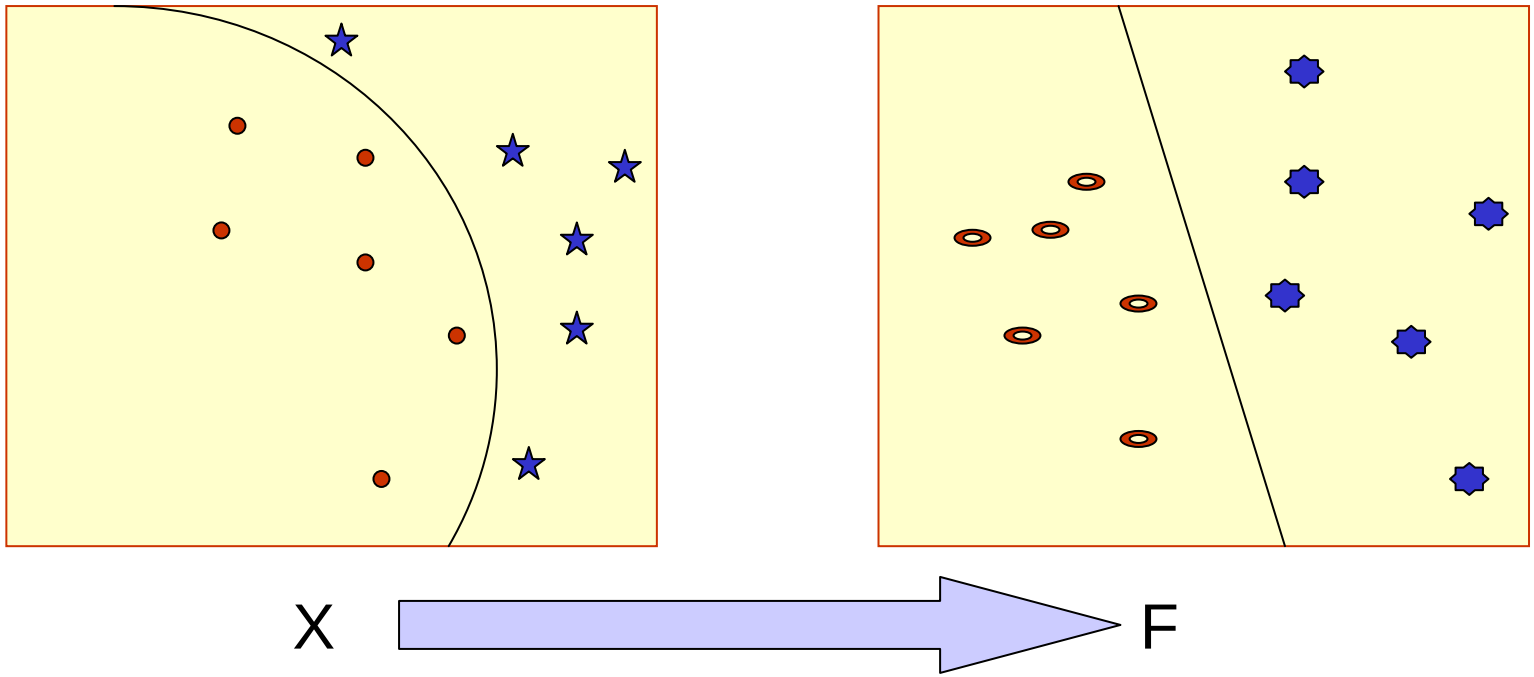


Figure 2. The idea of SV machines: map the training data nonlinearly into a higher-dimensional feature space via Φ, and construct a separating hyperplane with maximum margin there. This yields a nonlinear decision boundary in input space. By the use of a kernel function, it is possible to compute the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

Non-linear Separators



Useful URLs

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

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

      $a_i = a_i + 1$

      $b = b + y_i R^2$

Until no mistakes made within loop

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

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

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

		Learned threshold					Optimized threshold				
Class	Method	FP	FN	TP	TN	Cost	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.

15

		Learned threshold					Optimized threshold				
Class	Method	FP	FN	TP	TN	Cost	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.

16

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.

18

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 (NAC)
	YLR406C	RPL31B	FN	ribosomal protein L31B (L34B) (YL28)
	YLR075W	RPL10	FP	ribosomal protein L10
	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

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.

20

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).

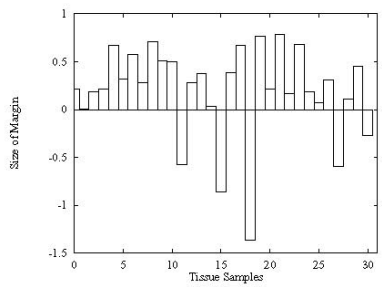
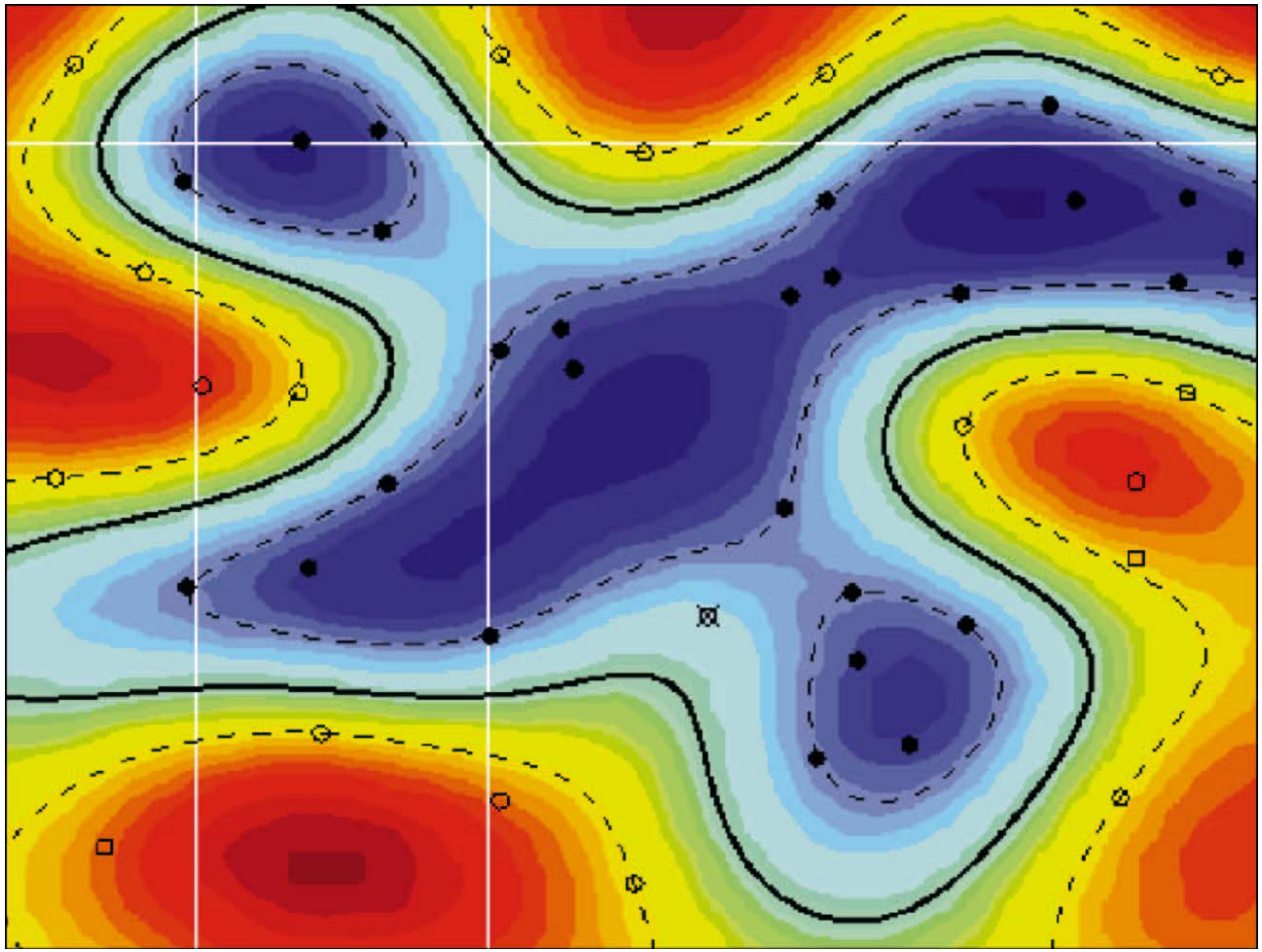


Figure 1: **SVM classification margins for ovarian tissues.** When classifying, the SVM calculates a margin which is the distance of an example from the decision boundary it has learned. In this graph, the margin for each tissue sample calculated using (10) is shown. A positive value indicates a correct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



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 facto sults.

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!)

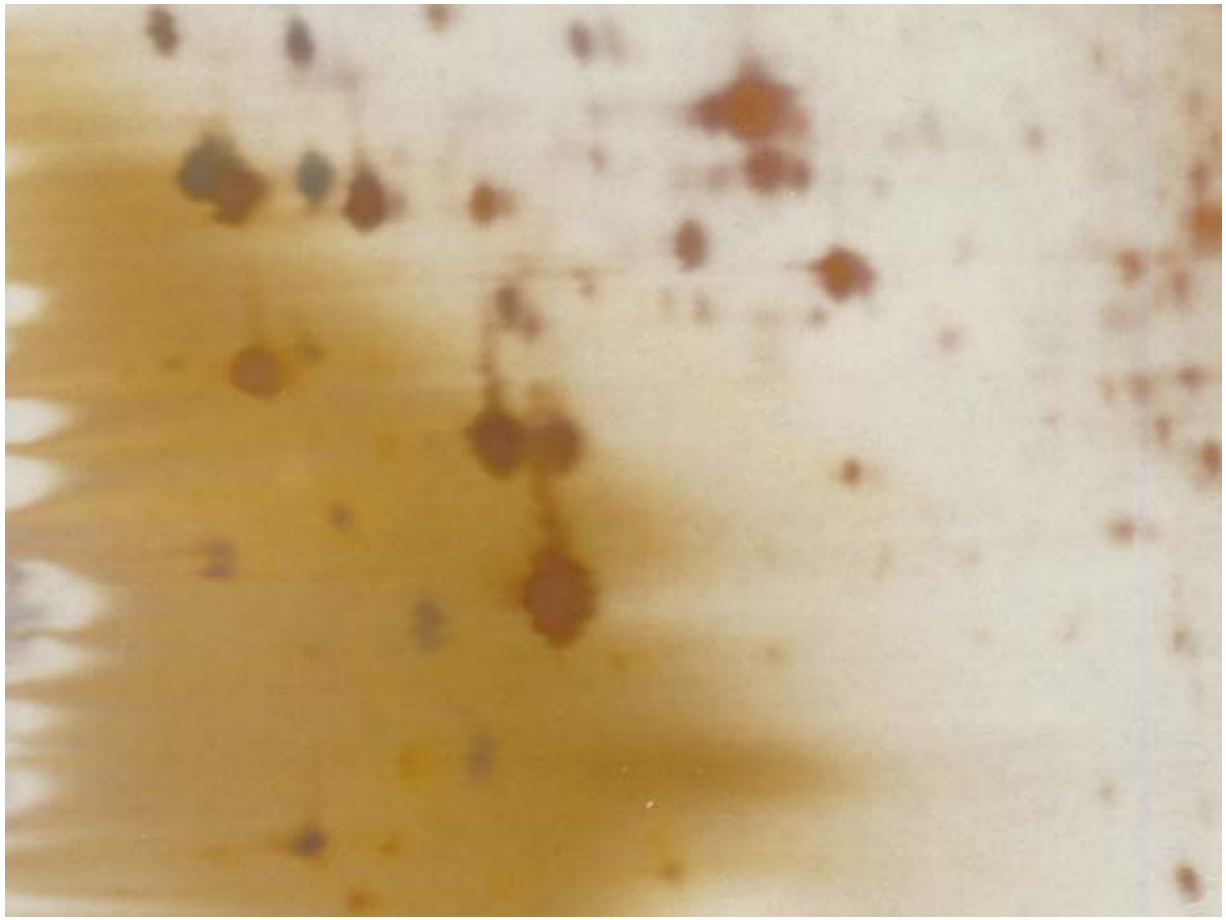
Genomics (Cont'd)

- 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

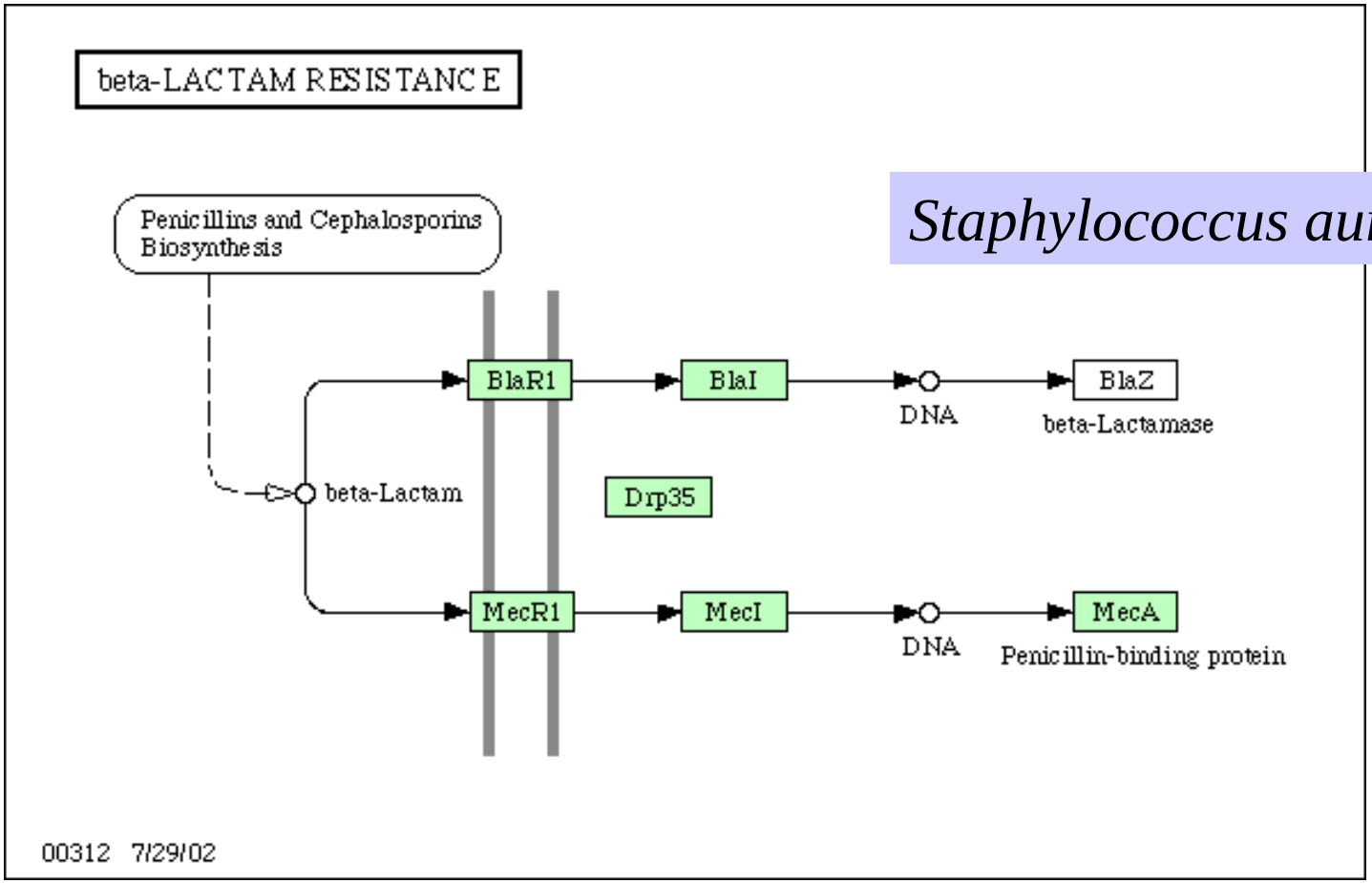
Databases for Comparative Genomics

- 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

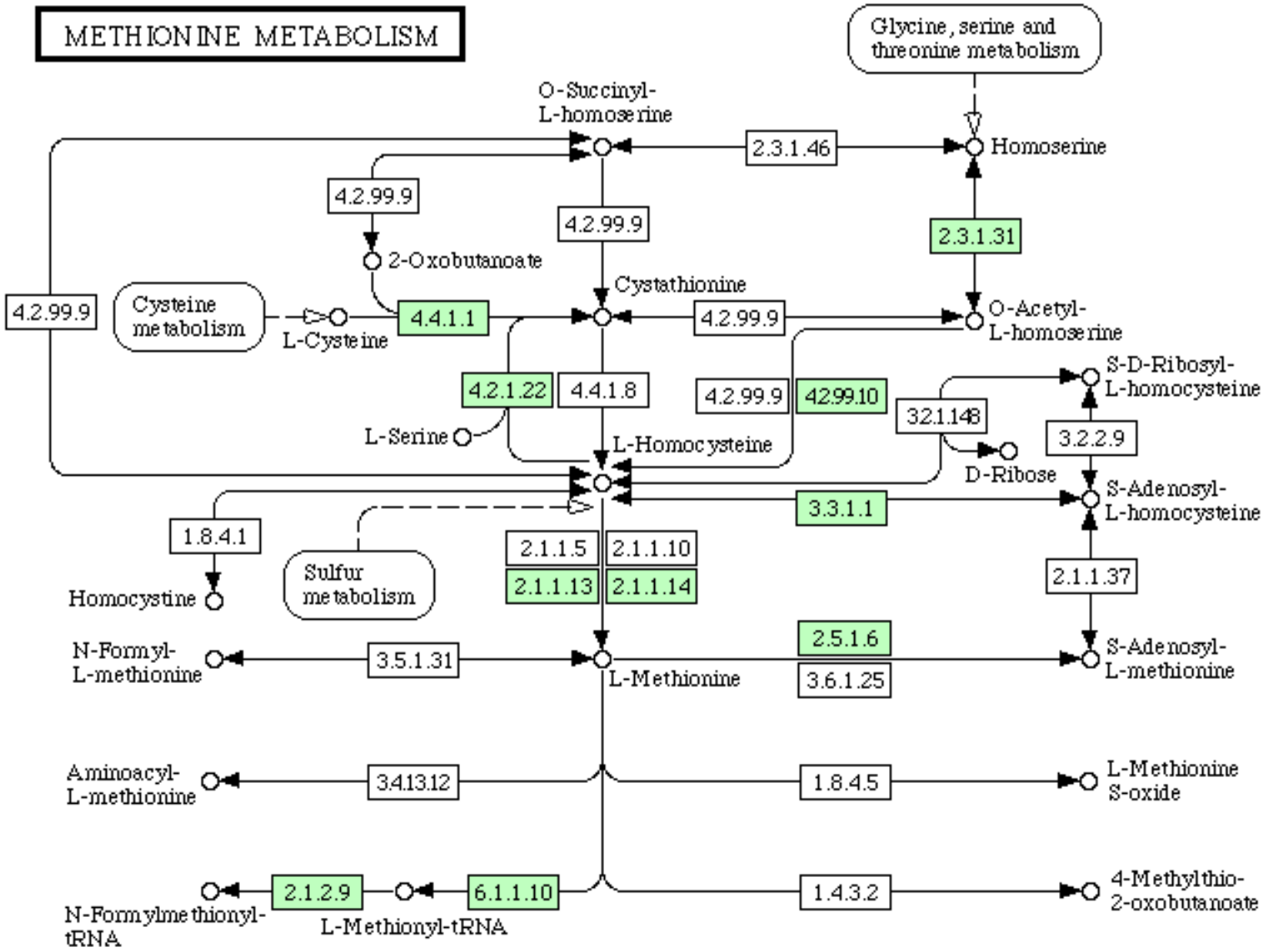
- Genes & Proteins act in concert and therefore form a complex network of dependencies.

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.

Start and Stop Codon Distribution

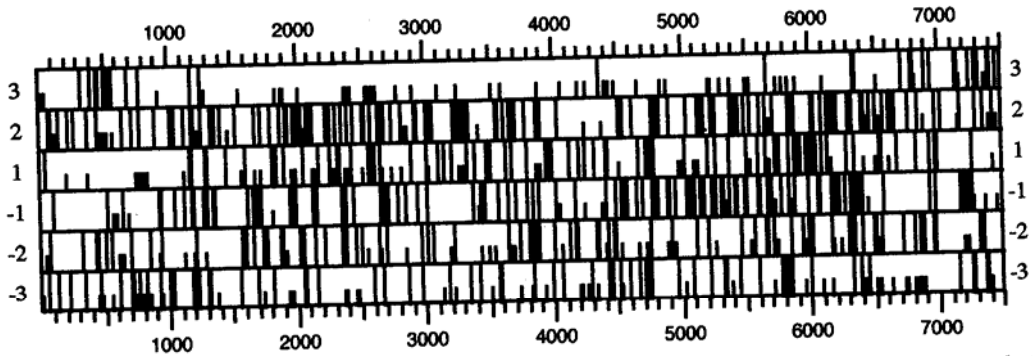


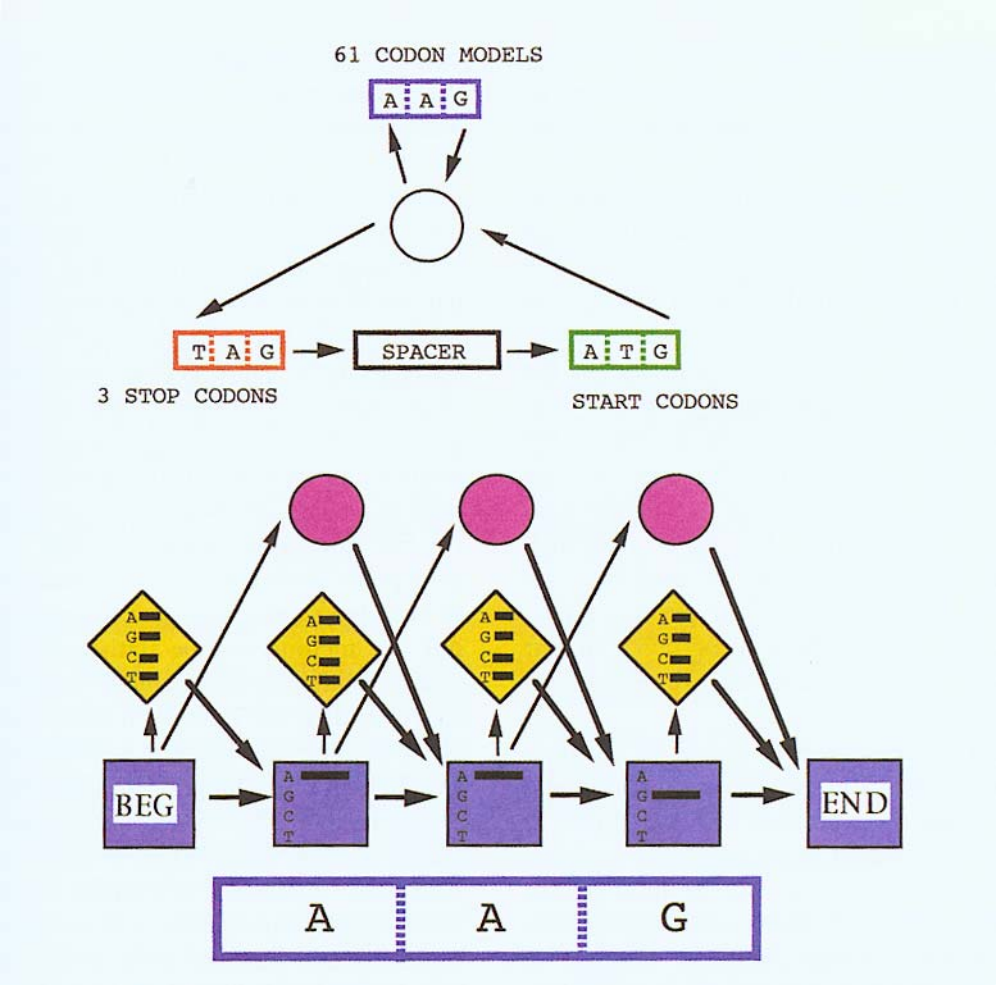
FIGURE 9.1. ORF map of a portion of the *E. coli lac* operon using the DNA STRIDER program (Marck 1988). Shown are AUG and termination codons as one-half and full vertical bars, respectively, in all six possible reading frames. The *lacZ* gene is visible as an ORF that runs from positions 1284 to 4355 in frame 3.

Genetic Code

First letter	Second letter: U		Second letter: C		Second letter: A		Second letter: G		Third letter
U	UUU UUC	Phenyl-alanine	UCU UCC UCA UCG	Serine	UAU UAC	Tyrosine	UGU UGC	Cysteine	U C A G
	UUA UUG	Leucine			UAA UAG	Stop codon Stop codon	UGA	Stop codon	
							UGG	Tryptophan	
C	CUU CUC CUA CUG	Leucine	CCU CCC CCA CCG	Proline	CAU CAC	Histidine	CGU CGC CGA CGG	Arginine	U C A G
					CAA CAG	Glutamine			
A	AUU AUC AUA	Isoleucine	ACU ACC ACA ACG	Threonine	AAU AAC	Asparagine	AGU AGC	Serine	U C A G
	AUG	Methionine; initiation codon			AAA AAG	Lysine	AGA AGG	Arginine	
G	GUU GUC GUA GUG	Valine	GCU GCC GCA GCG	Alanine	GAU GAC	Aspartic acid	GGU GGC GGA GGG	Glycine	U C A G
					GAA GAG	Glutamic acid			

Slide courtesy http://www.emc.maricopa.edu/faculty/farabee/BIOBK/code.gif

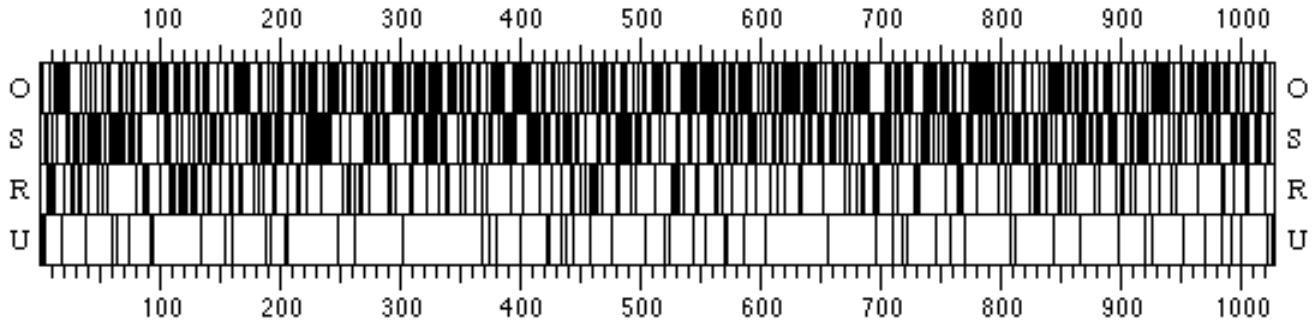
Recognizing Codons



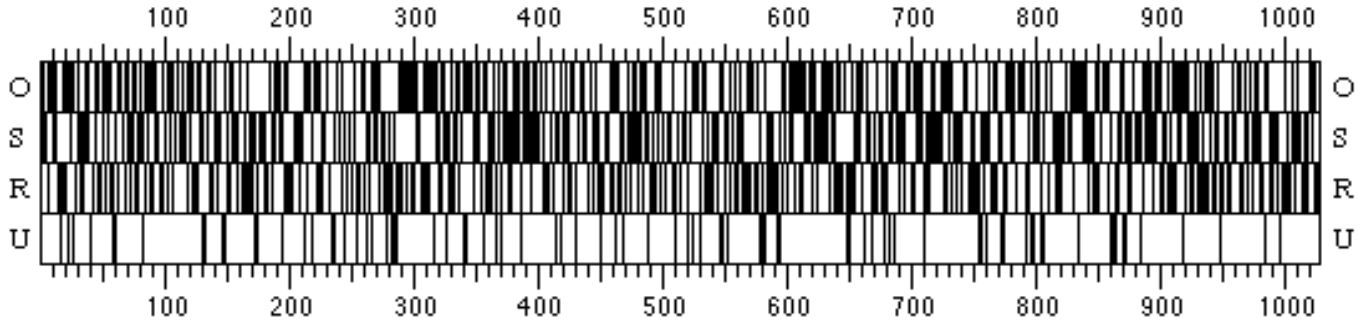
Codon Bias

- Some codons preferred over others.

O = optimal
S = suboptimal
R = rare
U = unfavorable



Frame Shift 1

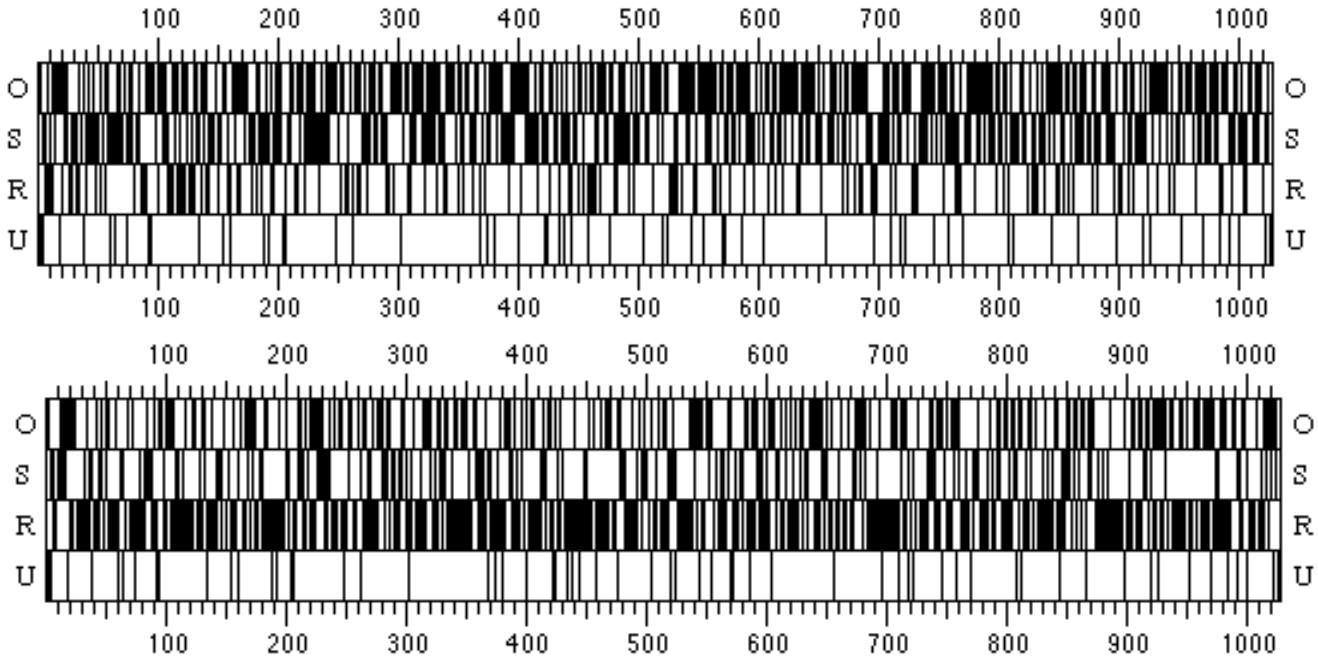


Frame Shift 2

Codon Bias

- Codon biases specific to organisms

O = optimal
S = suboptimal
R = rare
U = unfavorable



Same Frames; Different labeling of codon types (i.e., from yeast)

Eukaryotic Gene Prediction

- Complicated by introns & alternative splicing
- Exons/introns have different GC content.
- Many other measures distinguish exons/introns
- Software:
  - GENEPARSER Snyder & Stormo (NN)
  - GENIE Kulp, Haussler, Reese, Eckman (HMM)
  - GENSCAN Burge, Karlin (Decision Trees)
  - XGRAIL Xu, Einstein, Mural, Shah, Uberbacher (NN)
  - PROCRUSTES Gelfand (Formal Languages)
  - MZEF Zhang

Introns/Exons in *C. elegans*

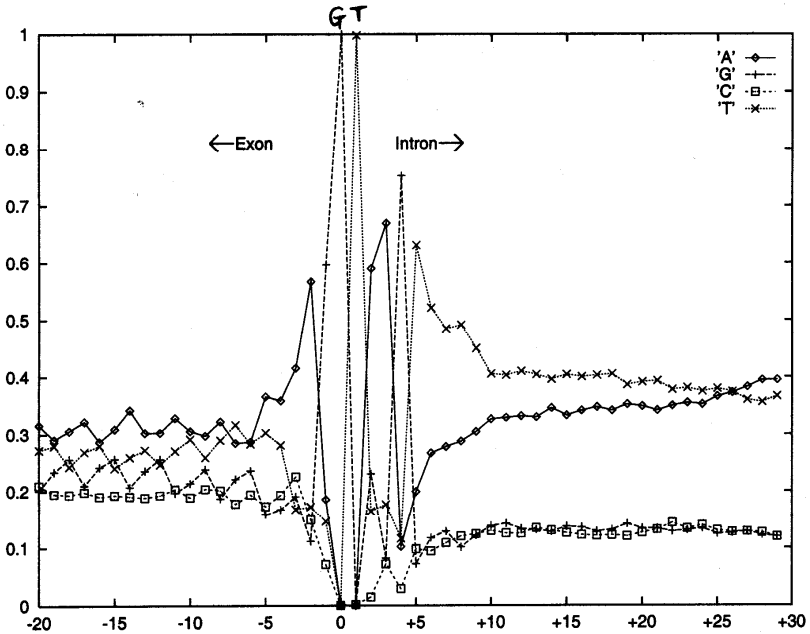


Figure 2: Profile of the same 5' collection but around a larger window.

A/T

G/C

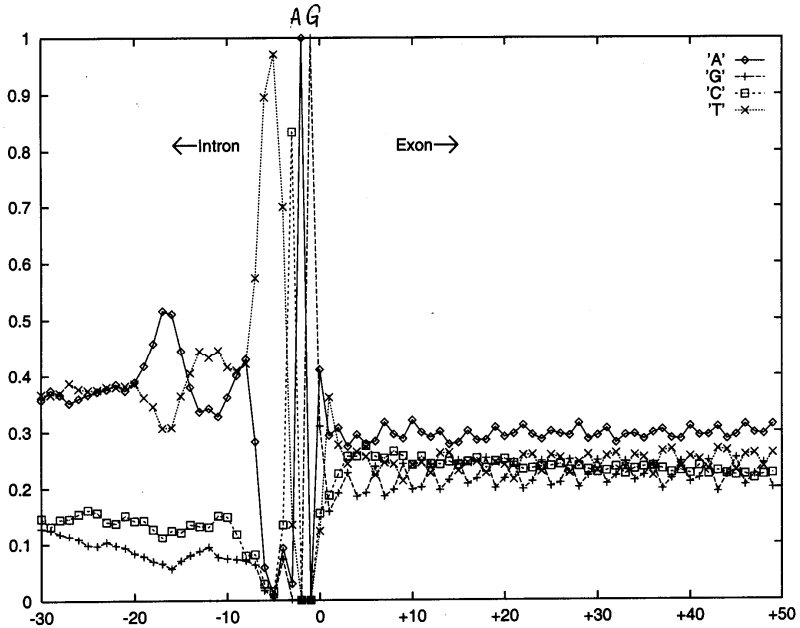
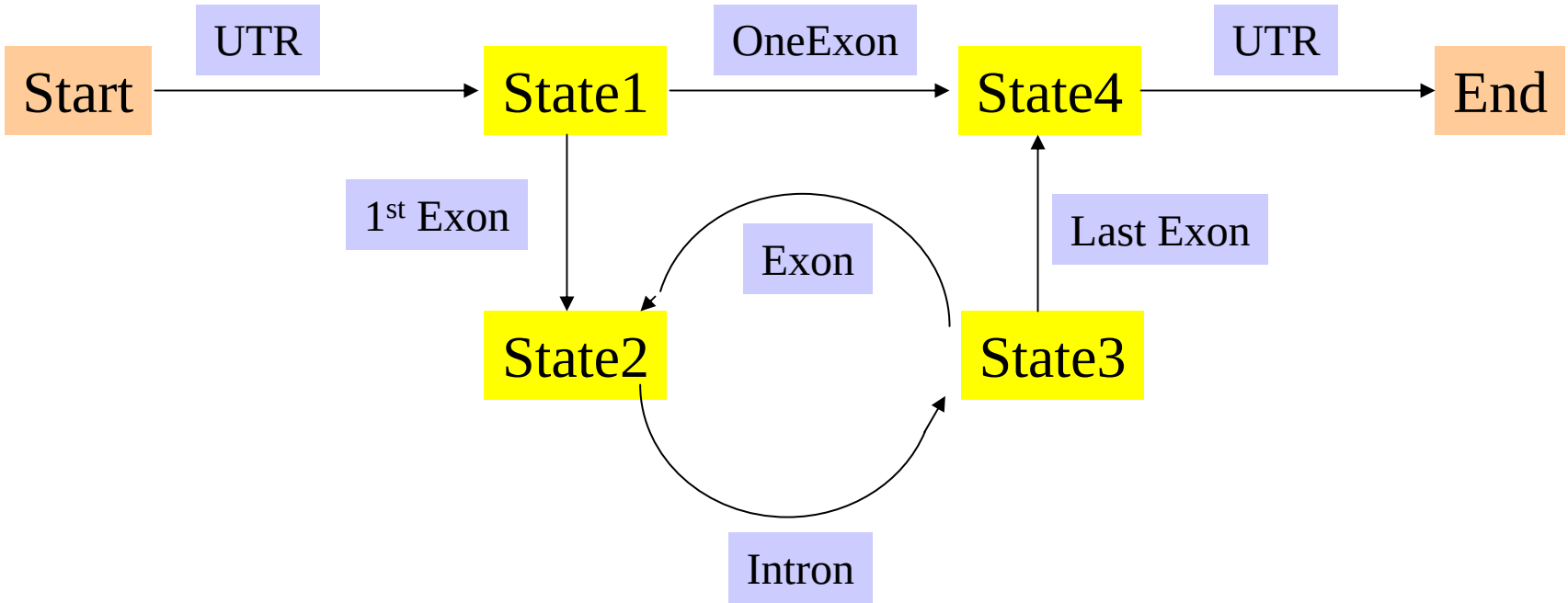


Figure 4: Profile of 8,192 sequences of length 80 around the 3' site. The first position in the exon is labeled 0.

- 8192 Introns in *C. elegans* : [**GT**...**AG**]
- Vary in lengths from 30 to over 600; Complexity varies

HMM structure for Gene Finding



Motifs in Protein Sequences

Motifs are combinations of secondary structures in proteins with a specific **structure** and a specific **function**. They are also called **super-secondary structures**.

Examples: Helix-Turn-Helix, Zinc-finger, Homeobox domain, Hairpin-beta motif, Calcium-binding motif, Beta-alpha-beta motif, Coiled-coil motifs.

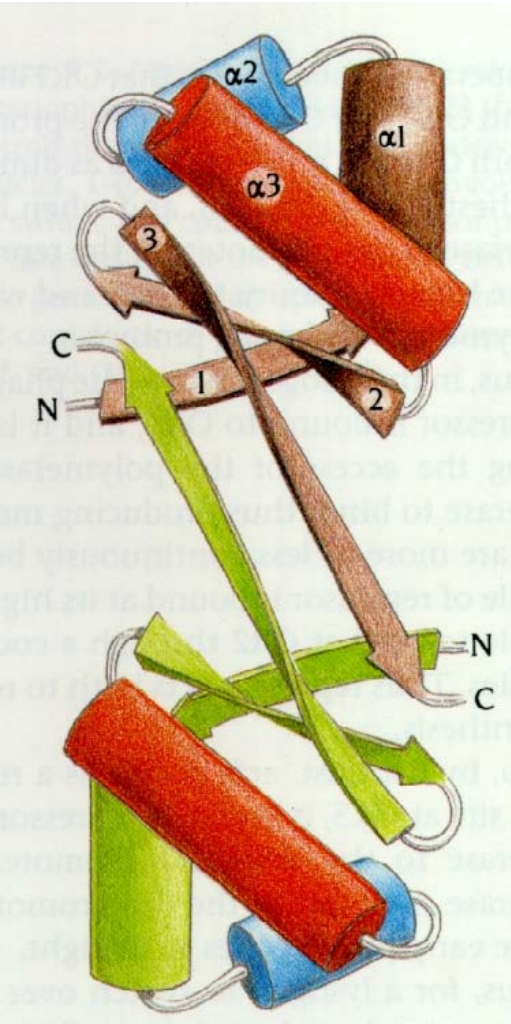
Several motifs may combine to form **domains**.

- Serine proteinase domain, Kringle domain, calcium-binding domain, homeobox domain.

Helix-Turn-Helix Motifs

- Structure
  - 3-helix complex
  - Length: 22 amino acids
  - Turn angle

- Function
  - Gene regulation by binding to DNA



Branden & Tooze

DNA Binding at HTH Motif

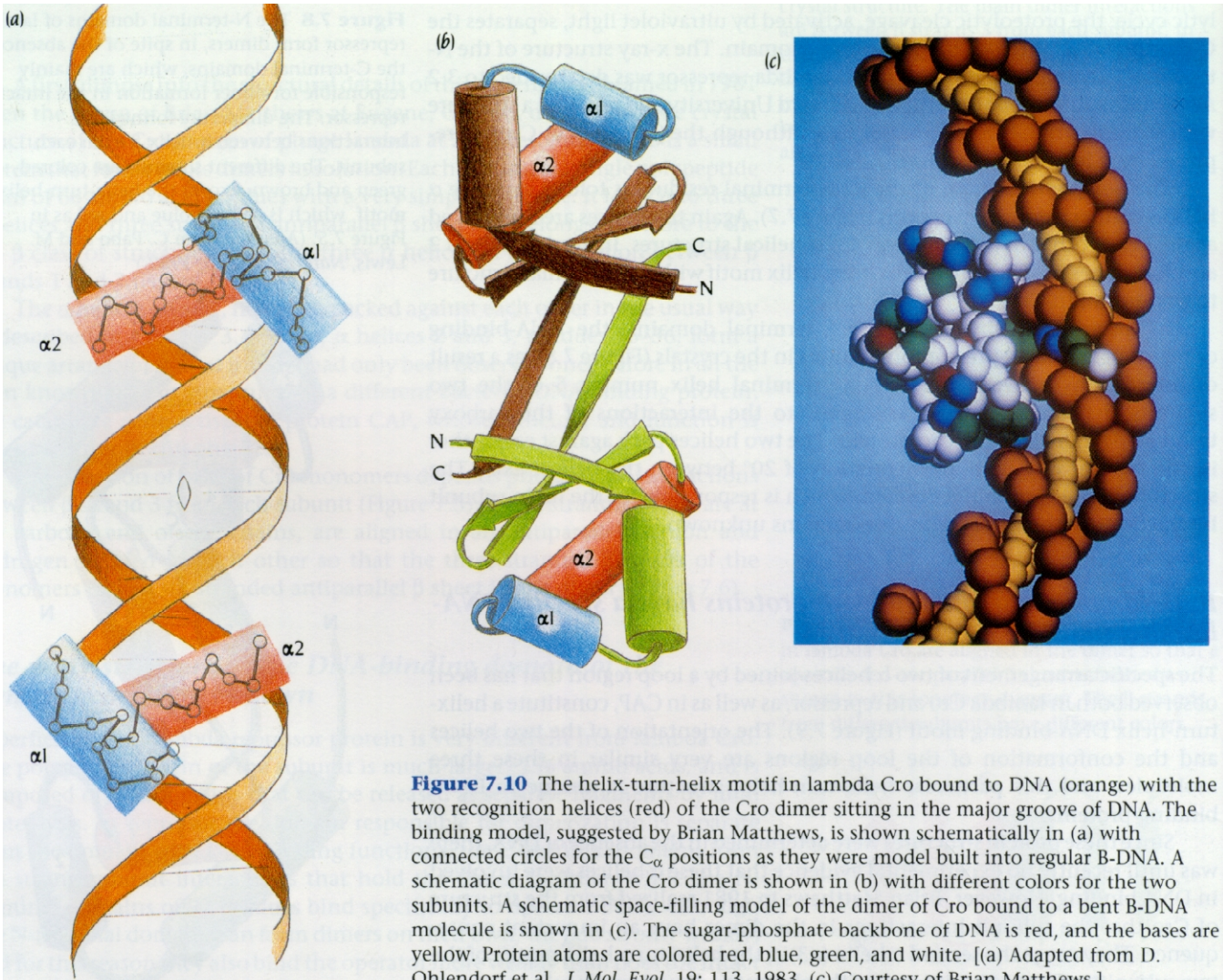


Figure 7.10 The helix-turn-helix motif in lambda Cro bound to DNA (orange) with the two recognition helices (red) of the Cro dimer sitting in the major groove of DNA. The binding model, suggested by Brian Matthews, is shown schematically in (a) with connected circles for the C_α positions as they were model built into regular B-DNA. A schematic diagram of the Cro dimer is shown in (b) with different colors for the two subunits. A schematic space-filling model of the dimer of Cro bound to a bent B-DNA molecule is shown in (c). The sugar-phosphate backbone of DNA is red, and the bases are yellow. Protein atoms are colored red, blue, green, and white. [(a) Adapted from D. Ohlendorf et al., *J. Mol. Evol.* 19: 113, 1983. (c) Courtesy of Brian Matthews.]

Branden & Tooze

HTH Motifs: Examples

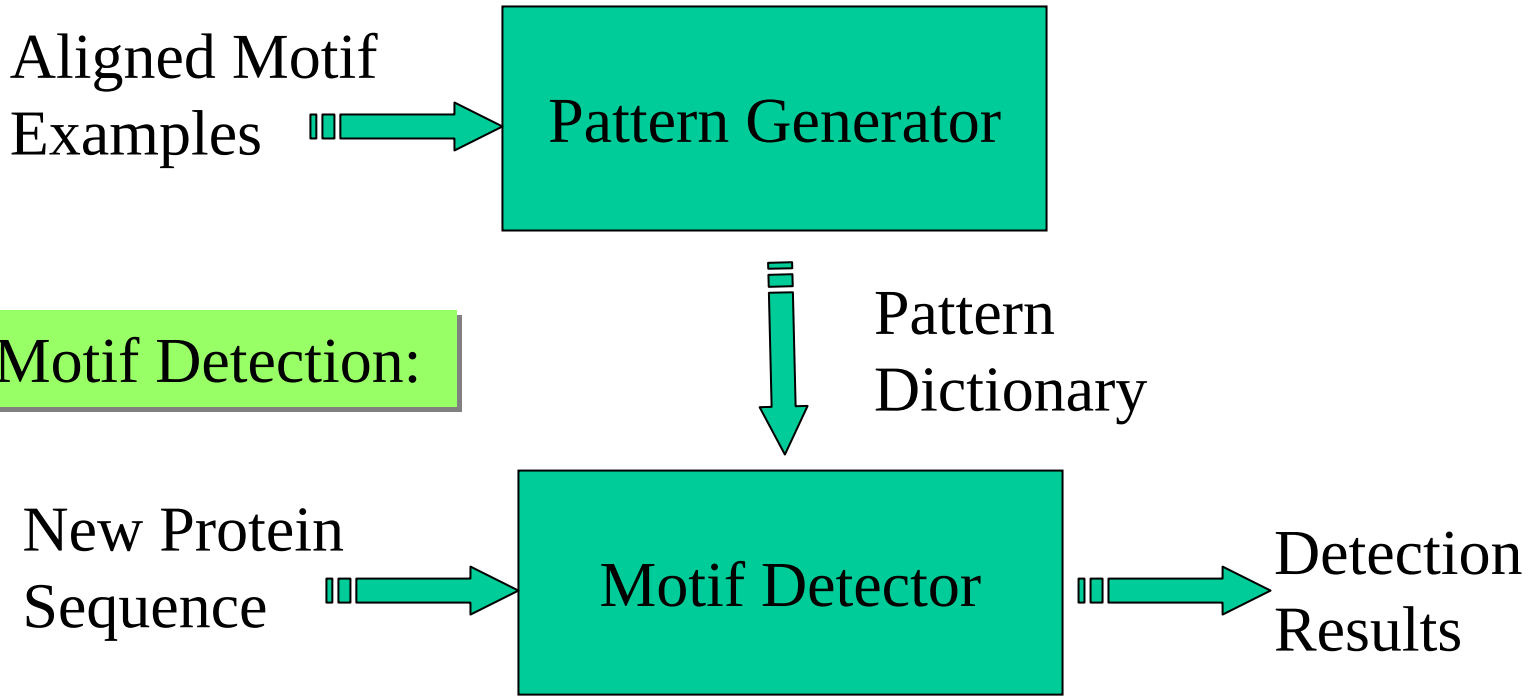
Loc	*Protein*	*Helix 2*									*Turn*				*Helix 3*								
	Name	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
14	**Cro**	F	G	Q	E	K	T	A	K	D	L	G	V	Y	Q	S	A	I	N	K	A	I	H
16	**434 Cro**	M	T	Q	T	E	L	A	T	K	A	G	V	K	Q	Q	S	I	Q	L	I	E	A
11	**P22 Cro**	G	T	Q	R	A	V	A	K	A	L	G	I	S	D	A	A	V	S	Q	W	K	E
31	**Rep**	L	S	Q	E	S	V	A	D	K	M	G	M	G	Q	S	G	V	G	A	L	F	N
16	**434 Rep**	L	N	Q	A	E	L	A	Q	K	V	G	T	T	Q	Q	S	I	E	Q	L	E	N
19	**P22 Rep**	I	R	Q	A	A	L	G	K	M	V	G	V	S	N	V	A	I	S	Q	W	E	R
24	**CII**	L	G	T	E	K	T	A	E	A	V	G	V	D	K	S	Q	I	S	R	W	K	R
4	**LacR**	V	T	L	Y	D	V	A	E	Y	A	G	V	S	Y	Q	T	V	S	R	V	V	N
167	**CAP**	I	T	R	Q	E	I	G	Q	I	V	G	C	S	R	E	T	V	G	R	I	L	K
66	**TrpR**	M	S	Q	R	E	L	K	N	E	L	G	A	G	I	A	T	I	T	R	G	S	N
22	**BlaA Pv**	L	N	F	T	K	A	A	L	E	L	Y	V	T	Q	G	A	V	S	Q	Q	V	R
23	**TrpI Ps**	N	S	V	S	Q	A	A	E	Q	L	H	V	T	H	G	A	V	S	R	Q	L	K

Basis for New Algorithm

- Combinations of residues in specific locations (may not be contiguous) contribute towards stabilizing a structure.
- Some reinforcing combinations are relatively rare.

New Motif Detection Algorithm

Pattern Generation:



Patterns

Loc	*Protein Name*	*Helix 2*									*Turn*				*Helix 3*								
		-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
14	**Cro**	F	G	Q	E	K	T	A	K	D	L	G	V	Y	Q	S	A	I	N	K	A	I	H
16	**434 Cro**	M	T	Q	T	E	L	A	T	K	A	G	V	K	Q	Q	S	I	Q	L	I	E	A
11	**P22 Cro**	G	T	Q	R	A	V	A	K	A	L	G	I	S	D	A	A	V	S	Q	W	K	E
31	**Rep**	L	S	Q	E	S	V	A	D	K	M	G	M	G	Q	S	G	V	G	A	L	F	N
16	**434 Rep**	L	N	Q	A	E	L	A	Q	K	V	G	T	T	Q	Q	S	I	E	Q	L	E	N
19	**P22 Rep**	I	R	Q	A	A	L	G	K	M	V	G	V	S	N	V	A	I	S	Q	W	E	R
24	**CII**	L	G	T	E	K	T	A	E	A	V	G	V	D	K	S	Q	I	S	R	W	K	R
4	**LacR**	V	T	L	Y	D	V	A	E	Y	A	G	V	S	Y	Q	T	V	S	R	V	V	N
167	**CAP**	I	T	R	Q	E	I	G	Q	I	V	G	C	S	R	E	T	V	G	R	I	L	K
66	**TrpR**	M	S	Q	R	E	L	K	N	E	L	G	A	G	I	A	T	I	T	R	G	S	N
22	**BlaA Pv**	L	N	F	T	K	A	A	L	E	L	Y	V	T	Q	G	A	V	S	Q	Q	V	R
23	**TrpI Ps**	N	S	V	S	Q	A	A	E	Q	L	H	V	T	H	G	A	V	S	R	Q	L	K

- Q1 G9 N20
- A5 G9 V10 I15

Pattern Mining Algorithm

Algorithm Pattern-Mining

Input: Motif length m, support threshold T,
list of aligned motifs M.

Output: Dictionary L of frequent patterns.

1. L_1 := All frequent patterns of length 1
2. **for** i = 2 **to** m **do**
3. C_i **:= Candidates**(L_{i-1})
4. L_i := Frequent candidates from C_i
5. **if** ($|L_i| <= 1$) **then**
6. **return** L as the union of all L_j , j <= i.

Candidates Function

L_3

G1, V2, S3
G1, V2, T6
G1, V2, I7
G1, V2, E8
G1, S3, T6
G1, T6, I7
V2, T6, I7
V2, T6, E8

C_4

G1, V2, S3, T6
G1, V2, S3, I7
G1, V2, S3, E8
G1, V2, T6, I7
G1, V2, T6, E8
G1, V2, I7, E8
V2, T6, I7, E8

L_4

G1, V2, S3, T6
G1, V2, S3, I7
G1, V2, S3, E8

G1, V2, T6, E8

V2, T6, I7, E8

Motif Detection Algorithm

Algorithm Motif-Detection

Input : Motif length m, threshold score T, pattern dictionary L, and input protein sequence P[1..n].

Output : Information about motif(s) detected.

```
1.  for each location i do
2.      S := MatchScore(P[i..i+m-1], L).
3.      if (S > T) then
4.          Report it as a possible motif
```

Experimental Results: **GYM 2.0**

Motif	*Protein Family*	*Number Tested*	*GYM = DE Agree*	*Number Annotated*	*GYM = Annot.*
HTH Motif (22)	Master	88	88 (100 %)	13	13
	Sigma	314	284 + 23 (98 %)	96	82
	Negates	93	86 (92 %)	0	0
	LysR	130	127 (98 %)	95	93
	AraC	68	57 (84 %)	41	34
	Rreg	116	99 (85 %)	57	46
	Total	675	653 + 23 (94 %)	289	255 (88 %)

Experiments

- Basic Implementation (Y. Gao)
- Improved implementation & comprehensive testing (K. Mathee, GN).
- Implementation for homeobox domain detection (X. Wang).
- Statistical methods to determine thresholds (C. Bu).
- Use of substitution matrix (C. Bu).
- Study of patterns causing errors (N. Xu).
- Negative training set (N. Xu).
- NN implementation & testing (J. Liu & X. He).
- HMM implementation & testing (J. Liu & X. He).