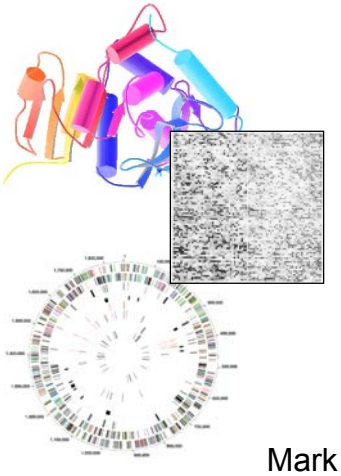
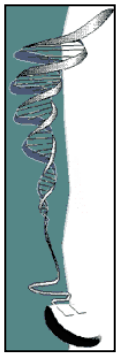
BIOINFORMATICS
Datamining

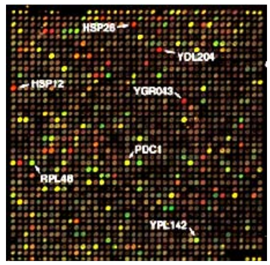

Mark Gerstein, Yale University
bioinfo.mbb.yale.edu/mbb452a

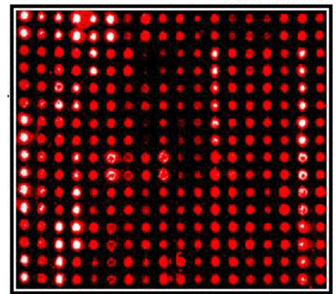
Large-scale Datamining

- Gene Expression
  - ◊ Representing Data in a Grid
  - ◊ Description of function prediction in abstract context
- Unsupervised Learning
  - ◊ clustering & k-means
  - ◊ Local clustering
- Supervised Learning
  - ◊ Discriminants & Decision Tree
  - ◊ Bayesian Nets
- Function Prediction EX
  - ◊ Simple Bayesian Approach for Localization Prediction

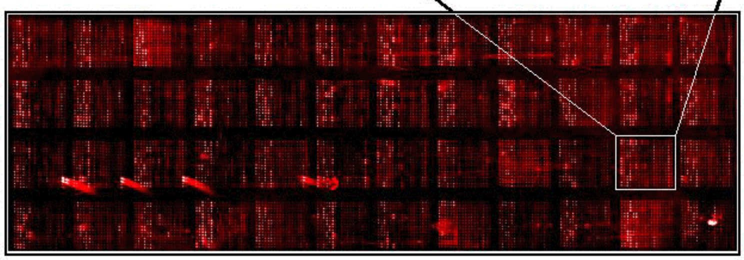
The recent advent and subsequent onslaught of microarray data

1st generation, Expression Arrays (Brown)





2nd gen., Proteome Chips (Snyder)



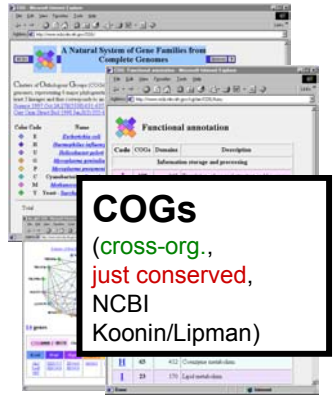
Gene Expression Information and Protein Features

Basics		Predictors												
		Sequence Features										Genomic Features		
		seq. length	Amino Acid Composition					How many times does the sequence have these motif features?				Abs. expr. Level (mRNA copies / cell)		Prot. Abun-dance
Yeast Gene ID	Sequence		A	C	D	W	Y	farn site	NLS	hdel motif	nuc2	signalp	tms1	Gene-Chip expt. from RY Lab	sage tag freq.	(1000 copies /cell)
YAL001C	M	1160	.08	.02	.06	.01	.04	0	1	0	1	0	0	0.3	0	?
YAL002W	K	1176	.09	.02	.06	.01	.04	0	0	0	0	0	1	0.2	?	?
YAL003W	K	206	.08	.02	.06	.01	.04	0	0	0	0	0	0	19.1	19	23
YAL004W	R	215	.08	.02	.06	.01	.04	0	0	0	0	0	0	?	0	?
YAL005C	V	641	.08	.02	.06	.01	.04	0	0	0	0	0	1	13.4	16	17
YAL007C	K	190	.08	.02	.06	.01	.04	0	0	0	0	1	4	2.2	8	?
YAL008W	H	198	.08	.02	.06	.01	.04	0	0	0	0	0	3	1.2	?	?
YAL009W	P	259	.08	.02	.06	.01	.04	0	2	0	0	0	3	0.6	?	?
YAL010C	M	493	.08	.02	.06	.02	.04	0	0	0	0	0	1	0.3	?	?
YAL011W	K	616	.08	.02	.06	.01	.04	0	8	0	1	0	0	0.4	?	?
YAL012W	G	393	.08	.02	.06	.01	.04	0	0	0	0	0	1	8.9	4	6.7
YAL013W	R	362	.08	.02	.06	.01	.04	0	0	0	0	0	0	0.6	?	?
YAL014C	G	202	.08	.02	.06	.01	.04	0	0	0	0	0	0	1.1	?	?
YAL015C	M	399	.08	.02	.06	.01	.04	0	1	0	0	0	0	0.7	0	1
YAL016W	K	635	.08	.02	.06	.01	.04	0	0	0	0	0	1	3.3	5	?
YAL017W	V	1356	.08	.02	.06	.01	.04	0	0	0	0	0	0	0.4	?	?
YAL018C	K	325	.08	.02	.06	.01	.04	0	0	0	0	0	4	?	?	?

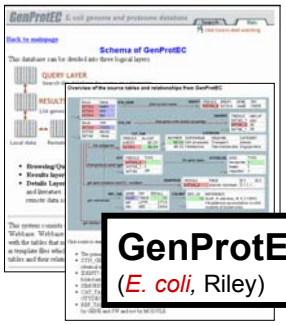
Yeast Gene ID	Cell cycle timecourse: t=0	t=1	t=2	t=3	t=4	t=5	t=6	t=7	t=8	t=9	t=10	t=11	t=12	t=13	t=14	t=15	t=16
YAL001C	5	3	4	4	5	4	3	5	5	3	5	7	9	4	4	4	5
YAL002W	8	4	2	3	4	3	4	5	5	3	4	4	6	4	5	4	3
YAL003W	70	73	91	69	105	52	112	88	64	159	106	104	75	103	140	98	126
YAL004W	18	12	9	5	5	3	6	4	4	3	3	5	5	4	5	4	6
YAL005C	39	38	30	13	17	8	11	8	7	8	6	8	8	7	9	8	14
YAL007C	15	20	32	20	21	19	29	19	16	22	20	26	23	22	25	16	17
YAL008W	9	6	7	1	3	2	4	2	2	3	3	4	4	3	3	2	3
YAL009W	6	2	4	3	5	3	5	5	5	3	4	6	6	4	4	3	5
YAL010C	11	6	4	5	6	4	7	8	7	4	5	6	7	5	6	6	6
YAL011W	6	5	4	4	8	5	8	8	6	6	5	6	6	7	6	5	6
YAL012W	29	26	25	27	53	26	43	36	25	28	23	28	31	29	34	23	29
YAL013W	7	9	6	5	14	6	12	14	10	9	9	9	10	9	8	6	10
YAL014C	12	13	10	8	10	10	12	13	12	14	11	11	11	10	11	9	12
YAL015C	19	18	14	10	14	12	17	17	14	13	11	13	16	11	14	12	13
YAL016W	15	20	20	102	20	20	30	22	18	19	18	20	21	21	23	16	16
YAL017W	14	3	3	4	8	5	6	6	5	5	8	9	10	6	5	4	7
YAL018C	4	2	2	2	1	1	2	2	2	1	2	1	2	2	1	2	1

...

Functional Classification

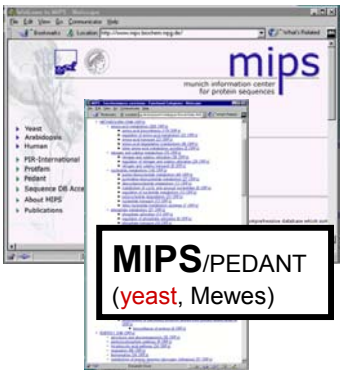


COGs
(cross-org.,
just conserved,
NCBI
Koonin/Lipman)

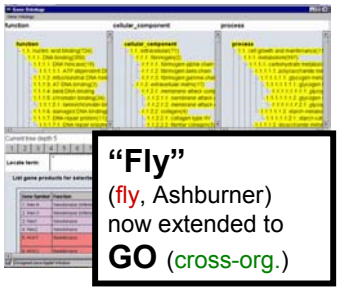


GenProtEC
(*E. coli,* Riley)

ENZYME
(SwissProt
Bairoch/
Apweiler,
just enzymes,
cross-org.)



MIPS/PEDANT
(yeast, Mewes)



"Fly"
(fly, Ashburner)
now extended to
GO (cross-org.)

Also:

Other SwissProt Annotation

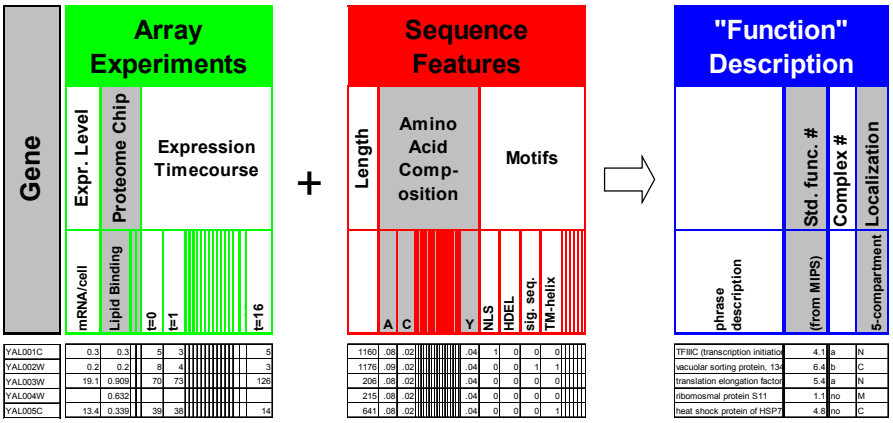
WIT, KEGG (just pathways)

TIGR EGAD (human ESTs)

SGD (yeast)

Core

Prediction of Function on a Genomic Scale from Array Data & Sequence Features



Different Aspects of function: molecular action, cellular role, phenotypic manifestation
Also: localization, interactions, complexes

6000+

Arrange data in a tabulated form, each row representing an example and each column representing a feature, including the dependent experimental quantity to be predicted.

	predictor1	Predictor2	predictor3	predictor4	response
G1	A(1,1)	A(1,2)	A(1,3)	A(1,4)	Class A
G2	A(2,1)	A(2,2)	A(2,3)	A(2,4)	Class A
G3	A(3,1)	A(3,2)	A(3,3)	A(3,4)	Class B

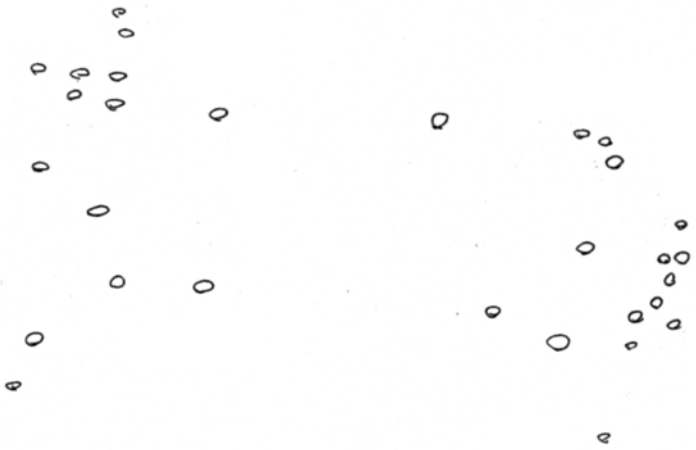
(adapted from Y Kluger)

Typical Predictors and Response for Yeast

Basics		Predictors																						Response		
		Sequence Features															Genomic Features									
		seq. length	Amino Acid Composition										How many times does the sequence have these motif features?							Abs. expr. Level (mRNA copies / cell)		Prot. Abun-dance	Cell cycle timecourse				Function		Localization
Yeast Gene ID	Sequence		A	C	D	E	F	W	Y	farn site	NLS	hdel motif	nuc2	signalp	tms1		Gene-Chip expt. from RY Lab	sage tag freq.	(1000 copies /cell)	t=0	t=1	t=15	t=16	function ID(s) (from MIPS)	function description	5-compartment
YAL001C	MNIFEMLRI	1160	.08	.02	.06			.01	.04	0	1	0	1	0	0		0.3	0	?	5	3	4	5	04.01.01;04.03	TFIIIC (transcription initi	N
YAL002W	KVFGRCELA	1176	.09	.02	.06			.01	.04	0	0	0	0	0	1		0.2	?	?	8	4	4	3	06.04;08.13	vacuolar sorting protein,	C
YAL003W	KMLQFNLRW	206	.08	.02	.06			.01	.04	0	0	0	0	0	0		19.1	19	23	70	73	98	126	05.04;30.03	translation elongation fac	N
YAL004W	RPDFCLEPP	215	.08	.02	.06			.01	.04	0	0	0	0	0	0	?		0	?	18	12	4	6	01.01.01	0	N
YAL005C	VINTFDGVA	641	.08	.02	.06			.01	.04	0	0	0	0	0	1		13.4	16	17	39	38	8	14	06.01;06.04;08	heat shock protein of HS	????
YAL007C	KKAVINGEQ	190	.08	.02	.06			.01	.04	0	0	0	0	1	4		2.2	8	?	15	20	16	17	99	????	????
YAL008W	HPETLVKVK	198	.08	.02	.06			.01	.04	0	0	0	0	0	3		1.2	?	?	9	6	2	3	99	????	????
YAL009W	PTLEWFLSH	259	.08	.02	.06			.01	.04	0	2	0	0	0	3		0.6	?	?	6	2	3	5	03.10;03.13	meiotic protein	????
YAL010C	MEQRITLKD	493	.08	.02	.06			.02	.04	0	0	0	0	0	1		0.3	?	?	11	6	6	6	30.16	involved in mitochondrial	????
YAL011W	KSFPEVVGK	616	.08	.02	.06			.01	.04	0	8	0	1	0	0		0.4	?	?	6	5	5	6	30.16;99	protein of unknown funct	????
YAL012W	GVQVETISP	393	.08	.02	.06			.01	.04	0	0	0	0	0	1		8.9	4	6.7	29	26	23	29	01.01.01;30.03	cystathionine gamma-lya	C
YAL013W	RTDCYGNVN	362	.08	.02	.06			.01	.04	0	0	0	0	0	0		0.6	?	?	7	9	6	10	01.06.10;30.03	regulator of phospholipid	N
YAL014C	GDVEKGKKI	202	.08	.02	.06			.01	.04	0	0	0	0	0	0		1.1	?	?	12	13	9	12	99	????	N
YAL015C	MTPAVTTYK	399	.08	.02	.06			.01	.04	0	1	0	0	0	0		0.7	0	1	19	16	12	13	11.01;11.04	DNA repair protein	N
YAL016W	KKPLTQEQL	635	.08	.02	.06			.01	.04	0	0	0	0	0	1		3.3	5	?	15	20	16	16	03.01;03.04;03	ser/thr protein phosphata	????

Represent predictors in abstract high dimensional space

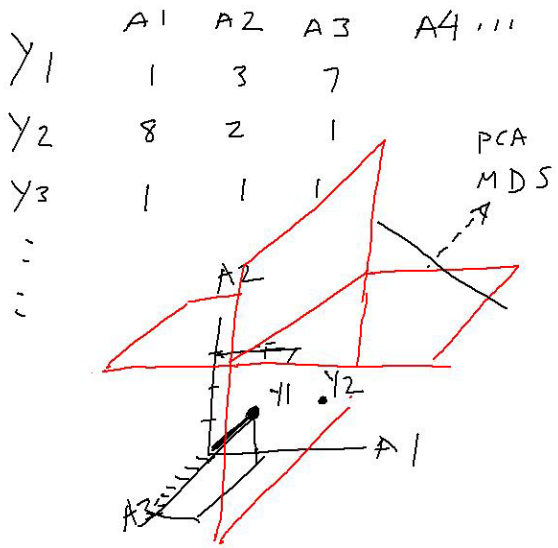
Core



"Tag" Certain Points

Core

Abstract high-dimensional space representation



Large-scale Datamining

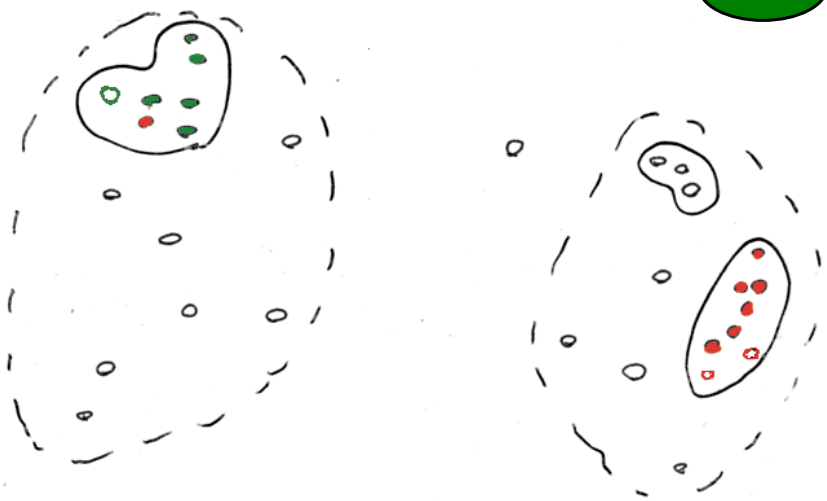
- Gene Expression
  - ◊ Representing Data in a Grid
  - ◊ Description of function prediction in abstract context
- Unsupervised Learning
  - ◊ clustering & k-means
  - ◊ Local clustering
- Supervised Learning
  - ◊ Discriminants & Decision Tree
  - ◊ Bayesian Nets
- Function Prediction EX
  - ◊ Simple Bayesian Approach for Localization Prediction

“cluster” predictors

Core

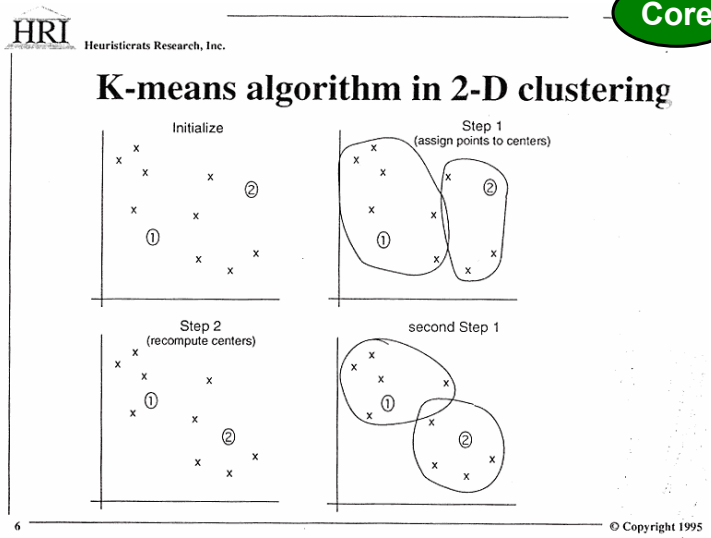
Use clusters to predict Response

Core



K-means

Core



K-means

Top-down vs. Bottom up

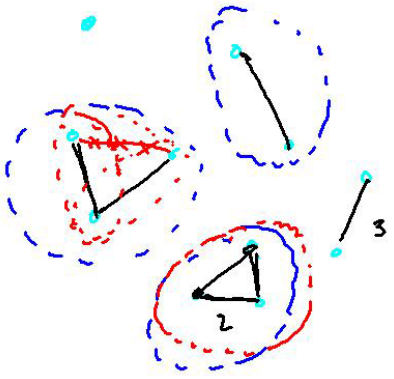
Top-down when you know how many subdivisions

k-means as an example of top-down

1) Pick ten (i.e. k?) random points as putative cluster centers.
2) Group the points to be clustered by the center to which they are closest.
3) Then take the mean of each group and repeat, with the means now at the cluster center.
4) I suppose you stop when the centers stop moving.

Bottom up clustering

Core



SINGLE LINK

MULTI LINK

(THRESHOLD)

BOTTOM UP

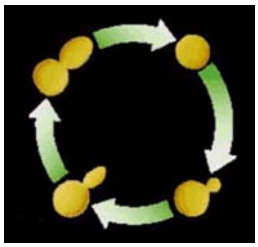
TOP-DOWN

K MEANS

Large-scale Datamining

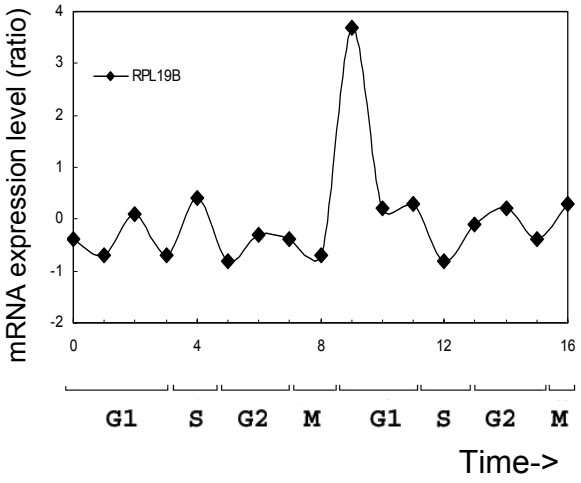
- Gene Expression
  - ◊ Representing Data in a Grid
  - ◊ Description of function prediction in abstract context
- Unsupervised Learning
  - ◊ clustering & k-means
  - ◊ Local clustering
- Supervised Learning
  - ◊ Discriminants & Decision Tree
  - ◊ Bayesian Nets
- Function Prediction EX
  - ◊ Simple Bayesian Approach for Localization Prediction

Clustering the yeast cell cycle to uncover interacting proteins



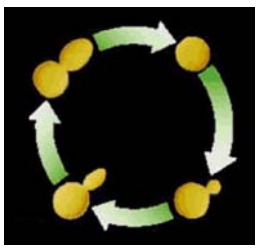
Extra

[Brown, Davis]

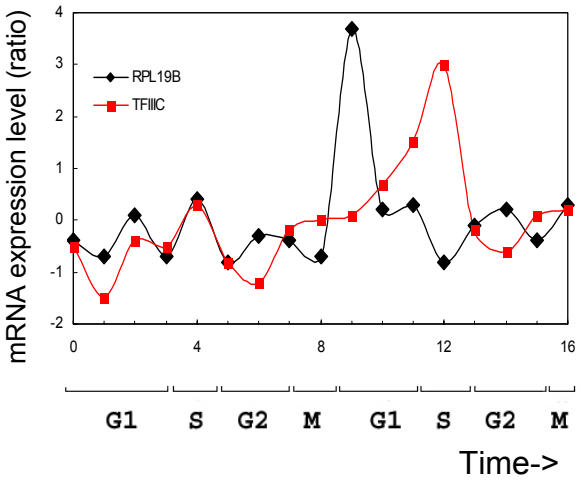


Microarray timecourse of
1 ribosomal protein

Clustering the yeast cell cycle to uncover interacting proteins

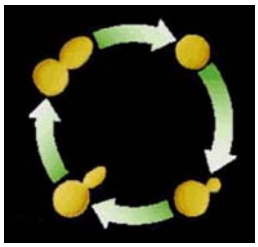


Extra



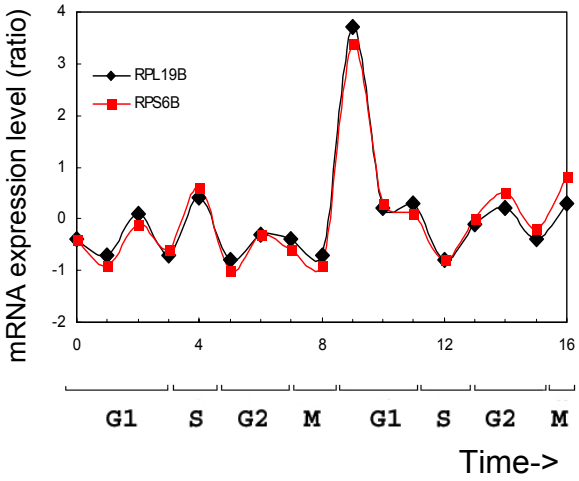
Random relationship from ~18M

Clustering the yeast cell cycle to uncover interacting proteins



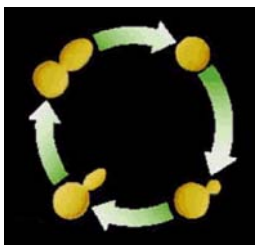
Extra

[Botstein; Church, Vidal]

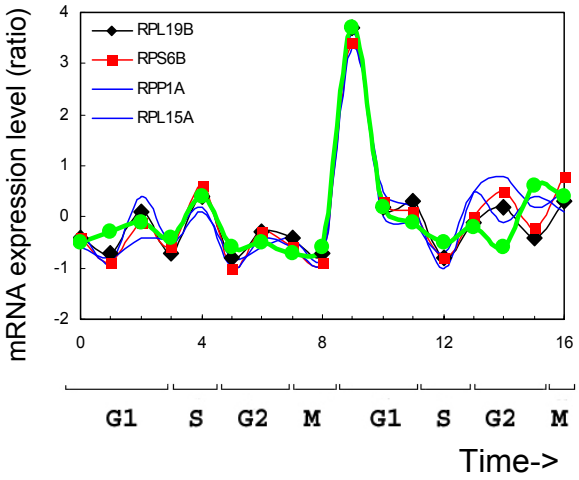


Close relationship from 18M
(2 Interacting Ribosomal Proteins)

Clustering the yeast cell cycle to uncover interacting proteins



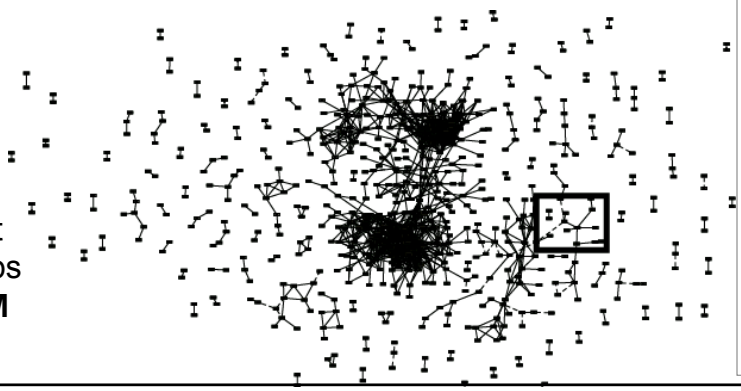
Extra



Predict Functional Interaction of
Unknown Member of Cluster

Global Network of Relationships

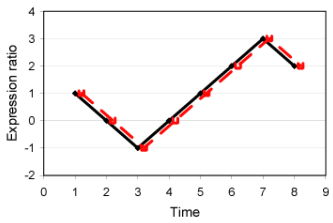
Core



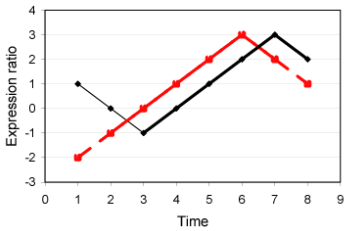
~470K significant relationships from **~18M** possible

Simultaneous

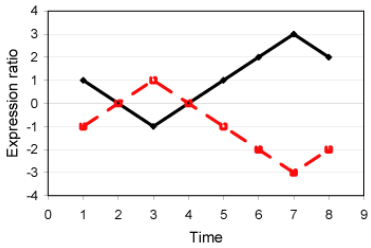
Traditional Global Correlation



Time-Shifted



Inverted



Local Clustering algorithm identifies further (reasonable) types of expression relation-ships

Core

[Church]

Local Alignment

Suppose there are n (1, 2, …, n) time points:

- The expression ratio is normalized in "Z-score" fashion;
- Score matrix: $S_{i,j} = S(x_i, y_j) = x_i \bullet y_j$;

Qian J. et al. Beyond Synexpression Relationships: Local clustering of Time-shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions. J. Mol. Biol. (2001) 314, 1053-1066

Local Alignment

Suppose there are n (1, 2, …, n) time points:

- Sum matrices $E_{i,j}$ and $D_{i,j}$:

$$E_{i,j} = \max(E_{i-1,j-1} + S_{i,j}, 0);$$

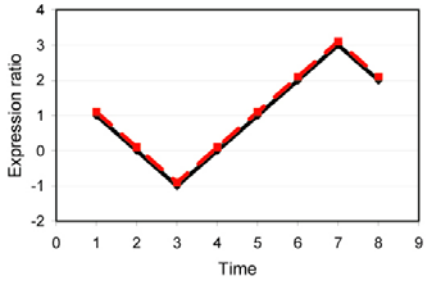
$$D_{i,j} = \max(D_{i-1,j-1} - S_{i,j}, 0);$$

- Match Score = $\max(E_{i,j}, D_{i,j})$

Qian J. et al. Beyond Synexpression Relationships: Local clustering of Time-shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions. J. Mol. Biol. (2001) 314, 1053-1066

Local Alignment

Simultaneous



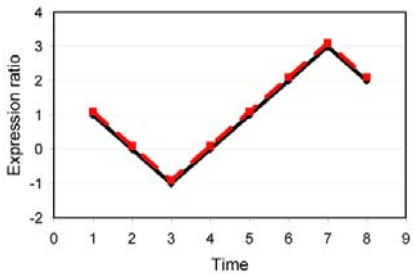
		1	0	-1	0	1	2	3	2
	0	0	0	0	0	0	0	0	0
1	0	1	0	-1	0	1	2	3	2
0	0	0	0	0	0	0	0	0	0
-1	0	-1	0	1	0	-1	-2	-3	-2
0	0	0	0	0	0	0	0	0	0
1	0	1	0	-1	0	1	2	3	2
2	0	2	0	-1	0	2	4	6	4
3	0	3	0	-3	0	3	6	9	6
2	0	1	0	-2	0	2	4	6	4

$S_{i,j} = x_i \bullet y_j$

Qian J. et al. Beyond Synexpression Relationships: Local clustering of Time-shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions. J. Mol. Biol. (2001) 314, 1053-1066

Local Alignment

Simultaneous



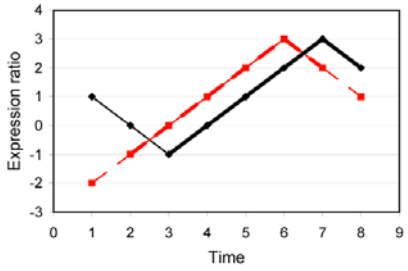
		1	0	-1	0	1	2	3	2
	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	2	3	2
0	0	0	1	0	0	0	1	2	3
-1	0	0	0	2	0	0	0	0	0
0	0	0	0	0	2	0	0	0	0
1	0	1	0	0	0	3	2	3	2
2	0	2	1	0	0	2	7	8	7
3	0	3	2	0	0	3	8	16	14
2	0	2	3	0	0	2	7	14	20

$E_{i,j} = \max(E_{i-1,j-1} + x_i \bullet y_j, 0)$

Qian J. et al. Beyond Synexpression Relationships: Local clustering of Time-shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions. J. Mol. Biol. (2001) 314, 1053-1066

Local Alignment

Time-Shifted



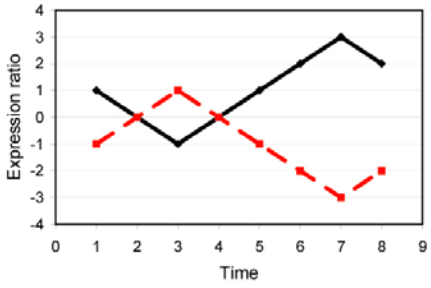
		-2	-1	0	1	2	3	2	1
	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	2	3	2	1
0	0	0	0	0	0	1	2	3	2
-1	0	2	1	0	0	0	0	0	2
0	0	0	2	1	0	0	0	0	0
1	0	0	0	2	2	2	3	2	1
2	0	0	0	0	4	6	8	7	4
3	0	0	0	0	3	10	15	14	10
2	0	0	0	0	2	7	16	19	16

$$E_{i,j} = \max(E_{i-1,j-1} + x_i \cdot y_j, 0)$$

Qian J. et al. Beyond Synexpression Relationships: Local clustering of Time-shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions. J. Mol. Biol. (2001) 314, 1053-1066

Local Alignment

Inverted

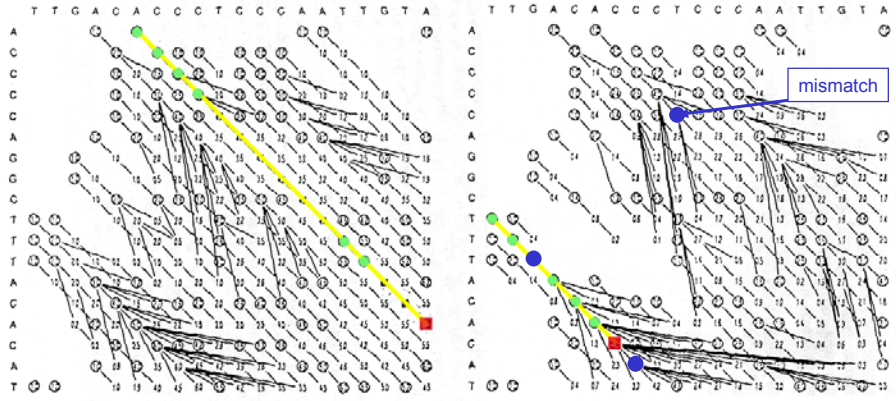


		-1	0	1	0	-1	-2	-3	-2
	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	2	3	2
0	0	0	1	0	0	0	1	2	3
-1	0	0	0	2	0	0	0	0	2
0	0	0	0	0	2	0	0	0	0
1	0	1	0	0	0	3	2	3	2
2	0	2	1	0	0	2	7	8	7
3	0	3	2	0	0	3	8	16	14
2	0	2	3	0	0	2	7	14	20

$$D_{i,j} = \max(D_{i-1,j-1} - x_i \cdot y_j, 0)$$

Qian J. et al. Beyond Synexpression Relationships: Local clustering of Time-shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions. J. Mol. Biol. (2001) 314, 1053-1066

Global (NW) vs Local (SW)Alignments

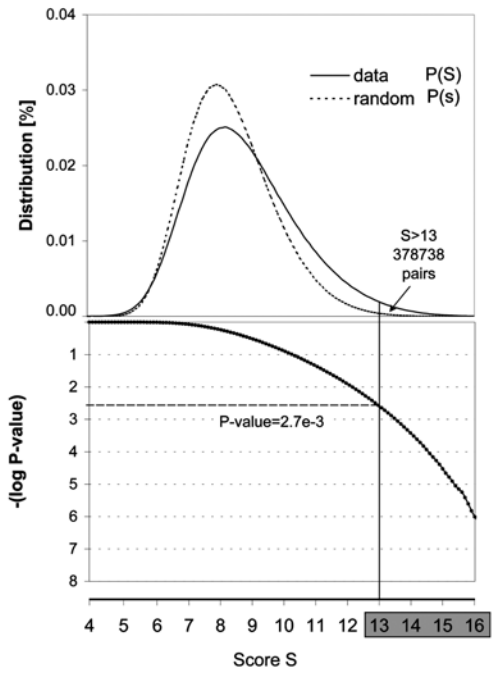



```
TTGACACCCTCCCAATTGTA...      T T G A C A C C...
||||      ||   |             | | - | | | | -
.....ACCCCAGGCTTTACACAT      T T T A C A C A...
1234444444566 67             1 2 1 2 3 4 5 4
                             0 0 4 4 4 4 4 8
```

Match Score = +1
Gap-Opening=-1.2, Gap-Extension=-.03
for local alignment Mismatch = -0.6

Adapted from D J States & M S Boguski, "Similarity and Homology," Chapter 3 from Gribskov, M. and Devereux, J. (1992). Sequence Analysis Primer. New York, Oxford University Press. (Page 133)

Statistical Scoring



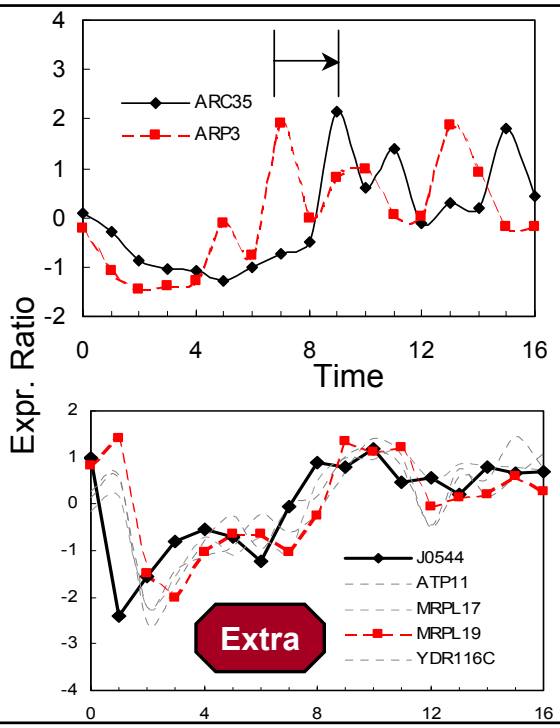
Examples time-shifted relationships

Suggestive

ARP3 : in actin remodelling cplx.
ARC35 : in same cplx. (required late in cell cycle)

Predicted

J0544 : **unknown** function
MRPL19: mito.ribosome



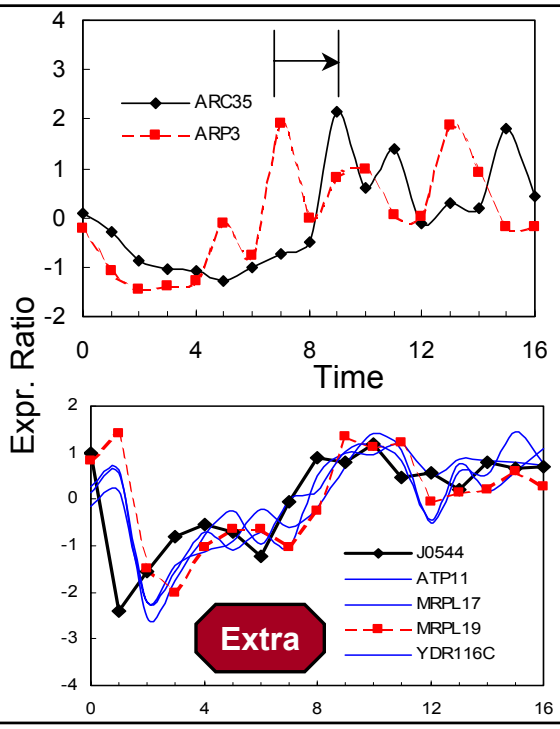
Examples time-shifted relationships

Suggestive

ARP3 : in actin remodelling cplx.
ARC35 : in same cplx. (required late in cell cycle)

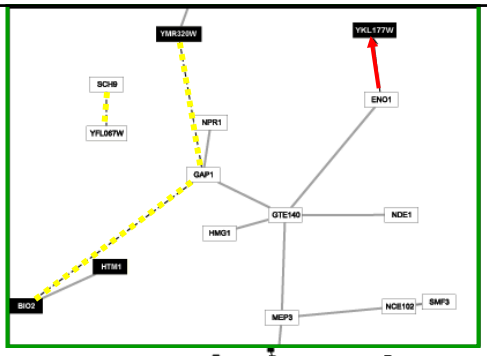
Predicted

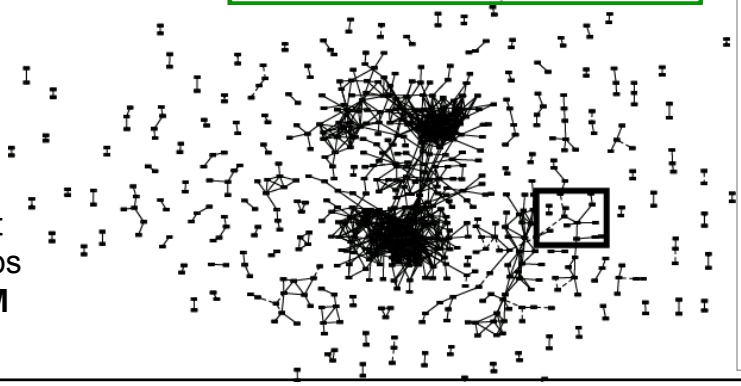
J0544 : **unknown** function
MRPL19: mito.ribosome



Global Network of 3 Different Types of Relationships

Simultaneous
Inverted
Shifted





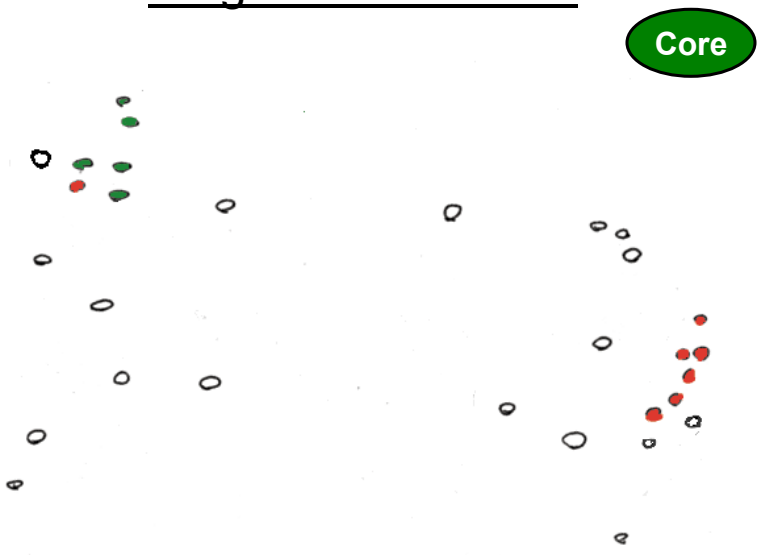
Extra

~470K significant relationships from **~18M** possible

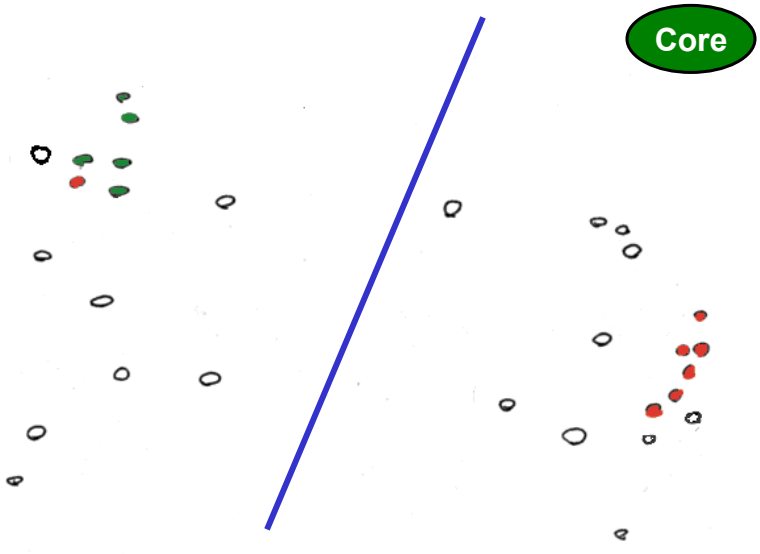
Large-scale Datamining

- Gene Expression
  - ◊ Representing Data in a Grid
  - ◊ Description of function prediction in abstract context
- Unsupervised Learning
  - ◊ clustering & k-means
  - ◊ Local clustering
- Supervised Learning
  - ◊ Discriminants & Decision Tree
  - ◊ Bayesian Nets
- Function Prediction EX
  - ◊ Simple Bayesian Approach for Localization Prediction

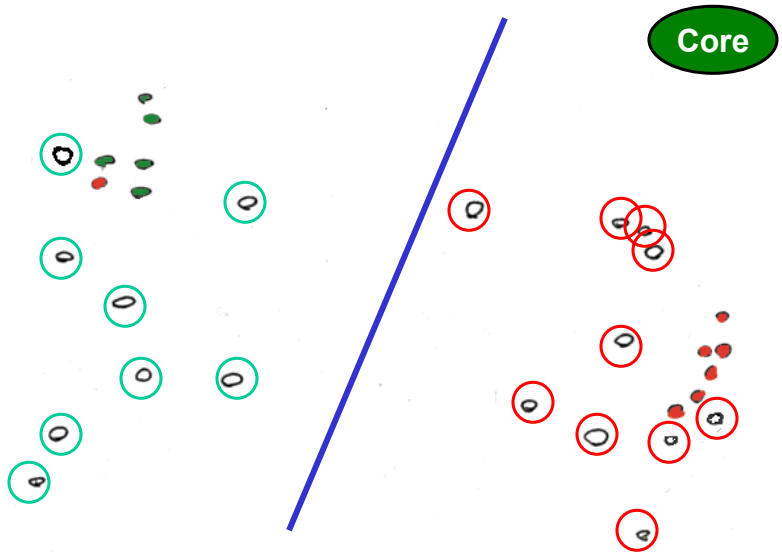
“Tag” Certain Points



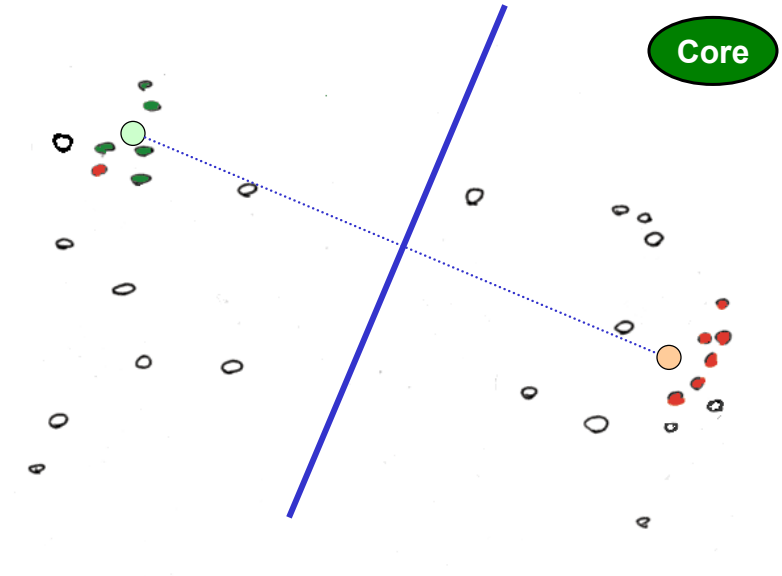
Find a Division to Separate Tagged Points



Extrapolate to Untagged Points



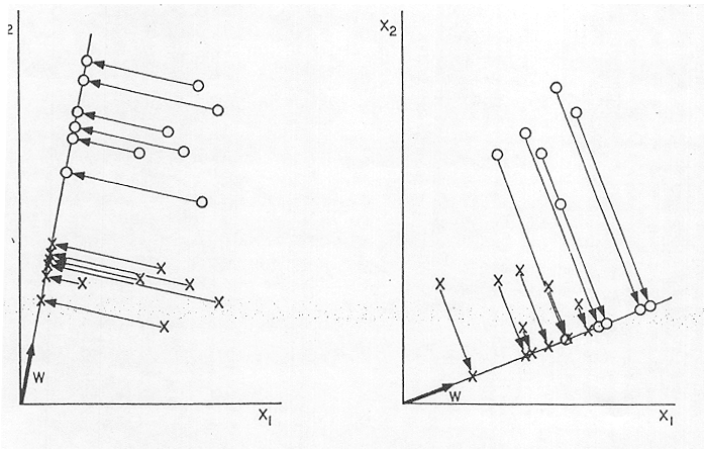
Discriminant to Position Plane



Fisher discriminant analysis

- Use the training set to reveal the structure of class distribution by seeking a linear combination
- $y = w_1x_1 + w_2x_2 + ... + w_nx_n$ which maximizes the ratio of the separation of the class means to the sum of each class variance (within class variance). This linear combination is called the first linear discriminant or first canonical variate. Classification of a future case is then determined by choosing the nearest class in the space of the first linear discriminant and significant subsequent discriminants, which maximally separate the class means and are constrained to be uncorrelated with previous ones.

Fischer's Discriminant



(Adapted from ???)

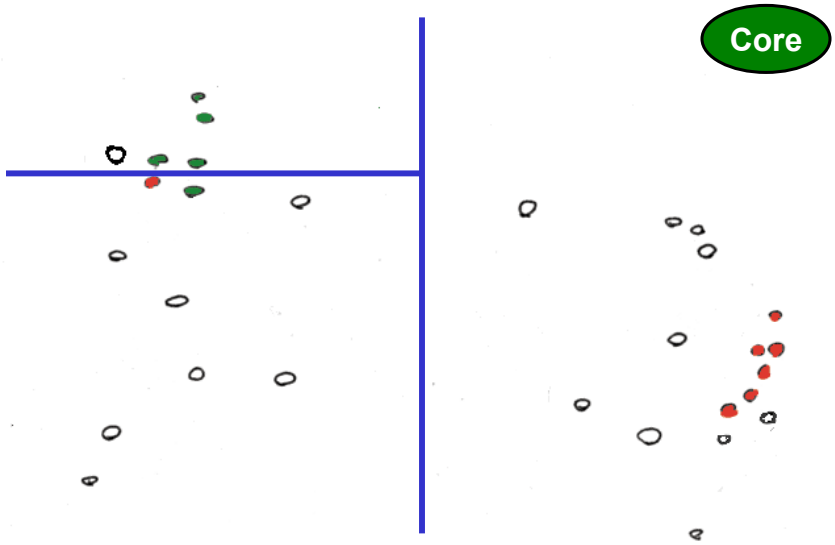
Fisher cont.

$$m_i = \vec{w} \cdot \vec{m}_i \qquad s_i^2 = \sum_{y \in Y_i} (y - m_i)^2$$

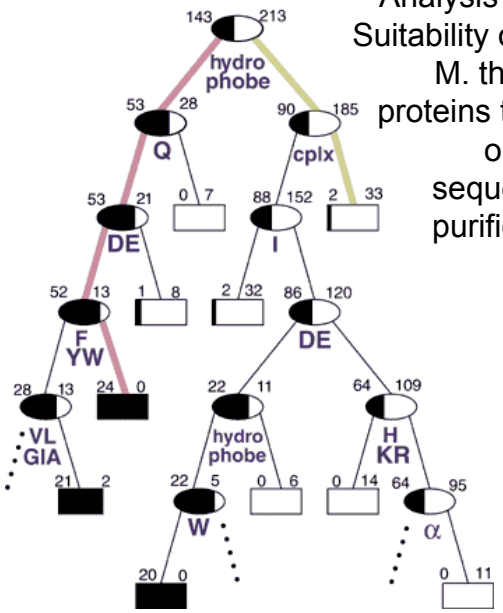
Solution of 1st variate

$$\vec{w} = S_W^{-1}(\vec{m}_1 - \vec{m}_2)$$

Find a Division to Separate Tagged Points



Retrospective Decision Trees



Analysis of the Suitability of 500 M. thermo. proteins to find optimal sequences purification

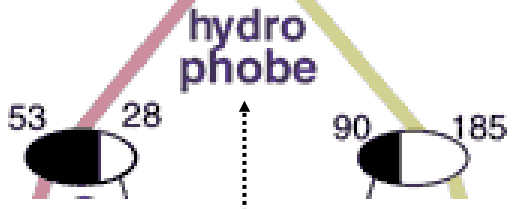
Core

Retrospective Decision Trees Nomenclature

Core

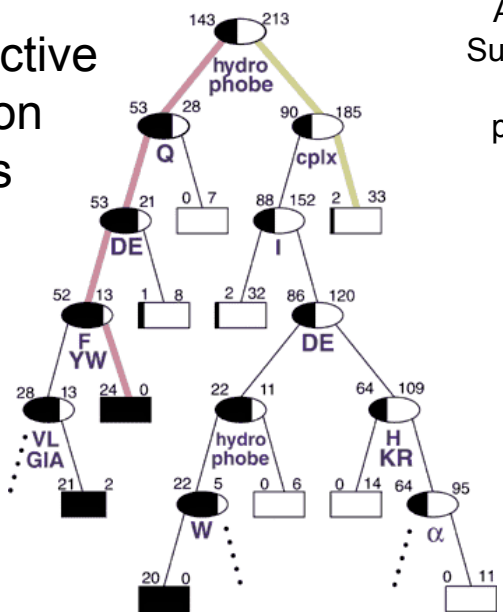
356 total

Not Expressible ⋯→ 143 213 ←⋯ **Express -ible**

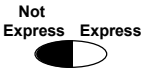


Has a hydrophobic stretch? (Y/N)

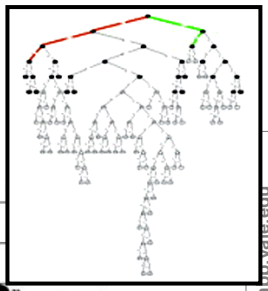
Retrospective Decision Trees

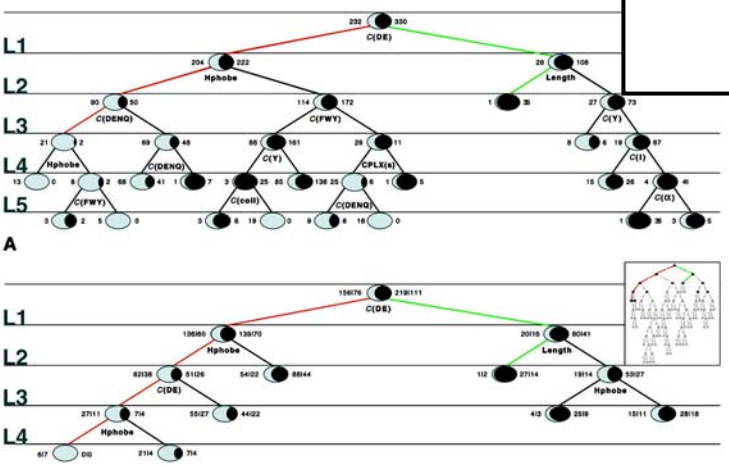


Analysis of the Suitability of 500 M. thermo. proteins to find optimal sequences purification



Overfitting, Cross Validation, and Pruning





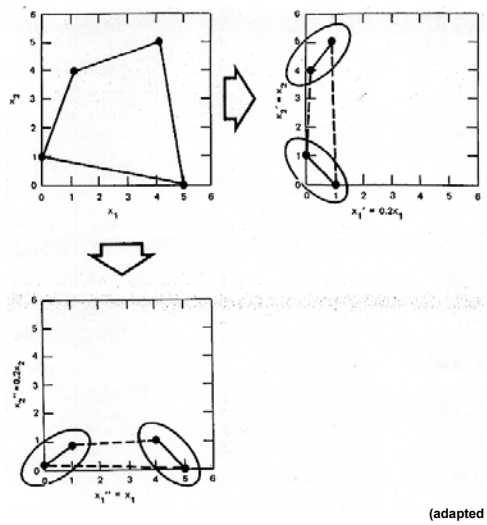
Decision Trees

- can handle data that is not linearly separable.
- A decision tree is an upside down tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a classification or *decision*. One classifies instances by sorting them down the tree from the root to some leaf nodes. To classify an instance the tree calls first for a test at the root node, testing the feature indicated on this node and choosing the next node connected to the root branch where the outcome agrees with the value of the feature of that instance. Thereafter a second test on another feature is made on the next node. This process is then repeated until a leaf of the tree is reached.
- Growing the tree, based on a training set, requires strategies for (a) splitting the nodes and (b) pruning the tree. Maximizing the decrease in average impurity is a common criterion for splitting. In a problem with noisy data (where distribution of observations from the classes overlap) growing the tree will usually over-fit the training set. The strategy in most of the cost-complexity pruning algorithms is to choose the smallest tree whose error rate performance is close to the minimal error rate of the over-fit larger tree. More specifically, growing the trees is based on splitting the node that maximizes the reduction in deviance (or any other impurity-measure of the distribution at a node) over all allowed binary splits of all terminal nodes. Splits are *not* chosen based on misclassification rate .A binary split for a continuous feature variable *v* is of the form *v<threshold* versus *v>threshold* and for a "descriptive" factor it divides the factor's levels into two classes. Decision tree-models have been successfully applied in a broad range of domains. Their popularity arises from the following: Decision trees are easy to interpret and use when the predictors are a mix of numeric and nonnumeric (factor) variables. They are invariant to scaling or re-expression of numeric variables. Compared with linear and additive models they are effective in treating missing values and capturing non-additive behavior. They can also be used to predict nonnumeric dependent variables with more than two levels. In addition, decision-tree models are useful to devise prediction rules, screen the variables and summarize the multivariate data set in a comprehensive fashion. We also note that ANN and decision tree learning often have comparable prediction accuracy [Mitchell p. 85] and SVM algorithms are slower compared with decision tree. These facts suggest that the decision tree method should be one of our top candidates to "data-mine" proteomics datasets. C4.5 and CART are among the most popular decision tree algorithms.

Optional: not needed for Quiz

(adapted from Y Kluger)

Effect of Scaling



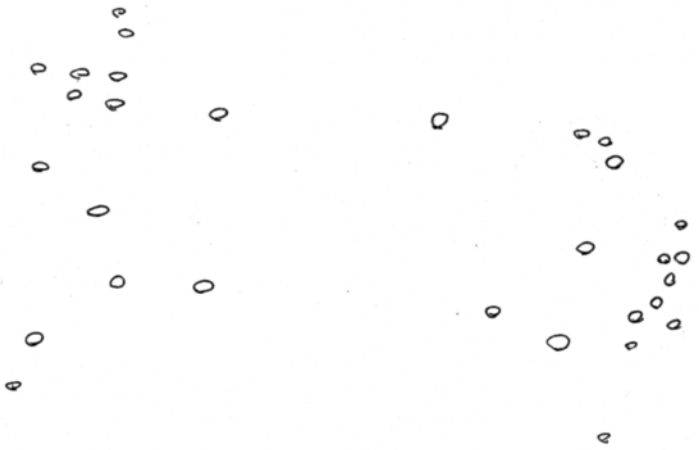
(adapted from ref?)

End of class 2002,12.01
(Bioinfo-13)
[started at beg. of datamining]

Large-scale Datamining

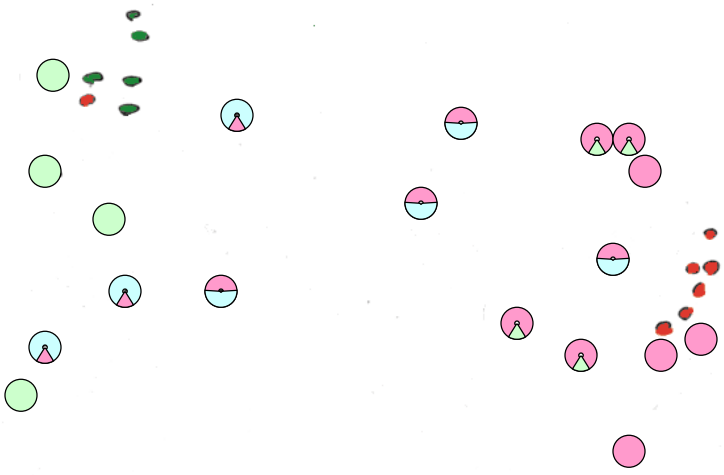
- Gene Expression
  - ◊ Representing Data in a Grid
  - ◊ Description of function prediction in abstract context
- Unsupervised Learning
  - ◊ clustering & k-means
  - ◊ Local clustering
- Supervised Learning
  - ◊ Discriminants & Decision Tree
  - ◊ Bayesian Nets
- Function Prediction EX
  - ◊ Simple Bayesian Approach for Localization Prediction

Represent predictors in abstract high dimensional space



Tagged Data

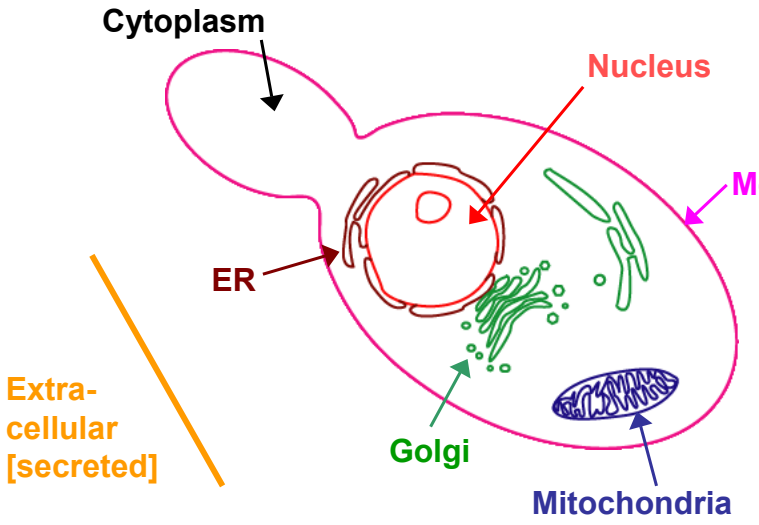
Probabilistic Predictions of Class



Large-scale Datamining

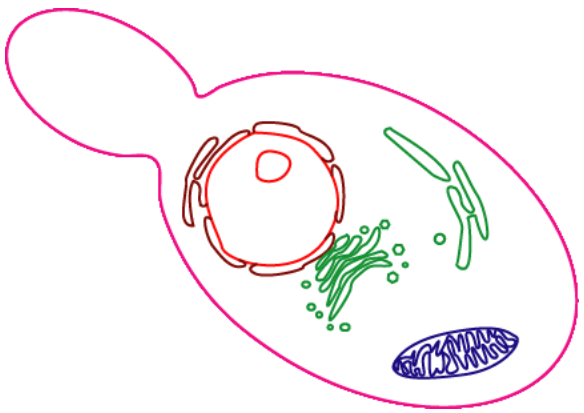
- Gene Expression
  - ◊ Representing Data in a Grid
  - ◊ Description of function prediction in abstract context
- Unsupervised Learning
  - ◊ clustering & k-means
  - ◊ Local clustering
- Supervised Learning
  - ◊ Discriminants & Decision Tree
  - ◊ Bayesian Nets
- Function Prediction EX
  - ◊ Simple Bayesian Approach for Localization Prediction

Subcellular Localization, a standardized aspect of function

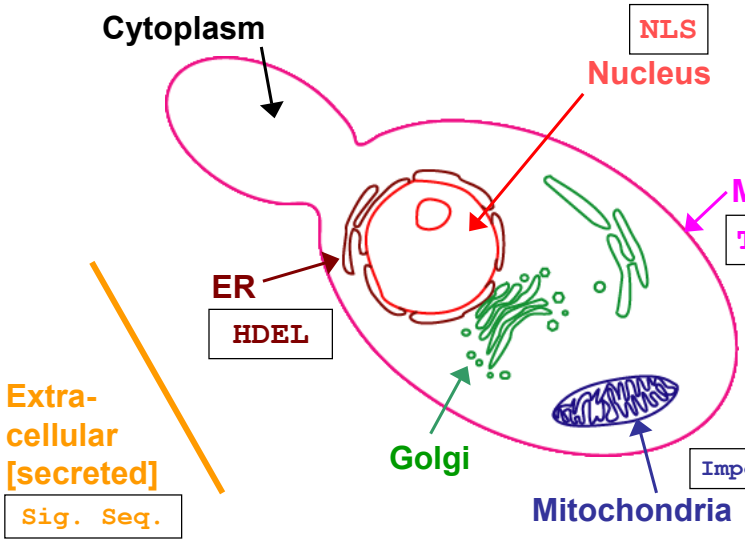


Subcellular Localization, Provides a simple goal for genome-scale functional prediction

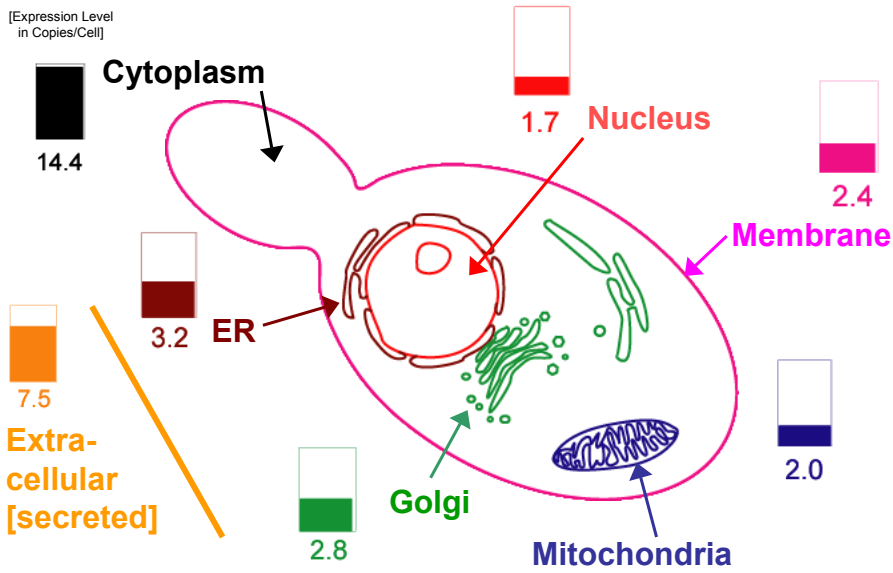
Determine how many of the ~6000 yeast proteins go into each compartment



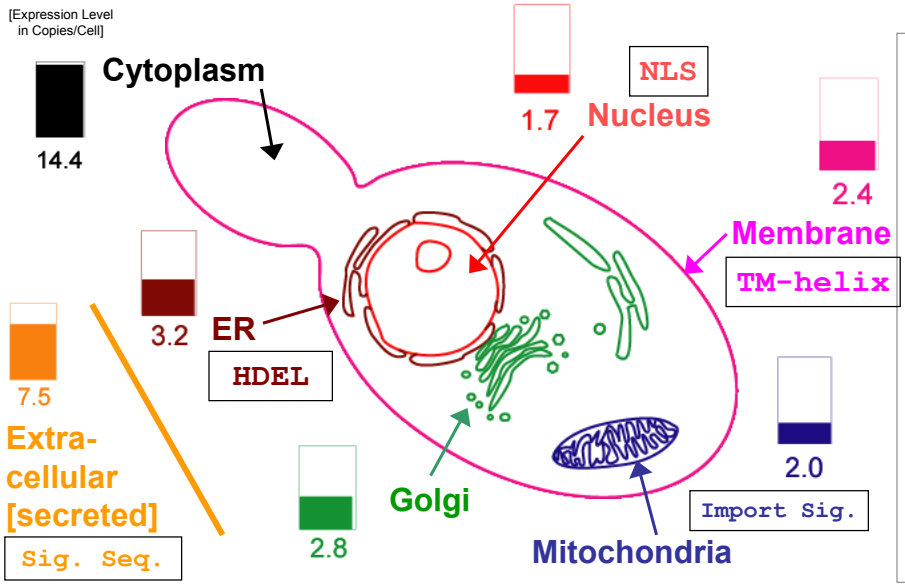
"Traditionally" subcellular localization is "predicted" by sequence patterns



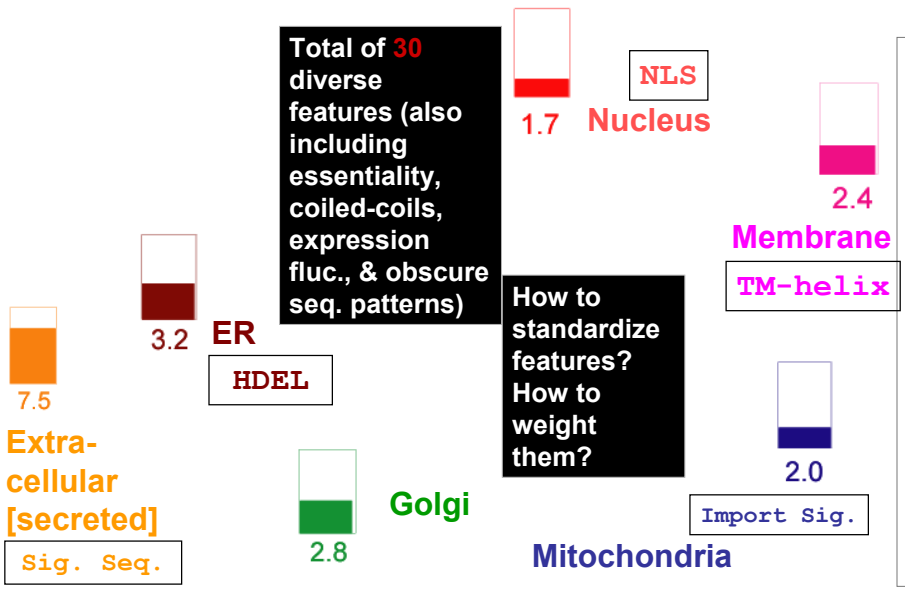
Subcellular localization is associated with the level of gene expression



Combine Expression Information & Sequence Patterns to Predict Localization



Issues in Combining Many Features

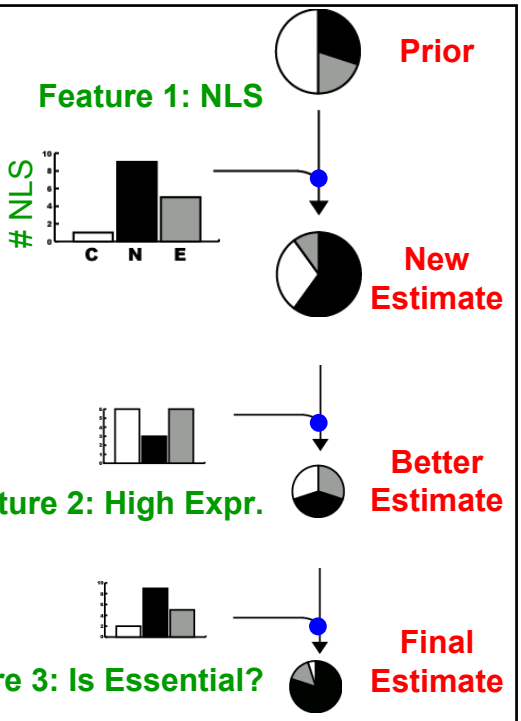


Bayesian System for Localizing Proteins

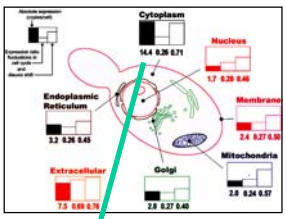
a) Everything in **standard probabilistic terms** (Handles indefinite proteins, 50% cyt., 50% nuc.)

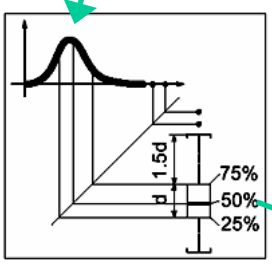
b) Sequentially **combine features** using **Bayes Rule** (**Feature** x **Prior** / Normalization)

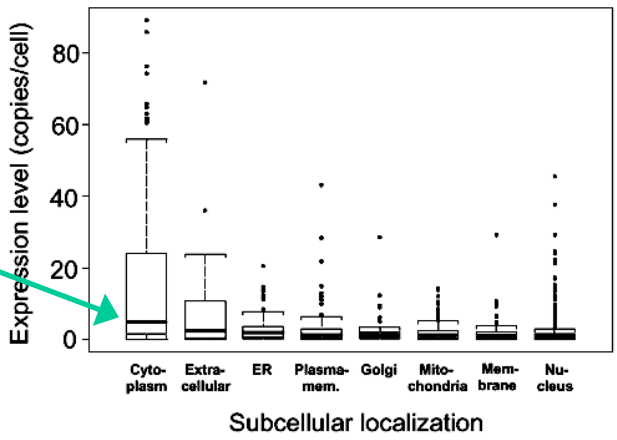
c) Final estimate naturally **weights** features



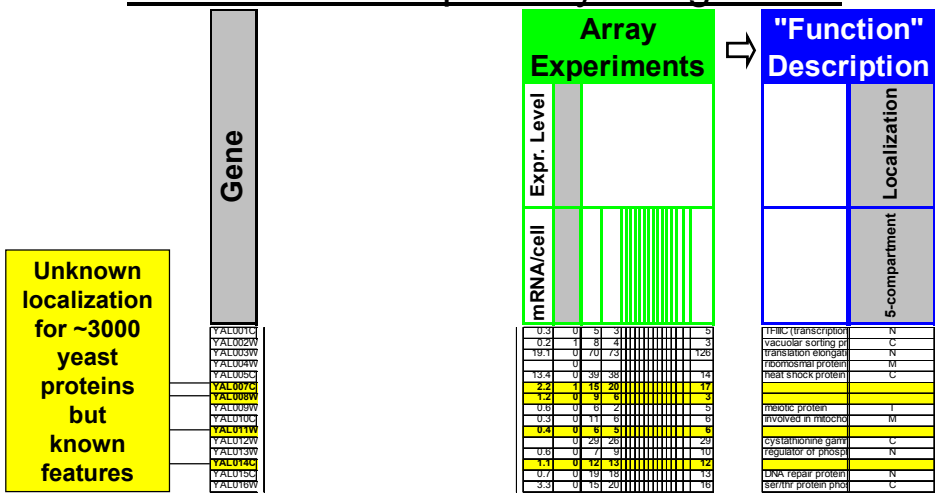
Distributions of Expression Levels





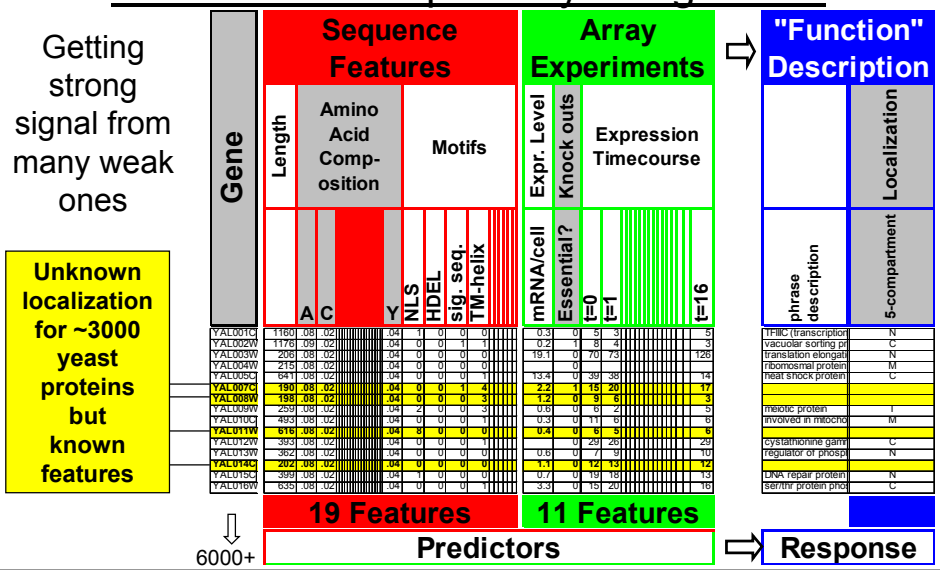


Integrate heterogeneous set of 30 diverse features to predict localization for uncharacterized part of yeast genome



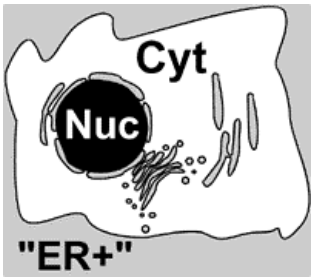
Integrate heterogeneous set of 30 diverse features to predict localization for uncharacterized part of yeast genome

Getting strong signal from many weak ones



Bayesian System for Localizing Proteins: Prior

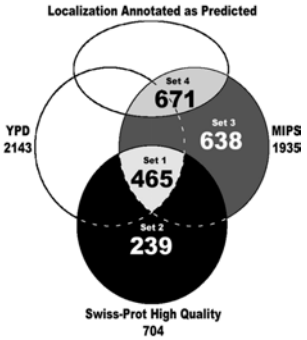
Simplified Cell: 3 (5) compartment

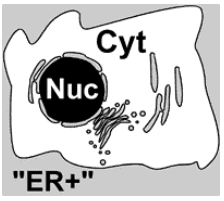


Prior probability distribution for a protein to be in each compartment

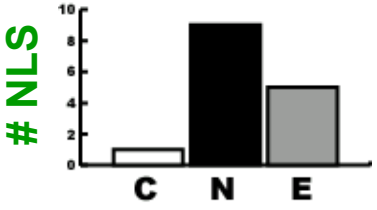
Bayesian System for Localizing Proteins: Features

Training Data: 1342 proteins with known localizations

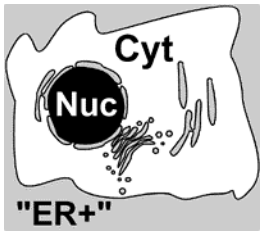




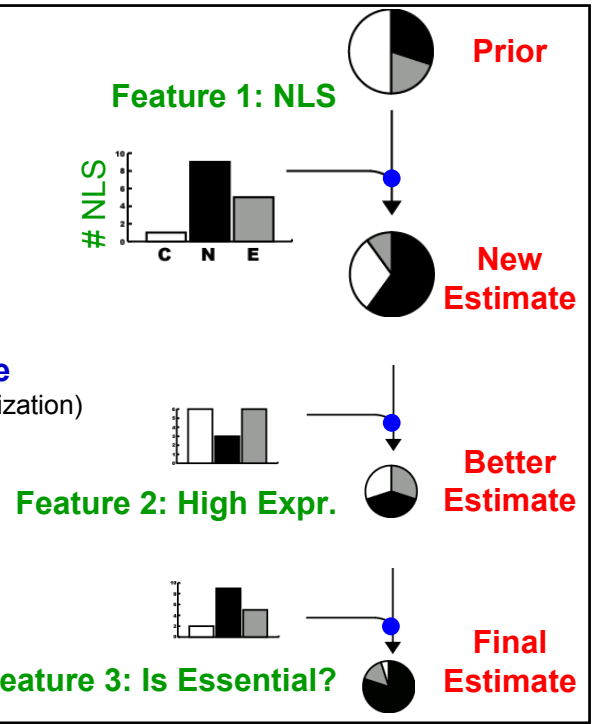
Tabulate occurrence of feature across comparments in training data for all 30 features



Bayesian System for Localizing Proteins: Bayes Rule



from **Bayes Rule**
(**Feature** x **Prior** / Normalization)



P(c|F) = P(F|c) P(c) / P(F)

P(c|F): Probability that protein is in class c given it has feature F

P(F|c): Probability in training data that a protein has feature F if it is class c

P(c): Prior probability that that protein is in class c

P(F): Normalization factor set so that sum over all classes c and ~c is 1 – i.e. P(c|F) + P(~c|F) = 1

Bayes Rule

This formula can be iterated with P(c) [at iter. i+1] <= P(c|F) [at iter. i]

$$C_{MAP} = \arg\max_{C_j \in \{C_1, C_2\}} P(c_j) \prod_{i=1}^{n} P(x_i \mid c_j)$$

Graphical Models - Microsoft Internet Explorer

File Edit View Favorites Tools Help

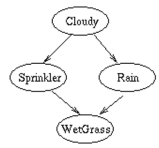
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss Links AT&T Best of the Web

Address C:\My Documents\Bayesian_summary.html Go

BN

example, in which all nodes are binary, i.e., have two possible values, which we will denote by T (true) and F (false).

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.8	0.2
T	0.2	0.8

C	P(R=F)	P(R=T)
F	0.5	0.5
T	0.9	0.1

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

We see that the event "grass is wet" (W=true) has two possible causes: either the water sprinker is on (S=true) or it is raining (R=true). The strength of this relationship is shown in the table. For example, we see that Pr(W=true | S=true, R=false) = 0.9 (second row), and hence, Pr(W=false | S=true, R=false) = 1 - 0.9 = 0.1, since each row must sum to one. Since the C node has no parents, its CPT specifies the prior probability that it is cloudy (in this case, 0.5).

The simplest conditional independence relationship encoded in a Bayesian network can be stated as follows: a node is independent of its ancestors given its parents,

Internet

Optional: not needed for Quiz

Yeast Tables for Localization Prediction

Basics		Predictors: Sequence Features												Predictors: Genomic Features								Response	Bayesian Localization						
		seq. length	Amino Acid Composition					How many times does the sequence have these motif features?						Abs. expr. Level (mRNA copies / cell)		Cell cycle timecourse				function		Localization	State Vector giving localization prediction					Collapsed Prediction	
Yeast Gene ID	Sequence		A	C	D	W	Y	farn site	NLS	hdel motif	nuc2	signalp	tms1	Gene-Chip expt. from RY Lab	sage tag freq.	t=0	t=1	t=15	t=16	function	function	5-compartment	C	N	M	T	E	Training	Extrapolation
YAL001C	M	1160	.08	.02	.06	.01	.04	0	1	0	1	0	0	0.3	0	5	3	4	5	04	TF	N	0%	100%	0%	0%	0%	N	
YAL002W	K	1176	.09	.02	.06	.01	.04	0	0	0	0	0	1	0.2	?	8	4	4	3	06	va	C	95%	3%	2%	0%	0%	C	
YAL003W	K	206	.08	.02	.06	.01	.04	0	0	0	0	0	0	19.1	19	70	73	98	126	05	tra	N	67%	33%	0%	0%	0%	C	
YAL004W	R	215	.08	.02	.06	.01	.04	0	0	0	0	0	0	?	0	18	12	4	6	01	0	N	41%	59%	0%	0%	0%	N	
YAL005C	V	641	.08	.02	.06	.01	.04	0	0	0	0	0	1	13.4	16	39	38	8	14	06	he	????	68%	32%	0%	0%	0%		C
YAL007C	K	190	.08	.02	.06	.01	.04	0	0	0	0	1	4	2.2	8	15	20	16	17	#	??	????	26%	43%	31%	0%	0%		-
YAL008W	H	198	.08	.02	.06	.01	.04	0	0	0	0	0	3	1.2	?	9	6	2	3	#	??	????	37%	60%	3%	0%	0%		-
YAL009W	P	259	.08	.02	.06	.01	.04	0	2	0	0	0	3	0.6	?	6	2	3	5	03	m	????	2%	98%	0%	0%	0%		N
YAL010C	M	493	.08	.02	.06	.02	.04	0	0	0	0	0	1	0.3	?	11	6	6	6	#	in	????	6%	90%	4%	0%	0%		N
YAL011W	K	616	.08	.02	.06	.01	.04	0	8	0	1	0	0	0.4	?	6	5	5	6	30	pr	????	28%	62%	10%	0%	0%		N
YAL012W	G	393	.08	.02	.06	.01	.04	0	0	0	0	0	1	8.9	4	29	26	23	29	01	cy	C	92%	5%	4%	0%	0%	C	
YAL013W	R	362	.08	.02	.06	.01	.04	0	0	0	0	0	0	0.6	?	7	9	6	10	01	re	N	0%	98%	0%	0%	1%	N	
YAL014C	G	202	.08	.02	.06	.01	.04	0	0	0	0	0	0	1.1	?	12	13	9	12	#	??	N	1%	96%	4%	0%	0%	N	
YAL015C	M	399	.08	.02	.06	.01	.04	0	1	0	0	0	0	0.7	0	19	18	12	13	11	D	N	4%	96%	0%	0%	0%	N	
YAL016W	K	635	.08	.02	.06	.01	.04	0	0	0	0	0	1	3.3	5	15	20	16	16	03	se	????	74%	26%	0%	0%	0%		C
YAL017W	V	1356	.08	.02	.06	.01	.04	0	0	0	0	0	0	0.4	?	14	3	4	7	#	??	????	0%	1%	99%	0%	0%		M
YAL018C	K	325	.08	.02	.06	.01	.04	0	0	0	0	0	4	?	?	4	2	2	1	#	??	????	0%	100%	0%	0%	0%		N

Large-scale Datamining

- Gene Expression
  - ◊ Representing Data in a Grid
  - ◊ Description of function prediction in abstract context
- Unsupervised Learning
  - ◊ clustering & k-means
  - ◊ Local clustering
- Supervised Learning
  - ◊ Discriminants & Decision Tree
  - ◊ Bayesian Nets
- Function Prediction EX
  - ◊ Simple Bayesian Approach for Localization Prediction