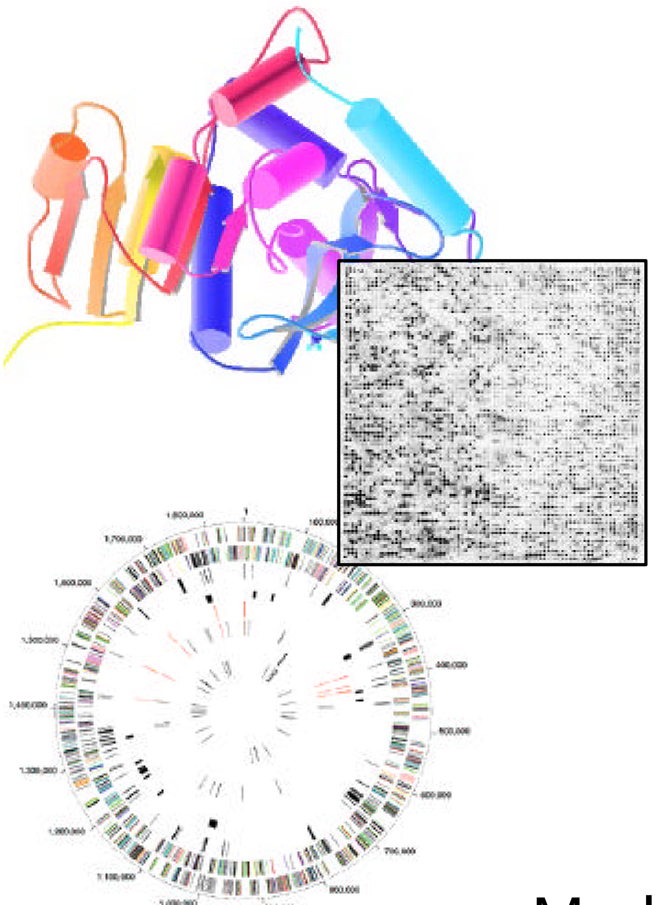
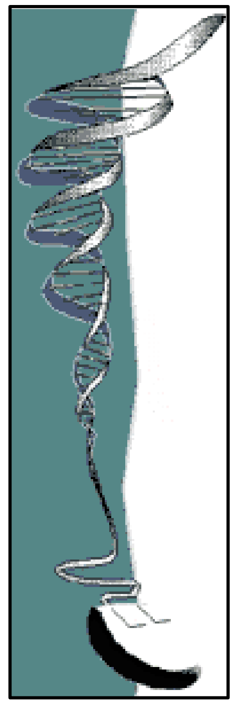
BIOINFORMATICS
Datamining





Mark Gerstein, Yale University

bioinfo.mbb.yale.edu/mbb452a

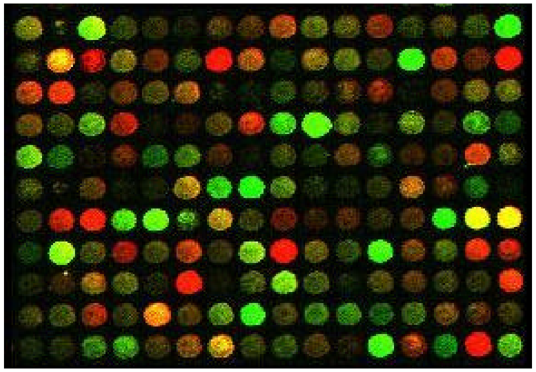
Large-scale Datamining

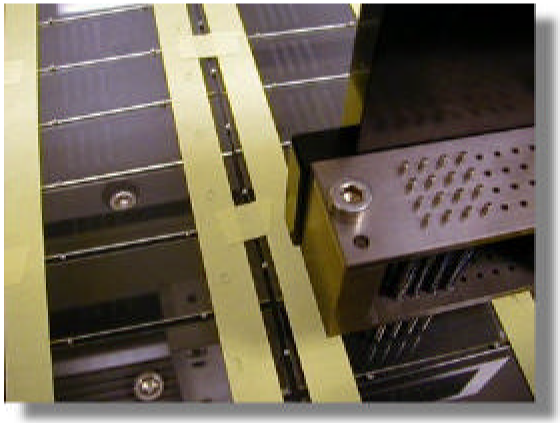
- Relating Gene Expression to Protein Features and Parts
- Supervised Learning: Discriminants
- Simple Bayesian Approach for Localization Prediction
- Unsupervised Learning: k-means
- Correlation of Expression Data with Function
- Overview of Issues in Datamining
- Overview of Methods of Supervised Learning
- Focus on Decision Trees
- Overview of Methods of Unsupervised Learning
- Cluster Trees, Evolutionary Trees

microarrays

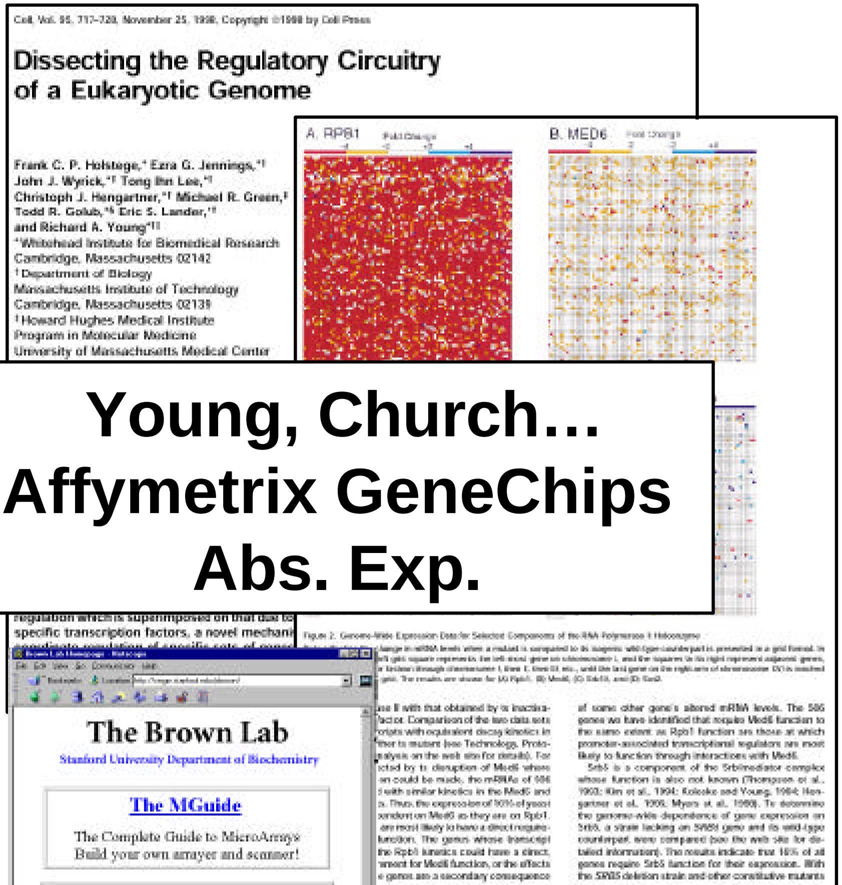
- Affymetrix
  - o Oligos
    - – Don’t have to know sequence

- Glass slides
  - ◊ Pat brown





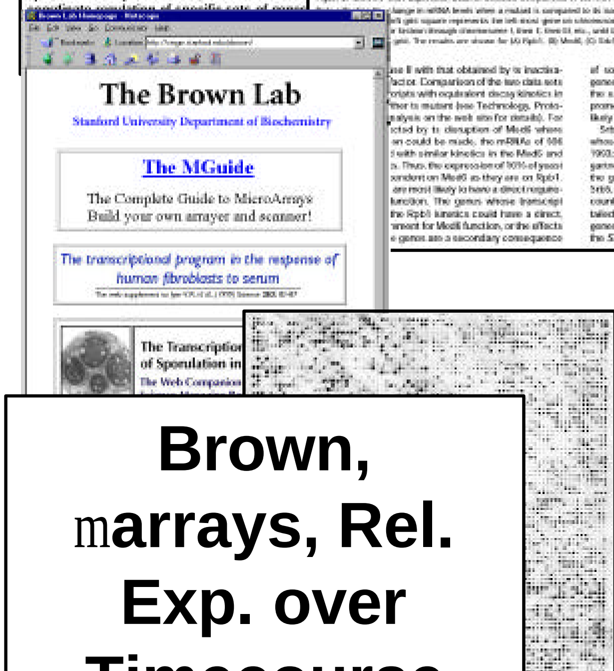
Gene Expression Datasets: the Yeast Transcriptome


Cell, Vol. 95, 717–728, November 25, 1998, Copyright ©1998 by Cell Press

Dissecting the Regulatory Circuitry of a Eukaryotic Genome

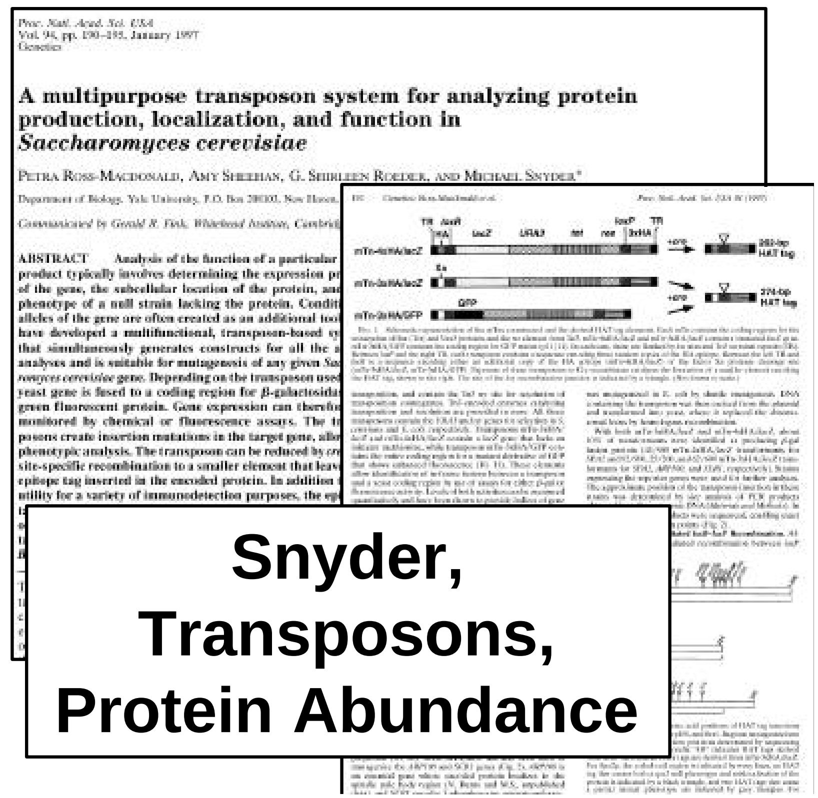
Frank C. P. Holstege, Ezra G. Jennings, John J. Wyrick, Tong Ihn Lee, Christoph J. Hengartner, Michael R. Green, Todd R. Golub, Eric S. Lander, and Richard A. Young

Young, Church… Affymetrix GeneChips Abs. Exp.



Brown, marrays, Rel. Exp. over Timecourse

Yeast Expression Data: 6000 levels!
Integrated Gene Expression Analysis System: X-ref. Parts and Features against expression data...


Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 190–195, January 1997
Genetics

A multipurpose transposon system for analyzing protein production, localization, and function in *Saccharomyces cerevisiae*

Petra Ross-Macdonald, Amy Sheehan, G. Shirleen Roeder, and Michael Snyder

Also:
SAGE (mRNA);
2D gels for Protein Abundance (Aebersold, Futcher)

Snyder, Transposons, Protein Abundance

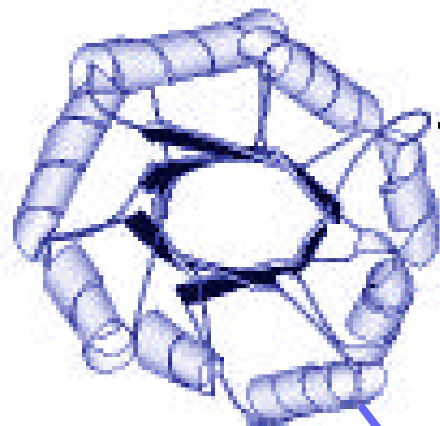
Gene Expression Information and Protein Features

Basics | **Predictors**: **Sequence Features** (Sequence, seq. length, Amino Acid Composition, How many times does the sequence have these motif features?) and **Genomic Features** (Abs. expr. Level (mRNA copies / cell), Prot. Abundance, Cell cycle timecourse)

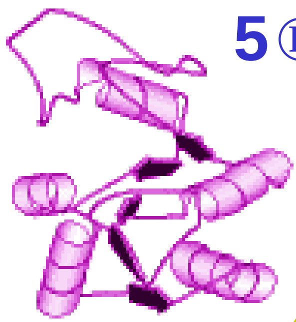
Yeast Gene ID	Sequence	seq. length	Amino Acid Composition: A	C	D	W	Y	farn site	NLS	hdel motif	nuc2	signalp	tms1	Abs. expr. Level (mRNA copies / cell): Gene-Chip expt. from RY Lab	sage tag freq.	Prot. Abundance (1000 copies /cell)
YAL001C	M	1160	.08	.02	.06	.01	.04	0	1	0	1	0	0	0.3	0	?
YAL002W	K	1176	.09	.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	?	?	?

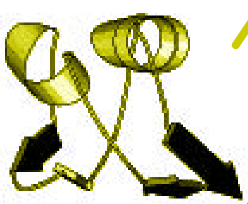
Yeast Gene ID	Cell cycle timecourse: t=0	t=1	t=2	t=3	t=4	t=5	t=6	t=7	t=8	t=9	t=10	t=11	t=12	t=13	t=14	t=15	t=16
YAL001C	5	3	4	4	5	4	3	5	5	3	5	7	9	4	4	4	5
YAL002W	8	4	2	3	4	3	4	5	5	3	4	4	6	4	5	4	3
YAL003W	70	73	91	69	105	52	112	88	64	159	106	104	75	103	140	98	126
YAL004W	18	12	9	5	5	3	6	4	4	3	3	5	5	4	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

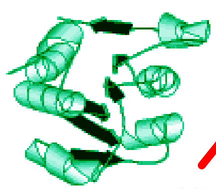
Common Parts: the Transcriptome



5®1





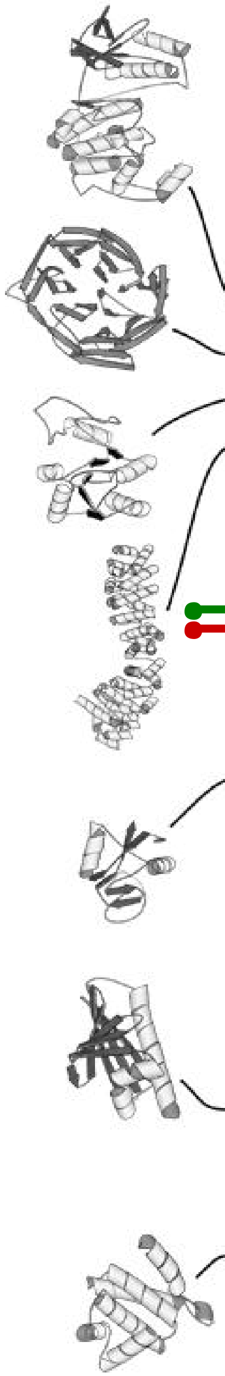


1®18

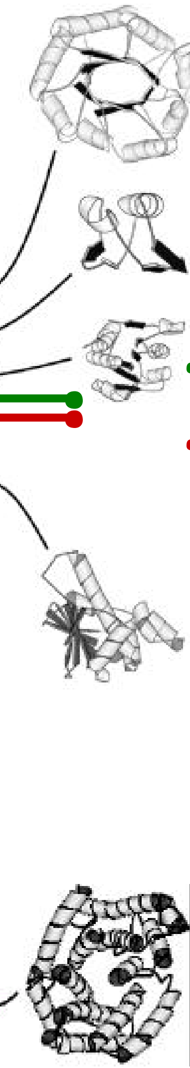
7®15

Fold	Fold Class	Rep. PDB	Composition: Genome [%]	Composition: Transcriptome [%]	Composition: Rel. Diff. [%]		Rank: Genome	Rank: Young	Rank: Samson	Rank: Church-a	Rank: Church-alpha	Rank: Church-gal	Rank: Church-heat	Rank: SAGE-GM	Rank: SAGE-L	Rank: SAGE-S
TIM barrel	α/β	1byb	4.2	8.3	+98	+	5	1	1	1	1	1	1	1	1	1
P-loop NTP hydrolases	α/β	1gky	5.8	5.2	-11	•	3	2	2	4	4	4	5	5	6	7
Ferredoxin like	α+β	1fxd	3.9	3.4	-14	•	6	3	7	11	9	8	10	4	10	11
Rossmann fold	α/β	1xel	3.3	3.3	0	•	8	4	3	3	3	2	2	19	15	9
7-bladed beta-propeller	β	1mda*	6.4	2.9	-55	-	2	5	4	5	6	6	7	9	9	16
aplha-alpha superhelix	α	2bct	4.4	2.7	-37	-	4	6	11	15	16	12	12	8	5	8
Thioredoxin fold	α/β	2trx	1.7	2.7	+63	+	14	7	6	8	2	5	4	11	10	6
G3P dehydrogenase-like	α+β	1drw†	0.2	2.7	+1316	+	81	8	12	2	5	3	3	35	19	30
beta grasp	α+β	1igd	0.6	2.6	+348	+	36	9	10	21	9	18	21	82	122	120
HSP70 C-term. fragment	multi	1dky	0.8	2.6	+231	+	31	10	16	17	11	16	12	48	25	56
Leu-zipper	α	1zta	3.8	2.1	-46	-	7	15	8	14	21	15	19	21	20	33
Protein kinases (cat. core)	multi	1hcl	6.8	1.6	-77	-	1	18	19	9	16	11	15	13	16	17
alpha/beta hydrolases	α/β	2ace	2.2	0.9	-62	-	10	32	31	25	26	21	23	26	26	26
Zn2/C6 DNA-bind. dom.	sml	1aw6	2.6	0.3	-89	-	9	75	94	27	50	32	40	48	39	50

Feature *F* is Folds, in particular the TIM-barrel (3.1)	Number of TIM-barrel fold matches in yeast genome	Number of matches with all folds in yeast genome	Genome composition of TIM-barrel fold matches	Number of TIM-barrel fold matches weighted by expression	Number of matches with all folds weighted by expression	Transcriptome composition of TIM-barrel fold matches	Relative enrichment of TIM-barrel matches in transcriptome
Spec. Num.	**65**	**1560**	**4.2%**	**389**	**4709**	**8.3%**	**97.8%**



Fold of	Freq.		Change					
	Genome	Transcriptome	CDC28	CDC15	Diauxic Shift	Sporulation	*E. coli* heat shock	Rep. PDB
Protein kinases (cat. core)	1	18	94	98	139	60	100	1p38
β-propeller	2	5	160	108	109	82	-	1mda
P-loop NTP hydrolases	3	2	100	88	91	57	39	1gky
α–α superhelix	4	6	136	90	110	44	55	2bct
TIM-barrel	5	1	58	57	39	24	91	1byb
Ferredoxin-like	6	3	135	61	63	70	144	1fxd
Rossmann fold	8	4	55	99	43	56	92	1xel
Ribonucleotide reductase (R1)	100	143	1	-	-	-	35	1rlr
ATPase dom. of HSP90	100	91	2	4	72	73	2	1ah6
Homing endonuclease-like	130	164	3	136	85	175	41	1af5
Aminoacid dehydrogenases; dim. dom.	-	-	4	169	121	3	51	1hup
DNA topo I (N-term)	-	-	175	1	148	126	-	1ois
DNA clamp	130	115	8	2	87	11	60	2pol
Metallothionein	100	14	89	3	33	12	-	1mhu
Phosphoenolpyruvate carboxykinase	130	190	51	26	1	98	169	1ayl
Citrate synthase	81	120	14	8	2	28	51	1csh
N-carbamoylsarcosine amidohydrolase	130	112	9	-	3	138	118	1nba
TBP-like	81	91	46	38	4	75	100	1bv1
5'-3' exonuclease	67	150	32	125	162	1	157	1tfr
α/α toroid	62	132	169	145	114	2	100	1gai
Cyclin-like	20	61	20	15	129	4	-	1vin
ATPase domain of GroEL	36	34	183	143	61	151	1	1aon
Head domain of GrpE	130	135	196	31	165	165	3	1dkg
HSP70 (C-term)	31	10	16	11	56	117	4	1dkz

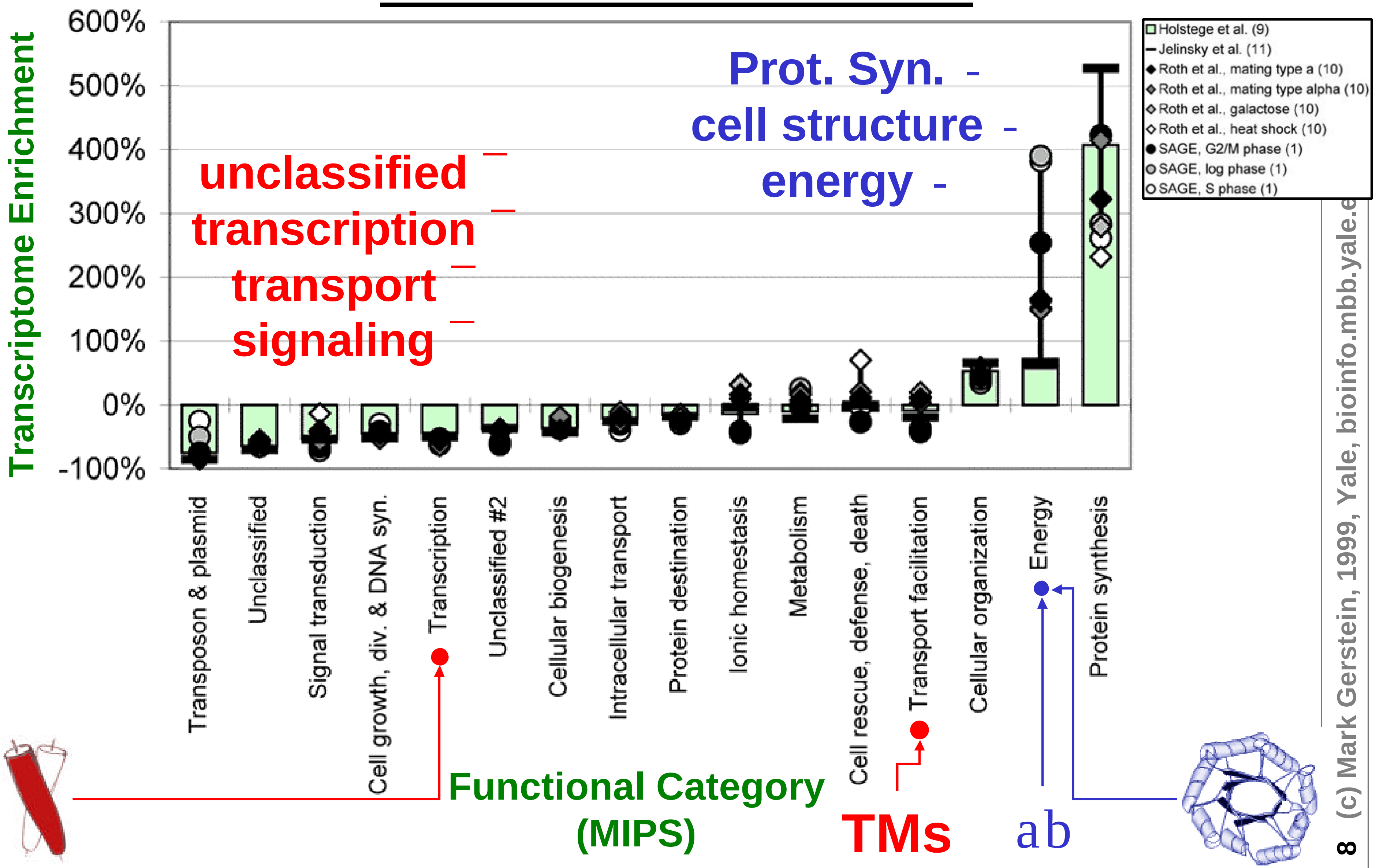


Common Folds

Folds that change a lot in frequency are not common

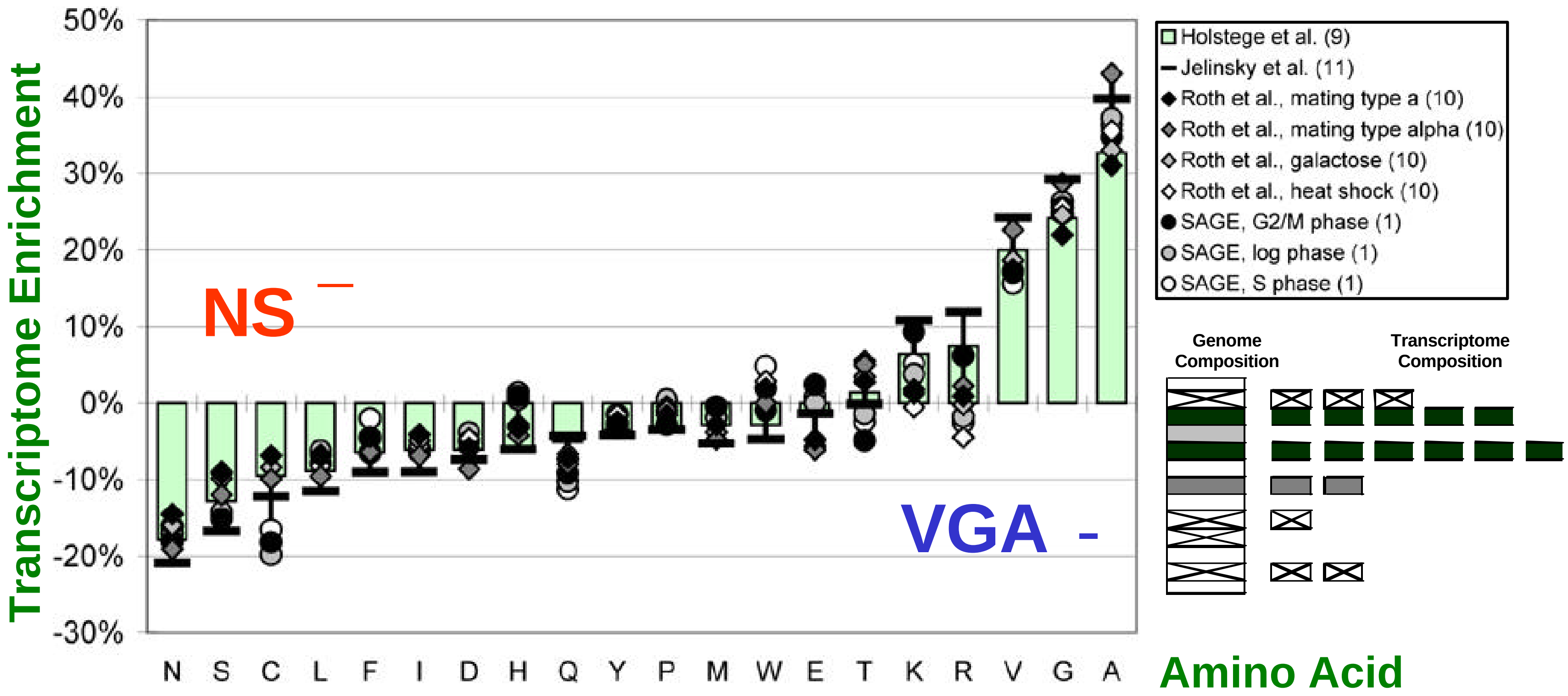
Changing Folds

Composition of Transcriptome in terms of Functional Classes

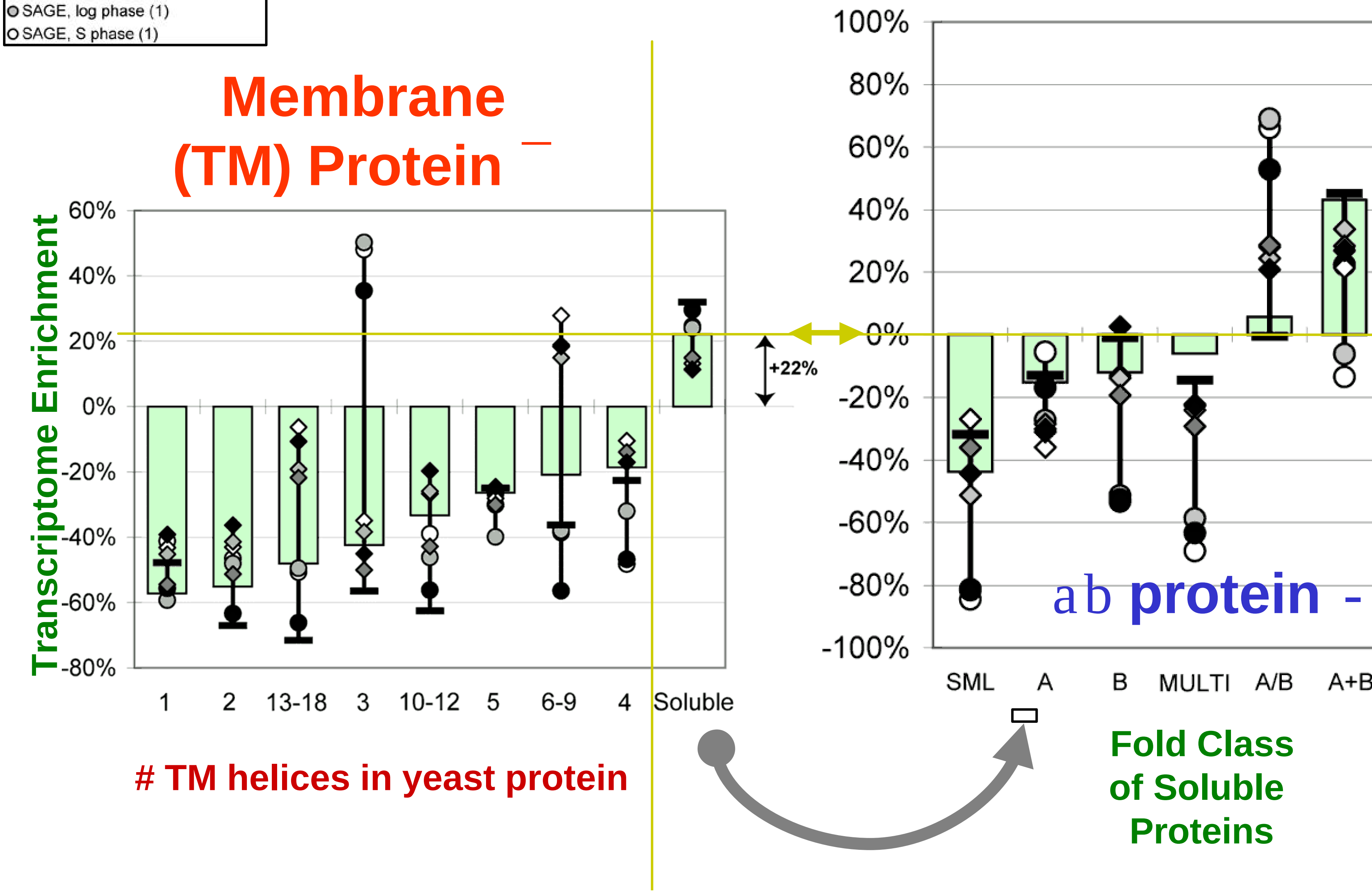


Composition of Genome vs. Transcriptome

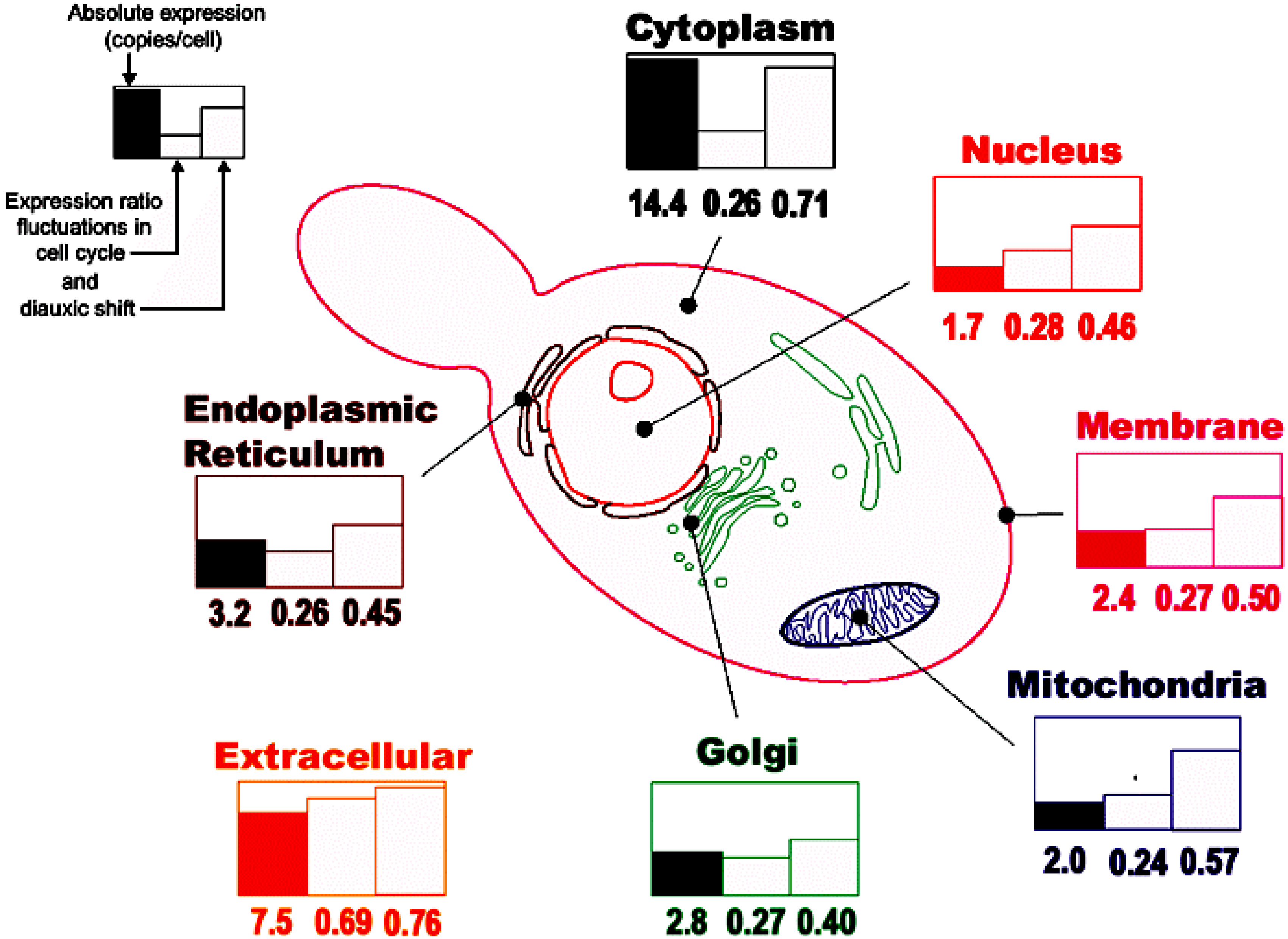
	$\sum_{\text{orf } i} n_i(F)$	$\sum_F \sum_{\text{orf } i} n_i(F)$	$G(F)$	$\sum_{\text{orf } i} e_i n_i(F)$	$\sum_F \sum_{\text{orf } i} e_i n_i(F)$	$T(F)$	$D(F)$
Feature *F* is Amino acids, in particular Ala	Number of Ala in yeast	Number of amino acids in yeast	Genome composition of Ala in yeast	Number of Ala weighted by expression	Number of amino acids weighted by expression	Transcriptome composition of Ala in yeast	Relative enrichment of Ala in transcriptome
Spec. Num.	**141890**	**2574876**	**5.5%**	**347807**	**4758441**	**7.3%**	**32.7%**
Feature *F* is Folds, in particular the TIM-barrel (3.1)	Number of TIM-barrel fold matches in yeast genome	Number of matches with all folds in yeast genome	Genome composition of TIM-barrel fold matches	Number of TIM-barrel fold matches weighted by expression	Number of matches with all folds weighted by expression	Transcriptome composition of TIM-barrel fold matches	Relative enrichment of TIM-barrel matches in transcriptome
Spec. Num.	**65**	**1560**	**4.2%**	**389**	**4709**	**8.3%**	**97.8%**



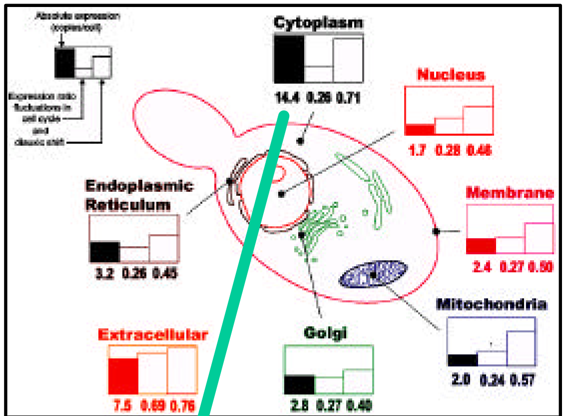
Composition of Transcriptome in terms of Broad Structural Classes

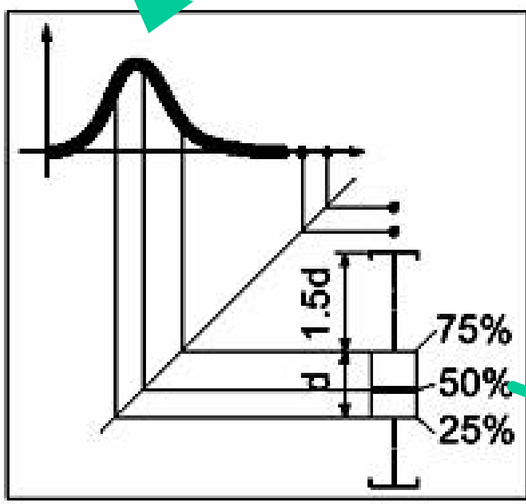


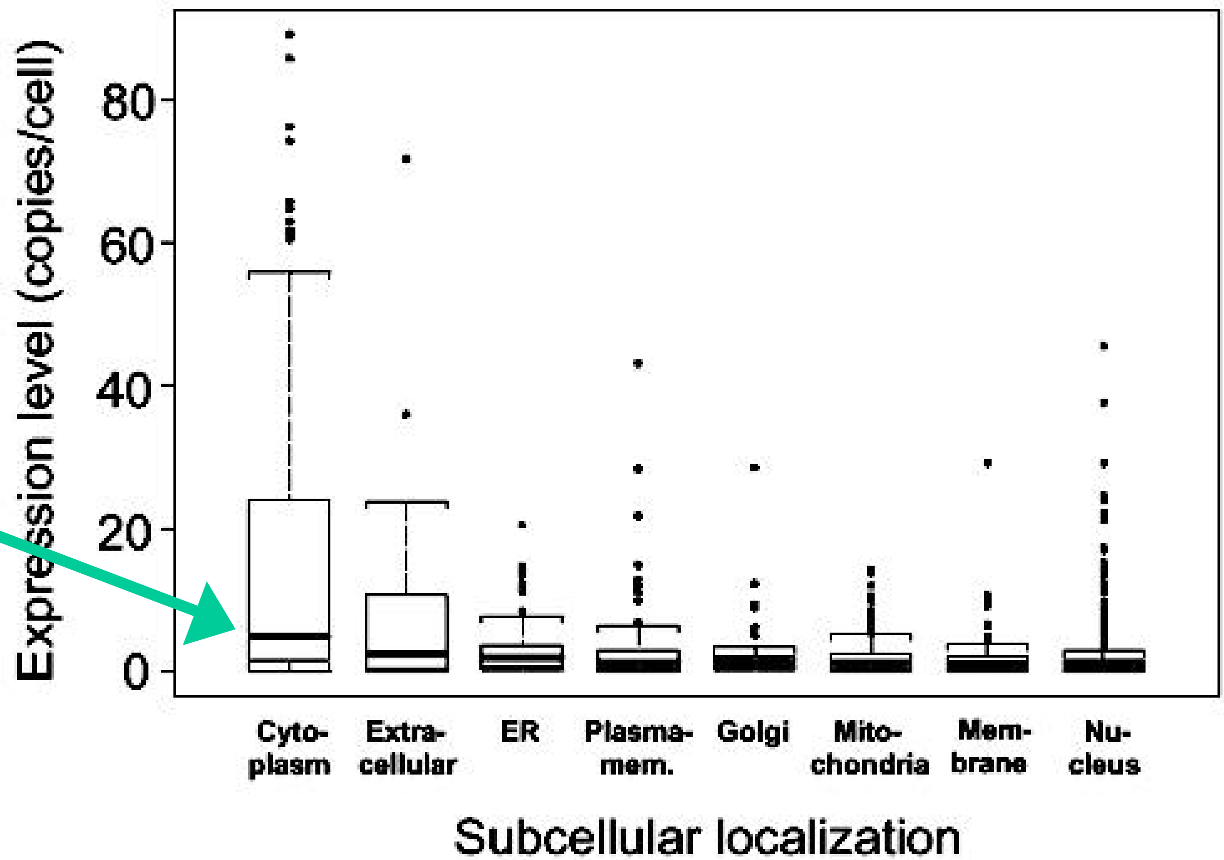
Expression Level is Related to Localization



Distributions of Expression Levels





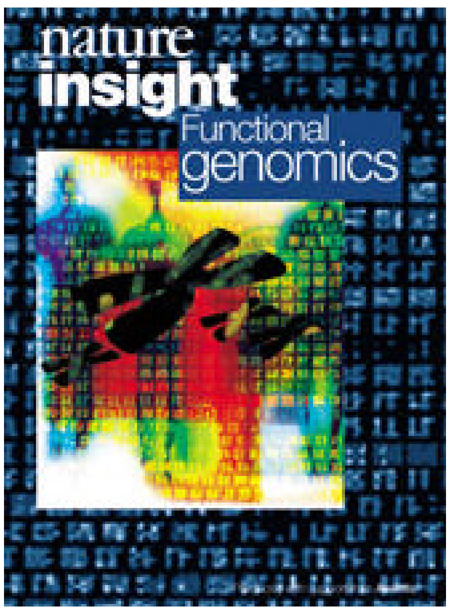


Large-scale Datamining

- Relating Gene Expression to Protein Features and Parts
- Supervised Learning: Discriminants
- Simple Bayesian Approach for Localization Prediction
- Unsupervised Learning: k-means
- Correlation of Expression Data with Function
- Overview of Issues in Datamining
- Overview of Methods of Supervised Learning
- Focus on Decision Trees
- Overview of Methods of Unsupervised Learning
- Cluster Trees, Evolutionary Trees

~6000 yeast genes with expression levels

but only ~2000 with localization….



insight review articles

Genomics, gene expression and DNA arrays

David J. Lockhart & Elizabeth A. Winzeler

Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, San Diego, California 92121, USA

Experimental genomics in combination with the growing body of sequence information promise to revolutionize the way cells and cellular processes are studied. Information on genomic sequence can be used experimentally with high-density DNA arrays that allow complex mixtures of RNA and DNA to be interrogated in a parallel and quantitative fashion. DNA arrays can be used for many different purposes, most prominently to measure levels of gene expression (messenger RNA abundance) for tens of thousands of genes simultaneously. Measurements of gene expression and other applications of arrays embody much of what is implied by the term (genomics); they are broad in scope, large in scale, and take advantage of all available sequence information for experimental design and data interpretation in pursuit of biological understanding.

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.

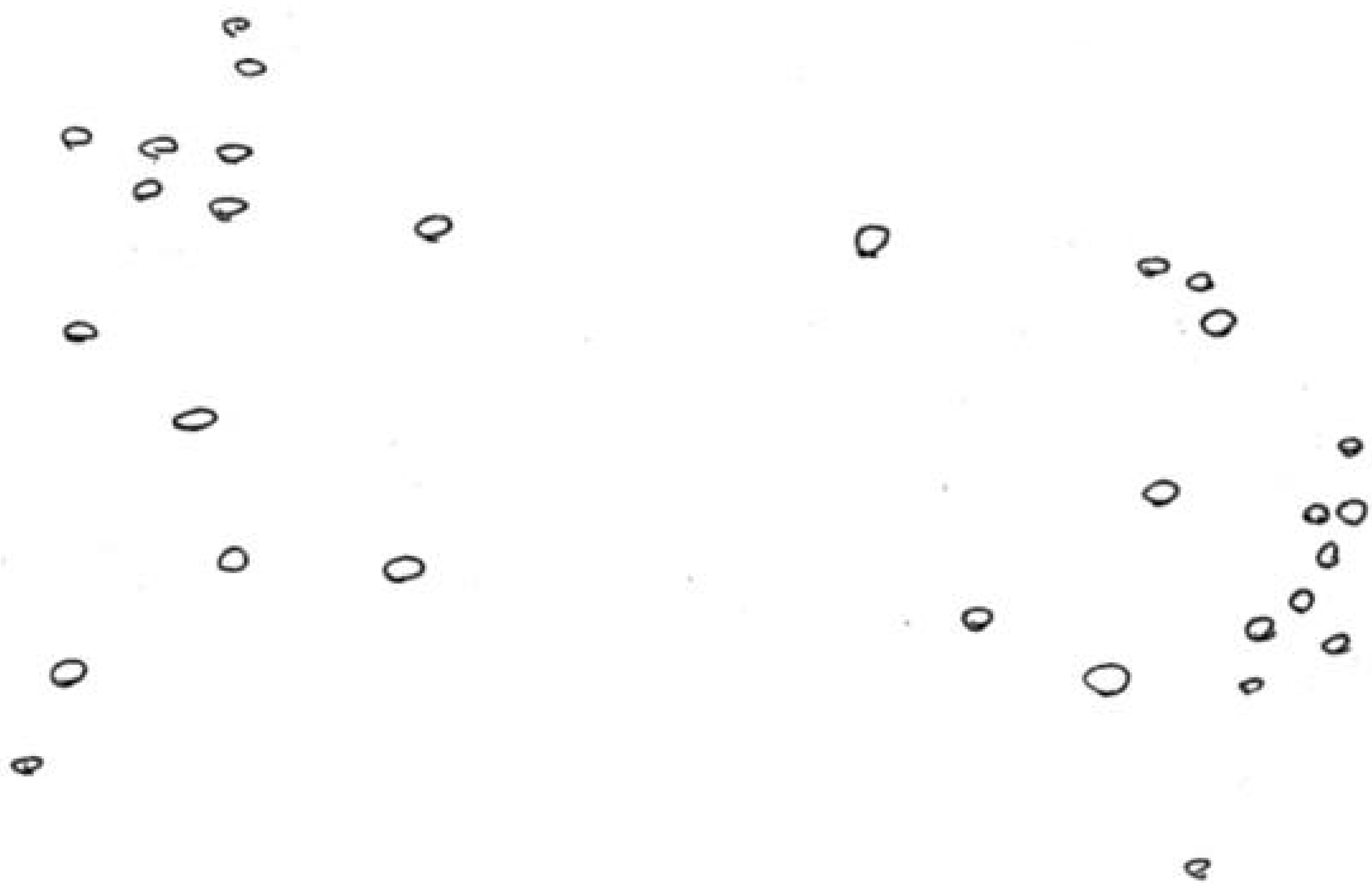
	predictor1	Predictor2	predictor3	predictor4	response
G1	A(1,1)	A(1,2)	A(1,3)	A(1,4)	Class A
G2	A(2,1)	A(2,2)	A(2,3)	A(2,4)	Class A
G3	A(3,1)	A(3,2)	A(3,3)	A(3,4)	Class B

(adapted from Y Kluger)

Typical Predictors and Response for Yeast

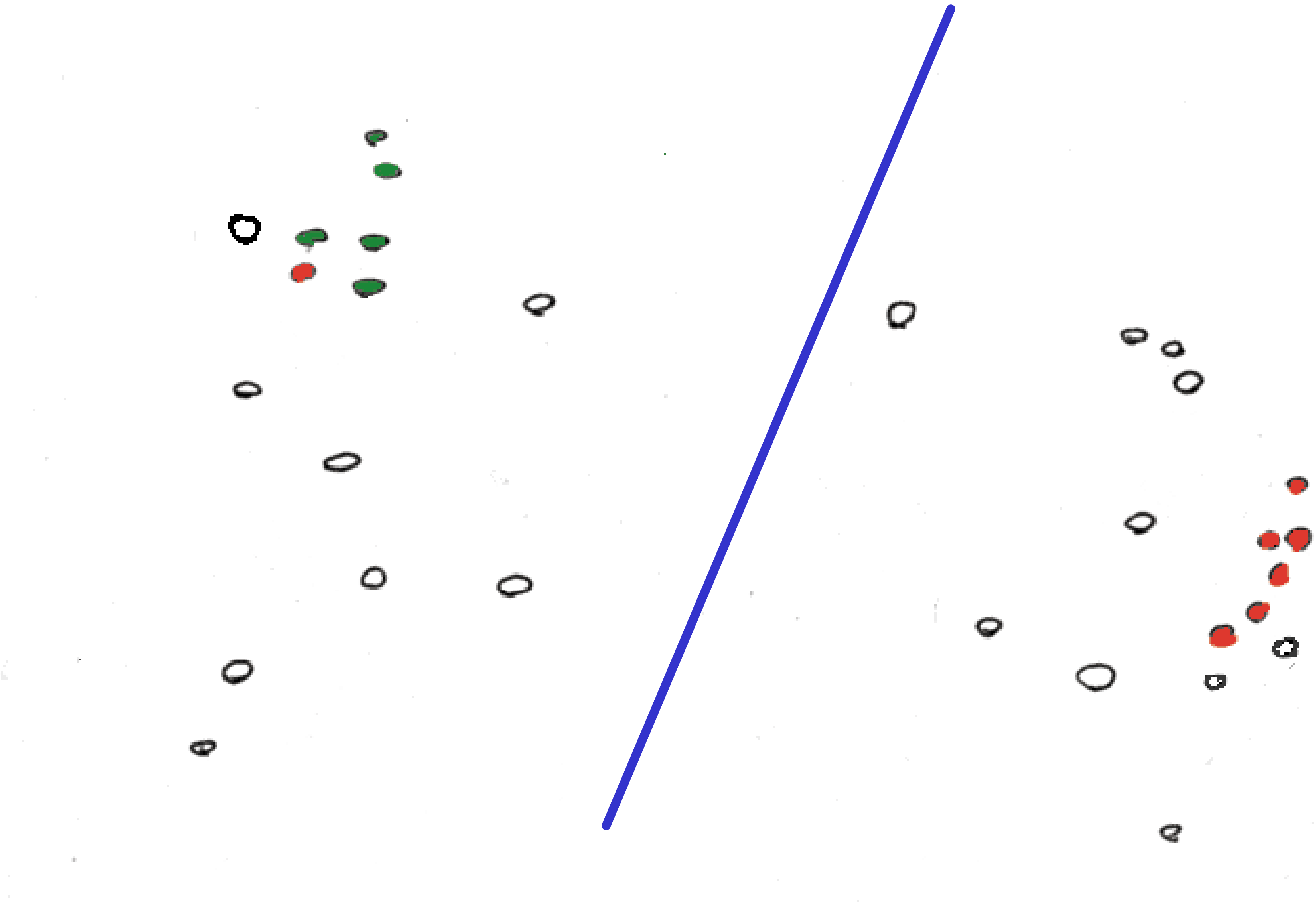
Basics		Predictors: Sequence Features											Predictors: Genomic Features							Response		
		seq. length	Amino Acid Composition					How many times does the sequence have these motif features?						Abs. expr. Level (mRNA copies / cell)		Prot. Abun-dance	Cell cycle timecourse			Function		Localization
Yeast Gene ID	Sequence		A	C	D	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 initia	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	18	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

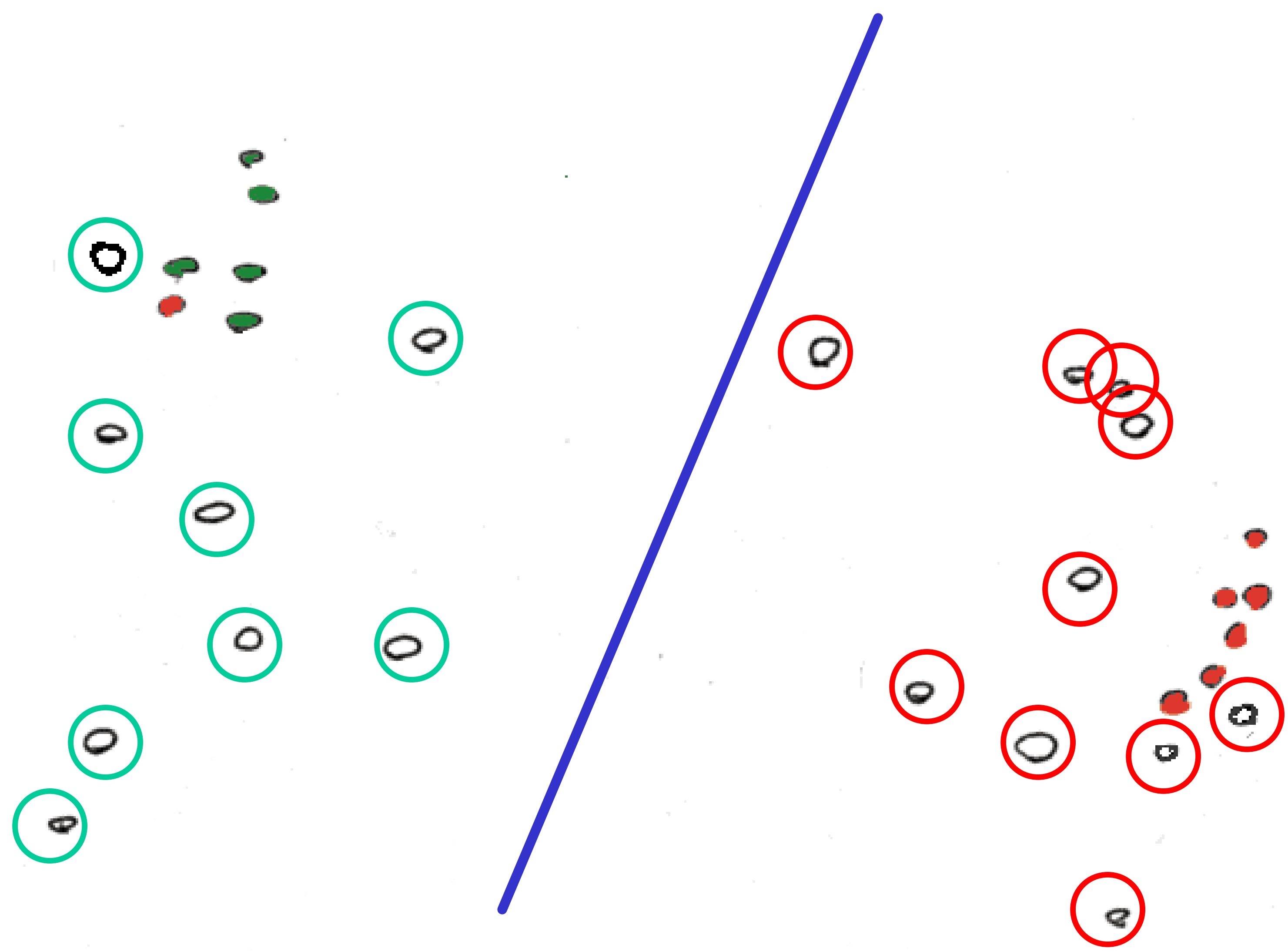


"Tag" Certain Points

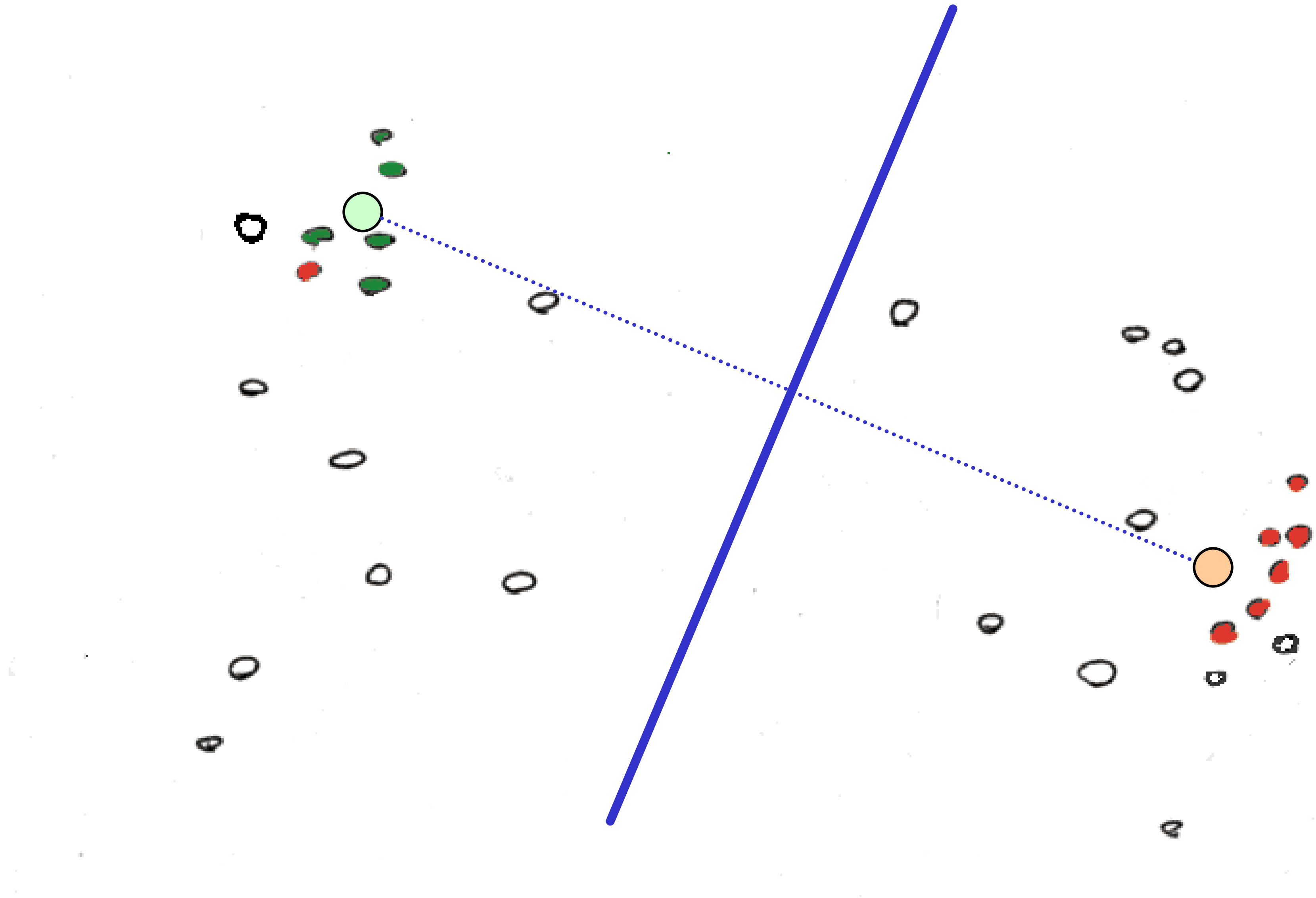
Find a Division to Separate Tagged Points



Extrapolate to Untagged Points



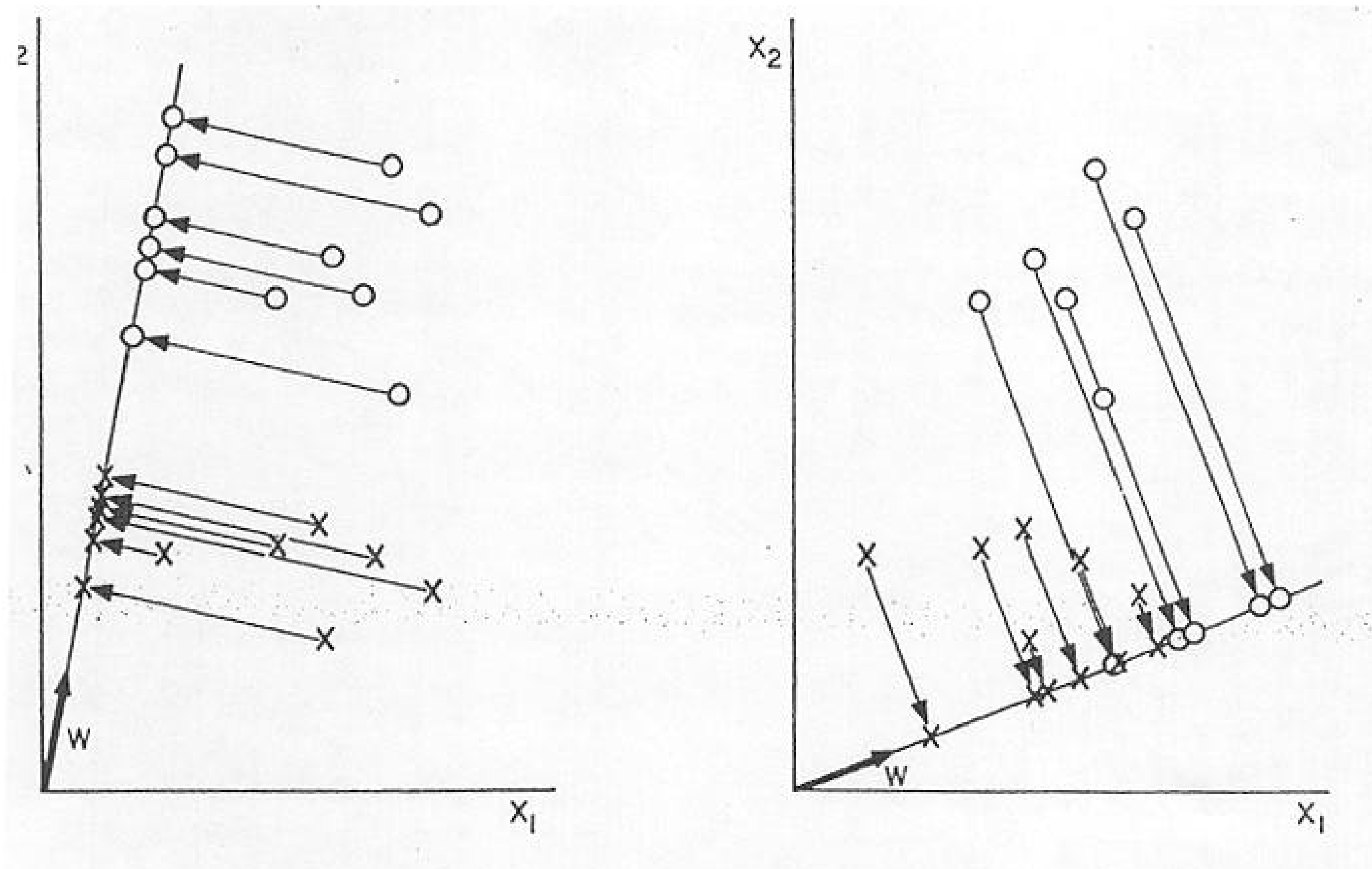
Discriminant to Position Plane



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

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

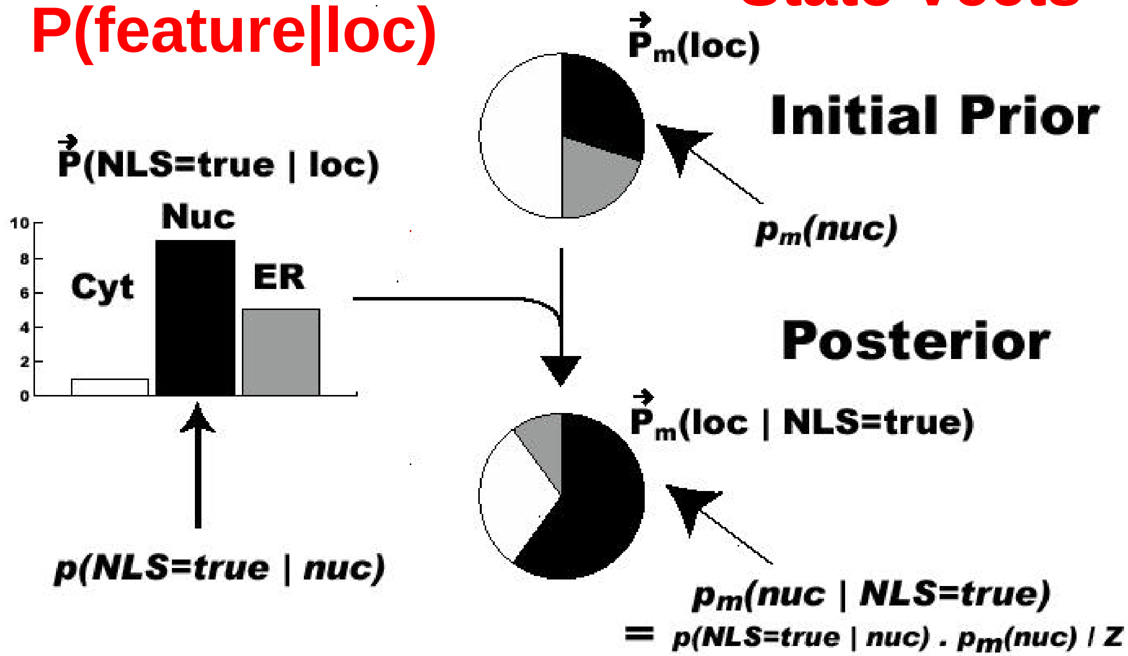
Large-scale Datamining

- Relating Gene Expression to Protein Features and Parts
- Supervised Learning: Discriminants
- Simple Bayesian Approach for Localization Prediction
- Unsupervised Learning: k-means
- Correlation of Expression Data with Function
- Overview of Issues in Datamining
- Overview of Methods of Supervised Learning
- Focus on Decision Trees
- Overview of Methods of Unsupervised Learning
- Cluster Trees, Evolutionary Trees

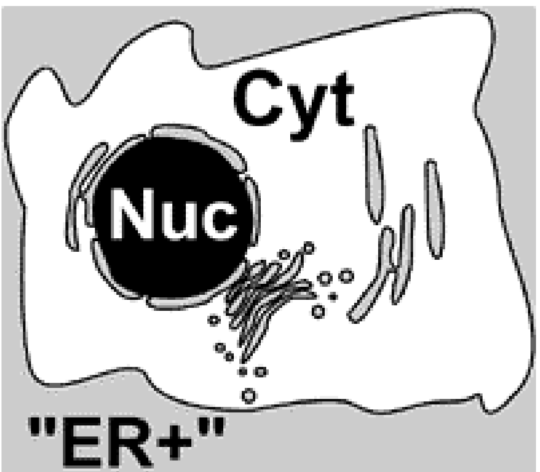
Bayesian System for Localizing Proteins

Feature Vects
P(feature|loc)

State Vects



loc=

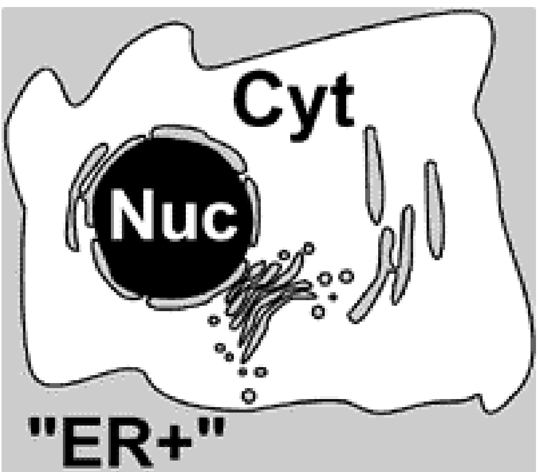


Represent localization of each protein by the state vector **P**(loc) and each feature by the feature vector P(feature|loc). Use Bayes rule to update.

18 Features: Expression Level (absolute and fluctuations), signal seq., KDEL, NLS, Essential?, aa composition

Bayesian System for Localizing Proteins

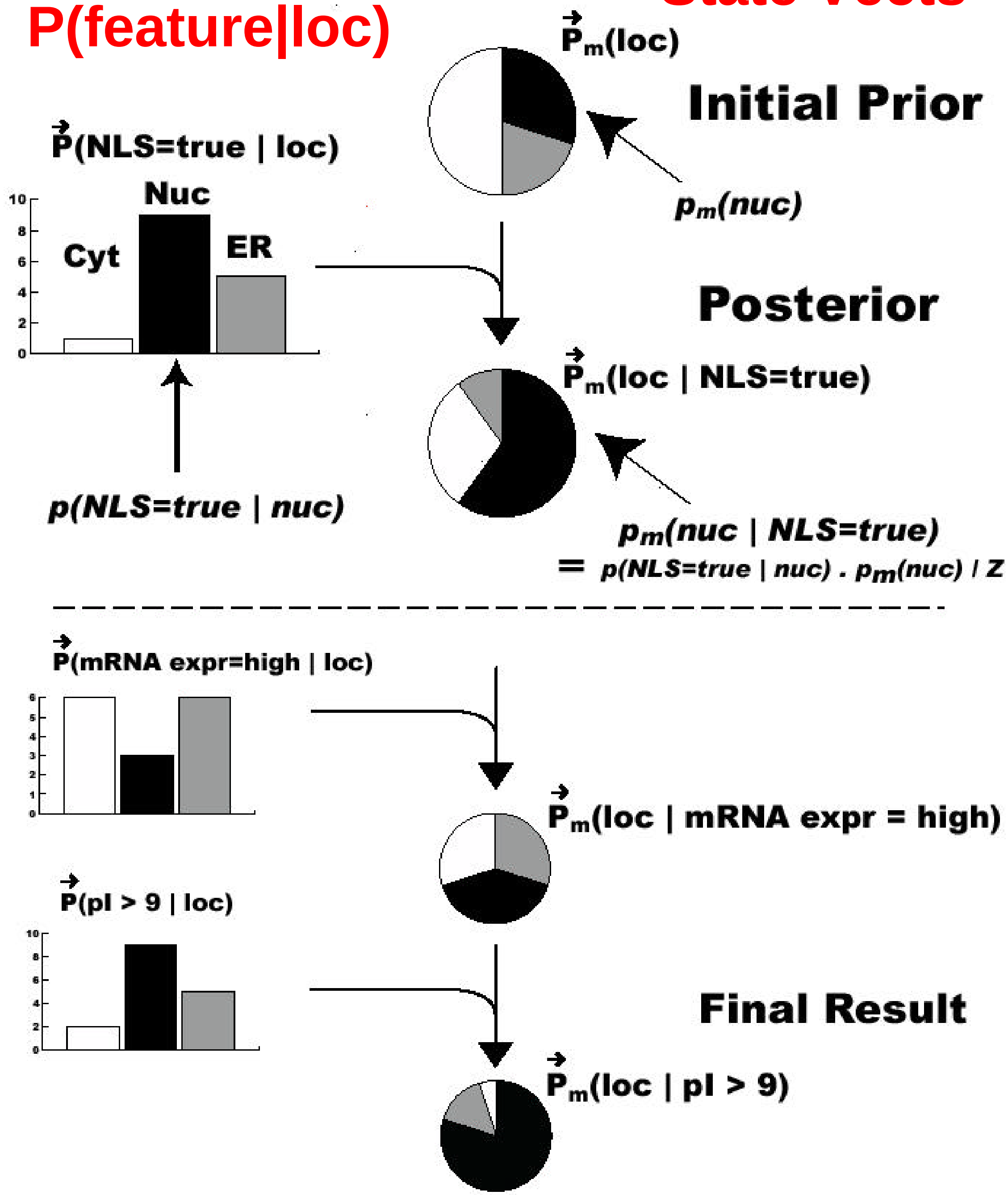
loc=



Represent localization of each protein by the state vector **P**(loc) and each feature by the feature vector P(feature|loc). Use Bayes rule to update.

Feature Vects P(feature|loc)

State Vects



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

Bayes Rule

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

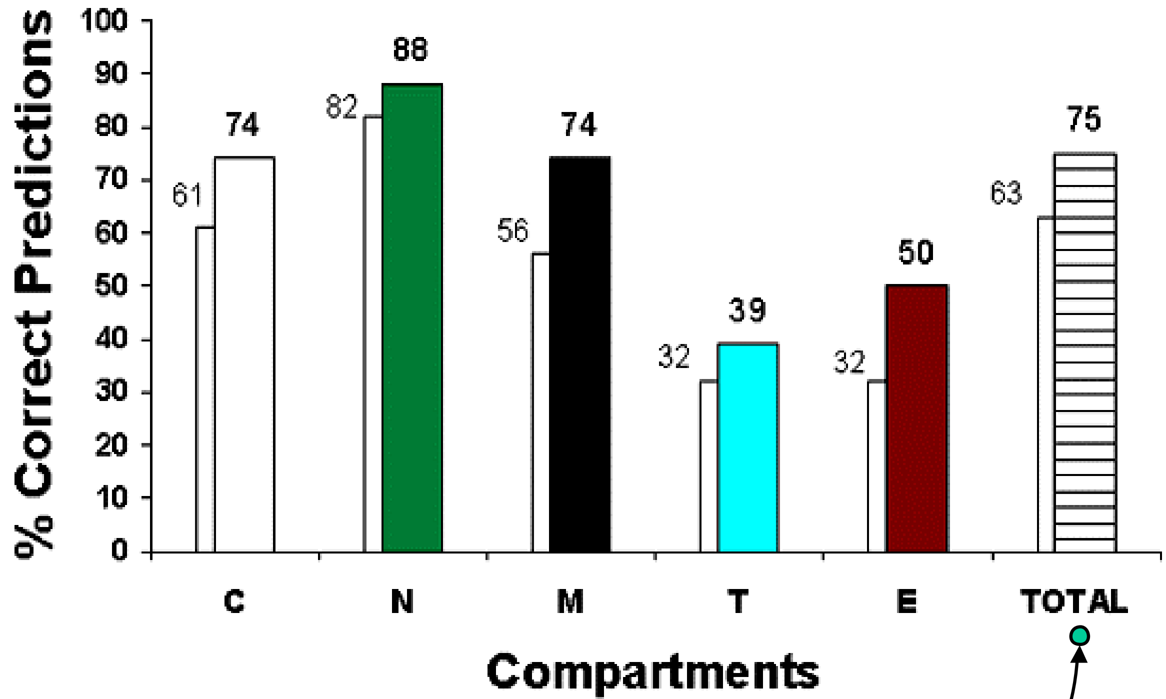
This formula can be iterated with
P(c) [at iter. i+1] <= P(c|F) [at iter. i]

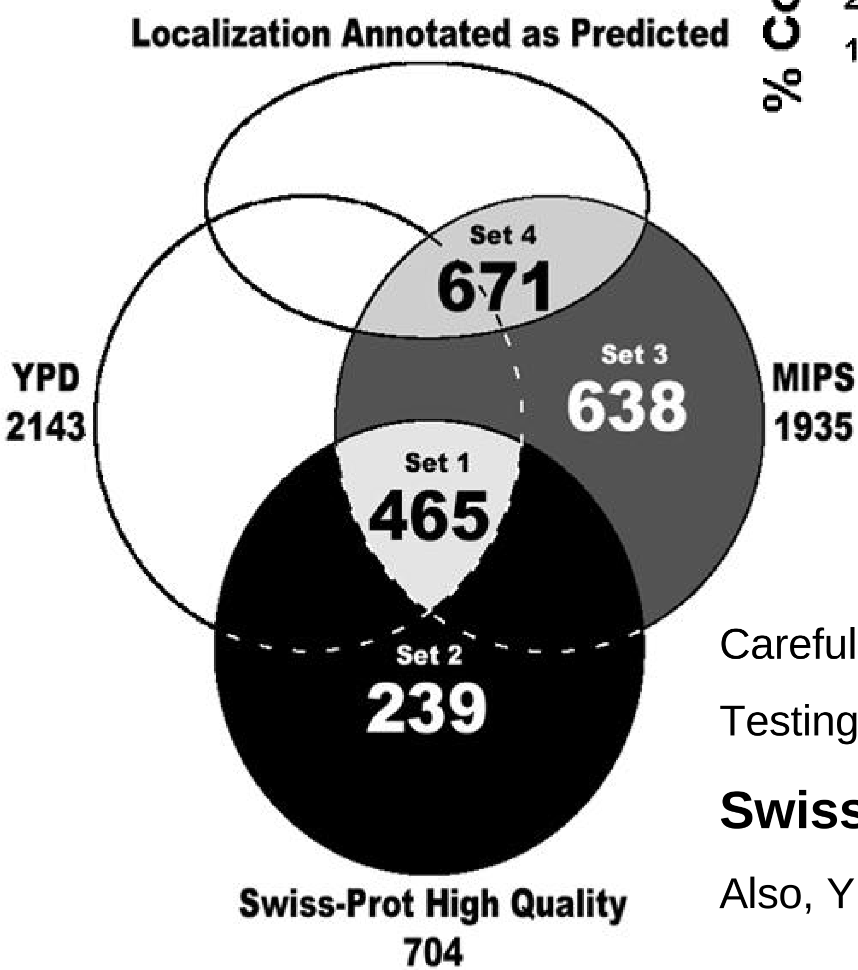
$$C_{MAP} = \arg\max_{C_j \in \{C_1, C_2\}} P(c_j) \prod_{i=1}^{n} P(x_i \mid c_j)$$

Yeast Tables for Localization Prediction

Basics		Predictors: Sequence Features												Predictors: Genomic Features						Response			Bayesian Localization						
			Amino Acid Composition					How many times does the sequence have these motif features?						Abs. expr. Level (mRNA copies / cell)		Cell cycle timecourse				unction		Localization	State Vector giving localization prediction					Collapsed Prediction	
Yeast Gene ID	Sequence	seq. length	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

Results on Testing Data





Individual proteins: 75% with cross-validation

Carefully clean training dataset to **avoid circular logic**

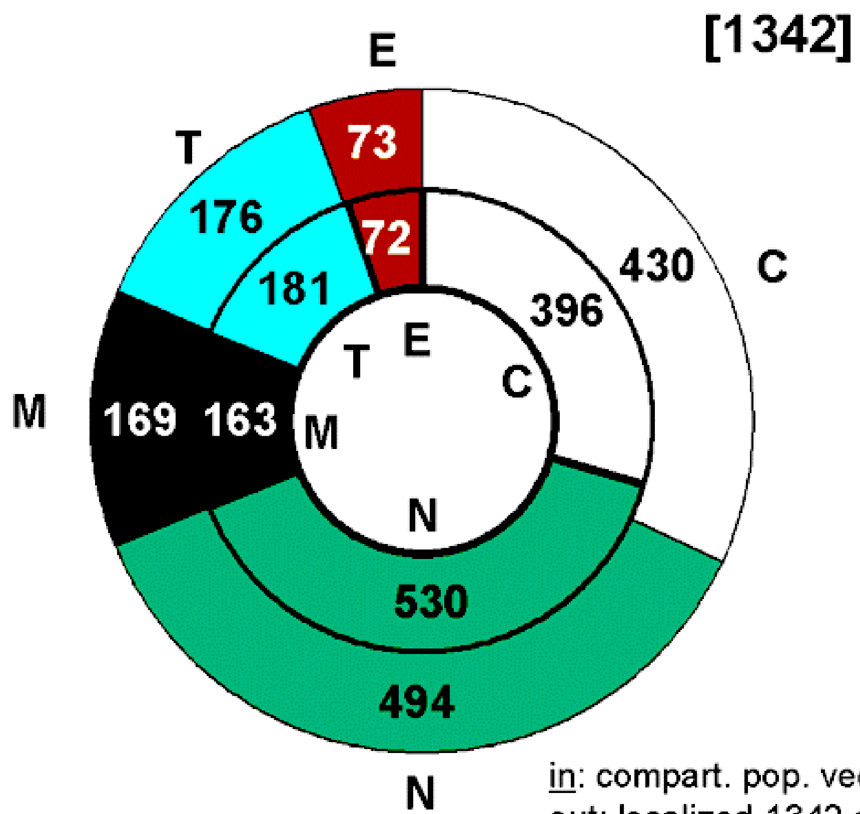
Testing, training data, Priors: ~2000 proteins from

Swiss-Prot Master List

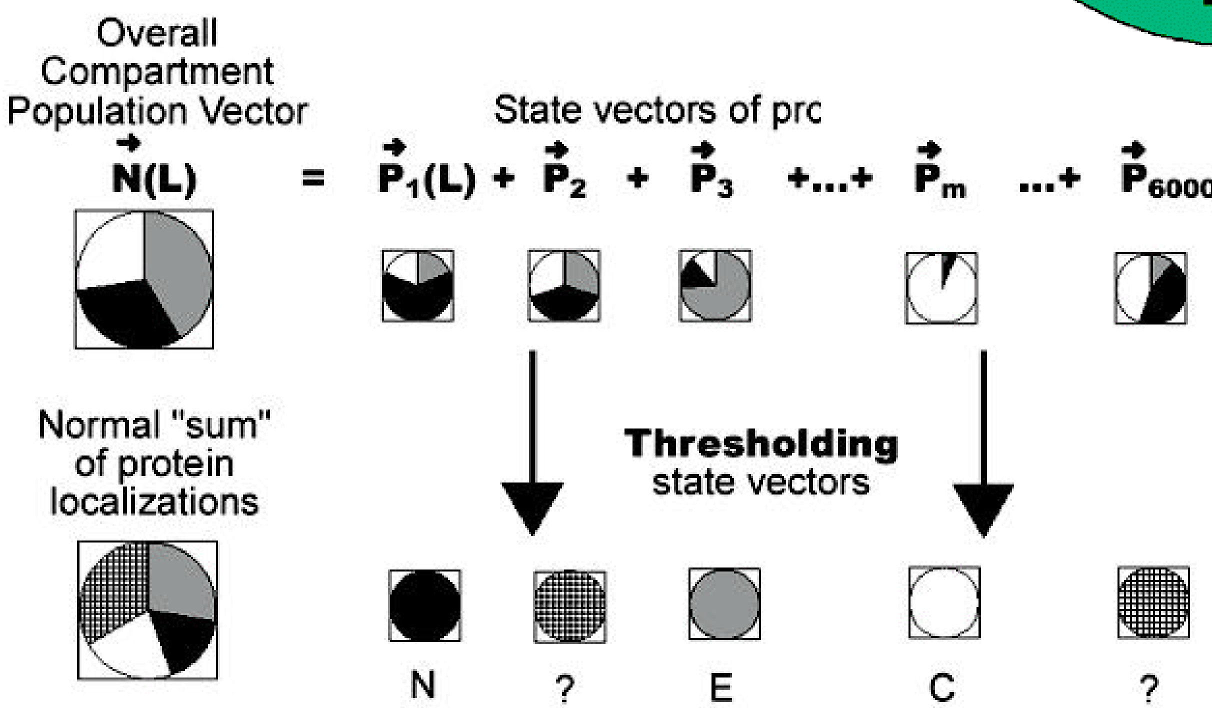
Also, YPD, MIPS, Snyder Lab

Results on Testing Data #2

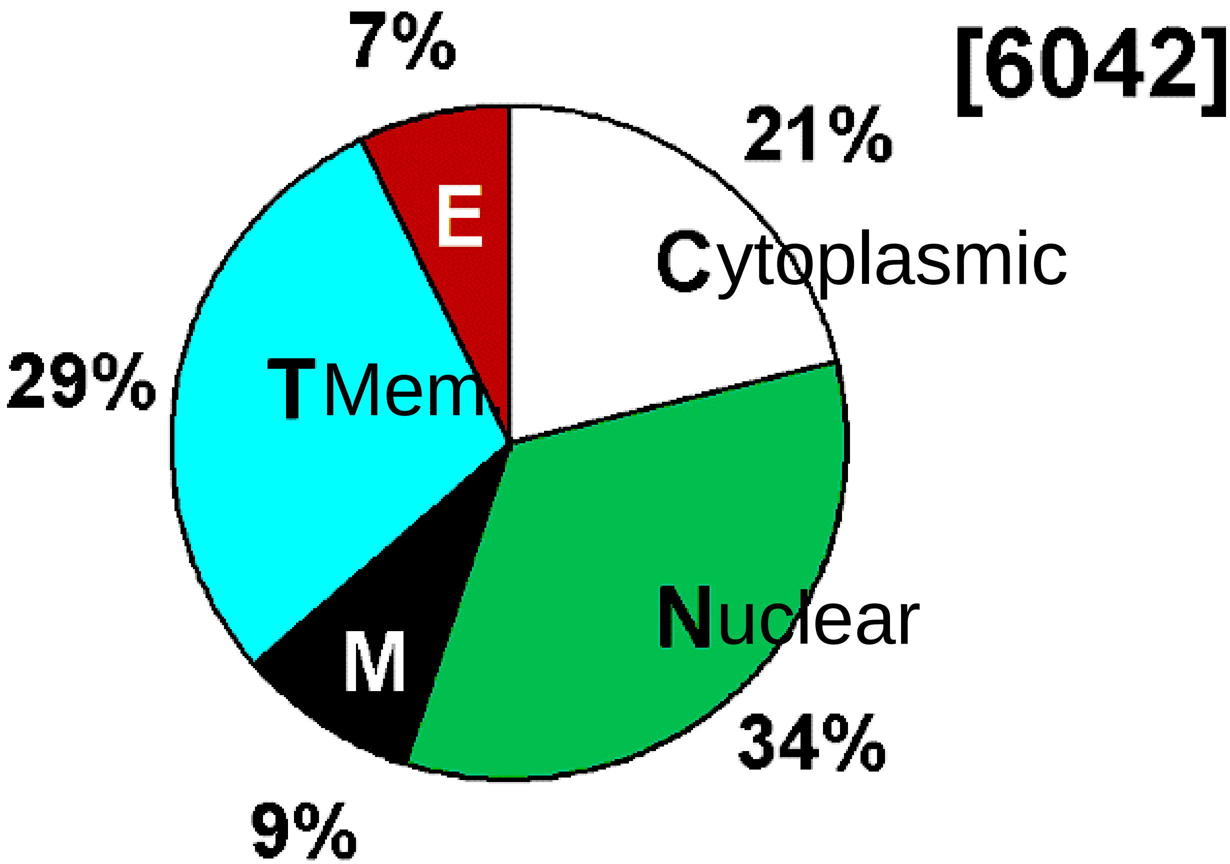
Compartment Populations. Like QM, directly sum state vectors to get population. Gives **96%** pop. similarity.



in: compart. pop. vector $\vec{N}$
out: localized-1342 expected



Extrapolation to Compartment Populations of Whole Yeast Genome: ~4000 predicted + ~2000 known



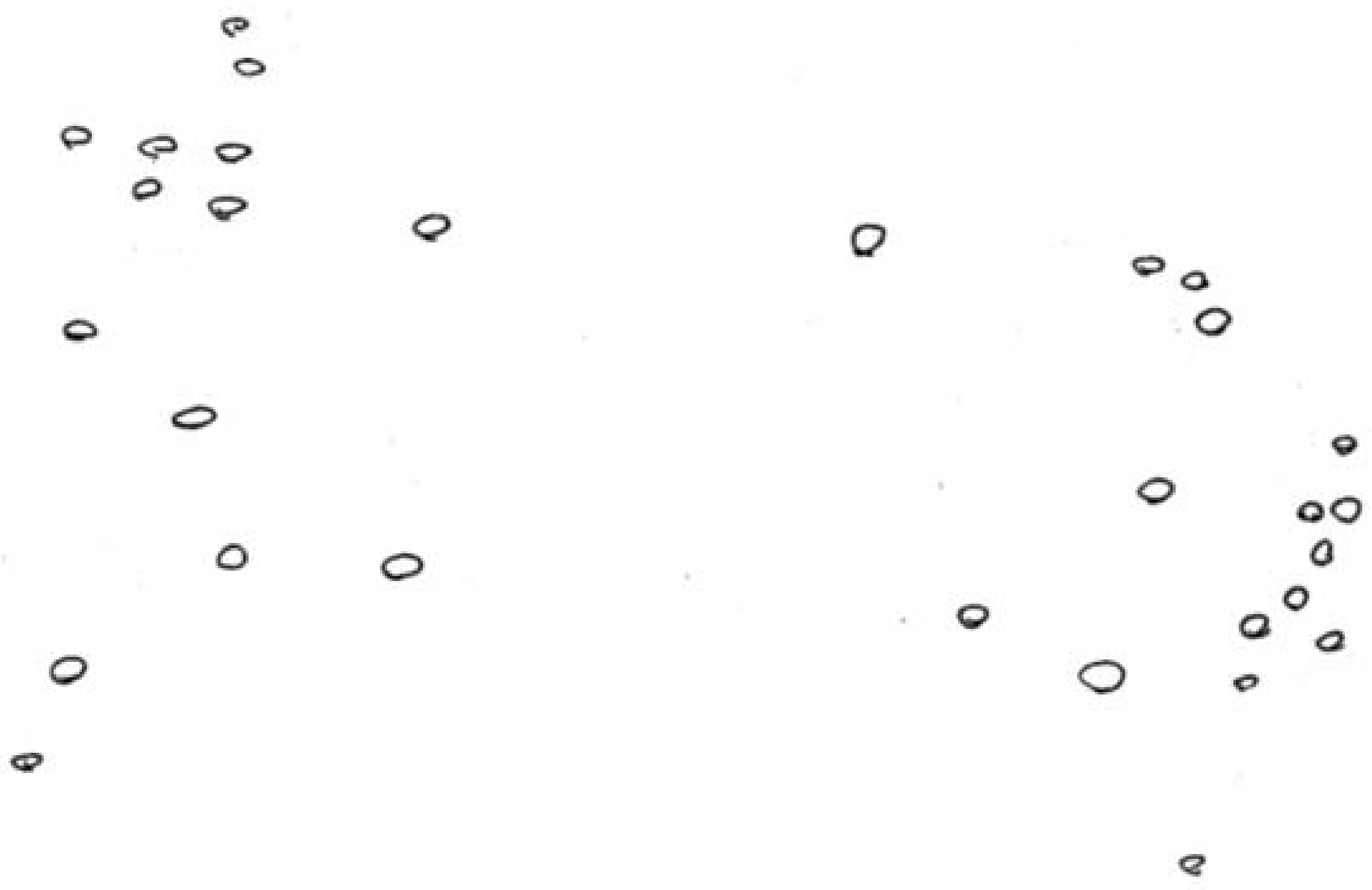
Large-scale Datamining

- Relating Gene Expression to Protein Features and Parts
- Supervised Learning: Discriminants
- Simple Bayesian Approach for Localization Prediction
- Unsupervised Learning: k-means
- Correlation of Expression Data with Function
- Overview of Issues in Datamining
- Overview of Methods of Supervised Learning
- Focus on Decision Trees
- Overview of Methods of Unsupervised Learning
- Cluster Trees, Evolutionary Trees

Typical Predictors and Response for Yeast

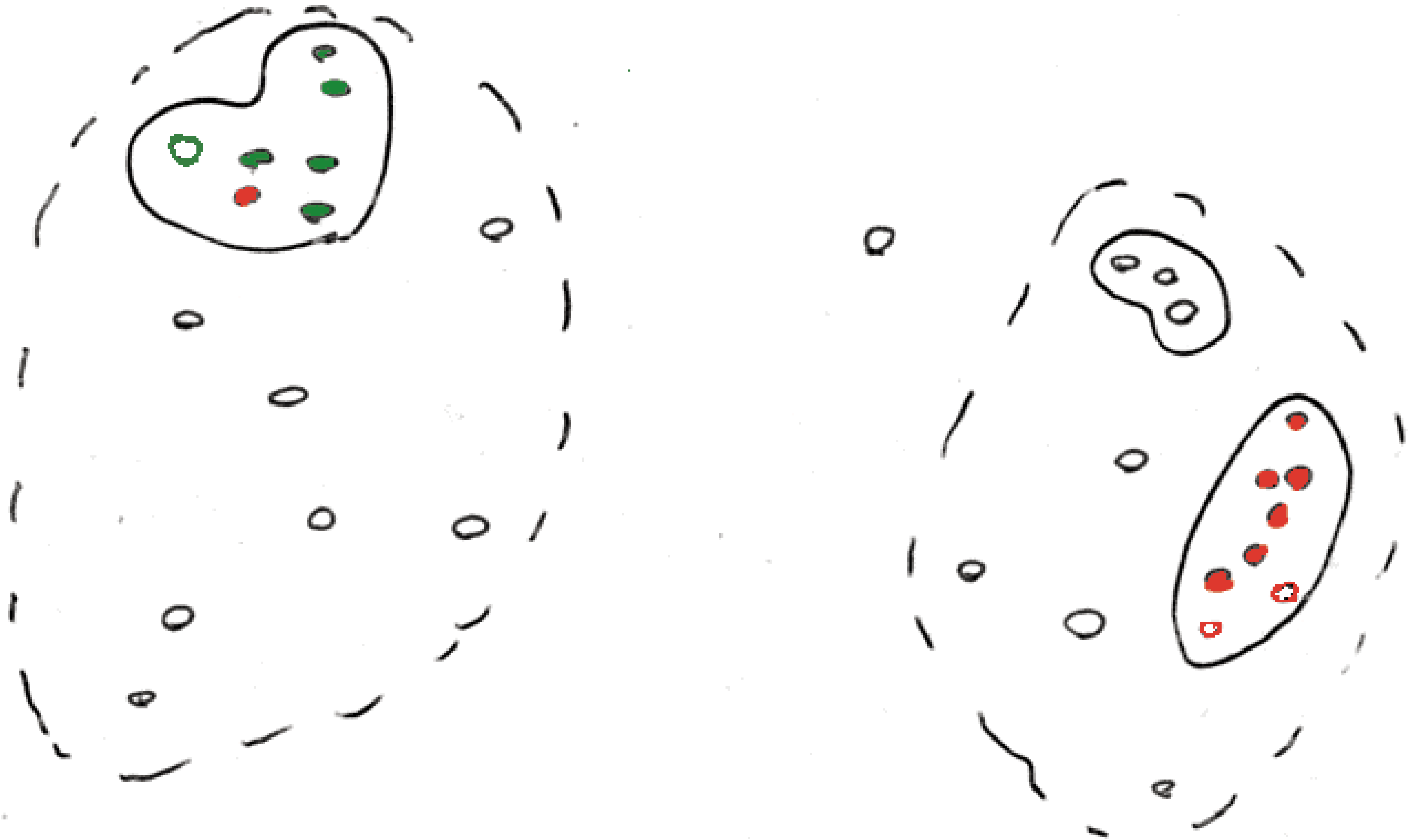
Basics		Predictors																			Response		
		Sequence Features												Genomic Features									
		seq. length	Amino Acid Composition					How many times does the sequence have these motif features?						Abs. expr. Level (mRNA copies / cell)		Prot. Abun-dance	Cell cycle timecourse				Function		Localization
Yeast Gene ID	Sequence		A	C	D	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	MNIFEMLRII	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 initia	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	KMLQFNLRWI	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	RPDFCLEPPY	215	.08	.02	.06	.01	.04	0	0	0	0	0	0	?	0	?	18	12	4	6	01.01.01	0	N
YAL005C	VINTFDGVAI	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	KKAVINGEQI	190	.08	.02	.06	.01	.04	0	0	0	0	1	4	2.2	8	?	15	20	16	17	99	????	????
YAL008W	HPETLVKVKI	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	MEQRITLKDY	493	.08	.02	.06	.02	.04	0	0	0	0	0	1	0.3	?	?	11	6	6	6	30.16	involved in mitochondrial	????
YAL011W	KSFPEVVGKT	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	RTDCYGNVNI	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	GDVEKGKKII	202	.08	.02	.06	.01	.04	0	0	0	0	0	0	1.1	?	?	12	13	9	12	99	????	N
YAL015C	MTPAVTTYKI	399	.08	.02	.06	.01	.04	0	1	0	0	0	0	0.7	0	1	19	18	12	13	11.01;11.04	DNA repair protein	N
YAL016W	KKPLTQEQLI	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

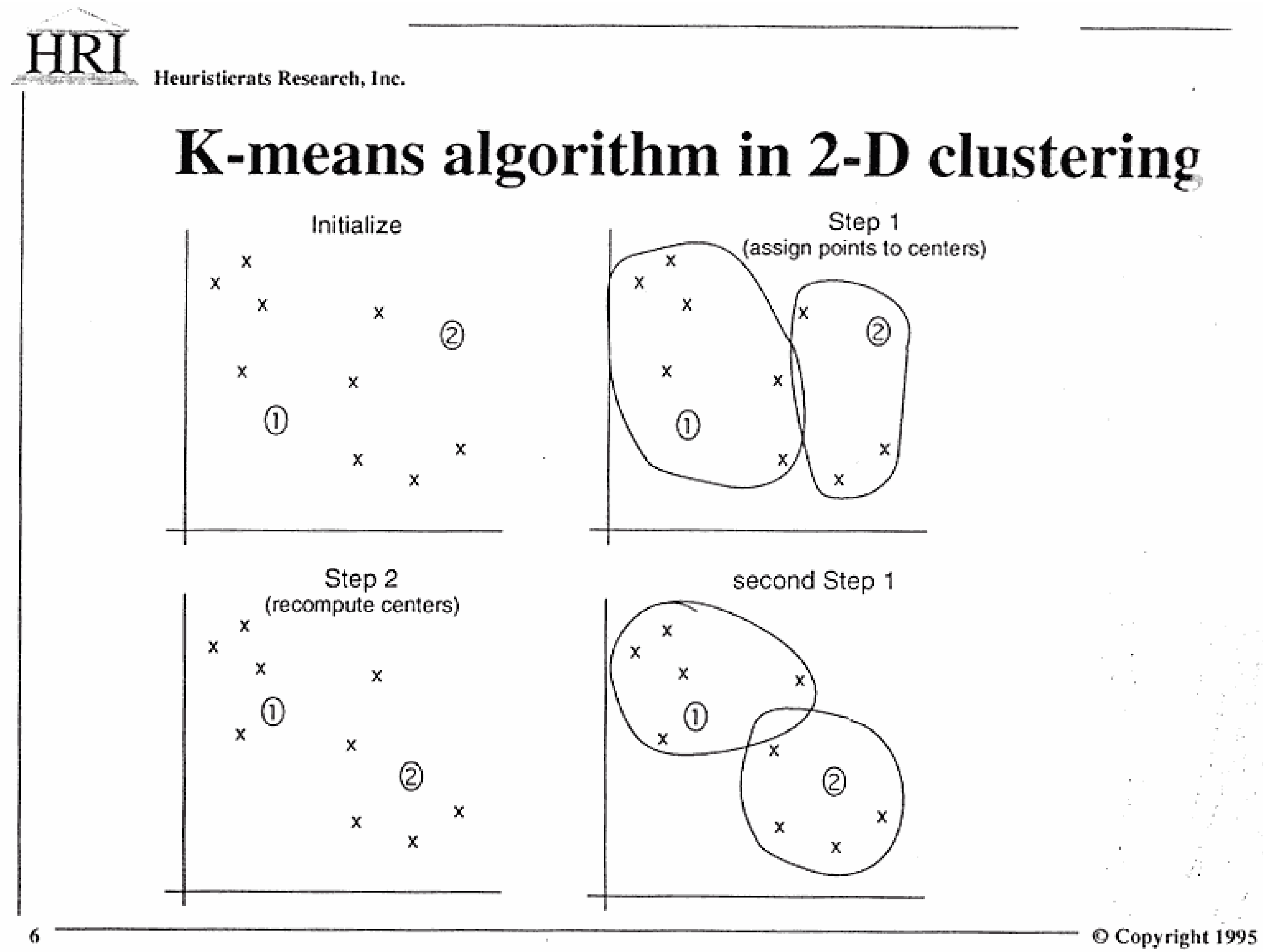


"cluster" predictors

Use clusters to predict Response



K-means



HRI
Heuristicrats Research, Inc.
K-means algorithm in 2-D clustering
Initialize
Step 1
(assign points to centers)
Step 2
(recompute centers)
second Step 1
6
© Copyright 1995

K-means

Top-down vs. Bottom up

Top-down when you know how many subdivisions

k-means as an example of top-down

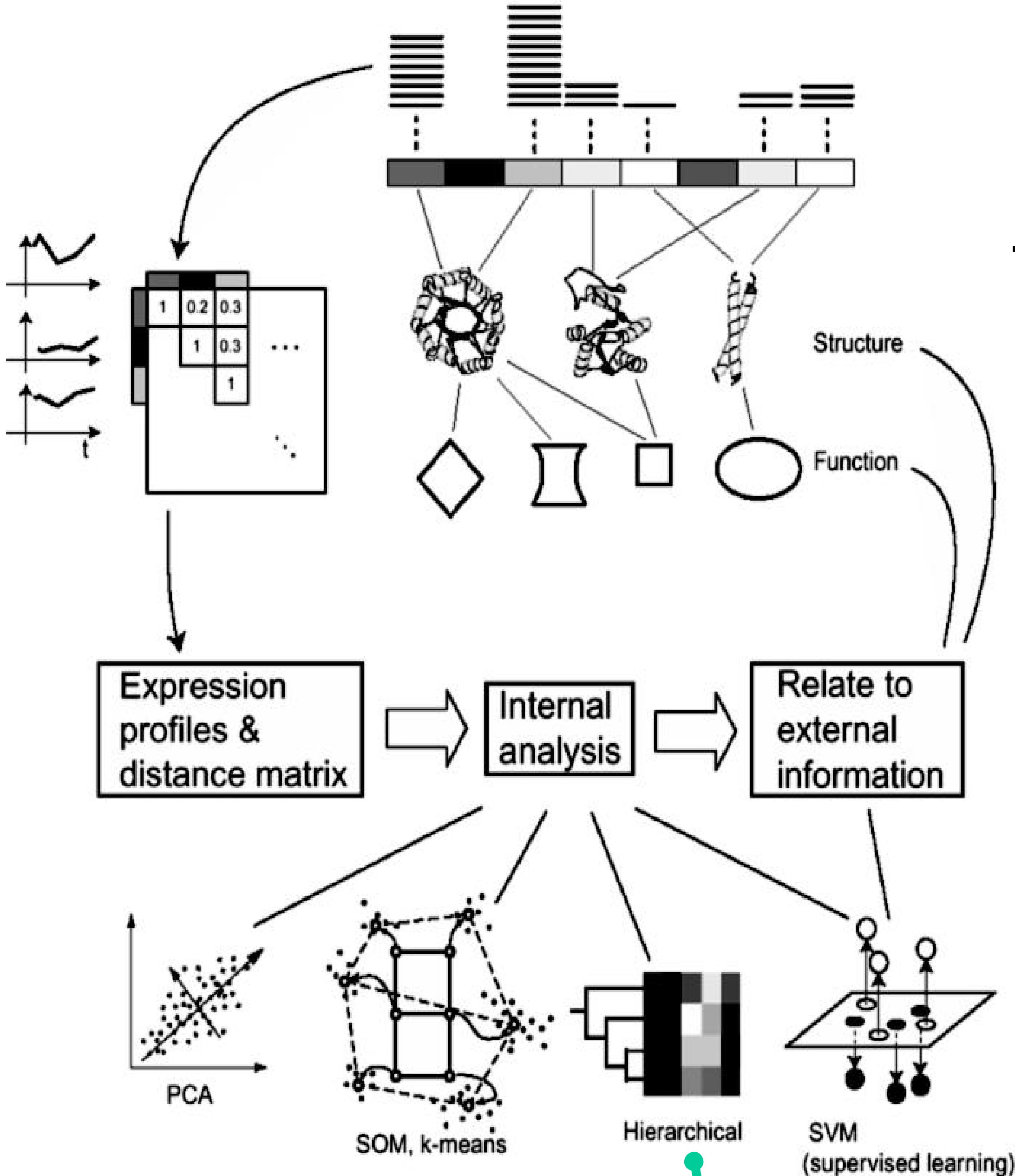
1) Pick ten (i.e. k?) random points as putative cluster centers.
2) Group the points to be clustered by the center to which they are closest.
3) Then take the mean of each group and repeat, with the means now at the cluster center.
4) I suppose you stop when the centers stop moving.

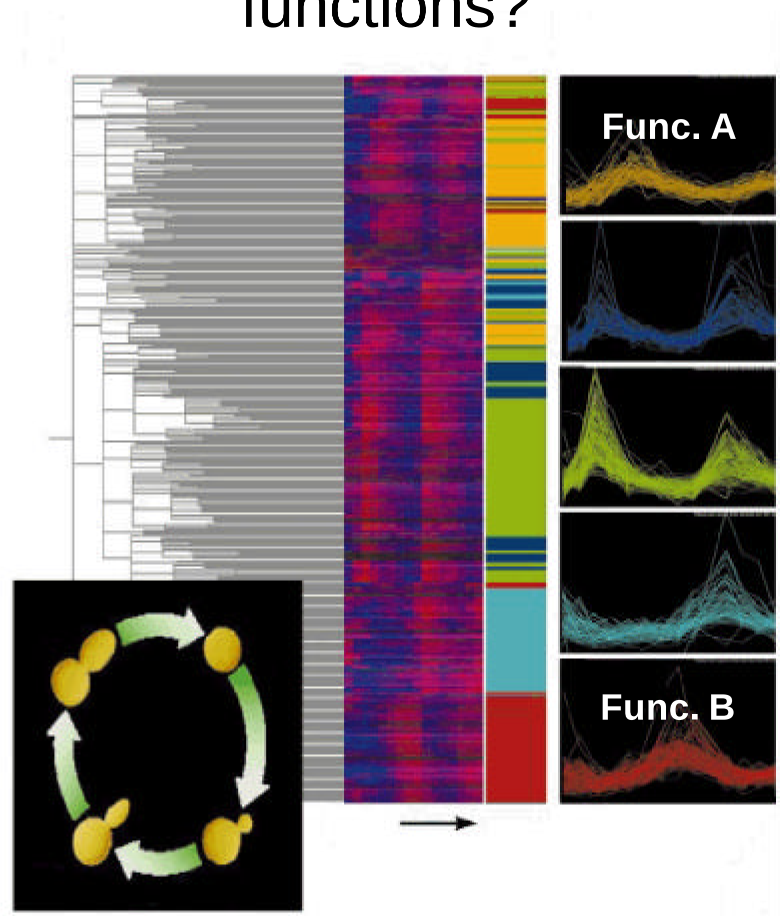
Large-scale Datamining

- Relating Gene Expression to Protein Features and Parts
- Supervised Learning: Discriminants
- Simple Bayesian Approach for Localization Prediction
- Unsupervised Learning: k-means
- Correlation of Expression Data with Function
- Overview of Issues in Datamining
- Overview of Methods of Supervised Learning
- Focus on Decision Trees
- Overview of Methods of Unsupervised Learning
- Cluster Trees, Evolutionary Trees

Do Expression Clusters Relate to Protein Function?

Can they predict functions?





- Clustering of expression profiles
- Grouping functionally related genes together (?)
- **Botstein (Eisen),** Lander, Haussler, and Church groups, Eisenberg

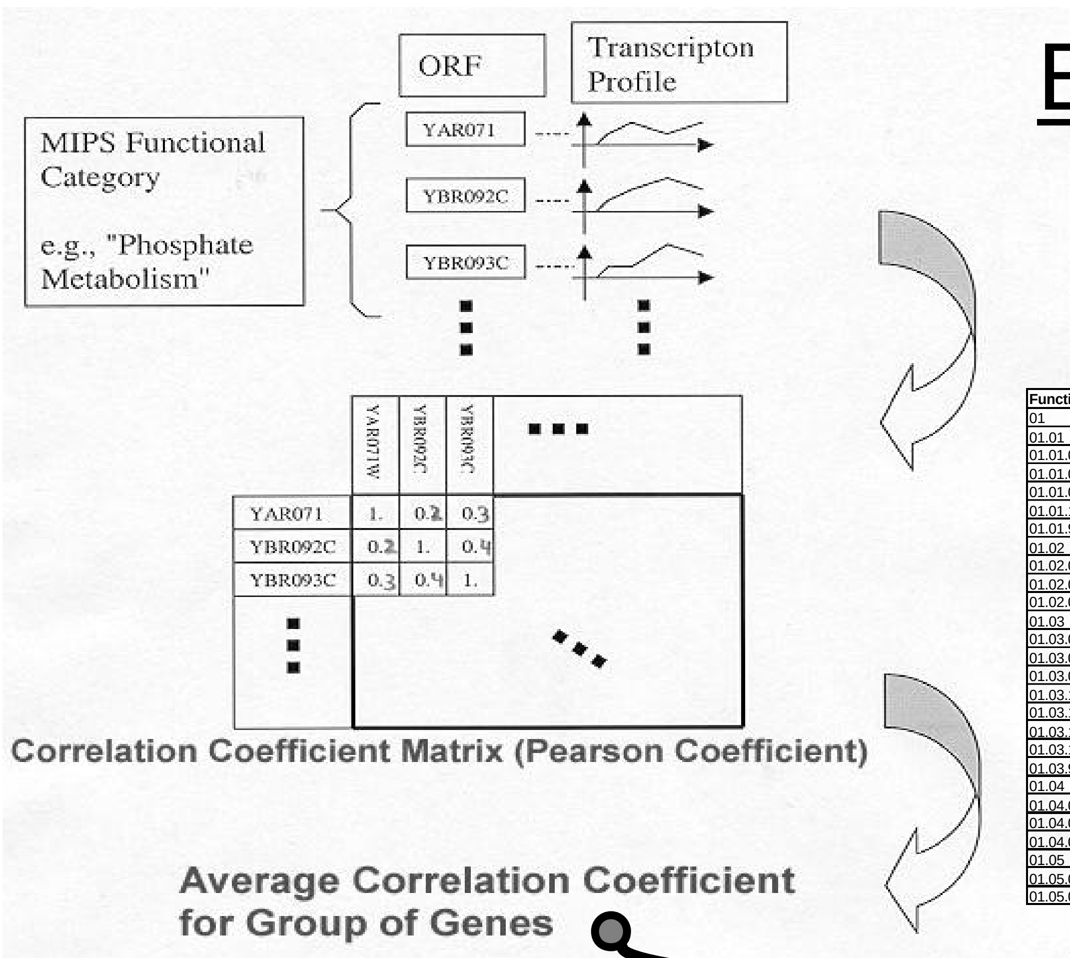
Information for Function Prediction

Basics		Predictors																				Response	
		ence Fea			Genomic Features																		
		seq. length	Amino Acid Composition	How man	Abs. expr. Level (mRNA copies / cell)		Prot. Abun-dance	Cell cycle timecourse														Function	
Yeast Gene ID	Sequence				Gene-Chip expt. from RY Lab	sage tag freq.	(1000 copies /cell)	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	function ID(s) (from MIPS)	function description
YAL001C	M	1160			0.3	0	?	5	3	4	4	5	4	3	5	5	3	5	7	9	4	04.01.01;04.03.0	TFIIIC (transcript
YAL002W	K	1176			0.2	?	?	8	4	2	3	4	3	4	5	5	3	4	4	6	4	06.04;08.13	vacuolar sorting
YAL003W	K	206			19.1	19	23	70	73	91	69	105	52	112	88	64	159	106	104	75	103	05.04;30.03	translation elonga
YAL004W	R	215			?	0	?	18	12	9	5	5	3	6	4	4	3	3	5	5	4	01.01.01	0
YAL005C	V	641			13.4	16	17	39	38	30	13	17	8	11	8	7	8	6	8	8	7	06.01;06.04;08.0	heat shock prote
YAL007C	K	190			2.2	8	?	15	20	32	20	21	19	29	19	16	22	20	26	23	22	99	????
YAL008W	H	198			1.2	?	?	9	6	7	1	3	2	4	2	2	3	3	4	4	3	99	????
YAL009W	P	259			0.6	?	?	6	2	4	3	5	3	5	5	5	3	4	6	6	4	03.10;03.13	meiotic protein
YAL010C	M	493			0.3	?	?	11	6	4	5	6	4	7	8	7	4	5	6	7	5	30.16	involved in mitoc
YAL011W	K	616			0.4	?	?	6	5	4	4	8	5	8	8	6	6	5	6	6	7	30.16;99	protein of unknow
YAL012W	G	393			8.9	4	6.7	29	26	25	27	53	26	43	36	25	28	23	28	31	29	01.01.01;30.03	cystathionine gar
YAL013W	R	362			0.6	?	?	7	9	6	5	14	6	12	14	10	9	9	9	10	9	01.06.10;30.03	regulator of phos
YAL014C	G	202			1.1	?	?	12	13	10	8	10	10	12	13	12	14	11	11	11	10	99	????
YAL015C	M	399			0.7	0	1	19	18	14	10	14	12	17	17	14	13	11	13	16	11	11.01;11.04	DNA repair prote
YAL016W	K	635			3.3	5	?	15	20	20	102	20	20	30	22	18	19	18	20	21	21	03.01;03.04;03.2	ser/thr protein ph
YAL017W	V	1356			0.4	?	?	14	3	3	4	8	5	6	6	5	5	8	9	10	6	99	????
YAL018C	K	325			?	?	?	4	2	2	2	1	1	2	2	2	1	2	1	2	2	99	????
YAL019W	A	1131			0.9	1	?	14	12	14	10	14	10	15	14	11	8	10	11	11	7	11.04;30.10	similarity to helica
YAL020C	M	333			0.7	1	?	6	3	4	3	3	2	3	3	2	2	2	3	3	3	30.04	alpha-tubulin sup
YAL021C	A	837			1.3	0	?	16	14	16	14	17	12	20	16	17	12	15	18	19	13	01.01.04;01.05.0	transcriptional re

Functional category number	Function	Average correlation	# ORFs
01	METABOLISM	0.1001	1005
01.01	amino-acid metabolism	0.1488	199
01.01.01	amino-acid biosynthesis	0.239	114
01.01.04	regulation of amino-acid metabolism	0.23	32

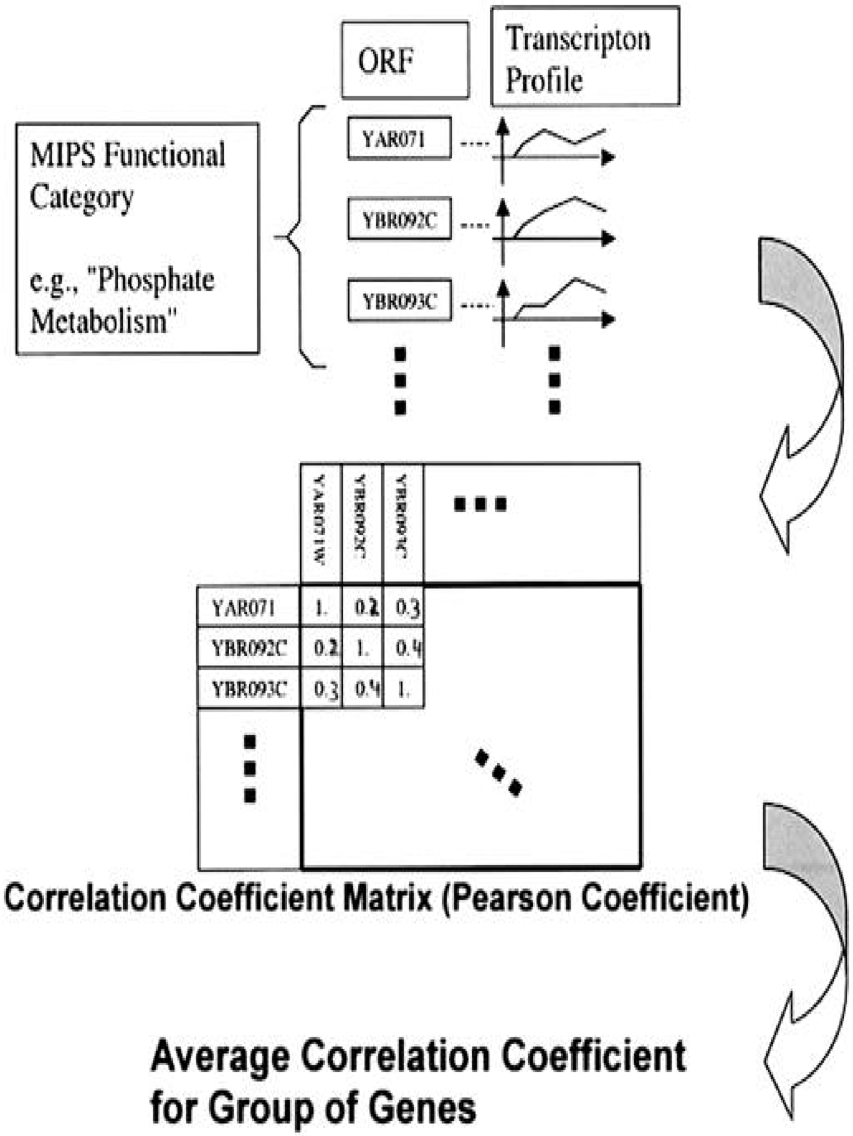
MIPS YFC: 66 bottom classes, 10 top classes
Average correlation of uncharacterized genes is 0.16
Similar to Botstein analysis.

Correlate with Expression Level with Functional Category



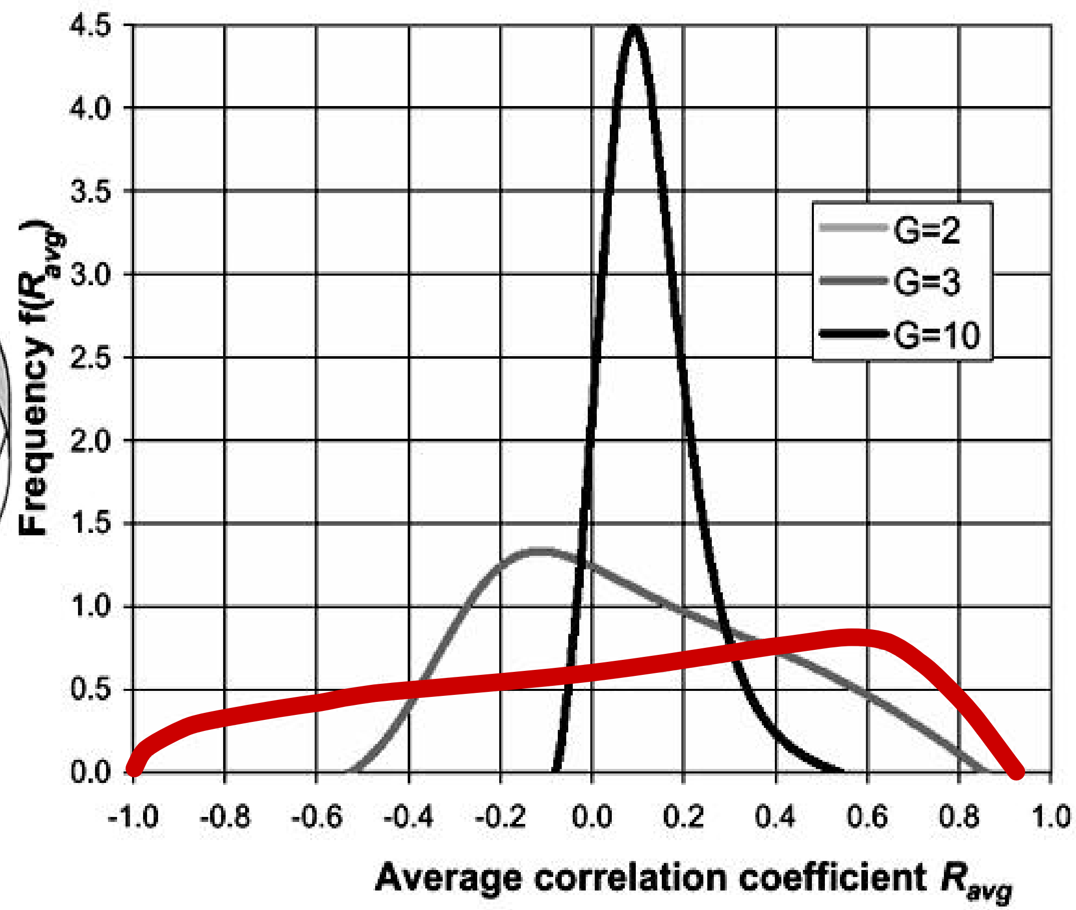
Functional category number	Function	Average correlation	# ORFs
01	METABOLISM	0.1001	1005
01.01	amino-acid metabolism	0.1488	199
01.01.01	amino-acid biosynthesis	0.239	114
01.01.04	regulation of amino-acid metabolism	0.23	32
01.01.07	amino-acid transport	0.1198	23
01.01.10	amino-acid degradation	0.0524	36
01.01.99	other amino-acid metabolism activities	0.2205	4
01.02	nitrogen and sulphur metabolism	0.1869	73
01.02.01	nitrogen and sulphur utilization	0.0726	37
01.02.04	regulation of nitrogen and sulphur utilization	0.3715	28
01.02.07	nitrogen and sulphur transport	0.2829	8
01.03	nucleotide metabolism	0.1708	134
01.03.01	purine-ribonucleotide metabolism	0.3639	42
01.03.04	pyrimidine-ribonucleotide metabolism	0.176	28
01.03.07	deoxyribonucleotide metabolism	0.1095	12
01.03.10	metabolism of cyclic and unusual nucleotides	0.2848	8
01.03.13	regulation of nucleotide metabolism	0.2696	13
01.03.16	polynucleotide degradation	0.2461	[illegible]
01.03.19	nucleotide transport	0.1187	12
01.03.99	other nucleotide-metabolism activities	-0.0328	7
01.04	phosphate metabolism	0.1348	31
01.04.01	phosphate utilization	[illegible]	13
01.04.04	regulation of phosphate utilization	[illegible]	8
01.04.07	phosphate transport	0.0724	10
01.05	carbohydrate metabolism	0.0779	409
01.05.01	carbohydrate utilization	0.075	256
01.05.04	regulation of carbohydrate utilization	0.1174	120

Distributions of Gene Expression Correlations, for All Possible Gene Groupings

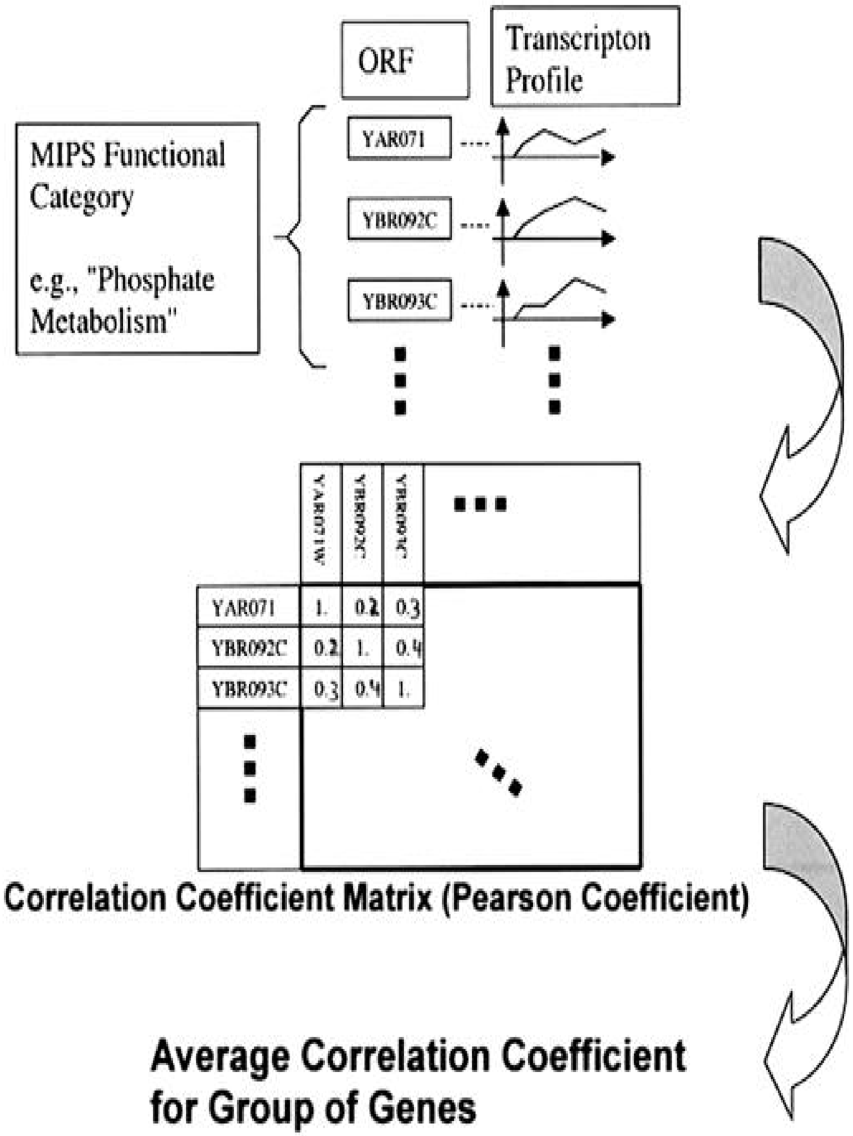


Sample for Diauxic shift Expt. (Brown),

Ex. $R_{avg,G=3}$ =
[R(gene-1,gene-3) + R(gene-1,gene-4)
+
R(gene-5,gene-7)] / 3

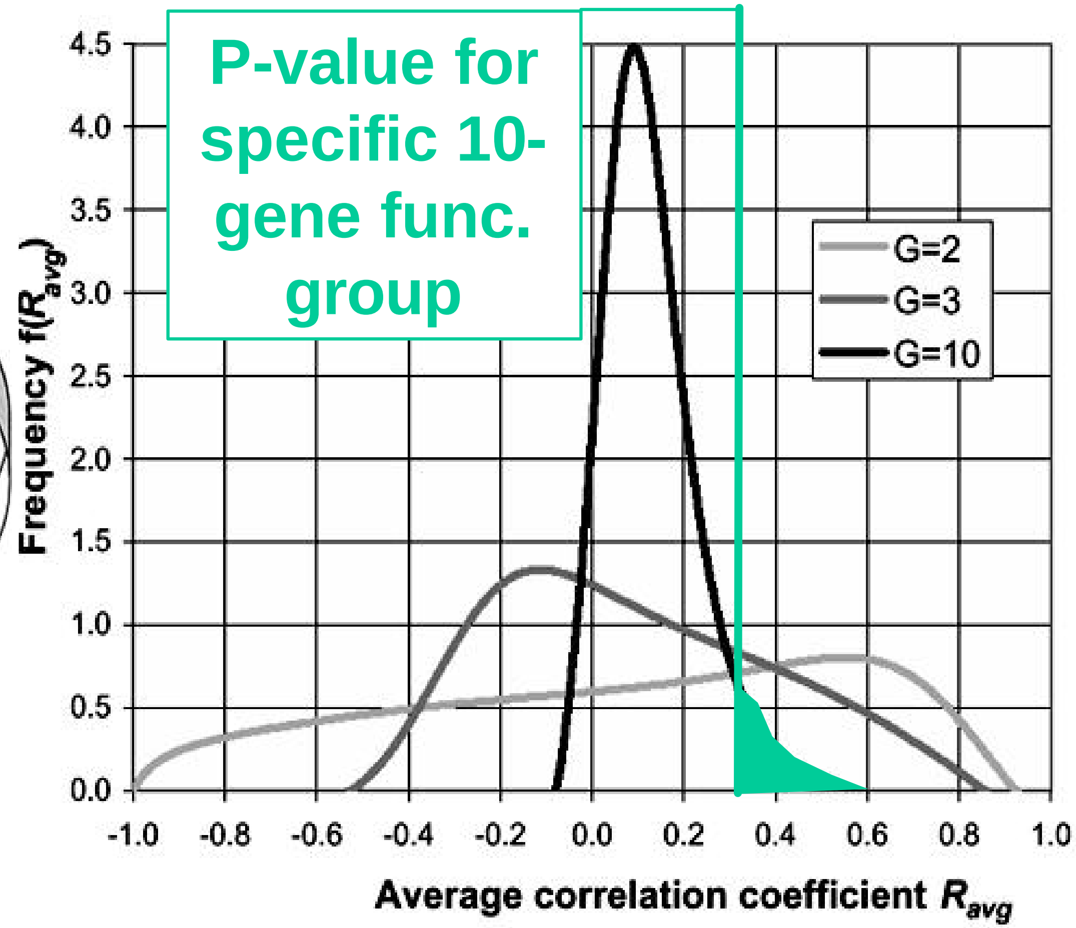


Distributions of Gene Expression Correlations, for All Possible Gene Groupings 2



Sample for Diauxic shift Expt. (Brown),

Ex. $R_{avg,G=3}$ =
[R(gene-1,gene-3) + R(gene-1,gene-4)
+
R(gene-5,gene-7)] / 3



MIPS category	Cell Cycle (CDC28)	Cell cycle (CDC15)	Diauxic shift	Spo-rulation
Cell growth, division & DNA syn.	>4	>4	>4	>4
Protein synthesis	>4	>4	>4	>4
Transcription	>4	>4	>4	1.6
Cellular organization	>4	>4	0.3	0.3
Energy	>4	>4	0.1	0.9
Cell rescue, defense, death	>4	>4	0	0
Intracellular transport	>4	>4	0	0
Ionic homeostasis	>4	>4	0	0.8
Metabolism	>4	>4	0	0
Transport facilitation	>4	>4	0	0
Signal transduction	2.5	1.6	0.1	0.6
Unclassifield	2.3	>4	0	0
Cellular biogenesis	2.0	>4	0.4	0.2
Protein destination	0.3	>4	0.2	0.6
Retrotransposon & plasmid	0	2.8	1.9	1.0

	Fraction of significant groups				Total # groups
	CDC28	CDC15	Diauxic Shift	Sporulation	
MIPS 1	63%	81%	19%	13%	16
MIPS 2	50%	63%	17%	13%	102
MIPS 3	23%	33%	5%	4%	73
"Energy" (2nd level)	40%	60%	20%	0%	10
SOM	93%	-	-	-	30
Hierarch. Clustering	80%				25

MIPS category	Cell Cycle (CDC28)	Cell cycle (CDC15)	Diauxic shift	Spo-rulation
Respiration	>4	>4	>4	3.4
TCA pathway	>4	>4	>4	0.6
Glycogen, trehalose metabolism	>4	>4	1.2	0.7
Glycolysis	>4	>4	0.9	2.1
Gluconeogenesis	3.7	>4	0.1	1.7
Glyoxylate cycle	1.6	0.7	3.0	2.3
Pentose-phosphate pathway	1.5	0.8	0	0.6
Fermentation	1.3	>4	0	2.2
Other energy generation activities	0.7	0.1	0.1	0.2
Beta-oxidation of fatty acids	0.5	0.4	0.4	0.2

Correlation:

Always Significant

Sometimes Significant (depends on expt.)

Never Significant

Based on Distributions, Correlation of Established Functional Categories, Computer Clusterings

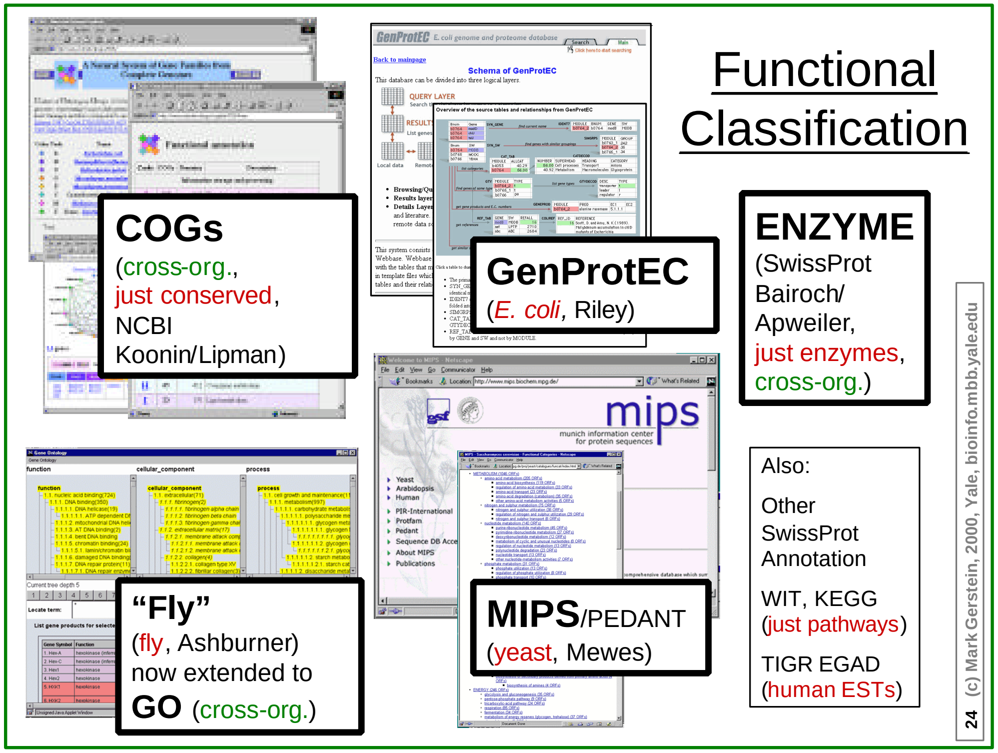
Can we define FUNCTION well enough to relate to expression?

Problems defining function:

Multi-functionality: 2 functions/protein (also 2 proteins/function)

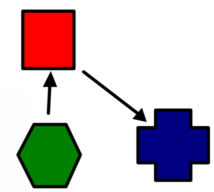
Conflating of Roles: molecular action, cellular role, phenotypic manifestation.

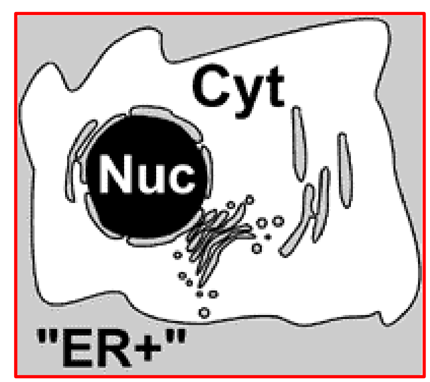
Non-systematic Terminology:
'suppressor-of-white-apricot' & 'darkener-of-apricot'



vs.

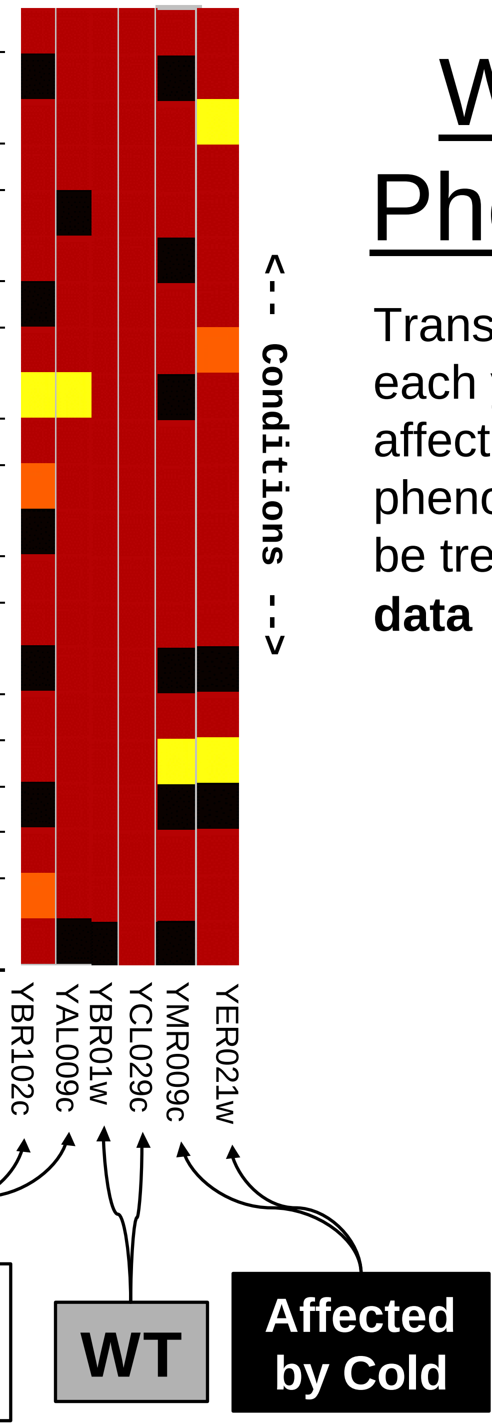
Fold, Localization, Interactions & Regulation are attributes of proteins that are much more clearly defined





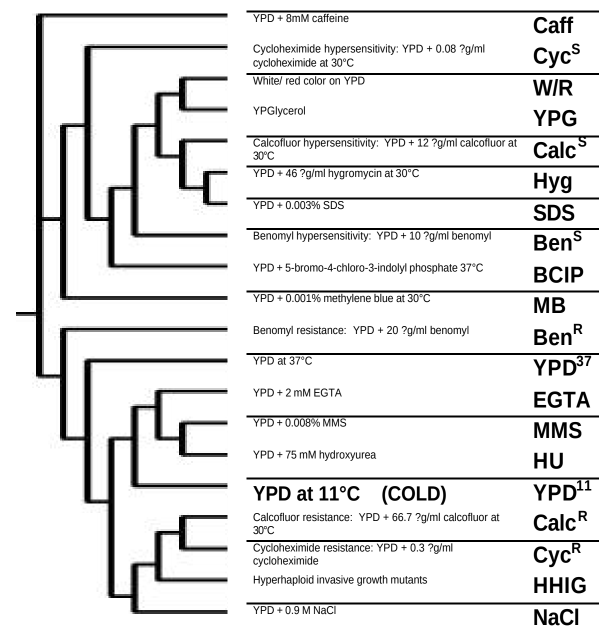
Whole Genome Phenotype Profiles

YPD + 8mM caffeine	Caff
Cycloheximide hypersensitivity: YPD + 0.08 ?g/ml cycloheximide at 30°C	Cyc^{S}
White/ red color on YPD	W/R
YPGlycerol	YPG
Calcofluor hypersensitivity: YPD + 12 ?g/ml calcofluor at 30°C	Calc^{S}
YPD + 46 ?g/ml hygromycin at 30°C	Hyg
YPD + 0.003% SDS	SDS
Benomyl hypersensitivity: YPD + 10 ?g/ml benomyl	Ben^{S}
YPD + 5-bromo-4-chloro-3-indolyl phosphate 37°C	BCIP
YPD + 0.001% methylene blue at 30°C	MB
Benomyl resistance: YPD + 20 ?g/ml benomyl	Ben^{R}
YPD at 37°C	YPD^{37}
YPD + 2 mM EGTA	EGTA
YPD + 0.008% MMS	MMS
YPD + 75 mM hydroxyurea	HU
YPD at 11°C (COLD)	**YPD^{11}**
Calcofluor resistance: YPD + 66.7 ?g/ml calcofluor at 30°C	Calc^{R}
Cycloheximide resistance: YPD + 0.3 ?g/ml cycloheximide	Cyc^{R}
Hyperhaploid invasive growth mutants	HHIG
YPD + 0.9 M NaCl	NaCl



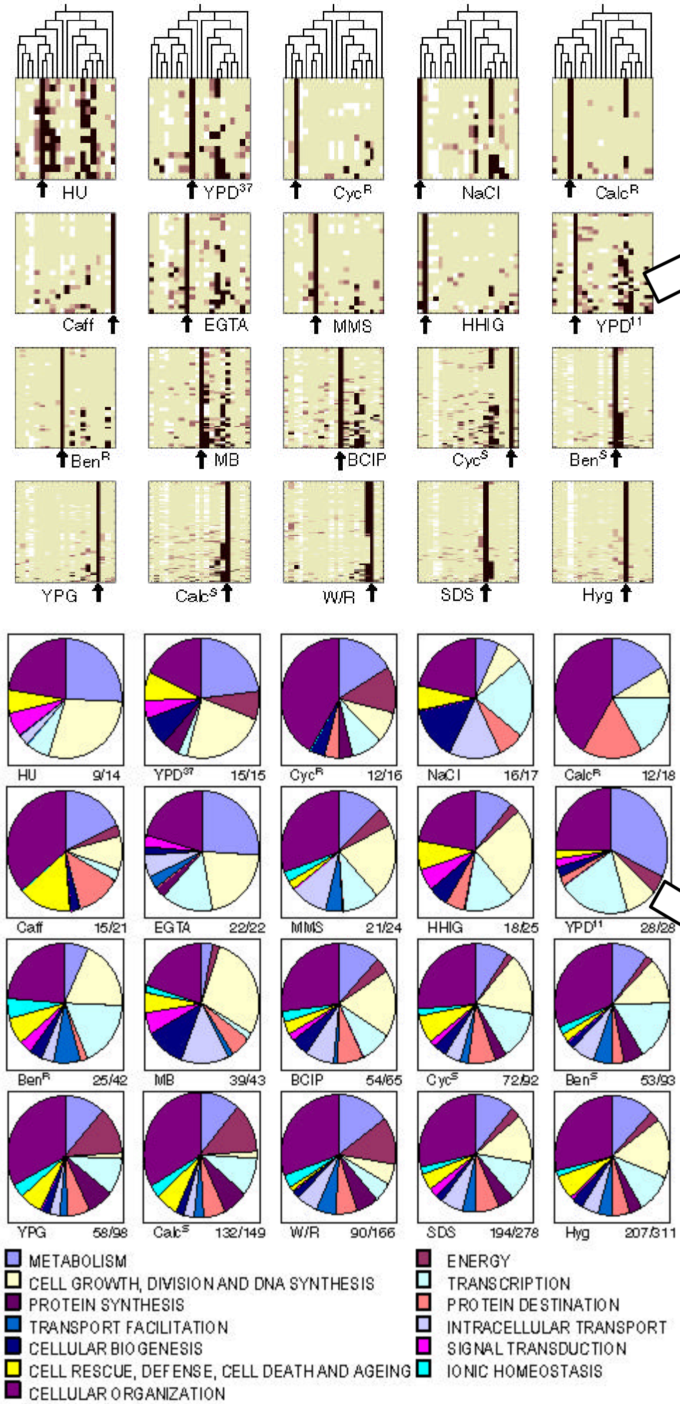
M Snyder

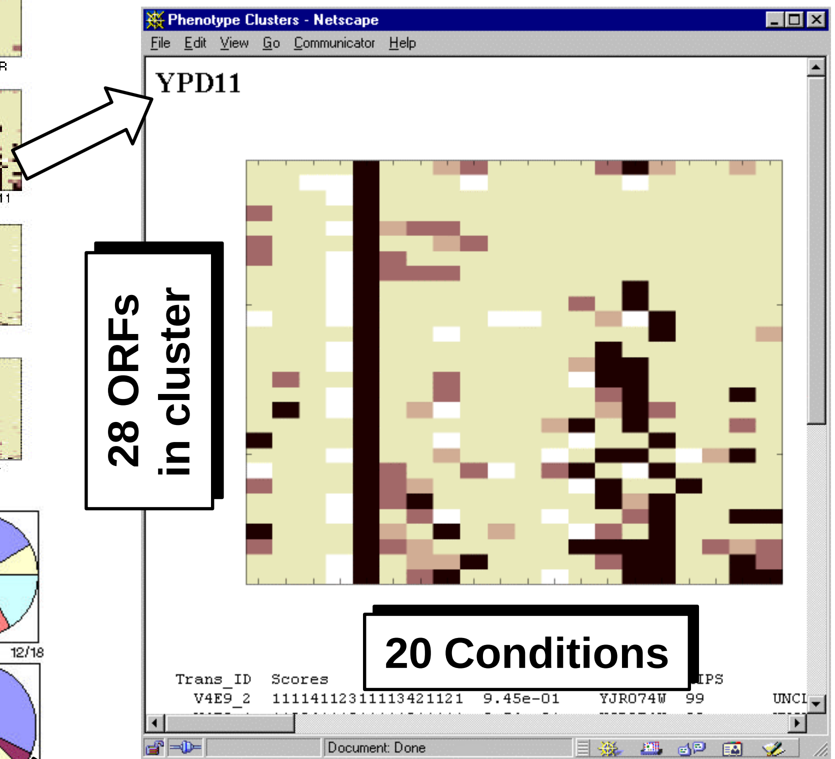
Transposon insertions into (almost) each yeast gene to see how yeast is affected in 20 conditions. Generates a phenotype pattern vector, which can be treated **similarly to expression data**



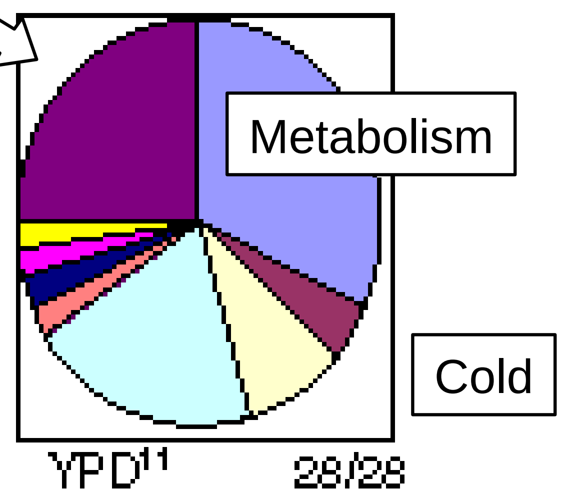
Clustering Conditions

Phenotype ORF Clustering





k-means clustering of ORFs based on “phenotype patterns,” cross-ref. to MIPs Functional Classes



Cluster showing cold phenotype (containing genes most necessary in cold) is enriched in metabolic functions

Large-scale Datamining

- Relating Gene Expression to Protein Features and Parts
- Supervised Learning: Discriminants
- Simple Bayesian Approach for Localization Prediction
- Unsupervised Learning: k-means
- Correlation of Expression Data with Function
- Overview of Issues in Datamining
- Overview of Methods of Supervised Learning
- Focus on Decision Trees
- Overview of Methods of Unsupervised Learning
- Cluster Trees, Evolutionary Trees

The remainder of this packet is purely **optional** material giving an overview of datamining methods

(some of this was adapted from Y Kluger)

Overview of Machine learning methods

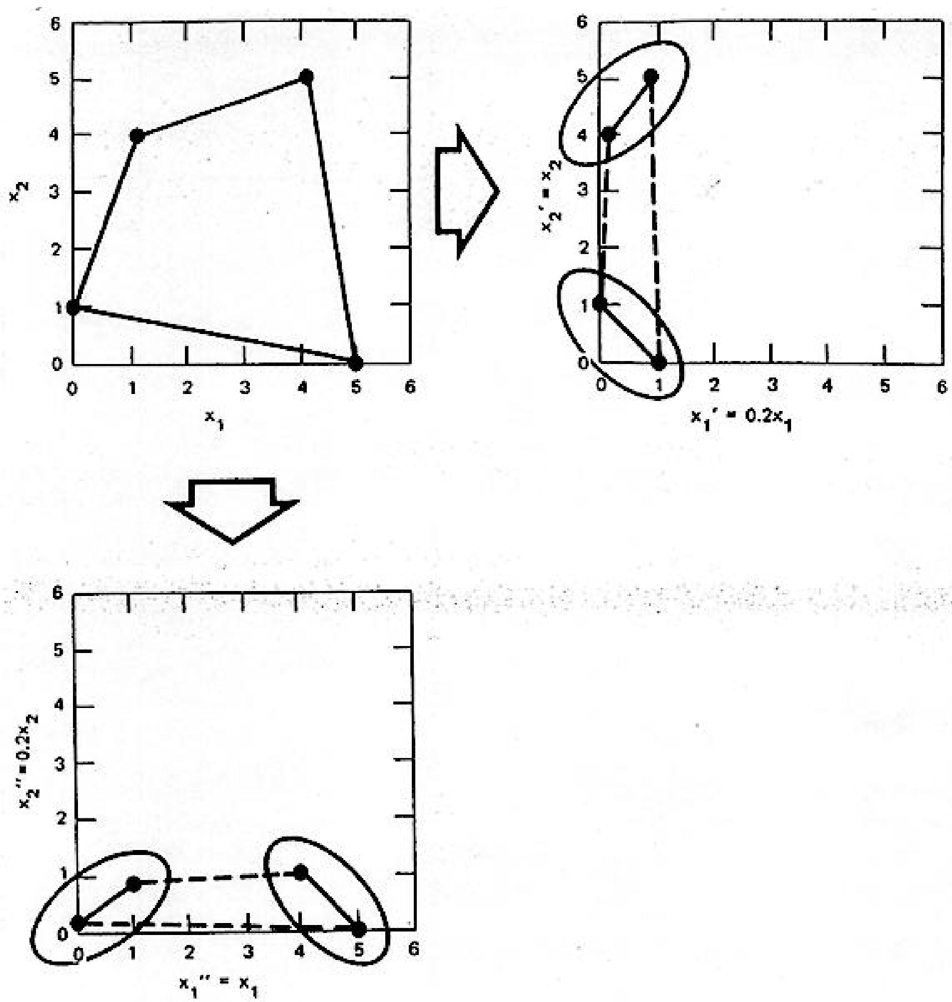
SUPERVISED

- Fisher discriminant analysis
- Statistical disc. analysis
- Logistic discrimination
- Nonlinear discrimination
- Support vector machines
- Decision trees
- Neural networks
- K nearest neighbors
- Bayesian networks

UNSUPERVISED

- K means
- Hierarchical
- Self Organizing Maps
- Spectral methods
  SVD, PCA, bi-clustering, normalized cuts
- Expectation Maximization
- Bayesian Network
- Multiscale analysis
- Ising-like models

Effect of Scaling



(adapted from ref?)

Data preparation and cleansing

- Feature manipulations: scaling, normalization, standardization, or numeric ←→discrete
- Strategy of handling missing values
- Choosing relevant discriminating features:

  expert, algorithms such as backward elimination and forward selection and/or by principal component analysis
- Removing outliers by visual inspection (could be too hard when the number of features is large) or by selecting them if several learning algorithms failed to classify them correctly and finally by inspecting these cases manually.

Get to know the parameters

of the various learning algorithms such as the k value in k-nearest-neighbors, pruning parameters in decision trees, the polynomial power and parameters related to minimization of error on the training set in SVM classification etc.

Choice of learning algorithms

- suitability to data size, data type (numeric, symbolic etc.) and data quality (noisy, inaccurate, missing values, etc.)

- The choice of a learning scheme also involves computational considerations such as time memory and operational simplicity
- degree of desired interpretability or output representation (decision trees are easy to communicate as opposed to neural networks.)

Assess performance of the learning algorithms on test sets

- cross validation, bootstrap, confusion matrix, various loss and cost functions and ROC (receiver operating characteristic) curves. Then, compare these algorithms by applying for instance statistical confidence bound tests on the algorithms' error rate distributions, or inspect the ROC curves obtained from cross validated learning schemes evaluations

Optional: not needed for Quiz
Optional: not needed for Quiz

(adapted from Y Kluger)

ROC Curve

- In our two-class classification task (soluble/insoluble), we can sort the proteins of a test set in descending order according to the probability that they are soluble as predicted by the learning model.
- ROC curve is constructed by going along the ranked list one step at a time and counting how many TP, FP, TN, and FN were accumulated up to that step.
- By changing the parameter of location in the list sorted in probability order, we can inspect at each point along the list the TP rate (TP/(TP+FN)) as a function of the FP rate (FP/(FP+TN)) up to that point. A worthy learning tool must yield a curve for which the TP-rate>FP-rate as opposed to the curve TP-rate=FP-rate generated by random (not-ranked) samples of different sizes taken from the test set (Note that at the curve's end points where none or all elements of the sorted list are taken into account TP-rate = FP-rate).

Optional: not needed for Quiz

(adapted from Y Kluger)

- The steeper the step-like (concave) curve near to the origin the better because the larger the coverage with high TP rate and low FP rate.
- A ROC curve based on one test set is jagged and in order to get a smoother and more reliable curve, one performs an N-fold cross validation. This is done by averaging over the TP-rates obtained from the N test datasets at each fixed point along the FP-rate axis (x axis). These fixed points along the FP-rate are determined by covering enough of the highest-ranked instances in the test datasets. The preferable learning tool is selected by taking the one with the lower FP rate at the desired coverage level of TP.
- Other measures used to evaluate false positives versus false negative tradeoff along the ranked list are Lift charts in which the TP are displayed against the subset size (TP+FP/(TP+FP+TN+FN)) and recall-precision curves where the TP rate (recall) is displayed against the precision (TP/(TP+FP+TN+FN)).

(adapted from Y Kluger)

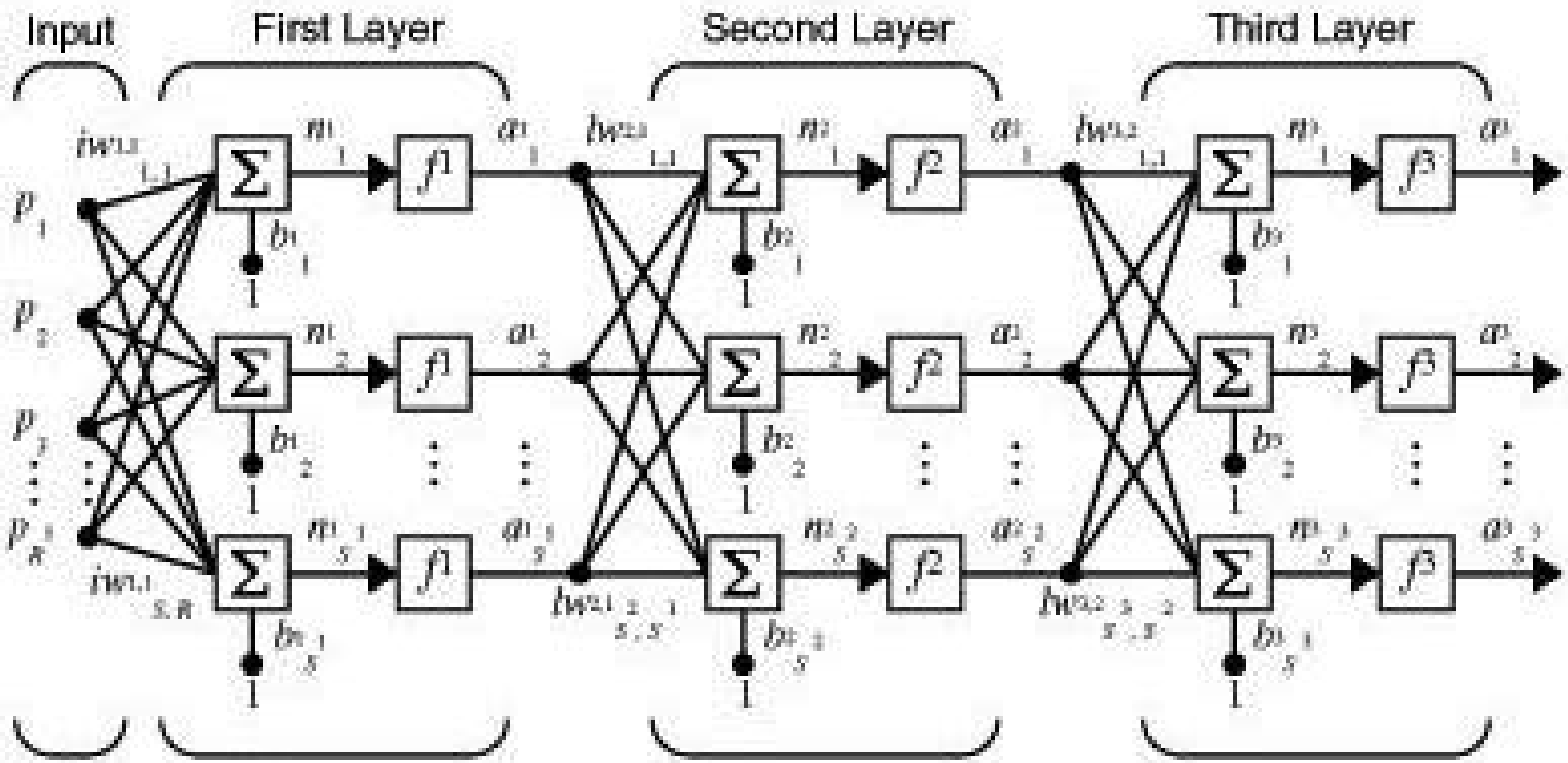
Optional: not needed for Quiz

Large-scale Datamining

- Relating Gene Expression to Protein Features and Parts
- Supervised Learning: Discriminants
- Simple Bayesian Approach for Localization Prediction
- Unsupervised Learning: k-means
- Correlation of Expression Data with Function
- Overview of Issues in Datamining
- Overview of Methods of Supervised Learning
- Focus on Decision Trees
- Overview of Methods of Unsupervised Learning
- Cluster Trees, Evolutionary Trees

Supervised Learners

ANN



$$a^1 = f^1(IW^{1,1}p + b^1)$$

$$a^2 = f^2(LW^{2,1}a^1 + b^2)$$

$$a^3 = f^3(LW^{3,2}a^2 + b^3)$$

$$a^3 = f^3(LW^{3,2}f^2(LW^{2,1}f^1(IW^{1,1}p + b^1) + b^2) + b^3)$$

Support Vector Machine (SVM)

- A sophisticated discriminant method that is capable of handling nonlinear class boundaries by transforming the original feature space to a new space, in which the non-linear class boundary is a hyperplane, and the new features are non-linear combinations of the original features.
- The number of features in the new space is larger than the number of the original features. Support vector machines overcome the shortcomings mentioned above: over-fitting (too many parameters to fit) and complexity (computational time for linear discriminant analysis is cubic in number of features.)
- If we assume that the classes of the dataset are linearly separable in the new space, their corresponding convex hulls (the tightest enclosing convex polygons connecting the data points of each class) do not overlap.

Optional: not needed for Quiz

(adapted from Y Kluger)

SVM cont.

- The discrimination task is then to find the maximum margin hyperplane defined as the hyperplane that is maximally distant from both convex hulls. This hyperplane also intersects the shortest line connecting such convex hulls midway. We call the cases that are closest to the maximum margin hyperplane support vectors. The minimum number of support vectors from each class is one, and they uniquely define the maximum margin hyperplane. A standard constrained quadratic optimization scheme is suitable for finding the support vectors and the parameters that determine the maximum margin hyperplane. Overfitting is unlikely because the maximum margin hyperplane is quite stable. This is because such hyperplane is determined by a small number of support vectors in a global fashion.

Optional: not needed for Quiz

(adapted from Y Kluger)

A solution for the complexity problem

- separate hyperplane of the standard linear discriminant analysis in terms of a weighted sum of an inner product of support vectors, with the feature vector x representing the example to be classified. This works because the standard linear discriminant problem of finding the solution (w*,b*) that minimizes ||w|| subject to

$$C_l(\vec{w}\cdot\vec{x}_l + b) \geq 1$$

can be written as

$$\vec{w}^* = \sum_l a_l C_l \vec{x}$$

where all the auxiliary variables alpha vanish excluding the samples that are the support vectors. Thus a new example x can be classified by the linear decision function

$$sign(\sum_l a_l C_l \vec{x}_l \cdot \vec{x} + b^*)$$

(adapted from Y Kluger)

Optional: not needed for Quiz

SVM4

- Substitution of the inner product in the sum by some power of this product is directly mapped to a polynomial nonlinear class boundary. Other functions of the inner product can be used for more complicated class boundaries.
- This key operation of the dot product between the support vectors and the test instances in the original lower dimensional space can be carried out before the nonlinear transformation to the new space. This allows using the optimization algorithm for finding the separating hyperplane of the new higher dimensional space in the original lower dimensional space. Therefore, the complexity is not as high as the one that results in applying standard discriminant analysis in the higher dimensional space, but is of the same order of magnitude as the one in the original feature space.

Optional: not needed for Quiz

(adapted from Y Kluger)

BN

Graphical Models - Microsoft Internet Explorer

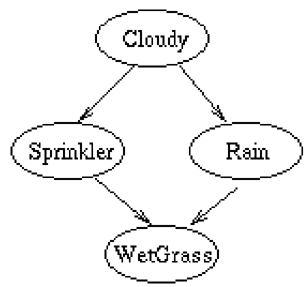
File Edit View Favorites Tools Help

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

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

Local Metods

- K nearest neighbors is a representative method of the instance-based learning approach. In this approach all the training instances are stored, and a distance function is used to determine which instances of the training set is closest to an unknown query instance. The distance between two instances with n dimensional feature vectors x and y is usually defined as the Euclidean distance between them.
- The k=1 nearest neighbor algorithm assigns to a query instance with feature vector y the class of the instance whose feature vector x is nearest to y.
- To increase stability it is better to take a larger value of k by assigning to the query instance the most common value among the k nearest training instances.

Optional: not needed for Quiz

(adapted from Y Kluger)

K nearest neighbors

- **Advantages**: simplicity, capability to approximate complex decision surfaces by a collection of simpler local decision surfaces in the vicinity of the query instance, and explicit conservation (storage) of all training set information.
- **Disadvantages**: strong sensitivity to the distance metric used and the fact that the features have different scales and therefore few of them can dominate others in determining a distance between the query and training set instances. Another difficulty is the fact that computation is done in query time rather than in advance.

Optional: not needed for Quiz

(adapted from Y Kluger)

Large-scale Datamining

- Relating Gene Expression to Protein Features and Parts
- Supervised Learning: Discriminants
- Simple Bayesian Approach for Localization Prediction
- Unsupervised Learning: k-means
- Correlation of Expression Data with Function
- Overview of Issues in Datamining
- Overview of Methods of Supervised Learning
- Focus on Decision Trees
- Overview of Methods of Unsupervised Learning
- Cluster Trees, Evolutionary Trees

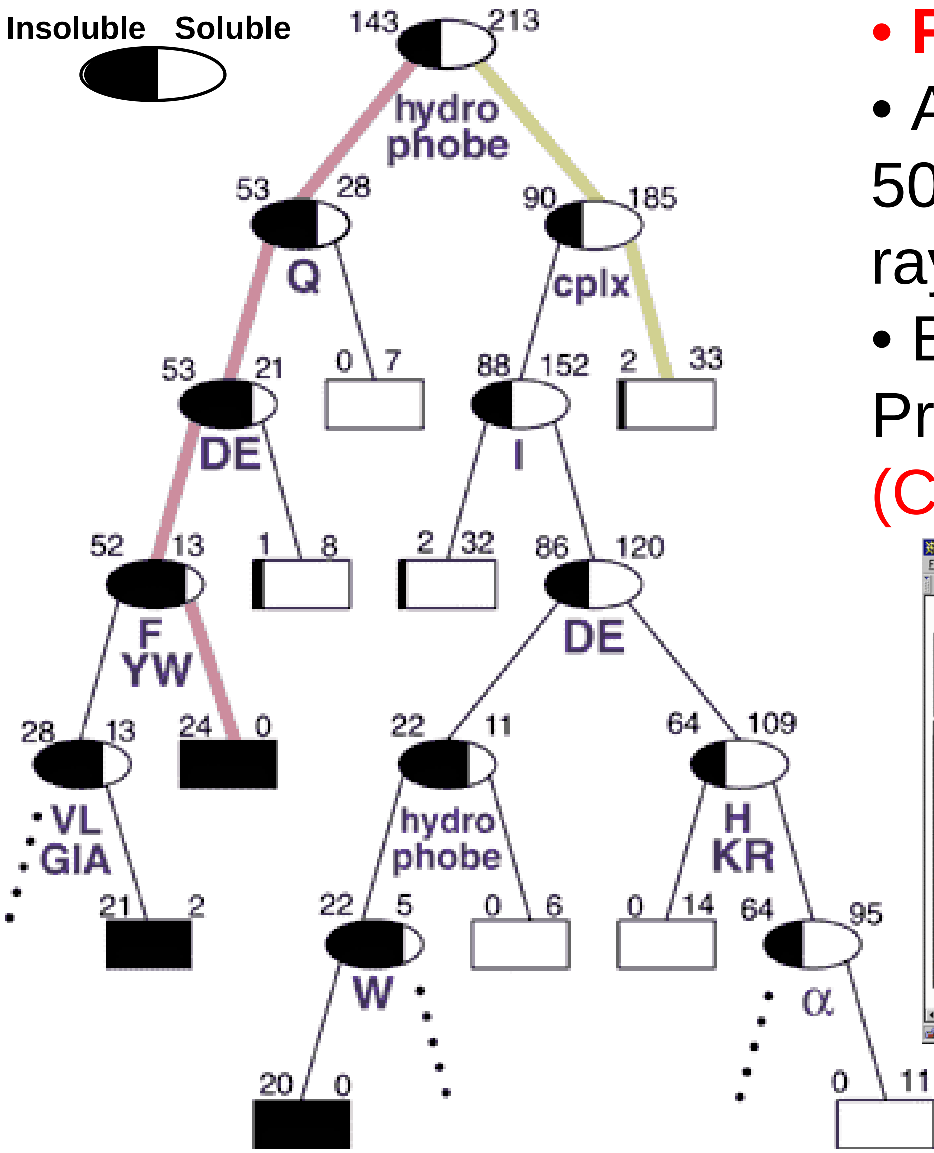
Decision Trees

- 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

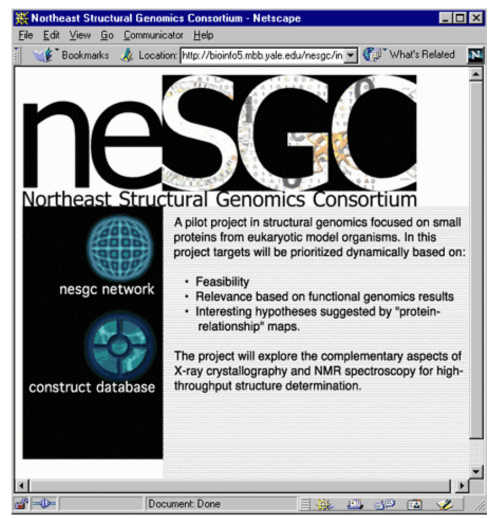
(adapted from Y Kluger)

Characterizing the Low-hanging Fruit for Experimental Structural Genomics



- **Retrospective Decision-Tree**
- Analysis of the Suitability of 500 M. thermo. proteins for X-ray/NMR work
- Based on results of Toronto Proteomics Group

(C Arrowsmith, A Edwards)



For example, proteins that fulfill the following sequence of four rules are likely to be insoluble: (1) have a hydrophobic stretch -- a long region (>20 residues) with average hydrophobicity less than -0.85 kcal/mole (on the GES scale); (2) Gln composition <4%; (3) Asp+Glu composition <17%; and (4) aromatic composition >7.5%. Conversely, proteins that do not have a hydrophobic stretch and have less than 27% of their residues in "low-complexity" regions are very likely to be soluble.

Trees

- devise prediction rules, screen the variables and summarize the multivariate dataset.
- nodes --ellipses (interior nodes) and rectangles (leaves) labeled by the more probable class (decision). Under each node-misclassification error proportion.
- Growing the tree requires (a) splitting the nodes and (b) pruning the tree. Maximizing the decrease in average impurity is a common criterion for splitting.
  noisy data- growing the tree will usually over-fit the training set.
  Most of the cost-complexity pruning algorithms--choose the smallest tree whose error rate performance is close to the minimal error rate of the over-fit larger tree.

Optional: not needed for Quiz

(adapted from Y Kluger)

Trees cont.

• Control parameters:
a) the threshold for splitting the node
b) minimal node size (default of 10) that can be further split
c) daughter node size must exceed a minimum (default of 5) for a split to be allowed

•Growing the trees is based on splitting the node that maximizes the reduction in deviance over all allowed binary splits of all terminal nodes. Splits are *not* chosen based on misclassification rate .A binary split for a continuous 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.

•Merge/split tree

Optional: not needed for Quiz

Advantages of tree-models

- easy to interpret and use when the predictors are a mix of numeric and nonnumeric (factor) variables
- invariant to scaling or re-expression of numeric variables.
- Compared with linear and additive models they are better in treating missing values and capturing non-additive behavior.
- They can also be used to predict nonnumeric dependent variables with more than two levels.
- ANN and decision tree learning often have comparable prediction accuracy and SVM algorithms are slower compared with decision tree.

Optional: not needed for Quiz

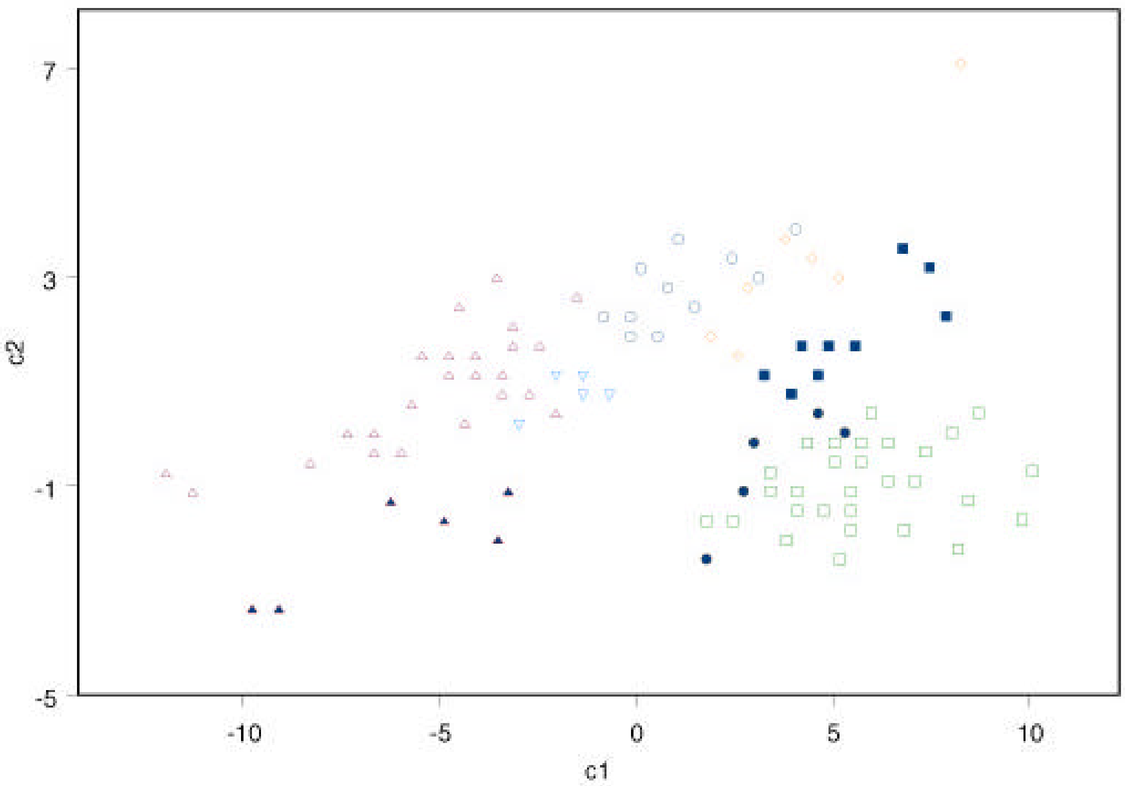
(adapted from Y Kluger)

Large-scale Datamining

- Relating Gene Expression to Protein Features and Parts
- Supervised Learning: Discriminants
- Simple Bayesian Approach for Localization Prediction
- Unsupervised Learning: k-means
- Correlation of Expression Data with Function
- Overview of Issues in Datamining
- Overview of Methods of Supervised Learning
- Focus on Decision Trees
- Overview of Methods of Unsupervised Learning
- Cluster Trees, Evolutionary Trees

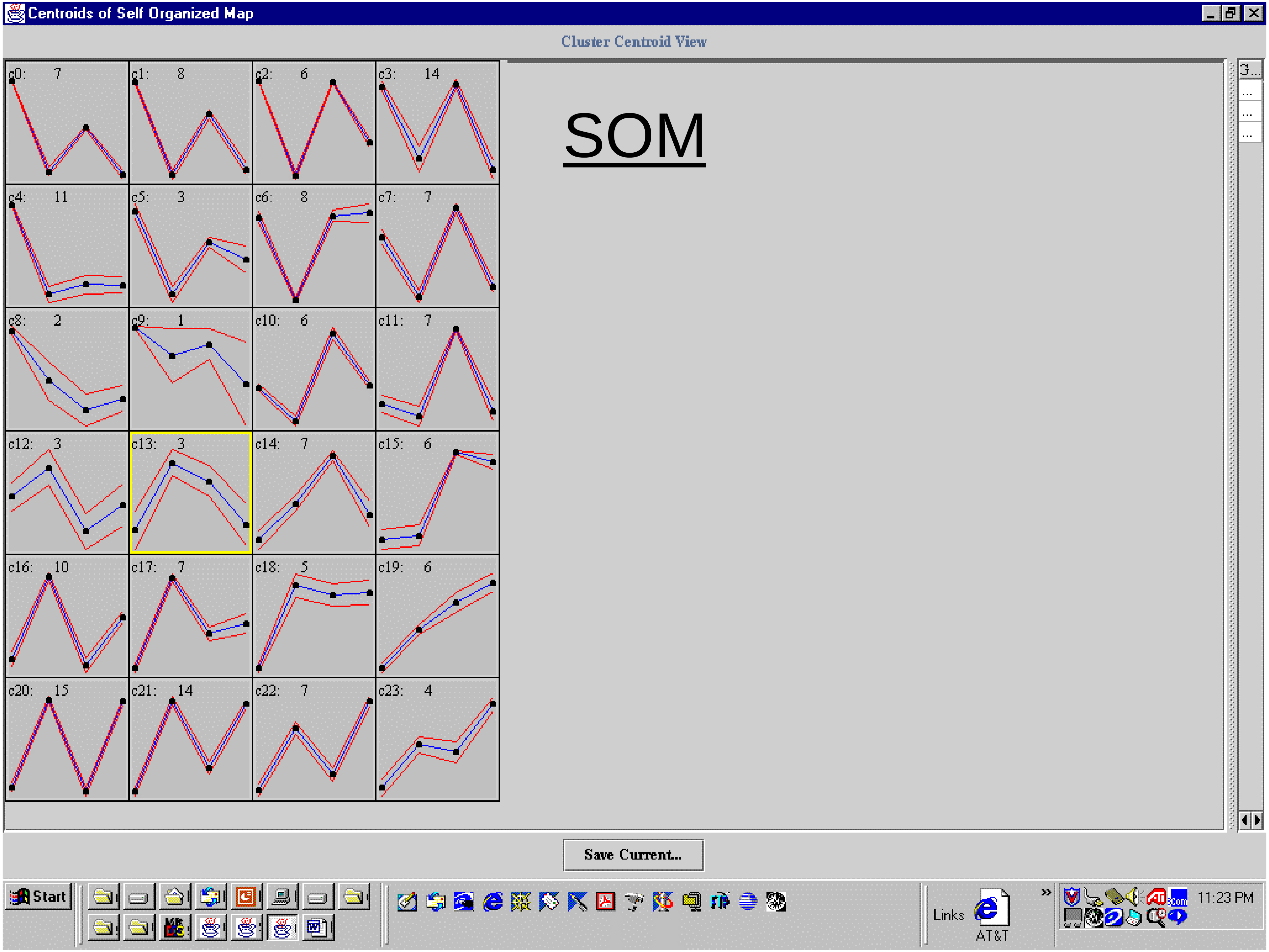
Unsupervised Learners

PCA

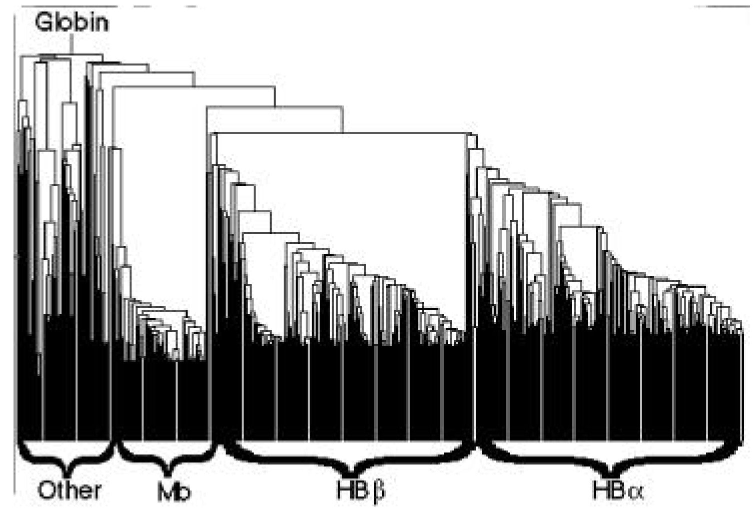


- principal components capture most of the variation of the data (95.2%). Each shape(color) belongs to a different ideal pattern.

(adapted from Y Kluger)


Centroids of Self Organized Map
Cluster Centroid View
c0: 7
c1: 8
c2: 6
c3: 14
c4: 11
c5: 3
c6: 8
c7: 7
c8: 2
c9: 1
c10: 6
c11: 7
c12: 3
c13: 3
c14: 7
c15: 6
c16: 10
c17: 7
c18: 5
c19: 6
c20: 15
c21: 14
c22: 7
c23: 4
SOM
Save Current...
Start
Links
AT&T
11:23 PM

Quickie Trees and Clustering



Top-down vs. Bottom up

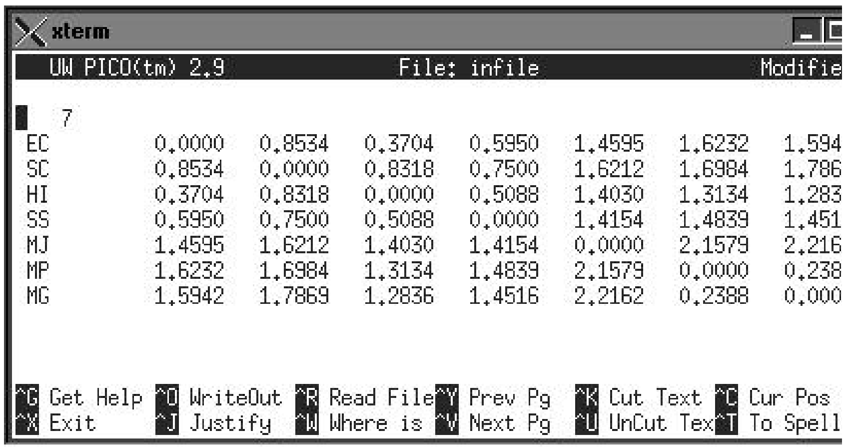
Top-down when you know how many subdivisions

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.

Methods of Building Trees from the bottom up

CHOOSE METHOD- Distance Based



Distance Methods

- Compute distance measures
- Build the tree from the table of distances

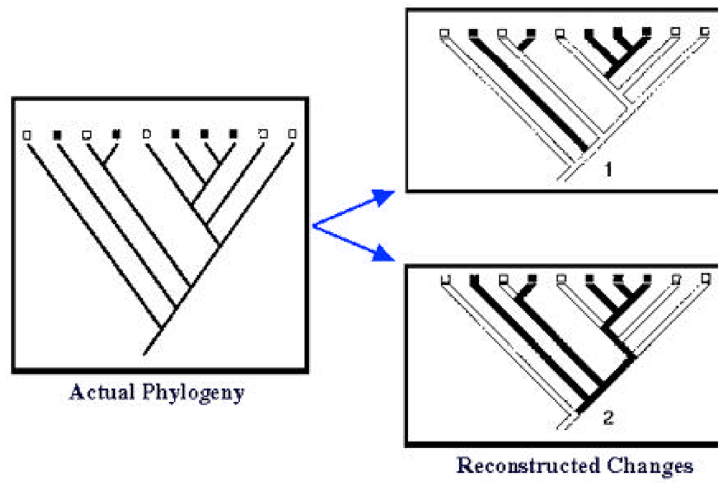
Assumptions

- A single coefficient of sequence similarity contains the information necessary to reconstruct the phylogeny
- May reduce the available information

Measuring Distances

- Compute all pairwise distances
- Correct for multiple substitution events
- Weight according to nucleotide substitution frequency
- Weight according to codon degeneracy
- Different measures presuppose different models of character evolution

CHOOSE METHOD- Parsimony



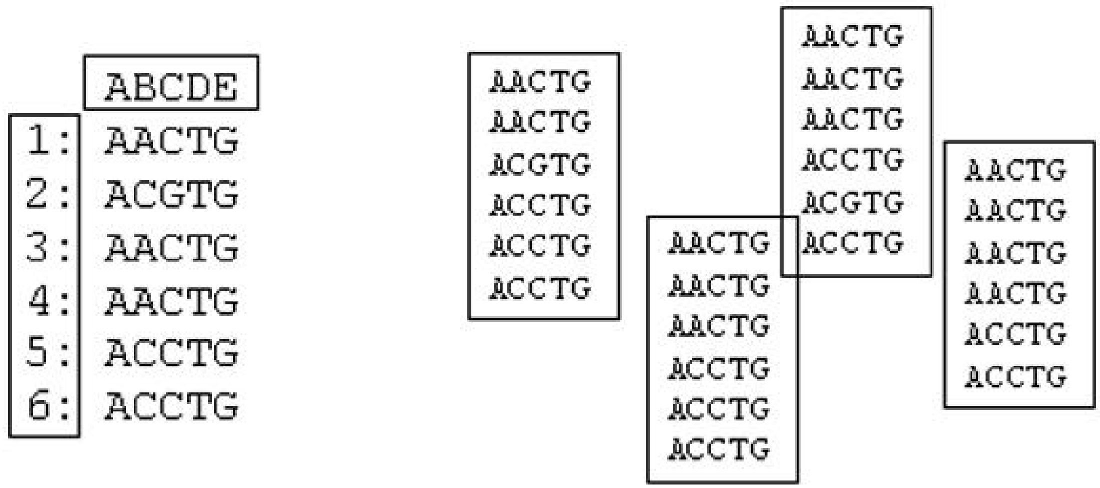
- Minimizing the number of changes at each node
- Requires greater computer resources than distance methods
- Depends on phylogenetically informative sites
- Retains all sequence information throughout the analysis

Problems:

- As the sequences diverge, the accuracy of the inference drops
- Long Edge Attraction
- Multiple islands of "almost the most parsimonious trees" can exist
- Requires greater computer resources than distance methods

Bootstrap to Test the Tree

ANALYZE TREE- Bootstrap

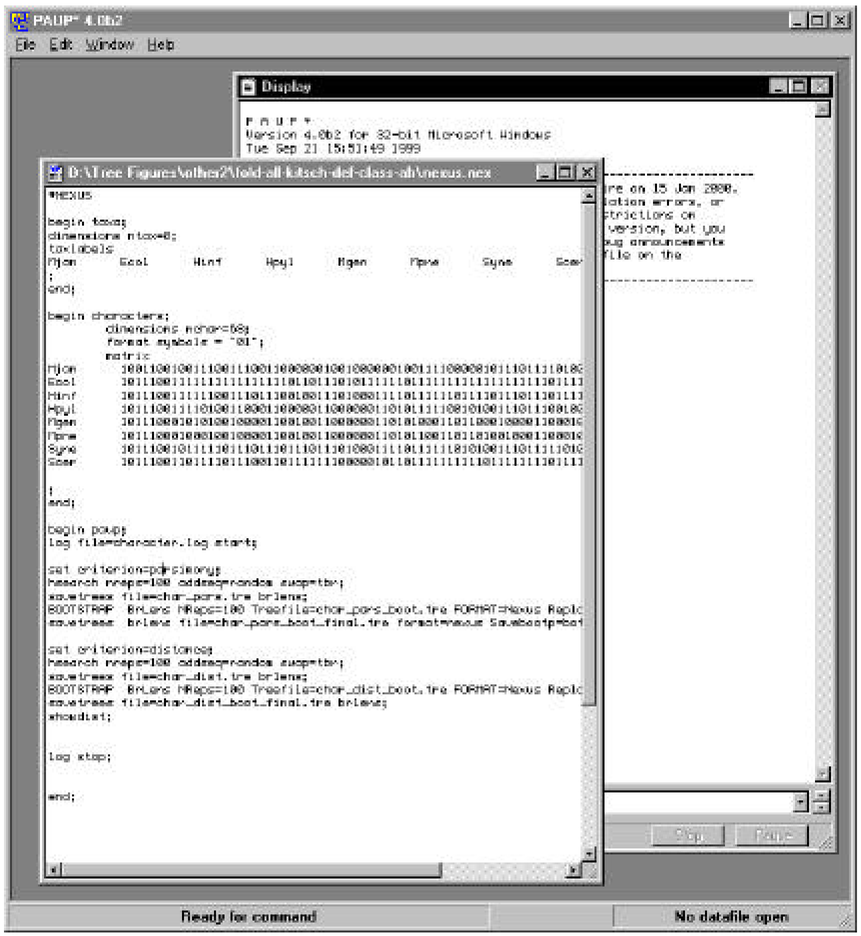


- Randomly resample the data with replacement, creating a new dataset that is then used to infer a phylogeny
- Generating replicate samples
- Observe tree topology
- Percentage of grouping
- Majority Rule Consensus

Popular Tree Program Systems

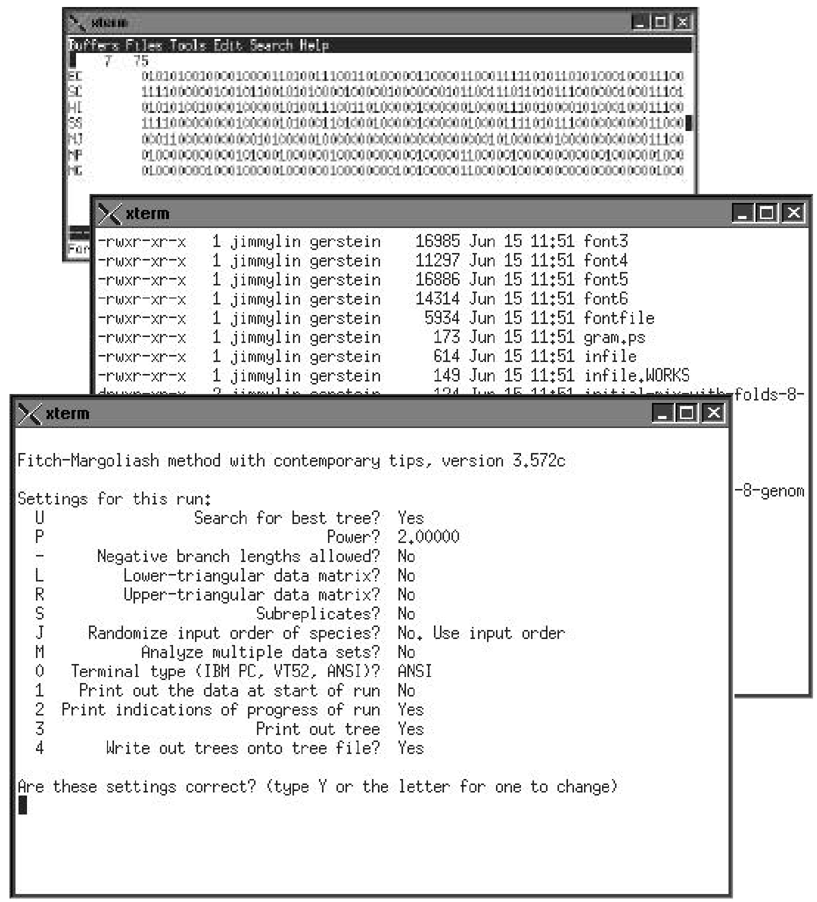
PREPARE THE DATA- PAUP

- Phylogenetic Analysis Using Parsimony
- David Swofford, Smithsonian
- Sophisticated parsimony program with a wide variety of options
  - Tree building algorithms
  - Weighting schemes
  - Resampling procedures

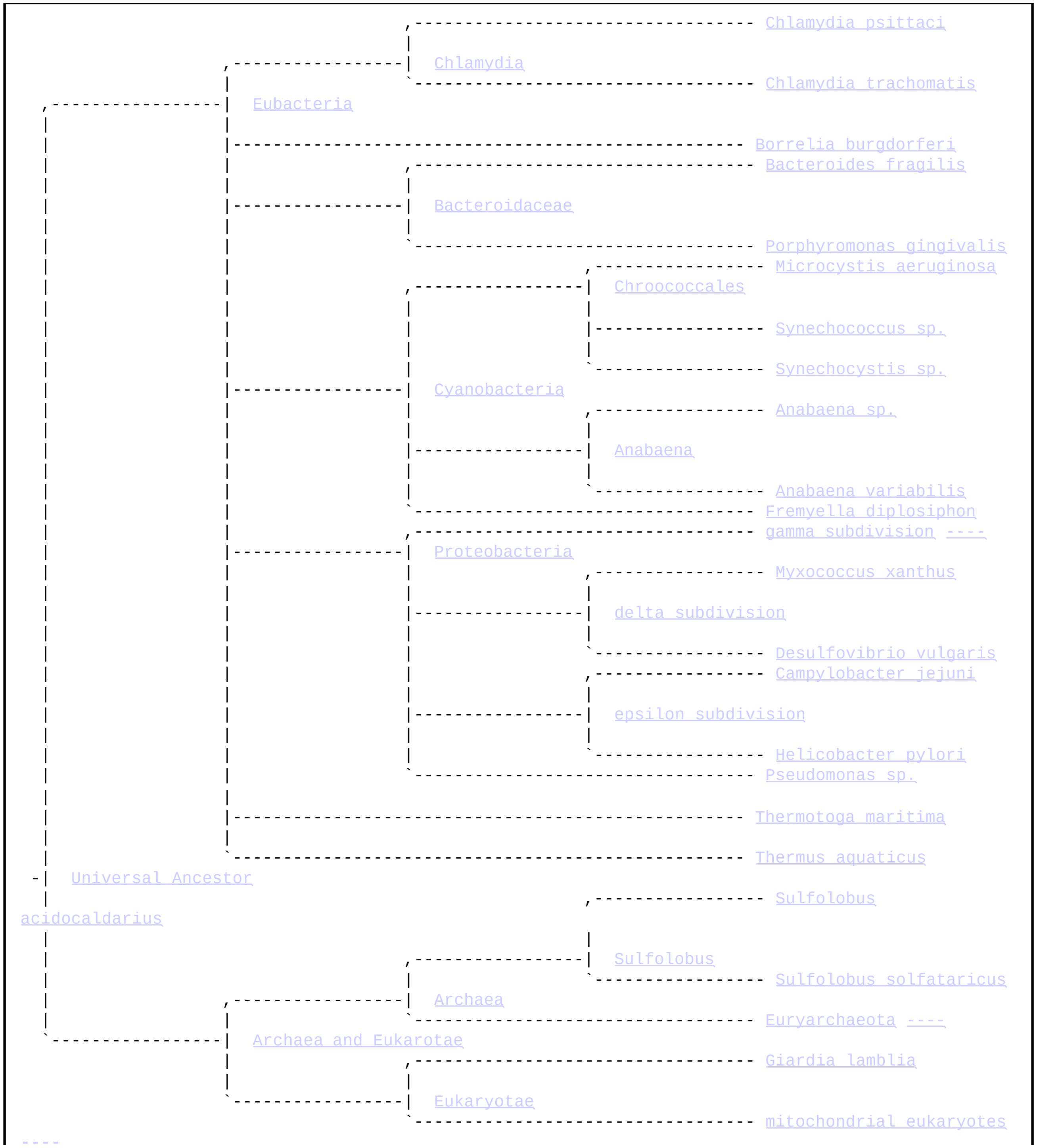


PREPARE THE DATA- Phylip

- J. Felsenstein, University of Washington
- A comprehensive set of phylogenetic inference programs
  - Maximum Likelihood
  - Parsimony
  - Distance
  - Single and multiple tree algorithms



Tree of Life


Chlamydia psittaci
Chlamydia
Chlamydia trachomatis
Eubacteria
Borrelia burgdorferi
Bacteroides fragilis
Bacteroidaceae
Porphyromonas gingivalis
Microcystis aeruginosa
Chroococcales
Synechococcus sp.
Synechocystis sp.
Cyanobacteria
Anabaena sp.
Anabaena
Anabaena variabilis
Fremyella diplosiphon
gamma subdivision ----
Proteobacteria
Myxococcus xanthus
delta subdivision
Desulfovibrio vulgaris
Campylobacter jejuni
epsilon subdivision
Helicobacter pylori
Pseudomonas sp.
Thermotoga maritima
Thermus aquaticus
Universal Ancestor
Sulfolobus
acidocaldarius
Sulfolobus
Sulfolobus solfataricus
Archaea
Euryarchaeota ----
Archaea and Eukarotae
Giardia lamblia
Eukaryotae
mitochondrial eukaryotes
