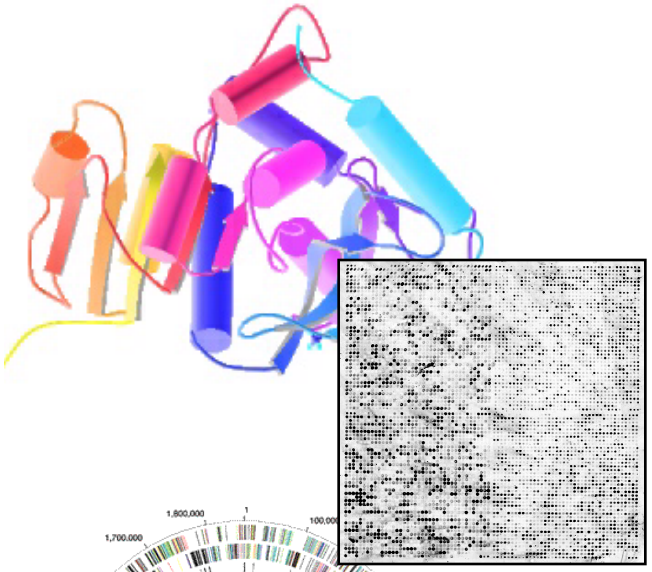
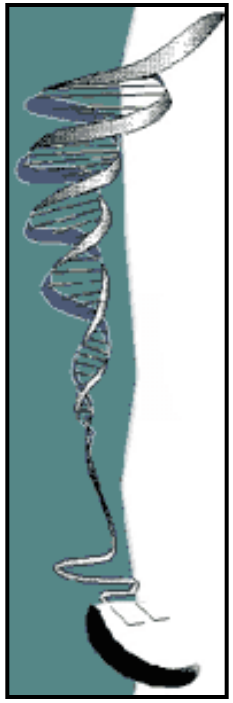
BIOINFORMATICS
Databases & Datamining I





Mark Gerstein, Yale University

gersteinlab.org/courses/452

(last edit in spring '11, including in-class changes & edits)

Databases & Datamining

- Building the data matrix
  - Generic
  - Gene centric
  - Genomic region
  - Network centric
- Databases

Building a Data Matrix

Generic

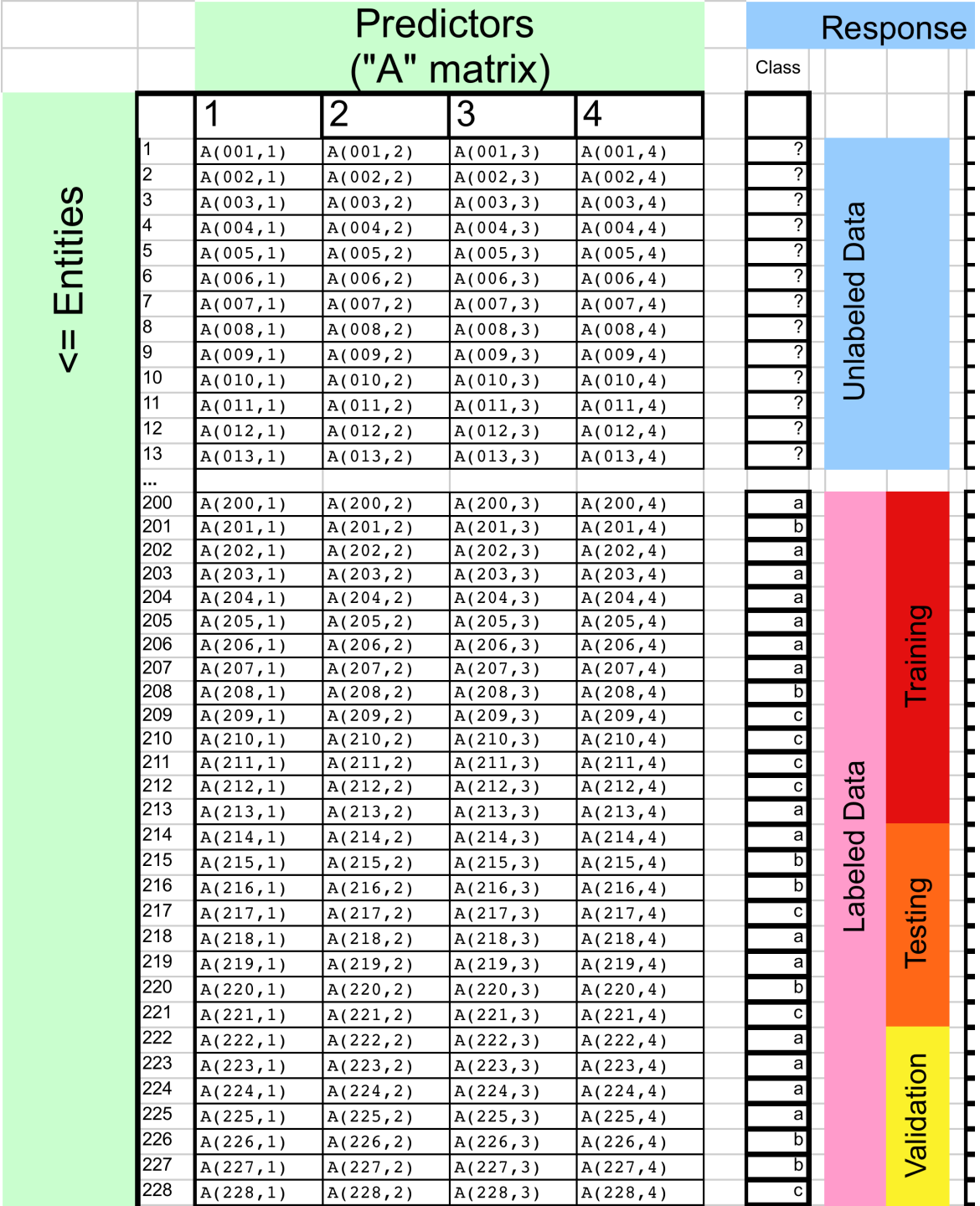
Generic Data Matrix

<= Entities	Features ("A" matrix)			
	1	2	3	4
1	A(001,1)	A(001,2)	A(001,3)	A(001,4)
2	A(002,1)	A(002,2)	A(002,3)	A(002,4)
3	A(003,1)	A(003,2)	A(003,3)	A(003,4)
4	A(004,1)	A(004,2)	A(004,3)	A(004,4)
5	A(005,1)	A(005,2)	A(005,3)	A(005,4)
6	A(006,1)	A(006,2)	A(006,3)	A(006,4)
7	A(007,1)	A(007,2)	A(007,3)	A(007,4)
8	A(008,1)	A(008,2)	A(008,3)	A(008,4)
9	A(009,1)	A(009,2)	A(009,3)	A(009,4)
10	A(010,1)	A(010,2)	A(010,3)	A(010,4)
11	A(011,1)	A(011,2)	A(011,3)	A(011,4)
12	A(012,1)	A(012,2)	A(012,3)	A(012,4)
13	A(013,1)	A(013,2)	A(013,3)	A(013,4)
...				

- Unsupervised analysis
  - Knn
  - PCA/SVD, CCA
  - Hierarchial clustering

Generic Data Matrix

- Supervised Classification Methods
  - SVMs & simple linear discriminant
  - Naive Bayes & more compex Bayes Nets
  - NNs
  - Decision Trees



Generic Data Matrix

- Regression
  - SVR

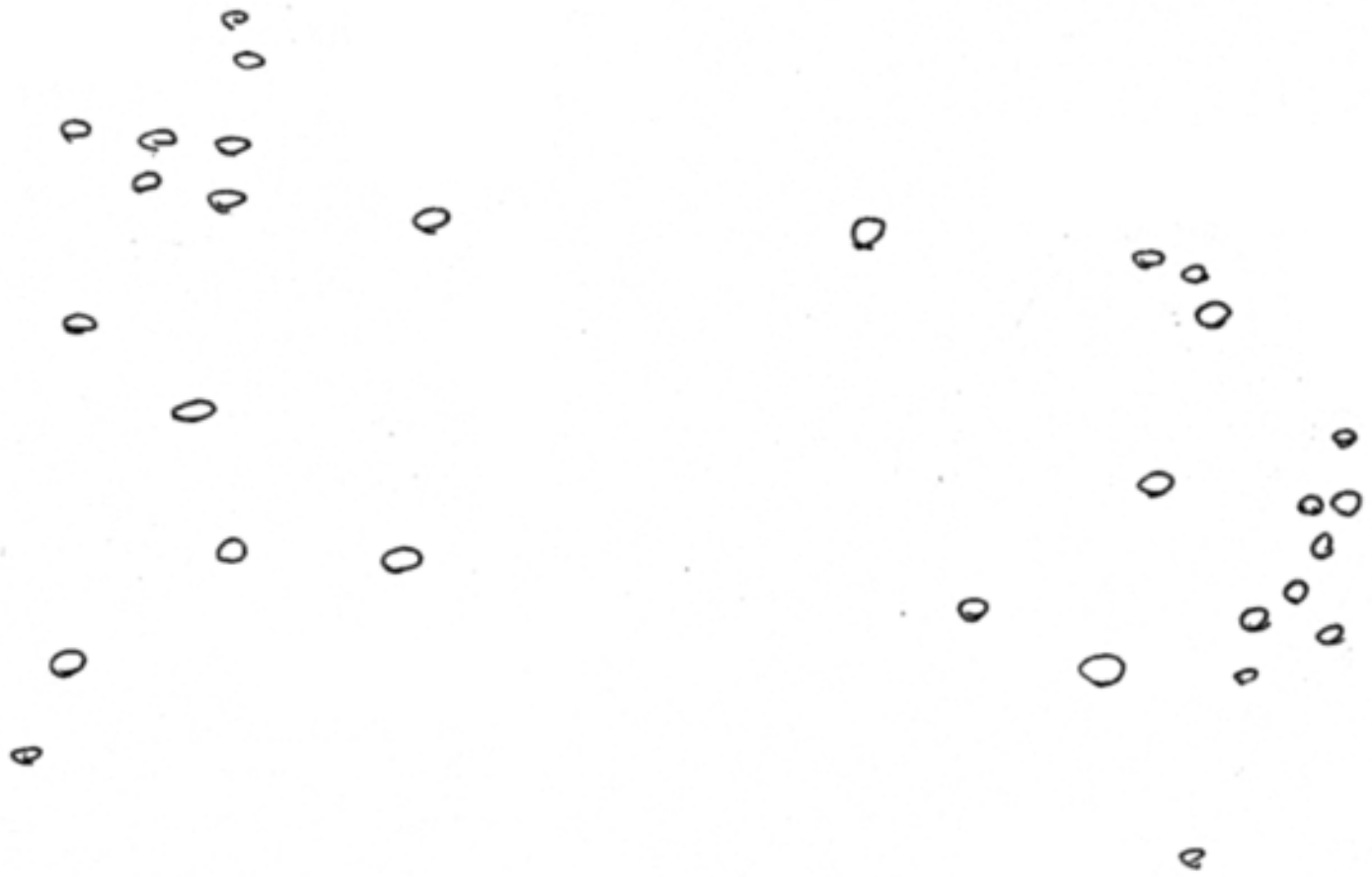
		Predictors ("A" matrix)				Response		
						Class		Value
<= Entities		1	2	3	4			
	1	A(001,1)	A(001,2)	A(001,3)	A(001,4)	?	Unlabeled Data	?
	2	A(002,1)	A(002,2)	A(002,3)	A(002,4)	?		?
	3	A(003,1)	A(003,2)	A(003,3)	A(003,4)	?		?
	4	A(004,1)	A(004,2)	A(004,3)	A(004,4)	?		?
	5	A(005,1)	A(005,2)	A(005,3)	A(005,4)	?		?
	6	A(006,1)	A(006,2)	A(006,3)	A(006,4)	?		?
	7	A(007,1)	A(007,2)	A(007,3)	A(007,4)	?		?
	8	A(008,1)	A(008,2)	A(008,3)	A(008,4)	?		?
	9	A(009,1)	A(009,2)	A(009,3)	A(009,4)	?		?
	10	A(010,1)	A(010,2)	A(010,3)	A(010,4)	?		?
	11	A(011,1)	A(011,2)	A(011,3)	A(011,4)	?		?
	12	A(012,1)	A(012,2)	A(012,3)	A(012,4)	?		?
	13	A(013,1)	A(013,2)	A(013,3)	A(013,4)	?		?
	...							
	200	A(200,1)	A(200,2)	A(200,3)	A(200,4)	a	Labeled Data	1
	201	A(201,1)	A(201,2)	A(201,3)	A(201,4)	b		1.3
	202	A(202,1)	A(202,2)	A(202,3)	A(202,4)	a		2
	203	A(203,1)	A(203,2)	A(203,3)	A(203,4)	a		1
	204	A(204,1)	A(204,2)	A(204,3)	A(204,4)	a		a
	205	A(205,1)	A(205,2)	A(205,3)	A(205,4)	a		6
	206	A(206,1)	A(206,2)	A(206,3)	A(206,4)	a		7
	207	A(207,1)	A(207,2)	A(207,3)	A(207,4)	a		0
	208	A(208,1)	A(208,2)	A(208,3)	A(208,4)	b		-1.3
	209	A(209,1)	A(209,2)	A(209,3)	A(209,4)	c		-5
	210	A(210,1)	A(210,2)	A(210,3)	A(210,4)	c		-1.2
	211	A(211,1)	A(211,2)	A(211,3)	A(211,4)	c		2
	212	A(212,1)	A(212,2)	A(212,3)	A(212,4)	c		4
	213	A(213,1)	A(213,2)	A(213,3)	A(213,4)	a		2
	214	A(214,1)	A(214,2)	A(214,3)	A(214,4)	a		1
	215	A(215,1)	A(215,2)	A(215,3)	A(215,4)	b		4
	216	A(216,1)	A(216,2)	A(216,3)	A(216,4)	b		-2
	217	A(217,1)	A(217,2)	A(217,3)	A(217,4)	c		-3
	218	A(218,1)	A(218,2)	A(218,3)	A(218,4)	a		-1
	219	A(219,1)	A(219,2)	A(219,3)	A(219,4)	a		2
	220	A(220,1)	A(220,2)	A(220,3)	A(220,4)	b		0
	221	A(221,1)	A(221,2)	A(221,3)	A(221,4)	c		6
	222	A(222,1)	A(222,2)	A(222,3)	A(222,4)	a		7
	223	A(223,1)	A(223,2)	A(223,3)	A(223,4)	a		0.2
	224	A(224,1)	A(224,2)	A(224,3)	A(224,4)	a		0
	225	A(225,1)	A(225,2)	A(225,3)	A(225,4)	a		0.3
	226	A(226,1)	A(226,2)	A(226,3)	A(226,4)	b		0.5
	227	A(227,1)	A(227,2)	A(227,3)	A(227,4)	b		-1.3
	228	A(228,1)	A(228,2)	A(228,3)	A(228,4)	c		-5

Training

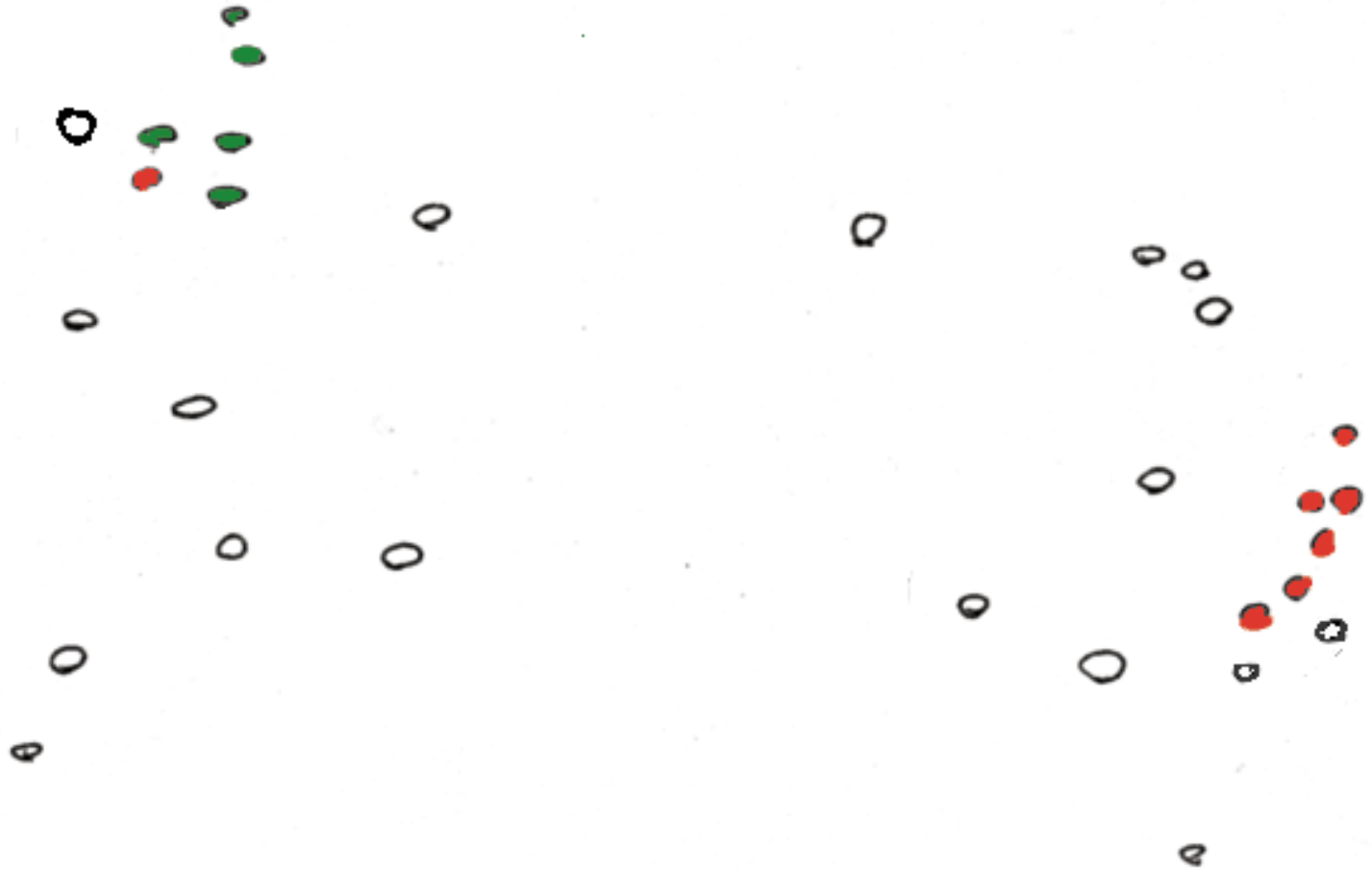
Testing

Validation

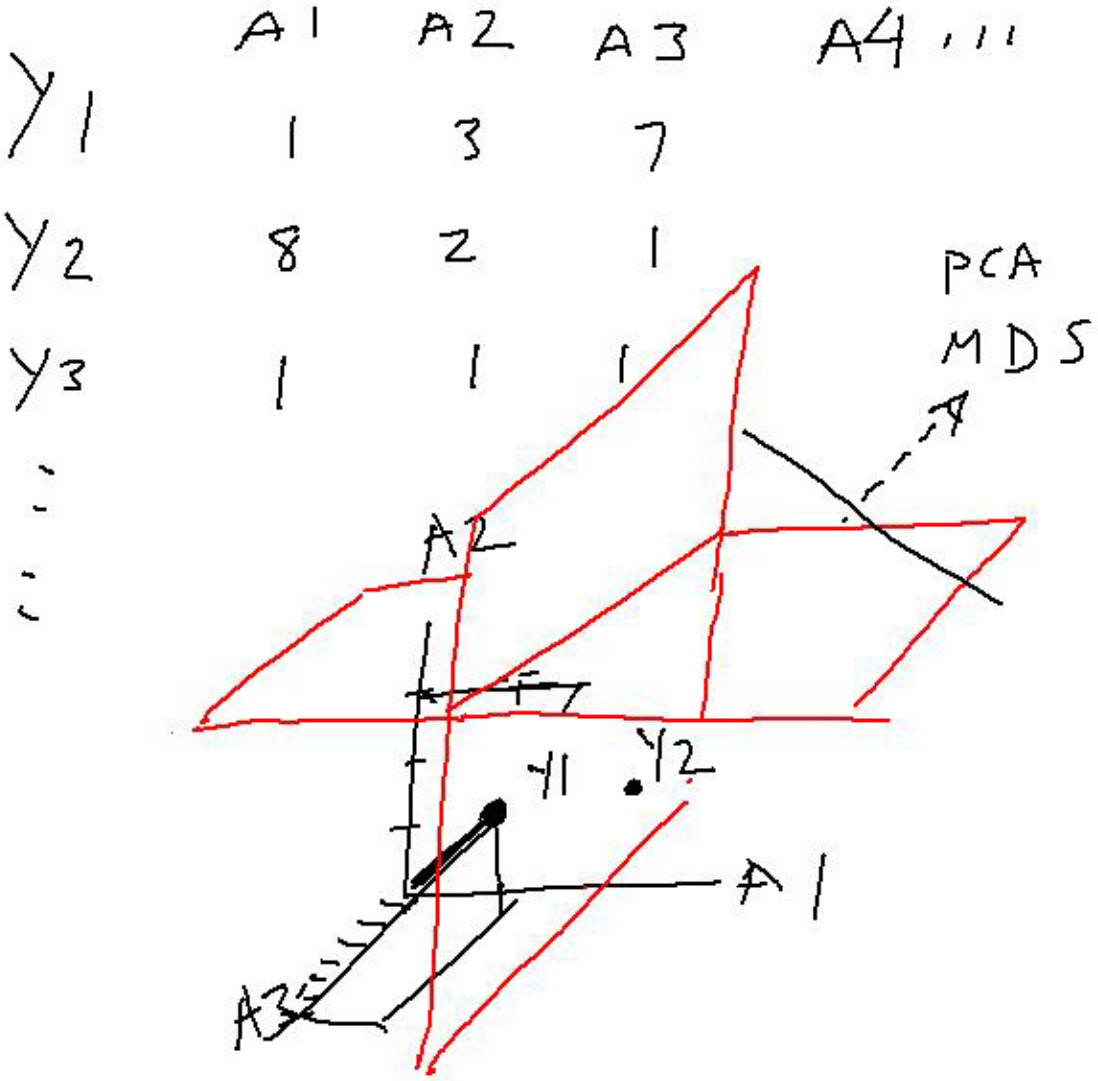
Represent predictors in abstract high dimensional space



“Tag” Certain Points



Abstract high-dimensional space representation



Building a Data Matrix

Gene Centric (yeast)

Gene Expression Information and Protein Features

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

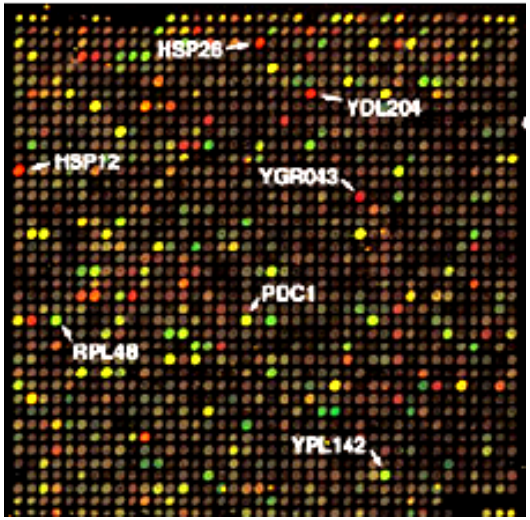
...

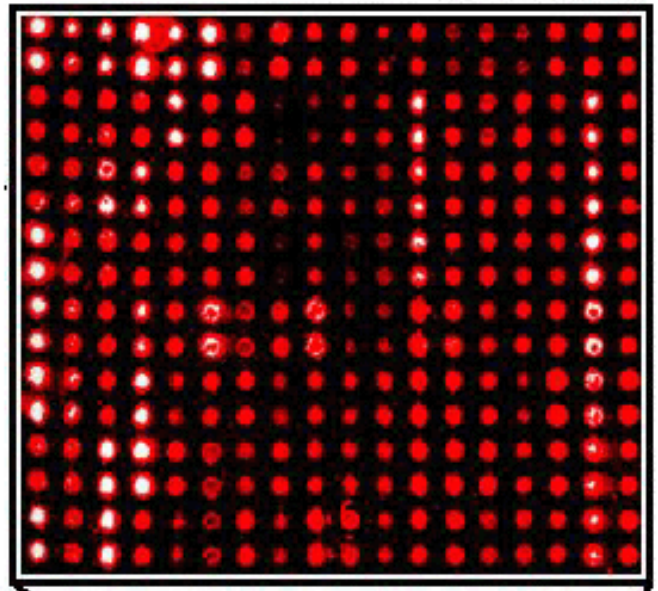
"Microarray" Data

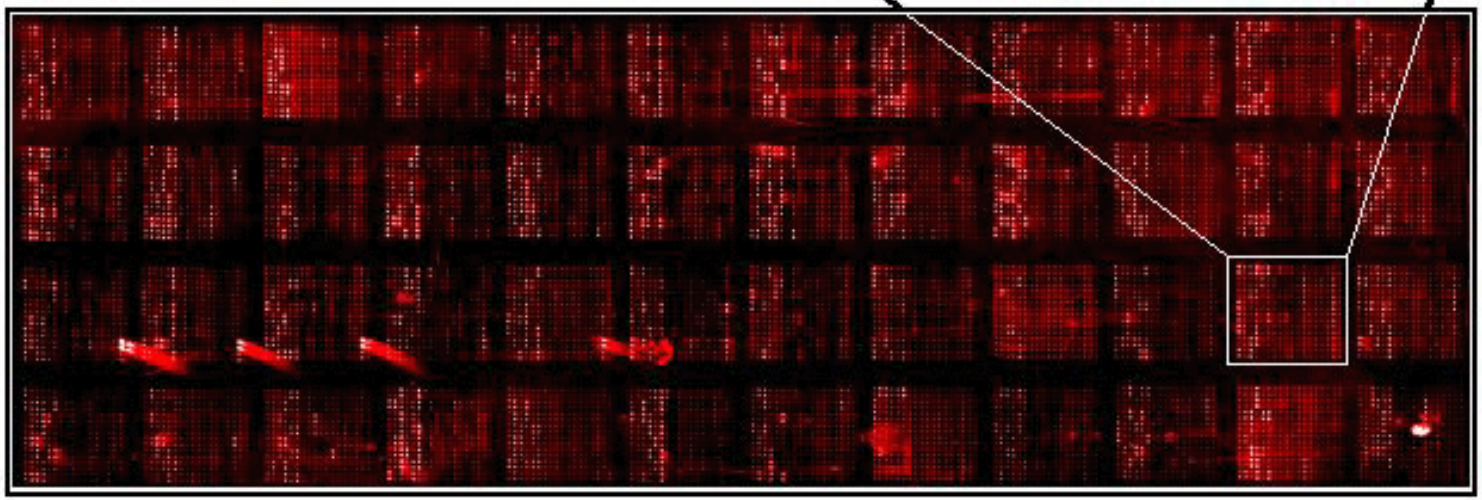
1st generation, Expression Arrays (Brown)

2nd gen., Proteome Chips (Snyder)

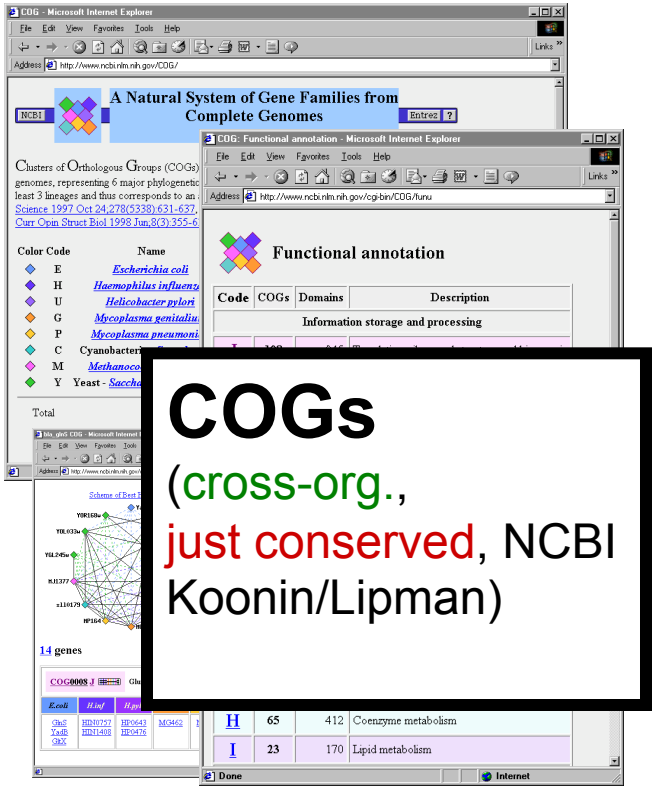
3rd gen., Readouts from nextgen seq. (RPKMS)



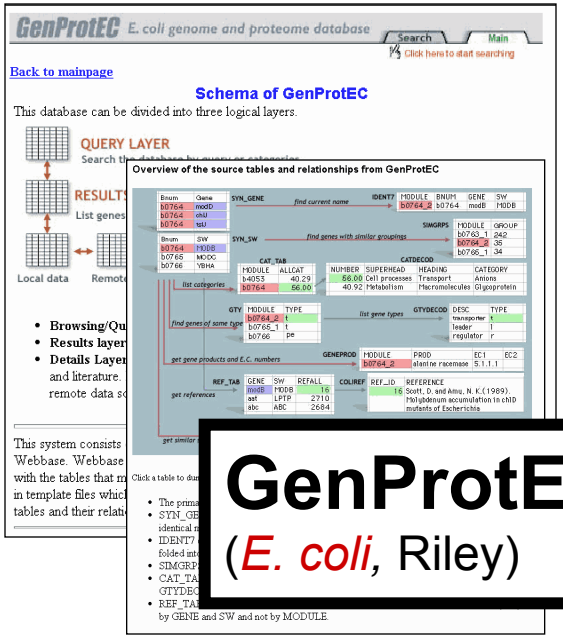




Functional Classification

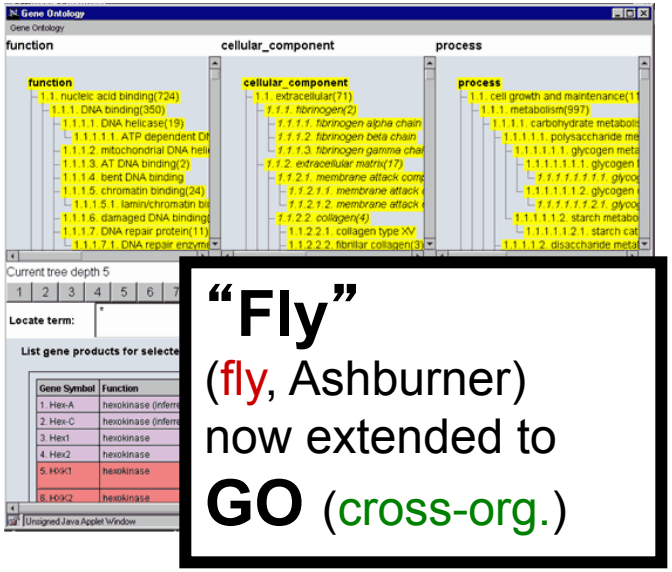


COGs
(cross-org., just conserved, NCBI Koonin/Lipman)

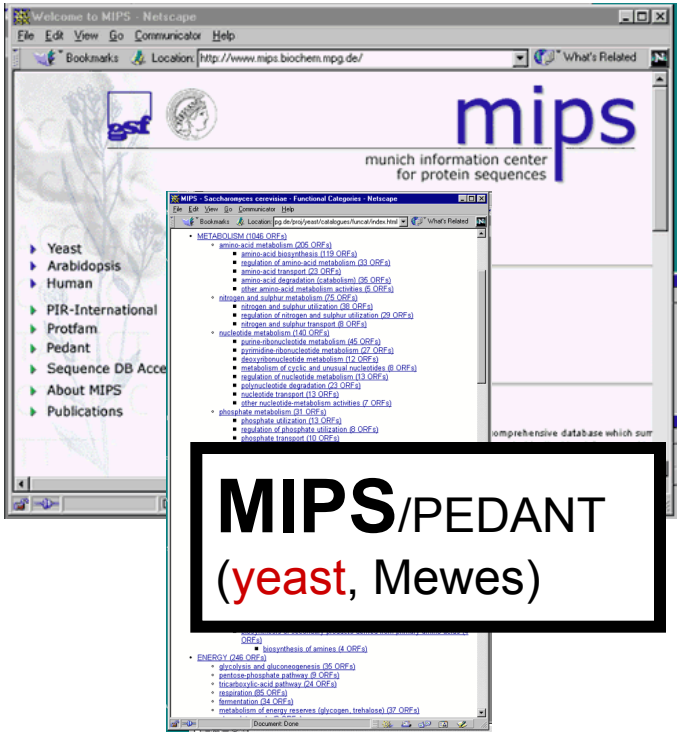


GenProtEC
(*E. coli*, Riley)

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



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



MIPS/PEDANT
(yeast, Mewes)

Also:

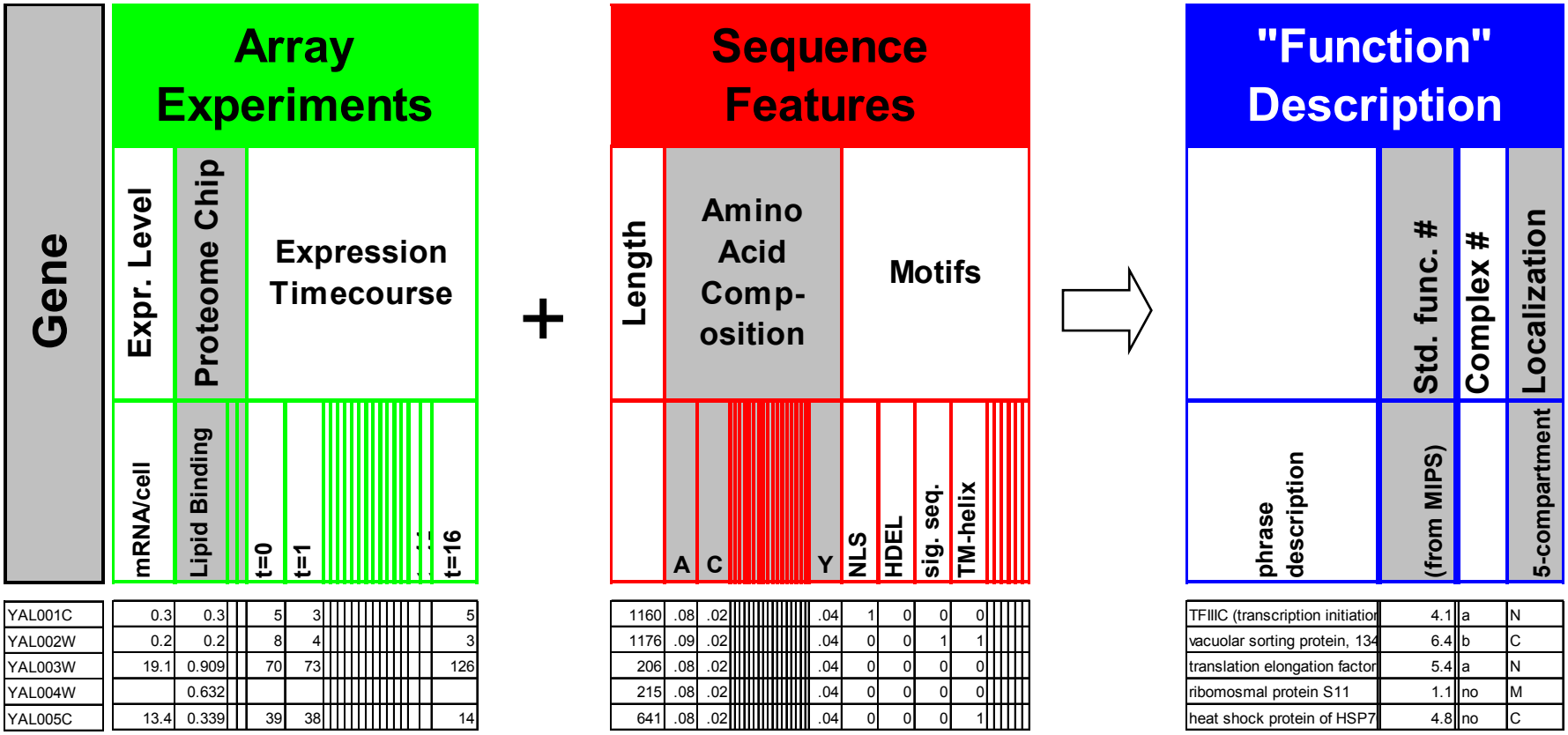
Other SwissProt Annotation

WIT, KEGG (just pathways)

TIGR EGAD (human ESTs)

SGD (yeast)

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



6000+

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

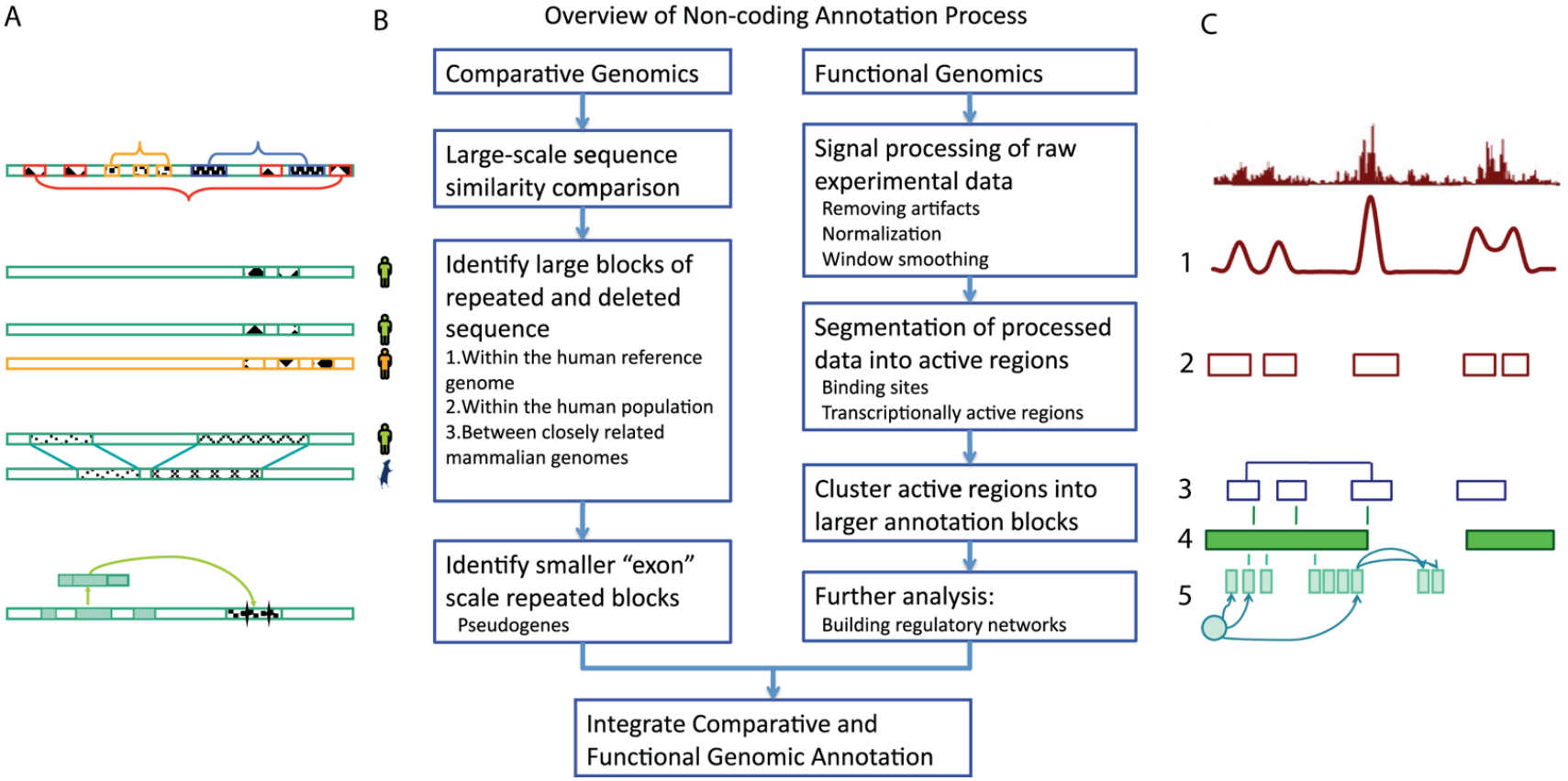
Typical Predictors and Response for Yeast (6000 genes X 1000 features)

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	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	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	????

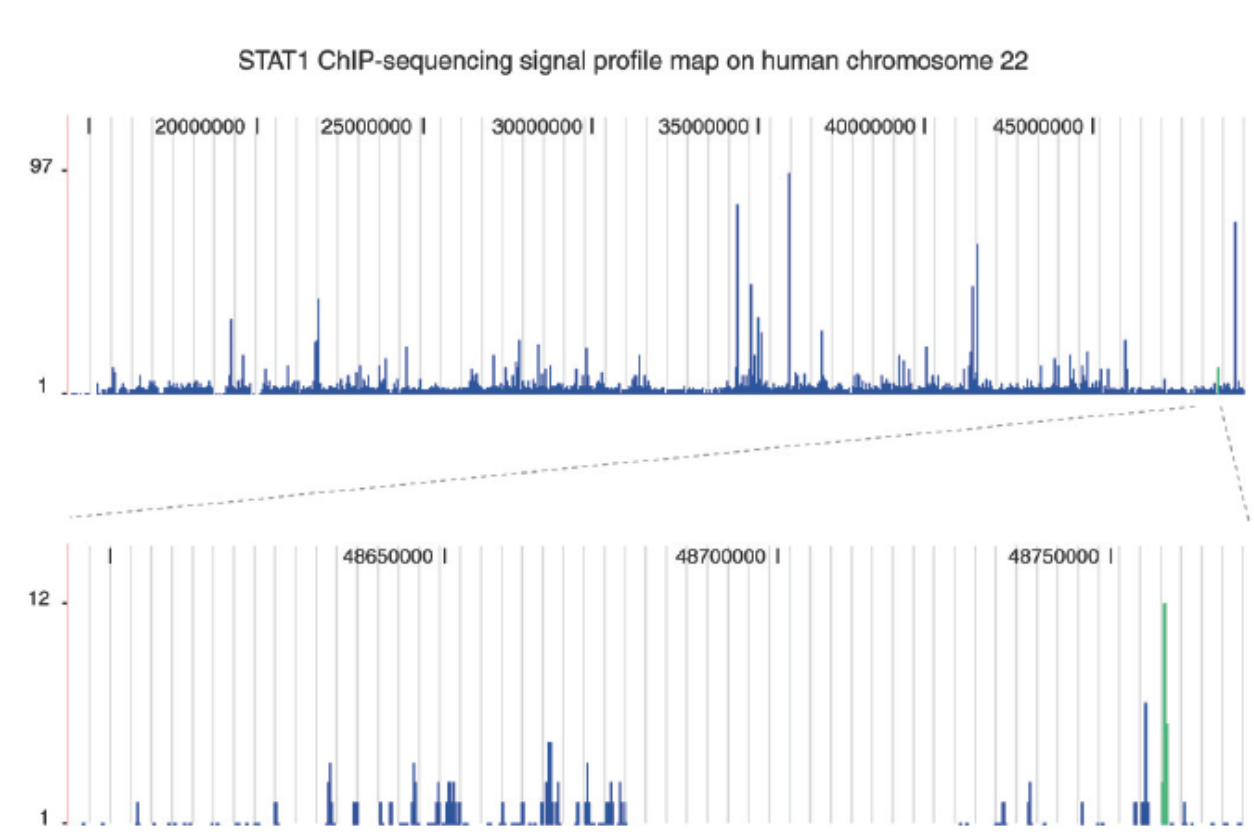
Building a Data Matrix

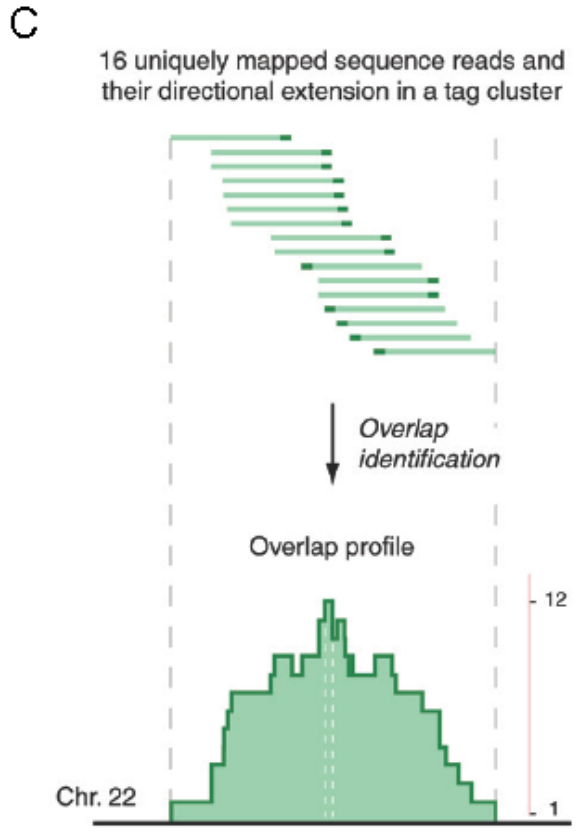
Genomic Region (Human Intergenic Regions)

Overview of Intergenic Annotation Flow



Representative Signal from Chip-Seq: creating a signal "track"



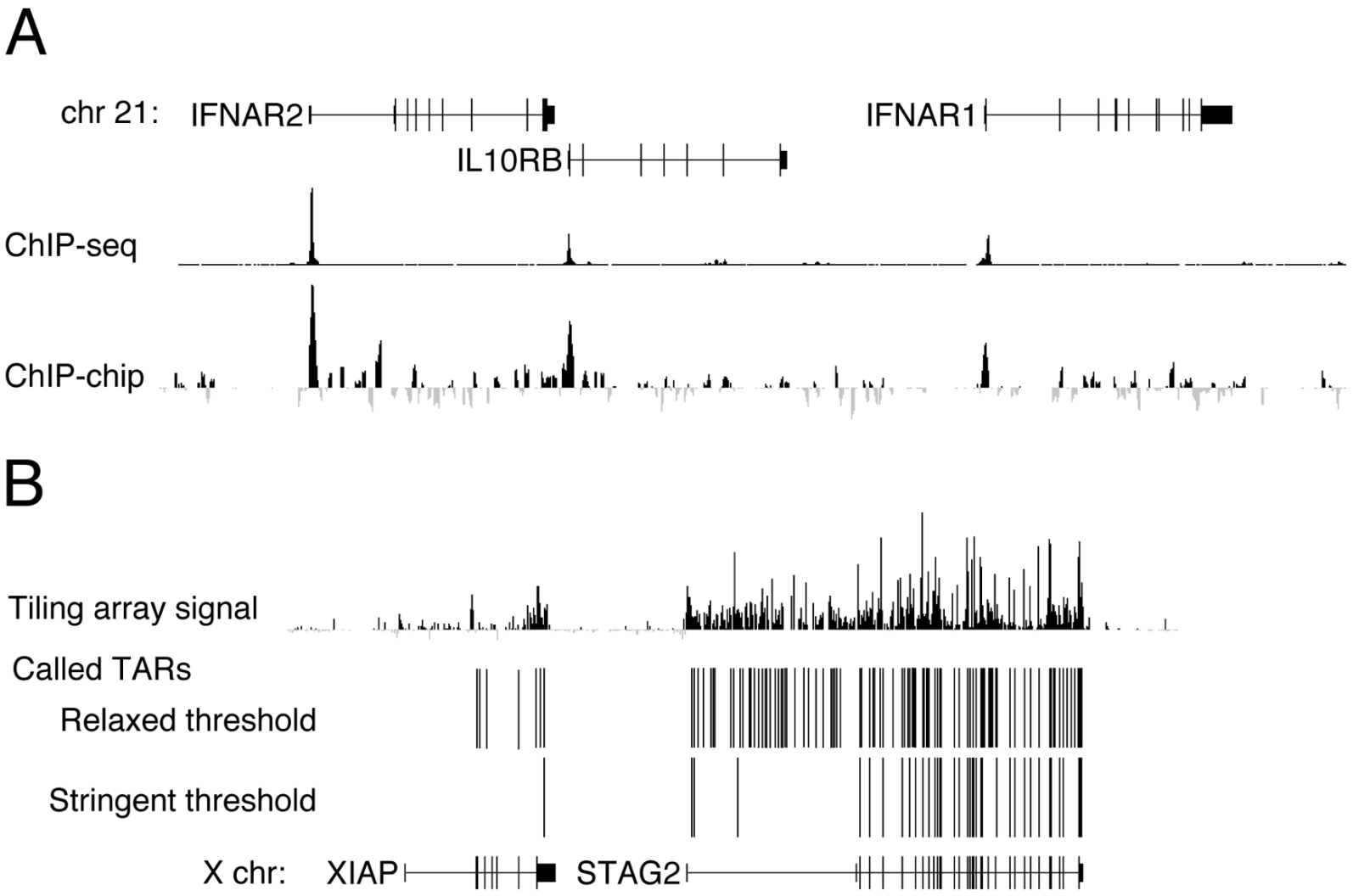


[Robertson et al., Nat. Meth. ('07); Zhang et al. PLOS Comp. Bio. (in revision, ’08)]

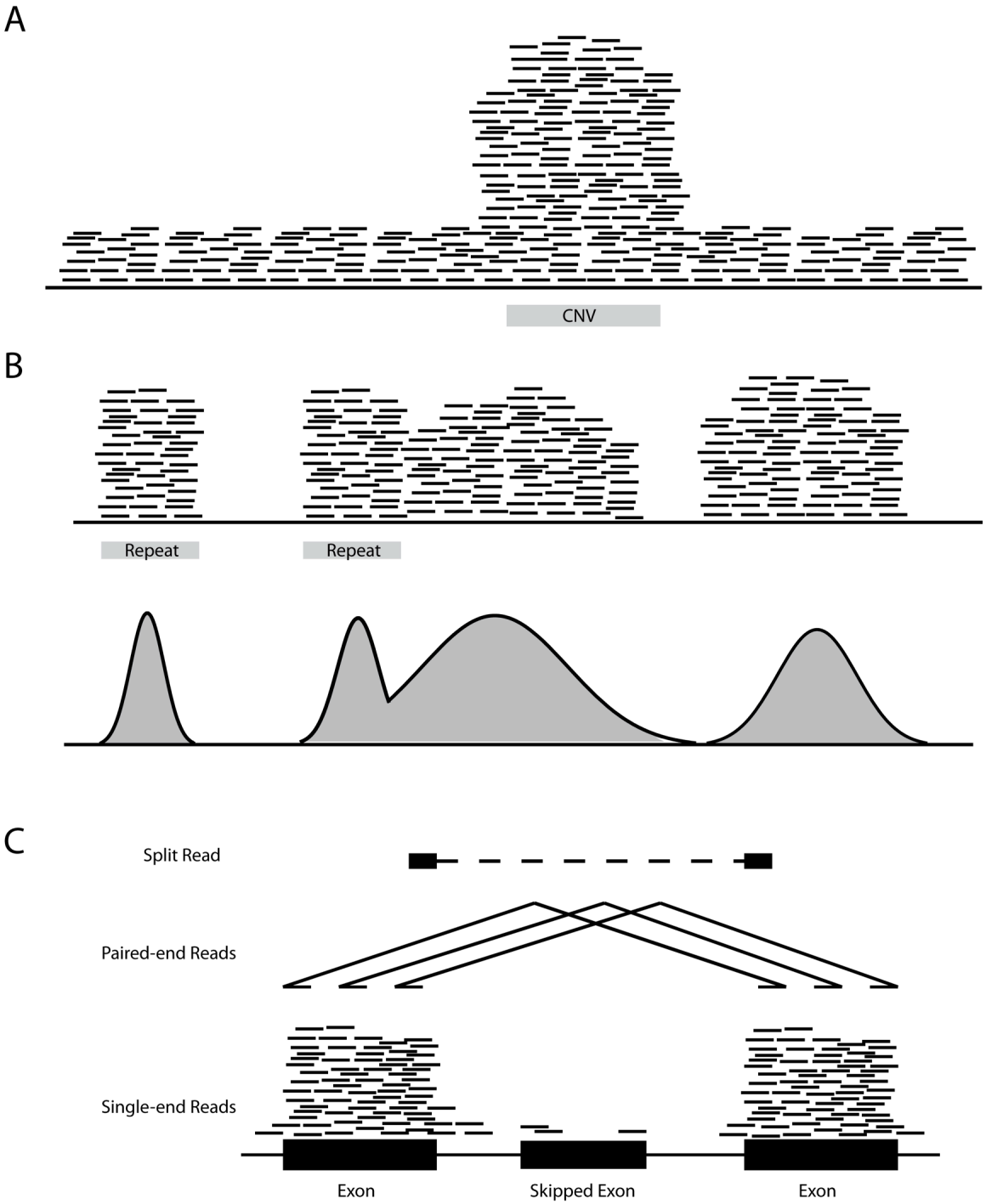
Segmentation of Raw Signal to Generate "Hits"

- Simple min-run, max-gap threshold
  - Threshold but demand that the "hit" last min-run and also jump over gaps smaller than max-gap
- Chip-seq programs that check (e.g. with bionomial test) for enrichment over the background
- HMM segmenters for broad regions

Issues in Signal Processing: Thresholding



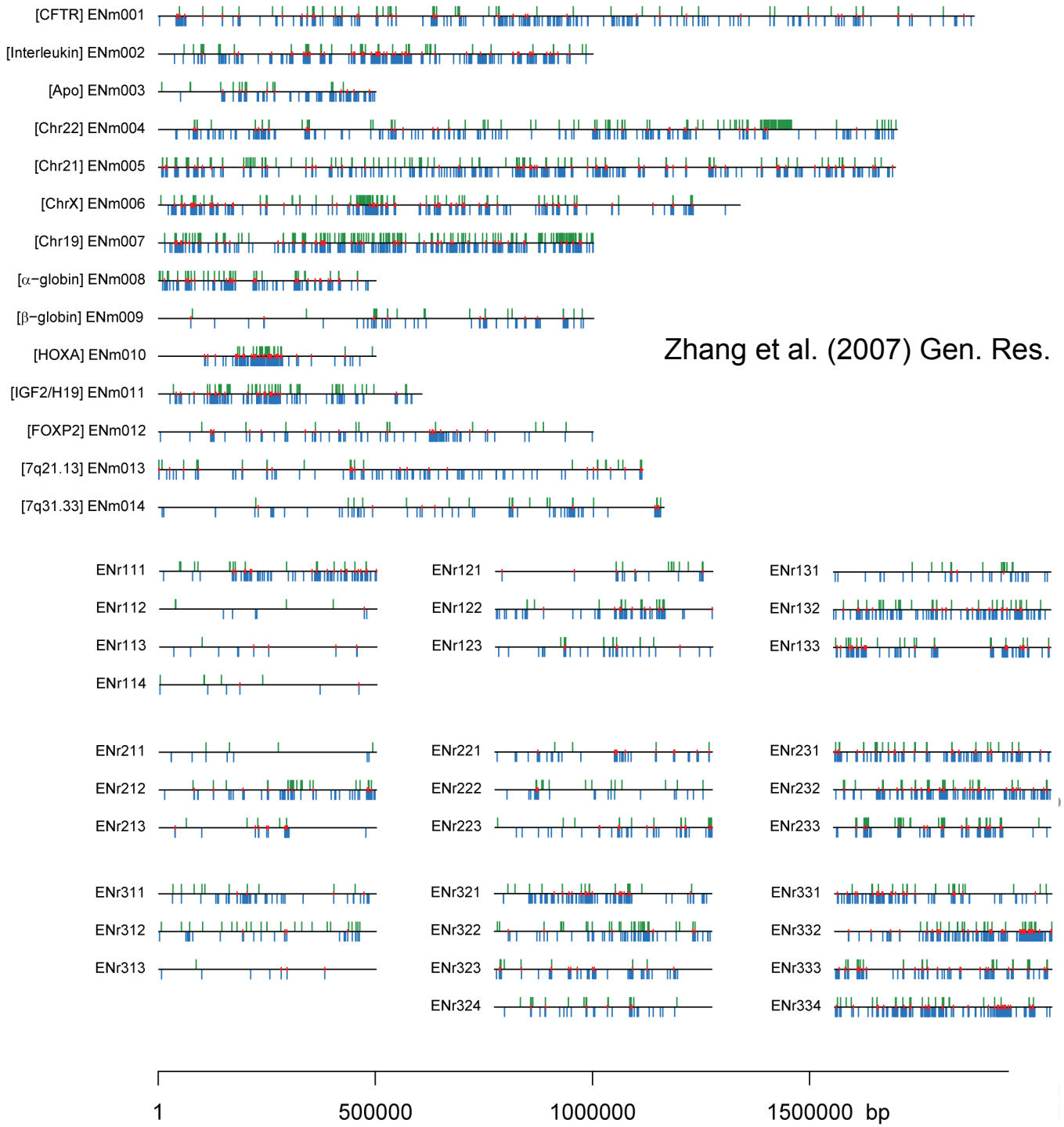
Issues in Signal Processing: Multi-mapping



Landscape of ENCODE Transcriptional Regulatory Elements

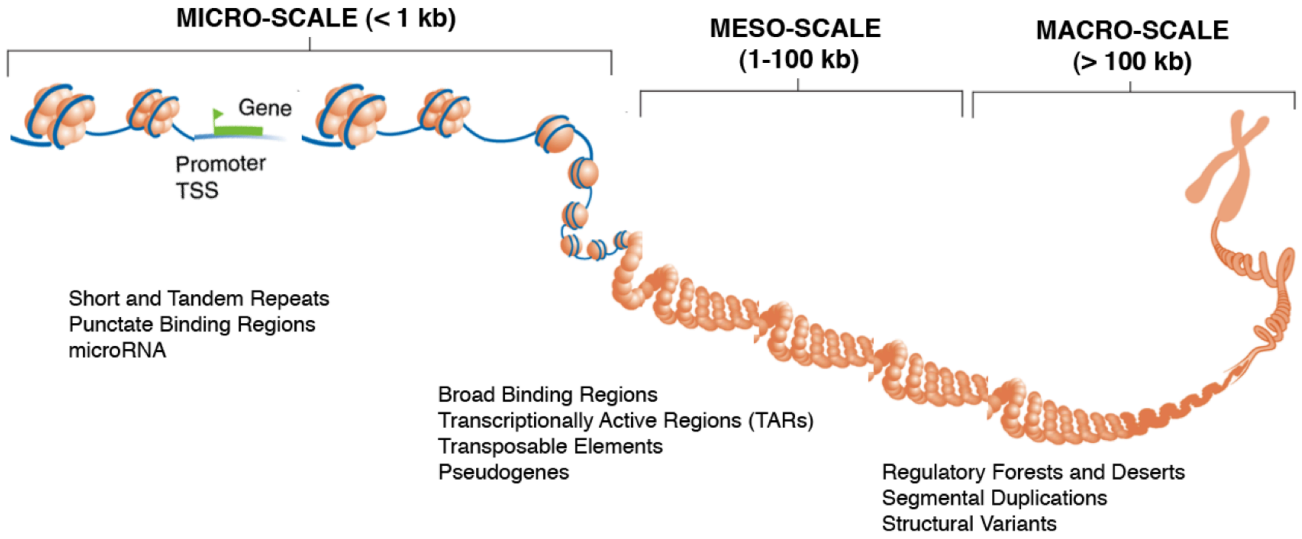
- Analyzed 105 lists of transcriptional regulatory elements in the encode regions
- 29 transcription factors, 9 cell lines, 2 time points
  - RNA Pol2
  - Histone modifications such as Ac & Me
  - Core promoters
  - Promoter proximal elements
  - Others such as enhancers, silencers, insulators, & response elements

UCSC browser



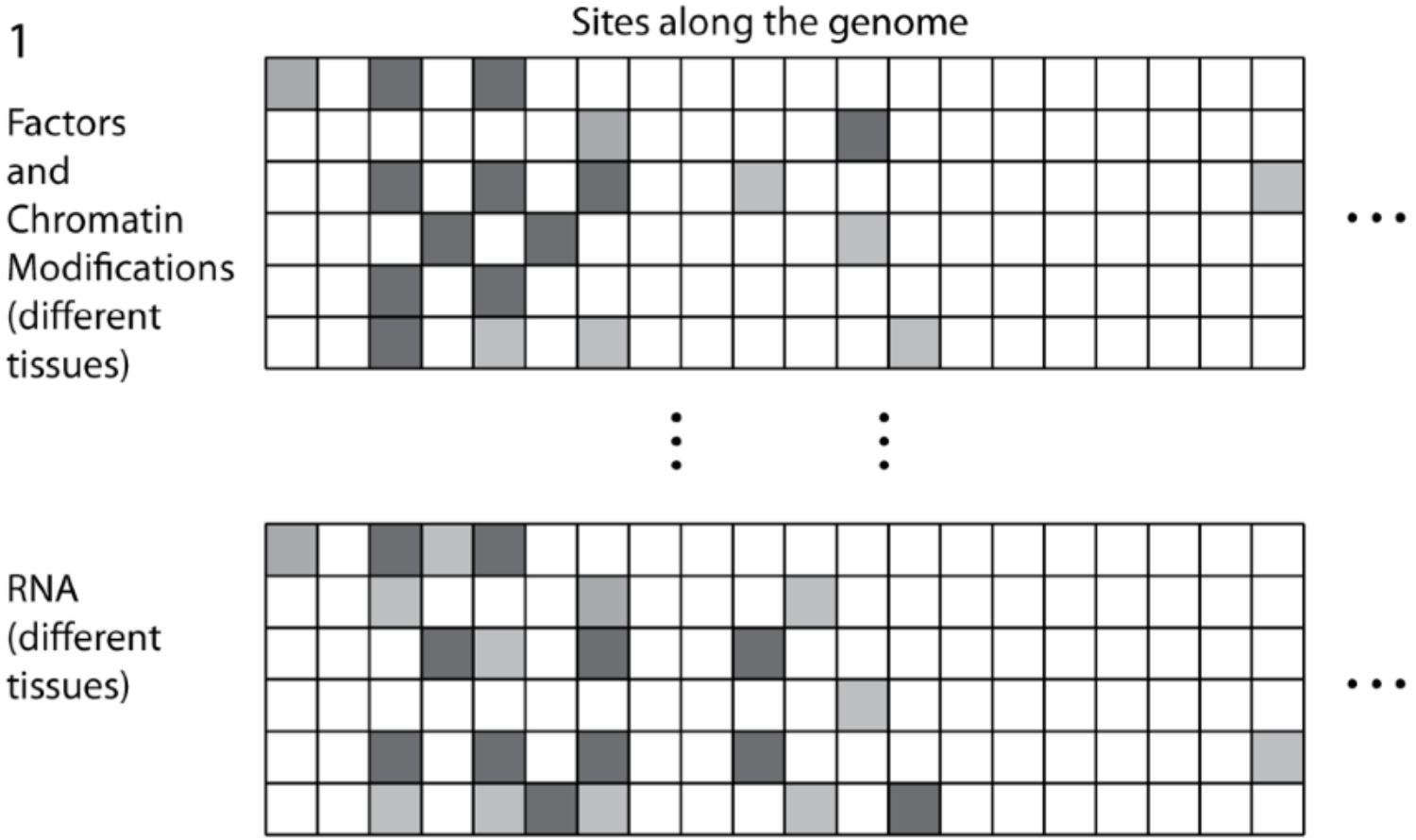
Zhang et al. (2007) Gen. Res.

Many types of Genomic Elements



	Length (nt)			Number of	Genome Coverage	
	Average	Longest	Scale	Items	(Mb)	(%)
Comparative Genomics						
Short and Tandem Repeats			Micro	795,016	56.1	1.81
Simple Repeat	63	2,961		415,917	26.1	0.84
Satellite	1,444	160,602		8,997	13.0	0.42
Low Complexity	46	2,023		370,102	17.0	0.55
DNA Transposons	215	3,625	Micro-Meso	459,524	98.6	3.17
Retrotransposons						
LINEs	426	8,505	Micro-Meso	1,490,241	634.6	20.4
Alu SINE Element	261	614	Micro	1,186,885	309.7	9.97
Pseudogenes			Micro-Meso			
Duplicated	6,607	181,882		2413	15.9	0.51
Processed	723	15,732		8303	6.0	0.19
Segmental Duplications	5,740	630 kb	Meso-Macro	26,469	151.9	4.89
Structural Variants	8,761	3.3 Mb	Meso-Macro	96,874	848.8	27.3
Functional Genomics						
Punctate Binding Sites						
STAT1	446	9,079	Micro-Meso	~2,300	1.0	0.03
CTCF	1,181	79,200	Meso	~35,000	41.4	1.33
H3K4me3	1,759	71,025	Meso	~62,000	110.2	3.55
Broad Binding Sites						
H3K36me3	4,518	380,076	Meso	~130,000	589	19.0
miRNA	89	150	Micro	718	0.063	0.00
TARs	72	1,854	Micro-Meso	644,200	46.7	1.50
Regulatory Forests	3,890	35,165	Meso-Macro	68,900	268	8.62
Regulatory Deserts	27,107	203,691	Meso-Macro	72,500	1970	63.4

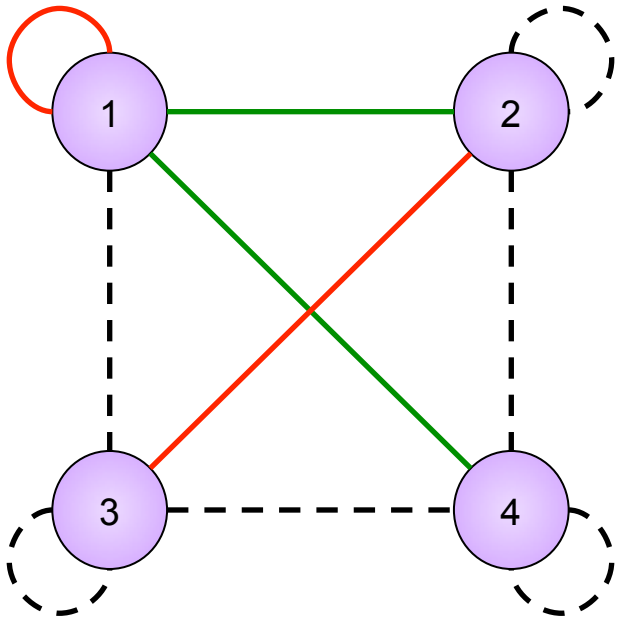
Structure of Genomic Features Matrix (1 M sites X 100 Features)

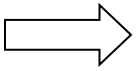


Building a Data Matrix

Network Centric

Training sets





	1	2	3	4
1	0	1	?	1
2	1	?	0	?
3	?	0	?	?
4	1	?	?	?

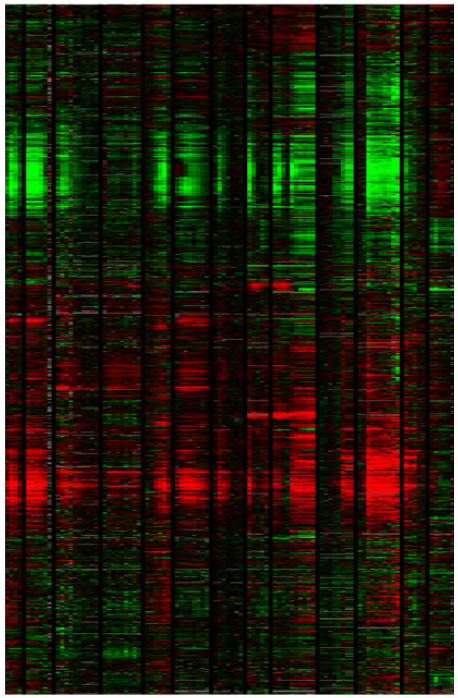
Known interactions

Known non-interactions

Unknown

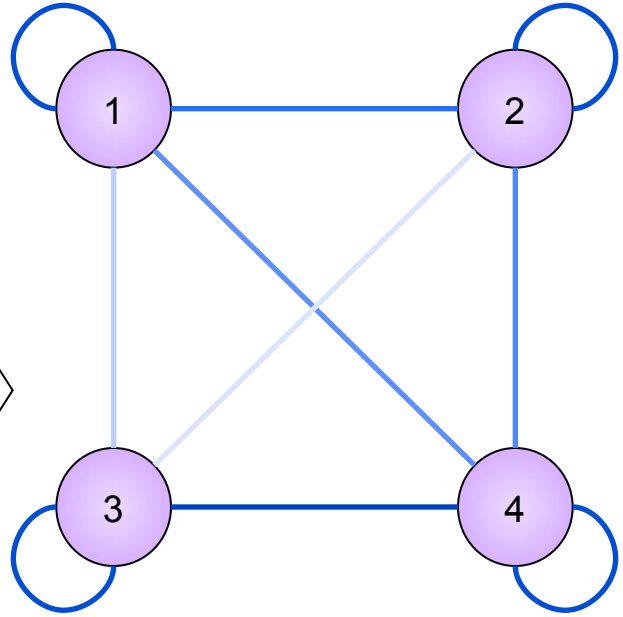
Network prediction: features

- Example 1: gene expression



Gasch et al., 2000

$x_1 = (0.2, 2.4, 1.5, \ldots)$
$x_2 = (0.8, 2.2, 1.5, \ldots)$
$x_3 = (4.3, 0.1, 7.5, \ldots)$
…
$sim(x_1, x_2) = 0.62$
$sim(x_1, x_3) = -0.58$
…

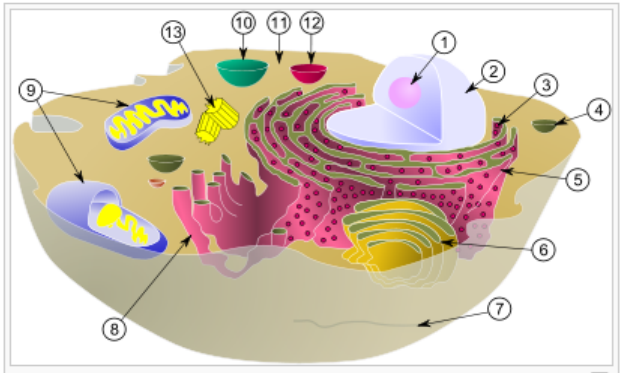


Similarity scale:

1 **-1**

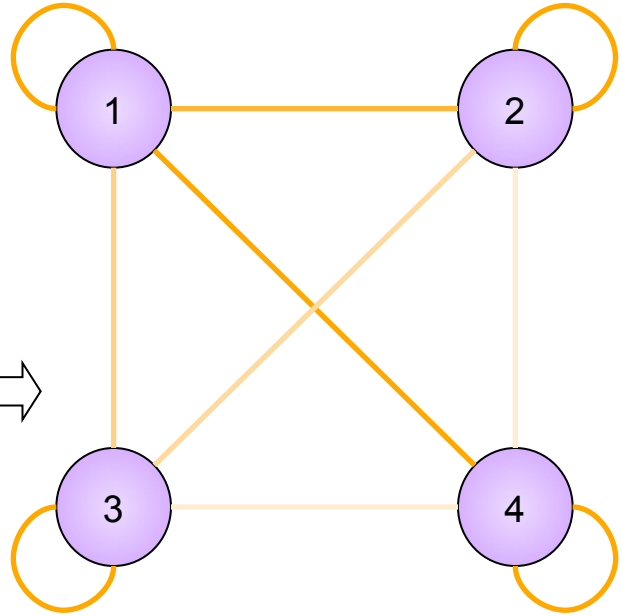
Network prediction: features

- Example 2: sub-cellular localization



http://www.scq.ubc.ca/wp-content/yeasttwohybridtranscript.gif

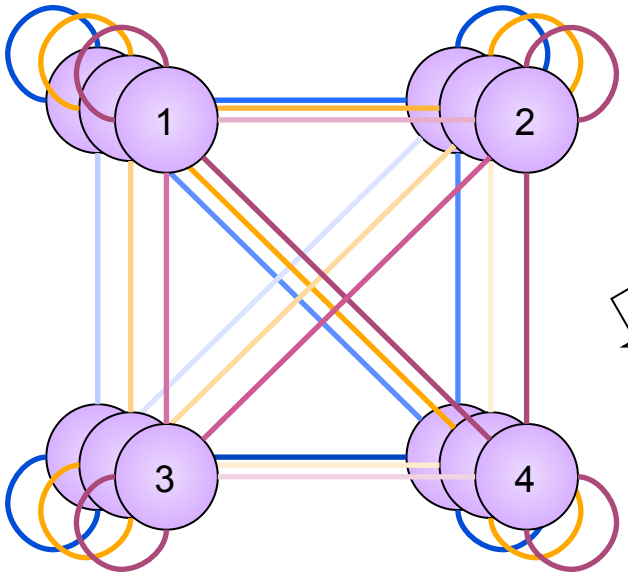
$x_1 = (1, 1, 0, 0, \ldots)$
$x_2 = (1, 1, 1, 0, \ldots)$
$x_3 = (1, 0, 1, 0, \ldots)$
…
$sim(x_1, x_2) = 0.81$
$sim(x_1, x_3) = 0.12$
…

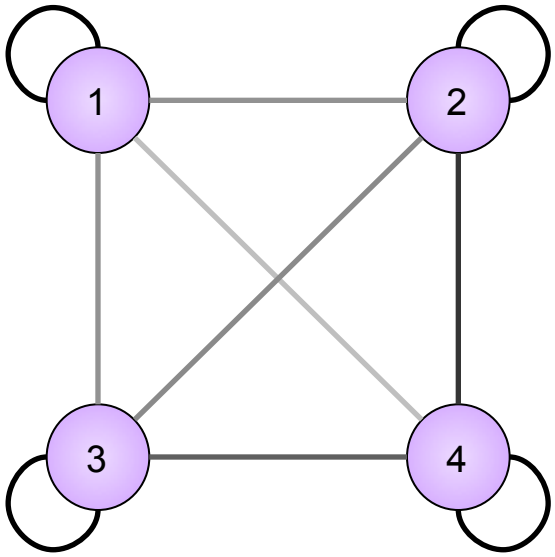


Similarity scale:

1 **-1**

Data integration & Similarity Matrix





	1	2	3	4
1	1.00	0.57	0.55	0.40
2	0.57	1.00	0.66	0.89
3	0.55	0.66	1.00	0.79
4	0.40	0.89	0.79	1.00

Adjacency Matrix

Structure of a Network Data Table (6000 * 5999/2 ~ 18 M pairs)

1st entity	2nd entity	Linked?	Feature 1	Feature 2
1	1	?	0.5	0.99
1	2	N	0.3	0.36
1	3	Y	0.1	0.43
1	4	Y	-1	0.20
...				
2	2	?	2	0.22
2	3	?	3	0.39
...			4	0.58
3	3	Y		
3	4	?	0.2	0.41
3	5	N	0.3	0.70
...				
99	99	Y	0.19	0.38
99	100	?	0.06	0.82
100	100	N	1.00	0.49

Databases

Useful tidbits to keep in mind when building the data table

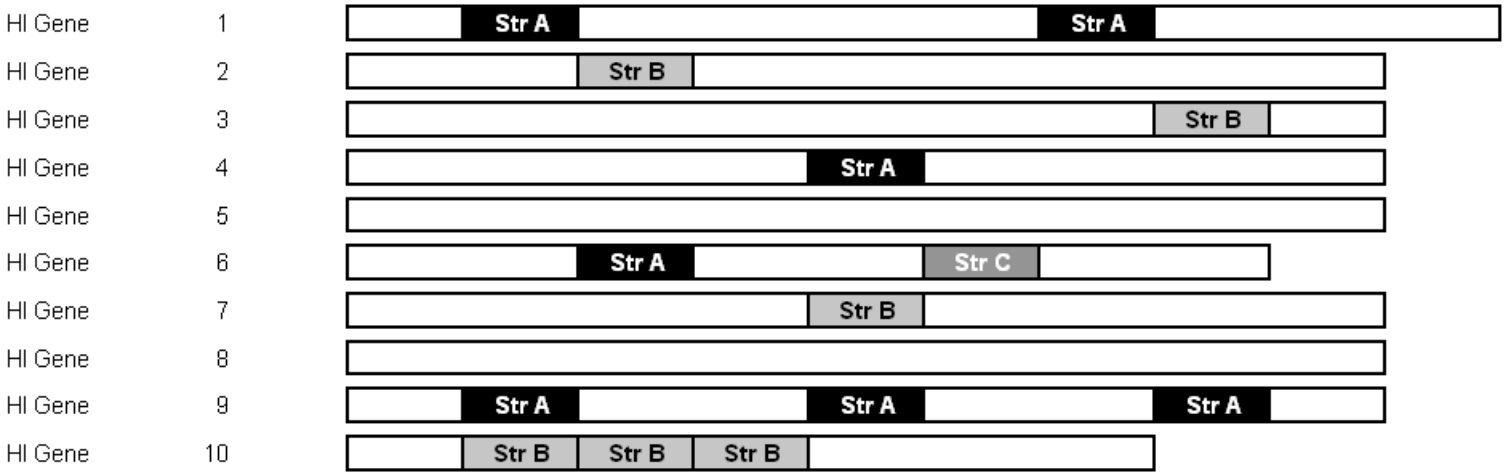
Structured Data

gid_	TrgStrt	TrgStop	did
HI0299	119	135	d193l__
HI0572	180	240	d1aba__
HI0989	56	125	d1aco_1
HI0988	106	458	d1aco_2
HI0154	2	76	d1acp__
HI1633	2	432	d1adea_
HI0349	1	183	d1aky__
HI1309	35	52	d1alo_3
HI0589	8	25	d1alo_3
HI1358	239	444	d1amg_2
HI1358	218	410	d1amy_2
HI0460	20	24	d1ans__
HI1386	139	147	d1ans__
HI0421	11	14	d1ans__
HI0361	285	295	d1ans__
HI0835	100	106	d1ans__

did_	fids
d2rs51_	1.002.007
d1imr__	1.010.002
d1pyib1	1.007.030
d1dxtd_	1.001.001
d181l__	1.004.002
d1vmoa_	1.002.044
d2gsq_1	1.001.031
d1etb2_	1.002.003
d1guha1	1.001.031
d1hrc__	1.001.003
d150lc_	1.004.002
d1dmf__	1.007.035
d1l19__	1.004.002
d1yrnc_	1.010.002
d1apld_	1.001.004
d1ndab2	1.003.004
d2rmai_	1.002.036

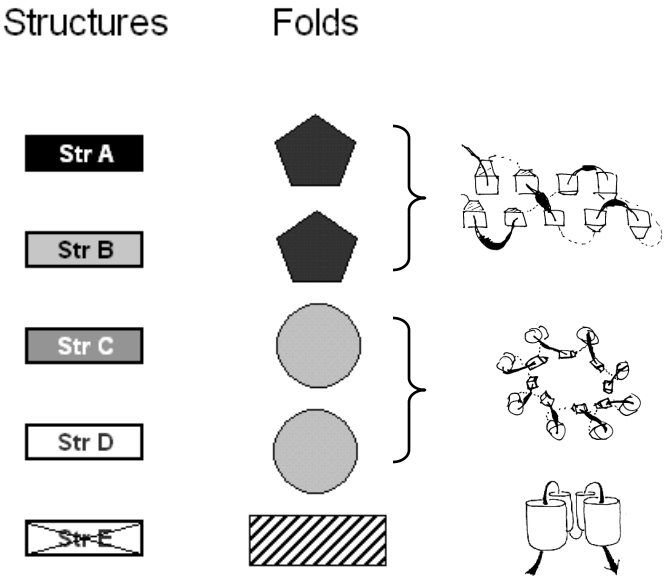
fid_	bestrep	N_minsp	N_scop	objname
1.001.001	d1flp__	8	340	Globin-like
1.001.002	d1hdj__	4	33	Long alpha-hairpin
1.001.003	d1ctj__	9	78	Cytochrome c
1.001.004	d1enh__	18	76	DNA-binding 3-helical bundle
1.001.005	d1dtr_2	1	3	Diphtheria toxin repressor (DtxR) dimeriz
1.001.006	d1tns__	1	2	Mu transposase, DNA-binding domain
1.001.007	d2spca_	1	2	Spectrin repeat unit
1.001.008	d1bdd__	1	4	Immunoglobulin-binding protein A modules
1.001.009	d1bal__	1	5	Peripheral subunit-binding domain of 2-ox
1.001.010	d2erl__	3	5	Protozoan pheromone proteins

Table Interpretation



Match Table: Ways Structures A, B, and C can match HI Genome

Structures have a limited number of folds, which have various characteristics



Structure of a Table

- Row
  - Entity, Tuple, Instance
- Column
  - Field
  - Attribute of an Entity
  - dimension
- Key
  - Certain Attributes (or combination of attributes) can uniquely identify an object, these are keys
- NULL
  - Variant Records

	key	key				
Table	**attr-a**	**attr-b**	**attr-c**	**attr-d**	**attr-e**	**attr-f**
tuple-1	**a1**	**b1**	c1	d1	e1	f1
tuple-2	**a2**	**b2**	c2	d2	e2	f2
tuple-3	**a3**	**b3**	c3	d3	e3	f3
tuple-4	**a4**	**b4**	c4	d4	e4	f4
tuple-5	**a5**	**b5**	c5	d5	e5	f5
tuple-6	**a6**	**b6**	c6	d6		
tuple-7	**a7**	**b7**	c7	d7		f7
tuple-8	**a8**	**b8**	c8	d8	e8	f8
tuple-9	**a9**	**b9**	c9	d9	e9	f9
tuple-10	**a10**	**b10**	c10	d10		f10
tuple-11	**a11**	**b11**	c11	d11	e11	f11
tuple-12	**a12**	**b12**	c12	d12	e12	f12
tuple-13	**a13**	**b13**	c13	d13	e13	f13
tuple-14	**a14**	**b14**	c14	d14	e14	f14

Joins

Foreign Key

Folds

Matches

gid_	TrgStrt	TrgStop	did	score
HI0299	119	135	d193l__	3.1
HI0572	180	240	d1aba__	0.0032
HI0989	56	125	d1aco_1	0.0049
HI0988	106	458	d1aco_2	4.4e-14
HI0154	2	76	d1acp__	1.2e-23
HI1633	2	432	d1adea_	0
HI0349	1	183	d1aky__	7.6e-36
HI1309	35	52	d1alo_3	1.1
HI0589	8	25	d1alo_3	1.8
HI1358	239	444	d1amg_2	0.002
HI1358	218	410	d1amy_2	0.00037
HI0460	20	24	d1ans_	1.8
HI1386	139	147	d1ans__	3.3
HI0421	**11**	**14**	**d1ans__**	**6.4**
HI0361	**285**	**295**	**d1ans__**	**8.2**
HI0835	**100**	**106**	**d1ans__**	**9.7**

Structures

did_	fid
d2rs51_	1.002.007
d1imr__	1.010.002
d1pyib1	1.007.030
d1dxtd_	1.001.001
d181l__	1.004.002
d1vmoa_	1.002.044
d2gsq_1	1.001.031
d1etb2_	1.002.003
d1guha1	1.001.031
d1hrc__	1.001.003
d150lc_	1.004.002
d1dmf__	1.007.035
d1l19__	1.004.002
d1yrnc_	1.010.002
d1ans__	**1.007.008**
d2rmai_	1.002.036

fid_	bestrep	N_hlx	N_beta	name
1.001.001	d1flp__	8	0	Globin-like
1.001.002	d1hdj__	4	0	Long alpha-hairpin
1.001.003	d1ctj__	9	0	Cytochrome c
1.001.004	d1enh__	2	0	DNA-binding 3-helical bundle
1.001.005	d1dtr_2	1	3	Diphtheria toxin repressor (DtxR) dimeriz
1.001.006	d1tns__	1	2	Mu transposase, DNA-binding domain
1.001.007	d2spca_	0	2	Spectrin repeat unit
1.001.008	d1bdd__	0	4	Immunoglobulin-binding protein A modules
1.007.008	**d1qkt__**	**4**	**3**	**Neurotoxin III (ATX III)**
1.001.010	d2erl__	3	5	Protozoan pheromone proteins

SQL

- SIMPLE Language for Building and Querying Tables
- CREATE a table
- INSERT values into it
- SELECT various entries from it (tuples, rows)
- UPDATE the values

- Example: How Many Globin Folds are there in E. coli versus Yeast?

SQL Select on a Single Table

	key	key				
Table	attr-a	attr-b	attr-c	attr-d	attr-e	attr-f
tuple-1	a1	b1	c1	d1	e1	f1
tuple-2	a2	b2	c2	d2	e2	f2
tuple-3	a3	b3	c3	d3	e3	f3
tuple-4	a4	b4	c4	d4	e4	f4
tuple-5	a5	b5	c5	d5	e5	f5
tuple-6	a6	b6	c6	d6		
tuple-7	a7	b7	c7	d7		f7
tuple-8	a8	b8	c8	d8	e8	f8
tuple-9	a9	b9	c9	d9	e9	f9
tuple-10	a10	b10	c10	d10		f10
tuple-11	a11	b11	c11	d11	e11	f11
tuple-12	a12	b12	c12	d12	e12	f12
tuple-13	a13	b13	c13	d13	e13	f13
tuple-14	a14	b14	c14	d14	e14	f14

- Select {columns} from {a table} where {row-selection is true}
- projection of a selection
- Sort result on a attribute
- Selection as an array lookup
  - $fid=$structure{$did}

SQL Select on Multiple Tables

Matches | **Structures**

	key	key						
Table 1	gid	TrgStrt	TrgStop	did		Table 2	did	fid
tuple-1	HI001	12	200	d1mbd__		tuple-i	d1lfg_1	1.007.006
tuple-2	HI002	15	231	d1hhba_		tuple-i	d1lfg_1	1.007.006
tuple-3	HI002	100	343	**d1lfg_1**		**tuple-i**	**d1lfg_1**	**1.007.006**
tuple-4	HI003	12	80	**d1lfg_1**		**tuple-i**	**d1lfg_1**	**1.007.006**
tuple-5	HI009	200	260	d1mba__		tuple-i	d1lfg_1	1.007.006
tuple-6	HI023	300	450	d2ubx__		tuple-i	d1lfg_1	1.007.006
tuple-7	HI045	2	89	d2lmg__		tuple-i	d1lfg_1	1.007.006
tuple-1	HI001	12	200	d1mbd__		tuple-ii	d1mba__	1.003.002
tuple-2	HI002	15	231	d1hhba_		tuple-ii	d1mba__	1.003.002
tuple-3	HI002	100	343	d1lfg_1		tuple-ii	d1mba__	1.003.002
tuple-4	HI003	12	80	d1lfg_1		tuple-ii	d1mba__	1.003.002
tuple-5	HI009	200	**260**	**d1mba__**		**tuple-ii**	**d1mba__**	**1.003.002**
tuple-6	HI023	300	450	d2ubx__		tuple-ii	d1mba__	1.003.002
tuple-7	HI045	2	89	d2lmg__		tuple-ii	d1mba__	1.003.002

- Select {columns} from {huge cross-product of tables} where {row-selection is true}
  - cross-product T(1) x T(2) builds a huge virtual table where every row of T(1) is paired with every row of T(2). Then perform selection on this.
- Select fid from matches,structures where gid=HI009 and matches.did = structures.did
- Joining as a double-lookup
  - ($bestrep, $N_hlx, $N_beta, $name) = $folds{ $structures{$did} }

ER-diagrams

Start gid structure

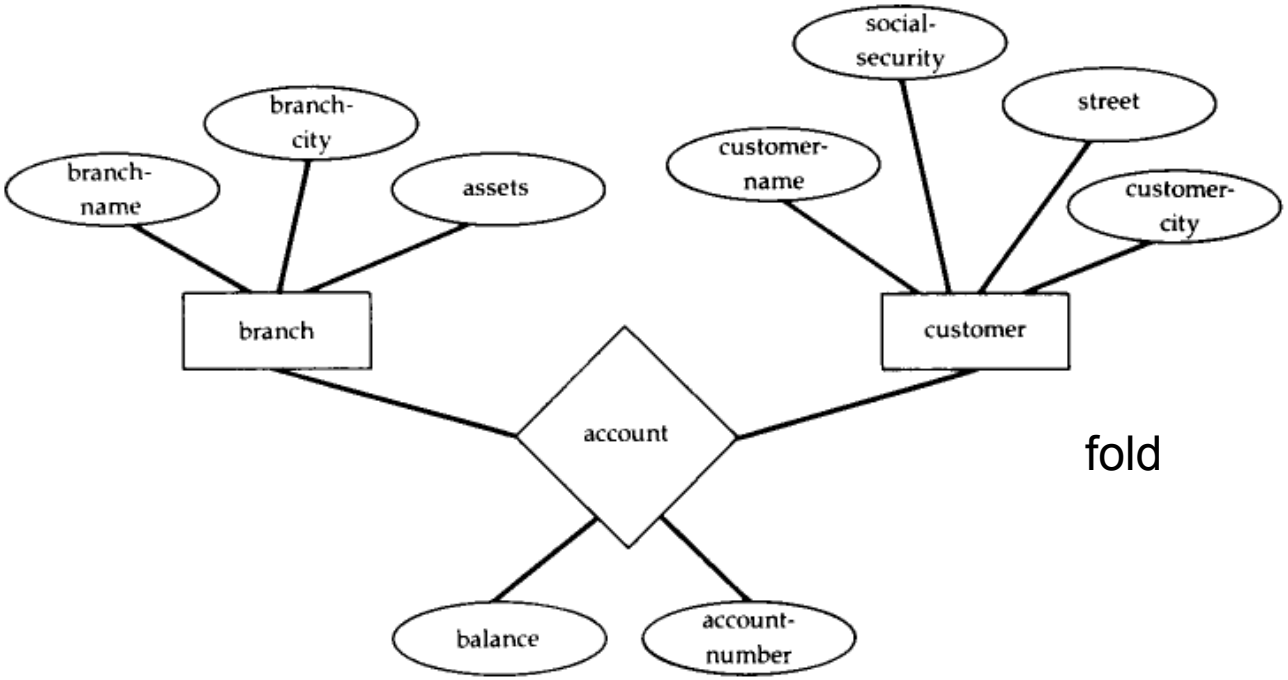


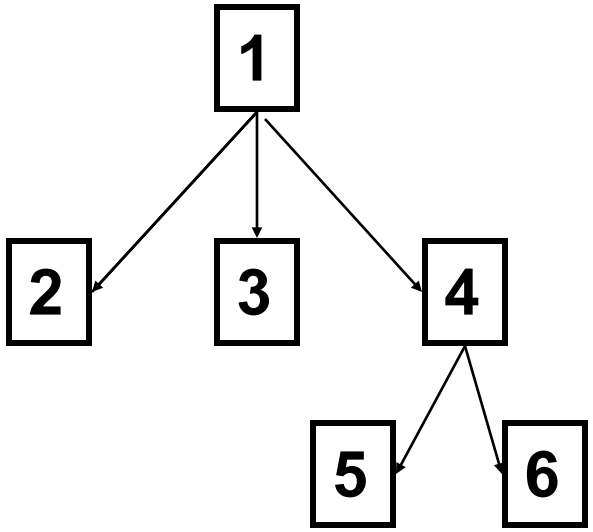
Figure 2.23 E-R diagram with *account* as a relationship set.

fold

- Korth & Silberschatz
  - branch <=> matches (gid-start +++ did)
  - customer <=> folds (fid +++)
  - linked by
    account <=> structures (did fid)

Complex Data Example: Encoding Trees in RDBs

Node	Parent
1	0
2	1
3	1
4	1
5	4
6	4



Node	Name
1	Organism
2	Bacteria
3	Archea
4	Eukarya
5	Metazoa
6	Plants

Join Gives Unnormalized Table

Joining Two or More Tables with a Select Query Gives a New, "Bigger" Table

gid_	TrgStrt	TrgStop	did	score	fid	N_hlx	N_beta	name
HI0299	119	135	d193l__	3.1	1.010.002	0	2	Spectrin repeat unit
HI0572	180	240	d1aba__	0.0032	1.002.045	1	2	Mu transposase, DNA-binding domain
HI0989	56	125	d1aco_1	0.0049	1.001.031	8	0	Globin-like
HI0988	106	458	d1aco_2	4.4e-14	1.001.031	8	0	Globin-like
HI0154	2	76	d1acp__	1.2e-23	1.001.031	8	0	Globin-like
HI1633	2	432	d1adea_	0	1.010.002	0	2	Spectrin repeat unit
HI0349	1	183	d1aky__	7.6e-36	1.001.031	8	0	Globin-like
HI1309	35	52	d1alo_3	1.1	1.007.008	4	3	Neurotoxin III (ATX III)
HI0589	8	25	d1alo_3	1.8	1.002.045	1	2	Mu transposase, DNA-binding domain
HI1358	239	444	d1amg_2	0.002	1.004.002	1	3	Diphtheria toxin repressor (DtxR)
HI1358	218	410	d1amy_2	0.00037	1.002.044	0	4	Immunoglobulin-binding protein A
HI0460	20	24	d1ans__	1.8	1.007.008	4	3	Neurotoxin III (ATX III)
HI1386	139	147	d1ans__	3.3	1.007.008	4	3	Neurotoxin III (ATX III)
HI0421	11	14	d1ans__	6.4	1.007.008	4	3	Neurotoxin III (ATX III)
HI0361	285	295	d1ans__	8.2	1.007.008	4	3	Neurotoxin III (ATX III)
HI0835	100	106	d1ans__	9.7	1.007.008	4	3	Neurotoxin III (ATX III)

Normalization

- What if Want to update Fold 1.007.008 to be "Neurotoxin IV"?
  - Many Updates
- So Good if Previously Normalized into Separate Tables
  - Eliminate Redundancy
  - Allow Consistent Updating

gid_	TrgStrt	TrgStop	did	score	fid	N_hlx	N_beta	name
HI0299	119	135	d193l__	3.1	1.010.002	0	2	Spectrin repeat unit
HI0572	180	240	d1aba__	0.0032	1.002.045	1	2	Mu transposase, DNA-binding domain
HI0989	56	125	d1aco_1	0.0049	1.001.031	8	0	Globin-like
HI0988	106	458	d1aco_2	4.4e-14	1.001.031	8	0	Globin-like
HI0154	2	76	d1acp__	1.2e-23	1.001.031	8	0	Globin-like
HI1633	2	432	d1adea_	0	1.010.002	0	2	Spectrin repeat unit
HI0349	1	183	d1aky__	7.6e-36	1.001.031	8	0	Globin-like
HI1309	35	52	d1alo_3	1.1	1.007.008	4	3	**Neurotoxin III** (ATX III)
HI0589	8	25	d1alo_3	1.8	1.002.045	1	2	Mu transposase, DNA-binding domain
HI1358	239	444	d1amg_2	0.002	1.004.002	1	3	Diphtheria toxin repressor (DtxR)
HI1358	218	410	d1amy_2	0.00037	1.002.044	0	4	Immunoglobulin-binding protein A
HI0460	20	24	d1ans__	1.8	1.007.008	4	3	**Neurotoxin III** (ATX III)
HI1386	139	147	d1ans__	3.3	1.007.008	4	3	**Neurotoxin III** (ATX III)
HI0421	11	14	d1ans__	6.4	1.007.008	4	3	**Neurotoxin III** (ATX III)
HI0361	285	295	d1ans__	8.2	1.007.008	4	3	**Neurotoxin III** (ATX III)
HI0835	100	106	d1ans__	9.7	1.007.008	4	3	**Neurotoxin III** (ATX III)

Normalization Example

Un-normalized

Name	City	Area-Code	Phone-Number
Charles	NY	212	345-6789
Mark	SF	415	236-8982
Jane	NY	212	567-2345
Jeff	SF	415	435-3535
Jack	Boston	617	234-9988

Normalized

Name	City	Phone-Number
Charles	NY	345-6789
Mark	SF	236-8982
Jane	NY	567-2345
Jeff	SF	435-3535
Jack	Boston	234-9988

City	Area-Code
NY	212
SF	415
Boston	617

Normalized Tables

Theory of Normaliz-ation

gid_	TrgStrt	TrgStop	did	score
HI0299	119	135	d193l__	3.1
HI0572	180	240	d1aba__	0.0032
HI0989	56	125	d1aco_1	0.0049
HI0988	106	458	d1aco_2	4.4e-14
HI0154	2	76	d1acp__	1.2e-23
HI1633	2	432	d1adea_	0
HI0349	1	183	d1aky__	7.6e-36
HI1309	35	52	d1alo_3	1.1
HI0589	8	25	d1alo_3	1.8
HI1358	239	444	d1amg_2	0.002
HI1358	218	410	d1amy_2	0.00037
HI0460	20	24	d1ans__	1.8
HI1386	139	147	d1ans__	3.3
HI0421	**11**	**14**	**d1ans__**	**6.4**
HI0361	**285**	**295**	**d1ans__**	**8.2**
HI0835	**100**	**106**	**d1ans__**	**9.7**

did_	fid
d2rs51_	1.002.007
d1imr__	1.010.002
d1pyib1	1.007.030
d1dxtd_	1.001.001
d181l__	1.004.002
d1vmoa_	1.002.044
d2gsq_1	1.001.031
d1etb2_	1.002.003
d1guha1	1.001.031
d1hrc__	1.001.003
d150lc_	1.004.002
d1dmf__	1.007.035
d1l19__	1.004.002
d1yrnc_	1.010.002
d1ans__	**1.007.008**
d2rmai_	1.002.036

fid_	bestrep	N_hlx	N_beta	name
1.001.001	d1flp__	8	0	Globin-like
1.001.002	d1hdj__	4	0	Long alpha-hairpin
1.001.003	d1ctj__	9	0	Cytochrome c
1.001.004	d1enh__	2	0	DNA-binding 3-helical bundle
1.001.005	d1dtr_2	1	3	Diphtheria toxin repressor (DtxR) dimeriz
1.001.006	d1tns__	1	2	Mu transposase, DNA-binding domain
1.001.007	d2spca_	0	2	Spectrin repeat unit
1.001.008	d1bdd__	0	4	Immunoglobulin-binding protein A modules
1.007.008	**d1qkt__**	**4**	**3**	**Neurotoxin III (ATX III)**
1.001.010	d2erl__	3	5	Protozoan pheromone proteins

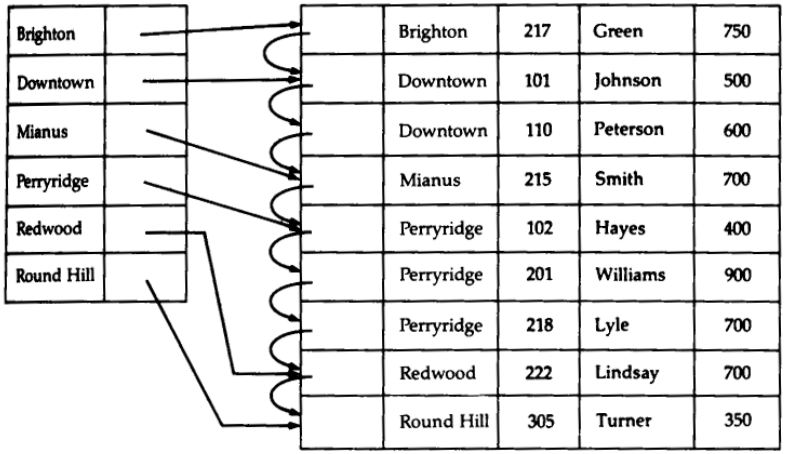
Query Optimization

- Get at the Data Quickly!!
- Indexes
- Hash Function Reproduce the Effect of Indexes
  - Rapidly Associate a Bucket with Each Key
- Joining 10 tables, which to do first?
  - Joining is slow so store some tables in unnormalized form
    - Speed vs Memory

Indexes Speed Access

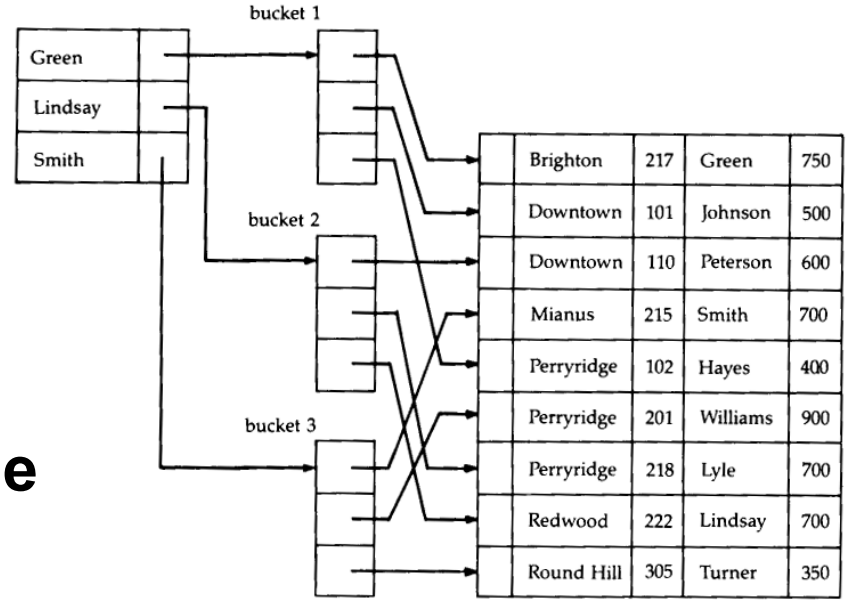
	Brighton	217	Green	750
	Downtown	101	Johnson	500
	Downtown	110	Peterson	600
	Mianus	215	Smith	700
	Perryridge	102	Hayes	400
	Perryridge	201	Williams	900
	Perryridge	218	Lyle	700
	Redwood	222	Lindsay	700
	Round Hill	305	Turner	350

No Index



Brighton		Brighton	217	Green	750
Downtown		Downtown	101	Johnson	500
Mianus		Downtown	110	Peterson	600
Perryridge		Mianus	215	Smith	700
Redwood		Perryridge	102	Hayes	400
Round Hill		Perryridge	201	Williams	900
		Perryridge	218	Lyle	700
		Redwood	222	Lindsay	700
		Round Hill	305	Turner	350

One Index

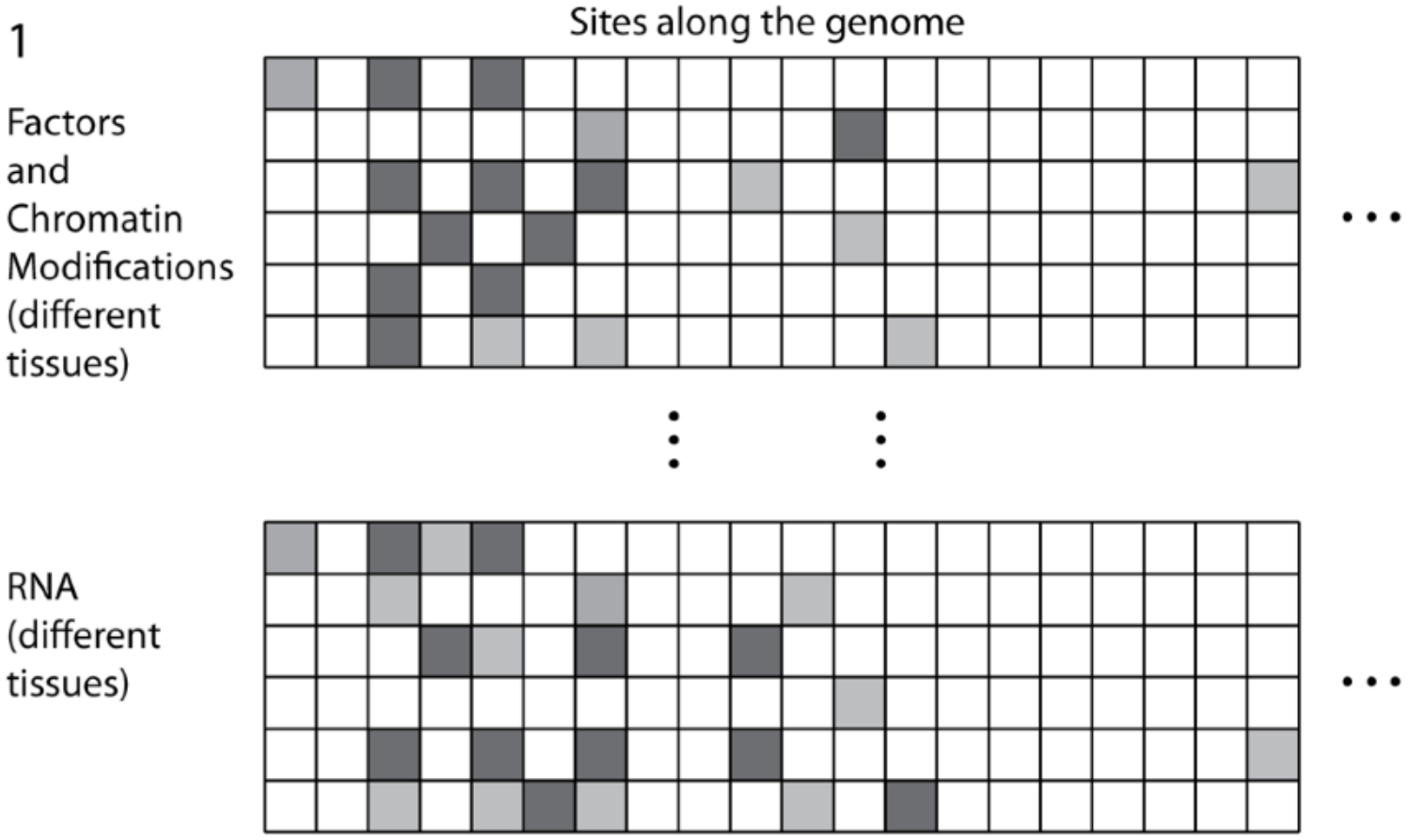


Double Index

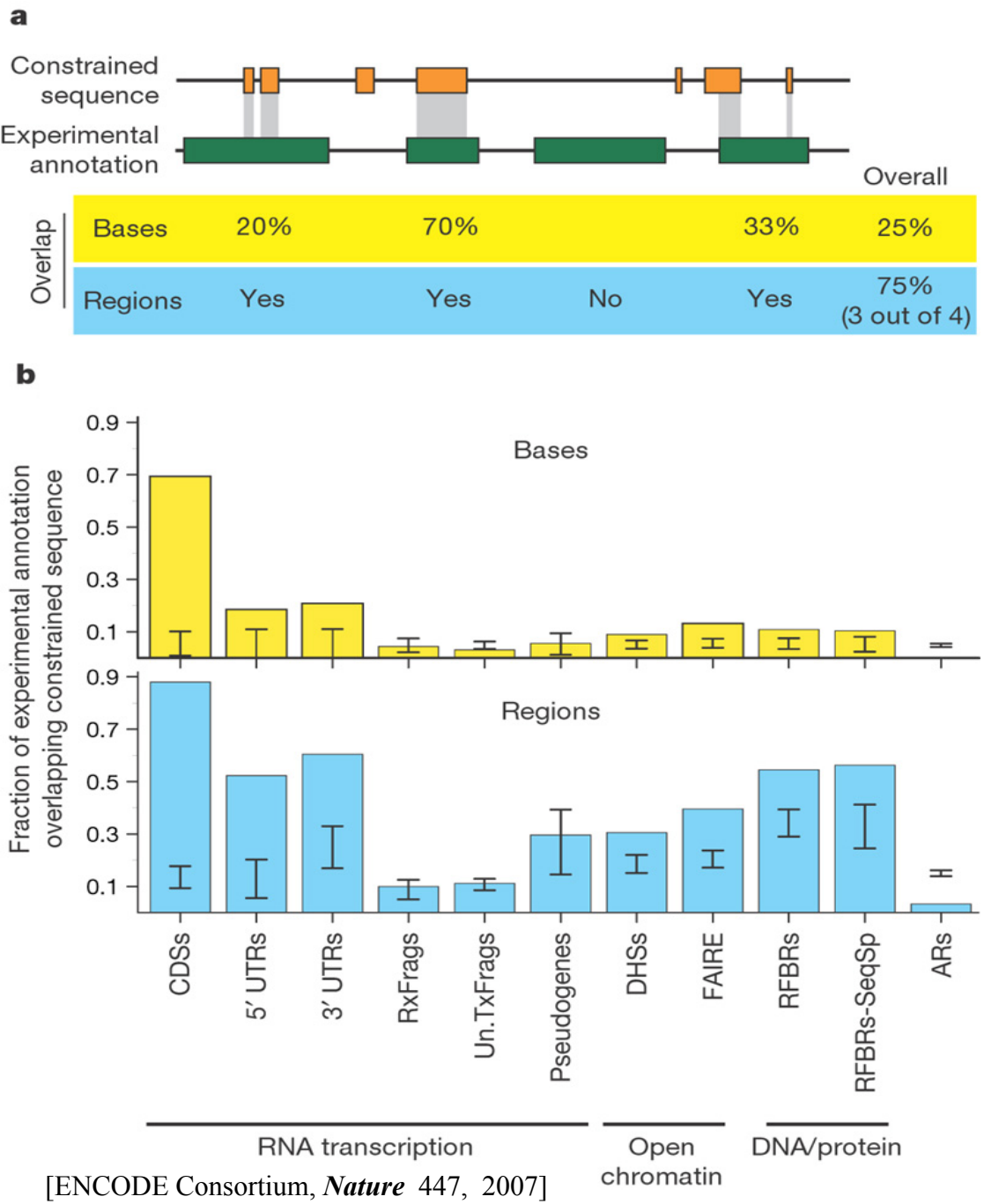
Unsupervised Mining

Simple "Overlaps"

Structure of Genomic Features Matrix



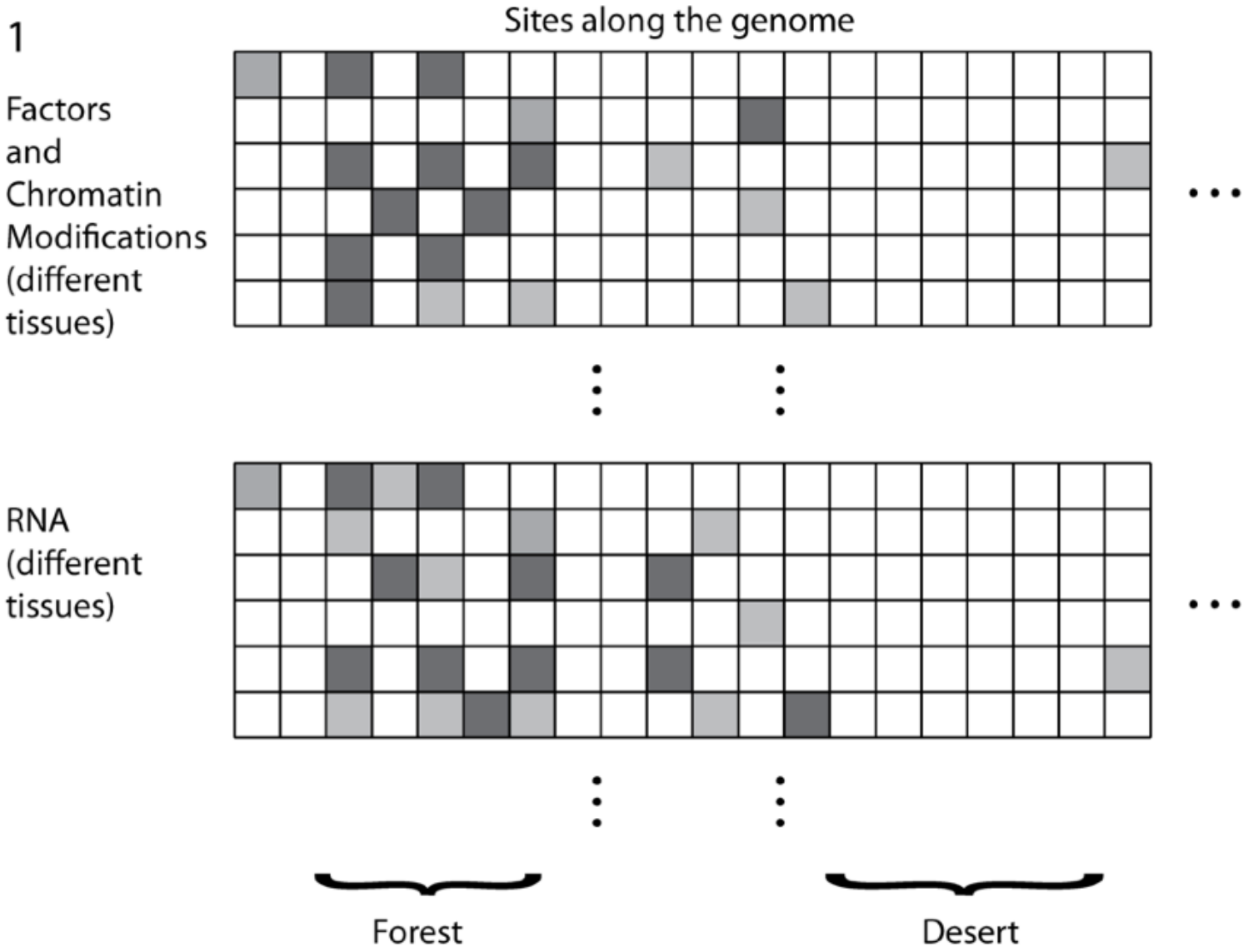
Example Overlap Analysis: Biochemically Active Regions Don't all Appear to be Under Constraint



[ENCODE Consortium, ***Nature*** 447, 2007]

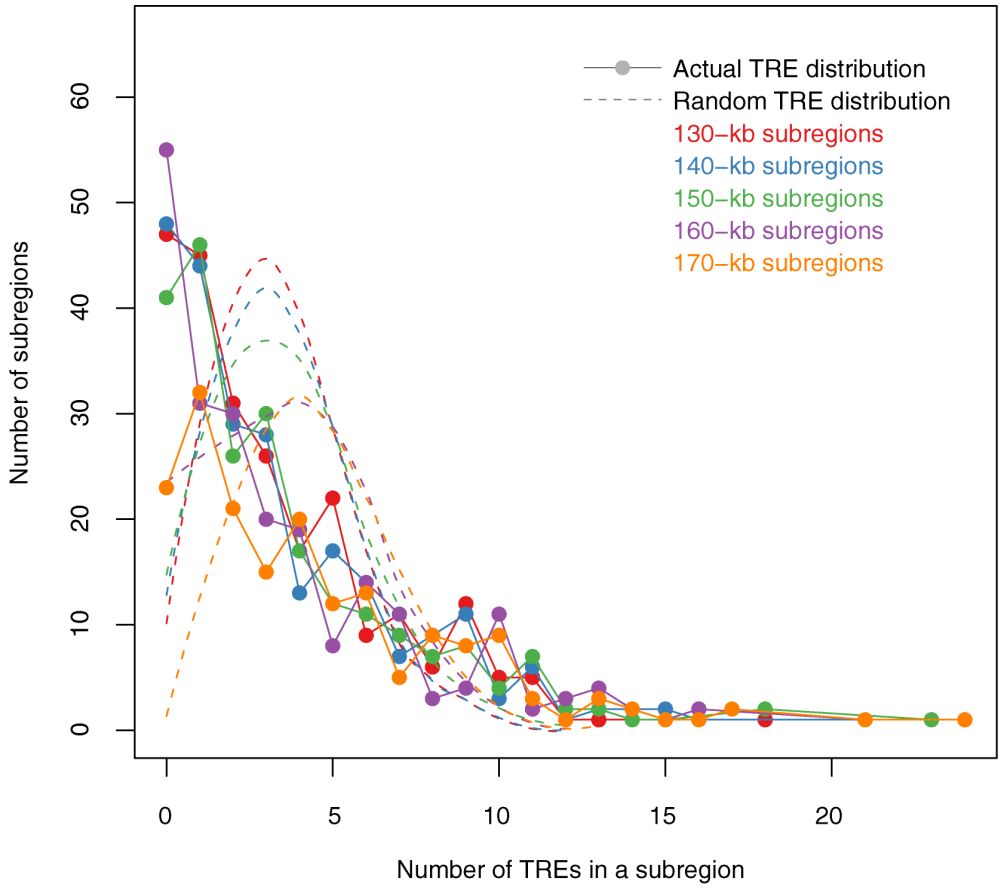
- Careful Randomization (GSC statistic)

Genomic Features Matrix: Deserts & Forests



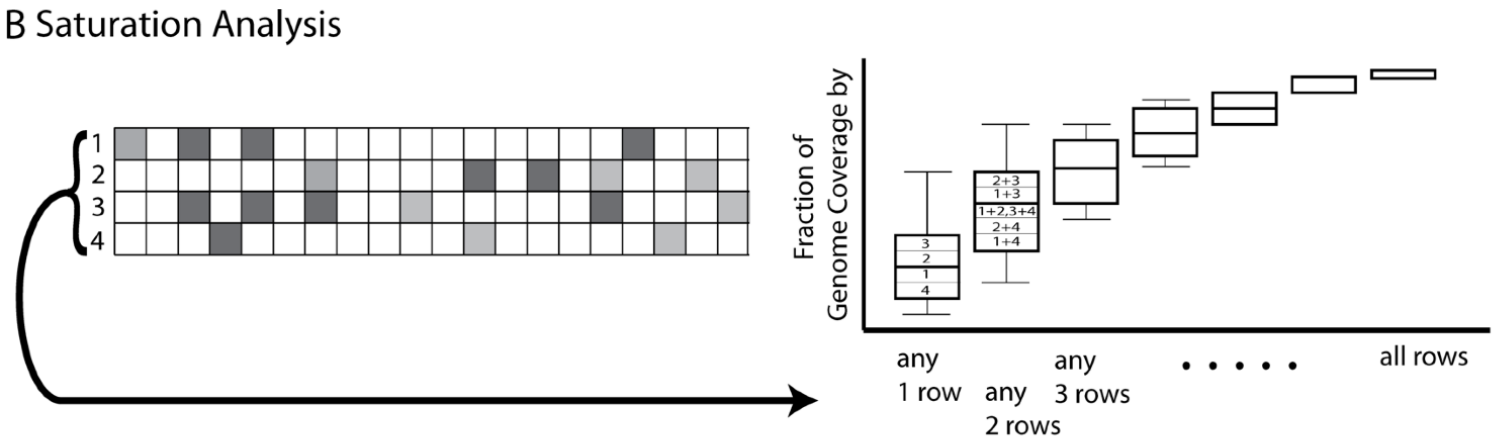
Non-random distribution of TREs

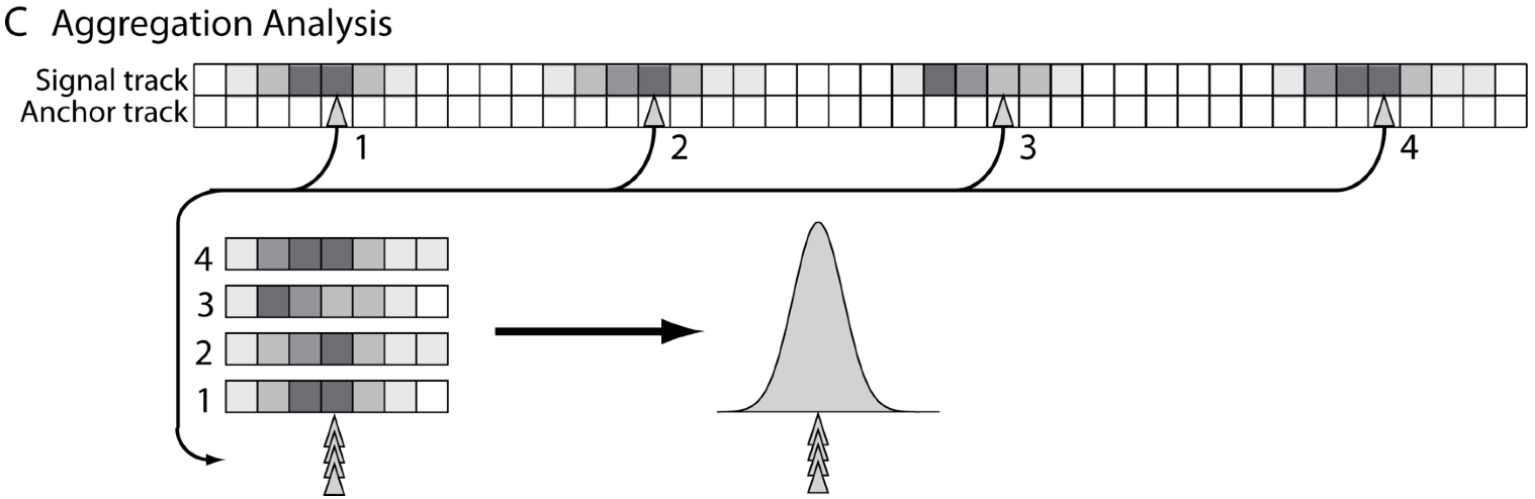
- TREs are not evenly distributed throughout the encode regions ($P < 2.2\times10^{-16}$).
- The actual TRE distribution is power-law.
- The null distribution is 'Poissonesque.'
- Many genomic subregions with extreme numbers of TREs.



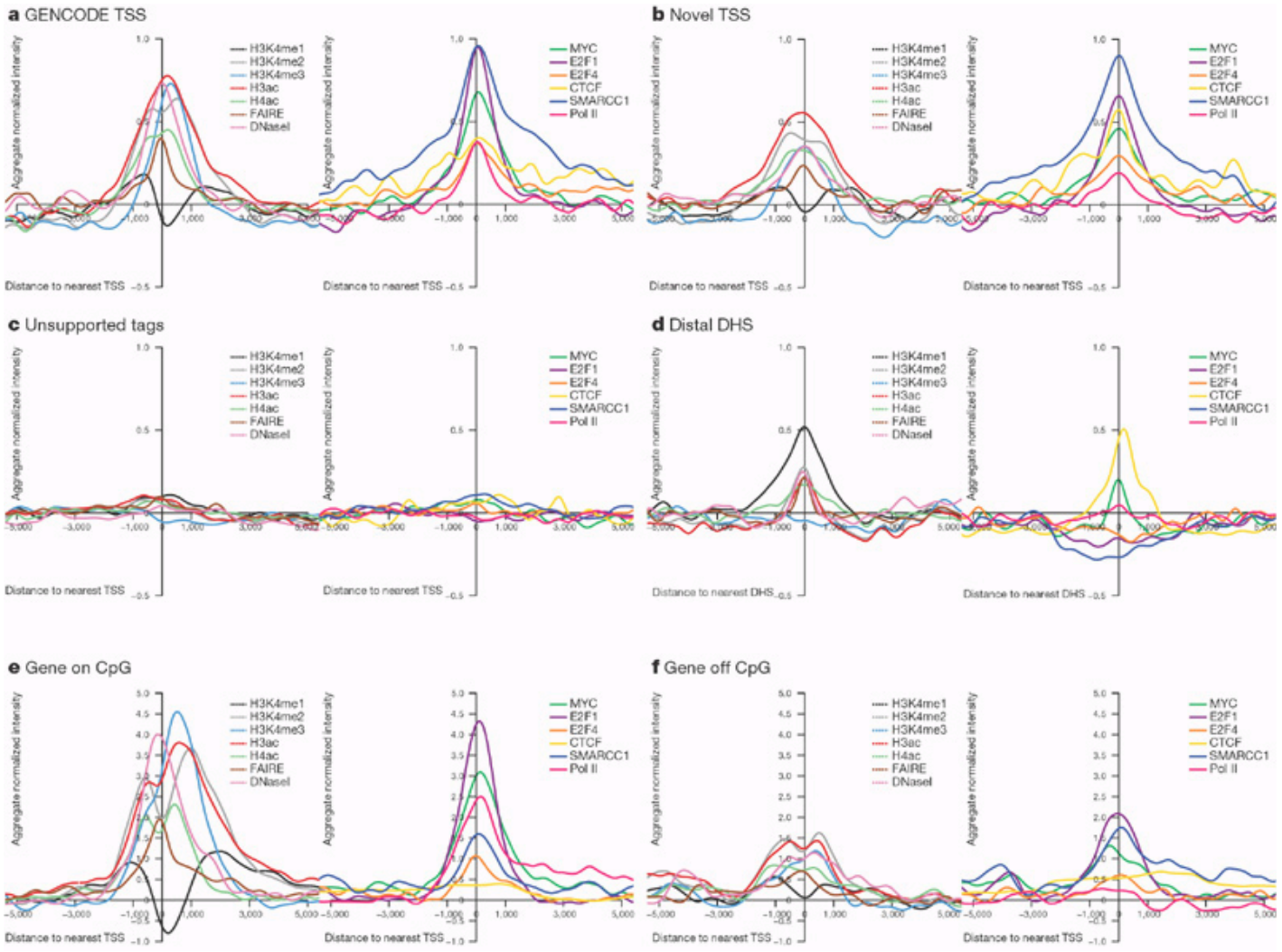
Zhang et al. (2007) Gen. Res.

Aggregation & Saturation





Aggregation Analysis

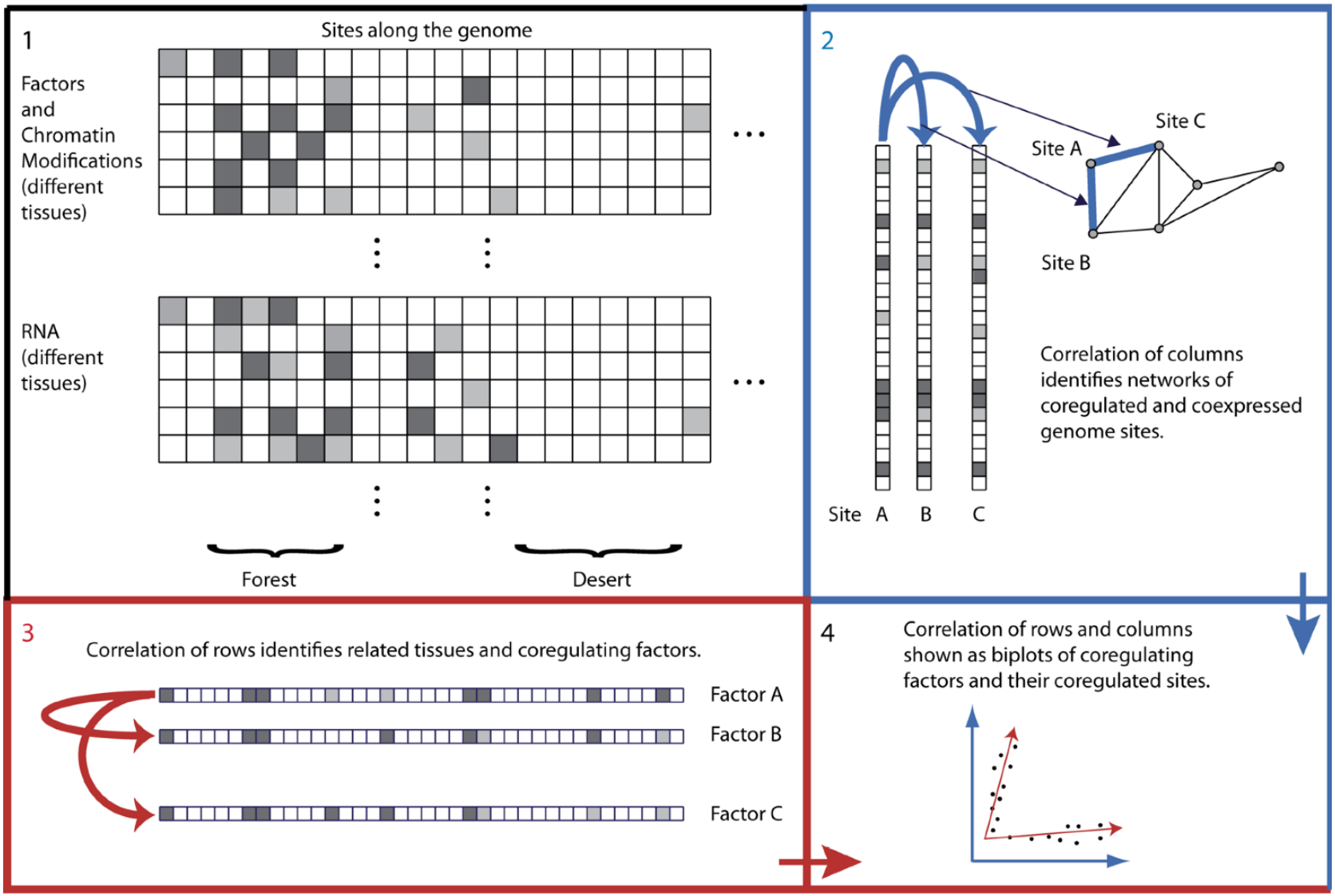


[ENCODE Consortium, ***Nature*** 447, 2007]

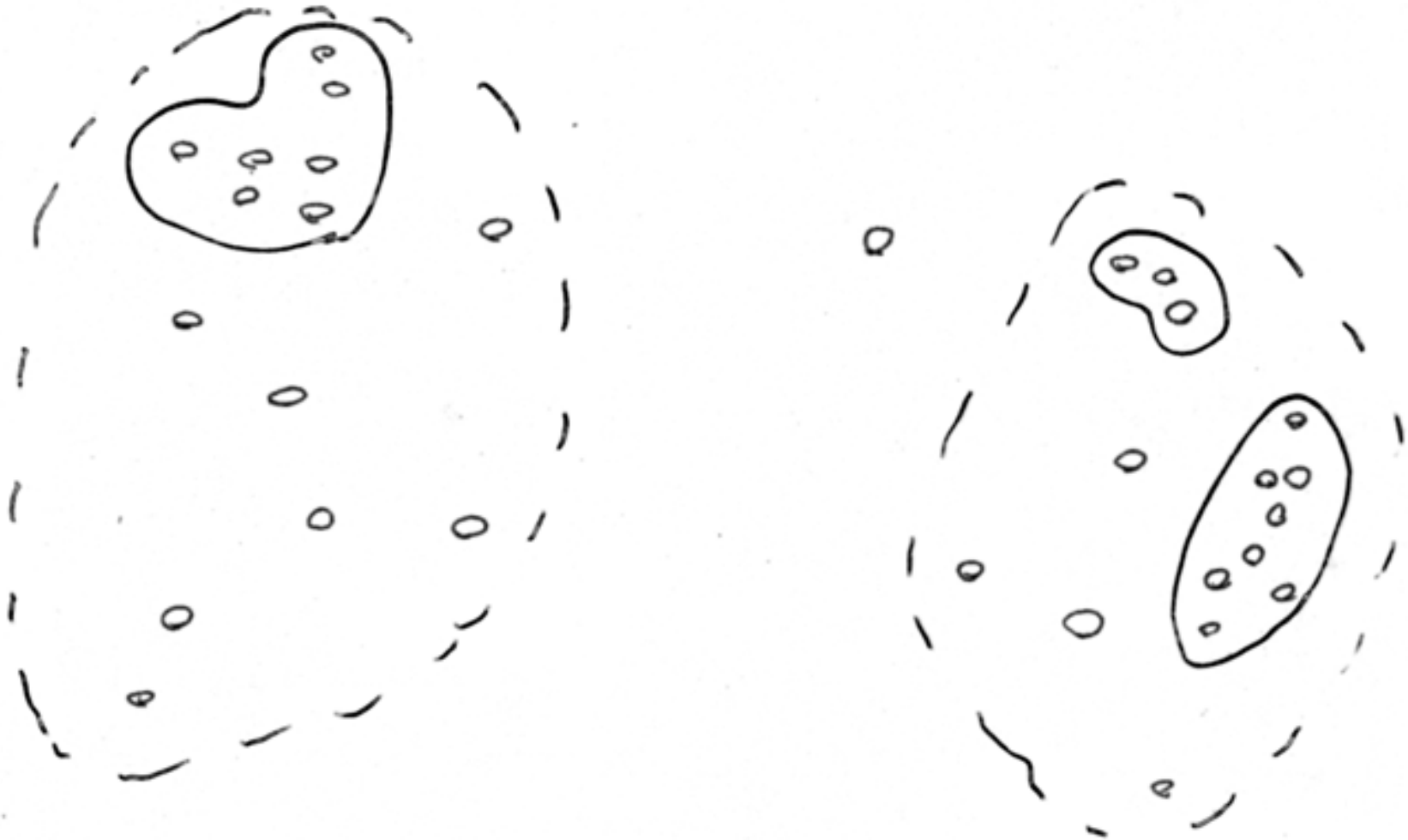
Unsupervised Mining

Clustering Columns & Rows of the Data Matrix

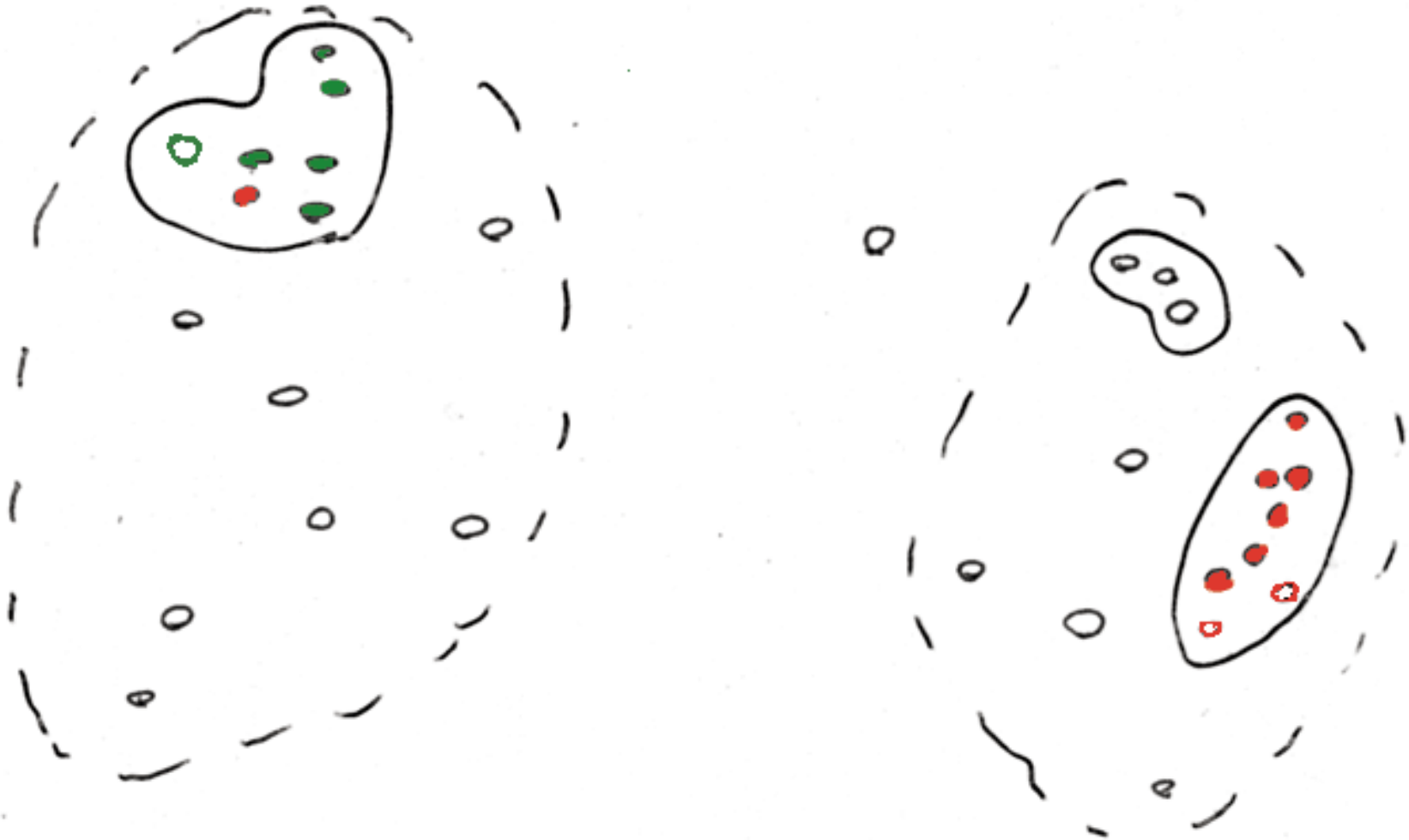
Correlating Rows & Columns



“cluster” predictors

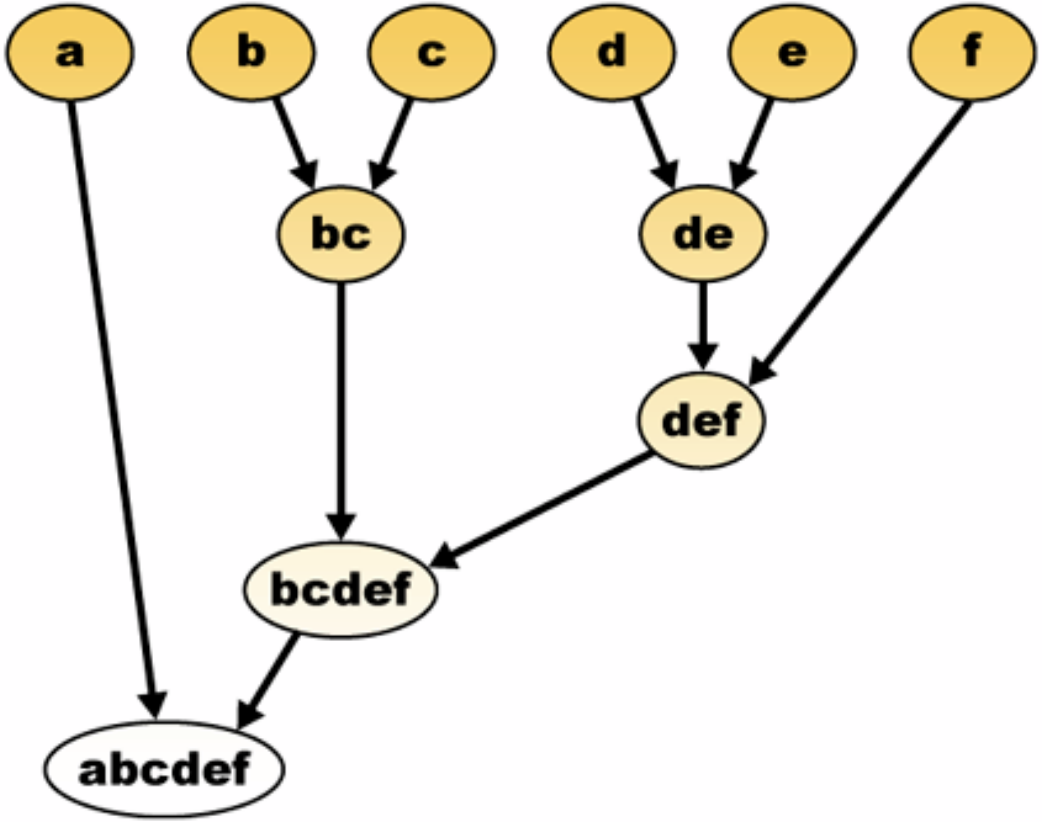


Use clusters to predict Response



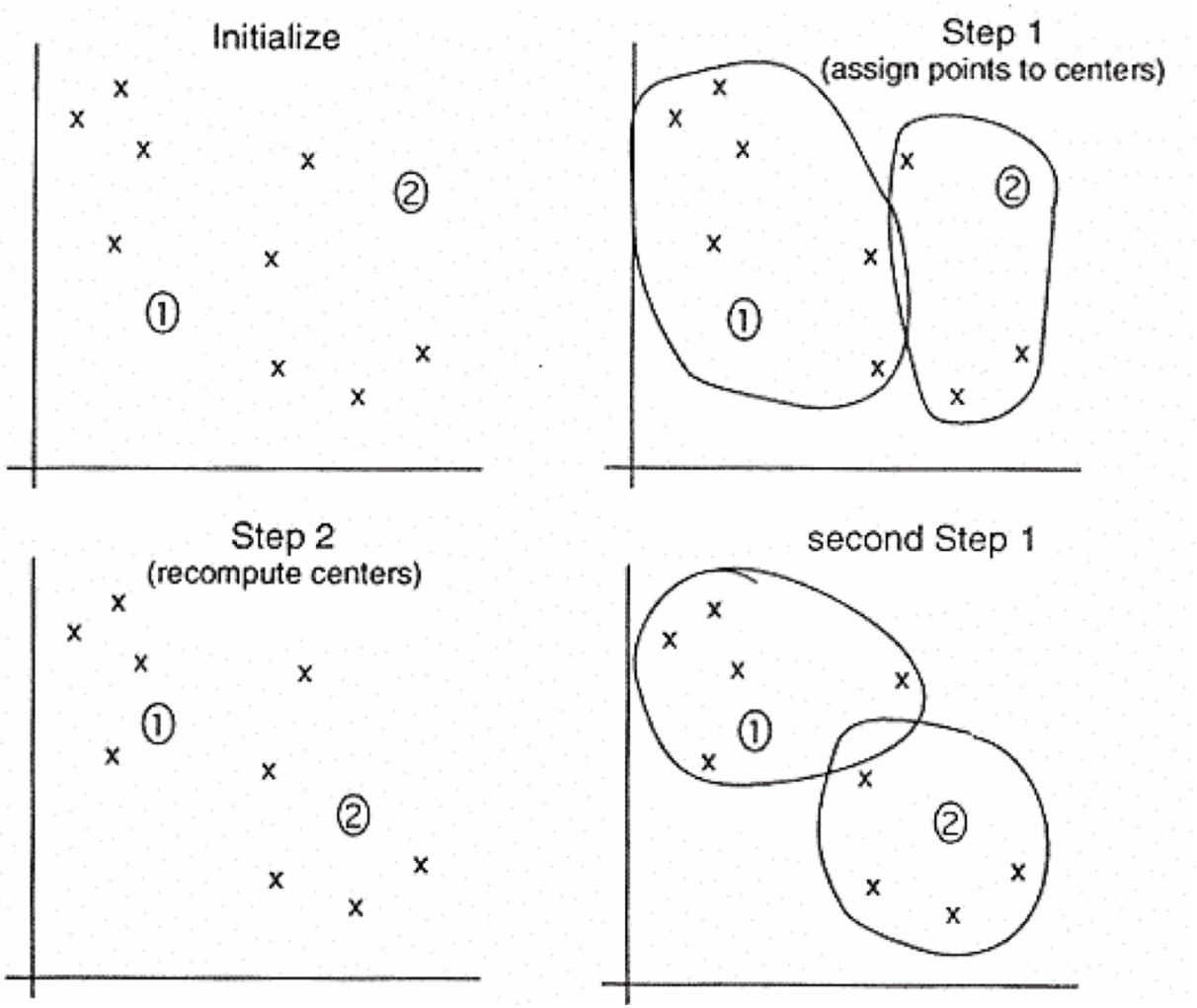
Agglomerative Clustering

- Bottom up
  v top down
  (K-means, know how many centers)
- Single or multi-link
  - threshold for connection?



http://commons.wikimedia.org/wiki/File:Hierarchical_clustering_diagram.png

K-means

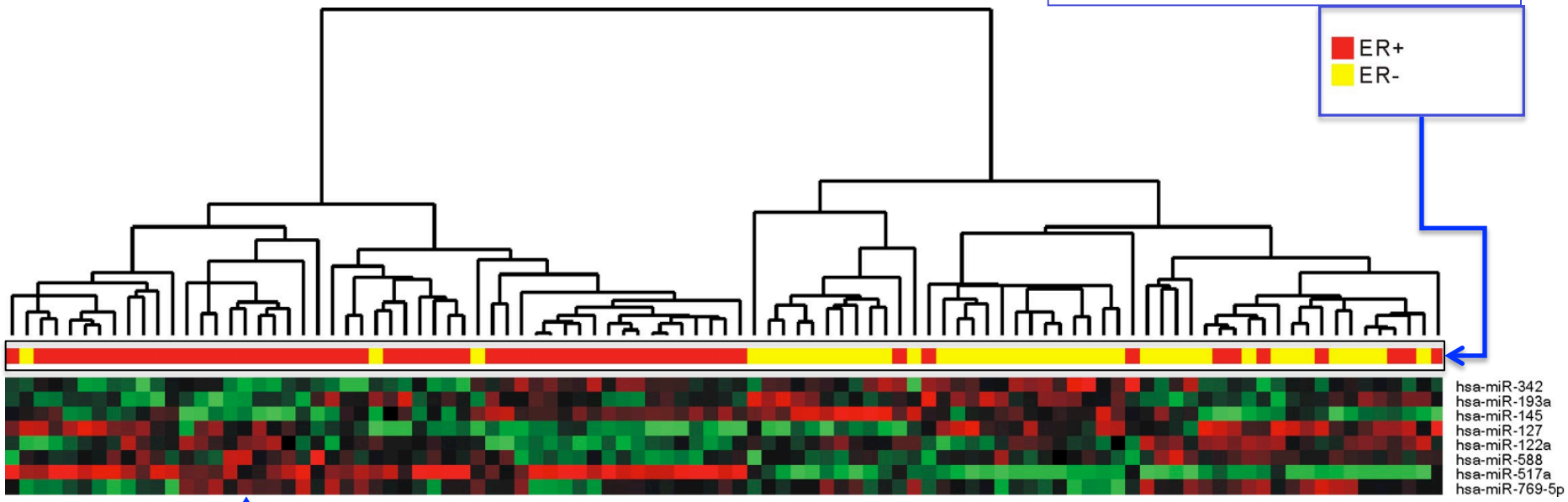


1) Pick ten (i.e. k?) random points as putative cluster centers.
2) Group the points to be clustered by the center to which they are closest.
3) Then take the mean of each group and repeat, with the means now at the cluster center.
4)Stop when the centers stop moving.

Using Genes to Cluster Cancer Samples

(3) Clustering based on RE score divides samples into 2 main types of cancer

(4) Clustering better than based on indiv. gene expression levels

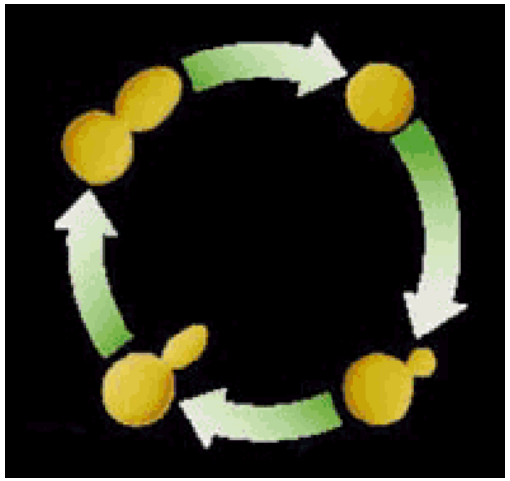


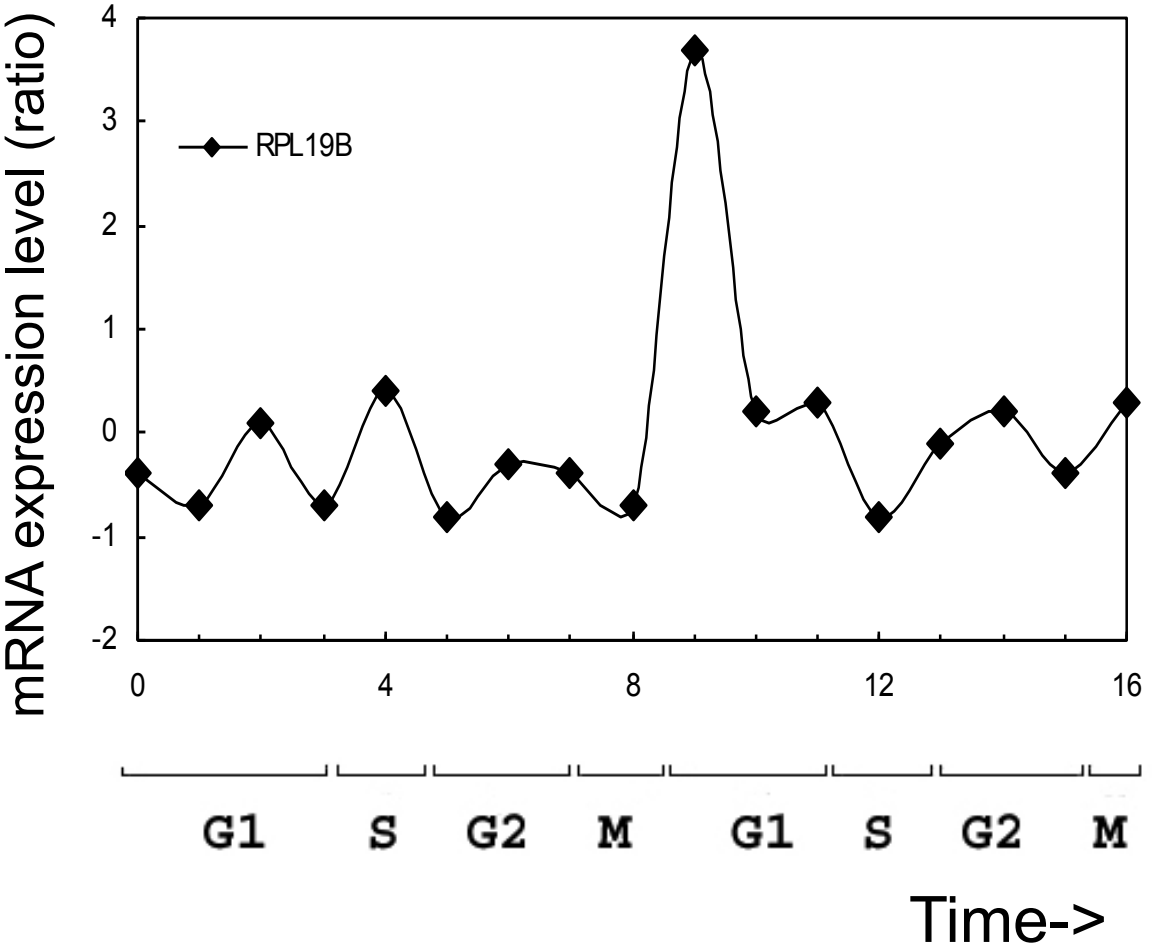
(1) RE-score profile for diff. miRNA in 1 cancer sample.
(2) Tabulate over many different breast cancer samples

Cheng et al., Genome Biology, 2009

Clustering the yeast cell cycle to uncover interacting proteins

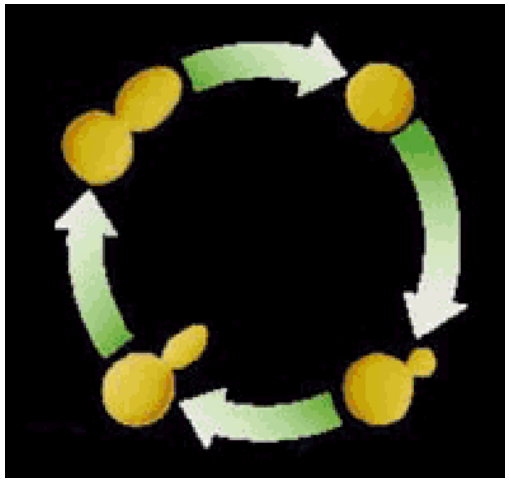
[Brown, Davis]

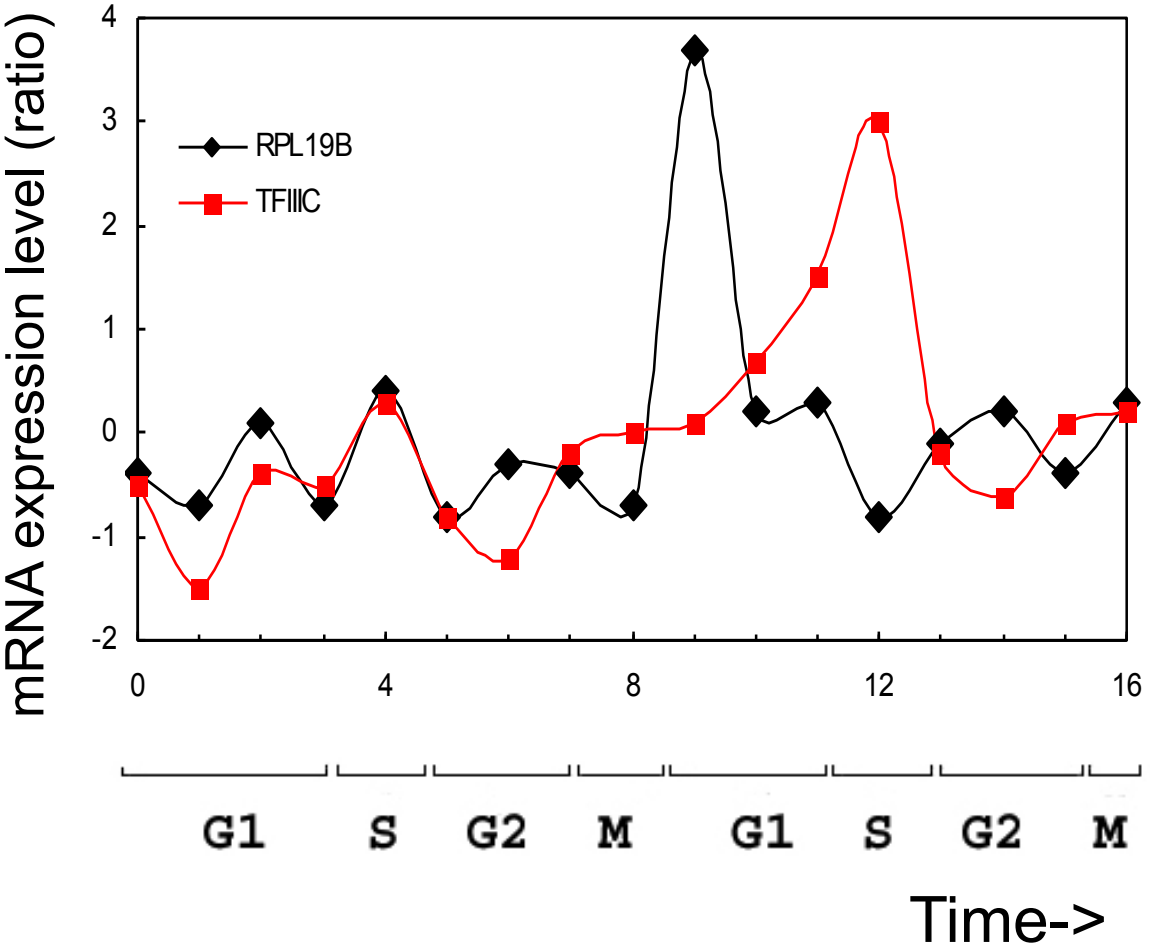




Microarray timecourse of
1 ribosomal protein

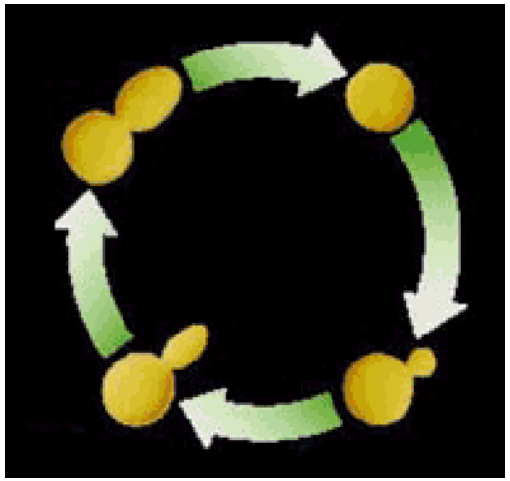
Clustering the yeast cell cycle to uncover interacting proteins

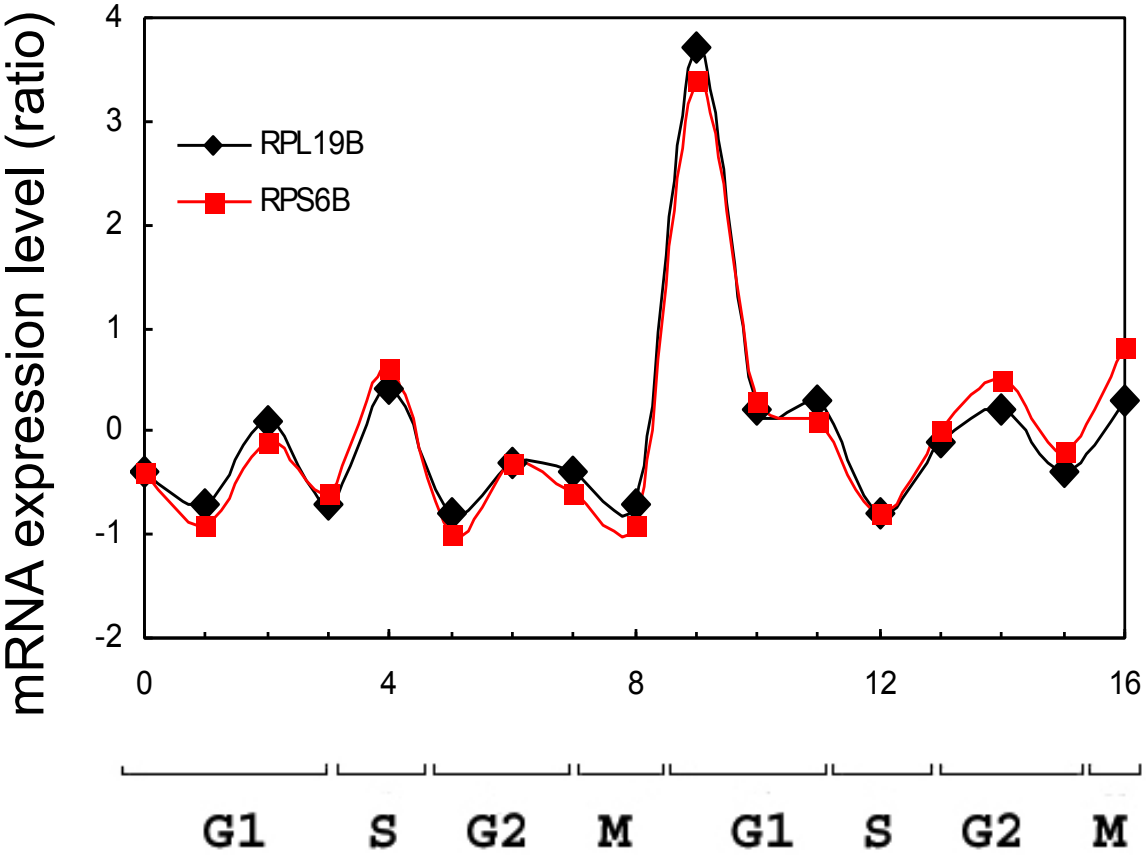




Clustering the yeast cell cycle to uncover interacting proteins

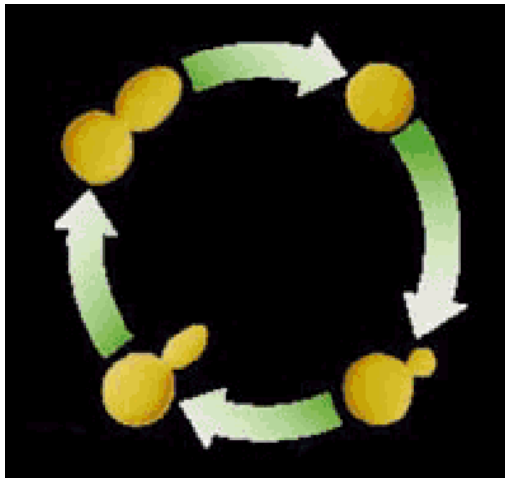
[Botstein; Church, Vidal]

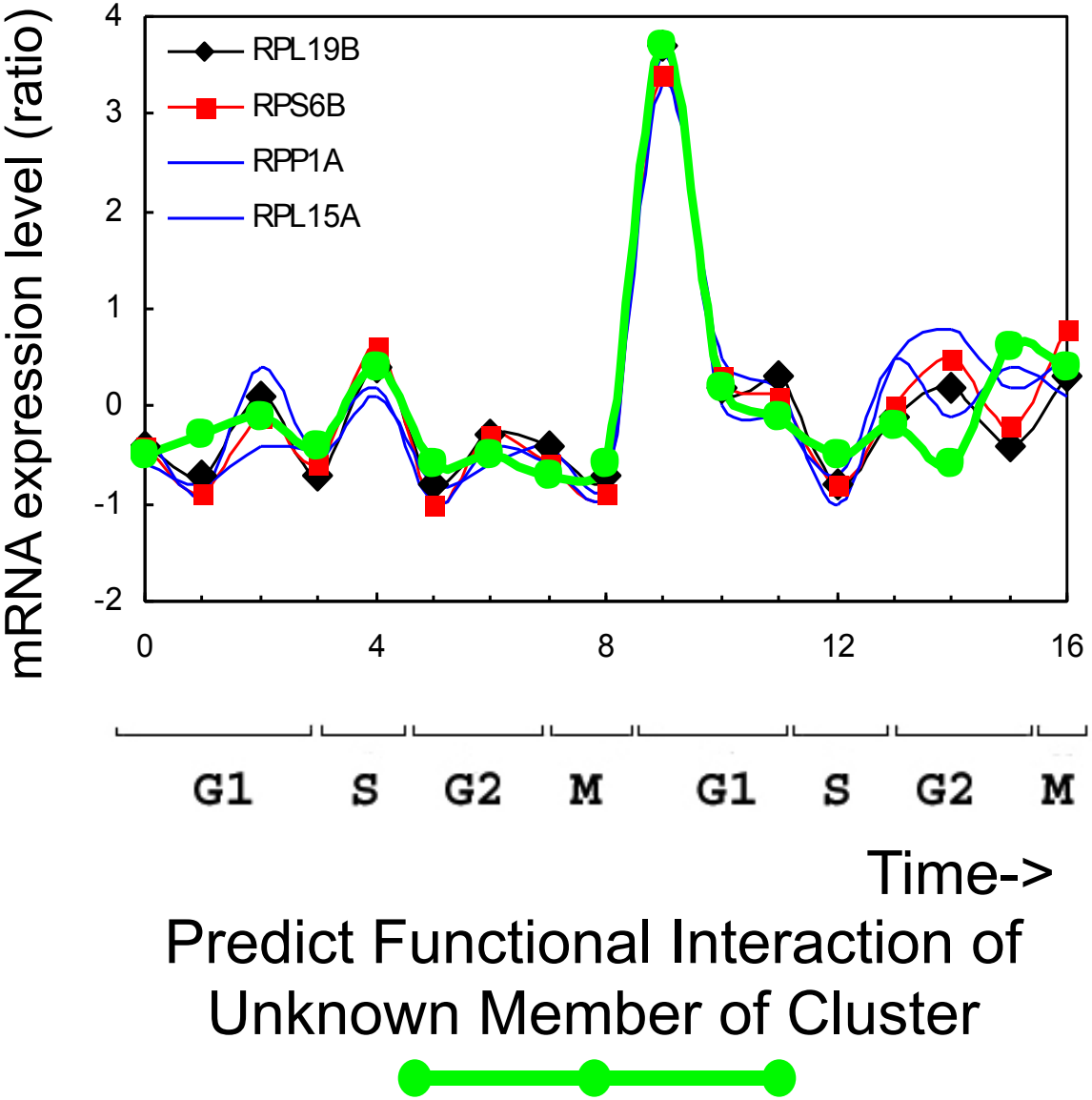




Close relationship from 18M
(2 Interacting Ribosomal Proteins)

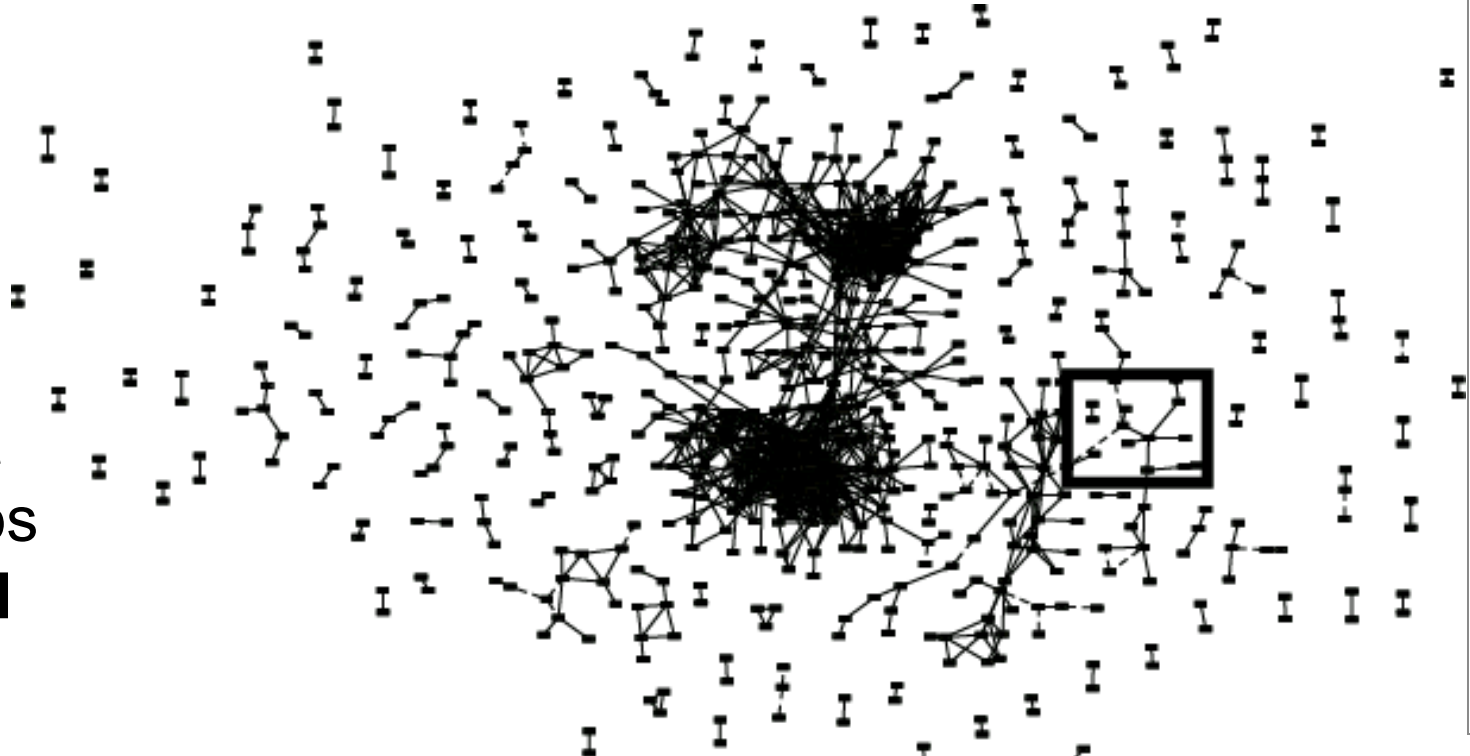
Clustering the yeast cell cycle to uncover interacting proteins





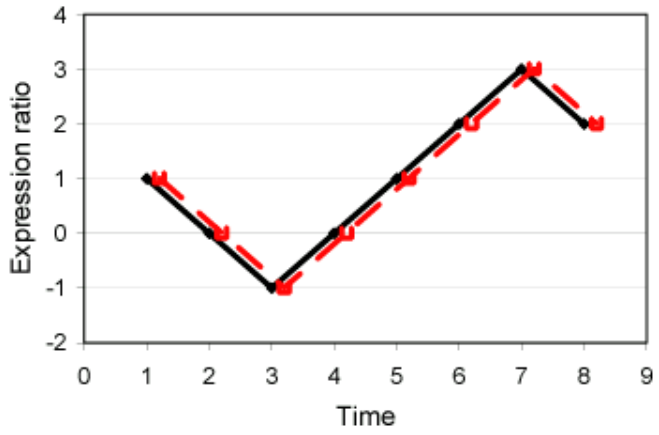
Global Network of Relationships

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

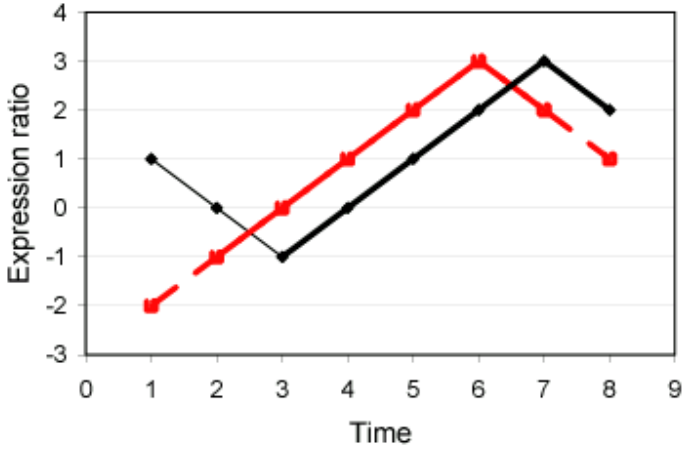


Simultaneous

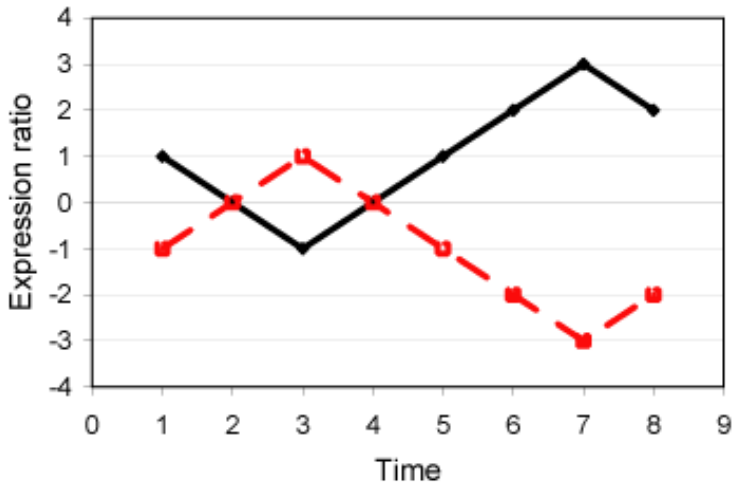
Traditional
Global
Correlation



Time-Shifted



Inverted

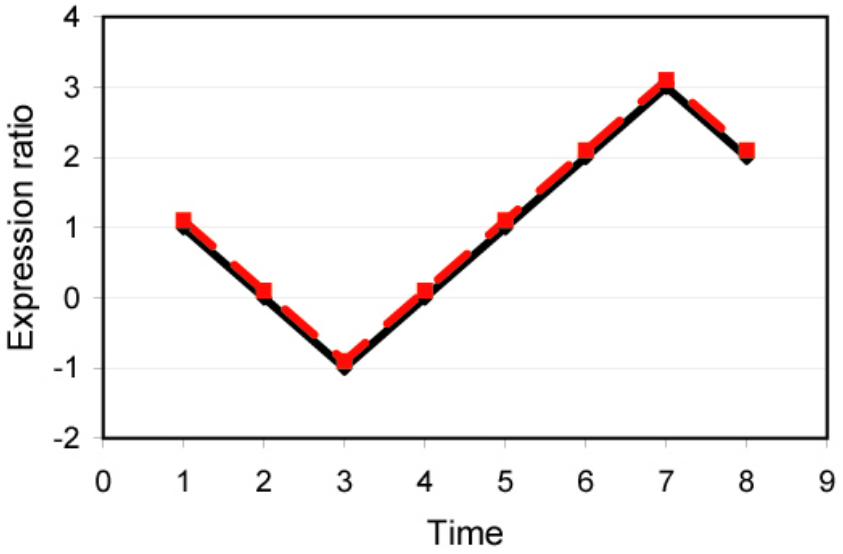


Local Clustering algorithm identifies further (reasonable) types of expression relation-ships

[Church]

Local Alignment

Simultaneous



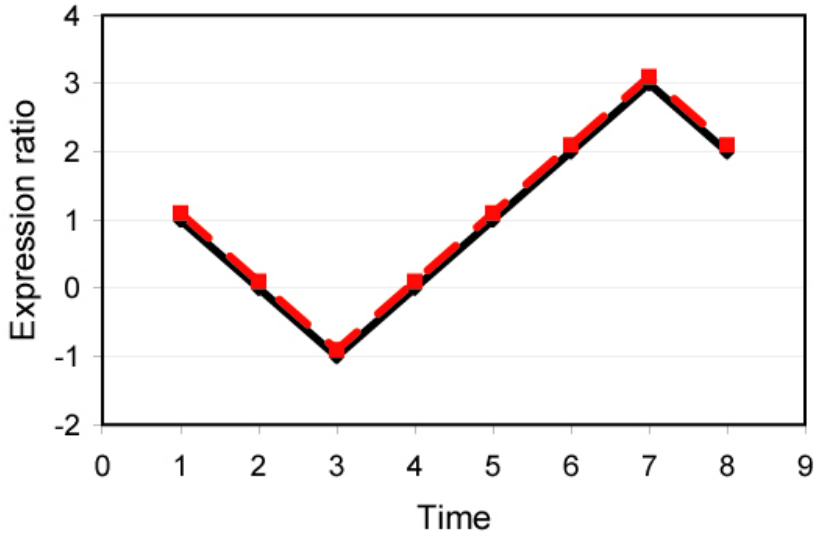
		1	0	-1	0	1	2	3	2
	0	0	0	0	0	0	0	0	0
1	0	1	0	-1	0	1	2	3	2
0	0	0	0	0	0	0	0	0	0
-1	0	-1	0	1	0	-1	-2	-3	-2
0	0	0	0	0	0	0	0	0	0
1	0	1	0	-1	0	1	2	3	2
2	0	2	0	-1	0	2	4	6	4
3	0	3	0	-3	0	3	6	9	6
2	0	1	0	-2	0	2	4	6	4

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

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



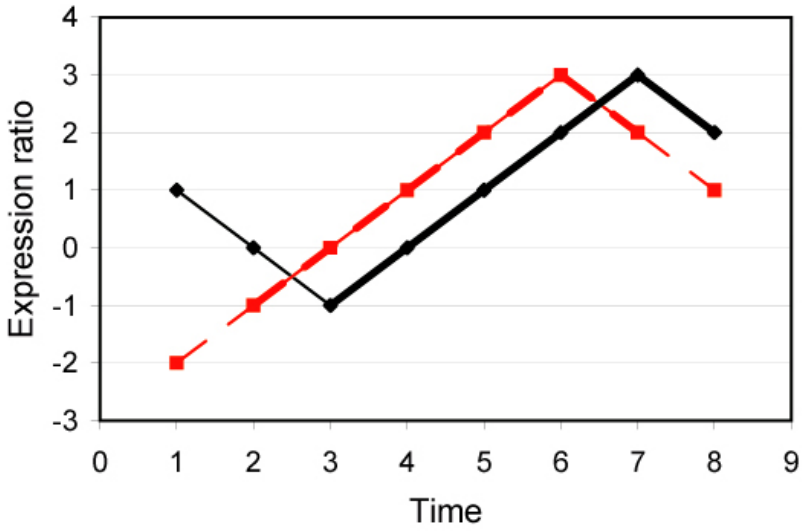
		1	0	-1	0	1	2	3	2
	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	2	3	2
0	0	0	1	0	0	0	1	2	3
-1	0	0	0	2	0	0	0	0	0
0	0	0	0	0	2	0	0	0	0
1	0	1	0	0	0	3	2	3	2
2	0	2	1	0	0	2	7	8	7
3	0	3	2	0	0	3	8	16	14
2	0	2	3	0	0	2	7	14	20

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

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

Time-Shifted



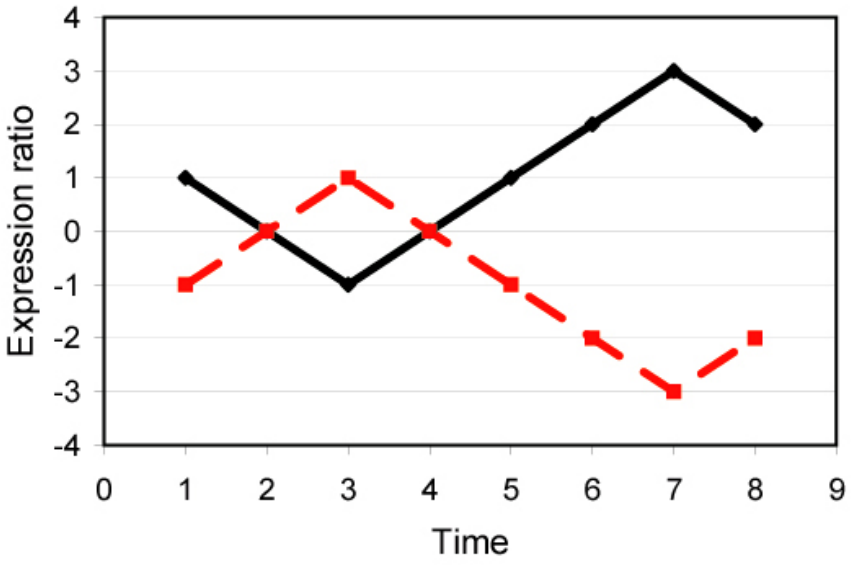
		-2	-1	0	1	2	3	2	1
	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	2	3	2	1
0	0	0	0	0	0	1	2	3	2
-1	0	2	1	0	0	0	0	0	2
0	0	0	2	1	0	0	0	0	0
1	0	0	0	2	2	2	3	2	1
2	0	0	0	0	4	6	8	7	4
3	0	0	0	0	3	10	15	14	10
2	0	0	0	0	2	7	16	19	16

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

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

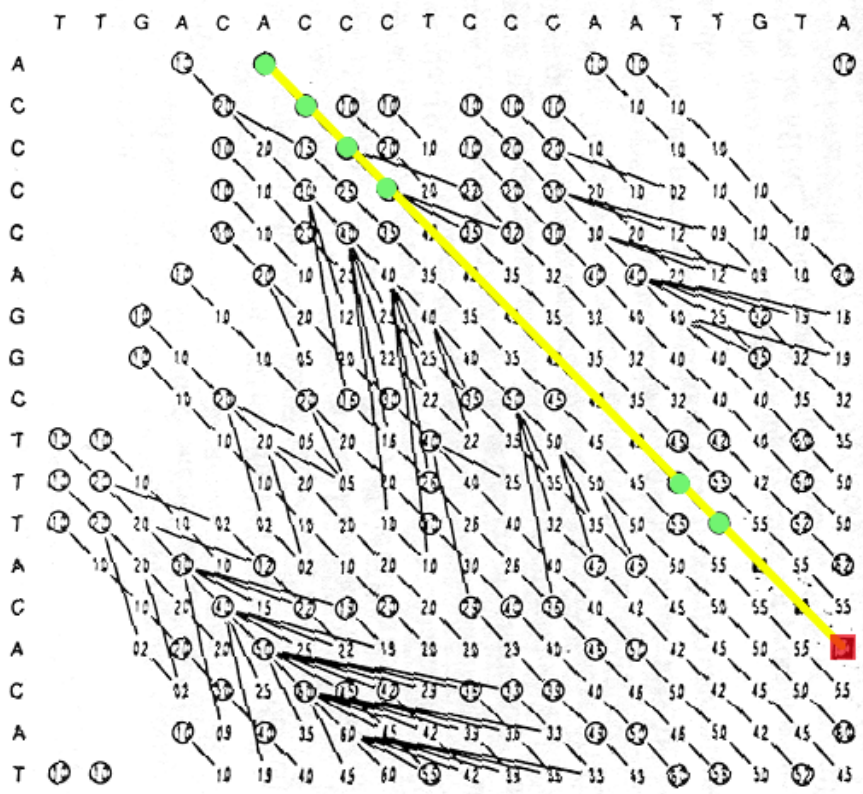


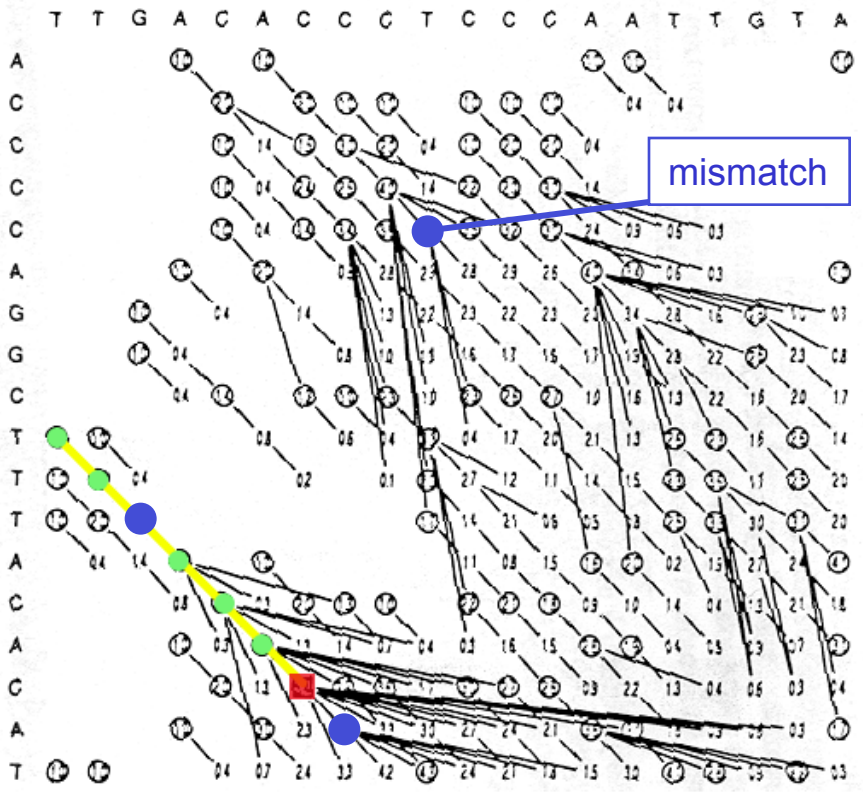
		-1	0	1	0	-1	-2	-3	-2
	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	2	3	2
0	0	0	1	0	0	0	1	2	3
-1	0	0	0	2	0	0	0	0	2
0	0	0	0	0	2	0	0	0	0
1	0	1	0	0	0	3	2	3	2
2	0	2	1	0	0	2	7	8	7
3	0	3	2	0	0	3	8	16	14
2	0	2	3	0	0	2	7	14	20

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

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...
||||  ||  |
.....ACCCCAGGCTTTACACAT
1234444444456667
```

```
T T G A C A C C...
| | - | | | | -
T T T A C A C A...
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