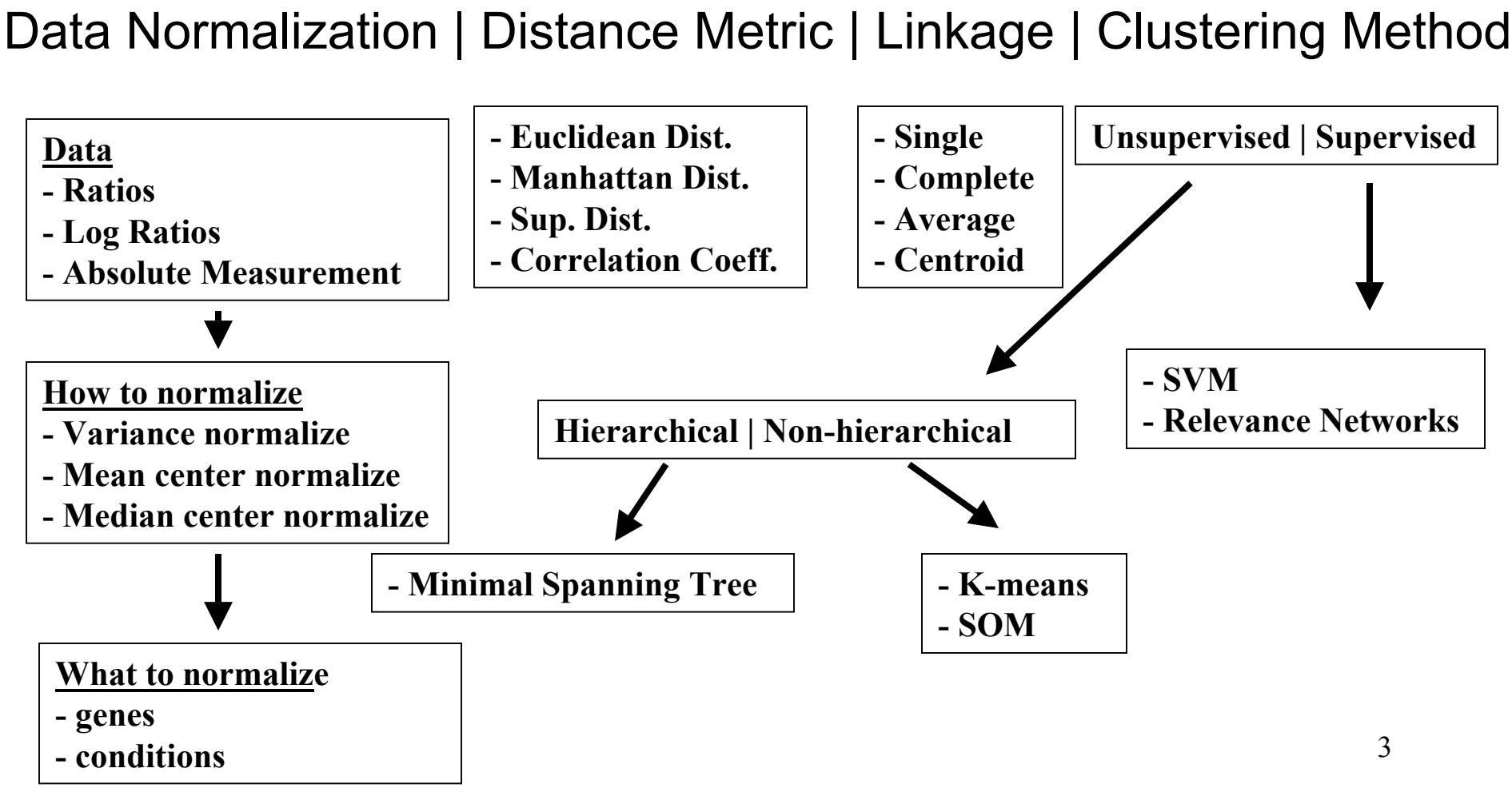
RNA1: Last week's take home lessons

- Integration with previous topics (HMM for **RNA structure**)
- Goals of molecular **quantitation** (maximal fold-changes, clustering & classification of genes & conditions/cell types, causality)
- Genomics-grade **measures** of RNA and protein and how we choose (SAGE, oligo-arrays, gene-arrays)
- Sources of random and systematic **errors** (reproducibility of RNA source(s), biases in labeling, non-polyA RNAs, effects of array geometry, cross-talk).
- **Interpretation** issues (splicing, 5' & 3' ends, editing, gene families, small RNAs, antisense, apparent absence of RNA).
- **Time series data:** causality, mRNA decay, time-warping

RNA2: Today's story & goals

- Clustering by gene and/or condition
- Distance and similarity measures
- Clustering & classification
- Applications
- DNA & RNA motif discovery & search

Gene Expression Clustering Decision Tree



(Whole genome) RNA quantitation objectives

RNAs showing maximum change
minimum change detectable/meaningful

RNA absolute levels (compare protein levels)
minimum amount detectable/meaningful

Classification: drugs & cancers

Network -- direct causality-- motifs

Clustering vs. supervised learning

K-means clustering
SOM = Self Organizing Maps
SVD = Singular Value decomposition
PCA = Principal Component Analysis

SVM = Support Vector Machine classification and Relevance networks
Brown et al. PNAS 97:262 Butte et al PNAS 97:12182
(http://www.pnas.org/cgi/content/full/97/1/262)
(http://www.pnas.org/cgi/content/full/97/22/12182)

Cluster analysis of mRNA expression data

By gene (rat spinal cord development, yeast cell cycle):

Wen *et al.*, 1998; Tavazoie *et al.*, 1999; Eisen *et al.*, 1998; Tamayo *et al.*, 1999
(http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=9419376&form=6&db=m&Dopt=b)
(http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=10391217&form=6&db=m&Dopt=b)
(http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=9843981&form=6&db=m&Dopt=b)
(http://www.pubmedcentral.nih.gov/b.cgi?pubmedid=10077610)

By condition or cell-type or by gene&cell-type (human cancer):

Golub, *et al.* 1999; Alon, *et al.* 1999; Perou, *et al.* 1999; Weinstein, et al 1997

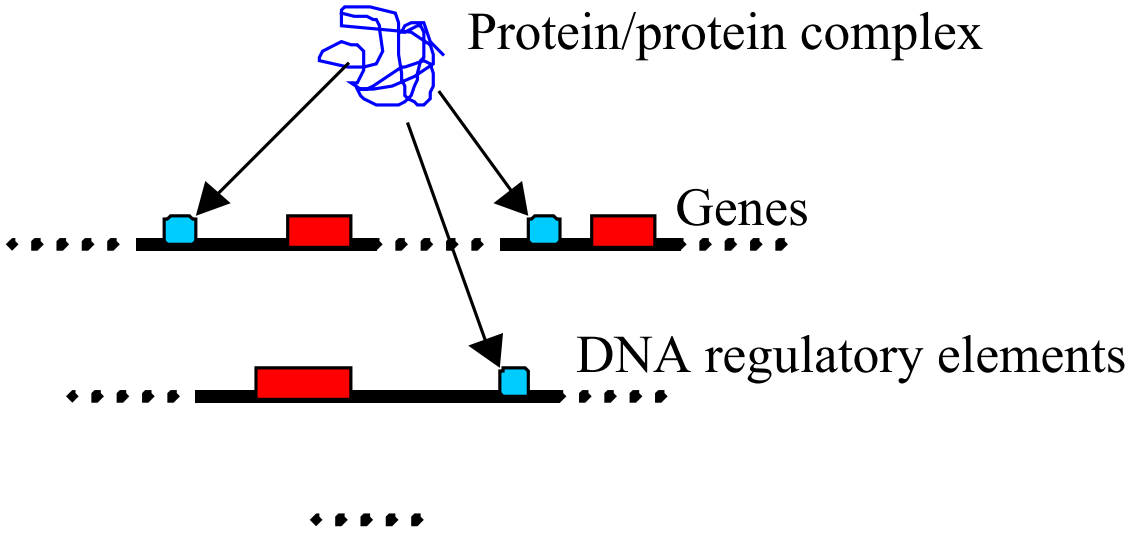
Cheng, ISMB 2000.
(http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&uid=10521349&db=m)
(http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=10359783&form=6&Dopt=b)
(http://www.pubmedcentral.nih.gov/b.cgi?pubmedid=10430922)

- Rana.lbl.gov/EisenSoftware.htm

Cluster Analysis

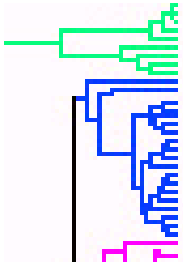
General Purpose: To divide samples into homogeneous groups based on a set of features.

Gene Expression Analysis: To find co-regulated genes.

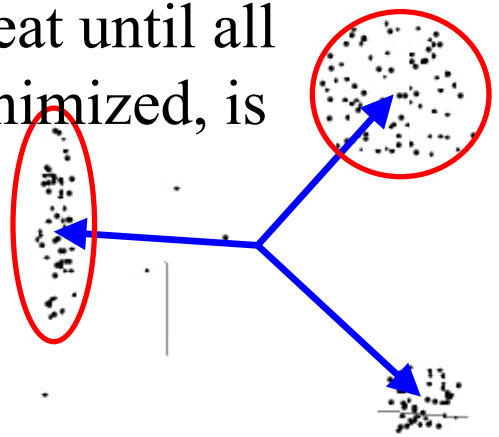


Clustering hierarchical & non-

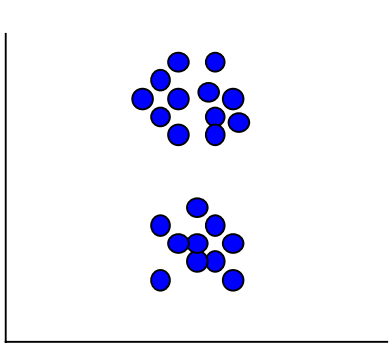
•**Hierarchical:** a series of successive fusions of data until a final number of clusters is obtained; e.g. Minimal Spanning Tree: each component of the population to be a cluster. Next, the two clusters with the minimum distance between them are fused to form a single cluster. Repeated until all components are grouped.

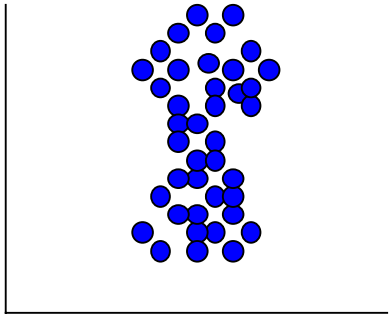


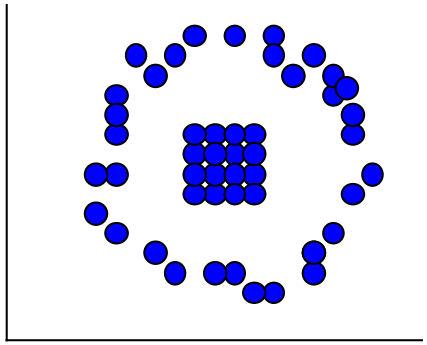
• **Non-:** e.g. K-mean: K clusters chosen such that the points are mutually farthest apart. Each component in the population assigned to one cluster by minimum distance. The centroid's position is recalculated and repeat until all the components are grouped. The criterion minimized, is the within-clusters sum of the variance.

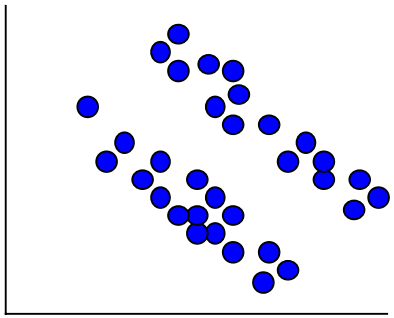


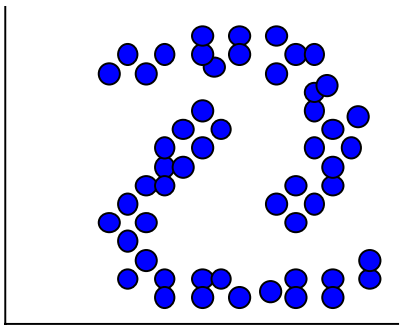
Clusters of Two-Dimensional Data

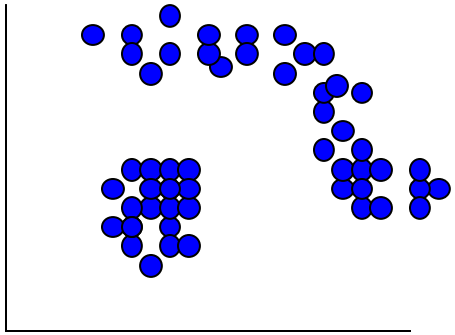












Key Terms in Cluster Analysis

- Distance measures
- Similarity measures
- Hierarchical and non-hierarchical
- Single/complete/average linkage
- Dendrogram

Distance Measures: Minkowski Metric

Suppose two objects x and y both have p features :

$$\boldsymbol{x} = (\boldsymbol{x}_1 \boldsymbol{x}_2 \cdots \boldsymbol{x}_p)$$

$$\boldsymbol{y} = (\boldsymbol{y}_1 \boldsymbol{y}_2 \cdots \boldsymbol{y}_p)$$

The Minkowski metric is defined by

$$\boldsymbol{d}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt[r]{\sum_{\boldsymbol{i}=1}^{\boldsymbol{p}} | \boldsymbol{x}_i - \boldsymbol{y}_i |^r}$$

Most Common Minkowski Metrics

1, $r = 2$ **(Euclidean distance)**

$$d(x, y) = \sqrt[2]{\sum_{i=1}^{p} | x_i - y_i |^2}$$

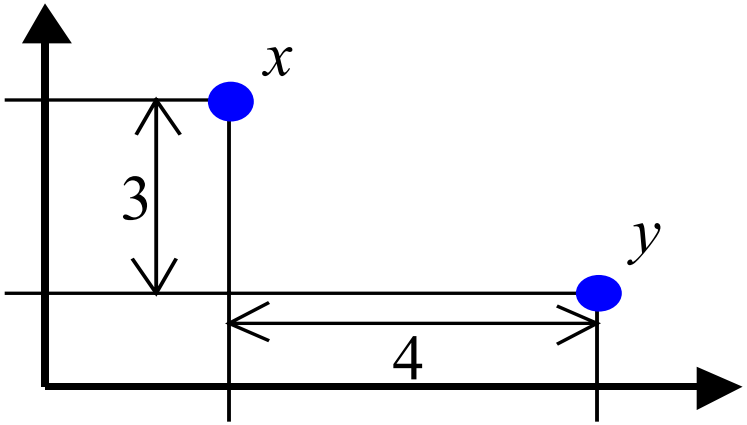
2, $r = 1$ **(Manhattan distance)**

$$d(x, y) = \sum_{i=1}^{p} | x_i - y_i |$$

3, $r = +\infty$ **("sup" distance)**

$$d(x, y) = \max_{1 \le i \le p} | x_i - y_i |$$

An Example



1, Euclidean distance : $\sqrt[2]{4^2+3^2}=5.$

2, Manhattan distance : $4+3=7.$

3, "sup" distance : $\max\{4,3\}=4.$

Manhattan distance is called *Hamming distance* when all features are binary.

Gene Expression Levels Under 17 Conditions (1-High,0-Low)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
GeneA	0	1	1	0	0	1	0	0	1	0	0	1	1	1	0	0	1
GeneB	0	1	1	1	0	0	0	0	1	1	1	1	1	1	0	1	1

Hamming Distance : $\#(01)+\#(10)=4+1=5$**.**

Similarity Measures: Correlation Coefficient

$$s(x,y) = \frac{\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i - \bar{x})^2 \times \sum_{i=1}^{p}(y_i - \bar{y})^2}}$$

where $\bar{x} = \frac{1}{p}\sum_{i=1}^{p} x_i$ **and** $\bar{y} = \frac{1}{p}\sum_{i=1}^{p} y_i$.

$$\left|s(x,y)\right| \leq 1$$

What kind of x and y give linear CC

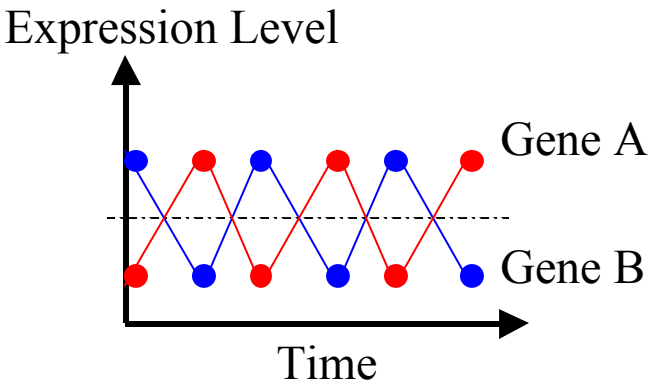
$(1)\ s(x,y)=1,$

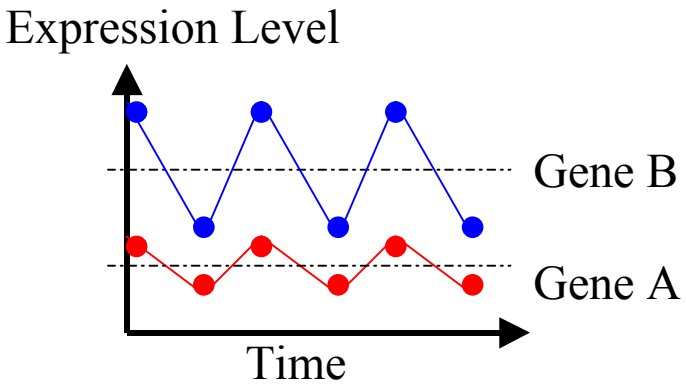
$(2)\ s(x,y)=-1,$

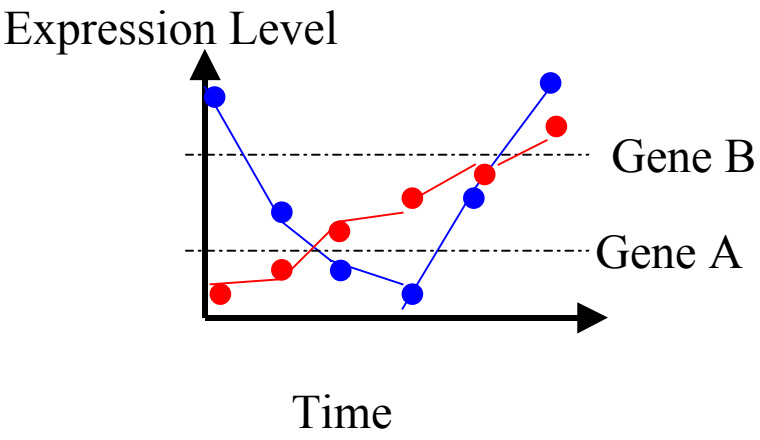
$(3)\ s(x,y)=0$

?

Similarity Measures: Correlation Coefficient







Hierarchical Clustering Dendrograms

See Alon *et al.* 1999

Hierarchical Clustering Techniques

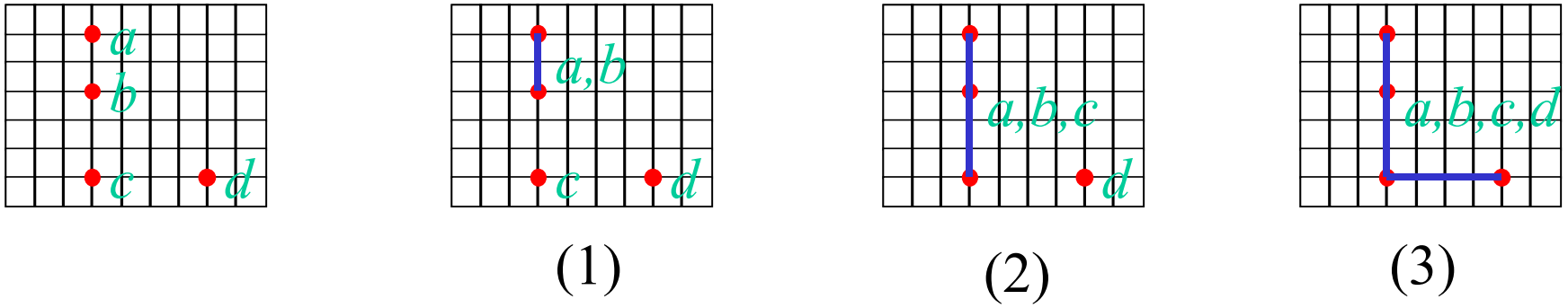
At the beginning, each object (gene) is a cluster. In each of the subsequent steps, two *closest* clusters will merge into one cluster until there is only one cluster left.

The distance between two clusters is defined as the distance between

- Single-Link Method / Nearest Neighbor: their closest members.
- Complete-Link Method / Furthest Neighbor: their furthest members.
- Centroid: their centroids.
- Average: average of all cross-cluster pairs.

Single-Link Method

Euclidean Distance



	b	c	d
a	2	5	6
b		3	5
c			4

	b	c	d
a	2	5	6
b		3	5
c			4

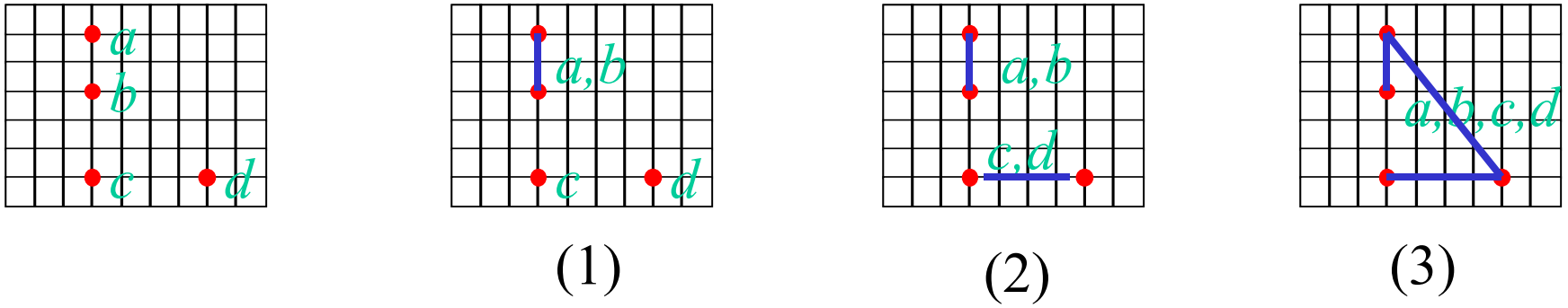
	c	d
a,b	3	5
c		4

	d
a,b,c	4

Distance Matrix

Complete-Link Method

Euclidean Distance



	b	*c*	*d*
a	2	5	6
b		3	5
c			4

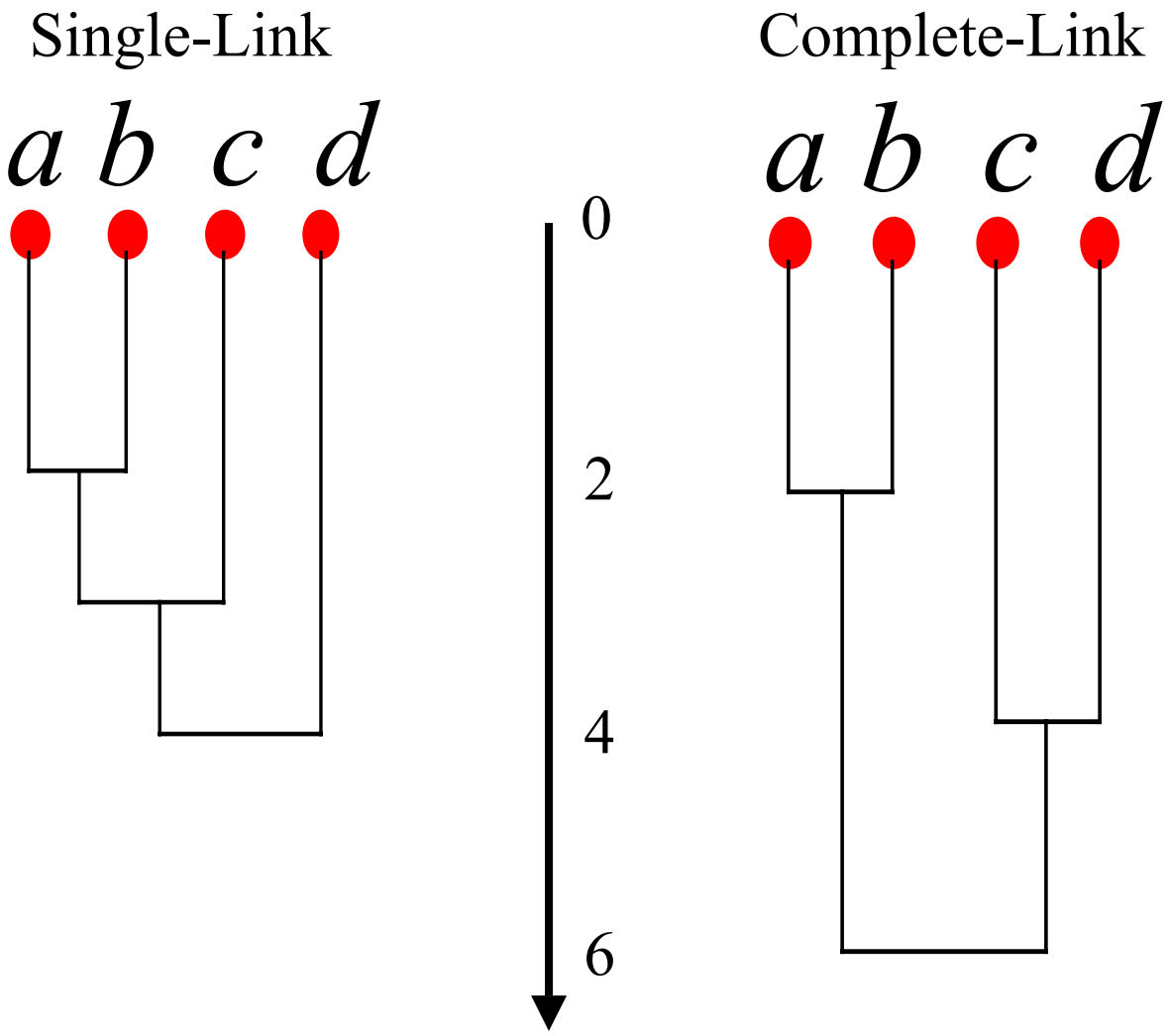
	b	*c*	*d*
a	2	5	6
b		3	5
c			4

	c	*d*
a*,*b	5	6
c		4

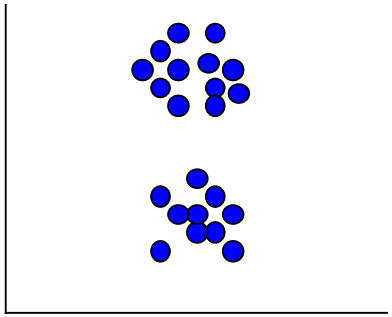
	c,*d*
a*,*b	6

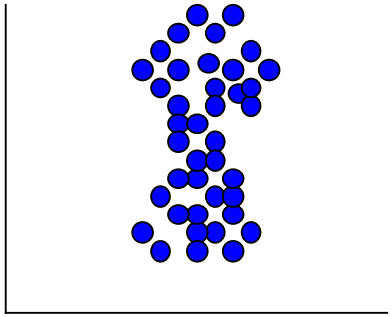
Distance Matrix

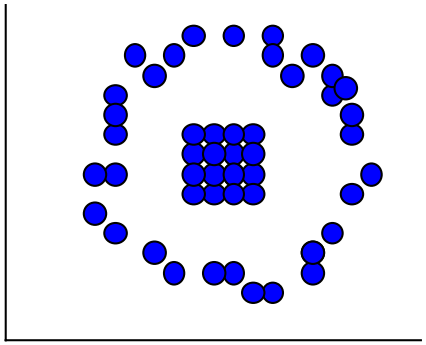
Dendrograms

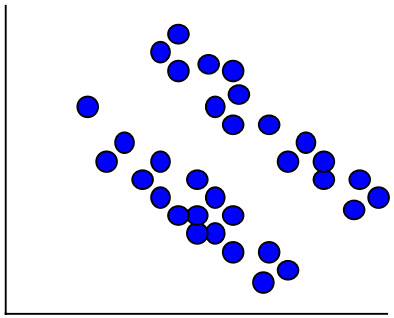


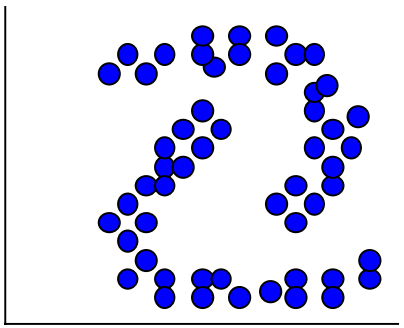
Which clustering methods do you suggest for the following two-dimensional data?

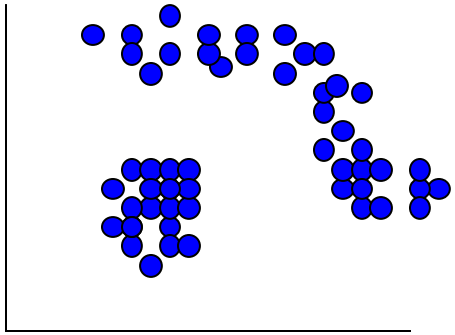








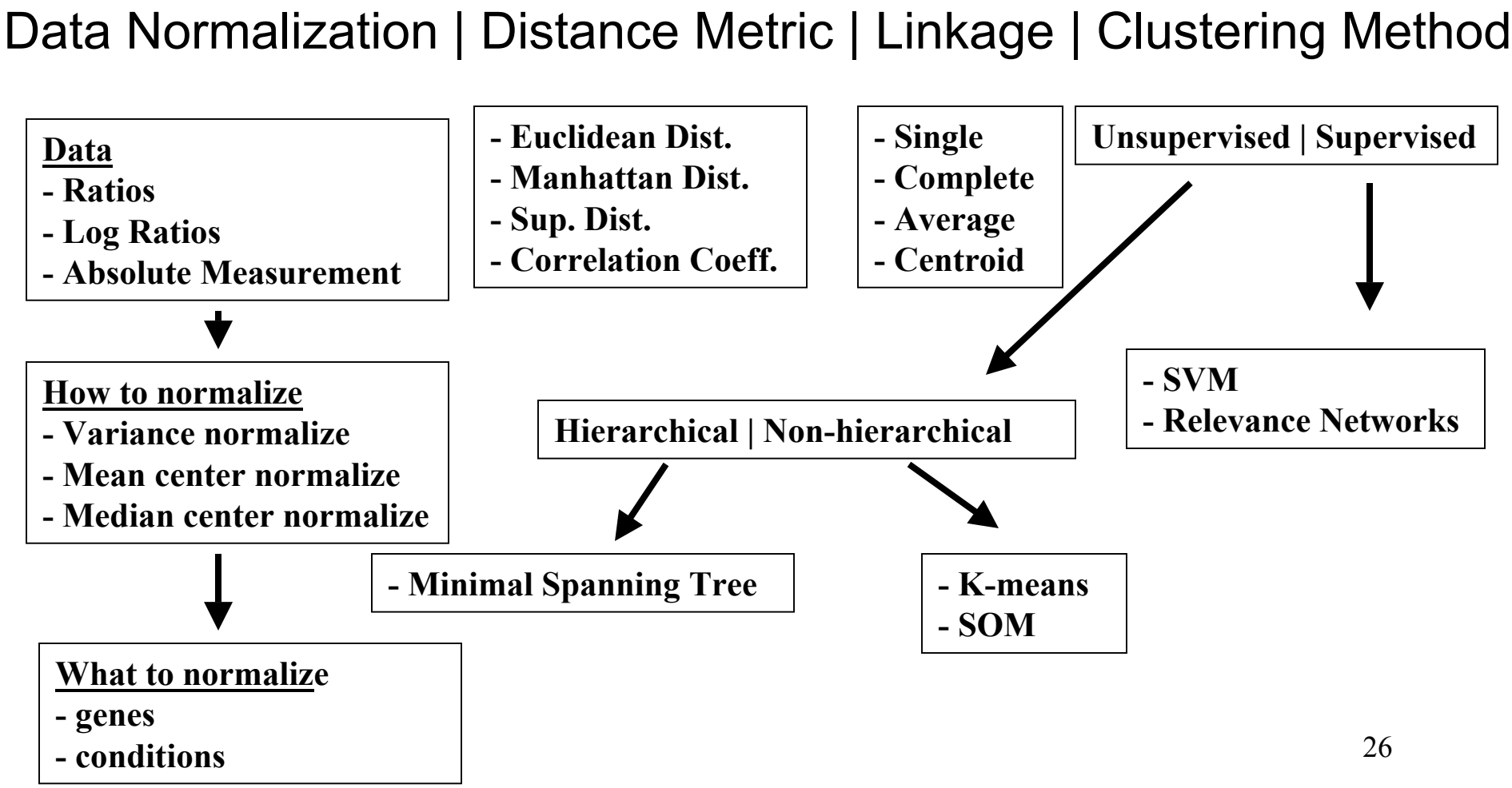




Graphical examples of hierarchical merging

See Nadler and Smith, Pattern Recognition Engineering, 1993

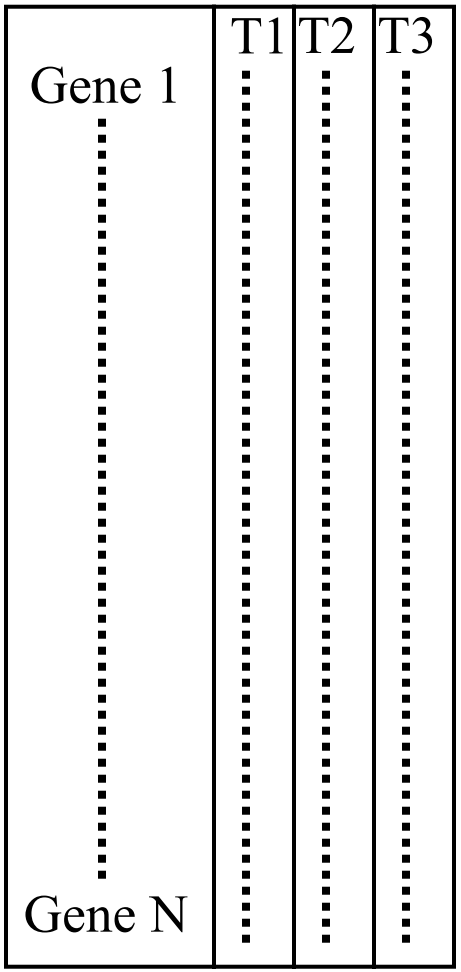
Gene Expression Clustering Decision Tree



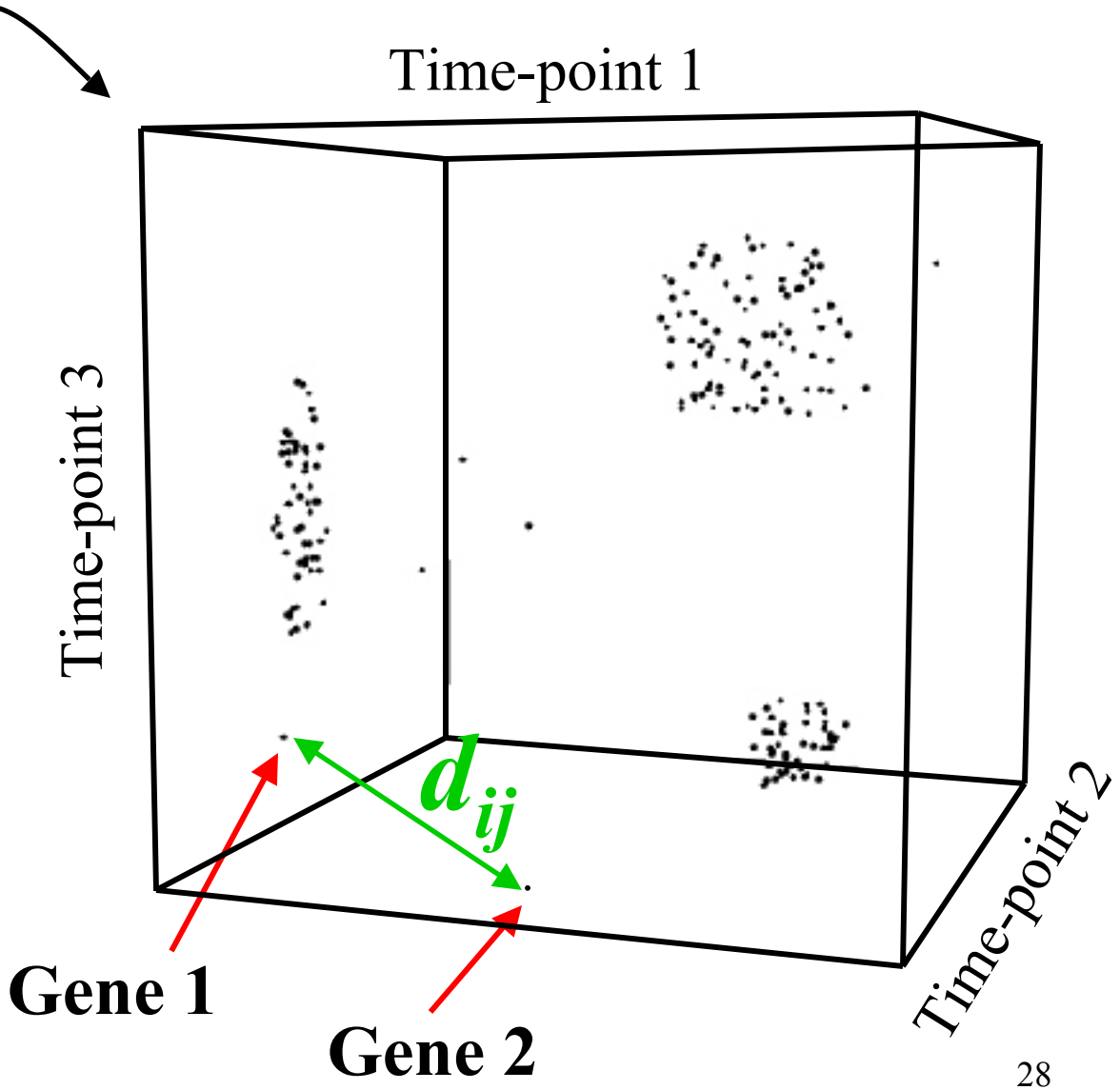
Identifying prevalent experession patterns (clusters)

See Tavazoie *et al. 1999 (http://arep.med.harvard.edu)*

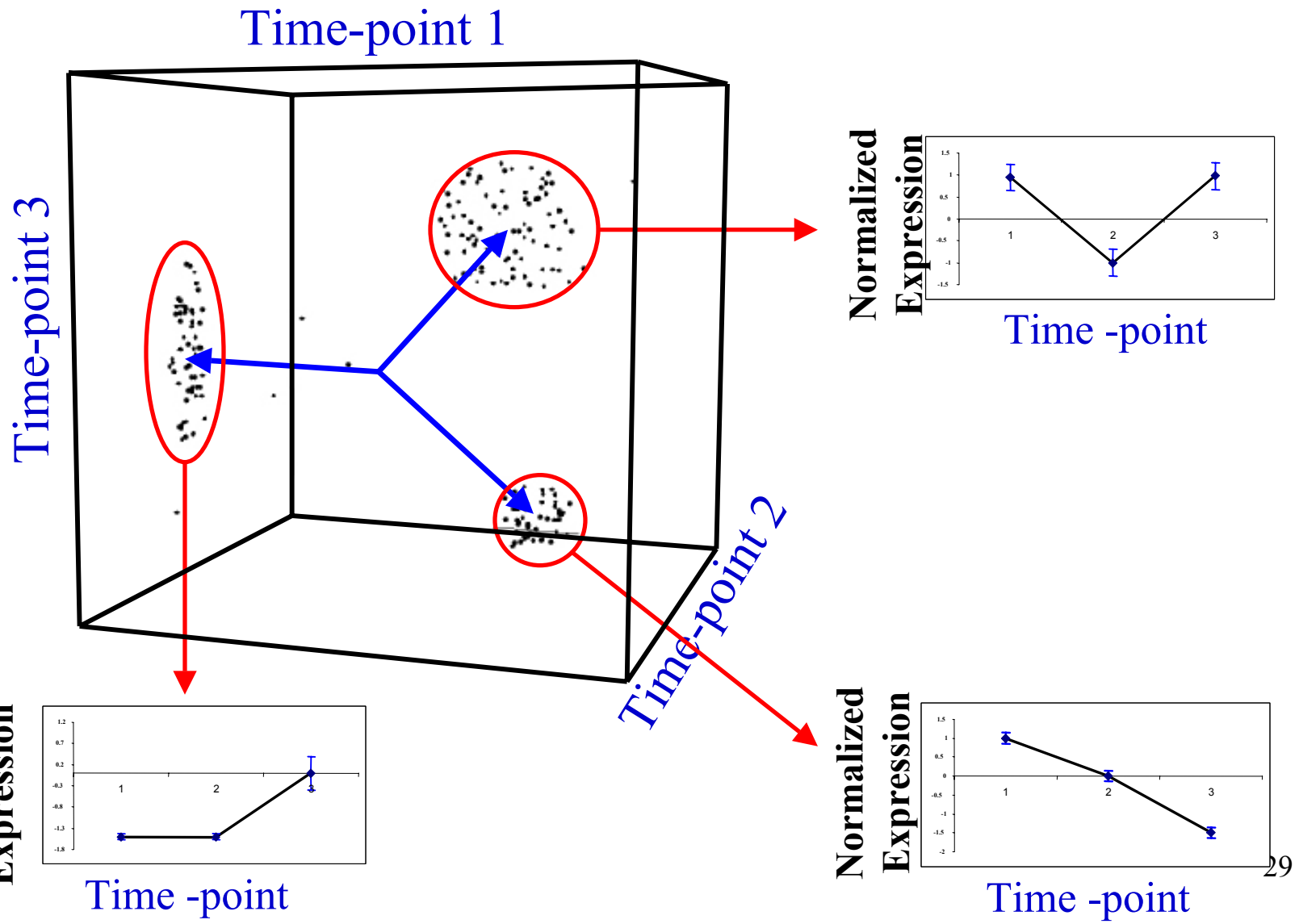
Representation of expression data



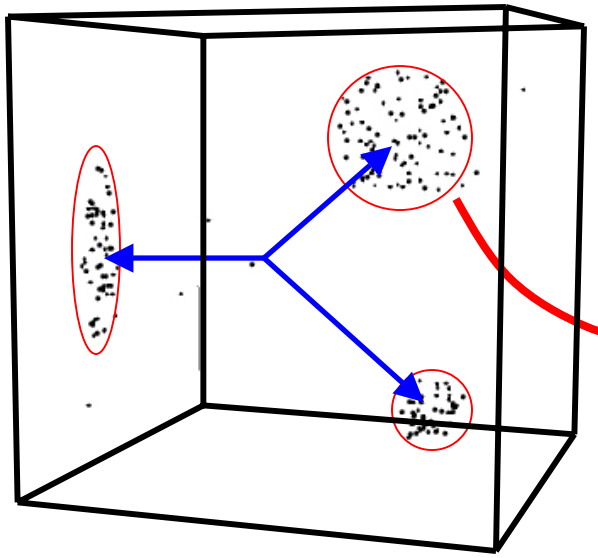
Normalized Expression Data from microarrays



Identifying prevalent expression patterns (gene clusters)



Cluster contents



Genes	MIPS functional category
gpm1	Glycolysis
HTB1	Nuclear Organization
RPL11A, RPL12B, RPL13A, RPL14A, RPL15A, RPL17A, RPL23A	Ribosome
TEF2	Translation
YDL228c, YDR133C, YDR134C, YDR327W, YDR417C, YKL153W, YPL142C	Unknown

See Eisen, *et al.* (1998): Fig.1 Cluster display of data

And Weinstein, *et al* (1997)

RNA2: Today's story & goals

- Clustering by gene and/or condition
- Distance and similarity measures
- Clustering & classification
- Applications
- DNA & RNA motif discovery & search

Motif-finding algorithms

- oligonucleotide frequencies
- Gibbs sampling (e.g. AlignACE)
- MEME
- ClustalW
- MACAW

Feasibility of a whole-genome motif search?

Genome:
(12 Mb)

Transcription control sites
(~7 bases of information)

- 7 bases of information (14 bits) ~ 1 match every 16000 sites.
- 1500 such matches in a 12 Mb genome (24 * 10^6 sites).
- The distribution of numbers of sites for different motifs is Poisson with mean 1500, which can be approximated as normal with a mean of 1500 and a standard deviation of ~40 sites.
- Therefore, ~100 sites are needed to achieve a detectable signal above background.

Sequence Search Space Reduction

- Whole-genome mRNA expression data: two-way comparisons between different conditions or mutants, clustering/grouping over many conditions/timepoints.
- Shared phenotype (functional category).
- Conservation among different species.
- Details of the sequence selection: eliminate protein-coding regions, repetitive regions, and any other sequences not likely to contain control sites.

Sequence Search Space Reduction

• Whole-genome mRNA expression data: two-way comparisons between different conditions or mutants, clustering/grouping over many conditions/timepoints.

• Shared phenotype (functional category).

• Conservation among different species.

• Details of the sequence selection: eliminate protein-coding regions, repetitive regions, and any other sequences not likely to contain control sites.

Motif Finding

AlignACE

(*Align*s nucleic *A*cid *C*onserved *E*lements)

• Modification of Gibbs Motif Sampling (GMS), a routine for motif finding in protein sequences (Lawrence, *et al*. Science 262:208-214, 1993).

• Advantages of GMS/AlignACE:
  • stochastic sampling
  • variable number of sites per input sequence
  • distributed information content per motif
  • considers both strands of DNA simultaneously
  • efficiently returns multiple distinct motifs

AlignACE Example

Input Data Set

300-600 bp of upstream sequence per gene are searched in *Saccharomyces cerevisiae*.

AlignACE Example

The Target Motif

5′- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAG**AAAAGAGTCA**GACATCGAAACATACAT ...*HIS7*

5′- ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCG**AAATGACTCA**ACG ...*ARO4*

5′- CACATCCAACGAATCACCTCACCGTTATCG**TGACTCACTT**TCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT ...*ILV6*

5′- TGCGAAC**AAAAGAGTCA**TTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC ...*THR4*

5′- ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATA**TGACTCATCC**CGAACATGAAA ...*ARO1*

5′- ATTGAT**TGACTCATTT**TCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA ...*HOM2*

5′- GGCGCCACAGTCCGCGTTTGGTTATCCGGC**TGACTCATTCTGACTCTTTT**TTGGAAAGTGTGGCATGTGCTTCACACA ...*PRO3*

AAAAGAGTCA
AAATGACTCA
AAGTGAGTCA
AAAAGAGTCA
GGATGAGTCA
AAATGAGTCA
GAATGAGTCA
AAAAGAGTCA
****

MAP score = 20.37 (maximum)

AlignACE Example

Initial Seeding

5′- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAA**TGAAAAATTC**ATGAGAAAAGAGTCA**GACATCGAAA**CATACAT …*HIS7*

5′- ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG …*ARO4*

5′- CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATC**GCCGAAGTGC**CATAAAAAATATTTTTT …*ILV6*

5′- TGCGAACAAAA**GAGTCATTAC**AACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC …*THR4*

5′- ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATA**GTAAATTGTC**ATGCATATGACTCATCCCGAACATGAAA …*ARO1*

5′- ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA …*HOM2*

5′- GGCG**CCACAGTCCG**CGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCAT**GTGCTTCACA**CA …*PRO3*

TGAAAAATTC
GACATCGAAA
GCACTTCGGC
GAGTCATTAC
GTAAATTGTC
CCACAGTCCG
TGTGAAGCAC
****

MAP score = -10.0

AlignACE Example

Sampling

TGAAAAATTC
GACATCGAAA
GCACTTCGGC
GAGTCATTAC
GTAAATTGTC
CCACAGTCCG
TGTGAAGCAC
****

How much better is the alignment with this site as opposed to without?

TCTCTCTCCA
TGAAAAATTC
GACATCGAAA
GCACTTCGGC
GAGTCATTAC
GTAAATTGTC
CCACAGTCCG
TGTGAAGCAC
****

AlignACE Example

Continued Sampling

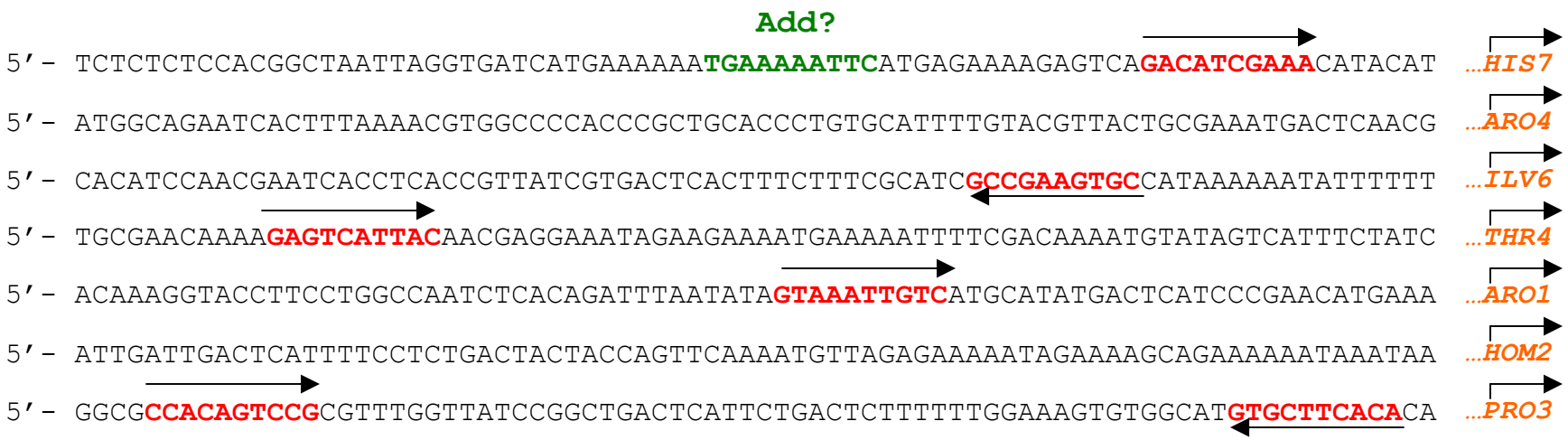
~~TGAAAAATTC~~
GACATCGAAA
GCACTTCGGC
GAGTCATTAC
GTAAATTGTC
CCACAGTCCG
TGTGAAGCAC
****

How much better is the alignment with this site as opposed to without?

ATGAAAAAAT
~~TGAAAAATTC~~
GACATCGAAA
GCACTTCGGC
GAGTCATTAC
GTAAATTGTC
CCACAGTCCG
TGTGAAGCAC
****

AlignACE Example

Continued Sampling



How much better is the alignment with this site as opposed to without?

GACATCGAAA	TGAAAAATTC
GCACTTCGGC	GACATCGAAA
GAGTCATTAC	GCACTTCGGC
GTAAATTGTC	GAGTCATTAC
CCACAGTCCG	GTAAATTGTC
TGTGAAGCAC	CCACAGTCCG
**********	TGTGAAGCAC
	**********

AlignACE Example

Column Sampling

How much better is the alignment with this new column structure?

```
GACATCGAAA          GACATCGAAAC
GCACTTCGGC          GCACTTCGGCG
GAGTCATTAC          GAGTCATTACA
GTAAATTGTC          GTAAATTGTCA
CCACAGTCCG          CCACAGTCCGC
TGTGAAGCAC          TGTGAAGCACA
**********          ********* *
```

AlignACE Example

The Best Motif

5′- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAG**AAAAGAGTCA**GACATCGAAACATACAT ...*HIS7*

5′- ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCG**AAATGACTCA**ACG ...*ARO4*

5′- CACATCCAACGAATCACCTCACCGTTATCG**TGACTCACTT**TCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT ...*ILV6*

5′- TGCGAAC**AAAAGAGTCA**TTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC ...*THR4*

5′- ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATA**TGACTCATCC**CGAACATGAAA ...*ARO1*

5′- ATTGAT**TGACTCATTT**TCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA ...*HOM2*

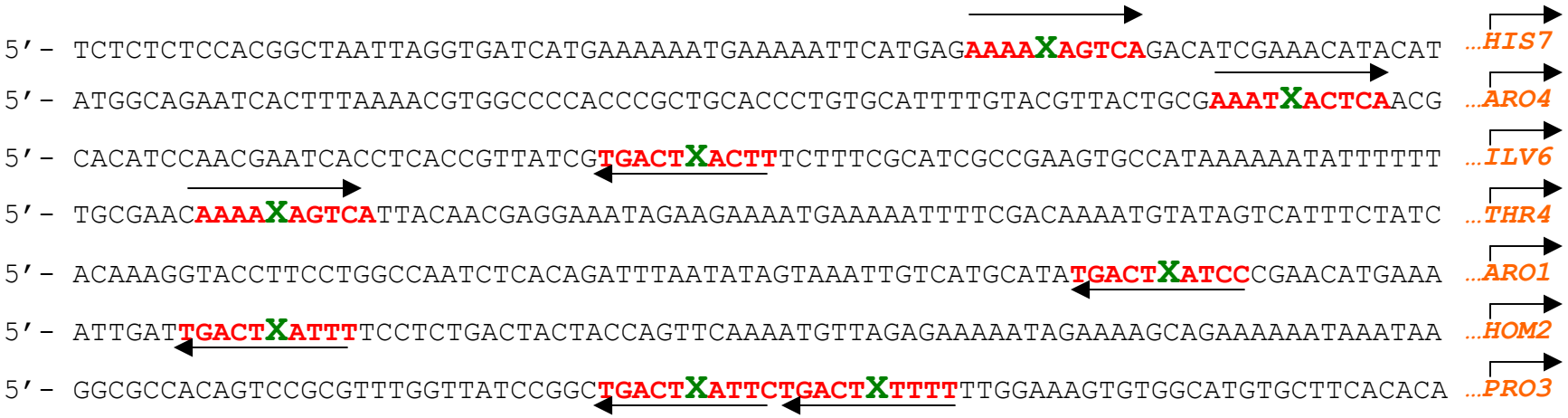
5′- GGCGCCACAGTCCGCGTTTGGTTATCCGGC**TGACTCATTCTGACTCTTTT**TTGGAAAGTGTGGCATGTGCTTCACACA ...*PRO3*

AAAAGAGTCA
AAATGACTCA
AAGTGAGTCA
AAAAGAGTCA
GGATGAGTCA
AAATGAGTCA
GAATGAGTCA
AAAAGAGTCA
****

MAP score = 20.37

AlignACE Example

Masking (old way)



AAAAGAGTCA
AAATGACTCA
AAGTGAGTCA
AAAAGAGTCA
GGATGAGTCA
AAATGAGTCA
GAATGAGTCA
AAAAGAGTCA
****

- Take the best motif found after a prescribed number of random seedings.
- Select the strongest position of the motif.
- Mark these sites in the input sequence, and do not allow future motifs to sample those sites.
- Continue sampling.

AlignACE Example

Masking (new way)

AAAAGAGTCA
AAATGACTCA
AAGTGAGTCA
AAAAGAGTCA
GGATGAGTCA
AAATGAGTCA
GAATGAGTCA
AAAAGAGTCA
****

- Maintain a list of all distinct motifs found.
- Use CompareACE to compare subsequent motifs to those already found.
- Quickly reject weaker, but similar motifs.

MAP Score

$$MAP = \log\Big[\prod_{j=1}^{C} \frac{\Gamma(\beta)}{\Gamma(F_j+\beta)} \prod_{b=1}^{4} \frac{\Gamma(F_{jb}+\beta_b)}{\Gamma(\beta_b)} \times \frac{B_{a,b}(N, T-N)}{B_{a,b}(0,T)} \times \prod_{b=1}^{4} G_b^{-F_b} \times \binom{W-2}{C-2}^{-1}\Big]$$

B,Γ = standard Beta & Gamma functions

N = number of aligned sites; T = number of total possible sites

F_{jb} = number of occurrences of base *b* at position *j* (F = sum)

G_b = background genomic frequency for base *b*

β_b = *n* x G_b for *n* pseudocounts (β = sum)

W = width of motif; C = number of columns in motif (W>=C)

MAP Score

MAP ≈ N log R

N = number of aligned sites
R = overrepresentation of those sites.

AlignACE Example: Final Results

(alignment of upstream regions from 116 amino acid biosynthetic genes in *S. cerevisiae*)

MAP score	Motif
188.3	
78.1	
20.6	
28.1	
117.5	
31.1	
73.4	
8.2	
19.3	
55.0	
89.4	
2.7	

GCN4

Indices used to evaluate motif significance

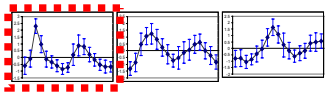
- Group specificity 
- Functional enrichment
- Positional bias
- Palindromicity
- Known motifs (CompareACE)

Searching for additional motif instances in the entire genome sequence

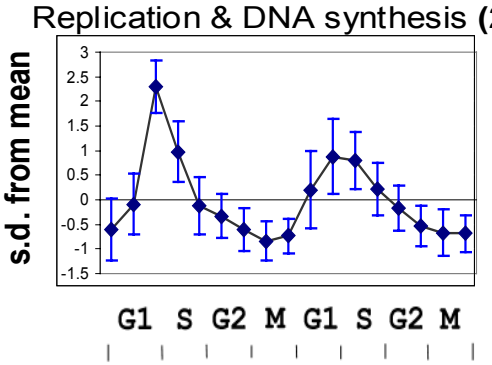
Searches over the entire genome for additional high-scoring instances of the motif are done using the **ScanACE** program, which uses the Berg & von Hippel weight matrix (1987).

$$E = \sum_{l=0}^{M} \ln\left[\frac{n_{lB} + 0.5}{n_{l0} + 0.5}\right]$$

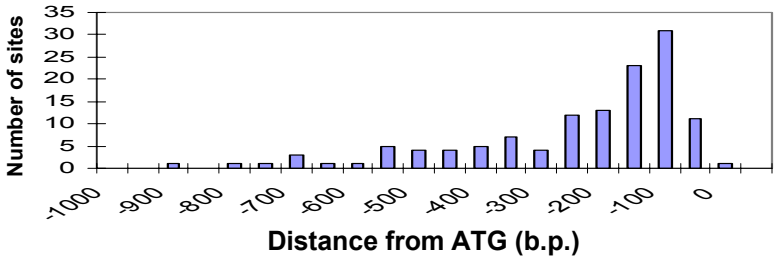
M = length of binding site motif
B = base at position l within the motif
n_{lB} = number of occurrences of base B at position l in the input alignment
n_{lO} = number of occurrences of the most common base at position l in the input alignment

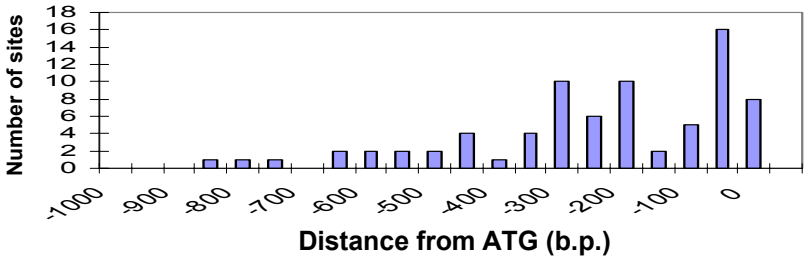


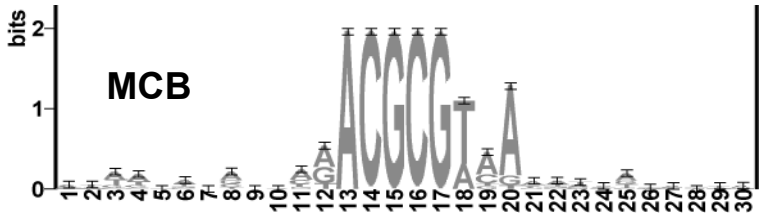
N = 186

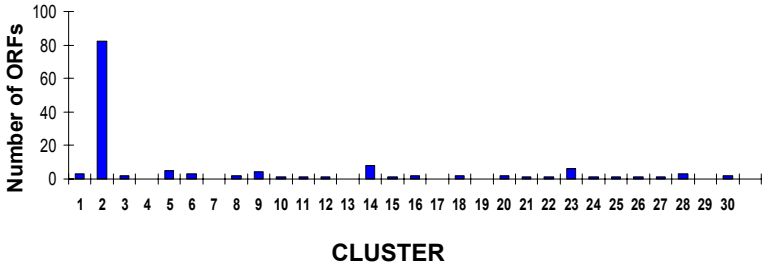


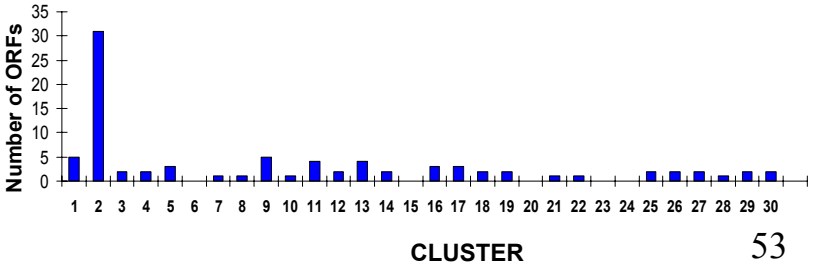
MIPS Functional category (total ORFs)	ORFs within functional category (k)	P-value $-\text{Log}_{10}$
DNA synthesis and replication (82)	23	16
Cell cycle control and mitosis (312)	30	8
Recombination and DNA repair (84)	11	5
Nuclear organization (720)	40	4



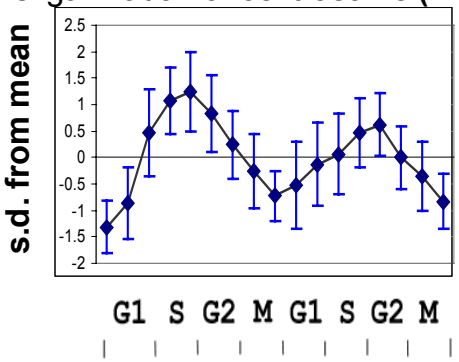


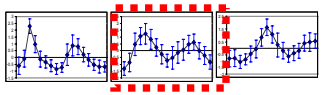






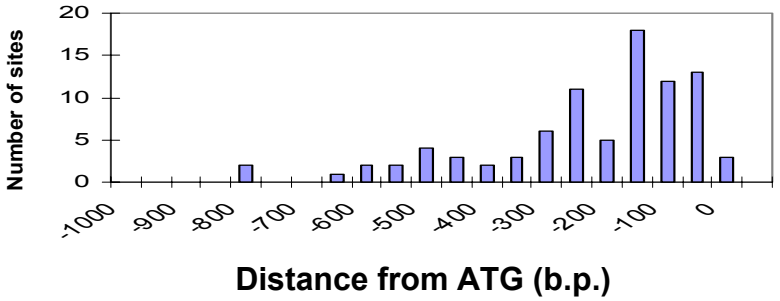
Organization of centrosome **(14)**

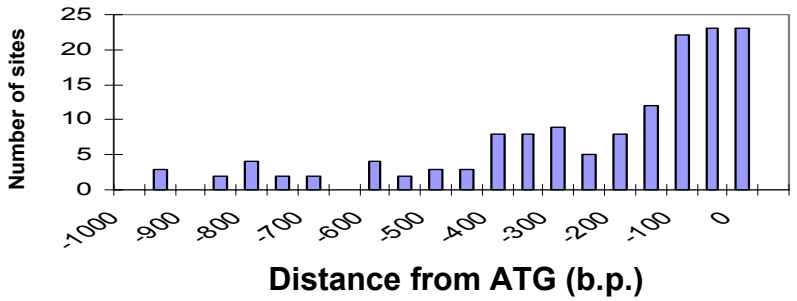


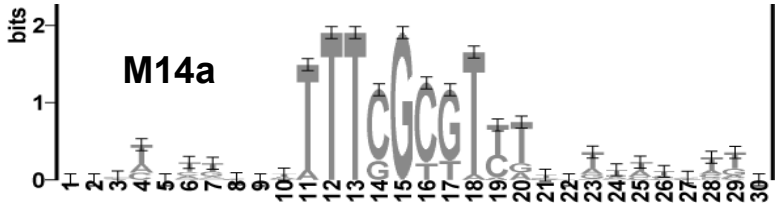


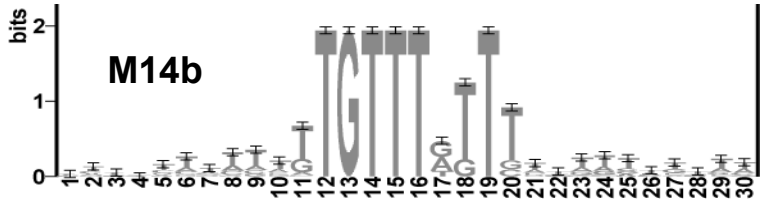
N = 74

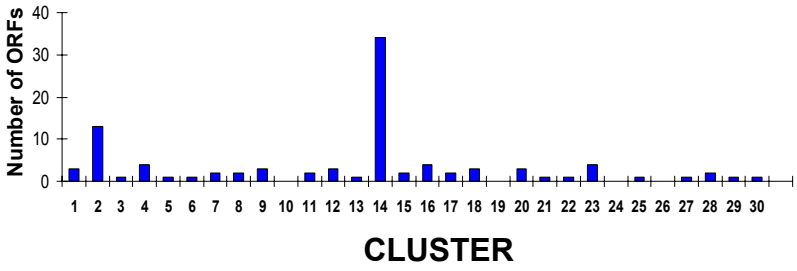
MIPS Functional category (total ORFs)	ORFs within functional category (k)	P-value $-\text{Log}_{10}$
Organization of centrosome (28)	6	6
Nuclear biogenesis (5)	3	5
Organization of cytoskeleton (93)	7	4*

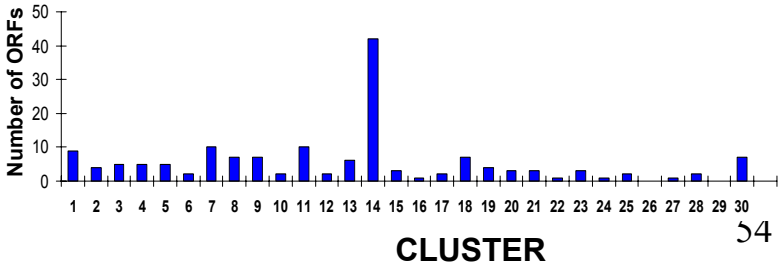


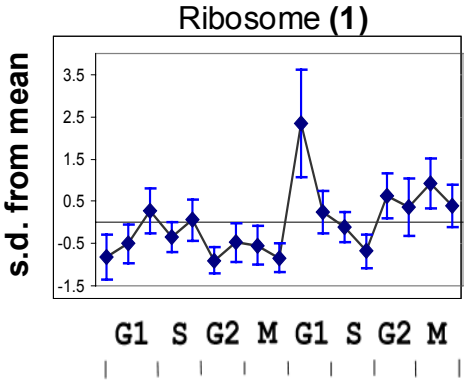






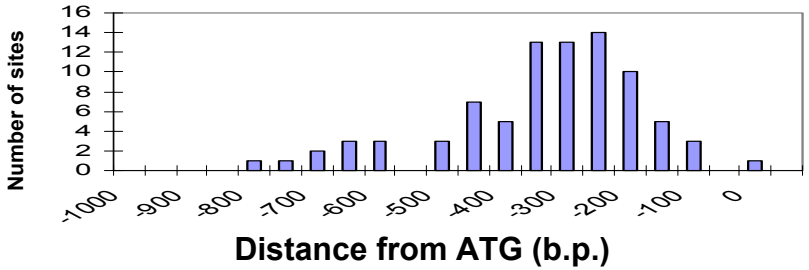


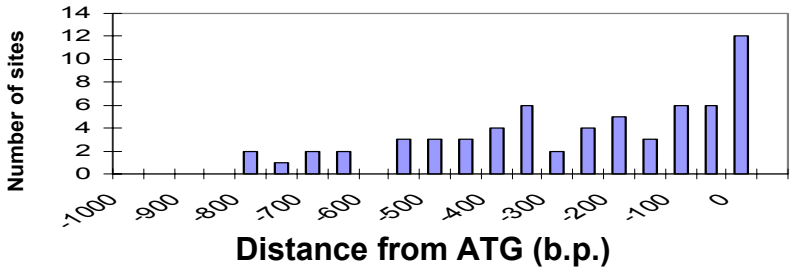


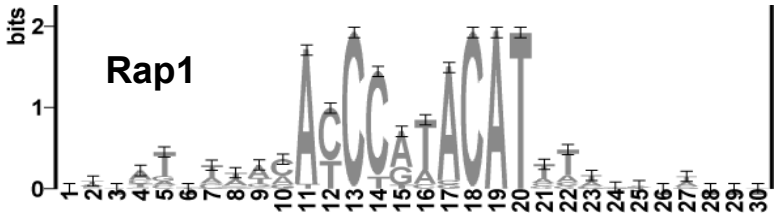


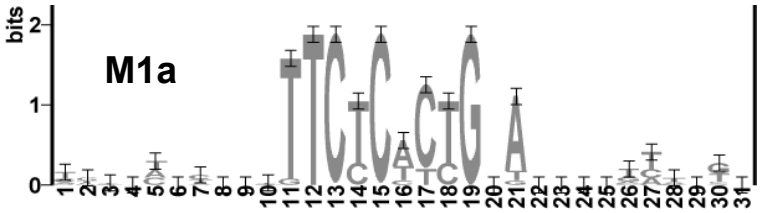
N = 164

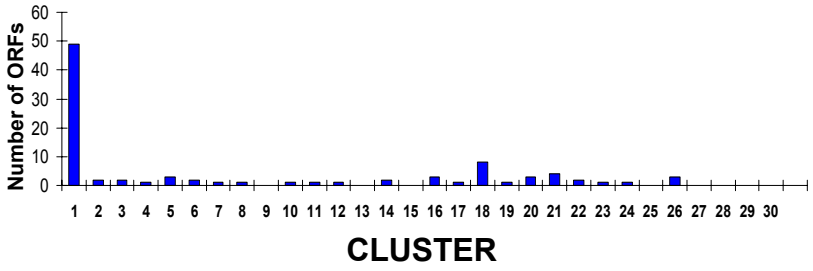
MIPS Functional category (total ORFs)	ORFs within functional category (k)	P-value -Log_{10}
Ribosomal proteins (206)	64	54
Organization of cytoplasm (555)	79	39
Organization of chromosome structure (41)	7	4

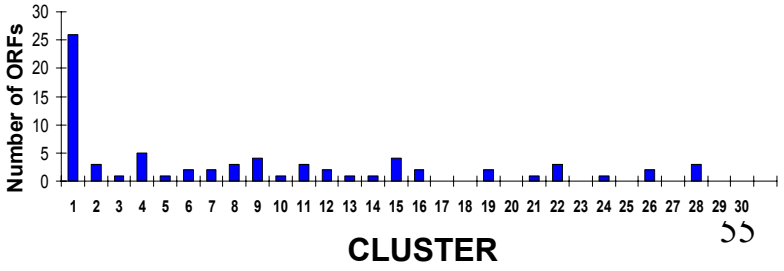




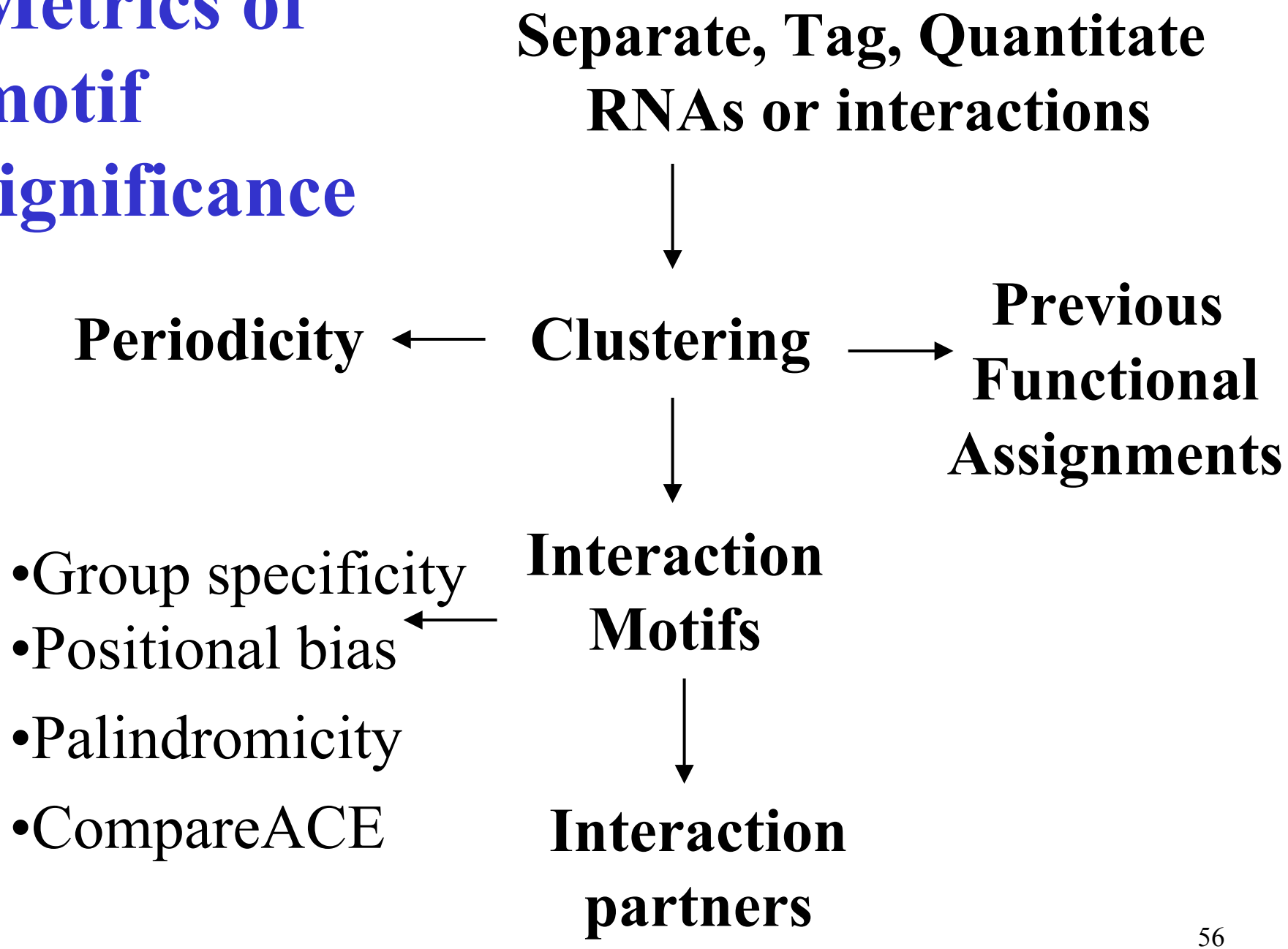








Metrics of motif significance



Functional category enrichment odds

N genes total; s1 = # genes in a cluster; s2= # genes in a particular functional category ("success"); p = s2/N; N=s1+s2-x
Which odds of exactly x in that category in s1 trials?

Binomial: sampling *with* replacement. **(Wrong!)**

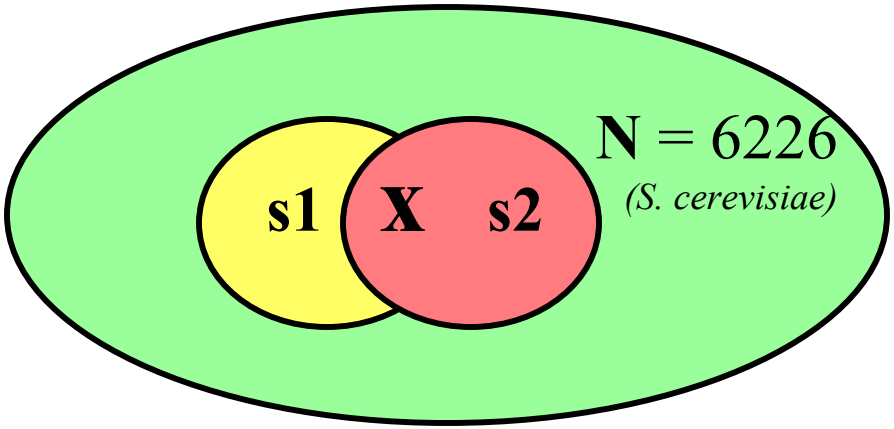
$$B = \binom{s1}{x} p^{x} (1 - p)^{(s1 - x)}$$

or Hypergeometric: sampling *without* replacement:
Odds of getting exactly x = intersection of sets s1 & s2:

$$H = \frac{\binom{s1}{x}\binom{N - s1}{s2 - x}}{\binom{N}{s2}}$$

Ref **(http://library.thinkquest.org/10030/statcon.htm)**

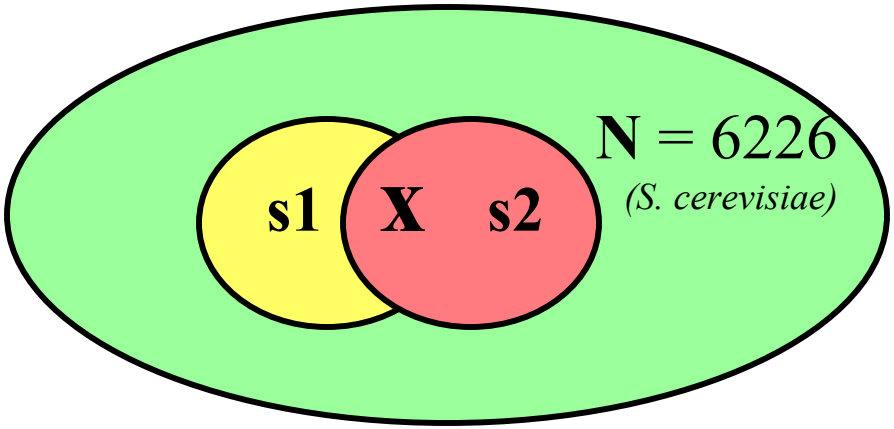
Functional category enrichment



N = Total # of genes (or ORFs) in the genome
s1 = # genes in the cluster
s2 = # genes found in a functional category
x = # ORFs in the intersection of these groups
(hypergeometric probability distribution)

$$S_{function} = \sum_{i=x}^{\min(s1,s2)} \frac{\binom{s1}{i}\binom{N-s1}{s2-i}}{\binom{N}{s2}}$$

Group Specificity Score (S_{group})



N = Total # of genes (ORFs) in the genome
s1 = # genes whose upstream sequences were used to align the motif (cluster)
s2 = # genes in the target list (~ 100 genes in the genome with the best sites for the motif near their translational starts)
x = # genes in the intersection of these groups

$$S_{group} = \sum_{i=x}^{\min(s1,s2)} \frac{\binom{s1}{i}\binom{N-s1}{s2-i}}{\binom{N}{s2}}$$

Positional Bias

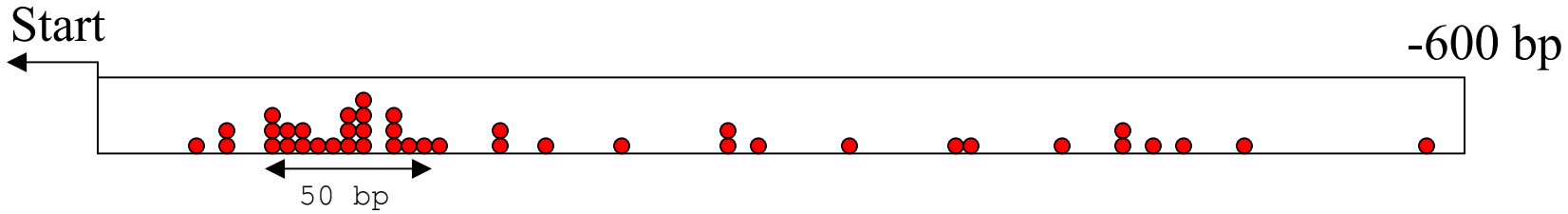
$$P = \sum_{i=m}^{t} \binom{t}{i} \left(\frac{w}{s}\right)^{i} \left(1-\frac{w}{s}\right)^{t-i} \qquad \text{(Binomial)}$$

t = number of sites within 600 bp of translational start from among the best 200 being considered

m = number of sites in the most enriched 50-bp window

s = 600 bp

w = 50 bp



Comparisons of motifs

- The **CompareACE** program finds best alignment between two motifs and calculates the correlation between the two position-specific scoring matrices
- Similar motifs: CompareACE score > 0.7

Clustering motifs by similarity

Cluster motifs using a similarity matrix
consisting of all pairwise CompareACE scores

motif **A**

motif **B**

motif **C**

motif **D**

CompareACE

	A	B	C	D
A	1.0	0.9	0.1	0.0
B		1.0	0.2	0.1
C			1.0	0.8
D				1.0

Hierarchical Clustering

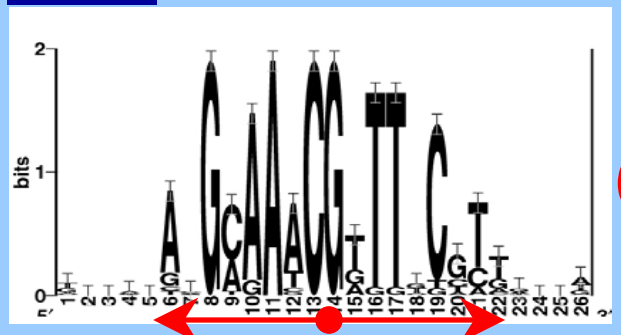
cluster 1: **A**, **B**

cluster 2: **C**, **D**

Palindromicity

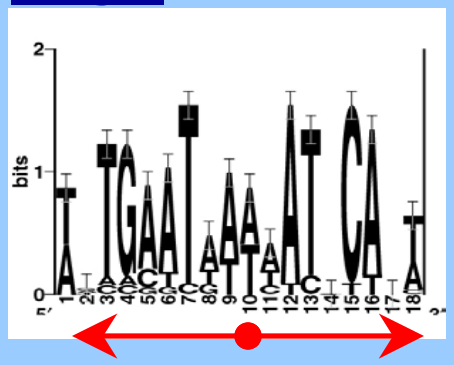
- CompareACE score of a motif versus its reverse complement
- Palindromes: CompareACE > 0.7
- Selected palindromicity values:

PurR



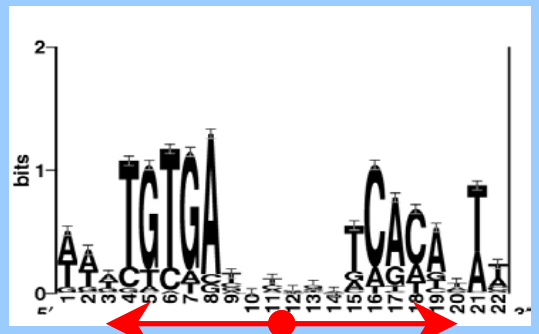
0.97

ArgR



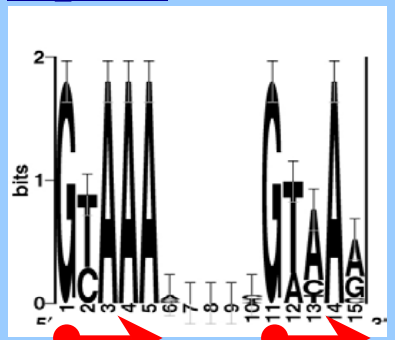
0.92

Crp



0.92

CpxR



0.39

S. cerevisiae AlignACE test set

- Functional categories (248 groups → 3313 motifs)
  - MIPS (135 goups)
  - YPD (17 groups)
  - names (96 groups)

- Negative controls (250 groups → 3692 motifs)
  - 50 each of randomly selected sets of 20, 40, 60, 80, or 100 genes

- Positive controls (29 groups)
  - Cold Spring Harbor website -- SCPD
  - 29 sets of genes controlled by a TF with 5 or more known binding sites

Most specific motifs

(ranked by S_{group})

Cluster	MAP	Spec	PosBias	Logo	Notes
1	231.8	3.0e-46	1.5e-07		Rap1
2	128.2	3.1e-32	3.0e-10		Rpn4
3	31.1	5.4e-23	1.6e-3		Gcn4
4	22.9	6.9e-20	2.2e-3		HSE
5	29.8	1.1e-15	9.0e-4		Mig1/STRE
6	17.4	1.3e-14	1.9e-4		Hap2,3,4
7	30.8	3.1e-14	9.3e-4		Cbf1
8	39.3	1.3e-13	3.5e-08		MCB
9	25.2	2.0e-13	3.5e-08		Lys14
10	19.3	2.1e-12	2.6e-3		Leu3
11	102.2	2.3e-12	2.0e-43		
12	17.1	2.7e-12	3.7e-3		
13	12.3	3.3e-12	2.0e-4		
14	20.6	1.0e-11	1.1e-2		Met31,32
15	29.3	1.2e-11	2.6e-4		ECB
16	24.6	1.4e-11	2.8e-4		Acr1
17	20.2	2.0e-11	3.2e-4		
18	28.0	1.1e-10	1.7e-4		CCA

Most positionally biased motifs

Cluster	MAP	Spec	PosBias	Logo	Notes
1	21.0	0.5	4.1e-175		
2	73.9	0.7	5.8e-92		AT repeats
3	28.3	0.08	1.4e-48		
4	22.3	3.0e-4	2.0e-43		SP11
5	23.8	3.3e-3	1.5e-35		Reb1
6	29.5	1.0e-3	2.7e-33		PAC
7	14.3	2.9e-3	1.5e-31		Abf1
8	26.7	0.95	1.2e-19		
9	32.6	2.2e-16	1.3e-19		GT repeats
10	125.4	9.5e-29	1.1e-14		Rpn4
11	12.5	8.1e-3	6.5e-11		
12	12.9	0.07	1.4e-10		
13	13.2	7.5e-06	7.0e-10		
14	10.5	9.7e-05	5.0e-09		MCB
15	13.0	0.11	5.4e-09		

Negative Controls

- 250 AlignACE runs on 50 groups each of 20, 40, 60, 80,and 100 orfs, resulting in 3692 motifs.
- Allows calibration of an expected false positive rate for a set of hypotheses resulting from any chosen cutoffs.

Example:

MAP > 10.0
Spec. < 1e-5

Functional Categories → 82 motifs (24 known)

Random Runs → 41 motifs

Computational identification of cis-regulatory elements associated with groups of functionally related genes in *S. cerevisiae* Hughes, et al JMB, 1999.

Positive Controls

- 29 transcription factors listed on the CSH web site have five or more known binding sites. AlignACE was run on the upstream regions of the corresponding genes.
- An appropriate motif was found in 21/29 cases.
- 5/8 false negatives were found in appropriate functional category AlignACE runs.
- False negative rate = ~ 10-30 %

Establishing regulatory connections

- Generalizing & reducing assumptions:
- Motif Interactions: (Pilpel et al 2001 Nat Gen) **(http://arep.med.harvard.edu/pdf/Pilpel01.pdf)**
- Which protein(s): in vivo crosslinking
- Interdependence of column in weight matrices: array binding (Bulyk et al 2001PNAS 98: 7158) **(http://arep.med.harvard.edu/pdf/Bulyk01.pdf)**

RNA2: Today's story & goals

- Clustering by gene and/or condition
- Distance and similarity measures
- Clustering & classification
- Applications
- DNA & RNA motif discovery & search