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

Data Normalization | Distance Metric | Linkage | Clustering Method

**Data**
- Ratios
- Log Ratios
- Absolute Measurement

How to normalize
- Variance normalize
- Mean center normalize
- Median center normalize

What to normalize
- genes
- conditions

- Euclidean Dist.
- Manhattan Dist.
- Sup. Dist.
- Correlation Coeff.

- Single
- Complete
- Average
- Centroid

Unsupervised | Supervised

Hierarchical | Non-hierarchical
- Minimal Spanning Tree

- SVM
- Relevance Networks

- K-means
- SOM
(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

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


Cheng, ISMB 2000.

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.
• 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:

$$x = (x_1 x_2 \cdots x_p)$$
$$y = (y_1 y_2 \cdots y_p)$$

The Minkowski metric is defined by

$$d(x, y) = r \sqrt[p]{\sum_{i=1}^{p} |x_i - y_i|^r}$$
Most Common Minkowski Metrics

1, $r = 2$ (Euclidean distance )

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

2, $r = 1$ (Manhattan distance)

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

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

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

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

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeneA</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GeneB</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**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 \( \bar{y} = \frac{1}{p} \sum_{i=1}^{p} y_i \).

\[ |s(x, y)| \leq 1 \]
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

Expression Level

Gene A

Gene B

Time

Expression Level

Gene B

Gene A

Time

Expression Level

Gene B

Gene A

Time
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

Distance Matrix
Complete-Link Method

Euclidean Distance

Distance Matrix
Dendrograms

Single-Link

Complete-Link
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

Data Normalization | Distance Metric | Linkage | Clustering Method

Data
- Ratios
- Log Ratios
- Absolute Measurement

- Euclidean Dist.
- Manhattan Dist.
- Sup. Dist.
- Correlation Coeff.

- Single
- Complete
- Average
- Centroid

Unsupervised | Supervised
- SVM
- Relevance Networks

How to normalize
- Variance normalize
- Mean center normalize
- Median center normalize

Hierarchical | Non-hierarchical
- Minimal Spanning Tree

What to normalize
- genes
- conditions

- K-means
- SOM
Identifying prevalent expression patterns (clusters)

See Tavazoie et al. 1999 (http://arep.med.harvard.edu)
Representation of expression data

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene N</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Normalized Expression Data from microarrays

Time-point 1

Time-point 2

Time-point 3

$di_j$
Identifying prevalent expression patterns (gene clusters)
Cluster contents

Genes
- gpm1
- HTB1
- RPL11A
- RPL12B
- RPL13A
- RPL14A
- RPL15A
- RPL17A
- RPL23A
- TEF2
- YDL228c
- YDR133C
- YDR134C
- YDR327W
- YDR417C
- YKL153W
- YPL142C

MIPS functional category
- Glycolysis
- Nuclear Organization
- Ribosome
- Translation
- 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?

- 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
*(Aligns nucleic Acid Conserved Elements)*


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

**The Target Motif**

5' - TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAATGAAAAATTCATGAGAAAAGAGTCA GACATCGAAACATACAT

5' - ATGGCAGAATCACTTTAAAAACGTGGCCCCACCCGCTGCAACCCTGTGCTATTGTACGTTACTGCAGAAATGACTCAACG

5' - CACATCCACGAATCACCCTACCATTTACCTGACTCACTTTCTTTCGCGATCCGCAAGTGACATCGGCAAAATGATAATATTTTTT

5' - TGCGAACAAAAGAGTCAATTACAAGGAAAATAGAAAGAAAATGAAAAATTTTTCGACAATAATGTATATAGTCAATTCTTATC

5' - ACAAGGCTACCTCTCTGGCCAATCTCACAGATTTAAATAGTAAATTGTCAATGATA TGACTCATCCGAACATGAA ARO1

5' - ATTGAT TGACTCATTTTCCTCTGACTCAATCACATTCAAAAAATGTGAAGGAAAATAGAAAGCAGAAAAATATAAAATAA HOM2

5' - GCCGCACAGTCCCGGTGGTTATCCGCGTGACTCATTTCTGACCTTTTTGGGAAATGTGCGCATGTGACTTCAACACA PRO3

**MAP score = 20.37 (maximum)**

aaaagagtca

aaatgacctca

aaatgacctca

aaatgacctca

aaatgacctca

aaatgacctca

***********
AlignACE Example
Initial Seeding

5’- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATTGAAAAATTCAGACATCGAAACATACAT [HIS7]
5’- ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG [ARO4]
5’- CACATCCAAACGAATCACCCTCACCGTTATCGGACTCATCTTTCTTTCGACATCGCGAACGAAAGTGACATAAAAAATATTTTTT [ILV6]
5’- TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC [THR4]
5’- ACAAGAGGTACCTTCTGCGCACACAGATTATATATAAGAATTTTTCGACAAATGGATAGTCATTTCTATC [ARO1]
5’- ATGATTGACTCATTTTCCTCTGACTCATCTACGATTATAAGAAATGGATAGTCATTTCTATC [HOM2]
5’- GCGCGCACAGTCCGGTTTGTTATCCTGGCTGACTCATCTCTTTTTGGAAGTGGGCATGTGCTTCACA [PRO3]

TGAAAAATTC
GACATCGAAA
GCACTTCGCC
GAGTCATTAC
GTAAATTGTC
CCACAGTCCG
TGTGAAGCCAC
**********

MAP score = -10.0
AlignACE Example

Sampling

Add?

5’- TCTCTCTTCCACGGCTAATTAGGTGATCATGAAAAAAATGAAAAATTCATGAGAAAGAGTCA
   GACATCGAAACATACAT
   TGCTACAGAATCACTTTAAAACGTGGCCCCACCCGCCCTGCACCTGTGCTATTTTGACGTATTCTGGCAATGACTCAACG

5’- ATGGCAGAATCACTTTAAAAACGTGGCCCCACCCGCCCTGCACCTGTGCTATTTTGACGTATTCTGGCAATGACTCAACG
   ARO4

5’- CACATCCAACGAATCACCTCACCGTTATCGTACTCAGTTTCTTTCGCAATTGGAAATCGAATGACTCAACG
   ILV6

5’- TGCGAACAAAAAGGTTCATTACAAAGGAGAATAGAGAAATGAAAAATTTCGACAAAATGTATAGTCATTTCTATC
   THR4

5’- ACAACGGATCCCTCTGGCCAATCTCAGATTAACTGATGTGGGTACATGCTATGACTCATCCACCAACATGAAA
   ARO1

5’- ATGGATGGACTCATTTCCTCTGACTCTACCCAGTTCCAATGTTAGAGAAAAATTAGAAAAAGCAGAACAAATAAATA
   HOM2

5’- GCCGCGACAGTCCCGCTTTTGGTTATCCGGCTGACTCTTCTTTTTTGAGAAGTGGGCTGACCTGCTTCACAGAAA
   PRO3

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

---

**TGAAAAATTC**
**GACATCGAAA**
**GCCCTTCCGC**
**GAGTCATTAC**
**GTAAATTTGTC**
**CCACAGTCCG**
**TGTGAAGCAC**

---

**TCTCTCTTCCA**
**TGAAAAATTC**
**GACATCGAAA**
**GCCCTTCCGC**
**GAGTCATTAC**
**GTAAATTTGTC**
**CCACAGTCCG**
**TGTGAAGCAC**

---

**********
AlignACE Example
Continued Sampling

Add? Remove

5'- TCTCTCTCCACGGCTAATTAGT GATC ATGAAAAATGAAAATTC ATGAGAAAAGAGTCA GACATCGAAA CATACAT

5'- ATGGCAGAATC A CTTTTAAAAACGTG G CCCCACCCGCACCTGTGCA TTTTGTACGT A TGC AATGACTCAACG

5'- CACATCCAACGAATC A CCTCACCAGTTATCGT GACTCACT TTTCTTTCG CATGCCGAAGTCACAATAAAAATA TTTT

5'- TGC AACAAAAAAGAGT CATTAC A AC GAGGAAATAGAAGAAATGAAAAATTTTCGACAAAATGTATAGTC ATTTCTATC

5'- ACAAAGGTACCTT C TTGGCAATC TACAGATTTTAATATA GTAAATTGTC ATG CATATG ACTCATC TCCGAACACATGA AA

5'- ATGTATGTGACT CATATTTTTCCCTG T CACTA C CAGTTCAAAAATTTAGAAAGA AAATGACAAGAAAATATAAAAT

5'- GCCG CCA CAGTC CGC GTTTG GTTTATCCCGG CTGACTCAT TCTTTTTTTGGAAGTGTG GCA TGT GCTT CACA

ATGAAAAAT

TGAAAAATT

GACATCGAAA

GCACTTCG GC

GAGT CATTAC

GTAAATTGTC

CCACAGT CCG

TGTGAAGC A C

************

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

ATGAAAAAT

TGAAAAATT

GACATCGAAA

GCACTTCG GC

GAGT CATTAC

GTAAATTGTC

CCACAGT CCG

TGTGAAGC A C

************
AlignACE Example
Continued Sampling

Add?

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

TGAAAAATTC
GACATCGAAA
GCACTTCGGC
GAGTCATTAC
GTAAATTGTC
CCACAGTCCG
TGTGAAGCAC

************

TGAAAAATTC
GACATCGAAA
GCACTTCGGC
GAGTCATTAC
GTAAATTGTC
CCACAGTCCG
TGTGAAGCAC

************
AlignACE Example

Column Sampling

5’- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAATGAAAAATTCATGAGAAAAGAGTCA
GACATCGAAAAACATACAT...HIS7
5’- ATGGCAGAATCACTTTAAACGTGGCCCACCACCCTGTGCATTTTGTACGTACTGCAATGACTCAACG
5’- CACATCCAACGAATCACCTCACCTACGGTTATCGTACTCACTTTCTTCTCGCATCGCCGAAGTGCACTATAAAAAATATTTTTT
5’- TGCGAACAAAAAGAGTCATTACAACGAGGAAATAGAAAGAAATGAAAAATTTTCGACAAAAATGTATAGTCATTTATATC
5’- ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATA
5’- GAGTCATTACAACGAGGAAATAGAAAGAAATGAAAAATTTTCGACAAAAATGTATAGTCATTTATATC

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

The Best Motif

5' - TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAATGAAAAATTCATGAGAAAAGAGTCA
5' - ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCCTGTGCATTTTGTACGTTACTGCGAAATGACTCA
5' - CACATCCAACGAATCACCACCGTTATCGTGACTCACCCTTTTTTGCATCGCCGAAATGCCCATAAAAATATTTTTT
5' - TGCGAACAAAAGAGTCAATTACAACGAGGAAATAGGAAAGAAATGAAAAATTTTCGACAAATGACTCATTTCTATC
5' - ACAAGGTACCTCTCTTGGCCAAATTCACAGATTTTAATATAGTAATATGCATGCAATTGACTCATTCCGAAACATGAAA
5' - ATGACTCATTTCCTCTGACTACTCGATTTAAATGTGAGGAAATGAAAAGCAGAAAAAATAAATA
5' - GCACGACAGTCCCGGTTGGTATTCGCTGTCCGCTGACTTCTTTTTGGAAGTGTGCGCATGCTTCACACAA

AAAAGAGTCA
AAATGACTCA
AGTGACTCA
AAAAGAGTCA
GGATGATGCA
AAATGACTCA
GAATGACTCA
AAAAGAGTCA
**********

MAP score = 20.37
• 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)**

- 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**

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

B, \Gamma = \text{standard Beta & Gamma functions}
N = \text{number of aligned sites}; T = \text{number of total possible sites}
F_{jb} = \text{number of occurrences of base } b \text{ at position } j \ (F = \text{sum})
G_b = \text{background genomic frequency for base } b
\beta_b = n \times G_b \text{ for } n \text{ pseudocounts} \quad (\beta = \text{sum})
W = \text{width of motif}; C = \text{number of columns in motif} \quad (W \geq C)
MAP Score

$$\text{MAP} \Leftarrow N \log R$$

N = number of aligned sites
R = overrepresentation of those sites.
AlignACE Example: Final Results

(MAP score)

<table>
<thead>
<tr>
<th>MAP score</th>
<th>Motif</th>
</tr>
</thead>
<tbody>
<tr>
<td>188.3</td>
<td>ATATAATAATA</td>
</tr>
<tr>
<td>78.1</td>
<td>CCACATTT</td>
</tr>
<tr>
<td>20.6</td>
<td>CCTACGTCG</td>
</tr>
<tr>
<td>28.1</td>
<td>TTTTTTTTT</td>
</tr>
<tr>
<td>117.5</td>
<td>GAATCA</td>
</tr>
<tr>
<td>31.1</td>
<td>TATATATATA</td>
</tr>
<tr>
<td>73.4</td>
<td>AACAAGA</td>
</tr>
<tr>
<td>8.2</td>
<td>CCGTGG</td>
</tr>
<tr>
<td>19.3</td>
<td>TGAAAAAA</td>
</tr>
<tr>
<td>55.0</td>
<td>AAAA</td>
</tr>
<tr>
<td>89.4</td>
<td>ATCGCTG</td>
</tr>
<tr>
<td>2.7</td>
<td></td>
</tr>
</tbody>
</table>

(Alignment of upstream regions from 116 amino acid biosynthetic genes in S. cerevisiae)
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
- \(n_{l0}\) = number of occurrences of the most common base at position \(l\) in the input alignment
Replication & DNA synthesis (2)

MIPS Functional category (total ORFs)       ORFs within functional category       P-value

| 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  |

N = 186
### Organization of centrosome (14)

**N = 74**

<table>
<thead>
<tr>
<th>MIPS Functional category (total ORFs)</th>
<th>ORFs within functional category</th>
<th>P-value $-\log_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization of centrosome (28)</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Nuclear biogenesis (5)</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Organization of cytoskeleton (93)</td>
<td>7</td>
<td>4*</td>
</tr>
</tbody>
</table>

**Number of ORFs**

#### CLUSTER

<table>
<thead>
<tr>
<th>Number of ORFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30</td>
</tr>
</tbody>
</table>

**Distance from ATG (b.p.)**

- **M14a**
- **M14b**

**Number of sites**

<table>
<thead>
<tr>
<th>Distance from ATG (b.p.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1000 -900 -800 -700 -600 -500 -400 -300 -200 -100 0</td>
</tr>
</tbody>
</table>

**Number of sites**

<table>
<thead>
<tr>
<th>Distance from ATG (b.p.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1000 -900 -800 -700 -600 -500 -400 -300 -200 -100 0</td>
</tr>
</tbody>
</table>

**Number of ORFs**

<table>
<thead>
<tr>
<th>Number of ORFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30</td>
</tr>
</tbody>
</table>
Ribosome (1)

MIPS Functional category (total ORFs)

<table>
<thead>
<tr>
<th>MIPS Functional category</th>
<th>ORFs within functional category</th>
<th>P-value -Log10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ribosomal proteins (206)</td>
<td>64</td>
<td>54</td>
</tr>
<tr>
<td>Organization of cytoplasm (555)</td>
<td>79</td>
<td>39</td>
</tr>
<tr>
<td>Organization of chromosome structure (41)</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>

N = 164
Metrics of motif significance

Separate, Tag, Quantitate RNAs or interactions

Clustering

Previous Functional Assignments

Periodicity

Interaction Motifs

Interaction partners

• Group specificity
• Positional bias
• Palindromicity
• CompareACE
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.

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

**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

\[ S'_{function} = \sum_{i=x}^{\min(s_1, s_2)} \binom{s_1}{i} \binom{N-s_1}{s_2-i} \binom{N}{s_2} \]

\( N = \) Total # of genes (or ORFs) in the genome
\( s_1 = \) # genes in the cluster
\( s_2 = \) # genes found in a functional category
\( x = \) # ORFs in the intersection of these groups

(hypergeometric probability distribution)

\( N = 6226 \) (\textit{S. cerevisiae})

\( s_1, s_2, N, x \)
Group Specificity Score ($S_{\text{group}}$)

$N = \text{Total # of genes (ORFs) in the genome}$

$s1 = \text{# genes whose upstream sequences were used to align the motif (cluster)}$

$s2 = \text{# genes in the target list (~100 genes in the genome with the best sites for the motif near their translational starts)}$

$x = \text{# genes in the intersection of these groups}$

$$S_{\text{group}} = \sum_{i=x}^{\min(s1, s2)} \frac{s1}{i} \frac{N-s1}{s_2-i} \frac{N}{s_2}$$

$N = 6226$ (S. cerevisiae)
Positional Bias

\[ P = \sum_{i=m}^{t} \binom{t}{i} \left( \frac{w}{s} \right)^i \left( 1 - \frac{w}{s} \right)^{t-i} \]  

\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
- \( s = 600 \text{ bp} \)
- \( w = 50 \text{ bp} \)
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

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 groups)
  - 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 $S_{\text{group}}$)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>MAP</th>
<th>Spec</th>
<th>PosBias</th>
<th>Logo</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>231.8</td>
<td>3.0e-46</td>
<td>1.5e-07</td>
<td>ATCCATACAT</td>
<td>Rap1</td>
</tr>
<tr>
<td>2</td>
<td>128.2</td>
<td>3.1e-32</td>
<td>3.0e-10</td>
<td>GGTGGCAAAA</td>
<td>Rpn4</td>
</tr>
<tr>
<td>3</td>
<td>31.1</td>
<td>5.4e-23</td>
<td>1.6e-3</td>
<td>aaAIgATCA</td>
<td>Gcn4</td>
</tr>
<tr>
<td>4</td>
<td>22.9</td>
<td>6.9e-20</td>
<td>2.2e-3</td>
<td>CIAGAA_TTC</td>
<td>HSE</td>
</tr>
<tr>
<td>5</td>
<td>29.8</td>
<td>1.1e-15</td>
<td>9.0e-4</td>
<td>ssp_T_CCcCC_Cc</td>
<td>Mig1/STRE</td>
</tr>
<tr>
<td>6</td>
<td>17.4</td>
<td>1.3e-14</td>
<td>1.9e-4</td>
<td>CCAAT_A</td>
<td>Hap2,3,4</td>
</tr>
<tr>
<td>7</td>
<td>30.8</td>
<td>3.1e-14</td>
<td>9.3e-4</td>
<td>CACGTGA_TTC</td>
<td>Cbf1</td>
</tr>
<tr>
<td>8</td>
<td>39.3</td>
<td>1.3e-13</td>
<td>3.5e-08</td>
<td>T_AGCGGT</td>
<td>MCB</td>
</tr>
<tr>
<td>9</td>
<td>25.2</td>
<td>2.0e-13</td>
<td>3.5e-08</td>
<td>TTTCC_GGgAT</td>
<td>Lys14</td>
</tr>
<tr>
<td>10</td>
<td>19.3</td>
<td>2.1e-12</td>
<td>2.6e-3</td>
<td>CCGgT_gGG</td>
<td>Leu3</td>
</tr>
<tr>
<td>11</td>
<td>102.2</td>
<td>2.3e-12</td>
<td>2.0e-43</td>
<td>TGAAAAATTT</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>17.1</td>
<td>2.7e-12</td>
<td>3.7e-3</td>
<td>T_AAcCCCgA</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>12.3</td>
<td>3.3e-12</td>
<td>2.0e-4</td>
<td>C_g_CCACg_Cg</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>20.6</td>
<td>1.0e-11</td>
<td>1.1e-2</td>
<td>_g_CCACACgIT</td>
<td>Met31,32</td>
</tr>
<tr>
<td>15</td>
<td>29.3</td>
<td>1.2e-11</td>
<td>2.6e-4</td>
<td>T_gC_AT_TGGgA</td>
<td>ECB</td>
</tr>
<tr>
<td>16</td>
<td>24.6</td>
<td>1.4e-11</td>
<td>2.8e-4</td>
<td>G_gC_gGGgC_</td>
<td>Acr1</td>
</tr>
<tr>
<td>17</td>
<td>20.2</td>
<td>2.0e-11</td>
<td>3.2e-4</td>
<td>CgGgAg_TTCgG</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>28.0</td>
<td>1.1e-10</td>
<td>1.7e-4</td>
<td>A_ggGAA_TTTGAA</td>
<td>CCA</td>
</tr>
</tbody>
</table>
## Most positionally biased motifs

<table>
<thead>
<tr>
<th>Cluster</th>
<th>MAP</th>
<th>Spec</th>
<th>PosBias</th>
<th>Logo</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21.0</td>
<td>0.5</td>
<td>4.1e-175</td>
<td>AAA_.._Aaa_aaA</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>73.9</td>
<td>0.7</td>
<td>5.8e-92</td>
<td>T_I_I_I_I_I_I_I_I</td>
<td>AT repeats</td>
</tr>
<tr>
<td>3</td>
<td>28.3</td>
<td>0.08</td>
<td>1.4e-48</td>
<td>T_I_I_I_I_I_I_I_I</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>22.3</td>
<td>3.0e-4</td>
<td>2.0e-43</td>
<td>T_S_A_A_A_A_A_T_T_T</td>
<td>SP11</td>
</tr>
<tr>
<td>5</td>
<td>23.8</td>
<td>3.3e-3</td>
<td>1.5e-35</td>
<td>T___CGG_GTAA_T</td>
<td>Reb1</td>
</tr>
<tr>
<td>6</td>
<td>29.5</td>
<td>1.0e-3</td>
<td>2.7e-33</td>
<td>T_G_GAT_GA_A</td>
<td>PAC</td>
</tr>
<tr>
<td>7</td>
<td>14.3</td>
<td>2.9e-3</td>
<td>1.5e-31</td>
<td>A_T_C_A___A_a_g</td>
<td>Abf1</td>
</tr>
<tr>
<td>8</td>
<td>26.7</td>
<td>0.95</td>
<td>1.2e-19</td>
<td>AAA_g___G_A_A_A_A</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>32.6</td>
<td>2.2e-16</td>
<td>1.3e-19</td>
<td>G_T_T_G_G_G_G</td>
<td>GT repeats</td>
</tr>
<tr>
<td>10</td>
<td>125.4</td>
<td>9.5e-29</td>
<td>1.1e-14</td>
<td>T_T_T_G_C_C_A_C</td>
<td>Rpn4</td>
</tr>
<tr>
<td>11</td>
<td>12.5</td>
<td>8.1e-3</td>
<td>6.5e-11</td>
<td>A_A_A_T__A_A_A_A_A</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>12.9</td>
<td>0.07</td>
<td>1.4e-10</td>
<td>T_I_I_I_I_I_I_I_I</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>13.2</td>
<td>7.5e-06</td>
<td>7.0e-10</td>
<td>T_A_T_A_T_A_T_A</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>10.5</td>
<td>9.7e-05</td>
<td>5.0e-09</td>
<td>T_A_C_G_C_G_T__C</td>
<td>MCB</td>
</tr>
<tr>
<td>15</td>
<td>13.0</td>
<td>0.11</td>
<td>5.4e-09</td>
<td>A_A_A__G_A_A_G</td>
<td></td>
</tr>
</tbody>
</table>
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 | Functional Categories | 82 motifs (24 known) |
| Spec. < 1e-5 | Random Runs | 41 motifs |

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 2001 PNAS 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