Motif Finding: Summary of Approaches
Lecture Outline

• Flashback: Gene regulation, the cis-region, and tying function to sequence
• Motivation
• Representation
  – simple motifs
  – weight matrices
• Problem: Finding motifs in sequences
• Approaches
  – enumerative (combinatorial)
  – statistical
• Comparison of approaches
• Higher Order Motifs and Approaches
Motif Finding Motivation

Clustering genes based on their expressions groups co-expressed genes

Assuming co-expressed genes are co-regulated, we look in their promoter regions to find conserved motifs, confirming that the same TF binds to them
Motifs vs Transcription Factor Binding Sites

• Motifs:
  – statistical or computational entities
  – predicted

• Transcription Factor Binding Sites (or more generally cis-regulatory elements)
  – biological entities
  – Real

• The hope is that TFBS are conserved, or otherwise significant computationally, so motifs can be used to find them
Finding Motifs in a Set of Sequences

GTGGCTGCACCACGTGTATGC...ACG
ACATCGCATCACGTGACCAGT...GAC
CCTCGCACGTGGTGGTACAGT...AAC
CTCGTTAGGACCATCACGTGA...ACA
GCTAGCCCACGTGGATCTTGT...AGA
Finding Motifs in a Set of Sequences

GGCTGCAC**CACGTG**TATGC...ACGATGTCTCGC
ATCGCAT**CACGTG**ACCAGT...GACATGGGACG
TCGC**CACGTG**GGTGGTACAGT...AACATGACTAAA
CGTTAGGACC**CACGTG**A...ACAATGAGAGCG
TAGCC**CACGTG**GATCTTTGT...AGAATGGCCCTAT
Finding Motifs in a Set of Sequences

TCTGCACACGTGTATGCA...ACGATGTCTTCGC
ATCGCATCACGTGACCATG...GACATGGACGGGC
GCCTCGCACGTGGTGACGTACAGT...AACATGAC
GGACCACACGTGA...ACAATGAGAGCG
GCTAGGCACACGTGGATCTTTGT...AGAATGGGCC

Protein binding
Phylogenetic Footprinting

• Finding overrepresented short sequences in cis-regions
• Based on multiple alignment but short sequences don’t have to be completely conserved
• Ex. FootPrinter (Blanchette and Tompa 2003)
Motif Finding Problem

Given \( n \) sequences, find a motif (or subsequence) present in many

This is essentially multiple alignment. The difference is that multiple alignment is global

- longer overlaps
- constant site sizes and gaps
- NP-complete!
Definition and Representation

• **Motifs:** Short sequences
• **IUPAC notation**
• **Regular Expressions**
  – consensus motif
    \[\text{ACGGGTA}\]
  – degenerate motif
    \[\{G|A\}\text{CGGGT}\{A|C\}\]

<table>
<thead>
<tr>
<th>Single-Letter Codes for Nucleotides</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Symbol</strong></td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>T</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>U</td>
</tr>
<tr>
<td>R</td>
</tr>
<tr>
<td>Y</td>
</tr>
<tr>
<td>M</td>
</tr>
<tr>
<td>K</td>
</tr>
<tr>
<td>S</td>
</tr>
<tr>
<td>W</td>
</tr>
<tr>
<td>H</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>V</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td>N</td>
</tr>
</tbody>
</table>
Position Specific Information

**Alignment Matrix (Profile)**

<table>
<thead>
<tr>
<th>Pos</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Position (Frequency) Weight Matrix**

<table>
<thead>
<tr>
<th>Pos</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>Conse</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.2</td>
<td>0</td>
<td>0.6</td>
<td>T</td>
</tr>
<tr>
<td>3</td>
<td>0.4</td>
<td>0.2</td>
<td>0.4</td>
<td>0</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0.4</td>
<td>0.6</td>
<td>0</td>
<td>G</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0.6</td>
<td>0.2</td>
<td>0.2</td>
<td>C</td>
</tr>
</tbody>
</table>
1. Use PWM to Find the Motif in any Sequence

Frequency Weight Matrix

<table>
<thead>
<tr>
<th>Pos</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>Conse</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.2</td>
<td>0</td>
<td>0.6</td>
<td>T</td>
</tr>
<tr>
<td>3</td>
<td>0.4</td>
<td>0.2</td>
<td>0.4</td>
<td>0</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0.4</td>
<td>0.6</td>
<td>0</td>
<td>G</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0.6</td>
<td>0.2</td>
<td>0.2</td>
<td>C</td>
</tr>
</tbody>
</table>

Given AAATC and the Weight Matrix of the data and for the background (i.e. prior), we want to calculate the joint probability

In general this is a lot of work, because of all possible ways a motif can depend on its sub-words.

E.g. TATTA=TAT.TA|TA.T.TA|T.A.T.T.A, etc.
2. Given Sequences Find Motifs

• Methods based on Position Weight Matrices (alignment)
  – Gibbs Sampling
  – Expectation Maximization

• Other Methods
  – HMMs
  – Bayesian methods
  – enumerative (combinatorial)
Simple Motif Finding

• Methods based on Position Weight Matrices (alignment)
  – Gibbs Sampling
  – Expectation Maximization

• Other Methods
  – HMMs
  – Bayesian methods
  – enumerative (combinatorial)
Popular Software:

- MEME (EM)
  http://meme.sdsc.edu/meme/website/intro.html
- AlignACE (Gibbs)
  http://atlas.med.harvard.edu/
- Cister (HMM)
  http://zlab.bu.edu/~mfrith/cister.shtml
- YMF (combinatorial)
- MITRA (combinatorial)
  http://www.cs.columbia.edu/compbio/mitra/
- NestedMICA
  http://www.sanger.ac.uk/Software/analysis/nmica
Overall Idea

• Enumerate motifs
• Score motifs based on their over-representation in all sequences
• The highest scoring ones, if occurring at surprising rates, are meaningful

Problems:
- How to enumerate?
- How to score motifs?
- What is surprise?
Using PWMs, main idea

- Capture the data in PWM
- Enumerate and score all patterns, $w$
  - suffix trees used to save space
- Update the PWM
- Scoring: over-representation

$$S = \frac{\text{observed frequency}}{\text{expected frequency}}$$

$w$ in given sequences  $w$ in genome
MEME

- Use Expectation-Maximization Algorithm to fit a two-component mixture model to the sequence data
  - Component 1 is the motif
  - Component 2 is the background

Algorithm:
- For each sequence $s_i$, (out of n)
  - Start with a random PWM, $P_i$ (i.e. alignment)
  - Score every segment of $s_i$ with $P_i$
  - Update $P_i$ = Sum all the scores with appropriate weights
  - Perform EM until there is a convergence

The best 100 scoring motifs are kept overall
Gibbs Sampler

• Use a simple leave-one-out sampling strategy

Algorithm
• Given n sequences, s1, s2, ..., sn
• Randomly initialize PWM (i.e. align)
• For each sequence $s_i$, take it out from the PWM
  - score each segment of $s_i$ with the rest of the sequences
  - put the sequence back
• Important feature: convergence
YMF: Enumeration

• Use a consensus model of motifs based on IUPAC alphabet
• Score motifs based on their significance of occurrence (vs. random)
• Clean up the found motifs to remove redundant motifs
Comparing the Methods

Tompa et al. (2005)

- Compared 13 different methods
- Used real sequences and searched for known binding sites (TRANSFAC)
  - 52 data sets + 4 negative controls
  - 4 organisms represented (fly, human, mouse and yeast)
- Scored methods based on confusion matrix statistics for the top motif observed
1. Assessing Method Performance

Score = Total overlap / Total span  \[\text{(Pevzner & Sze 2000)}\]

Score = 1, if span = overlap
Score = 0, if overlap = 0
2. Comparing Binary Predictors

Comparing a Model to Reality:

Measures of agreement:
- True Positives, True Negatives
- False Positives, False Negatives

Measures of accuracy:
- Accuracy = (TP+TN)/(TP+TN+FP+FN)
- Sensitivity = TP/(TP+FN)
- PPV=TP/(TP+FP)
- Specificity = TN/(FP+TN)
- Correlation coefficient:
  \[
  \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}}
  \]

Confusion matrix

<table>
<thead>
<tr>
<th>Actual (reality)</th>
<th>TP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Predictions (Y,N)</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

Type I error

Type II error

ROC Curves: Tradeoff between Sens. And Specificity (sens vs. 1-spec.)
### Table 2: Number of data sets for which each tool predicted no motif

<table>
<thead>
<tr>
<th>Tool</th>
<th>Total (56)</th>
<th>Fly (18)</th>
<th>Mouse (12)</th>
<th>Human (26)</th>
<th>Yeast (10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlignACE</td>
<td>32</td>
<td>7</td>
<td>5</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>ANN-Spec</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Consensus</td>
<td>37</td>
<td>4</td>
<td>3</td>
<td>26</td>
<td>4</td>
</tr>
<tr>
<td>GLAM</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Improbizer</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MEME</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>MEME3</td>
<td>14</td>
<td>0</td>
<td>5</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>QuickScore</td>
<td>20</td>
<td>2</td>
<td>4</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>SeSiMCMC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MITRA</td>
<td>13</td>
<td>7</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>MotifSampler</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>OligoPyd-analysis</td>
<td>23</td>
<td>1</td>
<td>5</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>Weeder</td>
<td>17</td>
<td>3</td>
<td>3</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>YMF</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

*The total number of data sets is given parenthetically in the column header.

### Table 3: Correlation coefficient (nCC) for all pairs of tools

<table>
<thead>
<tr>
<th></th>
<th>QuickScore</th>
<th>GLAM</th>
<th>SeSiMCMC</th>
<th>MITRA</th>
<th>Cursen</th>
<th>Improb</th>
<th>AlignACE</th>
<th>MotifSampler</th>
<th>MEME3</th>
<th>MEME</th>
<th>OligoPyd</th>
<th>ANN-Spec</th>
<th>YMF</th>
<th>Weeder</th>
</tr>
</thead>
<tbody>
<tr>
<td>QuickScore</td>
<td>0.019</td>
<td>0.028</td>
<td>0.042</td>
<td>0.045</td>
<td>0.025</td>
<td>0.062</td>
<td>0.038</td>
<td>0.068</td>
<td>0.072</td>
<td>0.072</td>
<td>0.074</td>
<td>0.064</td>
<td>0.061</td>
<td>0.069</td>
</tr>
<tr>
<td>GLAM</td>
<td>0.021</td>
<td>0.019</td>
<td>0.066</td>
<td>0.064</td>
<td>0.064</td>
<td>0.086</td>
<td>0.086</td>
<td>0.086</td>
<td>0.086</td>
<td>0.086</td>
<td>0.086</td>
<td>0.086</td>
<td>0.086</td>
<td>0.086</td>
</tr>
<tr>
<td>SeSiMCMC</td>
<td>0.056</td>
<td>0.059</td>
<td>0.079</td>
<td>0.064</td>
<td>0.064</td>
<td>0.083</td>
<td>0.071</td>
<td>0.091</td>
<td>0.081</td>
<td>0.088</td>
<td>0.088</td>
<td>0.088</td>
<td>0.088</td>
<td>0.088</td>
</tr>
<tr>
<td>MITRA</td>
<td>0.047</td>
<td>0.061</td>
<td>0.072</td>
<td>0.064</td>
<td>0.064</td>
<td>0.086</td>
<td>0.086</td>
<td>0.086</td>
<td>0.086</td>
<td>0.086</td>
<td>0.086</td>
<td>0.086</td>
<td>0.086</td>
<td>0.086</td>
</tr>
<tr>
<td>Cursen</td>
<td>0.086</td>
<td>0.066</td>
<td>0.083</td>
<td>0.077</td>
<td>0.056</td>
<td>0.082</td>
<td>0.089</td>
<td>0.114</td>
<td>0.120</td>
<td>0.120</td>
<td>0.120</td>
<td>0.120</td>
<td>0.120</td>
<td>0.120</td>
</tr>
<tr>
<td>Improb</td>
<td>0.065</td>
<td>0.069</td>
<td>0.083</td>
<td>0.072</td>
<td>0.056</td>
<td>0.082</td>
<td>0.089</td>
<td>0.113</td>
<td>0.120</td>
<td>0.120</td>
<td>0.120</td>
<td>0.120</td>
<td>0.120</td>
<td>0.120</td>
</tr>
<tr>
<td>AlignACE</td>
<td>0.088</td>
<td>0.088</td>
<td>0.086</td>
<td>0.085</td>
<td>0.085</td>
<td>0.111</td>
<td>0.068</td>
<td>0.097</td>
<td>0.102</td>
<td>0.102</td>
<td>0.102</td>
<td>0.097</td>
<td>0.097</td>
<td>0.097</td>
</tr>
<tr>
<td>MotifSampler</td>
<td>0.071</td>
<td>0.092</td>
<td>0.107</td>
<td>0.097</td>
<td>0.077</td>
<td>0.103</td>
<td>0.099</td>
<td>0.068</td>
<td>0.112</td>
<td>0.119</td>
<td>0.119</td>
<td>0.103</td>
<td>0.103</td>
<td>0.103</td>
</tr>
<tr>
<td>MEME3</td>
<td>0.089</td>
<td>0.094</td>
<td>0.092</td>
<td>0.102</td>
<td>0.074</td>
<td>0.102</td>
<td>0.093</td>
<td>0.109</td>
<td>0.124</td>
<td>0.126</td>
<td>0.126</td>
<td>0.096</td>
<td>0.096</td>
<td>0.096</td>
</tr>
<tr>
<td>MEME</td>
<td>0.091</td>
<td>0.090</td>
<td>0.100</td>
<td>0.102</td>
<td>0.077</td>
<td>0.091</td>
<td>0.095</td>
<td>0.120</td>
<td>0.120</td>
<td>0.120</td>
<td>0.120</td>
<td>0.094</td>
<td>0.094</td>
<td>0.094</td>
</tr>
<tr>
<td>OligoPyd</td>
<td>0.073</td>
<td>0.088</td>
<td>0.111</td>
<td>0.088</td>
<td>0.082</td>
<td>0.082</td>
<td>0.099</td>
<td>0.136</td>
<td>0.119</td>
<td>0.119</td>
<td>0.119</td>
<td>0.071</td>
<td>0.071</td>
<td>0.071</td>
</tr>
<tr>
<td>ANN-Spec</td>
<td>0.065</td>
<td>0.091</td>
<td>0.111</td>
<td>0.094</td>
<td>0.090</td>
<td>0.100</td>
<td>0.085</td>
<td>0.122</td>
<td>0.114</td>
<td>0.114</td>
<td>0.114</td>
<td>0.089</td>
<td>0.089</td>
<td>0.089</td>
</tr>
<tr>
<td>YMF</td>
<td>0.044</td>
<td>0.095</td>
<td>0.122</td>
<td>0.101</td>
<td>0.093</td>
<td>0.100</td>
<td>0.114</td>
<td>0.144</td>
<td>0.121</td>
<td>0.121</td>
<td>0.121</td>
<td>0.092</td>
<td>0.092</td>
<td>0.092</td>
</tr>
<tr>
<td>Weeder</td>
<td>0.052</td>
<td>0.169</td>
<td>0.167</td>
<td>0.167</td>
<td>0.167</td>
<td>0.167</td>
<td>0.167</td>
<td>0.167</td>
<td>0.167</td>
<td>0.167</td>
<td>0.167</td>
<td>0.167</td>
<td>0.167</td>
<td>0.167</td>
</tr>
</tbody>
</table>

*The primary tool is listed in the row header and the secondary tool in the column header. The score shown for the same tool on both axes (that is, along the main diagonal) is the individual nCC score from Figure 1. Numerical values are color-coded, ranging from dark blue (lower predictions) to red (better predictions).
Multi-site Motif

- Two-site: Dimer, dyad
- Gapped Motif
- In general, a motif is an ordered set of binding sites

Table 3 • Dimer alignment for MCM1 binding site

| .ACC . . . . AGGA . |
| .ACC . . . . GGAA |
| .. CCTA . . AGGA . |
| . ACCT . . . . AAGG . |
| .. CCT . . . . GGAA |
| .. CCTA . . . . GGAA |
| TACC . . . . AAGG . |
| . ACCT . . . . GGA |
| . ACCT . . . . AGGA |
| TACC . . . . GGA |
| TACC . . . . AGGA |
| . ACCT . . . . GGAA |
| TACC . . . . GGAA |
Higher Order Motifs

- Nature of course is more complicated...

- Combinatorial motifs: combinations of binding sites to which an interacting group of TFs binds
- More realistic, but difficult to look for
- Sinha, 2002
What is Nature Like?

Now that we are talking about realistic motifs, what is it that we know about them from biology?

- Combinatorial motifs are sets of simple motifs separated by a stretch of DNA
- Changing the order of the simple motifs within it doesn’t kill transcription, but changes it
- Changing the distance between the simple motifs usually kills transcription
- The distances between motifs are usually small (<20bp)
- The distance restriction is sometimes strict, and other times not
- Randomly distributed simple motifs do not activate transcription
Dependence of Simple Motif Pairs on Distance and Order Between Them

Ohmori et al., 1997
Finding Higher Order Motifs

Sinha (2002) reviews methods for finding higher order motifs, and groups the approaches based on their general relationship to simple motif finders:

- find simple motifs and discover patterns made of these
- start with simple motifs and build higher order ones
- find higher order motifs from scratch (e.g. Marsan and Sagot, 2000)
Models of Higher Order Motifs

- The set model \( \{M_1, M_2, ..., M_k\} \)
- Tuples with distance constraints  
  \((M_1, M_2, d_{12})\)
- Hidden Markov Model
- Boolean Combinations

Usually two step approaches:
  - Enumerate the motif models
  - Determine significance (Monte Carlo experiments)
Tricky Business

• All these models have a lot of parameters (e.g. distances between motifs)
• They depend on the initial choice of parameters and/or an initial set of simple motifs
• Using these tools is more of an art than science so far
Conclusions

• PWMs do well for simple motifs
• Combinatorial methods are probably doing better
• Should use all available tools to determine strong simple motifs
• Higher order motifs:
  – positive: knowing your biochemistry helps
  – negative: nobody knows the biochemistry fully!
References:

• Sinha and Tompa, Performance Comparison of Algorithms for Finding Transcription Factor Binding Sites, BIBE 2003
• Marsan and Sagot, Algorithms for Extracting Structured Motifs Using a Suffix Tree, JCB, v.7, 2000, 345-362
• Ohmori et al., Journal of Biological Chemistry, 1997
• Tompa et al (2005), Nat. Biotech