Projects

- Suggestion:
  - make a schedule
  - bite-size-pieces
- I’m happy to talk/listen/give (bad?) advice - send email
- Motif assessment:
  http://bio.cs.washington.edu/assessment/
AlignAce  (Roth, et al. 1998)

- Lawrence et al.: protein motifs
- Roth et al.: DNA regulatory motifs
- Differences:
  - Genomic background model,
    e.g. yeast Saccharomyces cerevisiae is 62% A-T
  - both strands used
  - overlapping sites prohibited
  - Multiple motifs: find best & mask
  - “MAP” scoring
Rocke & Tompa (Recomb ‘98)

- Gibbs, adapted for gapped motifs
- single “genomic” DNA sequence
Why Gaps

- Biology often tolerates diversity
- 2 similar TFs bind 2 similar sites
- Same TF binds 2 sites (perhaps one better than the other)
- Dimeric TFs often “don’t care” in middle & flexible
- TF and/or DNA may twist/bulge
A Gapped Motif

0 TAT < CCCCCCTCA   C   CTTCG G   CAGCTCCCCCCCATAA
1 ATC < CCCCCCTCA   C   TTTCG G   CAGCTCCCCCCCATAA
2 GTA < CCCCCCTCA GTG   CTTCGCG G   CAGCTCCCCCCCATAA
3 AAT < CCCCCCTCA GTG   TTTGCG G   CAGCTCCCCCCC TAA
Why gaps are hard

- Alignment
- Pairwise -- $O(n^2)$
- Multiple -- $O(n^k)$
- Gibbs/MEME/... require many alignments
- Scoring
R/T Approach - Scores

- WMM
- Relative entropy, aka expected LLR
- Score gaps like background, “minus a small penalty”
R/T Approach - Alignment

- Gibbs replaces 1 string per iteration
- Use pairwise alignment between new string and previously computed alignment of remaining k-1
- Actually align motif against whole genome - Time $O(\text{genome length} \times \text{motif width})$
R/T Approach- “Gibbs”

• discard 0-2 random strings at each iteration

• pick replacement greedily, not by sampling; avoid local max by random restarts (see Rocke’s thesis for more on this)
Test Data

- Haemophilus influenzae
- ~1.8 megabases
- Delete all protein-coding, leaves ~ 350 kb
- Concatenate, separated with markers
- Plus reverse complement, total ~ 700 kb
Figure 2: 160 trials of the basic algorithm on the noncoding genome vs. a random sequence

Motif width=10
A Motif + Context

<table>
<thead>
<tr>
<th>Position</th>
<th>Motif</th>
<th>Context</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>&lt; CGCCCTTTCA &gt;</td>
<td></td>
<td>118666.</td>
</tr>
<tr>
<td>1</td>
<td>&lt; CGCCCTTTCA &gt;</td>
<td></td>
<td>642660.</td>
</tr>
<tr>
<td>2</td>
<td>AAT</td>
<td>CGCCTTTTC &gt; AAA</td>
<td>425287.</td>
</tr>
<tr>
<td>3</td>
<td>ATC</td>
<td>CGCCC-TTCA &gt; TGA</td>
<td>330462.</td>
</tr>
<tr>
<td>4</td>
<td>TTG</td>
<td>CGCCC-TTCA &gt; CTA</td>
<td>558509.</td>
</tr>
<tr>
<td>5</td>
<td>AAC</td>
<td>CGCCCATTCA &gt; ATC</td>
<td>237890.</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>CGCCC-TTCA &gt; CGT</td>
<td>495353.</td>
</tr>
<tr>
<td>7</td>
<td>TCT</td>
<td>CGCCTTTTC &gt; TTG</td>
<td>34553.</td>
</tr>
<tr>
<td>8</td>
<td>&lt; CGCCCTTTTC &gt;</td>
<td></td>
<td>677174.</td>
</tr>
<tr>
<td>9</td>
<td>&lt; CGCCC-TTCA &gt;</td>
<td>GGG</td>
<td>222102.</td>
</tr>
</tbody>
</table>

Figure 1: A sample motif (score 16.6) produced by the basic algorit
Rewindowing

• After convergence, “rewindow” -- choose subset of rows and adjust left/right boundaries to maximize score.

• NP-hard? Use another greedy heuristic
Rewindowing

0 GGA < CGCCCTTTTCAAG > CGG at position 118663.
1 GGA < CGCCCTTTTCAAG > CGG at position 642657.
2 GCT < CGCCC–TTCAGGG > TTC at position 222099.
3 GGA < CGCCCTTTTCAAG > CGG at position 677171.
4 AAA < CGCCC–TTCACGT > AAT at position 495350.

Figure 3: The motif of Figure 1 after rewindowing (score 20.8)
Figure 4: 160 trials of the two-phase algorithm on the noncoding genome vs. a random sequence
A closer look at 35

- 6 almost perfectly identical regions of 5.3 kb, each 3 rRNA genes plus some tRNA genes
- 9% of genome but 50% of high-scoring motifs
- removed 80kb containing them & re-ran
Figure 5: 160 trials of the two-phase algorithm on the noncoding genome with long repeats removed vs. a random sequence
After Removal

0 TCG < GCAGCTCCCCCCCATAAATGG > GTG at position 449120.
1 TCG < GCAGCTCCCCCCCATAAATGG > GTG at position 448927.
2 GCG < ACAGCTCCCCCCCATAAATGG > GTG at position 232857.
3 GCG < CCAGCTCCC-CCGTAAACCGG > GTG at position 88280.

Figure 6: A sample motif (score 25) produced by two phases
More rewinding

<table>
<thead>
<tr>
<th></th>
<th>Original Sequence</th>
<th>New Sequence</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>TCG &lt; GCAGCTCCCCCACATAATGG &gt; GTG</td>
<td>at position 449120.</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>TCG &lt; GCAGCTCCCCCACATAATGG &gt; GTG</td>
<td>at position 448927.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>GCG &lt; ACAGCTCCCCCACATAATGG &gt; GTG</td>
<td>at position 232857.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>GCG &lt; CCAGCTCCCCCACATAATGG &gt; GTG</td>
<td>at position 88280.</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>TAT &lt; CCCCCCTCA--C--CTTCG-G-CAGCTCCCCCACATAATGGGAGAGCAAGAT &gt; TAG</td>
<td>at position 449105.</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>ATC &lt; CCCCCCTCA--C--CTTCG-G-CAGCTCCCCCACATAATGGGAGAGCAAGAT &gt; TAG</td>
<td>at position 448913.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>GTA &lt; TCCCCCTCAGTCCTTCGAGCTCCCCCACATAATGGGAGAGCAAGTT &gt; AAT</td>
<td>at position 232837.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>AAT &lt; CCCCCCTCAGTC--CTTCGCGCCAGCTCCCCTAATGGGAGAGCAAGGG &gt; ATC</td>
<td>at position 88262.</td>
<td></td>
</tr>
</tbody>
</table>

Figure 7: The motif of Figure 6 after seven phases (score 62)

0 & 1 identical for another 55 bases;
5 differences in next 44.
Probably not a TFBS, but not “random”
Summary

• Handles gaps

• avoids full multiple alignment by exploiting good partial alignment

• validation - null model for comparison

• look at data -
  • rewindowing

• rRNA cluster