TOOLS FOR MOTIF AND PATTERN SEARCHING

Prat Thiru
OUTLINE

• What are motifs?
• Algorithms Used and Programs Available
• Workflow and Strategies
• MEME/MAST Demo (online and command line)
Protein Motifs

(protein motifs that help bind to DNA major & minor grooves)

(a) Helix-turn-helix motif
(b) Zinc finger motif
(c) Leucine zipper motif

(A) [Image of protein structure]
(B) [Image of protein structure]
(C) [Image of protein structure]
DNA Motifs
MEME Output
Definitions

- **Motif**: Conserved regions of protein or DNA sequences
- **Pattern**: Qualitative description of a motif
  
  *eg. regular expression* C[AT]AAT[CG]X
- **Profile**: Quantitative description of a motif
  
  *eg. position weight matrix*
Patterns

• Regular Expression Symbols
  ➢ [ ] – OR eg. [GA] means G or A
  ➢ { } – NOT eg. {P,V} means not P or V
  ➢ ( ) – repeats eg. A(3) means AAA
  ➢ X or N or “.” – any

• Complex patterns representation difficult

• Loose frequency information
  eg. [AT] vs 20%A 80%T
Profiles

\[ M_{ij} = \log \left( \frac{p_j}{\bar{p}_j} \right) \]

- \( p_j \) = probability of amino acid \( j \) at position \( k \) in the profile
- \( \bar{p}_j \) = "background" probability of amino acid \( j \) in sequence

<table>
<thead>
<tr>
<th></th>
<th>0.02</th>
<th>0.95</th>
<th>0.01</th>
<th>0.59</th>
<th>0.51</th>
<th>0.40</th>
<th>0.44</th>
<th>0.39</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.09</td>
<td>0.02</td>
<td>0.03</td>
<td>0.13</td>
<td>0.20</td>
<td>0.11</td>
<td>0.05</td>
<td>0.10</td>
</tr>
<tr>
<td>C</td>
<td>0.10</td>
<td>0.01</td>
<td>0.00</td>
<td>0.15</td>
<td>0.13</td>
<td>0.12</td>
<td>0.08</td>
<td>0.09</td>
</tr>
<tr>
<td>G</td>
<td>0.79</td>
<td>0.03</td>
<td>0.96</td>
<td>0.13</td>
<td>0.15</td>
<td>0.37</td>
<td>0.43</td>
<td>0.42</td>
</tr>
</tbody>
</table>

motif T A T A A [AT] [AT] [AT]
Sequence Logos

HEM13  CCCATTTTTC
HEM13  TTTCGGTTC
HEM13  TCAATTGTTAG
ANB1   CTCATTTTTC
ANB1   TCCATTTTTC
ANB1   CCTATTTTTC
ANB1   TCCATTTTTC
ROX1   CCAATTTTTC

YCAATTGTTTC

A  002700000010
C  464100000505
G  000001800112
T  422087088261
Algorithms

- Enumeration
- Probabilistic Optimization
- Deterministic Optimization

1. Identify motifs
2. Build a consensus
Enumeration

- Exhaustive search: word counting method, count all $n$-mers and look for overrepresentation
- Less likely to get stuck in a local optimum
- Computationally expensive
  - YMF
  - Weeder
    - http://159.149.109.9/weederaddons/locator.html
Probabilistic Optimization

- Uses a Gibbs sampling approach
- One $n$-mer from each sequence is randomly picked to determine initial model. In subsequent iterations, one sequence, $i$, is removed and the model is recalculated. Pick a new location of motif in sequence $i$ iterate until convergence
- Assumes most sequences will have the motif
  - AlignAce
    http://atlas.med.harvard.edu/cgi-bin/alignace.pl
  - Gibbs Motif Sampler
    http://bayesweb.wadsworth.org/gibbs/gibbs.html
Deterministic Optimization

- Based on expectation maximization (EM)
- EM: iteratively estimates the likelihood given the data that is present
  
  I. **Expectation step:** Use current parameters (and observations) to reconstruct hidden structure
  
  II. **Maximization step:** Use that hidden structure (and observations) to re-estimate parameters

- MEME
  
  [http://meme.sdsc.edu](http://meme.sdsc.edu)
Multiple EM for Motif Elicitation
MEME

- Starting from a single site, EM alternates between assigning sites and updating motif model.
- Performs a single iteration for each \(n\)-mer in target sequences, selects the best motif from this site and then iterates only that one to convergence.
- Search space increases significantly with increasing number of sequence and/or sequence lengths.
# Programs Available

<table>
<thead>
<tr>
<th>Multi-purpose packages</th>
<th>Motif Scanning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TAMO</strong></td>
<td><strong>Ahab</strong></td>
</tr>
<tr>
<td>TAMO integrates several motif discovery programs. It includes support for motif scanning, scoring, evaluation of statistical significance, clustering, comparison, input/output, conversion between different motif representations, and visualization. <a href="http://tamo.rnd.mpi-magdeburg.mpg.de">http://tamo.rnd.mpi-magdeburg.mpg.de</a></td>
<td>The Ahab webservice allows users to scan for motifs in a set of sequences. Motifs may be user-specified or selected from a database of pre-defined matrices. <a href="http://gaspard.bio.nyu.edu/Ahab.html">http://gaspard.bio.nyu.edu/Ahab.html</a></td>
</tr>
<tr>
<td><strong>BEST</strong></td>
<td><strong>Clover</strong></td>
</tr>
<tr>
<td>BEST is a suite of four motif discovery tools integrated in a graphical user interface. BEST incorporates the BioOptimizer tool used to rank and improve the predictive power of the discovered motifs. <a href="http://webaster.cs.ucla.edu/~che/BEST/">http://webaster.cs.ucla.edu/~che/BEST/</a></td>
<td>Clover identifies overrepresented motifs in a set of sequences, based on a pre-compiled library of motif matrixes. <a href="http://clab.bu.edu/clover/">http://clab.bu.edu/clover/</a></td>
</tr>
<tr>
<td><strong>TOUCAN2</strong></td>
<td><strong>MAST</strong></td>
</tr>
<tr>
<td><strong>Expander</strong></td>
<td><strong>Monkey</strong></td>
</tr>
<tr>
<td>Expander is a tool for analyzing expression data. It can cluster genes, identify over-represented functional categories in clusters, and scan corresponding promoter regions for motifs. <a href="http://www.cs.tau.ac.il/~nh/Papers/">http://www.cs.tau.ac.il/~nh/Papers/</a></td>
<td>Monkey analyzes multiple sequence alignments to identify evolutionarily conserved matches to a motif. <a href="http://rana.ibr.gov/~aien/Monkey.htm">http://rana.ibr.gov/~aien/Monkey.htm</a></td>
</tr>
<tr>
<td><strong>MDScan</strong></td>
<td><strong>cisRED</strong></td>
</tr>
<tr>
<td><strong>BioProspector</strong></td>
<td><strong>OReGAno</strong></td>
</tr>
<tr>
<td>MDScan uses ChIP-chip enrichment ratio data to help the motif search. BioProspector is a Gibbs's sampling program. CompareProspector incorporates comparative genomics, biasing the search to regions of high conservation. <a href="http://www.seqmotifs.stanford.edu">http://www.seqmotifs.stanford.edu</a></td>
<td>cisRED is a database of conserved motifs and motif patterns obtained by genome scale motif discovery. OReGAno is a database of regulatory sites curated from the scientific literature. <a href="http://www.cisred.org/">http://www.cisred.org/</a> <a href="http://www.oreganno.org/">http://www.oreganno.org/</a></td>
</tr>
<tr>
<td><strong>Consensus</strong></td>
<td><strong>UCSC Genome Browser</strong></td>
</tr>
<tr>
<td><strong>PhyloCon</strong></td>
<td>UCSC Genome Browser is an online repository of genomic sequence, multiple sequence alignments, and annotation data. The browser includes tools for identifying conserved transcription factor binding sites. <a href="http://genome.ucsc.edu">http://genome.ucsc.edu</a></td>
</tr>
<tr>
<td><strong>Woodel</strong></td>
<td><strong>ENSEMBL</strong></td>
</tr>
<tr>
<td>An enumerative motif discovery program that performed well in a recent comparative analysis of fourteen algorithms. <a href="http://www.pesolelab.it/">http://www.pesolelab.it/</a></td>
<td>ENSEMBL is another online genomic sequence repository. Includes online tools for data mining as well as BLAST searches. <a href="http://www.ensembl.org/index.html">http://www.ensembl.org/index.html</a></td>
</tr>
<tr>
<td><strong>MEME</strong></td>
<td><strong>TRANSFAC</strong></td>
</tr>
<tr>
<td><strong>AlignACE</strong></td>
<td><strong>JASPAR</strong></td>
</tr>
<tr>
<td>A Gibbs sampling algorithm that can identify multiple motifs in a sequence set using an iterative masking procedure. <a href="http://sites.med.harvard.edu/">http://sites.med.harvard.edu/</a></td>
<td>Curated public database of transcription factor binding specificities represented as PWMs. <a href="http://jaspar.cgb.ki.se/">http://jaspar.cgb.ki.se/</a></td>
</tr>
</tbody>
</table>

*incomplete list

Programs Available: EMBOSS
Motif Searching
http://iona.wi.mit.edu/bio/tools/emboss/

- wordcount: Counts words of a specified size in a DNA sequence
- prophecy: Creates matrices/profiles from multiple alignments
- profit: Scan a sequence or database with a matrix or profile
Programs Available: EMBOSS
Pattern Searching

http://iona.wi.mit.edu/bio/tools/emboss/

- fuzznuc: Nucleic acid pattern search
- fuzzpro: Protein pattern search
Programs Available: Other

- Allegro (Expression)
  http://acgt.cs.tau.ac.il/allegro/

- CisGenome (ChIP-Seq)
  http://www.biostat.jhsph.edu/~hji/cisgenome
**Workflow and Strategies**

1. **Assemble input data.** Results may be improved by restricting the input to high-confidence sequences. Some algorithms achieve improved performance by using phylogenetic conservation information from orthologous sequences or information about protein DNA-binding domains.

2. **Choose several motif discovery programs for the analysis.** For recommended programs see Figure 3.

3. **Test the statistical significance of the resulting motifs.** Use control calculations to estimate the empirical distribution of scores produced by each program on random data.

4. **Clustering and post-processing the motifs.** Motif discovery analyses often produce many similar motifs, which may be combined using clustering. Phylogenetic conservation information may be used to filter out statistically significant, but non-conserved motifs that are more likely to correspond to spurious sequence patterns.

5. **Interpretation of motifs.** Algorithms exist for linking motifs to transcription factors and for combining motif discovery with expression data.

---

Further Reading

• Practical Strategies for Discovering Regulatory DNA Sequence Motifs
  http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.0020036

• How does DNA sequence motif discovery work?
  http://www.nature.com/nbt/journal/v24/n8/full/nbt0806-959.html

• MEME
MEME/MAST Demo

http://meme.sdsc.edu
MEME/MAST Demo
MEME Suite

Unaligned sequences → MEME Glam2 → Motifs

Motif database → Tomtom → Aligned motifs

Sequence database → FIMO MAST Beadstring MIAO MCAST Glam2Scan → Annotated sequences
MEME/MAST Demo
Command Line (on tak)

Usage: MEME (Find ungapped motifs in unaligned sequences)

eg. meme sample.fa -dna -maxw 10 -nmotifs 5 -mod zoops -pal -maxsize 1000000 -o sample_meme
meme <dataset> [optional arguments]
• <dataset> file containing sequences in FASTA format
• [-text] output in text format (default is HTML)
• [-dna] sequences use DNA alphabet
• [-protein] sequences use protein alphabet
• [-mod oops|zoops|anr] distribution of motifs
• [-nmotifs <nmotifs>] maximum number of motifs to find
• [-evt <ev>] stop if motif E-value greater than <evt>
• [-minw <minw>] minimum motif width
• [-maxw <maxw>] maximum motif width

For complete list of options enter “meme” at the command prompt
MEME/MAST Demo

Command Line (on tak)

Usage: MAST (Searches a sequence database for occurrences of known motifs )

eg. mast motifs.txt -d data.fa
mast <mfile> <database> [ optional arguments ... ]
• <mfile> file containing motifs to use; may be a MEME output file or similar file
• [-d <database> | -stdin] search sequences in <database> with motifs

For complete list of options enter “mast” at the command prompt