Motif Discovery Algorithms

Xiaohui S. Xie
University of California, Irvine

Course Information Update

• Grading
  – 20% Homework
  – 10% Lecture scribe notes
  – 20% Midterm exam
  – 50% Final project

• Course Prerequisites:
  – Programming skill (Perl/Python, Matlab/R)
  – Statistics and Calculus
Today’s Goals

• Gene regulation
• Motif discovery algorithms
  – Enumeration
  – Statistical significance
  – Expectation-Maximization
  – Gibbs Sampling

The Central Dogma

DNA → Transcription → RNA → Translation → Protein
Transcriptional Regulation

DNA binding proteins

Non-coding region

Activator

Repressor

Coding region (transcribed)

RNA transcript

Regulation in Eukaryotes

• Promotor
• Transcription Factors - TF
• Transcription Factor binding Sites - TFBS
• Cis-regulatory modules - CRM
• Transcription Start Site - TSS
• TATA boxes
• CG richness
• Phylogenetic Footprinting
• Combinatorial Interaction
• Enhancers

Motifs are fundamental units of gene regulation:
- What turns genes on (producing a protein) and off?
- When is a gene turned on or off?
- Where (in which cells) is a gene turned on?
- How many copies of the gene product are produced?

Specialized proteins (transcription factors) recognize these motifs.

What we know about regulatory motifs:
- Motifs are short (6-20 bp), sometimes degenerate
- Can contain any set of nucleotides (no ATG or other rules)
- Act at variable distances upstream (or downstream) from target gene (could be 100 Kb upstream or downstream)
- Human genome contains roughly 2000 motifs

Transcription Factor Binding Sites
- Gene regulatory proteins contain structural elements that can “read” DNA sequence “motifs”
- The amino acid – DNA recognition is not straightforward
- Experiments can pinpoint binding sites on DNA
Regulatory motif discovery

Promoter sequences for 15 genes

Method 1: Enumeration

List all potential motifs with a given length

For instance, 6-mer motifs

```
AAAAAA  0
AAAAAC  1
AAAAAG  2
AAAAAT  1
...
CACGTG  15
...
TTTTTT  1
TTTTTT  0
```

Total: $4^6 = 4096$ 6-mers
Regulatory motif discovery

CAAACTCCTGCACCCTGCTCTCAAGGAATTTCCCGCTCCTGCTTCTGAGTT
GGCTACAGATGTGTACACACACCAGCAACCAGCTGATTTCCCCCCTTTT
TTATCACCTGAGGCAAACGATTAGGAGAATTTCCCGCCTCTGTCTTCTGAGTT
GGCTACAGATGTGTACCACGGACCAACGAATTTCCCGCTCCTGCTTCTGAGTT
AGAAATGATGTTCCTAACCCTAAAGAGATGAGACCAACGCTGATTACAG
ACTCTTAAAGAAGACCCAGCTGATTTCCCACCTTT
TTATCACCTGAGGCAAACGATTAGGAGAATTTCCCGCCTCTGTCTTCTGAGTT
GGCTACAGATGTGTACCACGGACCAACGAATTTCCCGCTCCTGCTTCTGAGTT
AGAAATGATGTTCCTAACCCTAAAGAGATGAGACCAACGCTGATTACAG
ACTCTTAAAGAAGACCCAGCTGATTTCCCACCTTT
TTATCACCTGAGGCAAACGATTAGGAGAATTTCCCGCCTCTGTCTTCTGAGTT
GGCTACAGATGTGTACCACGGACCAACGAATTTCCCGCTCCTGCTTCTGAGTT
AGAAATGATGTTCCTAACCCTAAAGAGATGAGACCAACGCTGATTACAG
ACTCTTAAAGAAGACCCAGCTGATTTCCCACCTTT
TTATCACCTGAGGCAAACGATTAGGAGAATTTCCCGCCTCTGTCTTCTGAGTT
GGCTACAGATGTGTACCACGGACCAACGAATTTCCCGCTCCTGCTTCTGAGTT
AGAAATGATGTTCCTAACCCTAAAGAGATGAGACCAACGCTGATTACAG
ACTCTTAAAGAAGACCCAGCTGATTTCCCACCTTT
TTATCACCTGAGGCAAACGATTAGGAGAATTTCCCGCCTCTGTCTTCTGAGTT
GGCTACAGATGTGTACCACGGACCAACGAATTTCCCGCTCCTGCTTCTGAGTT
AGAAATGATGTTCCTAACCCTAAAGAGATGAGACCAACGCTGATTACAG
ACTCTTAAAGAAGACCCAGCTGATTTCCCACCTTT
TTATCACCTGAGGCAAACGATTAGGAGAATTTCCCGCCTCTGTCTTCTGAGTT
GGCTACAGATGTGTACCACGGACCAACGAATTTCCCGCTCCTGCTTCTGAGTT
AGAAATGATGTTCCTAACCCTAAAGAGATGAGACCAACGCTGATTACAG
ACTCTTAAAGAAGACCCAGCTGATTTCCCACCTTT
TTATCACCTGAGGCAAACGATTAGGAGAATTTCCCGCCTCTGTCTTCTGAGTT
GGCTACAGATGTGTACCACGGACCAACGAATTTCCCGCTCCTGCTTCTGAGTT
AGAAATGATGTTCCTAACCCTAAAGAGATGAGACCAACGCTGATTACAG
ACTCTTAAAGAAGACCCAGCTGATTTCCCACCTTT

Promoter sequences for 15 genes

How to measure significance?

Suppose we observe that among the $n$ promoter sequences, the motif occurs in $k$ of them.

How surprise is the observation?

1. Curate a set of control sequences (total number: $N$) that the motif is not enriched
2. Count the number of sequences that contain the motif ($K$)
Representation of motifs, PWM

| Site 1 | A G A T G G A T G G |
| Site 2 | T G A T T G A T G T |
| Site 3 | T G A T G G A T G G |
| Site 4 | A G A T T G A T C G |
| Site 5 | T G A T G G A T T G |
| Site 6 | T G A T G G A T T G |
| Site 7 | A G A T G G A T T G |

PWM represents frequencies of each base at each position in the motif *

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0</td>
<td>0.4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0.4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0.6</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
</tbody>
</table>

* These days, PWM/PSSM can correspond to the frequency matrix or a likelihood matrix.

Positional weight matrix representation

<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>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GTATCACCCGCGAGTGGTAT</td>
<td>ATACCCAGTGCGCTGATAC</td>
<td>TCAACACCGCCAGAGATAA</td>
<td>TTATCTCTCCGGCCTGTGTA</td>
<td>TAAACATCTGCCTGATAA</td>
<td>TATCACCCGCAAGAGATAA</td>
<td>TTAACCCCTGGCGCTGTGTA</td>
<td>CTAACCCGTGCGCGCGGTA</td>
<td>TCAACACCGCCAGAGATAA</td>
<td>TTATCACCCGCGAGTGGTAT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( W_{ij} \)

Lambda

cI/cro binding sites

Sequence Logo
The least variable positions likely are important for specifying the protein-DNA interaction. Therefore high information content = low sequence variation at that position.

Information Profile:

\[
\begin{array}{cccccccccccc}
\text{Position} & \text{bits} \\
\hline
1 & 1.0 & 2.0 & 2.0 & 2.0 & 1.1 & 2.0 & 2.0 & 2.0 & 1.3 \\
\end{array}
\]

= bit score of 15.9

Weight matrix, sequence logos

Very high frequency of false positives. A model for binding of MyoD will yield \(10^6\) binding sites, while only \(10^3\) might be real.