Regulatory Motif Finding

Lectures 12 – Nov 7, 2011
CSE 527 Computational Biology, Fall 2011
Instructor: Su-In Lee
TA: Christopher Miles
Monday & Wednesday 12:00-1:20
Johnson Hall (JHN) 022

Outline

- Biology background
- Computational problem
  - Input data
  - Motif representation
- Common methods
  - Enumeration
  - Expectation-Maximization (EM) algorithm
  - Gibbs sampling methods
Cell = Factory, Proteins = Machines

DNA

- Instructions for making the machines
- Instructions for when and where to make them
Transcriptional Regulation

- Regulatory regions are comprised of “binding sites”

- “Binding sites” attract a special class of proteins, known as “transcription factors”

- A TFBS can be located anywhere within the regulatory region (promoter region)

- Bound transcription factors can also inhibit DNA transcription
  - More realistic picture?

DNA Regulation

Source: Richardson, University College London
Gene Regulation

- Transcriptional regulation is one of many regulatory mechanisms in the cell

Transcriptional Regulation of Genes

- 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?
Structural Basis of Interaction

- **Key Feature:**
  - Transcription factors are not 100% specific when binding DNA because non-essential bases could mutate

- Not one sequence, but family of sequences, with varying affinities

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Affinity</th>
</tr>
</thead>
<tbody>
<tr>
<td>GACCG</td>
<td>0.54</td>
</tr>
<tr>
<td>GAGCG</td>
<td>0.48</td>
</tr>
<tr>
<td>GACTG</td>
<td>0.32</td>
</tr>
<tr>
<td>GACCA</td>
<td>0.25</td>
</tr>
<tr>
<td>GCCCG</td>
<td>0.11</td>
</tr>
<tr>
<td>GGCCTG</td>
<td>0.08</td>
</tr>
</tbody>
</table>
What is a motif?

- A subsequence (substring) that occurs in multiple sequences with a biological importance.

- Motifs can be totally constant or have variable elements.

- DNA motifs (regulatory elements)
  - Binding sites for proteins
  - Short sequences (5-25)
  - Up to 1000 bp (or farther) from gene
  - Inexactly repeating patterns

Motif Finding

- **Basic Objective:**
  - Find regions in the genome that transcription factors bind to

- Motivations
  - Understanding which TFs regulate which genes
  - Major part of the gene regulation

- Many classes of algorithms, differ in:
  - Types of input data
  - Motif representation
Input Data

- Single sequence

AGCATCAGCAGCACTCATCAGATACGACTCATGACATAGCCATGGGCTAAGTCGGGATGCGGATCAGCA
GATCGACATGACGCTCTACGATCATGACGAGAGGAGCGGACGCTCTACGATGACGAGGGAGCGCCGACGCTCTACGATCATGACGAGAGGAGCGGACGCTCTACGATGACGAGGGAGCGCCG

- Based on over-representation of short sequences

Random Sample

atgccgggat act gat acctt atttg gccccttctt gctt ataattgtg agagta gactctgtcgggcagctcttcttatttggcctaggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccgacccctattttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatactgggcataaggtaca
tgagtatcccgtgggaactatag cgtatctccgatatttg atagcgtatcagcgcctaatag ccgaaacct accttcttctcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gagatgggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gagatgggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gagatgggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgacttcgccatgctgatggagcatagctggttacgtagggaagccctaacggacttaatggcccacttagtccacttatagtcatacatgttcttgtgaatggatttttaactgagggcatagaccgcttggcgcacccaaattcagtgtgggcgagcgcaagcggttttggcccttgttagaggcccccgtactgatggaaactttcaattatgagagagctaatctatcgcgtgcgtgttcat aacttgagttggtttcgaaaatgctctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatttcaacgtatgccgaaccgaaagggaagctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccgtactgt gatgcgtatctcagctcggcggatctaatagcacgatcgtcttgctttgatcagcgatccg
Implanting Motif

AAAAAAAGGGGGGG

where is the Implanted Motif?

atgaccgggatactgataaaaaaaagggggggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg
acccctattttttagacacattagacctggaaaaaaaattgatgctacacatttctgatgtgaaatctgagtagcggcgcggtcga
agtctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggtgctctggt
Implanting Motif $AAAAAAAGGGGGGG$
with Four Mutations - (15,4)-motif

Where is the Motif???
Why Finding (15,4) Motif is Difficult?

Challenge Problem

- Find a motif in a sample of
  - 20 “random” sequences (e.g. 600 nt long)
  - each sequence containing an implanted pattern of length 15,
  - each pattern appearing with 4 mismatches as (15,4)-motif.
Input Data

- Single sequence
  
  ... AGCATCAGCAGCATCAGCATACGACTCGAGCATAGCCATGGGCTA CAGCA GATCGATCGAA CAGCA CG...

- Sequence + other data
  - Gene expression data
  - ChIP-chip
  - Others...

Identifying Motifs

- Genes are turned on or off by regulatory proteins (TFs).
- TFs bind to upstream regulatory regions of genes to either attract or block an RNA polymerase.
- So, multiple genes that are regulated by the same TF will have the same motifs in their regulatory regions.
- How do we identify the genes that are regulated by the same TF?
Sequence + Gene Expression Data

- Say that a microarray experiment showed that when gene X is knocked out, 20 other genes are not expressed.
  - How can one gene have such drastic effects?
- Say that 5 different genes are co-expressed across many experiments in a gene expression data.
  - These genes are likely to share the same binding sites.

daf-19 Binding Sites in *C. elegans*

- Motifs and transcriptional start sites

```
GTTGTAGCTGGTAC
GTTTCCATGGAAAC
GCTACCATGGCAAC
GTTACCATAGTAAC
GTTTCCATGGTAAC
```

-150 ——> -1
- che-2
- daf-19
- osm-1
- osm-6
- F02D8.3

source: Peter Swoboda
Input Data

- Single sequence
  
  ... AGCATCAGCAGCA CAGCATCATGACGACTCAGCATAGCCATGGGCTA CAGCA GATCGATCGAA CAGCA CG...

- Sequence + other data
  - Gene expression data
  - ChIP-chip
  - Others...

- Evolutionarily related set of sequences

Motif Finding

- **Basic Objective:**
  - Find regions in the genome that bind transcription factors

- Motivations
  - Understanding which TFs regulate which genes
  - Major part of the gene regulation

- Many classes of algorithms, differ in:
  - Types of input data
  - Motif representation
Structural Basis of Interaction

- **Key Feature:**
  - Transcription factors are not 100% specific when binding DNA
  - Not one sequence, but family of sequences, with varying affinities

```
GACCG 0.54
GAGCG 0.48
GACTG 0.32
GACCA 0.25
GGCCG 0.11
GGCTG 0.08
```

Motif Representation

- Structural discussion immediately raises difficulties

- Least expressive: **GACCG**
  - Single sequence

- Most expressive:
  - $4^k$-dimensional probability distribution
  - Independently assign probability for each of the possible k-mers*

---

* k-mer refers to a specific n-tuple of nucleic acid or amino acid sequences that can be used to identify certain regions within DNA or protein.
Motif Representation

- Standard Solution:
  - Position-Specific Scoring Matrix (PSSM)
  - Assuming independence of positions, assign a probability distribution for each position

- This is a too simple representation

Position Weight Matrix (PWM)

- Assign probability to (A,G,C,T) in each position

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.2</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>1</td>
<td>0.01</td>
<td>0.99</td>
<td>0.7</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>0.01</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.25</td>
<td>0</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>4</td>
<td>0.1</td>
<td>0.9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0.5</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0.2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Oversimplicity of PSSMs

- Assumes independence between positions

- ~25% of TRANSFAC motifs have been shown to violate this assumption
  - Two Examples: ADR1 and YAP6

Oversimplicity of PSSMs

- Assumes independence between positions

- Generates potentially unseen motifs
  - We need to model dependency between positions.
    - Revisit later
Outline

- Biology background
- Computational problem
  - Input data
  - Motif representation
- Common methods
  - Enumeration
  - Expectation-Maximization (EM) algorithm
  - Gibbs sampling methods

Finding Regulatory Motifs

- Given a collection of genes that are likely to be regulated by the same TFs,
- Find the TF-binding motifs in common
Identifying Motifs: Complications

- We do not know the motif sequence
- We do not know where it is located relative to the genes start
- Motifs can differ slightly from one gene to another
- How to discern it from “random” motifs?

Common Methods

- Problem statement:
  - Given a set of n promoters of n co-regulated genes, find a motif common to the promoters
  - Both the PWM (defined in page 30) and the motif sequences are unknown.

- Enumeration (simplest method)
  - Look at the frequency of all k-mers*

- EM algorithm (MEME)
  - Iteratively hone in on the most likely motif model

- Gibbs sampling methods (AlignAce, BioProspector)

---

* k-mer refers to a specific n-tuple of nucleic acid that can be used to identify certain regions within DNA or proteins.
Generating k-mers

- Example: \( k=5 \)

Motif Finding Using EM Algorithm

- MEME (Multiple EM for Motif Elucidation)
  - http://meme.sdsc.edu/meme/intro.html

- Expectation-Maximization
  - In each iteration, it learns the PWM model and identifies examples of the matrix (sites in the input sequences)

Identify binding locations for all PWMs

Optimize recognition preferences
Motif Finding Using EM Algorithm

- MEME works by iteratively refining PWMs and identifying sites for each PWM
  - 1. Estimate motif model (PWM)
    - Start with a k-mer seed (random or specified)
    - Build a PWM by incorporating some of background frequencies
  - 2. Identify examples of the model
    - For every k-mer in the input sequences, identify its probability given the PWM model.
  - 3. Re-estimate the motif model
    - Calculate a new PWM, based on the weighted frequencies of all k-mers in the input sequences
  - 4. Iteratively refine the PWMs and identify sites until convergence.

Example: MEME

- Find a 6-mer motif in 4 sequences
  - \( S_1: \) GGCTATTGCAGATGACGAGATGAGGCCCAGACC
  - \( S_2: \) GGATGACAATTATATAAAGGACGATAAGAGATGAC
  - \( S_3: \) CTAGCTCGTAGCTCGTTGAGATGCGCTCCCCGCTC
  - \( S_4: \) GATGACGGAGTATTTAAAGACTCGATGAGTTATACGA

  - 1. MEME uses an initial EM heuristic to estimate the best starting-point PWM matrix:
    - \( G: \) 0.26 0.24 0.18 0.26 0.25 0.26
    - \( A: \) 0.24 0.26 0.28 0.24 0.25 0.22
    - \( T: \) 0.25 0.23 0.30 0.25 0.25 0.25
    - \( C: \) 0.25 0.27 0.24 0.25 0.25 0.27
2. MEME scores the match of all 6-mers to current matrix

Here, just consider the underlined 6-mers. Although in reality all 6-mers are scored

\[ \text{GCTATTG} \text{CATATGCAG} \text{GATGAG} \text{GCGCGAG} \text{CC} \]

\[ \text{GGATGACA} \text{ATTATAAGGACCGT} \text{GATAAGAGATTAC} \]

\[ \text{CTAGCTCGTAGCTCGTGAGATGCCGCTGCCCCTC} \]

\[ \text{GATGACGGAAGTATTAAGACTCGATGAGTTATACGAGTATACGA} \]

3. Re-estimate the PWM based on the weighted contribution of all 6-mers.

\[
\begin{array}{cccccccc}
\text{G} & 0.29 & 0.24 & 0.17 & 0.27 & 0.24 & 0.30 \\
\text{A} & 0.22 & 0.26 & 0.27 & 0.22 & 0.28 & 0.18 \\
\text{T} & 0.24 & 0.23 & 0.33 & 0.23 & 0.24 & 0.28 \\
\text{C} & 0.24 & 0.27 & 0.23 & 0.28 & 0.24 & 0.24 \\
\end{array}
\]

4. MEME scores the match of all 6-mers to current matrix

\[ \text{GCTATTG} \text{CATATGCAG} \text{GATGAG} \text{GCGCGAG} \text{CC} \]

\[ \text{GGATGACA} \text{ATTATAAGGACCGT} \text{GATAAGAGATTAC} \]

\[ \text{CTAGCTCGTAGCTCGTGAGATGCCGCTGCCCCTC} \]

\[ \text{GATGACGGAAGTATTAAGACTCGATGAGTTATACGAGTATACGA} \]

5. Re-estimate the PWM based on the weighted contribution of all 6-mers.

\[
\begin{array}{cccccccc}
\text{G} & 0.40 & 0.20 & 0.15 & 0.42 & 0.24 & 0.30 \\
\text{A} & 0.30 & 0.30 & 0.20 & 0.24 & 0.46 & 0.18 \\
\text{T} & 0.15 & 0.30 & 0.45 & 0.16 & 0.15 & 0.28 \\
\text{C} & 0.15 & 0.20 & 0.20 & 0.16 & 0.15 & 0.24 \\
\end{array}
\]
6. MEME scores the match of all 6-mers to current matrix

$G\text{GCTATTG\text{CATATGACGA}\text{GATGAGG\text{CCCAGA}\text{GCC}}$  

$G\text{GATGAC\text{T}\text{TTATAAAAGGACCGT}\text{GATAAGAGATTAC}}$  

$\text{CTAGCTG\text{TAGCTGTTGAGATGCGCTCCCGCTC}}$  

$G\text{ATGAGCGGAGTATTAAAGACTCGATGAGTTATACG}$

Iterations continue until convergence

- Numbers do not change much between iterations

<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>0.85</td>
<td>0.05</td>
<td>0.10</td>
<td>0.80</td>
</tr>
<tr>
<td>A</td>
<td>0.05</td>
<td>0.60</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td>T</td>
<td>0.05</td>
<td>0.30</td>
<td>0.70</td>
<td>0.05</td>
</tr>
<tr>
<td>C</td>
<td>0.05</td>
<td>0.05</td>
<td>0.10</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Final motif

Gibbs Sampling

- References
  - AlignAce by Hughes et al. 2000
    http://atlas.med.harvard.edu/download/index.html,
  - BioProspector by Liu et al. 2001
    http://motif.stanford.edu/distributions/r

- Procedure
  - 1. Start by randomly choosing sites and creates an initial PWM matrix
  - 2. Sample other sites
    - Remove some set of matrix examples (sites)
    - Randomly choose other sites and calculate $P$ given matrix
    - If they have a high score to the matrix, keep the new site
  - 3. Iterate until convergence
Gibbs Sampling: Basic Idea

Current motif = PWM formed by circled substrings

Gibbs Sampling: Basic Idea

Delete one substring

Slides generously and unknowingly provided by S. Sinha, Urbana-Champaign CS Dept.
Gibbs sampling: Basic Idea

Try a replacement:
Compute its score,
Accept the replacement depending on the score.

Outline

- Biology background
- Computational problem
  - Input data
  - Motif representation
- Common methods
  - Enumeration
  - Expectation-Maximization (EM) algorithm
  - Gibbs sampling methods