Learning Sequence Motif Models Using Expectation Maximization (EM)

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2015
Colin Dewey
cdewey@biostat.wisc.edu
Goals for Lecture

the key concepts to understand are the following
• the motif finding problem
• using EM to address the motif-finding problem
• the OOPS and ZOOPS models
Sequence Motifs

• what is a sequence *motif*?
  – a sequence pattern of biological significance

• examples
  – DNA sequences corresponding to protein binding sites
  – protein sequences corresponding to common functions or conserved pieces of structure
Sequence Motifs Example

CAP-binding motif model based on 59 binding sites in E.coli

helix-turn-helix motif model based on 100 aligned protein sequences

Figure from Crooks et al., Genome Research 14:1188-90, 2004.
The Motif Model Learning Task

given: a set of sequences that are thought to contain occurrences of an unknown motif of interest

do:
  – infer a model of the motif
  – predict the locations of the motif occurrences in the given sequences
Why is this important?

- To further our understanding of which regions of sequences are “functional”
- DNA: biochemical mechanisms by which the expression of genes are regulated
- Proteins: which regions of proteins interface with other molecules (e.g., DNA binding sites)
- Mutations in these regions may be significant
Motifs and *Profile Matrices* (a.k.a. *Position Weight Matrices*)

- given a set of aligned sequences, it is straightforward to construct a profile matrix characterizing a motif of interest

![Diagram of shared motif with sequence positions and probability matrix]

- each element represents the probability of given character at a specified position
## Sequence logos

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1</td>
<td>0.5</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
<td>0.6</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>0.1</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>0.6</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>6</td>
<td>0.4</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>7</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.7</td>
</tr>
<tr>
<td>8</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

or

```
0.1 0.5 0.2 0.2
0.3 0.2 0.3 0.2
```

or

```
5'
A T G C
G C T A
C A T G
C T A G
T G A C
G T A C
T G A C
3'
```

- frequency logo
- information content logo
Motifs and Profile Matrices

• How can we construct the profile if the sequences aren’t aligned?
• In the typical case we don’t know what the motif looks like.
The Expectation-Maximization (EM) Approach


• EM is a family of algorithms for learning probabilistic models in problems that involve *hidden state*
• in our problem, the hidden state is where the motif starts in each training sequence
Overview of EM

• Method for finding the maximum likelihood (ML) parameters (θ) for a model (M) and data (D)

\[ \theta_{ML} = \arg \max_{\theta} P(D \mid \theta, M) \]

• Useful when
  – it is difficult to optimize \( P(D \mid \theta) \) directly
  – likelihood can be decomposed by the introduction of hidden information (Z)

\[ P(D \mid \theta) = \sum_{Z} P(D, Z \mid \theta) \]

  – and it is easy to optimize the function (with respect to θ):

\[ Q(\theta \mid \theta') = \sum_{Z} P(Z \mid D, \theta') \log P(D, Z \mid \theta) \]

(see text section 11.6 for details)
Applying EM to the motif finding problem

• First define the probabilistic model and likelihood function \( P(D | \theta) \)

• Identify the hidden variables (Z)
  – In this application, they are the locations of the motifs

• Write out the Expectation (E) step
  – Compute the expected values of the hidden variables given current parameter values

• Write out the Maximization (M) step
  – Determine the parameters that maximize the Q function, given the expected values of the hidden variables
Representing Motifs in MEME

• a motif is
  – assumed to have a fixed width, $W$
  – represented by a matrix of probabilities: $p_{c,k}$ represents the probability of character $c$ in column $k$

• also represent the “background” (i.e. sequence outside the motif): $p_{c,0}$ represents the probability of character $c$ in the background
Representing Motifs in MEME

- example: a motif model of length 3

\[
p = \begin{array}{cccccc}
A & 0.25 & 0.1 & 0.5 & 0.2 \\
C & 0.25 & 0.4 & 0.2 & 0.1 \\
G & 0.25 & 0.3 & 0.1 & 0.6 \\
T & 0.25 & 0.2 & 0.2 & 0.1 \\
\end{array}
\]

\[
p = \begin{array}{cccc}
0 & 1 & 2 & 3 \\
\hline
A & 0.25 & 0.1 & 0.5 & 0.2 \\
C & 0.25 & 0.4 & 0.2 & 0.1 \\
G & 0.25 & 0.3 & 0.1 & 0.6 \\
T & 0.25 & 0.2 & 0.2 & 0.1 \\
\end{array}
\]

background motif positions
Representing Motif Starting Positions in MEME

• the element $Z_{i,j}$ of the matrix $Z$ is an indicator random variable that takes value 1 if the motif starts in position $j$ in sequence $i$ (and takes value 0 otherwise)

• example: given DNA sequences of length 6, where $W=3$

$$
\begin{array}{cccc}
G & T & C & A & G & G \\
G & A & G & A & G & T \\
A & C & G & G & A & G \\
C & C & A & G & T & C \\
\end{array}
$$

$$
\begin{array}{cccc}
1 & 2 & 3 & 4 \\
seq1 & 0 & 0 & 1 & 0 \\
seq2 & 1 & 0 & 0 & 0 \\
seq3 & 0 & 0 & 0 & 1 \\
seq4 & 0 & 1 & 0 & 0 \\
\end{array}
$$
Probability of a Sequence Given a Motif Starting Position

\[ P(X_i \mid Z_{i,j} = 1, p) = \prod_{k=1}^{j-1} p_{c_k}, 0 \prod_{k=j}^{j+W-1} p_{c_k}, k-j+1 \prod_{k=j+W}^{L} p_{c_k}, 0 \]

- \( X_i \) is the \( i \)th sequence
- \( Z_{i,j} \) is 1 if motif starts at position \( j \) in sequence \( i \)
- \( c_k \) is the character at position \( k \) in sequence \( i \)
Sequence Probability Example

\[ X_i = \begin{array}{cccc}
G & C & T & G & T & A & G \\
\end{array} \]

\[ p = \begin{array}{cccc}
\text{A} & 0.25 & 0.1 & 0.5 & 0.2 \\
\text{C} & 0.25 & 0.4 & 0.2 & 0.1 \\
\text{G} & 0.25 & 0.3 & 0.1 & 0.6 \\
\text{T} & 0.25 & 0.2 & 0.2 & 0.1 \\
\end{array} \]

\[ P(X_i | Z_{i3} = 1, p) = \]

\[ p_{G,0} \times p_{C,0} \times p_{T,1} \times p_{G,2} \times p_{T,3} \times p_{A,0} \times p_{G,0} = \]

\[ 0.25 \times 0.25 \times 0.2 \times 0.1 \times 0.1 \times 0.25 \times 0.25 \]
Likelihood function

\[ P(D \mid p) = \prod_i P(X_i \mid p) \]

\[= \prod_i \sum_j P(X_i \mid Z_{ij} = 1, p)P(Z_{ij} = 1) \]

\[= (L - W + 1)^{-n} \prod_i \sum_j P(X_i \mid Z_{ij} = 1, p) \]

- This is the function that EM will (indirectly) optimize
Basic EM Approach

given: length parameter $W$, training set of sequences

$t=0$
set initial values for $p^{(0)}$
do
++$t$
re-estimate $Z^{(t)}$ from $p^{(t-1)}$ (E-step)
re-estimate $p^{(t)}$ from $Z^{(t)}$ (M-step)
until change in $p^{(t)} < \varepsilon$ (or change in likelihood is $< \varepsilon$)
return: $p^{(t)}$, $Z^{(t)}$
Warning: Notation Abuse!

- During the E-step, we compute the expected values of $Z$ given $p^{(t-1)}$.
- We denote these expected values by $Z^{(t)} = E[Z \mid p^{(t-1)}]$.
- For example:

\[
Z^{(t)} = \begin{pmatrix}
  1 & 2 & 3 & 4 \\
seq1 & 0.1 & 0.1 & 0.2 & 0.6 \\
seq2 & 0.4 & 0.2 & 0.1 & 0.3 \\
seq3 & 0.3 & 0.1 & 0.5 & 0.1 \\
\end{pmatrix}
\]
The E-step: Computing $Z^{(t)}$

• to estimate the starting positions in $Z$ at step $t$

$$Z_{i,j}^{(t)} = \frac{P(X_i \mid Z_{i,j} = 1, p^{(t-1)})P(Z_{i,j} = 1)}{\sum_{k=1}^{L-W+1} P(X_i \mid Z_{i,k} = 1, p^{(t-1)})P(Z_{i,k} = 1)}$$

• this comes from Bayes’ rule applied to

$$P(Z_{i,j} = 1 \mid X_i, p^{(t-1)})$$
The E-step: Computing $Z^{(t)}$

- assume that it is equally likely that the motif will start in any position

$$Z^{(t)}_{i,j} = \frac{P(X_i | Z_{i,j} = 1, p^{(t-1)})P(Z_{i,j} = 1)}{\sum_{k=1}^{L-W+1} P(X_i | Z_{i,k} = 1, p^{(t-1)})P(Z_{i,k} = 1)}$$
Example: Computing $Z^{(t)}$

$$X_i = \begin{array}{cccc}
G & C & T & G & T & A & G \\
A & 0.25 & 0.1 & 0.5 & 0.2 \\
C & 0.25 & 0.4 & 0.2 & 0.1 \\
G & 0.25 & 0.3 & 0.1 & 0.6 \\
T & 0.25 & 0.2 & 0.2 & 0.1
\end{array}$$

$$p^{(t-1)} = \begin{array}{cccc}
0 & 1 & 2 & 3 \\
A & 0.25 & 0.1 & 0.5 & 0.2 \\
C & 0.25 & 0.4 & 0.2 & 0.1 \\
G & 0.25 & 0.3 & 0.1 & 0.6 \\
T & 0.25 & 0.2 & 0.2 & 0.1
\end{array}$$

$$Z^{(t)}_{i,1} \propto P(X_i | Z_{i,1} = 1, p^{(t-1)}) = 0.3 \times 0.2 \times 0.1 \times 0.25 \times 0.25 \times 0.25 \times 0.25$$

$$Z^{(t)}_{i,2} \propto P(X_i | Z_{i,2} = 1, p^{(t-1)}) = 0.25 \times 0.4 \times 0.2 \times 0.6 \times 0.25 \times 0.25 \times 0.25$$

\[ \vdots \]

- then normalize so that

\[ \sum_{j=1}^{L-W+1} Z^{(t)}_{i,j} = 1 \]
The M-step: Estimating $p$

- recall $p_{c,k}$ represents the probability of character $c$ in position $k$; values for $k=0$ represent the background

$$p_{c,k}^{(t)} = \frac{n_{c,k} + d_{c,k}}{\sum_b (n_{b,k} + d_{b,k})}$$

$$n_{c,k} = \begin{cases} \sum_i \sum_{\{j|X_{i,j+k-1}=c\}} Z_{i,j}^{(t)} & k > 0 \\ n_c - \sum_{j=1}^W n_{c,j} & k = 0 \end{cases}$$

pseudo-counts

sum over positions where $c$ appears

total # of $c$'s in data set
Example: Estimating $p$

\[
\begin{align*}
&A C A G C A \\
Z_{1,1}^{(t)} &= 0.1, \quad Z_{1,2}^{(t)} = 0.7, \quad Z_{1,3}^{(t)} = 0.1, \quad Z_{1,4}^{(t)} = 0.1 \\
&A G G C A G \\
Z_{2,1}^{(t)} &= 0.4, \quad Z_{2,2}^{(t)} = 0.1, \quad Z_{2,3}^{(t)} = 0.1, \quad Z_{2,4}^{(t)} = 0.4 \\
&T C A G T C \\
Z_{3,1}^{(t)} &= 0.2, \quad Z_{3,2}^{(t)} = 0.6, \quad Z_{3,3}^{(t)} = 0.1, \quad Z_{3,4}^{(t)} = 0.1 \\
\end{align*}
\]

\[
\begin{align*}
p_{A,1}^{(t)} &= \frac{Z_{1,1}^{(t)} + Z_{1,3}^{(t)} + Z_{2,1}^{(t)} + Z_{3,3}^{(t)} + 1}{Z_{1,1}^{(t)} + Z_{1,2}^{(t)} + \ldots + Z_{3,3}^{(t)} + Z_{3,4}^{(t)} + 4} \\
p_{C,2}^{(t)} &= \frac{Z_{1,1}^{(t)} + Z_{1,4}^{(t)} + Z_{2,3}^{(t)} + Z_{3,1}^{(t)} + 1}{Z_{1,1}^{(t)} + Z_{1,2}^{(t)} + \ldots + Z_{3,3}^{(t)} + Z_{3,4}^{(t)} + 4} \\
\end{align*}
\]
The ZOOPS Model

• the approach as we’ve outlined it, assumes that each sequence has exactly one motif occurrence per sequence; this is the OOPS model

• the ZOOPS model assumes zero or one occurrences per sequence
E-step in the ZOOPS Model

• we need to consider another alternative: the $i$th sequence doesn’t contain the motif
• we add another parameter (and its relative)

$$\lambda = \frac{\gamma}{(L - W + 1)}$$

- prior probability of a sequence containing a motif
- prior probability that any position in a sequence is the start of a motif
E-step in the ZOOPS Model

\[ Z_{i,j}^{(t)} = \frac{P(X_i | Z_{i,j} = 1, p^{(t-1)}) \lambda^{(t-1)}}{P(X_i | Q_i = 0, p^{(t-1)}) (1 - \gamma^{(t-1)}) + \sum_{k=1}^{L-W+1} P(X_i | Z_{i,k} = 1, p^{(t-1)}) \lambda^{(t-1)}} \]

- \( Q_i \) is a random variable for which \( Q_i = 1 \) if sequence \( X_i \) contains a motif, \( Q_i = 0 \) otherwise

\[ Q_i = \sum_{j=1}^{L-W+1} Z_{i,j} \]

\[ P(X_i | Q_i = 0, p^{(t-1)}) = \prod_{j=1}^{L} p_{e,j,0}^{(t-1)} \]
M-step in the ZOOPS Model

- update $p$ same as before
- update $\gamma$ as follows:

$$\gamma^{(t)} \equiv (L - W + 1) \lambda^{(t)} = \frac{1}{n} \sum_{i=1}^{n} Q_{i}^{(t)}$$
Extensions to the Basic EM Approach in MEME

- varying the approach (TCM model) to assume zero or more motif occurrences per sequence
- choosing the width of the motif
- finding multiple motifs in a group of sequences
  - choosing good starting points for the parameters
  - using background knowledge to bias the parameters
Starting Points in MEME

• EM is susceptible to local maxima, so it’s a good idea to try multiple starting points
• insight: motif must be similar to some subsequence in data set

• for every distinct subsequence of length $W$ in the training set
  – derive an initial $p$ matrix from this subsequence
  – run EM for 1 iteration
• choose motif model (i.e. $p$ matrix) with highest likelihood
• run EM to convergence
Using Subsequences as Starting Points for EM

- set values matching letters in the subsequence to some value $\pi$
- set other values to $(1- \pi)/(M-1)$ where $M$ is the length of the alphabet
- example: for the subsequence TAT with $\pi = 0.5$

\[
p = \begin{pmatrix}
1 & 2 & 3 \\
A & 0.17 & 0.5 & 0.17 \\
C & 0.17 & 0.17 & 0.17 \\
G & 0.17 & 0.17 & 0.17 \\
T & 0.5 & 0.17 & 0.5 \\
\end{pmatrix}
\]
# MEME web server

Use this form to submit DNA or protein sequences to MEME. MEME will analyze your sequences for similarities among them and produce a description (motif) for each pattern it discovers.

## Data Submission Form

**Required**

- **Your e-mail address:**
  - [Input Field]
  - [Re-enter e-mail address]

- **Please enter the sequences which you believe share one or more motifs. The sequences may contain no more than 60000 characters, total, in any of a large number of formats.**
  - [Choose File]
  - [No file selected]
  - [Clear]

- **Enter the name of a file containing the sequences here:**
  - [Input Field]

- **How do you think the occurrences of a single motif are distributed among the sequences?**
  - [One per sequence]
  - [Zero or one per sequence]
  - [Any number of repetitions]

- **MEME will find the optimum width of each motif within the limits you specify here:**
  - [Minimum width ($\leq$ 2)]
  - [Maximum width ($\leq$ 300)]
  - [Maximum number of motifs to find]

## Options

- **Description of your sequences:**
  - [Input Field]

- **MEME will find the optimum number of sites for each motif within the limits you specify here:**
  - [Minimum sites ($\geq$ 2)]
  - [Maximum sites ($\leq$ 600)]

- **Shuffle sequence letters**

- **Perform discriminative motif discovery – Enter the name of a file containing “negative sequences”:**
  - [Choose file]
  - [No file selected]
  - [Clear]

- **Enter the name of a file containing a background Markov model:**
  - [Choose file]
  - [No file selected]
  - [Clear]

- **DNA-ONLY OPTIONS**
  - [Ignored for protein searches]
  - [Search given strand only]
  - [Look for palindromes only]

- **Start search**

## Version 4.9.0

*Powered by Opal*

http://meme.nbcr.net