Learning Sequence Motif Models Using Expectation Maximization (EM)

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2020
Daifeng Wang
daifeng.wang@wisc.edu

These slides, excluding third-party material, are licensed under CC BY-NC 4.0 by Mark Craven, Colin Dewey, Anthony Gitter and Daifeng Wang
Goals for Lecture

Key concepts

• the motif finding problem
• using EM to address the motif-finding problem
• the OOPS and ZOOPS models
Your genome is your genetic code book

<table>
<thead>
<tr>
<th>Book</th>
<th>Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapters</td>
<td>Chromosomes</td>
</tr>
<tr>
<td>Sentences</td>
<td>Genes</td>
</tr>
<tr>
<td>Words</td>
<td>Elements</td>
</tr>
<tr>
<td>Letters</td>
<td>Bases</td>
</tr>
</tbody>
</table>

Human
- 46 chromosomes
- ~20,000 – 25,000 genes
- ~Millions elements
- 4 unique bases (A, T, C, G), ~3 billion in total

https://goo.gl/images/vMaz4T
How to read sentences/genes for understanding book/genome?

<table>
<thead>
<tr>
<th>Book</th>
<th>Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapters</td>
<td>Chromosomes</td>
</tr>
<tr>
<td>Sentences</td>
<td>Genes</td>
</tr>
<tr>
<td>Words</td>
<td>Elements</td>
</tr>
<tr>
<td>Letters</td>
<td>Bases</td>
</tr>
</tbody>
</table>

“On most days, I enter the Capitol through the basement. A small subway train carries me from the Hart Building, where …”

- Key words
- Non-key words

Overhead, the ceiling forms a creamy white oval, with an American eagle etched in its center. Above the visitors’ gallery, the busts of the nation’s first twenty vice presidents sit in solemn repose.

And in gentle steps, one hundred mahogany desks rise from the well of the Senate in four horseshoe-shaped rows. Some of these desks date back to 1819, and atop each desk is a tidy receptacle for inkwells and quills. Open the drawer of any desk, and you will find within the names of the senators who once used it—Taft and Long, Stennis and Kennedy—scratched or penned in the senator’s own hand. Sometimes, standing there in

- Coding elements (Exon, 2%)
  - Become proteins carrying out functions
- Non-coding elements (98%)

https://goo.gl/images/vMaz4T
Grammar for book is clear but not for genome

<table>
<thead>
<tr>
<th>Book</th>
<th>Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapters</td>
<td>Chromosomes</td>
</tr>
<tr>
<td>Sentences</td>
<td>Genes</td>
</tr>
<tr>
<td>Words</td>
<td>Elements</td>
</tr>
<tr>
<td>Letters</td>
<td>Bases</td>
</tr>
</tbody>
</table>

- Key words
- Non-key words

- Set up “rules” in translating genomic codes to functions
- Broken rules -> Abnormal functions
- Unclear

- Coding elements
  - Non-coding elements
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

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.

• Each element represents the probability of given character at a specified position.
## Sequence Logos

<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>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>C</td>
<td>0.5</td>
<td>0.2</td>
<td>0.1</td>
<td>0.6</td>
<td>0.1</td>
<td>0.2</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0.2</td>
<td>0.2</td>
<td>0.6</td>
<td>0.5</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
<td>0.3</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

- Information content (IC) at position $n = \log_2(4) - \text{Entropy(position } n)$
- Entropy(position $n) = -P(n=A)\log_2 P(n=A) - P(n=T)\log_2 P(n=T) - P(n=C)\log_2 P(n=C) - P(n=G)\log_2 P(n=G)$

**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.
ChIP-chip experiment tells which probes are bound (though this protocol has been replaced by ChIP-seq)
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
Clustering
K-means

1) Pick two (i.e. k?) random points as putative cluster centers.
2) Group the points to be clustered by the center to which they are closest.
3) Then take the mean of each group and repeat, with the means now at the cluster center.
4) Stop when the centers stop moving.
Overview of EM

- Method for finding the maximum likelihood (ML) parameters ($\theta$) for a model (M) and data (D)

$$\theta_{ML} = \arg\max_{\theta} P(D | \theta, M)$$

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

$$P(D | \theta) = \sum_{Z} P(D, Z | \theta)$$

  - and it is easy to optimize the function (with respect to $\theta$):

$$Q(\theta | \theta^t) = \sum_{Z} P(Z | D, \theta^t) \log P(D, Z | \theta)$$

(see optional reading and text section 11.6 for details)
Convergence of the EM algorithm

Convergence of the EM algorithm

https://www.nature.com/articles/nbt1406
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

- MEME: **Multiple EM for Motif Elicitation**
- 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

- Data $D$ is a collection of sequences, denoted $X$
Representing Motifs in MEME

- Example: a motif model of length 3

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

- 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 where $L=6$ and $W=3$.
- Possible starting positions $m = L - W + 1$.

$$Z = \begin{bmatrix}
1 & 2 & 3 & 4 \\
0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0
\end{bmatrix}$$
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

$X_i = \text{G C T G T A G}$

$p = \begin{bmatrix}
A & 0.25 & 0.1 & 0.5 & 0.2 \\
C & 0.25 & 0.4 & 0.2 & 0.1 \\
G & 0.25 & 0.3 & 0.1 & 0.6 \\
T & 0.25 & 0.2 & 0.2 & 0.1 \\
\end{bmatrix}$

$P(X_i \mid Z_{i,3} = 1, p) = p_{G,0} \times p_{C,0} \times p_{T,1} \times p_{G,2} \times p_{T,3} \times p_{A,0} \times p_{G,0} =
0.25 \times 0.25 \times 0.2 \times 0.1 \times 0.1 \times 0.25 \times 0.25$
Likelihood Function

- EM (indirectly) optimizes log likelihood of observed data

\[ \log P(X \mid p) \]

- M step requires joint log likelihood

\[
\log P(X, Z \mid p) = \log \prod_i P(X_i, Z_i \mid p) \\
= \log \prod_i P(X_i \mid Z_i, p)P(Z_i \mid p) \\
= \log \prod_i \frac{1}{m} \prod_j P(X_i \mid Z_{i,j} = 1, p)^{Z_{i,j}} \\
= \sum_i \sum_j Z_{i,j} \log P(X_i \mid Z_{i,j} = 1, p) + n \log \frac{1}{m}
\]

See Section IV.C of Bailey’s dissertation for details
Basic EM Approach

given: length parameter $W$, training set of sequences $t=0$

set initial values for $p^{(0)}$

do

    $t = t + 1$

    - 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: $p^{(t)}$, $Z^{(t)}$
Expected Starting Positions

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

$$Z^{(t)} = \begin{pmatrix} 0.1 & 0.1 & 0.2 & 0.6 \\ 0.4 & 0.2 & 0.1 & 0.3 \\ 0.3 & 0.1 & 0.5 & 0.1 \end{pmatrix}$$
The E-step: Computing $Z^{(t)}$

- To estimate the starting positions in $Z$ at step $t$

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

- 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

$$P(Z_{i,j} = 1) = \frac{1}{m}$$

$$Z_{i,j}^{(t)} = \frac{\sum_{k=1}^{m} P(X_i \mid Z_{i,k} = 1, p^{(t-1)}) P(Z_{i,k} = 1)}{\sum_{k=1}^{m} P(X_i \mid Z_{i,k} = 1, p^{(t-1)}) P(Z_{i,k} = 1)}$$
Example: Computing $Z^{(t)}$

$$X_i = \text{G C T G T A G}$$

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

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

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

\[ \vdots \]

- Then normalize so that $\sum_{j=1}^{m} Z^{(t)}_{i,j} = 1$
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.

$$p^{(t)}_{c,k} = \frac{n_{c,k} + d_{c,k}}{\sum_{b \in \{A,C,G,T\}} (n_{b,k} + d_{b,k})}$$

- $n_{c,k}$ is the number of $c$'s at position $k$.
- $n_{c}$ is the total number of $c$'s in the data set.
- $Z_{i,j}^{(t)}$ is the sum over positions where $c$ appears.

For $k > 0$: $Z_{i,j}^{(t)} = \sum_{i} \sum_{\{j | X_{i,j+k-1} = c\}}$.

For $k = 0$: $n_{c} - \sum_{j=1}^{W} n_{c,j}$.

$$p^{(t)}_{c,k} = \frac{n_{c,k} + d_{c,k}}{\sum_{b \in \{A,C,G,T\}} (n_{b,k} + d_{b,k})}$$

pseudo-counts
Example: Estimating $p$

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

\begin{align*}
p^{(t)}_{\text{A},1} &= \frac{Z^{(t)}_{1,1} + Z^{(t)}_{1,3} + Z^{(t)}_{2,1} + Z^{(t)}_{3,3} + 1}{Z^{(t)}_{1,1} + Z^{(t)}_{1,2} \ldots + Z^{(t)}_{3,3} + Z^{(t)}_{3,4} + 4} \\
p^{(t)}_{\text{C},2} &= \frac{Z^{(t)}_{1,1} + Z^{(t)}_{1,4} + Z^{(t)}_{2,3} + Z^{(t)}_{3,1} + 1}{Z^{(t)}_{1,1} + Z^{(t)}_{1,2} \ldots + Z^{(t)}_{3,3} + Z^{(t)}_{3,4} + 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

---

[Diagram showing motifs in sequences with varying occurrences]
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)

\[ \lambda = \frac{\gamma}{(L - W + 1)} = \frac{\gamma}{m} \]

- Prior probability of a sequence containing a motif
- Prior probability that any position in a sequence is the start of a motif

- Possible starting positions $m = L - W + 1$
E-step in the ZOOPS Model

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

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

\[ Q_i = \sum_{j=1}^{m} Z_{i,j} \]

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

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

$$\gamma^{(t)} \equiv m\lambda^{(t)} = \frac{1}{n} \sum_{i=1}^{n} Q_i^{(t)}$$
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 $\pi = 0.7$

$$p = \begin{pmatrix} 1 & 2 & 3 \\ A & 0.1 & 0.7 & 0.1 \\ C & 0.1 & 0.1 & 0.1 \\ G & 0.1 & 0.1 & 0.1 \\ T & 0.7 & 0.1 & 0.7 \end{pmatrix}$$
MEME web server

http://meme-suite.org/