

# Learning Sequence Motifs Using Expectation Maximization (EM) and Gibbs Sampling

BMI/CS 776

[www.biostat.wisc.edu/~craven/776.html](http://www.biostat.wisc.edu/~craven/776.html)

Mark Craven

[craven@biostat.wisc.edu](mailto:craven@biostat.wisc.edu)

March 2002

## Announcements

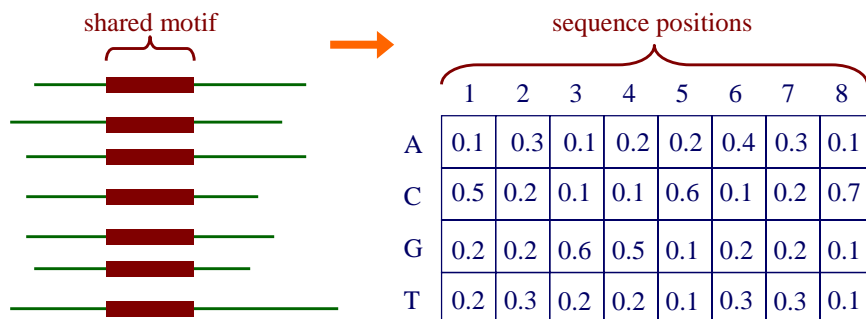
- HW due on Monday
- reading for next week
  - Chapter 3 of Durbin et al. (already assigned)
  - Krogh, An Introduction to HMMs for Biological Sequences
- is everyone's account set up?
- did everyone get the Lawrence et al. paper?

## Sequence Motifs

- what is a sequence *motif*?
  - a sequence pattern of biological significance
- examples
  - protein binding sites in DNA
  - protein sequences corresponding to common functions or conserved pieces of structure

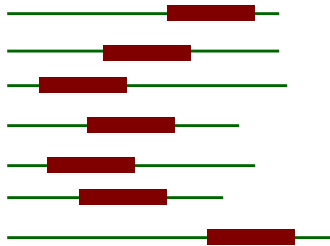
## Motifs and Profile Matrices

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



## Motifs and Profile Matrices

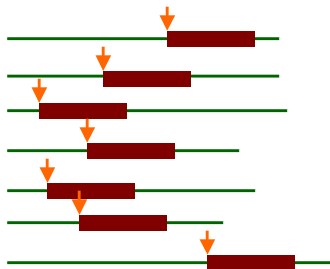
- how can we construct the profile if the sequences aren't aligned?
  - in the typical case we don't know what the motif looks like



- use an Expectation Maximization (EM) algorithm

## The 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



## The MEME Algorithm

- Bailey & Elkan, 1993
- uses EM algorithm to find multiple motifs in a set of sequences
- first EM approach to motif discovery: Lawrence & Reilly 1990

## Representing Motifs

- a motif is assumed to have a fixed width,  $W$
- a motif is represented by a matrix of probabilities:  $p_{ck}$  represents the probability of character  $c$  in column  $k$
- example: DNA motif with  $W=3$

$$p = \begin{array}{ccccc} & & 1 & 2 & 3 \\ \mathbf{A} & 0.1 & 0.5 & 0.2 & \\ \mathbf{C} & 0.4 & 0.2 & 0.1 & \\ \mathbf{G} & 0.3 & 0.1 & 0.6 & \\ \mathbf{T} & 0.2 & 0.2 & 0.1 & \end{array}$$

## Representing Motifs

- we will also represent the “background” (i.e. outside the motif) probability of each character
- $P_{c0}$  represents the probability of character  $c$  in the background
- example:

$$p_0 = \begin{array}{ll} \text{A} & 0.26 \\ \text{C} & 0.24 \\ \text{G} & 0.23 \\ \text{T} & 0.27 \end{array}$$

## Basic EM Approach

- the element  $Z_{ij}$  of the matrix  $Z$  represents the probability that the motif starts in position  $j$  in sequence  $I$
- example: given 4 DNA sequences of length 6, where  $W=3$

$$Z = \begin{array}{lcccc} & 1 & 2 & 3 & 4 \\ \text{seq1} & 0.1 & 0.1 & 0.2 & 0.6 \\ \text{seq2} & 0.4 & 0.2 & 0.1 & 0.3 \\ \text{seq3} & 0.3 & 0.1 & 0.5 & 0.1 \\ \text{seq4} & 0.1 & 0.5 & 0.1 & 0.3 \end{array}$$

## Basic EM Approach

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

set initial values for  $p$

do

re-estimate  $Z$  from  $p$  (E-step)

re-estimate  $p$  from  $Z$  (M-step)

until change in  $p < \epsilon$

return:  $p, Z$

## Basic EM Approach

- we'll need to calculate the probability of a training sequence given a hypothesized starting position:

$$\Pr(X_i | Z_{ij} = 1, p) = \underbrace{\prod_{k=1}^{j-1} p_{c_k,0}}_{\text{before motif}} \underbrace{\prod_{k=j}^{j+W-1} p_{c_k, k-j+1}}_{\text{motif}} \underbrace{\prod_{k=j+W}^L p_{c_k,0}}_{\text{after motif}}$$

$X_i$  is the  $i$ th sequence

$Z_{ij}$  is 1 if motif starts at position  $j$  in sequence  $i$

$C_k$  is the character at position  $k$  in sequence  $i$

## Example

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

		0	1	2	3
$p =$	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

$$\Pr(X_i | Z_{i3} = 1, p) =$$

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

$$0.25 \times 0.25 \times 0.2 \times 0.1 \times 0.1 \times 0.25 \times 0.25$$

## The E-step: Estimating $Z$

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

$$Z_{ij}^{(t)} = \frac{\Pr(X_i | Z_{ij} = 1, p^{(t)}) \Pr(Z_{ij} = 1)}{\sum_{k=1}^{L-W+1} \Pr(X_i | Z_{ik} = 1, p^{(t)}) \Pr(Z_{ik} = 1)}$$

- this comes from Bayes' rule applied to

$$\Pr(Z_{ij} = 1 | X_i, p^{(t)})$$

## The E-step: Estimating Z

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

$$Z_{ij}^{(t)} = \frac{\Pr(X_i | Z_{ij} = 1, p^{(t)}) \cancel{\Pr(Z_{ij} = 1)}}{\sum_{k=1}^{L-W+1} \Pr(X_i | Z_{ik} = 1, p^{(t)}) \cancel{\Pr(Z_{ik} = 1)}}$$

## Example: Estimating Z

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

		0	1	2	3
$p =$	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

$$Z_{i1} = 0.3 \times 0.2 \times 0.1 \times 0.25 \times 0.25 \times 0.25 \times 0.25$$

$$Z_{i2} = 0.25 \times 0.4 \times 0.2 \times 0.6 \times 0.25 \times 0.25 \times 0.25$$

⋮

- then normalize so that  $\sum_{j=1}^{L-W+1} Z_{ij} = 1$

## The M-step: Estimating $p$

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

$$p_{c,k}^{(t+1)} = \frac{n_{c,k} + d_{c,k}}{\sum_b (n_{b,k} + d_{b,k})}$$

pseudo-counts

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

total # of c's in data set

## Example: Estimating $p$

**A C A G C A**

$$Z_{1,1} = 0.1, Z_{1,2} = 0.7, Z_{1,3} = 0.1, Z_{1,4} = 0.1$$

**A G G C A G**

$$Z_{2,1} = 0.4, Z_{2,2} = 0.1, Z_{2,3} = 0.1, Z_{2,4} = 0.4$$

**T C A G T C**

$$Z_{3,1} = 0.2, Z_{3,2} = 0.6, Z_{3,3} = 0.1, Z_{3,4} = 0.1$$

$$p_{A,1} = \frac{Z_{1,1} + Z_{1,3} + Z_{2,1} + Z_{3,3} + 1}{Z_{1,1} + Z_{1,2} + \dots + Z_{3,3} + Z_{3,4} + 4}$$

## The EM Algorithm

- EM converges to a local maximum in the likelihood of the data given the model:

$$\prod_i \Pr(X_i | p)$$

- usually converges in a small number of iterations
- sensitive to initial starting point (i.e. values in  $p$ )

## MEME Enhancements to the Basic EM Approach

- MEME builds on the basic EM approach in the following ways:
  - trying many starting points
  - not assuming that there is exactly one motif occurrence in every sequence
  - allowing multiple motifs to be learned
  - incorporating Dirichlet prior distributions

## Starting Points in MEME

- 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

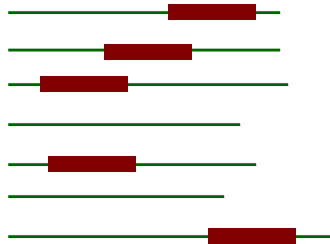
## Using Subsequences as Starting Points for EM

- set values corresponding to letters in the subsequence to  $X$
- set other values to  $(1-X)/(M-1)$  where  $M$  is the length of the alphabet
- example: for the subsequence TAT with  $X=0.5$

$$p = \begin{array}{c} \begin{array}{ccc} & 1 & 2 & 3 \\ \text{A} & 0.17 & 0.5 & 0.17 \\ \text{C} & 0.17 & 0.17 & 0.17 \\ \text{G} & 0.17 & 0.17 & 0.17 \\ \text{T} & 0.5 & 0.17 & 0.5 \end{array} \end{array}$$

## 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)

$\lambda$                       ■ prior prob that any position in a sequence is the start of a motif

$\gamma = (L - W + 1)\lambda$    ■ prior prob of a sequence containing a motif

## E-step in the ZOOPS Model

$$Z_{ij}^{(t)} = \frac{\Pr(X_i | Z_{ij} = 1, p^{(t)}) \lambda^{(t)}}{\Pr(X_i | Q_i = 0, p^{(t)}) (1 - \gamma^{(t)}) + \sum_{k=1}^{L-W+1} \Pr(X_i | Z_{ik} = 1, p^{(t)}) \lambda^{(t)}}$$

- here  $Q_i$  is a random variable that takes on 0 to indicate that the sequence doesn't contain a motif occurrence

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

## M-step in the ZOOPS Model

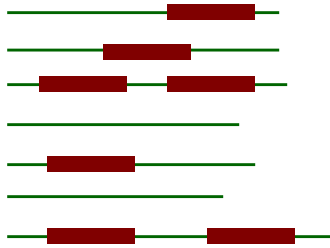
- update  $p$  same as before
- update  $\lambda, \gamma$  as follows

$$\lambda^{(t+1)} = \frac{\gamma^{(t+1)}}{(L-W+1)} = \frac{1}{n(L-W+1)} \sum_{i=1}^n \sum_{j=1}^m Z_{i,j}^{(t)}$$

- average of  $Z_{i,j}^{(t)}$  across all sequences, positions

## The TCM Model

- the TCM (two-component mixture model) assumes *zero or more* motif occurrences per sequence



## Likelihood in the TCM Model

- the TCM model treats each length  $W$  subsequence independently
- to determine the likelihood of such a subsequence:

$$\Pr(X_{ij} | Z_{ij} = 1, p) = \prod_{k=j}^{j+W-1} p_{c_k, k-j+1} \quad \text{assuming a motif starts there}$$

$$\Pr(X_{ij} | Z_{ij} = 0, p) = \prod_{k=j}^{j+W-1} p_{c_k, 0} \quad \text{assuming a motif doesn't start there}$$

## E-step in the TCM Model

$$Z_{ij}^{(t)} = \frac{\Pr(X_{i,j} | Z_{ij} = 1, p^{(t)})\lambda^{(t)}}{\underbrace{\Pr(X_{i,j} | Z_{ij} = 0, p^{(t)})}_{\text{subsequence isn't a motif}}(1 - \lambda^{(t)}) + \underbrace{\Pr(X_{i,j} | Z_{ij} = 1, p^{(t)})\lambda^{(t)}}_{\text{subsequence is a motif}}}$$

- M-step same as before

## Finding Multiple Motifs

- basic idea: discount the likelihood that a new motif starts in a given position if this motif would overlap with a previously learned one
- when re-estimating  $Z_{ij}$ , multiply by  $\Pr(V_{ij} = 1)$

$$V_{ij} = \begin{cases} 1, & \text{no previous motifs in } [X_{i,j}, \dots, X_{i,j+w-1}] \\ 0, & \text{otherwise} \end{cases}$$

- $V_{ij}$  is estimated using  $Z_{ij}$  values from previous passes of motif finding

## Gibbs Sampling

- a general procedure for sampling from the joint distribution of a set of random variables  $\Pr(U_1 \dots U_n)$  by iteratively sampling from  $\Pr(U_j | U_1 \dots U_{j-1}, U_{j+1} \dots U_n)$  for each  $j$
- application to motif finding: Lawrence et al. 1993
- can view it as a stochastic analog of EM for this task
- less susceptible to local minima than EM

## Gibbs Sampling Approach

- in the EM approach we maintained a distribution  $Z_i$  over the possible motif starting points for each sequence
- in the Gibbs sampling approach, we'll maintain a specific starting point for each sequence  $a_i$  but we'll keep resampling these

## Gibbs Sampling Approach

given: length parameter  $W$ , training set of sequences  
choose random positions for  $a$   
do  
  pick a sequence  $X_i$   
  estimate  $p$  given current motif positions  $a$  (update step)  
    (using all sequences but  $X_i$ )  
  sample a new motif position  $a_i$  for  $X_i$  (sampling step)  
until convergence  
return:  $p, a$

## Sampling New Motif Positions

- for each possible starting position,  $a_i = j$ , compute a weight

$$A_j = \frac{\prod_{k=j}^{j+W-1} p_{c_k, k-j+1}}{\prod_{k=j}^{j+W-1} p_{c_k, 0}}$$

- randomly select a new starting position  $a_i$  according to these weights