

# Learning Sequence Motif Models Using Expectation Maximization (EM)

BMI/CS 776

[www.biostat.wisc.edu/bmi776/](http://www.biostat.wisc.edu/bmi776/)

Spring 2018

Anthony Gitter

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

# Goals for Lecture

## Key concepts

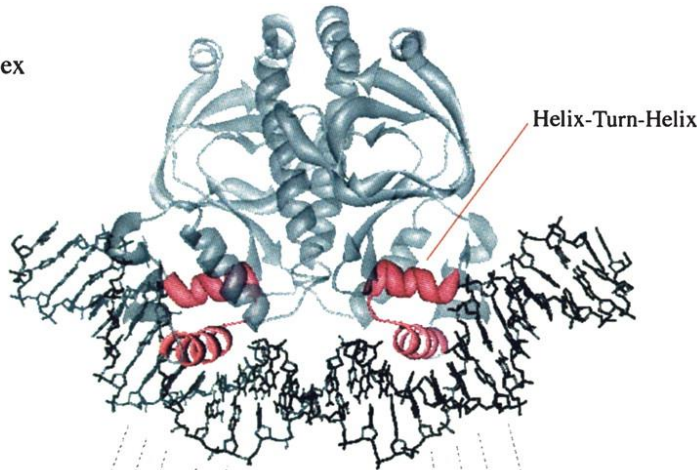
- the motif finding problem
- using EM to address the motif-finding problem
- the OOPS and ZOOPS models

# 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

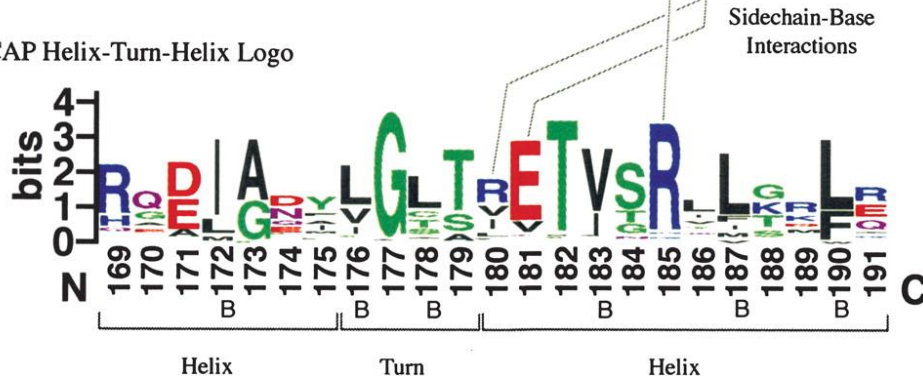
A CAP-DNA Complex



B CAP recognition site DNA Logo



C CAP Helix-Turn-Helix Logo



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

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

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:**

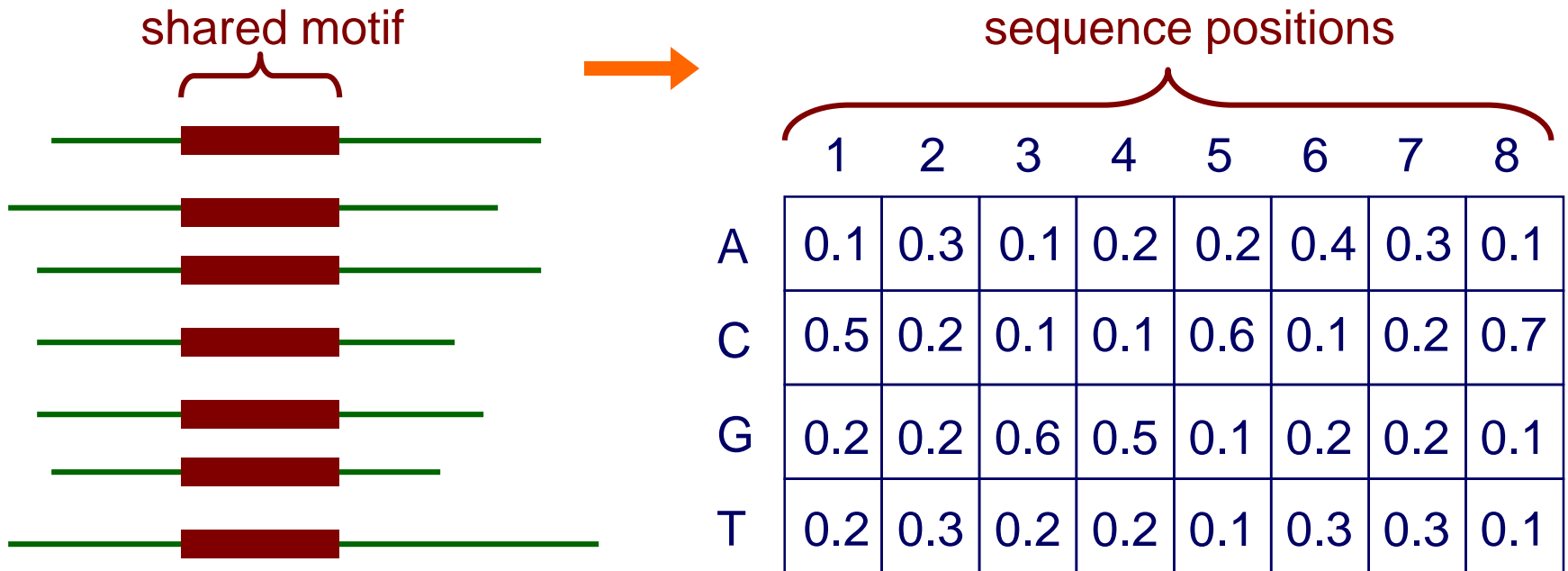
- infer a model of the motif
- predict the locations of the motif occurrences in the given sequences

# 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



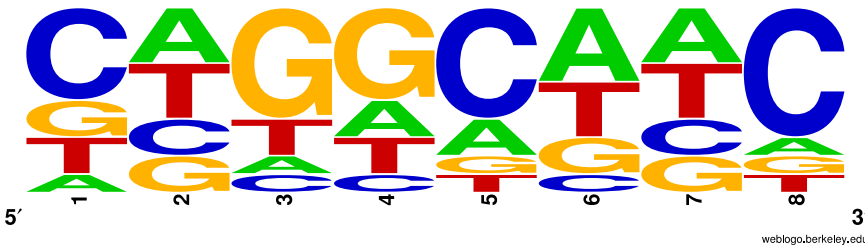
- Each element represents the probability of given character at a specified position

# Sequence Logos

|   | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   |
|---|-----|-----|-----|-----|-----|-----|-----|-----|
| A | 0.1 | 0.3 | 0.1 | 0.2 | 0.2 | 0.4 | 0.3 | 0.1 |
| C | 0.5 | 0.2 | 0.1 | 0.1 | 0.6 | 0.1 | 0.2 | 0.7 |
| G | 0.2 | 0.2 | 0.6 | 0.5 | 0.1 | 0.2 | 0.2 | 0.1 |
| T | 0.2 | 0.3 | 0.2 | 0.2 | 0.1 | 0.3 | 0.3 | 0.1 |

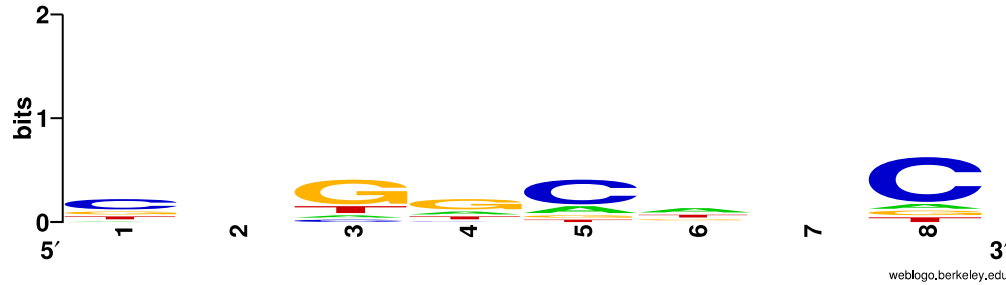


or



frequency logo

[weblogo.berkeley.edu](http://weblogo.berkeley.edu)



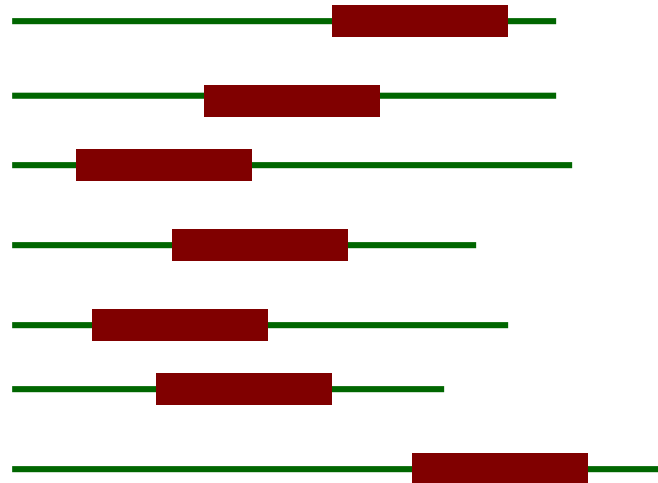
information content logo

[weblogo.berkeley.edu](http://weblogo.berkeley.edu)



# 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.



# Unaligned Sequence Example

- ChIP-chip experiment tells which probes are bound (though this protocol has been replaced by ChIP-seq)

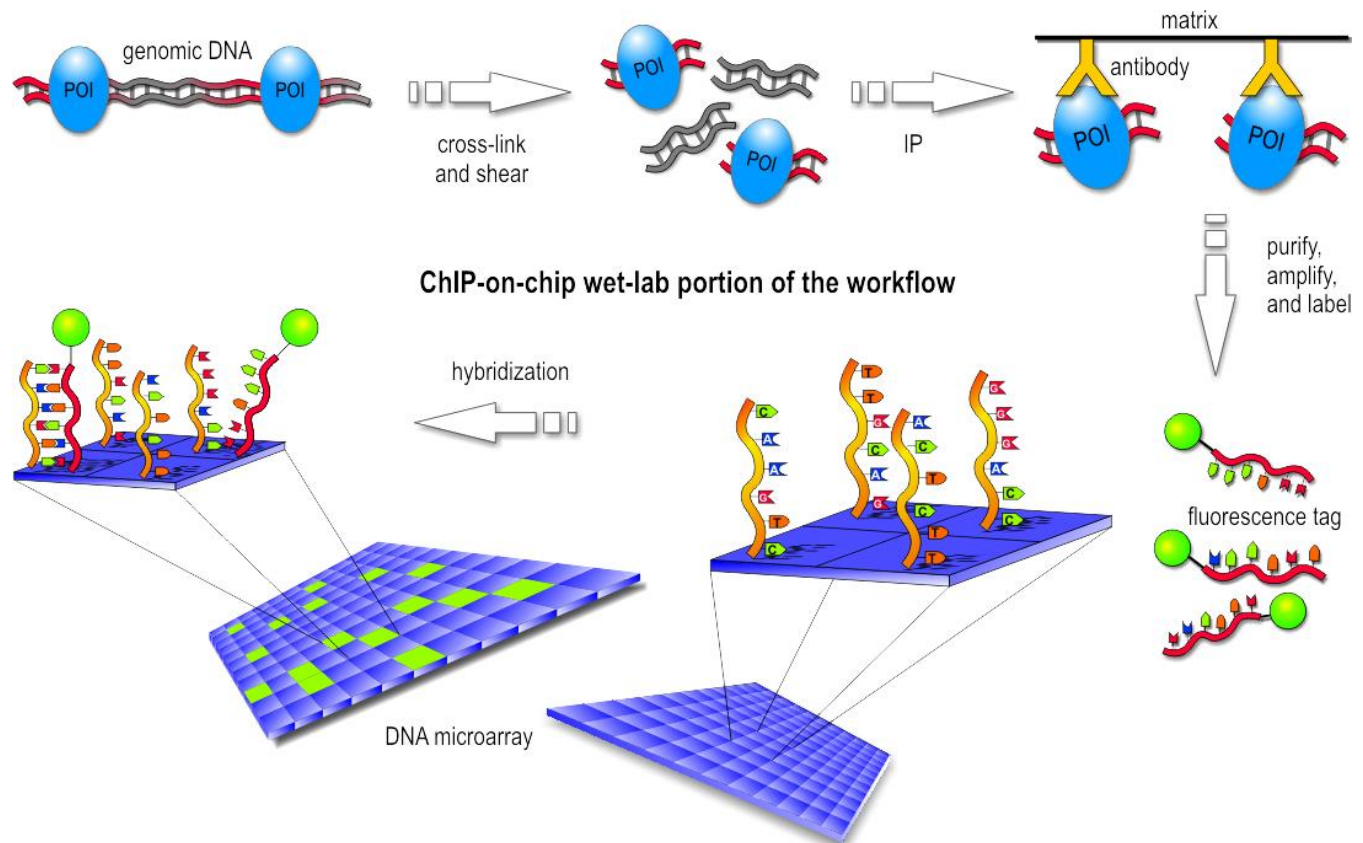
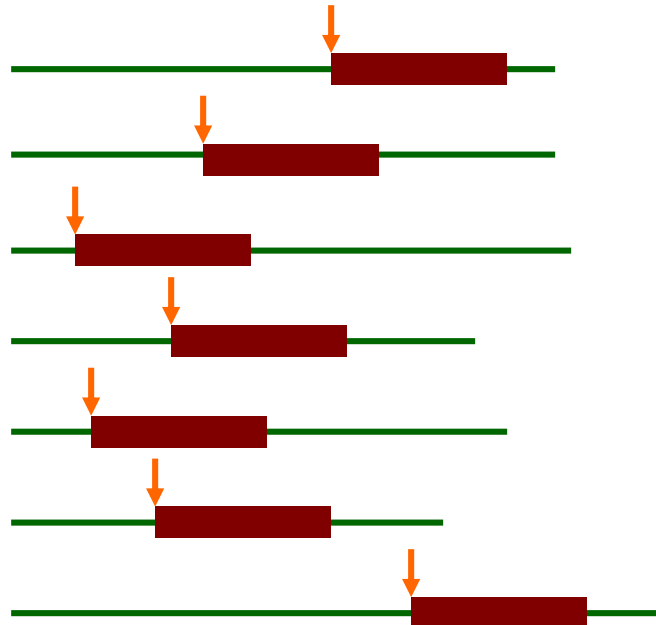


Figure from <https://en.wikipedia.org/wiki/ChIP-on-chip>

# The Expectation-Maximization (EM) Approach

[Lawrence & Reilly, 1990; Bailey & Elkan, 1993, 1994, 1995]

- 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 ( $\theta$ ) for a model (M) and data (D)

$$\theta_{ML} = \operatorname{argmax}_{\theta} P(D | \theta, M)$$

- 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  $\theta$ ):

$$Q(\theta | \theta^t) = \sum_Z P(Z | D, \theta^t) \log P(D, Z | \theta)$$

# 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: **M**ultiple **EM** for **M**otif **E**licitation
- 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

|   | 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 |

$p =$

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  $m = L - W + 1$

|       |   |   |   |   |   |      |   |   |   |
|-------|---|---|---|---|---|------|---|---|---|
| $Z =$ |   |   |   |   |   | 1    | 2 | 3 | 4 |
| G     | T | C | A | G | G | 0    | 0 | 1 | 0 |
| G     | A | G | A | G | T | 1    | 0 | 0 | 0 |
| A     | C | G | G | A | G | 0    | 0 | 0 | 1 |
| C     | C | A | G | T | C | 0    | 1 | 0 | 0 |
|       |   |   |   |   |   | seq1 |   |   |   |
|       |   |   |   |   |   | seq2 |   |   |   |
|       |   |   |   |   |   | seq3 |   |   |   |
|       |   |   |   |   |   | seq4 |   |   |   |



# Probability of a Sequence Given a Motif Starting Position



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

$X_i$  is the  $i$  th sequence

$Z_{i,j}$  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

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

$$p =$$

|   |      | 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 |   |

$$P(X_i | Z_{i,3} = 1, p) =$$

$$P_{G,0} \times P_{C,0} \times P_{T,1} \times P_{G,2} \times P_{T,3} \times P_{A,0} \times P_{G,0} =$$
$$0.25 \times 0.25 \times 0.2 \times 0.1 \times 0.1 \times 0.25 \times 0.25$$

# Likelihood Function

- EM (indirectly) optimizes log likelihood of observed data

$$\log P(X | p)$$

- M step requires joint log likelihood

$$\begin{aligned}\log P(X, Z | p) &= \log \prod_i P(X_i, Z_i | p) \\ &= \log \prod_i P(X_i | Z_i, p) P(Z_i | p) \\ &= \log \prod_i \frac{1}{m} \prod_j P(X_i | Z_{i,j} = 1, p)^{Z_{i,j}} \\ &= \sum_i \sum_j Z_{i,j} \log P(X_i | Z_{i,j} = 1, p) + n \log \frac{1}{m}\end{aligned}$$

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

$++t$

  re-estimate  $Z^{(t)}$  from  $p^{(t-1)}$                    (E-step)

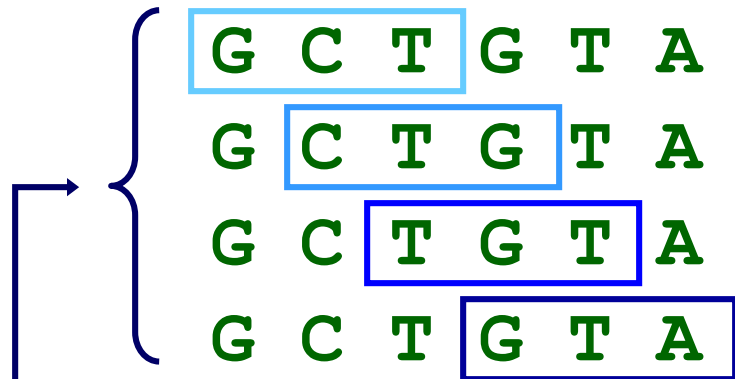
  re-estimate  $p^{(t)}$  from  $Z^{(t)}$                    (M-step)

  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:



|             |      | 1   | 2   | 3   | 4   |
|-------------|------|-----|-----|-----|-----|
| $Z^{(t)} =$ | 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 |

# The E-step: Computing $Z^{(t)}$

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

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

- This comes from Bayes' rule applied to

$$P(Z_{i,j} = 1 | X_i, p^{(t-1)})$$

# The E-step: Computing $Z^{(t)}$

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

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

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

# Example: Computing $Z^{(t)}$

$X_i =$  G C T G T A G

|             |   | 0    | 1   | 2   | 3   |
|-------------|---|------|-----|-----|-----|
| $p^{(t-1)}$ | 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 |

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

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

⋮

- 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

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

pseudo-counts

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

total # of  $c$ 's in data set

sum over positions where  $c$  appears

# Example: Estimating $p$

**A C A G C A**

$$Z^{(t)}_{1,1} = 0.1, Z^{(t)}_{1,2} = 0.7, Z^{(t)}_{1,3} = 0.1, Z^{(t)}_{1,4} = 0.1$$

**A G G C A G**

$$Z^{(t)}_{2,1} = 0.4, Z^{(t)}_{2,2} = 0.1, Z^{(t)}_{2,3} = 0.1, Z^{(t)}_{2,4} = 0.4$$

**T C A G T C**

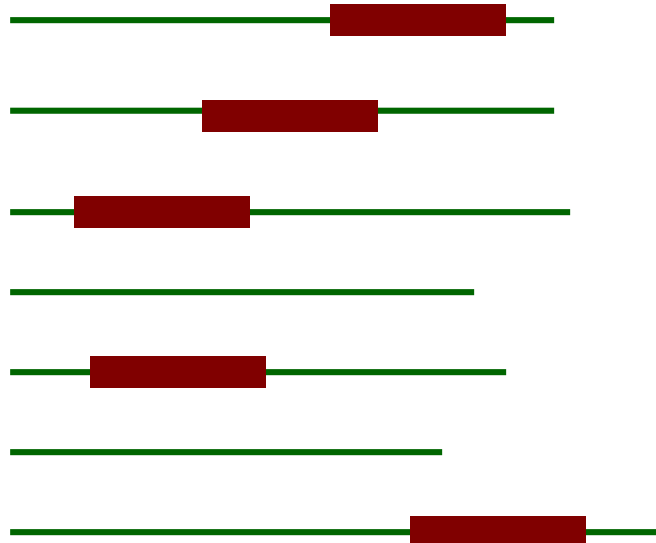
$$Z^{(t)}_{3,1} = 0.2, Z^{(t)}_{3,2} = 0.6, Z^{(t)}_{3,3} = 0.1, Z^{(t)}_{3,4} = 0.1$$

$$p^{(t)}_{A,1} = \frac{Z^{(t)}_{1,1} + Z^{(t)}_{1,3} + Z^{(t)}_{2,1} + Z^{(t)}_{3,3} + 1}{Z^{(t)}_{1,1} + Z^{(t)}_{1,2} \dots + Z^{(t)}_{3,3} + Z^{(t)}_{3,4} + 4}$$

$$p^{(t)}_{C,2} = \frac{Z^{(t)}_{1,1} + Z^{(t)}_{1,4} + Z^{(t)}_{2,3} + Z^{(t)}_{3,1} + 1}{Z^{(t)}_{1,1} + Z^{(t)}_{1,2} \dots + Z^{(t)}_{3,3} + Z^{(t)}_{3,4} + 4}$$

# 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)

$\gamma$

- prior probability of a sequence containing a motif

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

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

# E-step in the ZOOPS Model

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

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

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

$$P(X_i | Q_i = 0, p^{(t-1)}) = \prod_{j=1}^L p_{c_j,0}^{(t-1)} \quad P(Q_i = 0) = 1 - \gamma^{(t-1)}$$

# 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  $\pi = 0.7$

$$P = \begin{array}{c} \begin{array}{ccc} & 1 & 2 & 3 \\ \mathbf{A} & 0.1 & 0.7 & 0.1 \\ \mathbf{C} & 0.1 & 0.1 & 0.1 \\ \mathbf{G} & 0.1 & 0.1 & 0.1 \\ \mathbf{T} & 0.7 & 0.1 & 0.7 \end{array} \end{array}$$

# MEME web server

**MEME Suite 4.11.0**

- Motif Discovery
- Motif Enrichment
- Motif Scanning
- Motif Comparison
- Manual
- Guides & Tutorials
- Sample Outputs
- File Format Reference
- Databases
- Download & Install
- Help
- Alternate Servers
- ▼ Authors & Citing
  - Authors
  - Citing the MEME Suite
- Recent Jobs
- ↔ Previous version 4.10.2



**MEME**  
Multiple Em for Motif Elicitation

Version 4.11.0

MEME discovers novel, **ungapped** motifs (recurring, fixed-length patterns) in your sequences (sample output from sequences). MEME splits variable-length patterns into two or more separate motifs. See this Manual for more information.

### Data Submission Form

Perform motif discovery on DNA, RNA or protein datasets.

**Select the motif discovery mode**

Normal mode  Discriminative mode [?](#)

**Select the sequence alphabet**

Use sequences with a standard alphabet or specify a custom alphabet. [?](#)

DNA, RNA or Protein  Custom

**Input the primary sequences**

Enter sequences in which you want to find motifs. [?](#)

[?](#)

**Select the site distribution**

How do you expect motif sites to be distributed in sequences? [?](#)

**Select the number of motifs**

How many motifs should MEME find? [?](#)

**Input job details**

(Optional) Enter your email address. [?](#)

(Optional) Enter a job description. [?](#)

► **Advanced options**

Note: if the combined form inputs exceed 80MB the job will be rejected.

Version 4.11.0 [Please send comments and questions to: meme-suite@uw.edu](#) [Powered by Opal](#)

[Home](#) [Documentation](#) [Downloads](#) [Authors](#) [Citing](#)

<http://meme-suite.org/>