University of Wisconsin-Madison	Spring 2016
BMI/CS 776: Advanced Bioinformatics	Homework #1
Prof. Anthony Gitter	Due: Tue, Feb 16, 2016 11:59 PM

Assignment Goal

- Gain familiarity with the classic method, MEME, for discovering motifs within biological sequences.
- Consider extensions to the PWM model.

Part 1: MEME Implementation

Write a program, **`LearnMotif`**, that takes as input a set of DNA sequences and an integer **`W`** and learns an OOPS model for a motif of width **`W`**.

Use the EM algorithm called MEME to learn the OOPS model of the motif. You should calculate the log likelihood, $\log P(X|\theta)$, after each iteration (this can be efficiently computed using intermediate values generated during the E step) and stop the algorithm when this value is no longer increasing above a fixed threshold. You must implement multiple random starting points but do not need to implement the exhaustive subsequence approach described during lecture. Use simple pseudo-counts of ones ($d_{c,k} = 1$) when estimating the model parameters.

Your program should be callable from the command line as follows:

```
LearnMotif <sequences> W <model> <positions> <subsequences>
```

where

- **`<sequences>`** is a text file containing DNA sequences *one per line*. Assume each sequence contains exactly one motif. The sequences will all have the same length.
- **`W`** is the width of the motif to learn.
- **`<model>`** is the name of the text file into which the program will output the learned motif model (i.e., the probabilities for each nucleotide in each column) in a *tab-delimited format* with the background frequencies in the first column.
- **`<positions>`** is the name of the text file into which the program will output the predicted starting location of the motif in each sequence *one per line*.
- **`<subsequences>`** is the name of the text file into which the program will output the subsequences corresponding to the motif occurrence in each sequence *one per line*.

Input files and sample scripts can be downloaded from
`https://www.biostat.wisc.edu/bmi776/hw/hw1_files.zip`

If you are using a language that is not compiled to machine code such as Java, then you should make a small script called **`LearnMotif`** that accepts the command line arguments and invokes the appropriate source code and interpreter. Sample scripts are contained in the **`Hw1SampleScripts`** subdirectory.

Recall that only standard libraries for the chosen language may be used. Third-party libraries must be approved by the instructor or TA on the Piazza forum.

To test your program, you may use the file **`hw1_example.txt`** to discover the hidden consensus motif **`CGGAA`** and the file **`hw1_example2.txt`** to discover the consensus motif **`AACTGTGG`**.

Part 2: Motif Discovery

Use your **`LearnMotif`** program to discover the motif of width 8 hidden in sequences in the file **`hw1_hidden_motif.txt`**

Test that your program runs on the biostat server by running

```
  LearnMotif hw1_hidden_motif.txt 8 model.txt positions.txt subsequences.txt
```

from your **`handin`** directory and leaving the output files in the **`handin`** directory.

Part 3: Sequence Logo

Construct a sequence logo for the predicted motif sequences from Part 2 by using the **`WebLogo`** application (**`http://weblogo.berkeley.edu/`**). You can use the contents of the file **`subsequences.txt`** as input to the **`WebLogo`** application.

Part 4: Transcription Factor

Search the **`JASPAR`** transcription factor binding profile database (**`http://jaspar.genereg.net/`**) using the profile matrix that you learned in Part 2. Which transcription factor binds to this motif? You should be able to use the contents of the **`model.txt`** file for this search, after deleting the first (background) column.

Part 5: Alternative Motif Models

(A) The PWM P used by MEME to represent the motif model assumes independence among different columns of the matrix. One straightforward extension to P is a dinucleotide weight matrix (DWM)[1] D. $D_{c_1 c_2; ij}$ represents the probability of observing each pair of nucleotides c_1 and c_2 at positions i and j. Consider the following DWM and PWM models that recognize a motif of width 4.

[1] Siddharthan, R. (2010). Dinucleotide Weight Matrices for Predicting Transcription Factor Binding Sites: Generalizing the Position Weight Matrix. PLoS ONE, 5(3).

D	$i = 1, j = 2$	$i = 2, j = 3$	$i = 3, j = 4$
AA	0.01	0.01	0.43
AT	0.01	0.43	0.01
TA	0.43	0.43	0.43
TT	0.43	0.01	0.01
$otherwise$	0.01	0.01	0.01

P	$i = 1$	$i = 2$	$i = 3$	$i = 4$
A	0.1	0.4	0.4	0.7
C	0.1	0.1	0.1	0.1
G	0.1	0.1	0.1	0.1
T	0.7	0.4	0.4	0.1

Calculate $P(X|P)$ and $P(X|D)$ for `X = TATA` as well as $P(Y|P)$ and $P(Y|D)$ for `Y = TTTA`.

(B) In addition to recognizing real DNA binding sites, we would also like to have a motif model with high specificity, meaning it minimizes the number of false positive matches. Suppose we use a genome-wide epigenomic assay to identify regions of a genome that are unlikely to be bound by any transcription factor. Assume that there are 1,000,000,003 base pairs in this negative set (1 billion potential starting positions for a motif of width 4). For simplicity, further assume that the 1 billion base pairs occur in one contiguous block and that their background nucleotide frequencies are uniform:

	$background$
A	0.25
C	0.25
G	0.25
T	0.25

Define X_i to be the 4 nucleotides starting at position i in the negative set (that is, the subsequence of length 4 from X_i to X_{i+3}) and $P(X_i|P)$ and $P(X_i|D)$ to be the probability of observing that subsequence with the PWM and DWM, respectively. If this likelihood exceeds 0.07, we call X_i a motif match. How many motif matches are expected for the PWM model in **(A)**? For the DWM? Is the PWM or DWM preferable for minimizing false positive matches?

Show your work for partial credit.

(C) Suppose you extend MEME to identify PWM and DWM models. Drawing inspiration from MEME's approach for choosing the motif width, how will you decide whether to return the optimal PWM or DWM for a given set of input sequences? Will you return the PWM or DWM if the likelihoods of both models are nearly identical?

Submission Instructions

- Login to the biostat server **`mi1.biostat.wisc.edu`** or **`mi2.biostat.wisc.edu`** using your BMI (biostat) username and password.
- Copy any relevant files to the directory **`/u/medinfo/handin/bmi776/hw1/<USERNAME>`** where **`<USERNAME>`** is your BMI (biostat) username. For the programming part, make sure to submit the source code as well as any required files to run the program on the biostat server. For the rest of the assignment, compile all of your answers in a single file and submit as **`solution.pdf`**.