

Assignment Goals

- i. Gain familiarity with MEME and Gibbs sampling for discovering motifs within biological sequences.
- ii. Use sequence logos as a graphical tool for visualizing motifs.
- iii. Understand parameter estimation of motif models that incorporate prior knowledge.

Submission Instructions

- To turn in your assignment, log in to the server **mi1.biostat.wisc.edu** or **mi2.biostat.wisc.edu** using your BMI (biostat) username and password.
- Copy all relevant files to the directory

/u/medinfo/handin/bmi776/hw1/<USERNAME>

where **<USERNAME>** is your BMI (biostat) username. Submit all of your Python source code and test that it runs on the biostat server.

- For the rest of the assignment, compile all of your answers in a single file and submit as **solution.pdf**.
- Write the number of late days you used at the top of **solution.pdf**.
- For the written portions of the assignment, show your work for partial credit.

Part 1: MEME Implementation (50 points)

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

Implement the EM algorithm from MEME to learn the motif. You should calculate the likelihood $P(X|\theta)$, or the log likelihood $\log P(X|\theta)$, after each iteration and stop the algorithm when the change in this value is less than a fixed threshold (e.g., $1e-3$). Use the exhaustive subsequence approach described during lecture to choose the starting point for EM. Use a pseudocount of 1 ($d_{c,k} = 1$) for your parameter estimates.

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

```
python learn_motif.py \  
    --width=W \  
    --model=<model> \  
    --positions=<positions> \  
    --subseqs=<subsequences> \  
    <sequences>
```

where

- **<sequences>** is a text file containing DNA sequences one per line. 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 probabilities in the first column.
- **<positions>** is the name of the text file into which the program will output the predicted starting position of the motif occurrence in each sequence one per line. Use 0 for the first position in a sequence.
- **<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.

Example input files **example1.txt** and **example2.txt**, their corresponding output files, and the template **learn_motif.py** with argument parsing code can be found in the **hw1_files** directory. Your program will be evaluated on the example inputs and additional datasets that will be kept private.

Part 2: Sequence Logo (15 points)

Suppose that you have estimated the motif model in **example2_model.txt** by running MEME and want to convert it to an information content logo.

- (A) First, compute the height of the entire stack of characters at each position. Next, determine the height of individual characters at each position. Note that you should ignore the first, background column. (10 points)
- (B) Construct a logo for the same motif using the WebLogo application (<http://weblogo.threeplusone.com/create.cgi>). Use **example2_subseqs.txt** as input to the WebLogo application. Use PDF as the output format, select logo size as large and color scheme as classic. Save the logo as **logo.pdf** and submit it in your handin directory. The generated logo may be used for verifying your result in (A). (2 points)
- (C) Explain what information is conveyed by an information content logo. What can you say about the motif in **example2_model.txt**? (3 points)

Part 3: Gibbs Sampling (25 points)

Suppose that you are running the Gibbs sampling algorithm for motif finding on a set of 10 input sequences with a motif width of 4 nucleotides. Further, suppose that the current state of the Gibbs sampler is as shown below, with the motif occurrences indicated by the underlined nucleotides.

1	CATGTGAA
2	CAGCAGGG
3	ACCTCTTC
4	<u>CAGACATG</u>
5	ACCTATCG
6	GCGGCAGT
7	<u>GTGTAGTT</u>
8	CCAGGAAG
9	ATGACCGG
10	GGATAGTA

- (A) Suppose that in the next iteration of the sampler, sequence **5** is picked and the motif position in that sequence is to be resampled. Compute the parameter estimates, p , that are calculated in the predictive update step, given that sequence **5** has been selected. Use a pseudocount of 1 for your estimates and assume a standard PWM model that does not incorporate any prior knowledge. Be sure to give the estimates for both the PWM and background parameters. (10 points)
- (B) Given the parameter estimates you computed in (A), compute the probability that position 2 will be selected as the start of the motif occurrence in sequence **5** (which corresponds to a motif occurrence of **CCTA**). (5 points)
- (C) Repeat part (A), but with a palindromic motif model instead of a standard (non-palindromic) PWM model. (10 points)

Part 4: More on PWMs (10 points)

Consider the same set of input sequences in Part 3, with the motif occurrences indicated by the underlined nucleotides. Answer *one* of the following questions.

- (A) (*Motif width*) The motif has a width of 4 nucleotides. Consider extending the PWM to width 5. If you fix the motif starting positions and derive the new PWM and background parameters, will the log likelihood of the observed data increase, decrease, or stay unchanged under the new model?

More generally, consider how increasing the PWM width affects the log likelihood of an arbitrary set of sequences with known motif starting positions. Choose one of the following options and justify your answer by a formal argument or a mathematical proof.

- The log likelihood in the optimal PWM model with width $W+1$ **will always be greater than or equal to** the log likelihood in the optimal model with width W .
- The log likelihood in the optimal PWM model with width $W+1$ **will always be less than or equal to** the log likelihood in the optimal model with width W .
- The log likelihood in the optimal PWM model with width $W+1$ **could be greater than, equal to, or less than** the log likelihood in the optimal model with width W depending on the sequence data.

You may assume that every motif occurrence of width W can be extended. That is, the first (last) position of a motif occurrence is not at the start (end) of the sequence.

(B) (*Dirichlet mixture prior*) Recall that the four nucleotides A, C, G and T can be partitioned into two subsets of chemically similar nucleotides: the *purines* (A and G) and the *pyrimidines* (C and T). We could incorporate this knowledge into our standard PWM model by using a Dirichlet mixture prior with two components, one for purines and the other for pyrimidines. Suppose that the parameters for the two Dirichlet components are those given in the table below and that we assume a uniform prior distribution over the two components. Repeat Part 3(A), but with parameter estimates calculated using this Dirichlet mixture prior.

	$j = 1$ (purine)	$j = 2$ (pyrimidine)
$\alpha_A^{(j)}$	4	1
$\alpha_C^{(j)}$	1	4
$\alpha_G^{(j)}$	4	1
$\alpha_T^{(j)}$	1	4