CS313 Exercise 7 Cover Page

Name(s): _____

In the *TIME* column, please estimate the time you spent on the parts of this exercise. Please try to be as accurate as possible; this information will help us to design future exercises.

PART	TIME	SCORE
Exercise		

In this Exercise, you will be searching for motif instances (i.e., regulatory sites) in DNA sequences upstream of genes that may be similarly regulated. In particular, you will be searching for motif instances upstream of genes that are likely to be homologous, and you will be searching for motif instances upstream of genes that are likely to be co-regulated as evinced by gene expression data. You will be using both Expectation-Maximization and Gibbs motif sampling methods to identify potential motifs.

Task 1: Motif Discovery - Expectation Maximization Example

Suppose that we have the following four DNA sequences, and we believe the sequences share a regulatory site that is four nucleotides in length:

Sequence 1: G C T G A C T A
Sequence 2: T A C T A G G T
Sequence 3: C C T A C T G G
Sequence 4: A G G G T C T A

Using an expectation-maximization algorithm, determine a candidate regulatory site (with length of 4 nucleotides) for each of the four sequences. Rather than initially seed the algorithm with four randomly chosen 4-mers, one from each sequence, you should seed the algorithm with the first 4-mer in each sequence. That is to say, the algorithm should begin with the following four 4-mers: `GCTG`, `TACT`, `CCTA`, and `AGGG`. Using these four 4-mers as seeds, you should repeat the following two steps until convergence: (1) build a motif model from the current four 4-mers and (2) in each of the four sequences, find the most likely 4-mer, i.e., the 4-mer that corresponds to the motif model with highest likelihood.

For every repetition of the abovementioned two steps in the expectation-maximization algorithm, write below the motif model and the four 4-mer sequences with highest likelihood.

What is the consensus sequence for the motif that you found?

How many of the final four 4-mers match the consensus sequence?

How many of the final four 4-mers match one of the four initial 4-mer seeds?

Task 2: Motif Discovery - Information Content

Prior to completing this task please read the primer by D'haeseleer entitled "What are DNA sequence motifs?". The article can be found in `Exercise7` folder in the `download` folder on the CS server.

Suppose that we execute an expectation-maximization algorithm on four sequences and we find four regulatory sites, one in each sequence, and we generate a motif model, *M1*, from the four regulatory sites. Now suppose that we execute the expectation-maximization algorithm again on the same four sequences and find a different set of four regulatory sites, and we generate a motif model, *M2*, from these new four regulatory sites. How can we tell which motif model, *M1* or *M2*, is better?

The *information content* (also called *relative entropy*) of a motif model can be used to determine which motif model is better, *M1* or *M2*. Suppose we have the following motif model, which we will call *M*:

	1	2	3	4
A	A1	A2	A3	A4
C	C1	C2	C3	C4
G	G1	G2	G3	G4
T	T1	T2	T3	T4

The information content of position 1 in *M* is defined as

$$I_1 = \sum_{\beta \in \{A,C,G,T\}} \left(M_{\beta,1} \log_2 \frac{M_{\beta,1}}{Q_\beta} \right)$$

$$I_1 = M_{A,1} \log_2 \frac{M_{A,1}}{Q_A} + M_{C,1} \log_2 \frac{M_{C,1}}{Q_C} + M_{G,1} \log_2 \frac{M_{G,1}}{Q_G} + M_{T,1} \log_2 \frac{M_{T,1}}{Q_T}$$

$$I_1 = \mathrm{A1} * \log_2 \frac{\mathrm{A1}}{Q_A} + \mathrm{C1} * \log_2 \frac{\mathrm{C1}}{Q_C} + \mathrm{G1} * \log_2 \frac{\mathrm{G1}}{Q_G} + \mathrm{T1} * \log_2 \frac{\mathrm{T1}}{Q_T}$$

where Q_β is the background frequency of nucleotide β. For examples, in a genome with a GC-content of 50%, $Q_A = 0.25$ and $Q_C = 0.25$, and in a genome with a GC-content of 42%, $Q_A = 0.29$ and $Q_C = 0.21$.

Now, since I_1 is the information content of position 1 in *M*, I_2 and I_3 and I_4 are the information contents of positions 2 and 3 and 4 in *M*, respectively. The total information content of *M*, then, is given as the sum of the information content at each position:

$$I = I_1 + I_2 + I_3 + I_4$$

What is the information content of a motif model where A1=A2=A3=A4=Q_A and C1=C2=C3=C4=Q_C and G1=G2=G3=G4=Q_G and T1=T2=T3=T4=Q_T?

Assuming a background frequency of 25% for each of the four nucleotides, what is the information content of the final motif model that you found in Task1 (you may assume that the contribution to information content of $M_{\beta,i}$ is zero if $M_{\beta,i} = 0$)?

Task 3: Motif Discovery - Phylogenetic Footprinting

To start, download the `modA.txt` file from the `download` folder on the CS server.

Where does this file come from? We took a gene (**modA**) that codes for a molybdate transporter subunit in the bacterial organism *Escherichia coli* and BLASTed its sequence against the genomes of similar bacterial organisms. From the BLAST results, we selected 8 significant alignments from 8 different bacteria species, i.e., we selected 8 genes in other organisms which appear to be orthologous to the *E. coli* gene. We then retrieved the 500 nucleotides upstream (in front) of each of these 9 genes (the 8 putative orthologs plus the original gene).

You should submit these 9 sequences to the following Motif Finder:

http://cs.wellesley.edu/~btjaden/Motif/

We will use this program to search for motif instances, i.e., short regions of DNA that are common to the 9 sequences. Why do we expect common patterns in these sequences? If in fact the *genes* are homologous, then the regulatory mechanisms for the genes may be homologous. When running the program, you should fill in "16" for the motif width (we are searching for motifs of length 16).

Scan through the results. There may be no motif found or there may be multiple motifs found. If one or more motifs were found, search for the line starting with "Num Motifs:" to see how many instances of the motif were found. This line should report 9, if an instance of the motif was found in all nine input sequences. Run the program at least five different times, keeping track of the best motif that was found (the score of a motif can be found in the line starting with "Log Motif portion of MAP for motif…", where larger negative numbers represent better scoring motifs). Answer the questions below for the best motif that you found.

Is the motif instance identical in all 9 sequences?

What is the *consensus sequence* for this motif?

Do any of the 9 motif instances match the consensus sequence exactly?

Do you notice any special property of the consensus sequence (hint: Madam, I'm Adam)? Why might a regulatory site have this property? For further details, check out the following site:

http://www.cryst.bbk.ac.uk/PPS95/course/6_super_sec/super5.html

Task 4: Motif Discovery - Similarly Expressed Genes

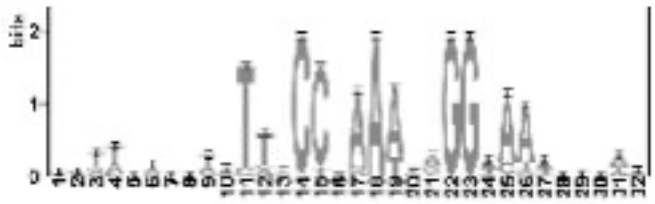
In this task, we have a set of 20 yeast genes which show evidence of being expressed in the cell under the same conditions at the same times. The evidence comes from gene expression assays, i.e., yeast gene expression data was clustered and these 20 genes clustered in the same group. Again, we extracted the upstream sequences for each of these 20 genes.

Download the `ECB.txt` file from the `download` folder on the CS server.

You should submit these 20 sequences to the Motif Finder. These 20 sequences come from the same genome, why do we expect common patterns in these sequences? Since the gene expression data suggests that the genes are co-expressed, i.e., they are expressed under the same conditions, the genes may be regulated by the same transcription factor and may contain common regulatory sites. When running the program, you should fill in "12" for the motif width. Run the program at least five different times, keeping track of the best motif that was found.

What is the *consensus sequence* for this motif? Do any of the 20 pattern instances match the consensus sequence exactly?

For any of the motifs that you found, does the *consensus sequence* or *motif model* look similar to that of the ECB (early cell cycle box) motif, or its reverse complement, which is a known motif in yeast (shown below)?



Now extract the upstream sequence (e.g., 1000 nucleotides) for the *rpn12* yeast gene and for the *hxk1* yeast gene. Add these two to the list of 20 sequences. Now re-run the program at least five times to search for a motif. Does the *rpn12* gene or the *hxk1* gene appear to contain a good match to the ECB (early cell cycle box) motif?