BMI/CS 776 Spring 2015
Homework #2

Prof. Colin Dewey

Due Thursday, February 26th, 2015 by 11:59pm

The goal of this assignment is to become more familiar with cis-regulatory module learning methodology.

To turn in your assignment, copy all relevant files to the directory:
`/u/medinfo/handin/bmi776/hw2/USERNAME`
where `USERNAME` is your account name for the BMI network.

1. As we saw from the Noto and Craven approach, the motif finding problem can be addressed through the use of HMMs. Consider the problem of finding a motif of width 5.

   (a) Draw the state transition diagram for an HMM that could represent the OOPS model of MEME.

   (b) Draw the state transition diagram for an HMM that represents the ZOOPS model of MEME.

   (c) Draw the state transition diagram for an HMM that models sequences that have zero, one, or multiple motif occurrences.

   (d) Recall that the EM algorithm for HMMs is called the Baum-Welch algorithm. Which values computed during the Baum-Welch algorithm are equivalent to the expected values of the Z_{ij} random variables computed by MEME? Justify your answer.

2. There are a number of logical and spatial features of cis-regulatory modules that are explicitly modeled by the Noto and Craven approach but that are not addressed by FIRE. In this problem, you will explore how to use the FIRE approach to address these features. Suppose that you have data set of gene promoters with each gene assigned to one of K clusters (e.g., a clustering from some gene expression data). Suppose that you run FIRE and determine that motifs A, B, and C all have significant mutual information with the clustering. For each of the following features, describe a pair of random variables of which you would compute the mutual information in order to determine if the feature is present.

(a) Assuming that motifs A and B are interacting, that they generally occur within a certain distance of each other.

(b) Assuming that motifs A and B and C are all interacting, that they generally occur in a specific order.

(c) That the three motifs function together in a cis-regulatory module with the logical combination (A AND (B OR C)).

3. The Noto and Craven approach performs binary classification of promoter sequences, i.e., it predicts whether each sequence is either a positive or negative example. Suppose that instead of having a labeled set of positive and negative sequences as training data, we have a set of sequences with each sequence assigned to one of K clusters (i.e., the FIRE setting).

   (a) Explain how to modify the Noto and Craven approach so that we could (in theory) learn a cis-regulatory module for each cluster and use the resulting model to perform *multi-way* classification (i.e., given a new sequence, we wish to predict the cluster to which it belongs).

   (b) Is the NOT operator still relevant in the multi-way classification scenario? Explain why or why not.

4. Suppose you have a set of genes partitioned into two clusters (0 and 1) and have noted the presence (1) or absence (0) of each of two motifs in the promoter of each gene. Let M_1, M_2, and C be the random variables indicating the presence/absence of motif 1, the presence/absence of motif 2, and the cluster assignment, respectively, for a gene. Suppose you observe the following frequencies for each of the possible configurations of motifs and cluster assignments:

M_1	M_2	C	frequency
0	0	0	0.36
0	0	1	0.04
0	1	0	0.02
0	1	1	0.08
1	0	0	0.09
1	0	1	0.01
1	1	0	0.08
1	1	1	0.32

   (a) Compute the mutual information of M_1 and C

   (b) Compute the mutual information of M_2 and C

(c) Compute the conditional mutual information of M_1 and C given M_2. For three random variables, X, Y, and Z, the conditional mutual information of X and Y given Z is defined as

$$I(X;Y|Z) = \sum_z P(z) \sum_x \sum_y P(x,y|z) \log_2 \left(\frac{P(x,y|z)}{P(x|z)P(y|z)} \right) \tag{1}$$

and can be interpreted as the expectation of the mutual information between X and Y given that Z is known.

(d) What does your answer from (c) tell you about the association between M_1 and C?