COMP 462 / 561 - Homework #3
Due on December 3rd 2014

Question 1 (40 points).

Consider the set of promoters associated to genes that are up-regulated upon increase of temperature, versus the set of promoters of genes that are not significantly affected. See data at:
http://www.cs.mcgill.ca/~blanchem/561/promoterPositive.fa
http://www.cs.mcgill.ca/~blanchem/561/promoterNegative.fa

Write a program to determine what is the 6-nucleotide consensus sequence (made of the extended alphabet A, C, G, T, A|C, A|G, A|T, C|G, C|T, G|T, A|C|G|T) that is the most likely to represent the transcription factor binding site that may be the common cause of the up-regulation of these genes. It is up to you to decide how to approach this problem. Don't go overboard here – a relatively simple approach should be able to get you to the answer. No need to implement any of the fancy algorithms that researchers have published in research papers – stick to the ideas and math we've seen in class.

Question 2 (60 points).

Choosing the appropriate type of model to represent the binding sites of a given transcription factor is not always obvious. Position weight matrices provide a very rich representation but may need a lot of known sites to be estimated correctly. Consensus sequences are less flexible but require fewer examples to be learned correctly. Here, we will describe a way to select the most appropriate model. A motif is a function f: Σ^6 -> {0,1}. Some motifs can be represented using consensus sequences. In that case, we have

f (w) = 1 if w matches a certain regular expression,
0 otherwise.

A motif can also be represented by a position weight matrix. In that case, we have

f(w) = 1 if the likelihood of w under a certain PWM is larger than t,
0 otherwise

for some threshold t.

The usefulness of a motif model depends on two things: (i) how accurately it can be learned from a limited number of training examples, and (ii) how accurately it can predict unseen sites. Consider a set of binding sites for transcription factor F: X1, X2, …, Xm, each of length k. Consider also a set of binding sites Y1…Yn (each of length k) of sites that are not bound by transcription factor F. The usefulness of a motif can be assessed by measuring its sensitivity and specificity using leave-one-out cross-validation (LOOCV):

```
TruePos = TrueNeg = 0
for i = 1…m do {
    Train function f on the positive set X1…Xi-1 Xi+1… Xm
    If f(Xi)=1 then  TruePos = TruePos + 1
}
Train function f on the entire positive set X1… Xm
for i = 1…n {
    If f(Yi)=0 then TrueNeg = TrueNeg + 1
}
Sensitivity = TruePos / m     /* Fraction of true positives correctly predicted */
Specificity = TrueNeg / n     /* Fraction of true negatives correctly predicted */
```

Two data sets will be studied:
Data set #1: http://www.cs.mcgill.ca/~blanchem/561/hw3/dataset1
Data set #2: http://www.cs.mcgill.ca/~blanchem/561/hw3/dataset2

a) (15 points) Write a program to assess the sensitivity and specificity of the following motif model and training procedure. Motifs are represented by consensus sequences on the alphabet A, C, G, T, A|C, A|G, A|T, C|G, C|T, G|T, A|C|G|T and are learned from a set of positive sites using the following rule:
If nucleotide x occurs in at least 90% of the positive sites at a given position, then the consensus is x. Otherwise, if nucleotides x and y together occur in at least 90% of the sites, then the consensus is x|y. Otherwise, the consensus is A|C|G|T. Note that this training procedure completely ignores the negative training set.
Calculate the sensitivity and specificity of this method on datasets #1 and #2.

b) (15 points) Write a program to assess the sensitivity and specificity of the following motif model and training procedure. Motifs are represented by position weight matrices trained using the following procedure:
M(i,j) = (# of sites with nucleotide i at position j) / (# of sites)

Remains the question of selecting the threshold above which a site will be predicted as positive. One way to do so is to choose threshold t to minimize the number of classification errors (false-positives + false-negatives) on the given training set of sites. What is the sensitivity and specificity of the method on dataset #1 and #2?

c) (15 points) Using the same method as described in (b), one can instead leave the threshold t as a free parameter that can be set by the user. Then, any choice of t will lead to a particular sensitivity and specificity. One can then draw the sensitivity-specificity curve, which is a curve showing, for every possible values of t, the resulting sensitivity and specificity. Generate and draw this curve for dataset #1 and dataset #2.

d) (15 points) One big problem with the type of training procedure for PWM seen in (b) is that if a nucleotide has never been observed at a given position of the positive sites, then any site that would contain it would be predicted as a negative (the likelihood ratio would be zero). The problem is caused by the fact that a probability distribution cannot be learned accurately from a finite sample. In particular, rare events are often unobserved in a small sample, even though their probability is not zero. One way of partially addressing this problem is to use pseudocounts, which consist of imaginary observations of A, C, G, and T. Specifically, in addition to the actual nucleotides observed in the real data, one would add a certain number x of fake observations:
M(i,j) = ((# of sites with nucleotide i at position j) + x) / (# of sites + 4 x)

Generate and draw the sensitivity-specificity curve for dataset #1 and dataset #2, for x=1. Is it better than that obtained in (c) ?