Problem Set 7

Due 11:59pm, Wednesday, November 24th

Your name:

Collaboration is allowed on this homework. You must hand in homework assignments individually. List the names of the people you worked with:

1. A particular gene has been shown to be regulated by two different transcription factors, an activator and a repressor. Recent experimental evidence has shown that the activator and the repressor bind very similar DNA motifs. Your co-workers have a mixed set of labeled motifs, including both activator and repressor binding sites. Using this set, they construct a frequency matrix (shown below) and the associated PSSM.

nt	1	2	3	4	5	6	7
A	0.1	0.25	0.5	0	0.25	0.5	0.8
T	0.05	0.25	0	0.5	0.25	0	0
G	0.05	0.25	0.25	0	0	0.5	0.1
C	0.8	0.25	0.25	0.5	0.5	0	0.1

   (a) After applying this PSSM to new unlabeled data, you find many sites that do not match the activator binding site or the repressor binding site. You take another look at the original training data and discover that all of the activator binding sites have ATC in positions 3, 4 and 5, respectively. The repressor binding sites have either a G or a C in position 3, always a C in position 4, and either an A or a T in position 5. Columns 1, 2, 6 and 7 are the same for both binding sites.

   Why does the PSSM assign high scores to motifs that are not known binding sites?

(b) In order avoid assigning high scores to non-binding motifs, you construct two separate PSSMs, one for the activator and one for the repressor. Show the frequency matrices for these PSSMs.

(c) One of your coworkers suggests that you design a Hidden Markov model (HMM1) that emits *both* the activator binding site and the repressor binding site. Draw a state diagram for a model that will assign high probabilities to the binding sites of the activator and the repressor, but not to other motifs. Show non-zero probability transitions as arcs in your state diagram and give the initial, transition, and emission probabilities for each state.

(d) Suppose you apply HMM1 to a new, unlabeled sequence and determine that it contains a high probability motif. What algorithm would you use to determine whether this motif is an activator binding site or a repressor binding site? Describe what output you would expect from this algorithm and how you would determine from that output whether the predicted motif is a binding site for the activator or for the repressor.

(e) A passive-aggressive co-worker comes to you with the sequences of upstream regions of several genes. They claim that there is a third transcription factor (TF3) that regulates this set of genes. They tell you the length of the TF3 binding site, but withhold any information about where it might bind. Your co-worker demands that you build a new model, HMM3, that will recognize the sequence motif of the TF3 binding site.

After you have designed the topology of the HMM3 model, what algorithm(s) would you use to determine the parameters of this model? Describe what input you would use, how you would apply this algorithm to that input, and what the output would be.

(f) Now that you have determined both the topology and the parameters of the HMM3 model, what algorithm would you use to find the actual TF3 binding site in each of the unlabeled sequences provided by your co-worker? Describe what the input would be, how you would apply this algorithm, what the output would be, and how you would determine the motif from the output.

(g) Your hyperactive co-worker comes running from the sequencer with yet another unlabeled sequence, an intergenic coding sequence that is 5' to a gene of interest. This co-worker wants to know which of the three transcription factors regulates this gene.

Using your two HMMs, you determine that the new sequence has only one high-probability motif. What algorithm(s) and method would you use to determine whether this motif is more likely to be a TF3 binding site or a binding site for the activator/repressor pair described in (a)? Describe how you would apply this algorithm, what the output would be, and how you would determine from that output whether the predicted motif is a TF3 binding site or a TF1/2 binding site. If it is a TF1/2 binding site, how would you determine whether it binds to the activator or the repressor?

2. *Hidden Markov model design:*

   (a) Some prokaryotes use non-canonical start codons, in addition to the canonical start, `AUG`. A study in *E. coli* reported that 3,544 genes used the canonical start codon, 612 used *`GUG`*, and 130 used *`UUG`*.

   Design an HMM to model start codons in *E. coli*. Give the topology of your model and the initial, transition, and emission probabilities. Use a pseudocount of $b = 1$ to account for codon variations not observed in the data. You may assume that insertions and deletions never occur within start codons.

(b) Many genes in *E. coli* have a *Shine-Dalgarno (SD)* sequence upstream of the start codon. The SD sequence contributes to translation initiation by binding to a complementary sequence in 16S rRNA. The canonical SD motif is `GGAGG`, although the variants `GGAG` and `GAGG` are observed in 16% and 12% of SD sequences, respectively. The SD motif is separated from the start codon by a variable length spacer of nucleotides (`n`): `GGAGGnn...nnAUG`. In *E. coli*, this spacer ranges from 4 to 6 nucleotides in length.

Design an HMM to model the SD sequence in *E. coli*. Your model should meet the following requirements:

- Sequences `GGAGG`, `GGAG` and `GAGG` should be emitted in the appropriate frequencies.
- Only SD sequences sequences should be emitted; that is, the model should not emit `GAG`, `AGG`, etc.
- Spacers of 4, 5, or 6 nucleotides should be emitted with equal probability.
- Nucleotide frequencies in the spacer region should be $(0.25, 0.25, 0.25, 0.25)$.
- The topology should include silent Start and End states.
- The emission frequencies should not use pseudocounts.
- The model should use the minimum number of states necessary to meet these requirements.

Give the topology of your model and the initial, transition, and emission probabilities. Note that there are several models that satisfy the above requirements. Just show one.

3. *Multiple sequence alignment using HMMs:* Your goal is to obtain a global multiple alignment of the following unlabeled sequences

   `RDAHK, RAHK, RDTYK, FARDAEHK`

   You decide to align the sequences by constructing a Profile HMM with six match states, M_1 to M_6. States M_0 and M_7 are the silent *start* and *end* states, respectively, and do not emit symbols.

   (a) What is the rationale for choosing six match states?

   (b) How many insertion states will there be in this model, assuming you use the canonical Profile HMM architecture? How many deletion states?

   (c) What estimation method will you use to determine the initial, transition and emission probabilities of this model? What algorithms are used in this method?

   (d) Once you have instantiated your model, you need to label your sequences. What algorithm(s) could you use to assign a state to each amino acid in each protein sequence?

(e) Suppose the paths through the HMM corresponding to the four sequences are

$M_0\ M_1\ M_2\ M_3\ D_4\ M_5\ M_6\ M_7$
$M_0\ M_1\ D_2\ M_3\ D_4\ M_5\ M_6\ M_7$
$M_0\ M_1\ M_2\ M_3\ D_4\ M_5\ M_6\ M_7$
$M_0\ I_0\ I_0\ M_1\ M_2\ M_3\ M_4\ M_5\ M_6\ M_7$

Give the alignment of the sequences specified by this labeling.

(f) Based on the alignment, would you perform "model surgery" on this profile HMM? Why or why not? If so, which states would you add and/or delete?

(g) If you do decide to perform model surgery, you will need to update your parameter values. What estimation method should you use to determine the new initial, transition, and emission probabilities? Give the name of the method and any equations you might need.