Solution Set 7

Due 1pm, Wednesday, Dec. 13th

This homework assignment is **optional**. If you choose to hand in this homework for credit, your score on this assignment will replace your lowest homework score, if this score is higher. This is in addition to dropping your lowest score on PS 0 to PS 6. To obtain credit for this assignment, you must hand it in by Wednesday at 1pm, either by email to comp-bio@cs.cmu.edu or in MI 646. **This is a hard deadline.** Solutions will be posted on Tuesday, shortly after 1:00 pm.

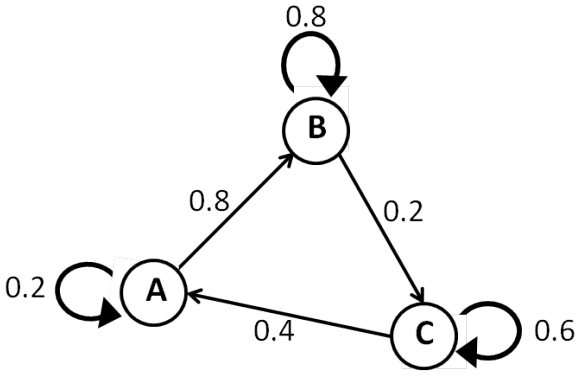
Collaboration is allowed on this homework. You must hand in homework assignments individually. List the names of the people you worked with:

Homework must be submitted by 1:00 pm in MI646 or electronically to comp-bio@cs.cmu.edu.

1. Define an HMM $\mathcal{H}$ with three states $\{A, B, C\}$ and alphabet $\{0, 1, 2\}$. The initial stable probabilities are $\pi_A = 1$ and $\pi_B = \pi_C = 0$. The transition and emission probabilities are as follows:

	A	B	C	0	1	2
A	0.2	0.8	0.0	0.8	0.2	0.0
B	0.0	0.8	0.2	0.0	0.6	0.4
C	0.4	0.0	0.6	0.2	0.0	0.8

   (a) Draw the state diagram of this HMM and show the transition probabilities.



(b) Give all state paths with non-zero probability for the sequence $O = 0, 1, 2, 0$.

$$AABC$$

$$ABBC$$

$$ABCC$$

$$ABCA$$

(c) What is $P(O)$? (Use the brute force approach, not the Forward algorithm.)

$$\begin{aligned}
P(AABC) &= 1.0 * .8 * .2 * .2 * .8 * .4 * .2 * .2 = 0.0004 \\
P(ABBC) &= 1.0 * .8 * .2 * .8 * .6 * .8 * .2 * .2 = 0.0025 \\
P(ABCC) &= 1.0 * .8 * .8 * .6 * .2 * .8 * .6 * .2 = 0.0074 \\
P(ABCA) &= 1.0 * .8 * .8 * .6 * .2 * .8 * .4 * .8 = 0.0197
\end{aligned}$$

$$\begin{aligned}
P(O) &= P(AABC) + P(ABBC) + P(ABCC) + P(ABCA) \\
&= 0.03
\end{aligned}$$

(d) What is the most probable path, Q^*? What is $P(O, Q^*)$, the probability of transitioning through this path and emitting O? (Again, you do not need to use the Viterbi algorithm here.)

$$P(ABCA) = .8 * .8 * .6 * .2 * .8 * .4 * .8 = 0.0197$$

(e) For a given a sequence, O, one might consider approximating $P(O)$, the probability that the model will emit O over all possible paths, by $P(O, Q^*)$, the probability that the model will emit O via the most likely path. For this particular HMM, would $P(O, Q^*)$ be a good approximation of $P(O)$? Explain your reasoning.

In this case, the probability of the most probable path is about two thirds of the total probability, which might be a reasonable approximation. In some cases, however, the $P(O|Q^*)$ is much lower than $P(O)$.

2. 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 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 unlabelled sequences, you find many sites that do not match either the activator or repressor binding sites. 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?

   The actual motifs of the activator and repressor exhibit mutually exclusive positional dependence. That is, in positions 3, 4 and 5, you only expect to see ATC for an activator or GCA, GCT, CCA, or CCT for a repressor, therefore with this pssm ACA could be scored very well, but would not correspond to either motif.

(b) In order avoid assigning high scores to non-binding motifs, you construct two separate PSSMs. Show the frequency matrices for these PSSMs. Why are two PSSMs required instead of just one?

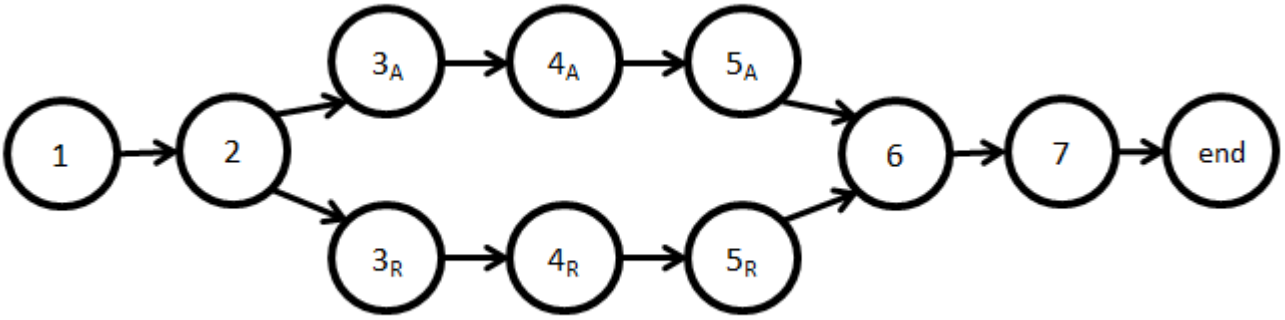
activator:

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

repressor:

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

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



	π	Transition prob											Emission prob			
		1	2	3_A	4_A	5_A	3_R	4_R	5_R	6	7	end	A	T	G	C
1	1	0	1	0	0	0	0	0	0	0	0	0	0.1	0.05	0.05	0.8
2	0	0	0	0.5	0	0	0.5	0	0	0	0	0	0.25	0.25	0.25	0.25
3_A	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
4_A	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
5_A	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
3_R	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0.5	0.5
4_R	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
5_R	0	0	0	0	0	0	0	0	0	1	0	0	0.5	0.5	0	0
6	0	0	0	0	0	0	0	0	0	0	1	0	0.5	0	0.5	0
7	0	0	0	0	0	0	0	0	0	0	0	1	0.8	0	0.1	0.1

(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 or a repressor binding site? Describe what output you would expect from this algorithm and how you would determine, given that output, whether the predicted motif is an activator or a repressor binding site.

Use Viterbi or Posterior Decoding to infer the states that emitted the sequence; that is, to label the data. The path chosen, either $3_A, 4_A, 5_A$ *or* $3_R, 4_R, 5_R$*, will indicate if it is an activator binding site or a repressor binding site.*

(e) A passive-aggressive co-worker comes to you with the sequences of several upstream regions of genes. He claims that there is a third transcription factor (TF3) that regulates this set of genes. He tells you the length of the TF3 binding site, but witholds any information about where it might bind. He 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.

Use Baum Welch to determine the parameters.

The input to the Baum Welch algorithm will be the unlabeled sequences provided by your co-worker. The output will be the emission and transition probabilities of the profile HMM.

(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 the 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.

Baum Welch estimates the parameters, but does not return labeled data. To find the location of the motif in each unlabeled sequence, use Viterbi or posterior decoding to infer the states that emitted the sequence; that is, to label the data.

The input to the Baum Welch algorithm will be the unlabeled sequences provided by your co-worker. The output will be labeled data in which a state is assigned to each symbol in each input sequence. The symbols that represent the motif are those symbols that are labeled with Match states.

(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. Your co-worker wants to know which of the three transcription factors regulates this gene.

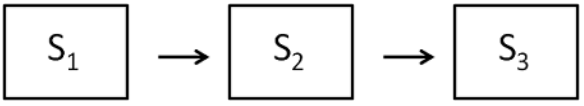
Using your two HMMs, you determine that the new sequence has only one high-probability motif. What algorithm would you use to determine whether this motif is more likely to be a TF3 binding site or a binding site for the activator/repressor pair described in (a)? Describe how you would apply this algorithm,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

Use the Forward or Backward algorithm to determine the probability of the new sequence for each of the two HMMs. Use a log odds ratio of the probability of the sequence being emitted from HMM1 and HMM3 to determine how much more likely it is to have been emitted by one of the models, compared with the other. If it is HMM1, then you need to use a strategy similar to that used in part **d** *to determine if it is the activator or repressor.*

3. *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*.

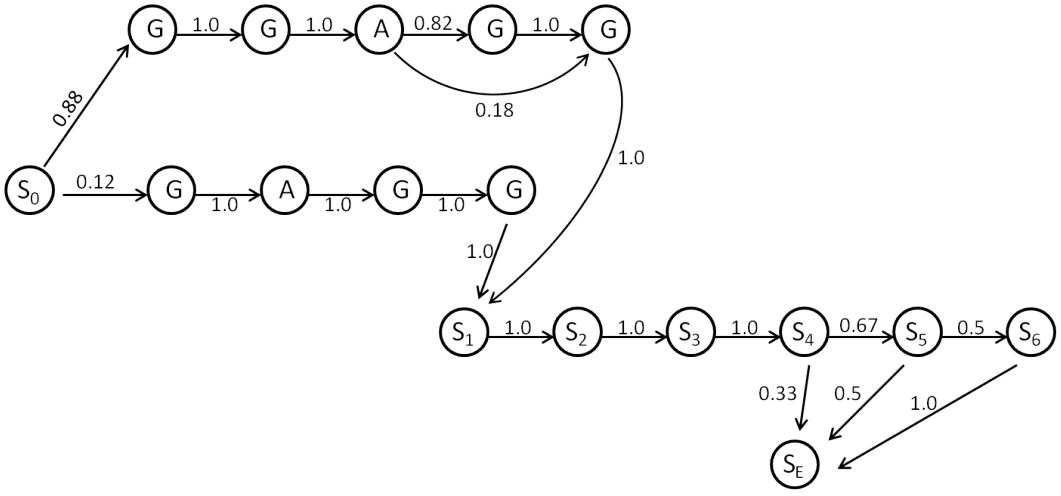
   Design an HMM to model start codons in *E. coli*. Give the topology of your model and the initial, transition, and emission probabilities. Use a pseudocount of $b = 1$ to account for codon variations not observed in the data. You may assume that insertions and deletions never occur within start codons.



$$
\begin{aligned}
\pi_{S_1} &= 1 \\
a_{S_1S_2} &= a_{S_2S_3} = 1 \\
e_{S_1}(A) &= \frac{3544+1}{4286+4} = 0.83 \\
e_{S_1}(G) &= \frac{612+1}{4286+4} = 0.14 \\
e_{S_1}(U) &= \frac{130+1}{4286+4} = 0.03 \\
e_{S_1}(C) &= \frac{0+1}{4286+4} = 2.3 \times 10^{-4} \\
e_{S_2}(U) &= 0.999 \\
e_{S_2}(\sigma) &= 2.3 \times 10^{-4},\ \sigma \in \{A, C, G\} \\
e_{S_3}(G) &= 0.999 \\
e_{S_3}(\sigma) &= 2.3 \times 10^{-4},\ \sigma \in \{A, U, G\}
\end{aligned}
$$

(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*. Give the topology of your model and the initial, transition, and emission probabilities. Your model should use the minimun number of states required to give rate appropriate sequences. Do not use pseudocounts. Your model should emit the sequences `GGAGG`, `GGAG` and `GAGG` in the appropriate frequencies. It should not emit `GAG`, `AGG`, etc. Your model should emit spacers of 4, 5, or 6 nucleotides, with equal probability. The frequencies of the nucleotides in the spacer region should be 0.25 for all four bases. Give your model silent Start and End states.



Initial probabilities:

$$\pi_i = \begin{cases} 1, & i = S_0 \\ 0, & otherwise. \end{cases}$$

Emission probabilities:

States labeled G (resp. A) emit G (resp. A) with probability 1.0. States S_0 and S_E are silent. For $1 \leq i \leq 6$,

$$\begin{aligned} e_{S_i}(A) &= 0.25 \\ e_{S_i}(G) &= 0.25 \\ e_{S_i}(U) &= 0.25 \\ e_{S_i}(C) &= 0.25. \end{aligned}$$

4. *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 *start* and *end* states, respectively, and do not emit symbols.

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

   The average sequence length is 5.5. Rounding up, that suggests a Profile HMM with 6 Match states (not including the silent start and end states.)
   A canonical Profile HMM with 6 Match states will have six Deletion states and seven Insertion states.

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

   Baum Welch, since the data are not labeled.

   (c) 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?

   The Viterbi Algorithm

(d) 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.

		M1	M2	M3	M4	M5	M6
-	-	R	D	A	-	H	K
-	-	R	-	T	-	H	K
-	-	R	D	A	-	Y	K
F	A	R	D	A	E	H	K

(e) 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?

It looks like column 4 is more likely to be an insertion than a match, since more than half of the sequences show deletions in column 4. This would mean we could remove states M_4, D_4 and I_4 from the model and connect the states in Column 3 directly to the states in Column 5 instead.

(f) 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.

In this case, model surgery has resulted in in a relatively small change to the topology of the Profile HMM; that is, only one column has been removed from the model. When the modification of the model is small, the parameters can be updated by relabeling the `E` *in Sequence 4 with state I_3 and removing D_4 from the state paths associated with Sequences 1 - 3.*
In general, when model surgery results in a change to a large fraction of the states in the model, you may obtain better results by retraining with the Baum Welch algorithm.