Problem Set 6

Due 11:59pm, Friday, November 5th

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. Position Specific Scoring Matrices

   In class, we learned how to construct a Position Specific Scoring Matrix (PSSM), representing a conserved pattern. Further, we saw that a PSSM can be used to score a subsequence, s, in an unlabeled sequence. This score can be viewed as a log likelihood ratio of the form

   $$\log_2 \frac{P(s|H_A)}{P(s|H_0)},$$

   where H_A is the hypothesis that s is an instance of the conserved pattern and H_0 is the hypothesis that s is a sample from the background distribution.

   We can use the same framework to model a pair of subsequences that represent the binding sites of two interacting proteins. These interactions must be specific; that is, each protein should bind its partner, but not other proteins with a similar fold. Interaction specificity is controlled by a small number of pairs of amino acids at specific sites, one in each protein. Substitutions at a given interacting site are constrained by the requirements of protein-protein interaction and will only be retained if a compensating substitution occurs at the corresponding site in the other protein. By comparing the multiple alignments of interacting pairs, it can be seen that these sites co-vary. This co-variance can be used to predict whether two newly discovered sequences will interact.

   The following alignments show the binding sites of four interacting pairs. Specificity-determining sites are annotated by a number at the top of the column. Columns labeled with the same number indicate interacting residue pairs.

```
                      12  3   4                      2  1  3  4
EnvK: RTLLMAGVSHDLRTPLTRIRLATEMMSEQ      EnvR: KILVVDDDMRLRALLERYLTEQGFQVR
CssK: RTLLQN-ISHDLKTPVTRIRLYTQSIKDG      CssR: TIYLVEDDDNLNELLTKYLENEGWNIT
LiaK: ERQRLARDLHDAVSQQLEAIVMMTSAVLE      LiaR: RVLLIDDHEMVRMVLAKFLEAQPDIEV
DegK: ERKRVSREIHDGPAQMLENVVMRSMLIER      DegR: NIVIIDDHQLFREVVKRILDFEPTFEV
RstK: KKQLIDGIAHELRTPLVRLRFRLEMSDN-      RstR: TIVFVEDDAEVGALIAAYLAKHDMQVT
-------------------------------------------------------------------------
              Protein 1                                Protein 2
```

(a) Construct a matrix giving the frequencies of amino acid pairs observed at the four pairs of sites (1, 2, 3, 4) associated with interaction specificity. Your frequency matrix should have four columns labeled 1, 2, 3, and 4 and one row for each pair of amino acids observed in the data. (A full frequency matrix would have 20^2 rows. To save space, only include rows for pairs of amino acids with non-zero frequencies.) Do not worry about pseudocounts.

(b) For each of the amino acid pairs in the matrix in part (a), calculate the probability of observing that amino acid pair by chance. Use the background amino acid frequencies given in the table below.

A	C	D	E	F	G	H	I	K	L
0.075	0.018	0.052	0.063	0.039	0.071	0.022	0.055	0.058	0.091

M	N	P	Q	R	S	T	V	W	Y
0.028	0.046	0.051	0.041	0.052	0.074	0.060	0.065	0.013	0.033

(c) Use the frequency and background matrices you constructed in (a) and (b) to build a propensity matrix representing the likelihood ratio for these co-varying sites.

(d) Construct the log odds scoring matrix for the co-varying sites.

(e) Use your log-odds matrix from (d) to calculate the score of the following pair of amino acid sequences, labeled with the specificity-determining sites.

```
                      12  3   4
YvrK: REEWIAGLSHDLKTPLTEIYVYSMMLESK
             2  1  3  4
YvrR: SILIVDDHKALVDVIKAVLEKEGYRNILDAASAEEAIPVV
```

(f) Calculate the score of the following pair of amino acid sequences, labeled with the specificity-determining sites.

```
                      12  3   4
WalK: RREFVANVSHELRTPLLEMRFYLESLAEG
             2  1  3  4
WalR: KILVVDDHKPFADLLEKNLRKEGYEVHCAHDGNEAVEMVE
```

(g) Which of the two protein pairs from parts **(e)** and **(f)** is more likely to be interacting pair, based on the results of your PSSM?

2. The Gibbs sampler:

   The following sequences contain a sequence motif that is the binding site of a small family of yeast transcription factors:

1:	G	A	G	A	T	A	G	G								
2:	G	T	T	G	A	T	A	A	A	C	A	C	A	T	A	
3:	A	A	G	C	G	G	A	T	T	A	T	T	C	C		
4:	T	T	A	C	G	A	T	A	G	C	C	G	A	T	G	T	A
5:	C	A	G	G	A	T	A	A	G	A	A	G	T	C	T		
6:	T	A	C	T	G	A	T	A	G	C	T	T	G	C			
7:	G	T	T	G	A	T	A	G	A	C	A	C	A	T	A		
8:	C	A	C	C	G	G	A	T	A	A	T	G	A	T			
9:	C	T	T	T	G	A	T	T	A	G	A	T	A	T	A	C	A
10:	G	C	T	G	T	C	A	C	G	A	T	T	G	C	C		

   Suppose you are using the Gibbs sampler to find a conserved pattern of length $w = 5$ in the $k = 10$ unlabeled sequences shown above. Suppose further that in the current iteration of the Gibbs sampler, $t^* = t_1$ and that the current estimate of the pattern is specified by the following offsets in the remaining nine sequences: $o_2 = 3$, $o_3 = 5$, $o_4 = 4$, $o_5 = 3$, $o_6 = 4$, $o_7 = 3$, $o_8 = 5$, $o_9 = 4$, $o_{10} = 8$. Remember, the offset o_i is the position *before* the first residue of the current estimate of the pattern in sequence t_i.

   (a) Show the ungapped local alignment specified by the starting positions $o_2 \ldots o_{10}$.

(b) Construct the frequency matrix with pseudocounts. Instead of using the pseudocount $b = 1$ that we used in class, use $b = \sqrt{k-1} \cdot p_i$, where k is the total number of sequences in the original input alignment (i.e., $k = 10$) and p_i is the background distribution. Assume $p_i = 0.25, i \in \{A, C, G, T\}$.

(c) Calculate the propensity matrix, $P[i, j]$, for the pattern using the same background frequencies, $p_i = 0.25, i \in \{A, C, G, T\}$.

(d) Calculate the probability $pdf[o]$ of each candidate offset, o, in t^*.

(e) What is the *second* most probable subsequence of t^*? What is the probability that this sequence will be selected as the new entry in A'?

(f) What are the values of the cumulative distribution function, $cdf[o]$, for each candidate offset, o, in t^*?

(g) Suppose that you generate a random number, r, between 0 and 1, assuming a uniform distribution and use the *cdf* to select o^*. If $r = 0.57$, what is the corresponding five letter subsequence in t_1?

(h) The *consensus sequence* of a multiple alignment is a sequence in which the i^{th} residue is the most common residue in the i^{th} column of the alignment. If there is a tie, then all most common residues are represented; e.g., `X/Y`. What is the consensus sequence for this binding site?

(i) Bonus question: In yeast, transcription factors that bind this site control genes that are involved in a particular metabolic process. What, generally speaking, is that process? (A short phrase is sufficient.)

3. Define an HMM $\mathcal{H}$ with three states $\{E_1, E_2, E_3\}$ and alphabet $\{a, b, c\}$. The transition, initial state, and emission probabilities are as follows:

	E_1	E_2	E_3
E_1	0.2	0.8	0.0
E_2	0.0	0.8	0.2
E_3	0.4	0.0	0.6
π_i	1	0	0
$e_i(a)$	0.8	0	0.2
$e_i(b)$	0.2	0.6	0
$e_i(c)$	0	0.4	0.8

This problem asks you to calculate the total probability of a sequence and the most probable path, given that sequence. Normally, we calculate these quantities using dynamic programming algorithms. With this very simple HMM, it is possible to calculate these quantities by enumerating all paths with non-zero probability, which is what we ask you to do here.

(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 $O = a, b, c, a$.

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

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

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