Problem Set 4

Due Wednesday, Dec 2nd, 2015

Collaboration is allowed on this homework. You must hand in homework assignments individually and list the names of the people you worked with. To get full credit, show your work.

Homework must be submitted by 4:30pm in MI646 or electronically to comp-bio@cs.cmu.edu.

1. **BLOSUM matrices 1**: In the BLOSUM framework, clustering is used to obtain a family of matrices corresponding to different levels of evolutionary divergence. In this problem, you will cluster the following aligned block at different thresholds:

   1: DSDQQD
   2: DSSQQD
   3: SSQQDD
   4: DDQQDD

   (a) Determine the percent identity between all possible pairs of sequences.

   (b) Cluster the sequences such that each sequence is at least 45% identical to at least one other sequence in the same cluster. In other words, every pair of sequences comprising members of two different clusters has less than 45% identity. How many clusters are there? Show the set of sequences in each cluster.
(c) Cluster the sequences such that each sequence in the cluster is at least 60% identical to one or more other sequences in the cluster. How many clusters are there? Show the set of sequences in each cluster.

(d) Cluster the sequences such that each sequence in the cluster is at least 80% identical to one or more other sequence in the cluster. How many clusters are there? Show the set of sequences in each cluster.

(e) Cluster the sequences such that each sequence in the cluster is at least 90% identical to at least one other sequence in the cluster. How many clusters are there? Show the set of sequences in each cluster.
(f) Cluster the sequences such that each sequence in the cluster is at least 95% identical to at least one other sequence in the cluster. How many clusters are there? Show the set of sequences in each cluster.

(g) If you had funding to construct three BLOSUM matrices based on the multiple alignment given above, which clustering thresholds would you pick? Why?
2. **BLOSUM matrices 2:** In this question, you will build BLOSUM substitution matrices for the three letter alphabet \( \Sigma = \{D, S, Q\} \).

   (a) Based on the 60% clusters that you constructed in Problem 1, calculate the observed substitution frequencies for \( DS \) and \( DQ \) (i.e., \( A_{DS}^{60} \) and \( A_{DQ}^{60} \)), using the BLOSUM method for counting pairs in clustered sequences.

   (b) Based on the 60% clusters that you constructed in Problem 1, calculate the expected amino acid frequencies, \( p_D^{60} \), \( p_S^{60} \), and \( p_Q^{60} \), using the BLOSUM method.
(c) Calculate the *expected* amino acid pair frequencies for DS and DQ (e.g., $E_{DS}^{60}$ and $E_{DQ}^{60}$), using the BLOSUM method.

(d) Use these frequencies to obtain the log odds matrix entries for DS and DQ (e.g., $S_{DS}^{60}$ and $S_{DQ}^{60}$), as defined in the BLOSUM framework.
(e) Based on the 80% clusters that you constructed in Problem 1, calculate the observed substitution frequencies for DS and DQ (i.e., \( A_{DS}^{80} \) and \( A_{DQ}^{80} \)), using the BLOSUM method for counting pairs in clustered sequences.

(f) Based on the 80% clusters that you constructed in Problem 1, calculate the expected amino acid frequencies, \( p_D^{80} \), \( p_Q^{80} \), and \( p_S^{80} \), using the BLOSUM method.
(g) Calculate the expected amino acid pair frequencies for DS and DQ (e.g., \(F^{80}_{DS}\) and \(F^{80}_{DQ}\)), using the BLOSUM method.

(h) Use these frequencies to obtain the log odds matrix entries for DS and DQ (e.g., \(S^{80}[DS]\) and \(S^{80}[DQ]\)), as defined in the BLOSUM framework.
(i) Compare your results for the 60% and 80% thresholds.

- For this data set, does $S[D, S]$ increase or decrease as the threshold increases?

- Does $S[D, Q]$ increase or decrease as the threshold increases?

- Based on your results, are you more likely to observe aspartic acid aligned with serine in conserved or diverged sequence families?

- Are you more likely to observe aspartic acid aligned with glutamine in conserved or diverged sequence families?
3. BLAST statistics: For ungapped alignments, the expected number of high scoring pairs (HSP’s) with score at least $S$ found in the alignment of two random sequences is

$$E = Kmne^{-\lambda S},$$

where $m$ and $n$ are the effective lengths of the sequences and $K$ and $\lambda$ are constants that can be derived from the theory and depend on the substitution matrix. We can define a “normalized” score

$$S' = \frac{\lambda S - \ln K}{\ln 2}. \quad (2)$$

Show algebraically that the number of HSP’s with score at least $S'$ is

$$E = mn2^{-S'} \quad (3)$$
4. *The Gibbs sampler:* Some researchers have discovered a new transcription factor. They have identified a number of possible binding sites but want to narrow down a specific binding sequence. You decide to use a Gibbs sampler to discover the most likely pattern.

```
1: T T A A A C A G G C A A T T C A G G
2: G T T A A G C A A C A C A T A
3: C A C C G A G G C A T G A T
4: C T T T G G G C A G A T A T A C A
5: G C T T G G C A G T C A C C C
```

Suppose you are using the Gibbs sampler to find a conserved pattern of length \( w = 5 \) in these sequences and that in the current iteration of the Gibbs sampler, \( t^* = t_1 \). Using the four following steps, construct the PSSM for the pattern specified by the following offsets: \( o_2 = 3 \), \( o_3 = 5 \), \( o_4 = 4 \), \( o_5 = 3 \).

(a) Show the ungapped local alignment specified by the offsets \( o_2 \ldots o_5 \). Don’t forget: the first symbol in the motif starts at position \( o + 1 \).
(b) Construct the frequency matrix using $b = 1$ as a pseudocount.

(c) Calculate the propensity matrix, $P[i, j]$, for the pattern using uniform background frequencies, $p_i = 0.25, i \in \{A, C, G, T\}$. 
(d) What is 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 is the cumulative distribution function, $cdf$? Plot the $cdf$. Your vertical axis should range from 0 to 1.

(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 $r = 0.35$, what is the corresponding five letter subsequence in $t_1$?