Seven11 Assignment 3

Due 11:59pm, Tuesday, November 9th

Your name:

This assignment is required for students taking the 12-unit course only.

In this assignment, you will build BLOSUM substitution matrices for the three letter alphabet $\Sigma = \{D, S, Q\}$, based on the aligned block from Problem Set 5, problem 3:

```
1: DSDQQD
2: DSSQQD
3: SSQQDD
4: DDQQDD
5: SDDDQS
6: SDSDQS
```

The procedure for constructing a BLOSUM matrix from a multiple alignment block is described in Section 3.3 of the class notes. You may also find it helpful to look at Ewens and Grant, Section 6.5.2: `http://www.cs.cmu.edu/~durand/03-711/Readings/eserdurand34.pdf`

Homework must be submitted by 11:59pm electronically to Canvas.

Collaboration is allowed on this homework. You must hand in homework assignments individually. List the names of the people you worked with, and cite all additional references you used to complete this assignment:

1. BLOSUM-45

   (a) Calculate the *observed* substitution frequencies for `DS` and `DQ` (i.e., $q^{45}_{\texttt{DS}}$ and $q^{45}_{\texttt{DQ}}$), using the BLOSUM method for counting pairs in clustered sequences and the 45% threshold clusters

1:	`DSDQQD`
2:	`DSSQQD`
3:	`SSQQDD`
4:	`DDQQDD`
5:	`SDDQQS`
6:	`SDSQDS`

   (b) Based on the 45% clusters, calculate the *expected* amino acid frequencies, p^{45}_D, p^{45}_S, and p^{45}_Q, using the BLOSUM method.

(c) Calculate the *expected* amino acid pair frequencies for `DQ` and `DS` (e.g., $E^{45}_{\texttt{DQ}}$ and $E^{45}_{\texttt{DS}}$), using the BLOSUM method.

(d) Use these frequencies to obtain the log odds matrix entries for `DQ` and `DS` (e.g., $S^{45}[\texttt{D,Q}]$ and $S^{45}[\texttt{D,S}]$), as defined in the BLOSUM framework.

2. BLOSUM-60

   (a) Calculate the *observed* substitution frequencies for `DS` and `DQ` (i.e., $q^{60}_{\texttt{DS}}$ and $q^{60}_{\texttt{DQ}}$), using the BLOSUM method for counting pairs in clustered sequences and the 60% threshold clusters

1:	`DSDQQD`
2:	`DSSQQD`
3:	`SSQQDD`
4:	`DDQQDD`
5:	`SDDQQS`
6:	`SDSQDS`

   (b) Based on the 60% clusters, calculate the *expected* amino acid frequencies, p^{60}_D, p^{60}_S, and p^{60}_Q, using the BLOSUM method.

(c) Calculate the *expected* amino acid pair frequencies for `DQ` and `DS` (e.g., $E^{60}_{\texttt{DQ}}$ and $E^{60}_{\texttt{DS}}$), using the BLOSUM method.

(d) Use these frequencies to obtain the log odds matrix entries for `DQ` and `DS` (e.g., $S^{60}[\texttt{D,Q}]$ and $S^{60}[\texttt{D,S}]$), as defined in the BLOSUM framework.

3. Compare your results for the 45% and 60% thresholds. For this data set, does $S[D, S]$ increase or decrease as the threshold increases? Does $S[D, L]$ increase or decrease as the threshold increases? How would you explain the trends you observe in terms of the processes of sequence evolution?