Problem Set 5

Due 11:59pm, Tuesday, October 26th

Collaboration is allowed on this homework. You may discuss the problems with your colleagues, but each student must prepare and submit a separate assignment. Please list the names of the people you worked with:

Please provide answers to all questions. In order to obtain full credit, explain your reasoning on each question and show all intermediate steps leading to your solution.

You may submit your assignments in any of the following ways:

- Download the assignment in pdf format and print it out. Write your solutions in the space provided. Scan your handwritten solution and upload it to Canvas. Attach additional pages, if needed.
- Download the assignment in pdf format. Use the commenting features in adobe or a similar tool to enter your solution directly on the pdf. Upload the completed assignment, annotated with your solutions.
- Download the assignment in latex format. Enter your solutions in latex format, compile the assignment, and turn in the resulting pdf.

1. PAM theory

   (a) Is the PAM-1 transition matrix symmetric? Justify your answer algebraically.

   (b) Is the PAM-1 log odds scoring matrix symmetric? Justify your answer algebraically.

(c) Verify that the rows of the PAM-1 transition matrix sum to one.

(d) Given a pair of sequences with one PAM of divergence, 99 out of 100 positions should be identical. Verify that $\sum_x p_x P^1[x, x] = 0.99$

2. Aspartic acid (D) and glutamic acid (E) are both acidic amino acids. They have a second carboxylic acid group and are polar and negatively charged under physiological conditions. The two acids are very similar; while aspartic acid has one methylene group ($-CH_2-$), glutamic acid has two. Their amides, asparagine (N) and glutamine (Q) are polar and uncharged. The side chains of the amides N and Q are the same as their respective acids D and E, except the second carboxylic acid group ($-COOH$) in the acid is a carboxamide group ($-CONH_2$) in the respective amide.

Acid	Amide
D	N
E	Q

For these four amino acids and the entries in the PAM30 matrix (available on the class website), which of the following amino acid alignments are you more likely to observe in closely-related sequences?

- An amino acid with the same charge.
- An acidic amino acid with its amide.

Show the evidence on which you base your answer.

3. 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
5:  SDDDQS
6:  SDSDQS
```

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

   (b) Cluster the sequences such that each sequence is at least 30% identical to one or more other sequences in the same cluster. In other words, every pair of sequences comprising members of two different clusters has less than 30% 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 45% 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 60% identical to one or more other sequences 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 75% identical to one or more 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 85% identical to at least one other sequence in the cluster. How many clusters are there? Show the set of sequences in each cluster.

(g) 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.

(h) If you had funding to construct four BLOSUM matrices based on the multiple alignment given above, which clustering thresholds would you pick? Why?

4. Both the PAM and the BLOSUM substitution matrix families are parametrized by evolutionary divergence.

   (a) Which represents a greater degree of divergence, BLOSUM90 or BLOSUM62? Explain your answer in terms of the process whereby BLOSUM matrices are constructed.

   (b) Which represents a greater degree of divergence, PAM120 or PAM70? Explain your answer in terms of the process whereby PAM matrices are constructed.

   (c) Which represents a greater degree of divergence, BLOSUM62 or PAM60? How do you know?