Solution 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. (*32 pts.*) PAM theory

   (a) (*8 pts.*) Is the PAM-1 transition matrix symmetric? Justify your answer algebraically.

   Suppose the PAM transition matrix is symmetric. Then, $P^1_{xy} = P^1_{yx}$*, which expands to*

$$\frac{0.01}{p_x} \frac{A_{xy}}{\sum_h \sum_{l \neq h} A_{hl}} = \frac{0.01}{p_y} \frac{A_{yx}}{\sum_h \sum_{l \neq h} A_{hl}}.$$

   Since the procedure for counting pairs in the PAM framework ensures that $A_{xy} = A_{yx}$*, the second term on the left hand side is equal to the second term on the right hand side, yielding*

$$\frac{0.01}{p_x} = \frac{0.01}{p_y}$$

   It is not generally true that $p_x = p_y$*, so the proposition is false: The PAM transition matrix is not symmetric.*

(b) (*8 pts.*) Is the PAM-1 log odds scoring matrix symmetric? Justify your answer algebraically.

$$S^1[x,y] \;=\; \lambda \log \frac{q_{x,y}}{p_x p_y} = \lambda \log \frac{p_x P^1[x,y]}{p_x p_y} = \lambda \log \frac{P^1[x,y]}{p_y}$$

and

$$S^1[y,x] = \lambda \log \frac{q_{y,x}}{p_x p_y} = \lambda \log \frac{p_y P^1[y,x]}{p_x p_y} = \lambda \log \frac{P^1[y,x]}{p_x}$$

So,

$$S^1[x,y] = S^1[y,x]$$

if and only if

$$\frac{P^1[y,x]}{p_x} = \frac{P^1[x,y]}{p_y}.$$

Note that

$$\begin{aligned}
\frac{P^1[x,y]}{p_y} &= \frac{0.01}{p_x} \frac{A_{xy}}{\sum_h \sum_{l \neq h} A_{hl}} \frac{1}{p_y} \\
&= \frac{1}{100 p_x p_y} \frac{A_{xy}}{\sum_h \sum_{l \neq h} A_{hl}} \\
&= \frac{1}{100 p_y p_x} \frac{A_{yx}}{\sum_h \sum_{l \neq h} A_{hl}} \\
&= \frac{0.01}{p_y} \frac{A_{yx}}{\sum_h \sum_{l \neq h} A_{hl}} \frac{1}{p_x} \\
&= \frac{P^1[y,x]}{p_x}
\end{aligned}$$

The PAM-1 log odds scoring matrix is symmetric.

(c) (*8 pts.*) Verify that the rows of the PAM-1 transition matrix sum to one.

We know

$$\begin{aligned} P^1[x,y] &= m_x \frac{A_{xy}}{\sum_{l \neq x} A_{xl}}, x \neq y \\ P^1[x,x] &= 1 - m_x \end{aligned}$$

Summing a row in the PAM 1 matrix we get

$$\begin{aligned} \sum_y P^1[x,y] &= P^1[x,x] + \sum_{y \neq x} P^1[x,y] \\ &= 1 - m_x + \sum_{y \neq x} m_x \frac{A_{xy}}{\sum_{l \neq x} A_{xl}} \\ &= 1 - m_x + m_x \frac{\sum_{y \neq x} A_{xy}}{\sum_{l \neq x} A_{xl}} \\ &= 1 - m_x + m_x \cdot 1 = 1 \end{aligned}$$

(d) (*8 pts.*) Given a pair of sequences with one PAM of divergence, 99 out of 100 positions should be identical. Verify that $\sum_x p_x P^1[x,x] = 0.99$

We know that $P^1[x,x] = 1 - m_x$.

$$\begin{aligned} \sum_x p_x P^1[x,x] &= \sum_x p_x (1 - m_x) \\ &= \sum_x p_x - \sum_x p_x m_x \\ &= 1 - \sum_x p_x m_x \\ &= 1 - \sum_x p_x \frac{0.01}{p_x} \frac{\sum_{y \neq x} A_{xy}}{\sum_h \sum_{l \neq h} A_{hl}} \\ &= 1 - 0.01 \left(\frac{\sum_x \sum_{y \neq x} A_{xy}}{\sum_h \sum_{l \neq h} A_{hl}} \right) \\ &= 1 - 0.01 = 0.99 \end{aligned}$$

2. (*8 pts.*) 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.

The average of the S^{30} scores for aligning D with E and N with Q is -0.5. This suggests that the observation of an acidic amino acid aligned with an acidic amino acid and an amide amino acid aligned with an amide amino acid is slightly less likely in related sequences with 30 PAMs divergence than expected by chance.

The average of the S^{30} scores for aligning D with N and E with Q is 1.5. In other words, the alignment of an acid with its amide (or vice versa) occurs more frequently than chance in related sequences with 30 PAMs divergence.

In closely related sequences, for these four amino acids, you are more likely to observe the alignment of an acid with its amide than an alignment of two amino acids with the same charge.

3. (*36 pts.*) 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) (*7 pts.*) Determine the percent identity between all possible pairs of sequences.

	[2]	[3]	[4]	[5]	[6]
[1]	0.83	0.5	0.5	0.33	0.17
[2]		0.5	0.5	0.17	0.33
[3]			0.67	0.17	0.17
[4]				0.17	0.17
[5]					0.83

(b) (*6 pts.*) 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.

There is one cluster that comprises all of the sequences in the data set.

(c) (*3 pts.*) 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.

There are two clusters

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

(d) (*3 pts.*) 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.

There are three clusters

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

(e) (*3 pts.*) 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.

There are four clusters

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

(f) (*3 pts.*) 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.

There are six clusters

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

(g) (*3 pts.*) 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.

There are six clusters

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

(h) (*8 pts.*) If you had funding to construct four BLOSUM matrices based on the multiple alignment given above, which clustering thresholds would you pick? Why?

BLOSUM45, BLOSUM60, BLOSUM75, and BLOSUM85.

The 30% threshold cannot be used because there is only one cluster at this threshold. At least two clusters are required to build a matrix.

The same clustering was obtained using a 95% threshold and an 85% threshold. Therefore, BLOSUM95 and BLOSUM85 matrices based on this block would be identical. Amino acid pairs obtained from these clusters are more representative of pairs observed in sequences that are 85% identical than of pairs observed in sequences that are 95% identical.

4. (*24 pts.*) Both the PAM and the BLOSUM substitution matrix families are parametrized by evolutionary divergence.

   (a) (*8 pts.*) Which represents a greater degree of divergence, BLOSUM90 or BLOSUM62? Explain your answer in terms of the process whereby BLOSUM matrices are constructed.

   The BLOSUM x matrices are constructed by clustering sequences such that each sequence in a cluster is at least x% identical to some other sequence in the cluster. When estimating amino acid substitution frequencies, only sequences from different clusters are compared. This means that the BLOSUM x matrix is derived from sequences that are as much as $(x-1)\%$ *identical. Therefore, smaller values of x correspond to more sequence divergence. In particular, BLOSUM62 represents greater divergence than BLOSUM90.*

   (b) (*8 pts.*) Which represents a greater degree of divergence, PAM120 or PAM70? Explain your answer in terms of the process whereby PAM matrices are constructed.

   The PAMN matrices are constructed to represent sequences that are N-PAMs apart; i.e., there are N substitutions per 100 residues. Therefore, larger values of N correspond to more sequence divergence. In particular,PAM120 represents greater divergence than PAM70.

   (c) (*8 pts.*) Which represents a greater degree of divergence, BLOSUM62 or PAM60? How do you know?

   BLOSUM62 corresponds to roughly 30% sequence identity, whereas PAM60 represents more than 60% identity, so BLOSUM62 represents greater divergence.