Problem Set 6

Due 3:00pm, Friday, November 2nd

Your name:

Collaboration is allowed on this homework. You must hand in homework assignments individually. List the names of the people you worked with:

*Homework must be submitted by 3pm in MI646.*
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) Given a pair of sequences with one PAM of divergence, 99 out of 100 positions should be identical. Verify that \( \sum x p(x)P^1[x, x] = 0.99 \).

(d) Verify that the rows of the PAM-1 transition matrix sum to one.
2. Substitution matrices and evolutionary divergence

(a) Consider the PAM30 and PAM250 matrices (shown on the web site). What is the average value on the diagonal of the PAM 30 matrix (i.e., the average of $S^{30}[x, x]$ over all values of $x$)?

(b) What is the average value on the diagonal of the PAM 250 matrix?

(c) Which average diagonal value is larger? How would you explain this in terms of the evolutionary divergence associated with each of the matrices?
(d) Which specific diagonal values are greater in PAM250 than in PAM30? That is, for which amino acids, $x$, is $S^{250}_{xx} > S^{30}_{xx}$? What does that suggest about the functional or structural properties of these amino acids?
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: AAAL
2: AAAA
3: AVVA
4: AVVL
5: VLLV
6: ALLV

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

(b) Cluster the sequences such that each sequence in a cluster is at least 45% identical to at least one other sequence in that cluster. 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 65% identical to at least one other sequence 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 85% identical to at least one 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 95% identical to at least one other sequence in the cluster. How many clusters are there? Show the set of sequences in each cluster.

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