1. Given a multiple sequence alignment (MSA) \( V \) of set \( S = \{ s_1, s_2, \ldots, s_n \} \) of sequences, the induced alignment of two sequences \( s_i \) and \( s_j \) in \( V \) is the two sequences corresponding to \( s_i \) and \( s_j \) in \( S \) (with any columns where both sequences have gap symbols being removed). For example, the induced alignment of the second and third sequences in the MSA given in Problem 1 is

\[
\begin{array}{cccc}
G & - & G \\
G & A & G \\
\end{array}
\]

Let \( A \) be an MSA of set \( X \) of sequences, and \( B \) be an MSA of set \( Y \) of sequences. We say that an MSA \( W \) of \( X \cup Y \) is consistent with both \( A \) and \( B \) if for every two sequences \( s_1, s_2 \in X \cup Y \):

- if \( s_1, s_2 \in X \), then the induced alignment of \( s_1, s_2 \) in \( W \) is identical to their induced alignment in \( A \), and
- if \( s_1, s_2 \in Y \), then the induced alignment of \( s_1, s_2 \) in \( W \) is identical to their induced alignment in \( B \).

Now, consider two sets of sequences \( X \) and \( Y \) and their MSA’s, \( A \) and \( B \), respectively.

(a) Describe how to compute an MSA \( W \) for \( X \cup Y \) that is consistent with both \( A \) and \( B \) if \( |X \cap Y| = 0 \).

(b) Describe how to compute an MSA \( W \) for \( X \cup Y \) that is consistent with both \( A \) and \( B \) if \( |X \cap Y| = 1 \).

(c) Discuss what challenges would you face an approach for computing an MSA \( W \) for \( X \cup Y \) that is consistent with both \( A \) and \( B \) if \( |X \cap Y| > 1 \).

2. An \( n \times n \) distance matrix \( M \) satisfies the four-point condition if for every \( i, j, k, l \in \{1, 2, \ldots, n\} \),

\[
M[i, j] + M[k, l] \leq \max\{M[i, k] + M[j, l], M[i, l] + M[k, j]\}.
\]

Prove that if \( M \) is additive, then \( M \) satisfies the four-point condition.

3. If two sequences differ by 65% of their positions and have evolved by a Jukes-Cantor model, what is the best estimate of the branch length between them?

4. Assume the Jukes-Cantor model. Let \( S_1 \) and \( S_2 \) be two sequences that differ at 10% of their sites.
(a) What is the branch length between $S_1$ and $S_2$?
(b) Assume $S_2$ evolves into $S_3$ by changing a completely different 10% of its sites. What is the branch length between $S_2$ and $S_3$?
(c) Compare the total branch lengths you obtained in (a) and (b) to the branch length you get when comparing $S_1$ and $S_3$ assuming they differ at 20% of their sites. Why the discrepancy?

5. What problem arises when reconstructing the evolutionary history of the following set of sequences using a distance-based method and assuming the Jukes-Cantor model of evolution? Suggest a reasonable way to fix the problem?

   $S_1$ AACCACACGTGTACAAACGT
   $S_2$ AACACACGTGTACGGACTT
   $S_3$ GTTTTACATAATGAAAGTCC
   $S_4$ GTTTCACCGTGTGAAAGTCC

6. The infinite sites model states that each site in a set of sequences mutates at most once during the evolutionary history of the set of sequences from their most recent common ancestor. The infinite alleles model states that whenever a site mutates, it changes to a new state not seen anywhere else (for that site) in the evolutionary tree. Are there any differences between the infinite sites and infinite alleles models? Given a tree leaf-labeled by a set of sequences, propose a method for testing whether the sequences evolved under the infinite alleles model, assuming you can use Fitch’s algorithm as a “black box”.

7. Informative sites are defined as columns in an alignment that favor one tree topology over another. Find the informative sites in the following alignment

   ATGTA
   TAGTA
   CGCTG
   GCCTG

Which tree will you find by the maximum parsimony method for this alignment?

8. Consider the following alignment:

   N I P E L M K T A N A D N G R E I A K K
   N I Q E L M K T A N A N H G R E I A K K
   P I E K L L Q T A S V E K G A A A A A K K
   P I A K L L A S A D A A A K G E A V F K K

(a) Construct a UPGMA tree from the Poisson corrected distances (under Poisson correction, $d_{ij} = -\ln(1 - p_{ij})$).
(b) Show that the distances defined by the UPGMA tree do not always coincide with the initial distances (the ones used to build the tree).
(c) Construct an unrooted NJ tree from the Poisson corrected distances.
(d) Show that the molecular clock property fails for any rooted tree derived from the tree you constructed in (c) by adding a root.