COSC494/594: Bioinformatics Computing
Homework #5 — Phylogenetic Tree Construction and Comparison
Due: November 14
Total Points: 30

Purpose

This assignment integrates algorithmic reasoning and bioinformatics tool use. You will move from implementing pairwise sequence comparison (Hamming distance) to building and interpreting phylogenetic trees with both manual and software-based methods.

Learning Outcomes

By the end of this assignment, you will be able to:

- Implement a simple sequence distance metric (Hamming distance).
- Construct and interpret distance matrices using UPGMA trees.
- Use PHYLIP for distance-based phylogenetic inference.
- Compare multiple sequence alignment (MSA) and tree construction methods.

What to Submit

Please submit all materials in a .tar/.zip folder named `hw5_<lastname>` via Canvas upload. Include:

```
hw5_<lastname>/
  ├── hamming_distance.py (or .cpp/.java)
  ├── upgma_tree.pdf (or image)
  ├── kitsch_tree.txt
  ├── neighbor_tree.txt
  ├── clustal_alignment.aln
  ├── other_alignment.aln (e.g., MUSCLE or MAFFT)
  ├── report.txt (summary and comments)
```

Assignment Tasks

1. Download datasets (0 points)

- Retrieve the small and large FASTA datasets from the course website.

2. Hamming Distance Program *(5 points)*

- Modify your global alignment code to compute the **Hamming distance** (mismatches only, ignoring indels) between all pairs of sequences in the **small dataset**.
- Output the resulting matrix to the console.
- Include your code, dataset, and instructions to compile and run.

3. Manual UPGMA Tree ***(8 points)***

- Using the small dataset, construct a distance-based tree **by hand** using the **UPGMA algorithm**.
- Submit as a scanned image, PDF, or clear photo labeled with sequence names.

4. PHYLIP Tree Construction ***(5 points)***

- Download and install the PHYLIP package.
- For the **large dataset**, compute a distance matrix using `DNADIST`.
- Generate trees using both `KITSCH` and `NEIGHBOR`.
- Save raw tree files and briefly **comment on differences**, if any.

5. Multiple Sequence Alignments (MSA) ***(12 points total)***

- Run **Clustal Omega** and **one additional tool** (e.g., MUSCLE or MAFFT) using EBI's MSA web portal.
- Save each alignment as plain text (`.aln`) and include in your submission folder.
- Write a short **report (report.txt)** addressing:
  - Are the resulting trees **similar** to those from KITSCH and NEIGHBOR? *(3 pts)*
  - Are the two MSA-based trees **similar to each other**? *(3 pts)*
  - Briefly describe any major differences or potential biological explanations. *(6 pts)*

Report and Rubric

Section	Description	Points
2	Hamming distance program & output	5
3	Manual UPGMA tree	8
4	PHYLIP trees (KITSCH & NEIGHBOR)	5
5	MSA analyses & interpretation	12
Total		**30**

Submission Notes

- All files must be organized and labeled clearly.
- Code should compile and run with standard tools (gcc, python3, javac, etc.).
- Include your name and email at the top of each file.
- Use consistent sequence names across all files for cross-reference.

Resources

- PHYLIP Documentation: https://evolution.genetics.washington.edu/phylip/doc/main.html
- EBI Clustal Omega: https://www.ebi.ac.uk/Tools/msa/clustalo/
- MUSCLE: https://www.ebi.ac.uk/Tools/msa/muscle/
- MAFFT: https://mafft.cbrc.jp/alignment/software/