Bilkent University
CS481 | CS583 - Bioinformatics Algorithms

Spring 2025
Deadline: **May 1, 2025**

Homework Assignment #4

GENERAL INSTRUCTIONS

FAILURE TO FULFILL ANY OF THE FOLLOWING ITEMS WILL RESULT IN A GRADE SCORE OF 0 (zero) WITHOUT ANY CHANCE OF REDEMPTION.

- You must **write your code yourself**. Sufficient evidence of plagiarism will be treated the same as for plagiarism or cheating.
- You **cannot** use any libraries (e.g., Biopython) that provide the algorithms that are required in this assignment. You should implement the algorithms yourselves.
- **C / C++, Python or Java** will be used as programming language. STL is allowed. The assignments are created with C / C++ in mind, so using other languages would require minimum tweaking.
- Your code must **compile**.
- Your code must **be complete**.
- Your code must **run on dijkstra.cs.bilkent.edu.tr** server.
- Your code must make use of **argument parsing**.
  Refer to: https://docs.google.com/presentation/d/1rU0DhBg6yVXbfEtNuK73pjc1yWSPG9OYnGNH-b-kklQ
- Submit your answers **ONLY** through the Moodle page.
- **Zip** your files and send them in only one zipped file.
- File name format `surname_name_hw#.zip`.
- The zip file must contain the following items:
  - All the source files.
  - `Makefile` to compile the source code and produce the binary. Even if you use Python, include this Makefile.
  - A `README.txt` file that briefly describes how your program works.
- All submissions must be made **strictly before the stipulated deadline**.
- A **bonus** will be given for the fastest code that solves the assignment successfully.

1) HELPFUL RESOURCES

You can refer to the following sources for help:

- Course Slides - 10
- https://en.wikipedia.org/wiki/Newick_format
- You can use iTOL online tool to visualize your output UPGMA trees.

2) UPGMA

Aim: In this assignment, you will implement UPGMA to construct a phylogenetic relationship. Using the Needleman-Wunsch algorithm, you will build a distance matrix for the DNA sequences of different species in a FASTA file. You can use your code from HW2.

2.1 Data

- **Input**
  1. A FASTA file containing the sequences. The file is given using the `-i` switch.
  2. The scoring function for the match, mismatch, and gap penalties. The scores are given in this order using the `-s` switch.
- **Output**
  1. UPGMA Tree (.txt): A text file containing the resulting UPGMA tree in Newick file format. Please see Helpful Resources for the format specifications. The switch `-t` specifies the file name.

2.2 Tasks

2.2.1 UPGMA

1. Perform pairwise alignment with NW on each pair given in the file with the specified scoring function.
2. Using the following formula, calculate the distance between each aligned pair:

   distance = number of mismatches + number of gaps
3. Construct the distance matrix.
4. Implement UPGMA using the distance matrix and output the UPGMA tree in Newick format.

See below for an example output format.

3) EXAMPLE

These examples show how the command line arguments and the outputs of your program should be:

```
user@dijkstra$ ./hw4 -i sequences.fasta -t upgma_tree.txt -s 1 -1 -1
UPGMA tree construction is done.
user@dijkstra$ cat upgma_tree.txt
(((A:0.2,B:0.3):0.3,(C:0.5,D:0.3):0.2):0.3,E:0.7):0.0;
```