Bilkent University
CS481 | CS583 - Bioinformatics Algorithms

Fall 2022

Homework Assignment #3

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.
- **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/1rU0DhBg6yVXbfEtNuK73pjc1yWSPG90YnGNH-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) GUIDE TREES

Aim: In this assignment, you will learn about Guide Trees using UPGMA and Neighbor Joining.

1.1 Overview

Given a set of sequences construct two guide trees. One using the UPGMA algorithm and another using Neighbor Joining Algorithm. The sequences to be used are in an input FASTA file. **Align the sequences first**, using any aligner you want (your own), and **then proceed to construct the tree**. Save the alignment output as Newick format in an output file.

1.2 Data

- **Input**
  1. A flag that indicates which tree to build. Given by the `-t` switch. The values can be `nj` or `upgma`.
  2. FASTA-formatted file containing all sequences. This file may include up to 25 sequences. Sequences can be multi-lined as seen in the previous HW. The file is given using the `-i` switch.
  3. The scoring function for the match, mismatch and gap penalties. The scores are given using the `-s` switch, separated by "`:`".
- **Output**
  1. A text file containing the tree generated in **Newick** format. (https://en.wikipedia.org/wiki/Newick_format. Given by the switch `-o`.

2) EXAMPLE 1

The numbers presented in the example are not necessarily correct. They just show how the program should output the information.

2.1 Input

```
user@dijkstra$ cat input.fasta
> SHEEP
MEPSLLMWRFFVFIVVPGCVTEACHDDPPSLRNA----------MFKVLRYE----VGTM
>MOUSE
MEPRLLMLGFLSLTIVPSCRAELCLYDPPEVPNA----------TFKALSYK----NGTI
>FELCA
MEPSLLLWGILTFVVVHGHVTELCDENPPDIQHA----------TFKALTYK----TGTM
>HUMAN
MDSYLLMWGLLTFIMVPGCQAELCDDDPPEIPHA----------TFKAMAYK----EGTM
```

2.2 Execution

```
user@dijkstra$ ./hw4 -t nj -i input.fasta -o output.tree -s 1:-1:-5
```

2.3 Output

```
user@dijkstra$ cat output.tree
((SHEEP:0.25000,FELCA:0.20652):0.03261,MOUSE:0.23913,HUMAN:0.17391);
```

3) EXAMPLE 2

The numbers presented in the example are not necessarily correct. They just show how the program should output the information.

3.1 Input

```
user@dijkstra$ cat input.fasta
>BOVIN
LPLPKVQC---FVFNVEYMNCTWNSSSEPQPNNLTLHYGYRNFNGDDKLQECGHYLFS--
>CANFA
LPLPEVQC---FVFNVEYMNCTWNSSSEPRPTNLTLHYWYKNSN-DDKVQECGHYLFS--
>RAT
NDCSHLKC---FYNSRANVSCMWSPEEALNVTSCHIHAK-SDMRHWNKTCELTPVRQASW
MOUSE
---------------------MAVEELQSIIKRCQILEE-HDFKEEDF----GLFQLAGQ
```

3.2 Execution

```
user@dijkstra$ ./hw4 -t upgma -i input.fasta -o output.tree -s 1:-1:-5
```

3.3 Output

```
user@dijkstra$ cat output.tree
((BOVIN:0.0648148,CANFA:0.0648148):0.445145,(RAT:0.411765,MOUSE:0.411765):0.445145):0.445145;
```

3.4 Extra Points!

Using figtree (`https://treeviewer.org/`, `https://drive.google.com/file/d/1RsCdrdjpcT244UESCKQK8y302h64KjdI`), generate a PNG of your output file. This can be done via BASH script, and included in the BUILD target of the MAKEFILE (hint `java -jar figtree.jar -help`).