Bilkent University

CS481 | cs583 - Bioinformatics Algorithms

Spring 2025
Deadline: **18 Apr 2025**

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 required 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 source files. **DO NOT** include binaries, executables, `__MACOSX` folders, `.git` directories, or input files.
  - `Makefile` to compile the source code and produce the binary. Even if you use Python, include this Makefile.
  - A `README` file briefly describes how your program works.
  - A ~~`report.txt`~~ `report.pdf` file that includes your plots and discussion.
- All submissions must be made **strictly before the stipulated deadline**.
- Your implementation must strictly follow the approach discussed in class. Implementations based on external algorithms, libraries, or sources are not allowed.

1) MULTIPLE SEQUENCE ALIGNMENT

Aim: In this assignment, you will learn about multiple sequence alignment using pair-wise alignment and center-star alignment.

1.1 Data

- **Input**
  1. A FASTA file containing **MULTIPLE**, **MULTILINE** sequences. The file is given using the `-i` switch.
  2. The Affine Gap Penalty scores for the match, mismatch, opening gap, and gap penalties. The scores are given using the `-s` switch, separated by "`:`".
- **Output**
  1. A text file containing multiple sequence alignment in PHYLYP format (https://scikit.bio/docs/latest/generated/skbio.io.format.phylip.html. Given by the switch `-o`.

1.2 Tasks

Perform multiple sequence alignment using the Center-Star algorithm. The input sequences will be provided in a FASTA file, and your program should output the aligned sequences in PHYLIP format. After implementing the algorithm, evaluate your program using multiple randomly generated datasets. Conduct two sets of experiments:

- Vary the number of sequences while keeping sequence lengths similar—for example, use datasets with 4, 8, 16, and 32 sequences, each approximately 1,000 bases long.
- Keep the number of sequences fixed (e.g., 16 sequences) while varying the sequence lengths—for example, 100, 1,000, 10,000, and 100,000 bases.

You may benchmark your experiments with **/bin/time -v** command.

These setups are provided as a starting point. If you find that the results are not sufficiently informative or consistent, feel free to choose alternative values for the number of sequences or sequence lengths that better demonstrate the behavior and performance of your implementation.

For each experiment, measure and record the execution time and memory usage of your program. Plot your results and write a brief analysis discussing whether the observed outcomes matched your expectations. If not, explain any discrepancies or interesting findings you encountered.

Your executable should be **hw3**. If you implement it in Python, make it **hw3.py**.

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
>gi|3212
MGIKGLTGLLSENAPKCMKDHEMKTLFGRKVAIDASMSIYQFLIAVRQQE
>gi|3211
MGIKGLTQVIGDTAPTAIKENEIKNYFGRKVAIDASMSIYQFLIAVRSE
>gi|3210
MGIKGLTQTRGDTAPTAIKEIKNYFGRKVVIDASMSIYQFLIAVRSGET
```

2.2 Execution

```
user@dijkstra$ ./hw3 -i input.fasta -o output.phy -s 5:-4:-16:-4
```

2.3 Output

```
user@dijkstra$ cat output.phy
     3    51
gi|3211    MGIKGLTGLLSENAPKCMKDHEMKTLFGRKVAIDASMSIYQFLIAVRQQE-
gi|3212    MGIKGLTQVIGDTAPTAIKENEIKNYFGRKVAIDASMSIYQFLIAVRS-E-
gi|3210    MGIKGLTQTRGDTAPTAIKE--IKNYFGRKVVIDASMSIYQFLIAVRSGET
```