
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) 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 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 multiple sequence alignment in PHYLIP format (<http://scikit-bio.org/docs/0.2.3/generated/skbio.io.phylip.html>). Given by the switch `-o`.

1.2 Tasks

Perform Multiple Sequence Alignment using Center-Star Algorithm. The sequences to be aligned are in the input FASTA file. Save the alignment output as PHYLIP format in an output file.

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

```
1 user@dijkstra$ cat input.fasta
2 >S1-----a
3 MPE
4 >S2-----b
5 MKE
6 >S3__c
7 MSKE
8 >S4-----d
9 SKE
```

2.2 Execution

```
1 user@dijkstra$ ./hw3 -i input.fasta -o output.phy -s 1:-1:-5
```

2.3 Output

```
1 user@dijkstra$ cat output.phy
2      4      4
3 S1-----aM-PE
4 S2-----bM-KE
5 S3__c      MSKE
6 S4-----dS-KE
```

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

```

1 user@dijkstra$ cat input.fasta
2 >gi|3212
3 MGIKGLTGLLSENAPKCMKDHEMKTFLGRKVAIDASMSIYQFLIAVRQQDGQMLMNESGDVTSHLMGFFY
4 RTIRMVDHGIKPCYIFDGKPELKGSVLAKRFARREEAKEGEEAAKETGTAEDVDKLARRQVRVTREHNE
5 ECKKLLSLMGIPVVTAPGEAEAQCAELARAGKVYAAGSEMDTLTFHSPILLRHLTFSEAKKMPISEIHL
6 DVALRDLEMSMDQFIELCILLGCDYLEPCKGIGPKTALKLMREHGTLGKVVHIRGKMAEKAEEIKAAAD
7 EEAEAEAEAEKYSDPENEEGGETMINSDGEEVPAPSKPKSPKKKAPAKKKKIASSGMQIPEFWPWEEAK
8 QLFLKPDVVNGDDLVLWEWKQPDTEGLVEFLCRDKGFNEDRVRAGAAKLSKMLAAKQQGRLDGFFTVKPK
9 PAAKDAGKGKGDTKGEKRKAEEKGAAKKTKK
10 >gi|3211
11 MGIKGLTQVIGDTAPTAIKENEIKNYFGRKVAIDASMSIYQFLIAVRSEGAMLTSADETTSHLMGIFYR
12 TIRMVDNGIKPVYVFDGKPPDMKGELTKRAEKREEASKQLVLATDAGDAVEMKMNKRLVKVNGHTDE
13 CKQLLTLMGIPYVEAPCEAEAQCAALVKAGKVYATATEDMDSLTFGNSVLLRYLTYSEAKKMPIKEFHLD
14 KILDGLSYTMEFIDLCLMGLCDYCDTIKIGAKRAKELIDKHRCIEKVIENLDTKKYTPENWPYQEAR
15 RLFKTPDVADAETLDLKWTPDEEGLVKFMC GDKNFNEERIRSGAKKLCKAKTGQTQGRLD SFFKVLPS
16 KPSTPSTPASKRKVGCIYFLYF

```

3.2 Execution

```

1 user@dijkstra$ ./hw3 -i input.fasta -o output.phy -s 1:-1:-5

```

3.3 Output

```

1 user@dijkstra$ cat output.phy
2      2      454
3 gi|3211      MGIKGLTGLL SENAPKCMKD HEMKTFLGRK VAIDASMSIY QFLIAVRQQD
4 gi|3212      MGIKGLTQVI GDTAPTAIKE NEIKNYFGRK VAIDASMSIY QFLIAVRSE -
5
6      GQMLMNESGD VTSHLMGFFY RTIRMVDHGI KPCYIFDGKP PELKGSVLAK
7      GAMLTSADEG TTSHLMGIFY RTIRMVDNGI KPVYVFDGKP PDMKGELTK
8
9      RFARREEAKE GEEAAKETGT AEDVDKLARR QVRVTREHNE ECKKLLSLMG
10     RAEKREEASK QLVLATDAGD AVEMKMNKR LVKVNKGHTD ECKQLLTLMG
11
12     IPVVTAPGEA EAQCAELARA GKVYAAGSED MDTLTFHSPI LLRHLTFSEA
13     IPYVEAPCEA EAQCAALVKA GKVYATATED MDSLTFGSNV LLRYLTYSEA
14
15     KKMPISEIHL DVALRDLEMS MDQFIELCIL LGCDYLEPCK GIGPKTALKL
16     KKMPIKEFHL DKILDGLSYT MDEFIDLCLM LGCDYCDTIK GIGAKRAKEL
17
18     MREHGTLGKV VEHIRGKMAE KAEEIKAAAD EEAEAEAEAE KYSDPENEE
19     IDKHRCIEKV IENL-----D -----TK KY-TVPEN--
20
21     GGETMINSDG EEVPAPSKPK SPKKKAPAKK KKIASSGMQI PEFWPWEEAK
22     -----WPYQEAR
23
24     QLFLKPDVVN GDDLVLWEWKQ PDTEGLVEFL CRDKGFNEDR VRAGAAKLSK
25     RLFKTPDVAD AETLDLKWTPQ PDEEGLVKFM CGDKNFNEER IRSGAKKLCK
26
27     MLAAKQQGRL DGFFTVKPKPE PAAKDAGKGK GKDTKGEKRK AEEKGAAKKK
28     AKTGQTQGRL DSFFKVLPP--SSKPS----TPSTPASKRK V---GCIYFLYF
29
30     TKK-
31     FLYF

```