Kuiming Zhou

1327380

# CIS 6060 — Assignment #2

Fall 2025

**Due:** *Wednesday, Feb. 26, 2025 at 23:59*
Total Points: 50

Please submit your assignment solutions as **one zip** file (named as YOUR/UoG/ID_a2.zip, e.g. 1234567_a2.zip) to Dropbox under Assignment 2 before the due date.

## Exercise 1 Sequence Alignment (7)

1. (1pt) Which of the following scoring matrices would be most appropriate to find distant homology? Justify your answer.
   (a) PAM 30 and BLOSUM 90
   (b) PAM 30 and BLOSUM 45
   (c) PAM 250 and BLOSUM 90
   [✓ (d) PAM 250 and BLOSUM 45]

2. (a) (1pt) With a match score of 3, and mismatch score of −1, and using a gap penalty of −2, what is the score of the following alignment?

$$\begin{array}{c} TCTCCTGACCCC \\ TCT----AGTCC \end{array}$$

   (b) (1pt) If we keep the same match score and mismatch score, but use a gap open penalty of −5 and a gap extention penalty of −1, what is the score of the previous alignment?

   (c) (4pts) Using the scoring system used in part (b), which of the following alignments is the best? How do you know?

$$\begin{array}{c} TCTCCTGACCCC \\ TCT--AGTCC-- \end{array}$$

$$\begin{array}{c} TCTCCTGACCCC \\ TCT----AGTCC \end{array}$$

$$\begin{array}{c} TCTCCTG-ACCCC \\ TCT--AGT-CC-- \end{array}$$

$$\begin{array}{c} TCT---CCTGACCCC \\ TCTAGTCC------- \end{array}$$

### Submission

Submit the answers in a PDF file with calculation steps included for all the sub-questions in question 2.

## Exercise 2    Dot Plot (3 pts)

Draw a dot plot between the sequences `ATGACGGCTA` and `AATGCGTCT`. Use the following parameters: match = +1, mismatch = −1, window size = 3, and threshold = 1. You may create a matrix to represent the dot plot, as the example shown below.

| | A | T | G | A | C | G | G | C | T | A |
|---|---|---|---|---|---|---|---|---|---|---|
| A | | | | | | | | | | |
| A | | | | | | | | | | |
| T | | | | | | | | | | |
| G | | | | | | | | | | |
| C | | | | | | | | | | |
| G | | | | | | | | | | |
| T | | | | | | | | | | |
| C | | | | | | | | | | |
| T | | | | | | | | | | |

### Submission

Submit the answers in the PDF file you have created.

## Exercise 3    Database, Sequence Alignment (20 pts)

1. Which organism is more closely related to humans—the mouse or the blue whale? Intuition might say the mouse, but then again, the blue whale is also a mammal. Here, you are going to use mitochondrial 线粒体 DNA sequences and sequence alignments to answer this question.

   (a) (3pts) From NCBI, download the complete *mitochondrial* reference genome sequence for the three organisms, *Homo sapiens*, *Balaenoptera musculus* and *Mus musculus*. Save the files in FASTA format, Make sure that all three sequences are of similar lengths (should be between 16000 and 17000 bp).

   **Submission**: In the PDF file, write down in the **header lines** of the three FASTA files. (Hint: Think about which database on NCBI you are going to use, and also you can limit your results by indicating the species when searching.)

   (b) (3pts) Use the `needle` program (Needleman-Wunsch introduced in lecture, with default parameters) to perform three separate pairwise global sequences alignments and record the results based on the following submission requirement.

   **Submission**: In the PDF file, include the top part of the output page (all the information before the actual alignment).

   For example, the result may include:

```
############
Aligned_sequences: 2
1:
2:
Matrix:
Gap_penalty:
Extend_penalty:
Length:
Identity:
Similarity:
Gaps:
```

Score:
############

(c) (2pts) Based on your alignment(s), which two sequences are more similar? What information led you to this answer? What conclusion can you draw from the alignment(s)? Note that the results of this sequence alignment do not *prove* that one organism or the other is more closely related to human; potentially, a different conclusion could have been reached if a different genomic sequence (othe than *mitochondrial* reference genome) had been used.

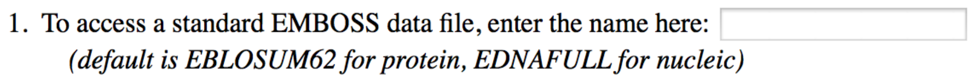
**Submission**: Write down your answer in the PDF file.

2. Search NCBI for a particular protein found mainly in muscle tissue in *Homo sapiens* (NP_001128711), a homolog in *Mus Musculus* (NP_034829), and a homolog in *Xenopus laevis* (NP_001080702, a frog).

(a) (3pts) Using Needleman-Wunsch, and the default parameters, what is the score, and the percent identity when aligning each protein to the others? (Note: You will need to conduct three separate pairwise alignments on these sequences, and report the answers respectively.)

(b) (9pts) For the online program (both Smith-Waterman and Needleman-Wunsch), the scoring matrix can be changed from PAM10 to PAM500 in increments of 10.

Just change this field on the websites:



1. **To access a standard EMBOSS data file, enter the name here:** ▭
**_(default is EBLOSUM62 for protein, EDNAFULL for nucleic)_**

by typing in "EPAM10" or the appropriate one available (note the 'E').

Using Needleman-Wunsch, align the above human protein (NP_001128711) with the mouse protein (NP_034829) using PAM10, PAM100 and PAM200. Do the same with the human protein (NP_001128711) and the *Ruminococcoides intestinale* protein (bacteria, accession WP_015522884). In the result page, track down the (alignment) scores and write them down in your submission (6ptss). What conclusions can you draw about the use of different matrices (3pts)?

**Submission**: Write down your answer in the PDF file. Tables are preferred.

## Exercise 4 Programming Question (20 pts)

When a segment of DNA is sequenced by a sequencing machine, the result is often placed in a FASTQ file. FASTQ files are similar to FASTA files, but also contain information about the "quality" of each base. The "quality" of a given base indicates the level of confidence that the base is actually correct. (Sequencing machines are not 100% accurate!). High quality scores indicate confidence that the base is correct, while low quality scores indicate a higher likelihood that the base is incorrect. Quality scores range from 0 to 40. However, in a FASTQ file, each quality score is in fact represented by a single character. To get the quality score associated with each character, 33 is subtracted from the ASCII value of that character (type `man ascii` or visit http://en.wikipedia.org/wiki/ASCII for tables of ASCII values). Thus, a quality score of 0 is represented by the character '!' (ASCII value 33), while a quality score of 40 is represented by the character 'I' (ASCII value 73).

An example of a FASTQ file is as follows:

```
@my_sequence
GCCAGCAGCCGCGGTAACACGTAGGCACCA
+
DD;?BDCCC=CBC>CB<BB=:8-.+/IC=D
```

The first line of the file begins with an @ sign, and contains the name of the sequence. The second line contains the sequence in its entirety (the sequence may not be spread over more than one line, as in a FASTA file). The third line consists only of a + sign. The fourth line contains characters representing the quality score associated with each corresponding base. There is one quality score for each base in the sequence. In this example, the quality scores of the first five nucleotides (G, C, C, A, and G) are 35, 35, 26, 30, and 33, respectively.

In this question, you will write a Python script called `quality_checker.py`. The script will prompt the user for the (name and path of a) FASTQ file they would like to check, and a quality score threshold. Your script is to perform the following tasks:

1. Read in the FASTQ file. Although FASTQ files in general may contain more than one sequence, you may assume that the input file contains only a single sequence.

2. Check to make sure that all of the bases in the sequence portion of the file are valid. Valid characters are [acgtACGT]. If any of the characters are not valid, your script should print a message to this effect to the standard output, and then exit.

3. Check to make sure that all of the characters in the quality score portion of the file are valid. A character is NOT valid if (ASCII value of character minus 33) is less than 0 or greater than 40. If any characters are not valid, your script should print a message to this effect to the standard output, and then exit.

4. If both the sequence and the quality scores are valid, then your script is to check whether the **average** quality value is equal to or greater than the quality score threshold specified by the user. If the average is less than the threshold, then your script should output, "<name of input file> does not meet the specified quality threshold". Otherwise, your script should output, "<name of input file> meets the specified quality threshold".

### Submission

1. Submit the source code of this question as `quality_checker.py`. Make sure to include your name and student number as comments in the file.

2. Submit the documentation as `README_quality_checker.txt`. It should include information on how to run your code, any dependency or libraries needed, etc..

3. Submit the **screenshot** of testing your script using the two FASTQ files attached with this assignment: `toy_seq.fastq` and `5_OHara_1.fsatq` (downloaded from https://zenodo.org/records/3736457 and it is the first sequence from the original file `5_OHara_S2B_trnL_2019_minq7.fsatq`).

(a) (1pt) With a match score of 3, and mismatch score of $-1$, and using a gap penalty of $-2$, what is the score of the following alignment?

3 3 3 -2 -2 -2 -2 3 -1 -1 3 3

$$TCTCCTGACCCC$$

$$TCT{-}{-}{-}{-}AGTCC$$

$3+3+3-2-2-2-2+3-1-1+3+3$

$= 9-8+3-2+6 = 8$

(b) (1pt) If we keep the same match score and mismatch score, but use a gap open penalty of $-5$ and a gap extention penalty of $-1$, what is the score of the previous alignment?

3 3 3 -5 -1 -1 -1 3 -1 -1 3 3

$TCTCCTGACCCC$

$TCT{-}{-}{-}{-}AGTCC$

$3+3+3-5-1-1-1+3-1-1+3+3$

$= 9 - 8 + 3 - 2 + 6 = 8$

(c) (4pts) Using the scoring system used in part (b), which of the following alignments is the best? How do you know?

3 3 3 -5 -1 -1 3 -1 3 3 -5 -1
*TCTCCTGACCCC* = 9-7 +3-1 +6-6
*TCT−−AGTCC−−* = 4

The best.

3 3 3 -5 -1 -1 -1 3 -1 -1 3 3 9-8 +3-2+6
*TCTCCTGACCCC* = 8
*TCT−−−−AGTCC*

3 3 3 -5 -1 -1 3 -5 -1 3 3 -5 -1 9 - 7 +3- 6+ 6 - 6
*TCTCCTG−ACCCC* = -1
*TCT−−AGT−CC−−*

3 3 3 -5 -1 -1 3 3 -5 -1 -1 -1 -1 -1 -1
*TCT−−−CCTGACCCC* 9-7+6-11
*TCTAGTCC−−−−−−−* = -3

As the second one has only one consecutive gap.

## Exercise 2 Dot Plot (3 pts)

Draw a dot plot between the sequences ATGACGGCTA and AATGCGTCT. Use the following parameters: match = +1, mismatch = −1, window size = 3, and threshold = 1. You may create a matrix to represent the dot plot, as the example shown below.

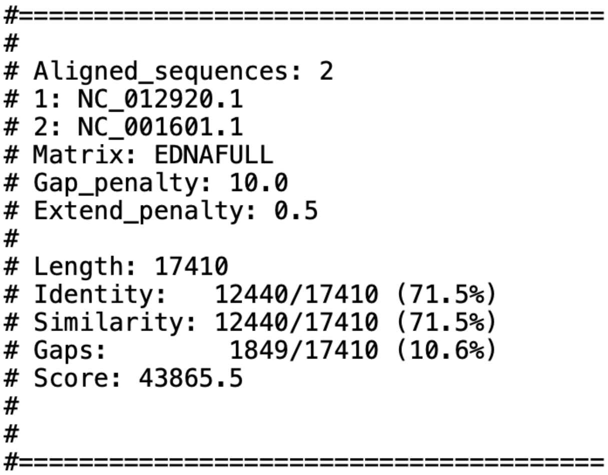
| | A | T | G | A | C | G | G | C | T | A |
|---|---|---|---|---|---|---|---|---|---|---|
| A | | | | | | | | | | |
| A | ● | | | ● | | | | | | |
| T | | ● | | | | ● | | | | |
| G | | | | ● | | | ● | | | |
| C | | | | | ● | | | | | |
| G | | | ● | | | ● | | | | |
| T | | | | | | | ● | | | |
| C | | | | | | | | | | |
| T | | | | | | | | | | |

3

1) a)

>NC_012920.1 Homo sapiens mitochondrion, complete genome
>NC_001601.1 Balaenoptera musculus mitochondrion, complete genome
>NC_005089.1 Mus musculus mitochondrion, complete genome

1) b)



```
#=======================================
#
# Aligned_sequences: 2
# 1: NC_012920.1
# 2: NC_001601.1
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 17410
# Identity:   12440/17410 (71.5%)
# Similarity: 12440/17410 (71.5%)
# Gaps:        1849/17410 (10.6%)
# Score: 43865.5
#
#
#=======================================
```

human & blue whale

```
#=======================================
#
# Aligned_sequences: 2
# 1: NC_012920.1
# 2: NC_005089.1
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 17911
# Identity:   11828/17911 (66.0%)
# Similarity: 11828/17911 (66.0%)
# Gaps:        2954/17911 (16.5%)
# Score: 40515.0
#
#
#=======================================
```

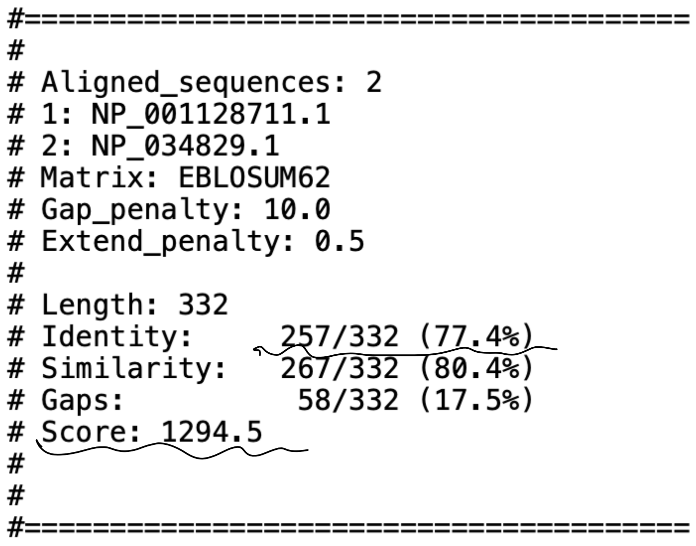
human & house mouse

```
#=======================================
#
# Aligned_sequences: 2
# 1: NC_001601.1
# 2: NC_005089.1
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 17578
# Identity:   12080/17578 (68.7%)
# Similarity: 12080/17578 (68.7%)
# Gaps:        2455/17578 (14.0%)
# Score: 42955.5
#
#
#=======================================
```

blue whale & house mouse

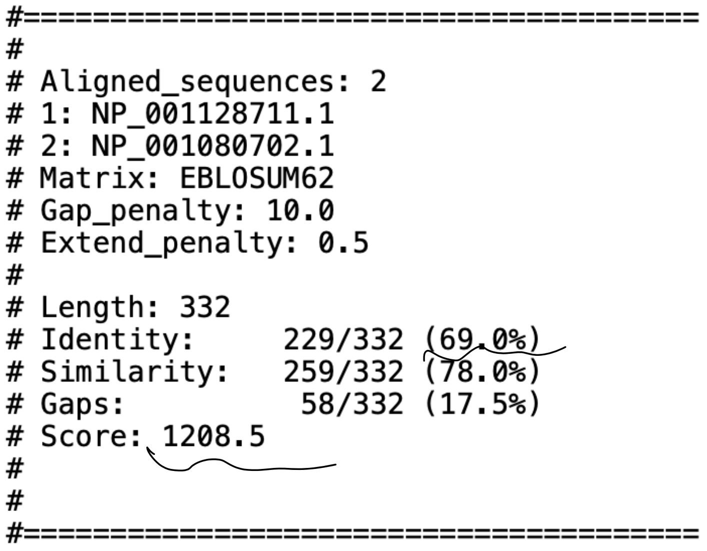
Based on the alignment, human and blue whale are more similar while comparing mitochondrial DNA sequences. As they are having 71.5% identity and highest score of 43865.5

2. a.



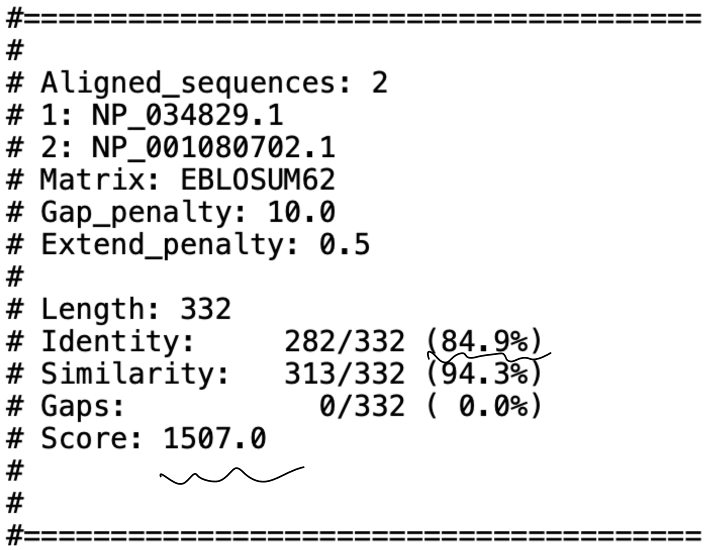
```
#=======================================
#
# Aligned_sequences: 2
# 1: NP_001128711.1
# 2: NP_034829.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 332
# Identity:     257/332 (77.4%)
# Similarity:   267/332 (80.4%)
# Gaps:          58/332 (17.5%)
# Score: 1294.5
#
#
#=======================================
```

human
& house mouse



```
#=======================================
#
# Aligned_sequences: 2
# 1: NP_001128711.1
# 2: NP_001080702.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 332
# Identity:     229/332 (69.0%)
# Similarity:   259/332 (78.0%)
# Gaps:          58/332 (17.5%)
# Score: 1208.5
#
#
#=======================================
```

house mouse
&
frog



```
#=======================================
#
# Aligned_sequences: 2
# 1: NP_034829.1
# 2: NP_001080702.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 332
# Identity:     282/332 (84.9%)
# Similarity:   313/332 (94.3%)
# Gaps:           0/332 ( 0.0%)
# Score: 1507.0
#
#
#=======================================
```

2. b.

human v.s. mouse

| | |
|---|---|
| PAM 10 : | 1984.5 |
| PAM 100 : | 1453.5 |
| PAM 200 : | 1521.5 |

human v.s. bacteria.

| | |
|---|---|
| PAM 10 : | 94. |
| PAM 100 : | 319.5 |
| PAM 200 : | 491 |

when comparing human and mouse, as they are not that divergent, the lower pam generates higher score. while comparing human and bacteria, as they are so divergent, the higher pam will get higher score.