

GS 373 Homework 1

Due April 12th before 1:30 PM on Canvas

- 100 points: 4 bioinformatics questions (80 points), 1 programming assignment (20 points).
- Submit answers to everything except the programming assignment in a Microsoft Word document or PDF and upload that to Canvas. Type answers immediately below each question.
- Submit the programming assignment as a separate .py file onto Canvas that can be directly run by Python.
- Justify all of your answers. Some questions may require information from Friday's lecture.
- Question 4 requires external reading which is provided on Canvas

Bioinformatics Questions

1. (15 points) Consider two very short DNA sequences, CA and GC. **Write out every possible alignment of these two sequences that has at most one gap in each sequence.** Remember that a gap aligned with another gap is non-informative.
2. (15 points) Below is a possible alignment of two longer sequences. **Score this alignment using a linear gap penalty of -4 and the substitution matrix provided below.** Please show your arithmetic.

substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

gaps = -4

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

3. (25 points) Consider two homologous DNA sequences, GATTC and CCATG. **Use the Needleman-Wunsch algorithm to find the optimal local alignment between these two sequences.** Use a linear gap penalty of -4 and the substitution matrix provided below. The dynamic programming matrix is already outlined below, you just need to fill it according to the algorithm. *Be sure to write out your final alignment!*

Note: You will need to fill out both the scoring matrix values **and** the traceback arrows. You can either complete the matrix on paper and include a photo or create the table and add in arrows using PowerPoint or another program.

substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

gaps = -4

		G	A	T	T	C
	0					
C						
C						
A						
T						
G						

4. (25 points) An important genome sequencing study from 1998 (Andersson *et al.*, 1998) is posted on both the canvas and quiz section websites. This paper describes the genome sequence of the parasitic bacterial species *Rickettsia prowazekii* and compares it to the mitochondrial genome of the single-celled eukaryote *Reclinomonas americana*. **Read the introduction, the methods section, and the results sections titled “Reductive evolution” and “Mitochondrial affinities” and answer the questions below.** Don’t worry about getting every detail – focus on the big picture.
- (8 points) Describe two findings reported in the paper that required sequence alignment analysis.
 - (7 points) Which of the various mutational processes described (gene loss via pseudogene conversion, genome rearrangement, gene inversion) would cause major difficulties for aligning the bacterial and mitochondrial genomes using Needleman-Wunsch? Explain why.
 - (5 points) Imagine you are doing a global alignment of large segments of two genomes, and you suspect there has been a lot of gene loss in one or both genomes. How might that affect the scoring parameters you choose for the alignment?

(20 points) Programming problem:

Write a Python program that performs various calculations on a pair of integers and reports the results. Your program should perform the following steps:

- 1) **Take two integers as input and print their values.**

This step can be accomplished using the following lines of code (included in the starter Python file `homework1_skh.py`). They tell the computer to take two inputs from the command line, assign their values as integers to the variables named `a` and `b`, and then print the values of `a` and `b`.

```
# GS 373 HW1 by SKH
```

```
import sys    # This line imports a module that lets us load variables
```

```
a = int(sys.argv[1])    # Assigns the first number to 'a'
b = int(sys.argv[2])    # Assigns the second number to 'b'
print("a is", a)
print("b is", b)
```

```
# Add your code below...
```

For example, if you wanted to provide the numbers 9 and 4 as input to the program, you would run the script from the command line with the command: `python homework1_skh.py 9 4`

This command would assign the variable **a** the value of 9 and the variable **b** the value of 4.

- 2) **Calculate and print the sum of the two integers.**
- 3) **Check whether the first integer is even or odd and print either “A is even” or “A is odd”.**
Hint: the operator “%” (called the modulo operator) returns the remainder of one number divided by another. For example, `6 % 3 = 0`.
- 4) **Check which of the two integers is larger. Set the variable `a` to the larger integer and set the variable `b` to the smaller integer. Print out the values of `a` and `b` to confirm the ordering.**
Hint: You may want to create a third value in the case where you need to switch the values of **a** and **b**.

Notes:

- I have included a starter python file you may use to start your program (homework1_skh.py). Below is the code in the starter file:

```
# GS 373 HW1 by SKH
```

```
import sys    # This line imports a module that lets us load variables
```

```
a = int(sys.argv[1])    # Assigns the first number to 'a'
b = int(sys.argv[2])    # Assigns the second number to 'b'
print("a is", a)
print("b is", b)
```

```
# Add your code below...
```

- Label each of the above steps in your program with a comment (#).
- Upload your program as a separate .py file on Canvas. **Please name your .py file in the form homework1_***.py where the stars are your initials.**
- Here is an example output from the completed program using the sample call:
python homework1_skh.py 4 9

```
a is 4
b is 9
sum is 13
a is even
a is 9
b is 4
```

(5 optional points) Feedback:

The following questions are not required but answering them will give you an additional 5 points.

- a) Which aspects of the first week of class and quiz section did you find most challenging?
- b) How much time did you spend on this assignment? How much of that was on the programming problem?