Name:

Date of mid-term exam: Thursday, October 12
Duration: 2 hours from 10AM to 12PM

The exam has three sections. The first two are worth 28 points each while the last one is worth 24 points – 80 points total. For section 1 you will need your laptop to connect to the server and do some work in the terminal. You may use your class notes and examples from previous lectures to solve the problems presented here.

Linux, the Terminal, and Python programming (28 points total)

1. You type *pwd* and it tells you that you are in the *Downloads* folder in your home directory (*~/Downloads*). You want to navigate to the *Desktop* that is also located in your home directory (*~/Desktop*).

Which command are you going to use to change directories? (1 pt)

cd

How do you use a relative path to navigate from *Downloads* to the *Desktop*? (1 pt)

cd ../Desktop

How do you use an absolute path to navigate from *Downloads* to the *Desktop*? (1 pt)

cd ~/Desktop

2. Use ssh to connect to the server at IP 168.123.185.34. Navigate to the folder *Midterm* in your home directory. In the folder you will find two files, *File1* and *File2*.

What file type is *File1*? (2 pt)

Fasta

What file type if *File2*? (2 pt)

FastQ

What are the owner/user of the file, the group and everyone else's permission for the two files? What command did you use to figure this out? (3 pts)
File1: Owner: read/write Group: read/write Everyone else: read
File2: Owner: read/write/execute Group: read/write Everyone else: read
ls -l
Now that you have determined the file types of each rename each file with a descriptive extension. In addition, set the permissions of both files to read only for user/owner,

group, and everyone else to prevent overwriting the file by accident. What command(s) did you use to achieve this? (3 pts)
mv File1 File1.fasta
mv File2 File2.fastq

How many sequences are contained in *File1*? What command did you use to determine the number of sequences? (1 pt)

2135; grep –c '>' File1.fasta

The *Midterm* folder contains a script (*MyScript*). What language is the script written in? (1 pt)

Python

Make the script *MyScript* executable and run it, giving it an appropriate input file as an argument (*MyScript file*). Write your commands below. What does the script do? (3 pts)
chmod 744 MyScript OR chmod u+x MyScript
./MyScript File1.fasta
The script translates nucleotide sequences using the first 2 reading frames
Take the first 300 sequences from *File2* and put them into a new file named *File2_First300* and put that file into a new directory named *Q3* in *Midterm*. What command(s) did you use to achieve these tasks? (3 pts)
mkdir Q3
head –n 1200 File2.fastq > Q3/File2_First300.fastq

How would you transfer the file *File2_First300* to your laptop over the network? Write down the exact command you would use – you may try and execute your command to verify that it works. (1 pt)

scp user@168.123.185.34:~/Midterm/Q3/File2_First300.fastq .

3. On Linux/Unix operating systems every program can produce output to 2 different streams that are usually printed to the display.

What are the names of these streams? (2 pts)

stdout (standard out – stream 1) and stderr (standard error – stream 2)

You can use the output of one program that is printed to the display as input for another program. What special character would you use to string multiple programs together in the terminal? What is the name for this character? (1 pt)
| pipe
How do you redirect the output of a program to a file? What special character do you use to achieve this? (1 pt)
> or >>

Name:

4. Python is a _____ programming language. (1 pt)

Choose one: *compiled* or ***interpreted***

5. You assign a variable in the manner below in Python. What type of variable is *MyVar*? You may use the ipython shell to come up with an answer to this question. (1 pt)

MyVar = '33.2'

Looks like a float at first glance but the single quotes indicate that this is a string.

Sequence Alignment (28 points)

6. One fundamental assumption underlying sequence alignments is that the sequences to be aligned are homologous. What does it mean for sequences to be homologous? (1 pt)

The sequences share common ancestry

7. Homology comes in different flavors with some genes being orthologs while other are paralogs. What is the difference between orthology and paralogy? What may lead to paralogy? You can draw a tree diagram to make your point. (3 pts)

Orthologs share ancestry through speciation.

Paralogs are the result of gene duplications within a genome.

8. We know that certain mutations in DNA sequences are less frequent than others. In particular, changes from a pyrimidine to a purine are less common than changes from purine to purine or pyrimidine to pyrimidine. (2 pts)

Changes from purine to pyrimidine are referred to as **transversions**

Changes from purine to purine are referred to as **transitions**

9. There are pairwise and multiple sequence alignments. How many sequences are aligned in a pairwise alignment and a multiple sequence alignment? (1 pt)

pairwise alignment → 2; multiple sequence alignments → equal or greater than 3

Name:

10. What are the 3 steps in a progressive multiple sequence alignment? Provide a sketch of the procedure if necessary. (3 pts)

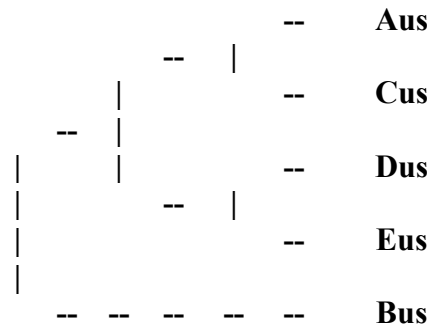
1) **Calculate distance matrix**
2) **Build guide tree from distance matrix**
3) **Progressive alignment according to guide tree**
    a. **Perform all pairwise alignments**
    b. **Alignment progresses from leaves to root**

11. What is a major drawback of progressive alignments? How do iterative alignments differ and address this issue? (2 pts)

- **Not guaranteed to find global optimum**
- **Mistakes made early on will be propagated to final result**

12. Given the distance matrix below (where 0 would mean two sequences are identical and 1 means that the sequences are completely different), what would the guide tree for that alignment look like?
(3 pts)

	Aus	Bus	Cus	Dus	Eus
Aus	0				
Bus	0.81	0			
Cus	0.18	0.92	0		
Dus	0.54	0.83	0.61	0	
Eus	0.46	0.78	0.64	0.27	0



13. What is the difference between a global and a local alignment? (2 pts)

Global: alignment is carried out from beginning to end of the sequence to find the best possible alignment.

Local: finds the local regions with high level of similarity

Name:

14. Below is the scoring matrix for a global alignment. Using the traceback indicated by the arrows, align the two sequences and report the alignment (indicate gaps using -). (3 pts)

	_	A	G	A	C	T	A	G	T	T	A	C
_	0	-5	-10	-15	-20	-25	-30	-35	-40	-45	-50	-55
C	-5	-3	-8	-13	-6	-11	-16	-21	-26	-31	-36	-41
G	-10	-6	4	-1	-6	-9	-12	-9	-14	-19	-24	-29
A	-15	0	-1	14	9	4	1	-4	-9	-14	-9	-14
G	-20	-5	7	9	9	6	3	8	3	-2	-7	-12
A	-25	-10	2	17	12	7	16	11	6	1	8	3
C	-30	-15	-3	12	26	21	16	11	11	6	3	17
G	-35	-20	-8	7	21	23	20	25	20	15	10	12
T	-40	-25	-13	2	16	29	24	20	33	28	23	18

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

15. Using the Smith-Waterman algorithm, perform a local alignment of the following 2 sequences and report the longest alignment you find. Show your work for all 3 steps of the alignment process on the next page (page 6)! (6 pts)

Match = +1
Mismatch = -2
Gap = -2
Sequence 1: TCAGTTGCC
Sequence 2: AGGTTG

Before you start: what do you think the best alignment will look like?

Both sequences share GTTG

Name:

Smith-Waterman Alignment:

Apply the match, mismatch and gap costs from the previous page! (example and solution courtesy of Avril Coghlan)

Match = +1
Mismatch = -2
Gap = -2

		T	C	A	G	T	T	G	C	C
	0	0	0	0	0	0	0	0	0	0
A	0	0	0	1	0	0	0	0	0	0
G	0	0	0	0	2	0	0	1	0	0
G	0	0	0	0	1	0	0	1	0	0
T	0	1	0	0	0	2	1	0	0	0
T	0	1	0	0	0	1	3	1	0	0
G	0	0	0	0	1	0	1	4	2	0

Alignment:

```
G T T G
| | | |
G T T G
```

(Pink traceback)

Name:

16. A BLAST search produces the following two local alignments. Using the provided scores, which alignment is better? (2 pts)

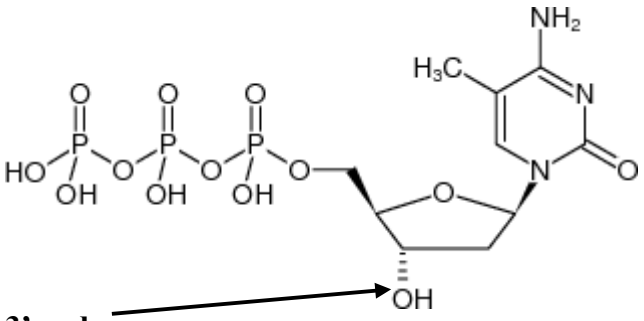
GACTTGC	**GA-TTGC**	Match = +3
|| || |	**|| ||||**	Mismatch = -2
GATTTAC	**GATTTGC**	Gap = -4

Alignment in bold is better since it receives a cumulative score of 14 (6 matches + 1 gap) while the second alignment receives a score of 11 (5 matches + 2 mismatches).

Genetics, Sequencing, and Assembly (24 points total)

17. DNA is extended from the **5'** end to the **3'** end. (1 pt)

18. Indicate the 3' end in the deoxynucleotide below. How would you modify this molecule to prevent DNA strand elongation? (2pts)



3' end

changing the hydroxyl group (OH) at the 3' end to a single hydrogen (H) would prevent extension of the growing DNA strand

19. A Sanger sequencing reaction produces the following population of DNA molecules, each of which is labeled with a fluorescent dye at the 3' end. What is the sequence that can be reconstructed from this pool of DNA sequences? (3 pt)

```
5'- AGGCTGACC -3'
5'- AGGCTGACCT -3'
5'- AGGCTGACCTG -3'
5'- AGGCTGACCTGA -3'
5'- AGGCTGACCTGAA -3'
5'- AGGCTGACCTGAAT -3'
```

During sequencing the strands would be separated on a gel matrix and only the last nucleotide of each strand would be read/identified. The resulting sequence would be:
C T G A A T

Name:

20. Two identical copies of DNA are created in the cell through **replication**. In the lab, we can duplicate DNA artificially using **PCR**. **Transcription** is the process that makes RNA based on a DNA template strand. **Translation** describes the process of creating proteins from RNA molecules. (4 pt)

a) translation
b) transcription
c) replication
d) PCR (Polymerase Chain Reaction)

21. Codons encode amino acids in genes and messenger RNAs. How long is a codon in DNA/RNA nucleotides? (1 pt)

3

22. You have a file with sequence data, one of which is the sequence below. What type of molecule was this sequence most likely derived from? (1 pt)

AGGCTCAATGGCTCATGTTTTTTTTTTTTTT

The poly-T tail indicates that the complementary sequence in the sequencing reaction was most likely a messenger RNA (mRNA) molecule.

23. Next generation sequencing data can be assembled using *de Bruijn* graphs that rely on overlaps between short words of length *k* derived from the sequences to be aligned.

Which 4-mers are contained in the following sequence? (2 pts)

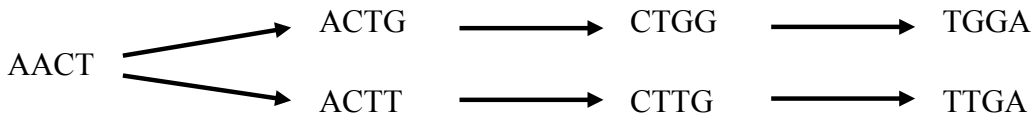
AACTCGA

AACT, ACTC, CTCG, TCGA

Draw the graph that connects these k-mers using an overlap of k-1. (2 pts)

Name:

24. Reconstruct the sequences from the k-mer graph below. (4 pts)



AACTGGA
AACTTGA

Name 2 scenarios that could lead to the fork in the assembly graph above. (2 pts)

- **Sequencing error**
- **Mutation (single nucleotide polymorphism) in a diploid organism**

Assuming only one of the sequences represents the real sequence encoded in the cell, what type of information could help you decide which sequence is the real one? (2 pt)

- **Information on number of reads supporting one path versus the other**
- **Sequencing reads that span the full length of this portion of the graph (eg, an observed sequencing read supports the lower path but no read supports the upper path – remember that paths may exist in the kmer graph that do not necessarily have a read supporting them)**