Final test Name:

Introduction to Bioinformatics, 2016 Neptun id:

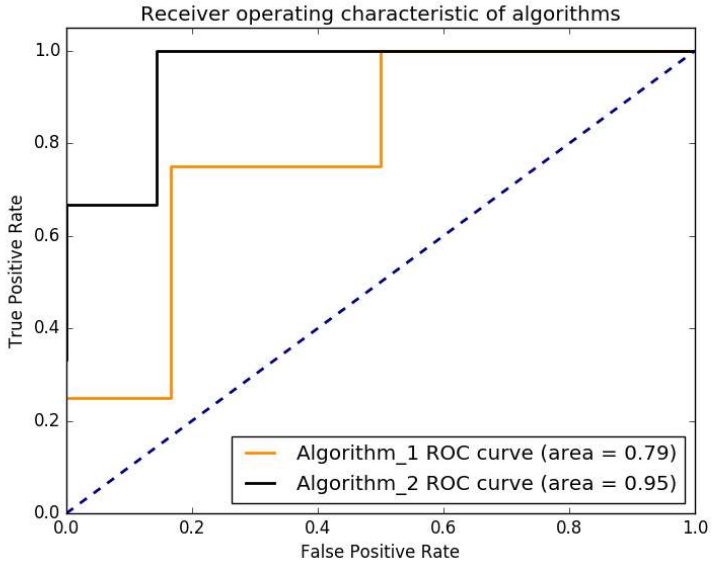
The point value of each question is listed in parentheses.
Total points: 100

1. **True or False? (23p)**

	T	F
The PAM matrix was constructed from handmade multiple alignments.	**x**	
A multiple alignment can be used to highlight conserved positions in a domain of several proteins.	**x**	
A three-dimensional structure can be used to validate a multiple alignment.	**x**	
The lower the E-value, the more significant your hit is.	**x**	
The grep unix utility is useful for filtering to lines of a text file.		**x**
Stockholm format is a multiple alignment format preferred by Pfam.	**x**	
Pair-end libraries are heavily used in assembly projects.	**x**	
SPAdes program uses a modified de Bruijn graph for assembling a genome.	**x**	
An advantage of doing phylogenetic analysis in the nucleotide level is that mutations are easier to follow here.	**x**	
Maximum likelihood method is an example for distance based phylogenetics.		**x**
The flying ability of bats and birds is an example for homoplasy.	**x**	
Mutations are changes in the DNA sequence that occur in more than 1% of the population.		**x**
Oncogene-addiction means that the tumor needs a specific (malfunctioning) signal transduction pathway in order to maintain itself.	**x**	
Functional genomics approaches are used to identify connections between genotype and phenotype in genomic level.	**x**	
Gro-seq is used for identifying all mRNAs in the cell at the time of the experiment.		**x**
MACS is a software tool for ChIP-seq data analysis.	**x**	
Operation Taxonomic Units (OUTs) are constructed from clustering 16SrRNA reads.	**x**	
In a dotplot, larger words show matching regions more clearly.	**x**	
We do not use multiple alignment in cases where point mutations are not the predominant cause of variations.	**x**	
A multiple alignment cannot be decomposed into pairwise alignments.		**x**
The NCBI accession number of a protein is not identical with the UniProtKB ID of the same protein, but its GI number is.		**x**
A nucleotide sequence can be the query of a blastx search.	**x**	
The size of the database does not affect the E value of a BLAST search.		**x**

2. **Given two ranking, where the positive elements are colored. Plot the ROC curve (False-positive rates (FPR) against the true-positive-rates (TPR))! (5p) Which one is the better ranker? Calculate the AUC value (3p).**

Algorithm_1	Algorithm_2
1.00E-23	1.00E-64
1.00E-12	1.00E-63
1.00E-9	1.31E-45
1.20E-9	1.02E-54
1.09E-2	1.03E-47
1.00E-1	1.00E-26
1.00E1	1.00E-22
3.00E1	1.02E-11
1.30E2	1.70E-06
1.91E2	1.65E-05



Name:
Neptun id:

3. Give the Burrows-Wheeler Transformed form of the following sequence! Give the Rank and Occupancy tables for the sequence! Which process requires these tables? Briefly describe your calculations! (6p)

sequence: AATGCATC

Burrows-Wheeler Transformed form: C$CATGTAA

1	A	A	T	G	C	A	T	C	$
2	$	A	A	T	G	C	A	T	C
3	C	$	A	A	T	G	C	A	T
4	T	C	$	A	A	T	G	C	A
5	A	T	C	$	A	A	T	G	C
6	C	A	T	C	$	A	A	T	G
7	G	C	A	T	C	$	A	A	T
8	T	G	C	A	T	C	$	A	A
9	A	T	G	C	A	T	C	$	A

end sequence symbol
shift the sequence one by one

reorder the table into alphabetic order of the rows:

if the first column is the same, then check second, third, etc. columns

new index										old index
0	$	A	A	T	G	C	A	T	C	2
1	A	A	T	G	C	A	T	C	$	1
2	A	T	C	$	A	A	T	G	C	5
3	A	T	G	C	A	T	C	$	A	9
4	C	$	A	A	T	G	C	A	T	3
5	C	A	T	C	$	A	A	T	G	6
6	G	C	A	T	C	$	A	A	T	7
7	T	C	$	A	A	T	G	C	A	4
8	T	G	C	A	T	C	$	A	A	8

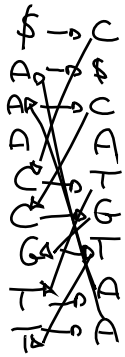
this is the BW transformed column

Rank table: the new index of the 1. appearance of a character in the 1. ~~row~~ of the ordered sequence shifts

	A	C	G	T
RANK	1	4	6	7

first occurance of A in the first column

What can we do with the BW transformed form?
-> Reconstruct original sequence



1) Start with $ symbol
2) Search first C in first column
3) Search 1. T in 1. col.
4) Search for 2nd A in 1st col. etc.
-> if it is a second A in the 2nd column, then I look for the second A in the 1st column

Occupancy table: rolling sums of the BWT form of the sequence

new index	A	C	G	T
0	0	1	0	0
1	0	1	0	0
2	0	2	0	0
3	1	2	0	0
4	1	2	0	1
5	1	2	1	1
6	1	2	1	2
7	2	2	1	2
8	3	2	1	2

this row means that 1 A, 2 C and 1 T is encountered in the last column so far

string searching requires the Rank and Occupancy tables,

4. What is the correct order of steps in RNA sequencing workflow? (6p)

a: normalizing expression levels;

b: NGS sequencing of the libraries;

c: fragment size selection;

d: visualization of differential gene expression between the samples;

e: collecting the samples;

f: isolate RNAs;

g: fragmentation of RNAs;

h: calculating expression levels;

i: break the cells;

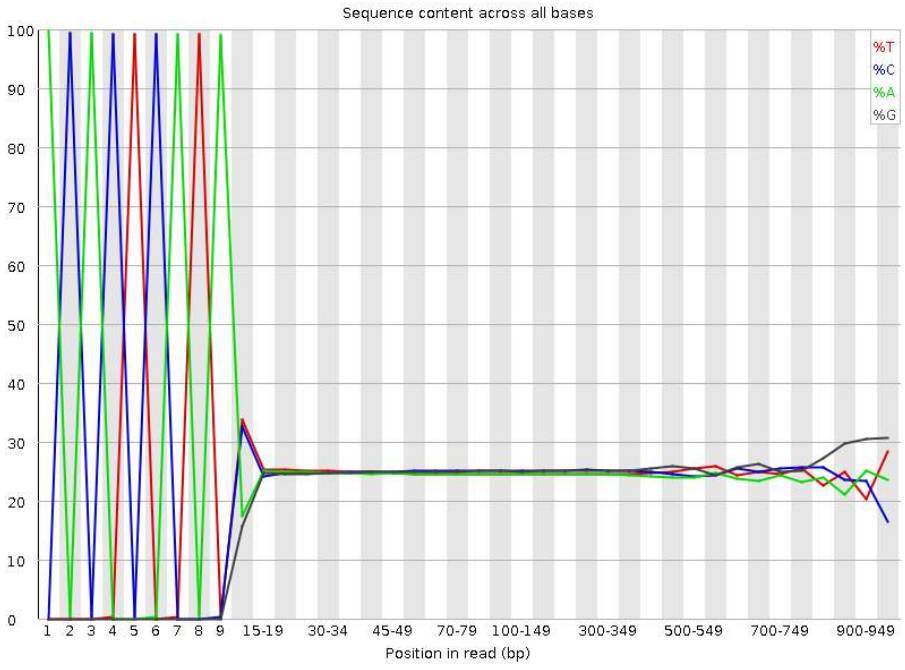
j: cDNA synthesis;

k: mapping to reference genome or transcriptome;

l: adding linker sequences

1	e
2	i
3	f
4	g
5	j
6	c
7	l
8	b
9	k
10	h
11	a
12	d

- **You sequenced a new pseudomonas strain. You checked the per base sequence content of the reads. Explain what you see in the picture and try to explain the results (3p). How would you improve the quality in that particular case? (2p)**



- It should be a line
- general: low quality at the beginning of reads; possibly adapters;
- Low quality in the endings
- Trimming in the beginnings (remove first 9-14 bases) and endings (>750)

5. **You made two different assemblies of the pseudomonas genome. Which one is the better based on the N50 statistics? Define what the N50 value is (2p), then calculate it for the two assembly (3p) (briefly describe your calculation)**

Length of contigs:

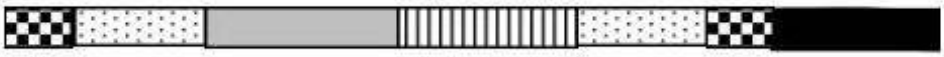
Genome 1: 420 280 240 110 87 13 10 6 3

Genome 2: 390 310 108 85 65 62 58 52 50 42 38

6. **Draw a dot-plot for the following two protein sequences! Briefly explain how you created the graph. (5p)**

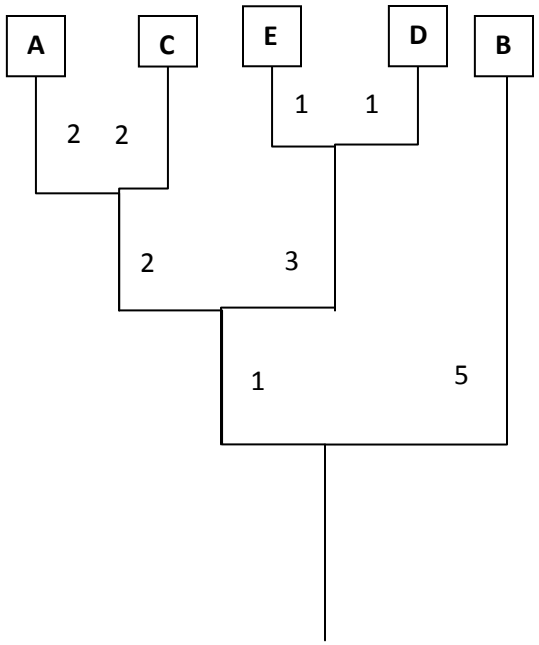
S1:

S2:



7. Given the following distance matrix draw a dendrogram based on the UPGMA method and briefly explain your calculation! (5p)

	A	B	C	D
B	10			
C	4	10		
D	8	10	8	
E	8	10	8	2



8. **Align the two sequences GMLVAI and GLVV using the Needleman-Wunsch algorithm (global alignment) using the following scoring matrix and a penalty -1 for indels. Briefly describe your calculation. (6p)**

	A	G	I	L	M	V
A	4					
G	0	6				
I	-1	-4	4			
L	-1	-4	2	4		
M	-1	-3	1	2	5	
V	0	-3	3	1	1	4

(r: I arrived from the left neighboring cell, d: I arrived from the upper neighboring cell, rd: I arrived diagonally)

		G	M	L	V	A	I
	0	-1r	-2r	-3r	-4r	-5r	-6r
G	-1d	**6rd**	**5r**	4r	3r	2r	1r
L	-2d	5d	8rd	**9rd**	8r	7r	6r
V	-3d	4d	7d	9rd	**13rd**	**12r**	11r
V	-4d	3d	6d	8d	13rd	13rd	**15rd**

Score 15, alignment
GMLVAI
G_LV_V

9. **Calculate the log odds ratio of residues Glutamine (Q) and Tyrosine (Y) from the following multiple alignment (4p):**

A V Q T M Y
Q Y Y H M D
A M S H M Q
N V H Y M D

<f(Y)=4, f(Q)=3, f(Y Q)=2, M(Y/Q)= log(f(Y Q)/(f(y)*f(Q)))=log(2/(3*4))=log(1/6)= -0.78>

10. Order the steps of making a multiple alignment from pairwise alignments using an iterative approach (5p).

1	2	3	4	5
e	**b**	**d**	**a**	**c**

a) Construct a guide tree from the matrix containing the pairwise comparison values, using a clustering algorithm
b) Record the resulting similarity scores
c) Perform multiple alignment in growing order of scores; the most similar are aligned first, the others are saved for later
d) From this, calculate how sequences are related to each other (the more similar [pairs with high scores] are easier to align)
e) Compare every single sequence to every other sequence, using pairwise sequence alignment

11. Highlight the difference between these database types. (3p) Write an example next to 3 of them. (3p)

primary vs. secondary
full sequences(GeneBank) derived entities, like domain sequences (PFAM,PDB,RefSec)

raw vs. annotated
only sequence, some predicted function(trEMBL) provided with additional info (UniProtKB)

comprehensive vs. specialized
all species species specific

12. Fill the table about the different data description methods! Decide if the following data types are structured, unstructured or composition-type descriptions! For structured descriptions define the entities and relationships! (8p)

protein sequence, cysteine containing, H_2O, protein multiple alignment, phylogenic tree of a group

structured			unstructured	composition-type description
	entity	**relationship**		
protein sequence	amino acids	sequence vicinity		
			cysteine containing	
				H_2O
protein multiple alignment	position of the amino acids	sequential vicinity		
phylogenic tree of a group	species/sequences	evolutionary relations		

13. Write an example of a nucleotide sequence in concatenated FASTA format. (1p)

>first sequence
cggctatcgatcgatcgatcgcatatttatataagctagctcgatcgctagctagttagagatcgatagactag
>second sequence
actagatcgatagctatagctagatcgatcgatcattagctatactcgatactcgatcgatcagta

14. Fill in the missing words/expressions! (4p)

Metagenomics based on 16SrRNA consists of the following main steps:
a) _____ amplification of reference gene with general primers
b) Clustering the reads into groups, these groups are called: _____ c) Building a _____ from the clusters
d) Identifying known and unknown nodes
e) Determining the _____ based on the number of reads in a group.

PCR, Operational Taxonomic Units (OTUs), phylogenetic tree, metagenome compositions

15. Fill in the missing words/expressions! (3p)

A sequence group is a cluster, connected by significant similarities, and it can be described with a _____, a _____ or a _____. (write at least three)

multiple alignment, a common motif, a tree, a frequency matrix, graf