Summer 2023 Name:

BIN – Bioinformatics – 7.9.2023

Q1	Q2	Q3	Q4	Q5	Exam (50)	Labs (50)	Total (100)

Instructions: The written exam takes 120 minutes. Answer directly below the questions, use free sheets only when necessary. You can use the front page as well. Be as detailed as possible, answer in a structured way rather than in free text.

Question 1 *(10 points) Sequence assembly*

Explain the basic principles of DNA sequencing and subsequent sequence assembly by answering the questions below.

(a) (2 points) Explain the terms and connections between them: DNA fragment, read, contig, gap, scaffolding, coverage.

(b) (2 points) Name the basic methods of DNA sequencing and compare them in terms of speed, cost, error rate and read length.

(c) (2 points) Explain why in vitro DNA sequencing methods cannot read the entire DNA sequence of a human cell at once.

(d) (2 points) Define the problem of assembling a DNA sequence biologically and then as a mathematical problem. Name the difference. Put this problem in the correct complexity class, explain what is the key parameter that the complexity estimate is based on.

(e) (2 points) Suggest a straightforward heuristic solution to the above mathematical problem based on greedy search. Estimate its complexity.

Question 2 *(10 points) Markov sequential models*

There is a set of genomic sequences {ACGGAGA, CGTTGACA, ACTGAA, CCGTTCAC}. Your task is to build a profile hidden Markov model (HMM) for this set.

(a) (2 points) Explain the purpose of profile HMM. Explain in math expressions, demonstrate on an example. Name at least two ways of its utilization.

(b) (1 point) Define the problem of profile HMM learning formally (what distribution do we learn?, what is the learning criterion?).

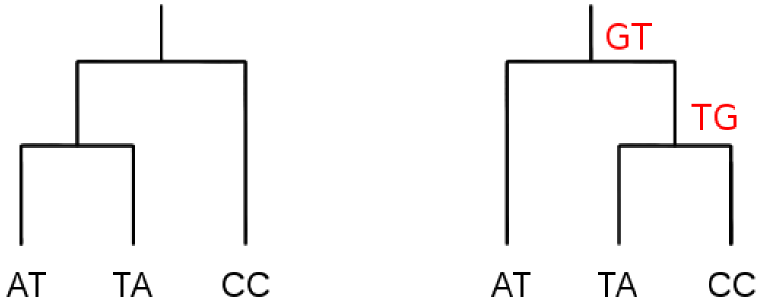
(c) (2 points) Compare two distinct ways of profile HMM learning. The first way starts with multiple sequence alignment, the second approach omits this step. Name advantages and disadvantages of both the approaches. Which approach is more common?

(d) (3 points) Describe the method of progressive tree alignment (e.g., the algorithm CLUSTALW). Outline its application on the given set of 4 sequences. Consider the trivial alignment score which counts +1 for match in a pair of symbols and -1 for mismatch/gap insertion. Describe the principle, the detailed alignment would be too difficult.

(e) (2 points) Draw a profile HMM whose structure matches the multiple sequence alignment that originated in the previous step. Let the alignment be {ACG–GAGA, -CGTTGACA, AC-T-GA-A, CCGTTCAC-}.

Question 3 *(10 points) Phylogenetic trees*

Based on the parsimony method, decide whether a left or right phylogenetic tree is more appropriate for a given triplet of sequences. Part of the solution is finding the optimal sequences for the internal nodes of the tree on the left, for the tree on the right these sequences have already been found. Assume that the sequences are aligned and consider independence between residues, i.e. positions in the sequences. Rate substitutions between nucleotides according to the price matrix below. Points will be awarded for:



	A	C	G	T
A	0	0.8	0.2	0.9
C	0.8	0	0.6	0.5
G	0.2	0.6	0	0.1
T	0.9	0.5	0.1	0

(a) (4 points) use of the weighted parsimony algorithm,

(b) (4 points) adding sequences to the internal nodes of the tree on the left,

(c) (2 points) evaluating both trees and deciding which one is better.

Question 4 *(10 points) Gene expression*

Discuss the issues of RNA sequencing and gene expression quantification.

(a) (2 points) Explain what DNA sequencing and RNA sequencing have in common and different.

(b) (2 points) What is the typical output of RNA sequencing and what is it for us?

(c) (2 points) What is the Poisson distribution and the negative binomial distribution and what are they generally used for?

(d) (1 point) How do the above two distributions apply to the processing of RNA sequencing output?

(e) (3 points) Define a generalized linear model working with a Poisson distribution that will allow to decide on the differential expression of a transcript between a group of healthy and diseased people. Consider the presence of another explanatory variable, patient age.

Question 5 *(10 points) RNA secondary structure*

You are tasked with modeling the secondary structure of ribonucleic acid. Discuss the following questions.

(a) (2 points) Compare the chain structure of RNA and DNA. Name the differences and try to explain what they result from.

(b) (2 points) What is the secondary structure of RNA? Why is it important in RNA and what is it used for? What other levels of RNA description do you know?

(c) (2 points) What type of grammar can be used to describe the secondary structure of RNA and under what conditions? Explain, give an example of grammar.

(d) (2 points) Describe how free energy minimization is used to predict RNA secondary structure.

(e) (2 points) Compare the procedure described above with the Nussinov algorithm based on dynamic programming (assumptions, complexity, success).