Summer 2023 Name:

BIN – Bioinformatics – 22.6.2023

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.

Otázka 1 *(10 points) Sequence alignment.*

Align the following pair of sequences: TTAG and AAGA.

(a) (3 points) First, do a global alignment. Use dynamic programming (Needleman-Wunsch algorithm). Consider a linear penalty for a gap in the alignment (-2 per position), simplify the substitution matrix to score +1 for a symbol match at a position and -1 for a mismatch.

(b) (1 point) How many solutions are there? List them.

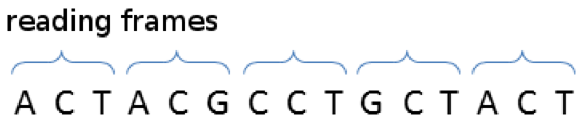
(c) (2 points) Does the alignment change somehow if we are not looking for global but local alignment? Indicate the algorithm change (Smith-Waterman) and perform alignment.

(d) (2 points) Explain how the linear penalty function used so far is unrealistic. Describe the enhancements to the penalty function used and explain the implications in terms of the complexity of the alignment task.

(e) (2 points) Give an example of an affine gap penalty for which the global alignment found ad a) will change. Leave the other alignment parameters unchanged. What will the new alignment be?

Otázka 2 *(10 points) Markov sequential models*

We will construct a 5th order inhomogeneous Markov chain for the sequence shown below while working with a reading frame and codons of length 3.



(a) (3 points) Draw the simple version of the model without regularization.

(b) (1 point) What is its disadvantage?

(c) (3 points) Outline the structure of the regularized model (you do not need to enumerate all the probabilities, show just a few of them).

(d) (3 points) Discuss the number of states and edges the model will have in both regularized and non-regularized versions.

Otázka 3 *(10 points) Phylogenetic trees*

Consider unrooted phylogenetic trees formed by the maximum parsimony method. Assume that you have an algorithm that evaluates the parsimony of a particular candidate phylogenetic tree (e.g., the Sankoff-Cedergren weighted parsimony algorithm). Your task is to design an algorithm that finds the optimal phylogenetic tree topology for a given set of organisms of size n.

(a) (2 points) Verbally define an unrooted phylogenetic tree. Draw an example and a counterexample.

(b) (2 points) How many different unrooted trees are there in the given task? For what sizes of n can this set be searched exhaustively?

(c) (2 points) Use pseudocode to describe at least one method of heuristic search of a set of unrooted trees (recommended method: nearest neighbor interchange).

(d) (3 points) Describe in pseudocode at least one method of searching a set of unrooted trees that guarantees finding the optimal tree and is more efficient than exhaustive search (recommended method: branch and bound).

(e) (1 point) What determines the effectiveness of the method chosen ad d?

Otázka 4 *(10 points) Gene Ontology*

Your task is to perform *enrichment analysis* of gene expression data using *Gene Ontology (GO)*.

(a) (2 points) Explain the term gene ontology. What is its purpose? What are its 3 basic domains? What types of relationships does it define between the concepts contained in the ontology?

(b) (1 point) What are annotations in gene ontology and how are they created?

(c) (2 points) What is gene ontology-based enrichment analysis? What are its basic steps? What can it help us with compared to the usual analysis of gene expression data?

(d) (2 points) Define in more detail the enrichment analysis method called over representation analysis (ORA). Which statistical test can be applied in it?

(e) (2 points) Define gene set enrichment analysis (GSEA) in more detail. The description does not have to be completely formal, however, the explanation of the concept of enrichment score and the way of its statistical evaluation is important.

(f) (1 point) Explain the principal difference between the two types of enrichment analysis described above. What are their advantages and disadvantages?

Otázka 5 *(10 points) Searching for sequence motifs*

Your task is to search for sequence motifs in the specified set of sequences:

s_1:	C	T	G	C	A	G
s_2:	G	A	T	A	C	G
s_3:	G	G	C	T	A	A

(a) (1 point) Define what a sequence motif is.

(b) (1 point) Explain how a motif can be formally described. Consider both deterministic and statistical motifs.

(c) (2 points) Find the strongest motif of length 3 in the above-defined set of sequences. Specify the information content per motif symbol and write the motif logo (consensus logo = information content logo).

(d) (1 point) Explain the general principle of Gibbs sampling.

(e) (3 points) Describe how you would use Gibbs sampling to search for sequence motifs. Explain the basic concepts, give the pseudocode of the method.

(f) (1 point) Explain when you can stop sampling in the given task.

(g) (1 point) What is the advantage and disadvantage of Gibbs sampling in motif search compared to the EM algorithm?