Summer 2017 Name:

a6m33bin, b4m36bin – Bioinformatics – 29.5.2017

Q1	Q2	Q3	Q4	Q5	Total (50)

Instructions: A 90 minutes test. A detailed and structured answers welcome. No literature allowed.

Question 1 *(10 points) Sequence alignment*

(a) (1 point) What is the purpose of substitution matrix in sequence alignment?

(b) (3 points) What is the meaning of the individual entries in the matrix (a formal definition or formula needed here)?

(c) (3 points) How would you construct such a matrix (write a short pseudocode)?

(d) (2 points) Why there are more substitution matrices for the same set of amino acids (e.g., BLOSUM62, BLOSUM80, BLOSUM45)?

(e) (1 point) Which one is the right one?

Question 2 *(10 points) Sequence assembly*

You just made the first step of sequencing by hybridization. You obtained the following set of 3-mers $M = \{TGA, GTG, GGT, CGG, GGA, AGG\}$.

(a) (4 points) Name an efficient method suitable for subsequent sequence assembly. What is the underlying graph theory behind the method (explain the terms and their relationship)? What is the time complexity of the method?

(b) (2 points) Decide whether M can make a spectrum of a sequence. In other words, determine whether it can be a complete set of k-mers appearing in a certain sequence. Justify formally.

(c) (2 points) If the previous answer says yes, construct such a sequence. If no, provide the least extension of M to make the answer positive. Construct the sequence.

(d) (2 points) How many solutions does the previous task have? If you did not provide them in the previous answer, extend it.

Question 3 *(10 points) Phylogenetic trees*

There are four biological species defined by the following set of aligned characteristic sequences {AAA,CCC,CCT,AGC}. The main task is to construct a phylogenetic tree. Work with a simplified evolutionary model in which the distance between a pair of species equals the number of mismatches in their sequences.

(a) (3 points) Enumerate the UPGMA assumptions. Decide whether the UPGMA assumptions are met in this particular task. What would be the consequences of violation of these assumptions?

(b) (4 points) Provided that UPGMA will construct a consistent phylogenetic tree, construct the tree. If not, change the smallest number of symbols in the sequences to meet the assumptions and construct the tree. Do not forget about the time axis.

(c) (3 points) Comment the resulting tree, pay attention to the internal nodes of the tree.

Question 4 *(10 points) Statistical microarray analysis*

(a) (2 point) Explain the design and purpose of microarrays.

(b) (2 points) What is differential gene expression and how do we detect it?.

(c) (2 points) What is multiple comparison problem and how does it relate to microarray statistical analysis? Give an example.

(d) (2 points) Describe the solution to the above mentioned problem through family-wise error rate and false discovery rate? What is the principal difference between them?

(e) (2 points) Name and explain at least one of the particular methods for multiple comparison adjustment.

Question 5 *(10 points) Structural bioinformatics*

Describe the main idea and assumptions of the *threading* method for protein structure prediction. Define the state space being searched in this method. Characterize the objective function being minimized while searching (the exact formula is not needed here, it is sufficient to name and explain its additive components).