Name, SURNAME and ID ⇒

Middle East Technical University
Department of Computer Engineering

CENG 734

Advanced Topics in Bioinformatics

Fall '2006-2007

Final Exam

- **Duration: 100** minutes.
- **Exam:**
  - This is a **closed book**, **closed notes** exam. The use of any reference material is strictly forbidden.
- **About the exam questions:**
  - The points assigned for each question are shown in parenthesis next to the question.
  - For *True-False* type questions, put your results in the boxes provided.
- **This exam consists of *4* pages including this page. Check that you have them all!**
- **GOOD LUCK !**

Question 1	
Question 2	
Question 3	
Total ⇒	

1 **(20 pts)**

For the following 10 statements, indicate whether the statement is *true* or *false* by marking the corresponding box with **T** or **F**, respectively (2 points each).

i. Finding the optimal solution to a multiple alignment of n sequences has an exponential running time complexity.

ii. In *iterative* approaches to the multiple sequence alignment problem, you start with a pairwise alignment and add new sequences to the intermediate alignment until all the sequences are aligned.

iii. Prediction of β-sheets in secondary structure prediction is difficult because of non-local interactions.

iv. Secondary structure prediction is a difficult problem even when we know the 3D (i.e., tertiary) structure of the protein.

v. Optimal structural alignment of two proteins has the same running time complexity as the running time complexity of optimal sequence alignment.

vi. Multiple structural alignment can be used to detect conserved substructures in a protein family.

vii. In a microarray experiment, spots with missing values are due to experimentation errors.

viii. Hidden Markov models are probabilistic models.

ix. In order to use support vector machines for a classification problem, you have to represent your objects as numerical vectors.

x. A protein interaction network is a graph that encodes proteins as nodes. An edge between two nodes describes the amount of sequence similarity between the respective protein sequences.

2 **(40 pts)**

You are given a weighted protein network of *H. sapiens* which contains 10000 proteins and 50000 nodes. The weight on an edge denotes the confidence that the two proteins interact and it is between 0 and 1. Your job is to analyze this network and prepare a report to a biologist. What would you do to analyze this network? What type of analysis tools you would use? How would you visualize it? What statistical measures for nodes/edges would you compute? Do you think the analysis report you prepare is useful for the biologist?

3 **(40 pts)**

Imagine you are working in a research lab that designs drugs. Your boss tells you that there is a new disease discovered in a rural area that effects sheep. He wants you to find a cure to this disease as soon as possible using the related bioinformatics techniques you have learned this semester. You will start working on the problem as soon as the tissues from the sick sheep are delivered to the lab. Describe how you will proceed step by step, which techniques you will use for what purposes. Assume that you have all the technical resources and an adequate number of personnel that can assist you. Feel free to comment on anything that may be related to this question.