Name, SURNAME and ID ⇒

Middle East Technical University
Department of Computer Engineering

CENG 465

Introduction to Bioinformatics

Spring '2008-2009

Midterm Exam

- **Duration: 120** minutes.
- **Exam:**
  - This is a **closed book**, **closed notes** exam. The use of any reference material is strictly forbidden.
  - No attempts of cheating will be tolerated. In case such attempts are observed, the students who took part in the act will be prosecuted.
- **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 *8* pages including this page. Check that you have them all! GOOD LUCK !**

Question 1	
Question 2	
Question 3	
Question 4	
Question 5	
Question 6	
Question 7	
Total ⇒	

1 (16 pts)

□

For the following 8 statements, indicate whether the statement is *true* or *false* by marking the corresponding box with **T** or **F**, respectively (2 points each).

- The lower the *z-score* the more statistically significant is the observed outcome. □

- The length of a pairwise alignment (i.e., the total length of match, mismatch, insertion, and deletion columns) cannot be greater than the length of the longer sequence. □

- The number of children of any node in a suffix tree cannot be greater than the number of characters in the respective alphabet including the $ character. □

- A BLAST query may possibly miss some biologically significant alignments between the query sequence and the database sequences. □

- In global (i.e., no free terminal gaps) pairwise alignment with dynamic programming, the highest value in the partial scores table is always at the lower right corner of the table. □

- A Hidden Markov Model (HMM) may have more than one start state. □

- One can use the dynamic programming pairwise alignment algorithm to align a *sequence* against a *profile* with a simple modification of the scoring function. However, alignment of two *profile*s is not possible using dynamic programming. □

- It is not possible to construct a hidden Markov model for a set of protein sequences without performing a multiple alignment of those sequences first. □

2 **(14 pts)**

Provide a 1-2 sentence description for each of the following terms.

(a)(3 pts) Orthologous proteins

(b)(3 pts) Semi-global pairwise alignment

(c)(4 pts) E-value of a BLAST hit

(d)(4 pts) Secondary structure (in the context of proteins)

3 **(15 pts)**

(a)(8 pts) Fill out the dynamic programming table for determining the optimum **local alignment** between the DNA sequences GGACTA and AAGGC. Assume that a match is scored +3 and that mismatches and gaps are penalized -1 each.

	-	G	G	A	C	T	A
-							
A							
A							
G							
G							
C							

(b)(7 pts) What is the optimum local alignment corresponding to the table in part (a) and what is its score? Show the alignment below and also show a traceback of the alignment on the table in part (a).

$\boxed{4}$ **(10 pts)**

Using the pigeon hole principle one can state that if there are 49 students in a class, it is certain that there will be at least one group of 5 students born on the same month of the year. In other words the probability of observing 5 students having the same birth month in a group of 49 students is 1. What about the E-value? Is it 1?

Compute the expected value (E-value) of this random event. In other words, in a class of 49 students, how many distinct, but possibly overlapping, groups of 5 same birth month students is expected to be observed? You do not need to find the exact number, writing down to correct mathematical expression is enough.

5 **(15 pts)**

(a)(10 pts) Construct a suffix tree for the following string: **accacgcg$**. Show the individual steps of construction.

(b)(5 pts) Suffix trees are suitable for exact matches. However, describe shortly how the profile ACc/g (i.e., A and C in the first two positions and C or G in the last position) could be searched in a suffix tree. You do not need to do the actual search in the tree in part (a). Just describe how it could be done.

6 (15 pts)

Consider the multiple sequence alignment of 5 DNA sequences given below:

```
ACG-AT
ACA-AT
TCAGAT
TCGTAT
GCG-AA
```

Draw a profile hidden Markov model for these five sequences. You may use any number of match/insert/delete states you want. You may omit some states of the HMM if you believe that they are not required for the sequences given above. Give the emission probabilities at each state and the transition probabilities between every state.

Notes: A column in the MSA is considered a match state if the majority of the rows in that column is non-gap. **Do not** use pseudocounts when computing the emission or transition probabilities.

7 **(15 pts)**

(a)(8 pts) Consider the greedy approach for multiple sequence alignment in which the multiple alignment is built in $k-1$ steps combining two alignments at each step. Compare the greedy approach to the **star** alignment approach in terms of accuracy and running time. Justify your answers. You do not need to give the actual time complexity expressions.

(b)(7 pts) Why does the ClustalW technique generally generate better multiple alignments compared to the star alignment technique?