

Name, SURNAME and ID ⇒

KEY

● 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 *profiles* 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.

Incorrect student answers marked with a X



2 (14 pts)



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

(a)(3 pts) Orthologous proteins

Orthologous proteins are proteins in different species that evolved from a common ancestral gene by speciation. Orthologs usually retain the same function in the course of evolution.

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

Alignment where terminal gaps are not penalized. (a.k.a. "free end gaps").

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

Expected number of sequences in the database that would produce the same or better alignment score by chance.

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

Regular structural patterns of local segments, usually formed by Hydrogen bonds. Secondary structure elements are classified into: Alpha Helix, Beta Sheet, and Turns and Loops.

Correct answers are marked with ✓

Incorrect " " " " " X

Partially correct answers are given their partial grades, e.g. (2)

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.

\* missing internal zeros  
→ (-3)

	-	G	G	A	C	T	A
-	0	0	0	0	0	0	0
A	0	0	0	3	2	1	3
A	0	0	0	3	2	1	4
G	0	3	3	2	2	1	3
G	0	3	6	5	4	3	2
C	0	2	5	5	8	7	6

(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).

G G A C } Alignment score is 8  
G G - C

\* You should specifically show the local alignment, not alignment of the whole sequences. (-2)

\* You should specify the alignment score. (-2)  
 ↳ if global alignment is given (-1)  
 ↳ if local alignment is correct (-1)



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.

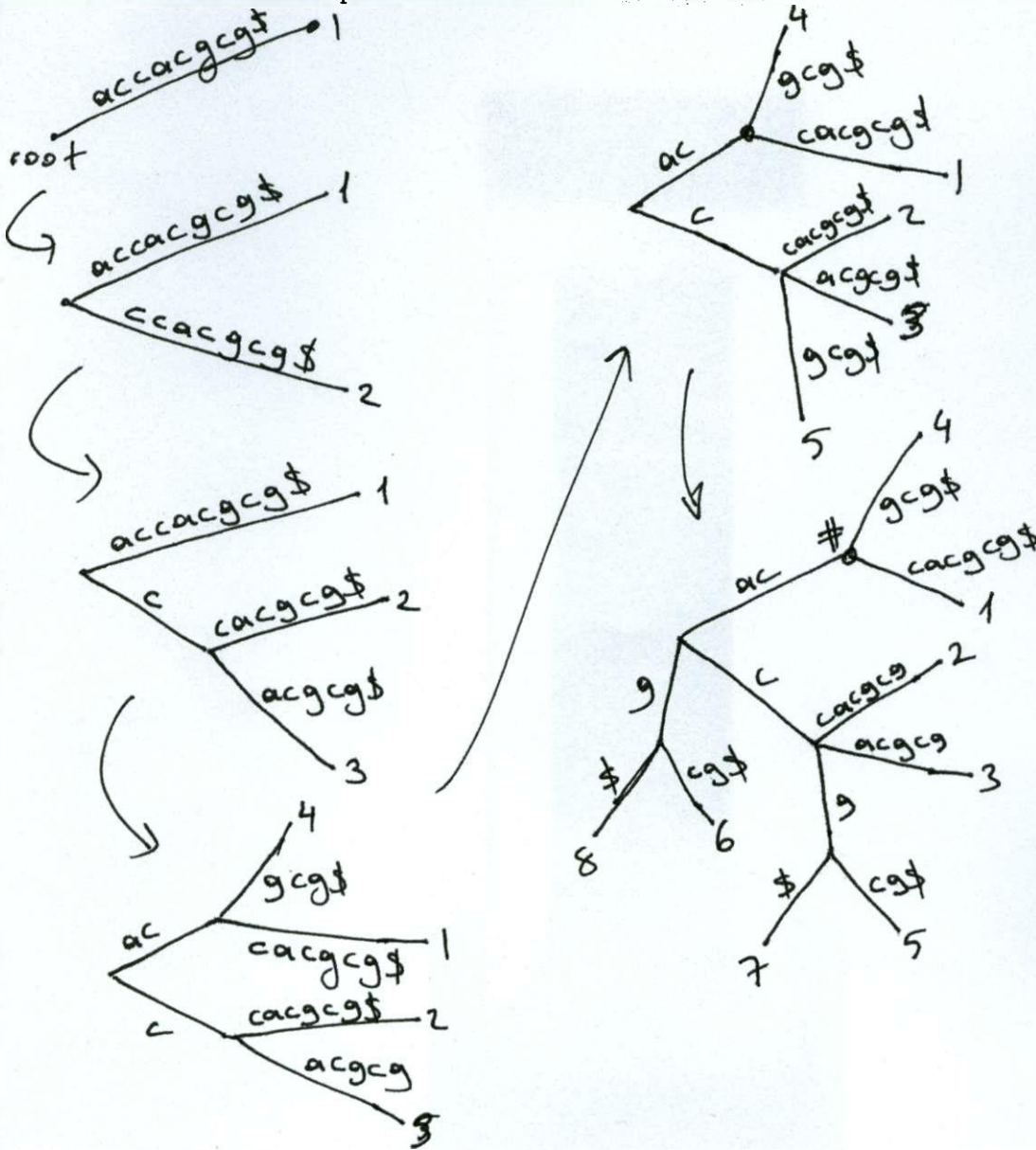
$$\text{Probability of 5 students being born in the same month} = \frac{1}{12} \cdot \frac{1}{12} \cdot \frac{1}{12} \cdot \frac{1}{12} = \left(\frac{1}{12}\right)^4$$

$$\text{Number of different groups of 5 students among 49 students} = \binom{49}{5}$$

$$\begin{aligned} \Rightarrow \text{E-value} &= \left(\frac{1}{12}\right)^4 \cdot \binom{49}{5} \\ &= \left(\frac{1}{12}\right)^4 \cdot \frac{49!}{44! 5!} \end{aligned}$$



(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.

The "exact" part of query is searched as usual. For inexact characters, we follow each path that match each of these alternative characters.

→ Search "AC" and from the resulting node (marked with # above) search for both g & c, matching edges leading to leaf nodes 1 and 4.

(→ Note that the search for exact part may result in end) (within an edge label, which can be handled similarly.)



6 (15 pts)

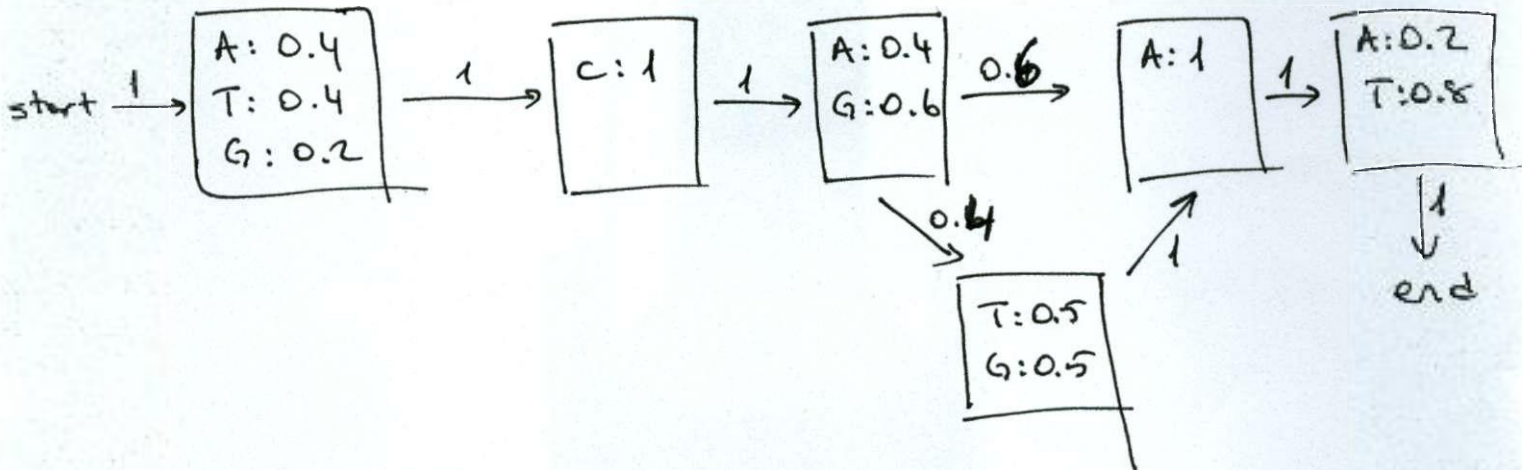


Consider the multiple sequence alignment of 5 DNA sequences given below:

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.

Greedy approach has a higher running time, because at each iteration an all-to-all pairwise alignment is run.

Greedy approach is expected to have higher accuracy because "star" alignment uses a single central sequence, whereas the greedy approach finds the most similar sequences at each step. If we had two distinct sequence groups in the input, "star" alignments "single center" or "single reference" strategy will lead to erroneous alignments for which the central sequence does not belong to.

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

Because ClustalW makes use of natural groupings in the input sequences, whereas star alignment assumes a uniform group around a central sequence which may be incorrect for many input sequences.