Name, SURNAME and ID ⇒ SOLUTIONS

Middle East Technical University, Northern Cyprus Campus
Computer Engineering Program

CNG 465

Introduction to Bioinformatics

Spring '2012-2013

Midterm Exam

- **Duration: 120** 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.
- **This exam consists of *6* pages including this page. Check that you have them all! GOOD LUCK !**

Question 1	20
Question 2	30
Question 3	10
Question 4	40
Total ⇒	100

1 (20 pts)

20

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

(a)(4 pts) Dynamic programming

An algorithmic technique where the solution can be written recursively and smaller solutions stored in a table.

(b)(4 pts) Position Specific Scoring Matrix

A 20xn matrix for a profile of length n which shows the score of aligning a specific amino acid type for each column of the profile separately.

(c)(4 pts) Statistical significance

How likely the solution is different from a random event.

(d)(4 pts) Homolog

Two sequences that are evolved from the same ancestor.

(e)(4 pts) Genome

The collection of all genes of an organism.

2 **(30 pts)**

30

Given the following partial scores table for the **local alignment** of two protein sequences, answer the following questions. Assume that a scoring matrix such as BLOSUM is used. In other words, match scores and mismatch penalties of different types of amino acids are different. However, assume that all match scores are positive and all mismatch and gap penalties are negative.

	-	N	P	A	G	D	E	M
-	0	0	0	0	0	0	0	0
P	0	0	6	2	0	0	0	0
G	0	1	2	4	7	3	0	0
D	0	0	0	1	2	13	9	5
V	0	0	0	0	0	9	9	5
E	0	0	0	0	0	5	10	6

(a)(15 pts) What are the match/mismatch scores for the following amino acid pairs? If the data in the table is not sufficient to determine a value, please indicate it by writing *insufficient data* next to the corresponding item. You do not need to show your reasoning. Just write the values. (3 pts each)

- S(G,A) −2 (6→4)
- S(V,E) −4 (13→9)
- S(P,G) Insufficient data
- S(D,A) −1 (2→1)
- S(G,G) 5 (2→7)

(b)(4 pts) What is the gap penalty? Does this alignment use a *linear* or an *affine* gap penalty model?

−4. Linear gap penalty is used.

13 → 9 → 5 shows linear penalty.

(c)(11 pts) What is the score of the optimum local alignment? Show the best local alignment below and also show a traceback of the alignment path on the partial scores table above.

13. Shown on the table.

P A G D

P − G D

↘ including the last part of seqs in the alignment

−5

3 **(10 pts)**

10

Suppose that we have a modified deck of cards in which we only have the cards with face values up to 5. We have all the four houses of cards, which results in a total of $5 \times 4 = 20$ cards in our deck. In other words, we have: 4 aces (i.e., ones), 4 twos, 4 threes, 4 fours, and 4 fives. Suppose that we draw 10 cards randomly one by one from this deck by replacement, i.e., after we draw a card and note its value, we put the card back into the deck. What is the E-value of observing two cards of the same face value (e.g., an ace of diamonds and an ace of spades) one after another in this draw of 10 cards. E.g., *ace* after another *ace*, or *two* after *two*, etc.

-5 pts

Probability of two cards having same face value $= \frac{1}{5}$

# of cases the event can be observed $= 9$

$E = p \cdot n = \frac{1}{5} \cdot 9 = \frac{9}{5}$

10/15 -2 pts

4 (40 pts)

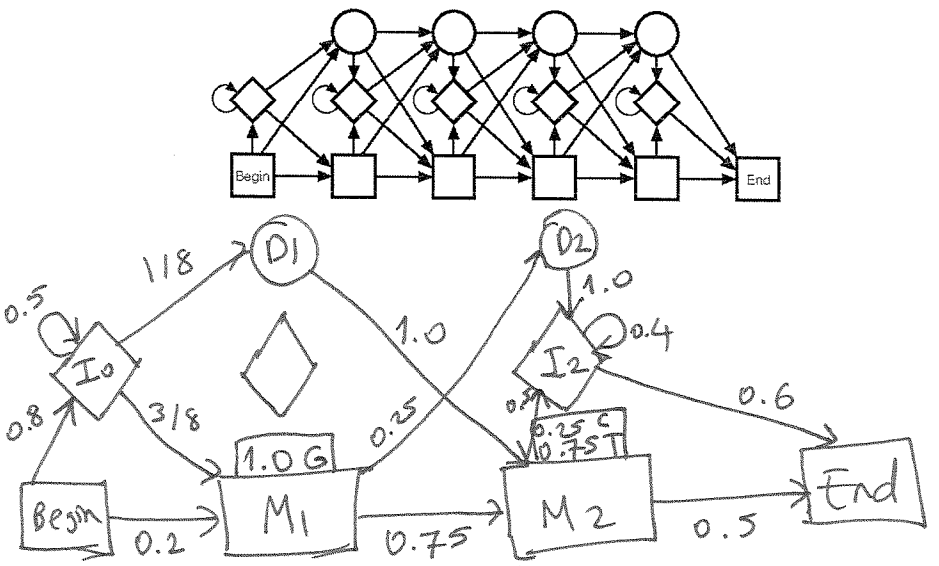
40

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

----G-A- → part c
-GAT-T---
T---GTG--
TG-TG-GAA
--A-GC---
----GTC--
**

(a)(20 pts) Draw a profile hidden Markov model for these five sequences using the profile HMM structure given below. The columns of the alignment that correspond to match states are shown with an $*$ at the bottom row. You may omit some states of the original profile HMM, if they are not visited by the sequences given above. Show only the observed transitions between states. Give the emission probabilities at match states and the transition probabilities between relevant states in your HMM.

Notes: Do not use pseudocounts when computing the emission or transition probabilities. You do not need to write the emission probabilities for insertion states.



seq1 Begin → I0 → I0 → I0 → D1 → M2 → End
seq2 Begin → I0 → M1 → M2 → I2 → End
seq3 Begin → I0 → I0 → I0 → M1 → D2 → I2 → I2 → I2 → End
seq4 Begin → I0 → M1 → M2 → End
seq5 Begin → M1 → M2 → I2 → End

(b)(5 pts) What are the sequence of states of the HMM that are visited to emit the sequence, "TGTGGAA", in the alignment. (Note: do not use the Viterbi algorithm for this. This information is contained in the alignment.)

Begin → I_0 → I_0 → I_0 → M_1 → D_2 → I_2 → I_2 → I_2 → End

(c)(15 pts) Using the Viterbi algorithm determine the most likely sequence of states that could have generated the sequence, "GA". Show the partial probability table. What is the probability of the best path?

When computing probabilities do NOT use "log" of the probabilities. During multiplications you may use scientific notation and round to 2 digits after the decimal point to represent small numbers, e.g., 0.34E-6 or 0.21E-2 to indicate 0.000000342321 and 0.00206784, respectively. Use the rounded numbers in the subsequent computations to make your computations easier.

	–	G	A
Begin	1.0	0.0	0.0
I_0	0.0	0.8·0.25 = 0.2	0.2·0.5·0.25 = 0.025
I_2	0.0	0.0	0.05·1.0·0.25 = 0.0125
D_1	0.0	0.2·1/8 = 0.025	0.025·1/8 = 0.003125
D_2	0.0	0.2·025 = 0.05	0.0
M_1	0.0	0.2	0.0 →
M_2	0.0	0.0	0.0 →
End	0.0	0.0	$I_2(A)$·0.6 = 0.0125·0.6 = 0.0075

Because of emission probabilities

Begin → M_1 → D_2 → I_2 → End

↓ Best path.

Best path