Department of Biosciences and Bioengineering

Bioinformatics: BT 302

Mid Semester Examination (18 September 2024)

[Duration: 2 hrs] **Answer All the Questions** [Total marks: 40]

Note:

- **All problems should be worked out in a systematic manner explaining the steps involved.**
- **Answers should be short and up to the point. Ambiguous and circuitous answers will attract penalty.**
- **Please return both the question and the answerscripts after the exam**

1)

(a) Calculate the compositional complexity (K) for the following nucleotide sequences: (i) GCGACT and (ii) TATATA. Which of these shows low complexity based on K? [2 marks]

(b) In BLAST, two sequences A (250 residues) and B (5×10^7 residues) are aligned with an E-value=0.05. Calculate the normalized score (S_b) in bits. [2 marks]

(c) From the normalized score calculated above, determine the raw score while using the BLOSUM 45 matrix with $\lambda = 0.203$ and $K = 0.041$. [2 marks]

(d) For the same alignment scored using BLOSUM 45 and BLOSUM 62, the scores are as follows: S_{B45}=650 bits and S_{B62}=619 bits. What do you infer from these scores about the scoring scheme? [2 marks]

2)

	123456
1	GTCKQT
2	ATCRNM
3	TSCRNA
4	SACENG
5	SDCEQA
6	SECENL

For the multiple sequence alignment shown on the left, clustering based on 50% sequence identity produces three clusters and each of these clusters are shown enclosed by a box. Calculate the BLOSUM log odds score in bits for A to T substitution ($S_{A,T}$) . [5 marks]

3) (a) While the PAM substitution matrix is symmetric, the mutation probability matrix is not. Why? [2 marks]

(b) Calculate $\mathbf{M}_{A,A}$ for the following alignment, where M is the mutation probability matrix: [3 marks]

EDGE
ANGE

(c) Name two substitution matrices from the following that will be most suitable for aligning distantly related sequences? Justify.
(i) BLOSUM30, (ii) BLOSUM45, (iii) BLOSUM62, (iv) BLOSUM80, (v) PAM30, (vi) PAM70, (vii) PAM100 and (viii) PAM250 [2 marks]

4)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	B	Z
C	12																					
S	0	2																				
T	-2	1	3																			
P	-3	1	0	6																		
A	-2	1	1	1	2																	
G	-3	1	0	-1	1	5																
N	-4	1	0	-1	0	0	2															
D	-5	0	0	-1	1	2	2	4														
E	-5	0	0	-1	0	0	1	3	4													
Q	-5	-1	-1	0	0	-1	1	2	2	4												
H	-3	-1	-1	0	-1	-2	2	1	1	3	6											
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6										
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5									
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6								
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5							
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6						
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4					
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9				
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10			
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17		
B	-4	0	0	-1	0	0	2	3	2	1	1	-1	1	-2	-2	-3	-2	-5	-3	-5	2	
Z	-5	0	-1	0	0	-1	1	3	3	3	2	0	0	-2	-2	-3	-2	-5	-4	-6	2	3

PAM 250

$K = 0.09; \lambda = 0.229$

Two sequences that are about 250 amino acids long are aligned by the Smith-Waterman local alignment algorithm using the PAM 250 matrix (log odds score) and a high gap score to omit gaps from the alignment and that the following alignment within the sequences is found:

RKRKRKRK
RKDERKDE

(a) What is the log odds score of the alignment?
[2 marks]

(b) What is the probability that an alignment of two random sequences of the same length could achieve such a score? Is this alignment significant? Justify
[3 marks]

5)

	Column 1	Column 2	Column 3	Column 4	Column 5
Amino acid	Frequency	Frequency	Frequency	Frequency	Frequency
N	0.6	0.1	0.1	0.1	0.1
K	0.05	0.7	0.1	0.05	0.1
C	0.1	0.1	0.5	0.2	0.1
D	0.1	0.1	0.1	0.7	0
E	0	0.1	0	0.1	0.8

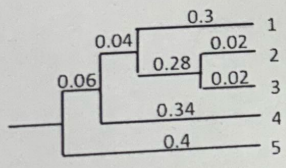
(a) Calculate the log odds score in bits for each table position shown above
[2 marks]

(b) Align the PSSM with the sequence **DENKCDEKD** and calculate the log odds score for the matrix to match that position. Which sequence start position when aligned with the matrix has high log odds score and whether it shows the presence of the conserved motif in the sequence?
[3 marks]

6) Perform the sequence alignment using (i) Needleman-Wunch algorithm for the following sequences:
Sequence 1: **INDIANA** Sequence 2: **INDIGO**.
Use the following scoring scheme – Match = 5; Mismatch = -2; INDEL = -6. The matrix-fill and trace-back should be shown separately. The logic of every step should be clearly mentioned.
[5 marks]

7) (a) In multiple sequence alignment by CLUSTAL-W, E-value is not reported. What is the reason for this? [2 marks]
(b) For aligning the two partial alignments (1-3 and 4-5), the guide tree and the scoring scheme are given below. Calculate the weighted average score as implemented in CLUSTAL-W. [3 marks]

(1) ....**A**...
.
(2) ....**E**...
.
(3) ....**D**...
.



Score	A	D	E	I	V
A		-2	-4	5	8
D			9	-3	-5
E				-7	-5
I					6
V					

(4) ....**I**...
.
(5) ....**V**...
.

→ THE END ←