BT-502

BIOINFORMATICS

Time Allotted: 3 Hours

Full Marks: 70

The questions are of equal value.
The figures in the margin indicate full marks.
Candidates are required to give their answers in their own words as far as practicable.

GROUP A
(Multiple Choice Type Questions)

1. Answer all questions. $10 \times 1 = 10$

(i) What does the term 'LOCUS' explain in GenBank flat file?

(A) accession number
(B) length of molecule
(C) type of molecule
(D) all of these

(ii) The meaning of the E-value in BLAST is

(A) The probability that the query sequence and the subject sequence come from the same organism.
(B) The probability that the query sequence and the subject sequence are homologous.
(C) The expected number of generated sequences that would have the observed alignment (or better)
(D) The inverse of the similarity between the query sequence and the subject sequence.

(iii) Biologically significant similar sequences have

(A) higher E value
(B) lower bit score
(C) higher bit score
(D) lower raw score

(iv) What is the function of ReadSeq program?

(A) reading the sequence
(B) aligning two sequences
(C) construct a tree
(D) convert the sequence one format to another format

(v) In Perl, the name of the scalar variable starts with

(A) @ (B) #
(C) $ (D) %

(vi) Which of the following represent a file opened in write mode?

(A) open(MYFILE, "File1.txt") (B) open(MYFILE,">File1.txt")
(C) open(MYFILE,">>File1.txt") (D) none of these

(vii) Which one of the following is not a structure analysis application of bioinformatics?

(A) nucleic acid structure prediction (B) protein structure prediction
(C) protein structure classification (D) metabolic pathway modeling

(viii) The specificity of a gene prediction program is given by which of the following formulae?

(A) $S_p = TP/(TP + FN)$ (B) $S_p = TP/(TP + FP)$
(C) $S_p = TP/\sqrt{(TP + FN)}$ (D) $S_p = TP \cdot TN - FP \cdot FN$

(ix) PAM is constructed by

(A) Lipman (B) Chou-Fasman
(C) Dayhoff (D) Waterman

(x) A Hidden Markov Model (HMM) has better predictive power than profiles because

(A) its probability modeling is worse
(B) it is able to differentiate between insertion and deletion states
(C) a single gap penalty score determines insertion of deletion
(D) all of these

GROUP B
(Short Answer Type Questions)

Answer any *three* questions. 3×5 = 15

2. Write down organizational structure and three specific mission of NCBI. 2+3

3. State the Lipinski's rule of five. 5

4. Mention the differences i) PAM and BLOSUM ii) Progressive and iterative alignment method. 2×2.5

5. Write down the function of the following terms in PERL with examples: 5
   i) Reverse
   ii) Scalar
   iii) Defined
   iv) Pop
   v) Shift

6. What do you mean by BLOCKs? Calculate Hamming and Levenshtein distance for the given sequence. 1+4
   Seq1: aggat and Seq2: gacat (for Hamming)
   Seq3: agtcc and Seq4: cgctca (for Levenshtein)

GROUP C
(Long Answer Type Questions)

Answer any *three* questions. 3×15 = 45

7. What do you mean by Coding Statistics? Using the definition of coding statistics find out S is coding or non coding? Given that S = AGGACC, C1 = AGG, C2 = ACC F(AGG) = 0.022 and F(ACC) = 0.038 [Hints P(S) = F(CI)·F(C2) where P(S), probability of S codes for protein, $P_0(S)$ = $F_0(C1)·F_0(C2)$ where $P_0(S)$, probability of S having non coding sequence]. Write down the difference between prokaryotic and eukaryotic gene finding approach. Mention the names of any four major method used for gene prediction. 2+8+2+3

[Turn over]

8. (a) Write a program in perl to count the number of residue for A, T, G, C of a given sequence. 5

(b) Write a program to display the complement of the given DNA sequence. 3

(c) Write a program to find out the TATA box and its position in a given sequence. 5

(d) Define Bioperl. 2

9. What is Structure Based Drug Design (SBDD)? Define Scoring function used in docking. What do you mean by agonist, antagonist and Icm50? What is fold recognition method for protein structure prediction? Describe first order energy minimization approach. Give an example of an Molecular Dynamics programme which is widely use for Protein dynamics. 3+2+3+2+4+1

10.(a) What are the stabilizing forces in proteins? What are the regular secondary structures in proteins? Give one example of a super secondary structure. What are the general tertiary structures? Briefly distinguish the nature of the interactions in these tertiary structures. 2+2+3

(b) Define protein secondary structure prediction as *implied in bioinformatics*. What are the specific applications of correct identification of secondary structural elements (SSE's) in bioinformatics? Name 3 algorithms used for protein secondary structure prediction and 1 used for protein tertiary structure prediction. 2+3+3

11.(a) What is sequence alignment? What is midnight zone, twilight zone and safe zone of protein sequence alignment? Define sequence homology, identity and similarity. 2+3+3

(b) Align these two following sequences by Needleman-Wunsch Algorithm given that the score for match = 4, mismatch = 1 and gap = –2. 5+1+1

Seq1: AGGTCTAGGA, Seq2: AGTCTAG. What is the full from of BLAST? What is the relation between raw score and bit score in BLAST?