Name : .....

Roll No. : .....

Invigilator's Signature : .....

CS/B. Tech (BT)/SEM-5/BT-503/2011-12

2011
BIOINFORMATICS-I

Time Allotted : 3 Hours *Full Marks* : 70

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. Choose the correct alternatives for any *ten* of the following : $10 \times 1 = 10$

   i) The main difference between Pfam-*A* and Pfam-*B* is that
      a) Pfam-*A* is manually curated while Pfam-*B* is automatically curated
      b) Pfam-*A* uses HMMs while Pfam-*B* does not
      c) Pfam-*A* incorporates data from SMART and PROSITE while Pfam-*B* does not
      d) Pfam-*A* provides full length protein alignments while Pfam-*B* aligns protein fragments.

   ii) PROSITE is a
      a) database of protein structures
      b) database of interacting proteins
      c) database of protein motifs
      d) search tool.

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

a) its probability modelling is worse

b) it is able to differentiate between insertion and deletion states

c) a single gap penalty score determines insertion or deletion

d) all of these.

iv) Archaea and bacteria have relatively small genomes with size range and gene density as

a) 20 to 30 Mbp 40%

b) 0·5 to 10 MBp 90%

c) 20 to 30 MBp 50%

d) 0·5 to 10 MBp 20%

v) The full form of EMBOSS is

a) European Molecular Biology Open Software Suite

b) Extended Molecular & Bioinformatics Open Software Suite

c) Eastern Molecular Biology Open Software Suite

d) European Molecular Biology Open Service Suite.

vi) Any sub-routine defined in Perl should start with

a) sub-routine b) sub

c) subroutine d) none of these.

vii) In Linux, rm command is used to

a) move files from one location to another

b) delete files

c) display the current directory

d) create a new directory.

viii) Which of the following is a valid scalar variable in PERL ?

a) dna l seq b) dna!seq

c) 1 dnaseq d) ! dnaseq.

ix) Which of the following amino acids is least mutable according to the PAM scoring matrix ?

a) Alanine b) Glutamine

c) Methionine d) Cysteine.

x) A global alignment algorithm (such as Needleman-Wunsch algorithm) is guaranteed to find an optimal alignment. Such an algorithm

a) puts the two proteins being compared into a matrix and finds the optimal score by exhaustively searching every possible combination of alignments.

b) puts the two proteins being compared into a matrix and finds the optimal score by iterative recursions

c) Puts the two proteins being compared into a matrix and finds the optimal alignment by finding optimal sub-paths that define the best alignment(s)

d) can be used for proteins but not for DNA sequences.

xi) You have a DNA sequence. You want to know which protein in the main protein database ("nr", the nonredundant database) is most similar to some protein encoded by your DNA. Which program should you use ?

a) blastn
b) blastp
c) blastx
d) tblastn
e) tblastx

xii) 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.

GROUP – B

(Short Answer Type Questions)

Answer any *three* of the following.

$3 \times 5 = 15$

2. What is sequence alignment ? State its utilities. 2 + 3
3. What is the objective of gene prediction ? What is GENSCAN? Mention the use of 'tr' operator in PERL program. 2 + 1 + 2
4. What is the importance of DOT plot ? Why is sometimes stringency used ? 3 + 2

5. Write notes on any *two* of the following briefly : $2 \times 2\frac{1}{2}$

   a) Taxonomy browser

   b) PMC

   c) COG

   d) Pubmed

6. What is the full form of PERL ? What intrinsic properties of PERL programming language make it suitable for biological sequence analysis tasks ? With what operating system was PERL first used ? 1 + 3 + 1

7. a) Biocomputing tool development is at the foundation of all bioinformatics analysis. Name 4 applications of such tools in the application areas of sequence, structure and function analysis.

   b) Why is motif discovery important for sequence analysis ? 3 + 2

GROUP – C

(Long Answer Type Questions)

Answer any *three* of the following. $3 \times 15 = 45$

8. a) What are the different types of variables used in PERL ?

   b) Mention the use of "tr" operator in PERL program.

   c) Write a program to report how GC rich some sequence is [Hint : Percentage of G and C in the DNA].

   d) With the same sequence write another to find out its complementary sequence. 3 + 2 + 6 + 4

9. a) What is GBFF and mention the role of LOCUS and DEFINITION terms used in the mentioned sequence file.

   b) Mention the steps of BLAST program in sequence alignment. Mention the relationship between *E* and *P* value.

   c) Define Masking and mention role of masking in reducing the signal to noise ratio

(2 + 4) + (4 + 2) + (1 + 2)

10. Write short notes on any *five* of the following : 5 × 3

   a) Pfam

   b) Bio-Perl

   c) EMBOSS

   d) FASTA

   e) PROSITE

   f) Vi-Editor

   g) OMIM

   h) Pubmed

11. a) Diagrammatically represent a second order HMM for prokaryotic gene prediction that includes statistical modelling for start codons, stop codons and the coding region. Explain the diagram.

   b) Why is a correlation coefficient a better index/parameter of gene finding accuracy ? Use quantitative logic in your answer. 8 + 7

12. a) Define motifs and domains in protein sequences.

    b) Why is identification of protein and domains an important part of biological sequence determination ?

    c) What are the specific bioinformatics tools and algorithms upon which identification of motifs/domains are based ?

    d) Use a regular expression representation of a motif to explain exact and fuzzy matching mechanisms of matching a regular of a motif with a query sequence.

3 + 4 + 4 + 4

13. a) Define shell. Mention what shell does from an input file and mention the different types of execution of it.

    b) What does Vi-editor stand for ? Mention briefly the modes of operation of Vi-editor. Mention one alternate editor for UNIX environments. Mention the commands for the following operations :

    i) Move cursor down on line

    ii) Move forward one screen

    iii) Insert text before cursor

    iv) Writing out modified file to file named in original invocation.

(2 + 1 + 3) + (1 + 3 + 1 + 4)