Name: Abhinav

Roll No: 002282C80019

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY KOTTAYAM
Department of Computer Science and Engineering
First Mid-Semester Examination- September 2023

ISC211: Introduction to Bioinformatics

Date&Time: 04-09-2023, 02:30-04:00 PM
Course Instructor: Dr. Manu Madhavan, Dr. Sujamol S

Max Marks: 50
Batch: III Sem CSE

Provide precise and concise answers

PART-A

(Each question carries 2 marks, $2 \times 10 = 20$ marks)

1. Differentiate between DNA and RNA?
2. What is the role of following bio-molecules:(a)tRNA (b) mRNA
3. What is the fundamental difference between a gene and a genome?
4. What do you understand by degeneracy of genetic codes?
5. What are some common types of biological data that bioinformaticians work with?
6. Given the following DNA sequence in upstream, identify the corresponding *reverse complement sequence* and mRNA sequence.
   ATGCCTAGT
7. Write a python code to find the reverse complement of a DNA sequence, *without using any special* packages.
8. What is the role of RNA Polymerise in protein production process?
9. What is the significance of di-ester bonds?
10. How does DNA complementarity play a crucial role in ensuring the accuracy of gene replication

PART-B

(Each question carries 5 marks, $5 \times 4 = 20$ marks)

11. Imagine you are a bioinformatician working in a healthcare setting. A patient presents with a complex and rare genetic disorder. Describe how you would use bioinformatics tools and techniques to assist in diagnosing the patient's condition and developing a personalized treatment plan. Please provide specific examples of data sources, algorithms, and technologies you would utilize, as well as the potential challenges you might encounter in this process.
12. Discuss the steps involved in the flow genetic information explained by central dogma of molecular biology.
13. The $GC-content$ of a set of DNA strings is given by the percentage of symbols in the string that are C or G. For example, the $GC-content$ of $AGCTATAG$ is 37.5%. Note that the reverse complement of any DNA string has the same $GC-content$. Given atmost n DNA sequences, write a program to compute the $GC-content$
14. Given two DNA sequences s and t of equal length, the **Point Mutation** between s and t, denoted $\mu(s,t)$, is the number of corresponding symbols that differ in s and t. For example:
    s : $GAGCCTACTAACGGGAT$
    t : $CATCGTAATGACGGCCT$
    $\mu(s,t) = 7$
    Write a program (in any programming language of your choice) to count the point mutation between two given DNA sequence.

PTO

PART-C

(Answer any one question, $10 \times 1 = 10$ marks)

15. Apply ORF finder algorithm to find the ORFs and the respective translation from the following sequence: 5'-AAATGGCGCGCTGGTGGATTTAGGTAACCACACATGCGCT-3'. Genetic code is given as Figure 1. (10 Marks)

first letter	second letter: U	second letter: C	second letter: A	second letter: G	third letter
U	UUU, UUC: Phe; UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr; UAA stop; UAG stop	UGU, UGC: Cys; UGA stop; UGG Trp	U C A G
C	CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His; CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg	U C A G
A	AUU, AUC, AUA: Ile; AUG: Met	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn; AAA, AAG: Lys	AGU, AGC: Ser; AGA, AGG: Arg	U C A G
G	GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp; GAA, GAG: Glu	GGU, GGC, GGA, GGG: Gly	U C A G

Figure 1: Optimal Alignment matrix

OR

16. Discuss the significance of biological databases in biological research and highlight some key types of biological databases, their contents, and their applications. Explain how these databases facilitate data retrieval, analysis, and integration in various aspects of life sciences, from genomics to structural biology. (10 Marks)

Roll No: 2022BCS0019

Name: Abhinav Bhagwat

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY KOTTAYAM

Department of Computer Science and Engineering

End-Semester Examination- November 2023

ISC211: Introduction to Bioinformatics

Duration: 3 hrs | Max Marks: 100

Course Instructors: Dr. Suja, Dr. Manu | Batch: CSE (2022 Admn)

Answer All Questions

1. (a) State the difference between global and local alignment. **(4 Marks)**

   (b) Explain the working of BLAST based on your knowledge of sequence alignment. **(4 Marks)**

   (c) Can you explain the differences between alpha-helices and beta-sheets in protein secondary structure? **(4 Marks)**

   (d) How do peptide bonds contribute to the formation of the protein secondary structure? **(4 Marks)**

   (e) Explain the concept of scoring matrices for aligning amino acid sequences. Briefly explain how PAM is derived? **(4 Marks)**

2. (a) Using Needleman–Wunsch algorithm, construct the partial alignment score table for the following two sequences, using the scoring parameters: *match score= +5, mismatch* score =-3, gap penalty =-1. What is the optimal *global alignment between these* sequences? Analyse the time and space complexities of the algorithm.

   ACAGTCGAACG

   ACCGTCCG **(12 Marks)**

   (b) A palindromic sequence is a nucleic acid sequence in a double-stranded DNA or RNA molecule whereby reading in a certain direction (e.g. 5' to 3') on one strand is identical to the sequence in the same direction (e.g. 5' to 3') on the complementary strand. Write a program to check whether the given sequence is palindromic or not. Give one example. **(8 Marks)**

3. Answer the following questions in the context of Phylogenetic trees.

   (a) Differentiate between character based and distance based approaches for phylogenetic tree reconstruction **(5 marks)**

   (b) What are the limitations of UPGMA algorithm? How Neighbor-joing approach addresses these issues? **(5 marks)**

   (c) Apply maximum parsimony approach to construct phylogenetic tree for the following set of sequences.

   OTU-1: GGGGGGTGCC

   OTU-2: GGGAGTTCCA

   OTU-3: GGATAGTGGA

   OTU-4: GATCATTACG **(10 marks)**

4. (a) A maximal repeat of a string s is a repeated substring t of s having two occurrences $t1$ and $t2$ such that $t1$ and $t2$ cannot be extended by one symbol in either direction in s and still agree.

For example, "**AG**" is a maximal repeat in "TAGTTAGCGAGA" because even though the first two occurrences of "**AG**" can be extended left into "**TAG**", the first and third occurrences differ on both sides of the repeat; thus, we conclude that "**AG**" is a maximal repeat. Note that "**TAG**" is also a maximal repeat of "TAGTTAGCGAGA", since its only two occurrences do not still match if we extend them in either direction.

Now, given a DNA string s of length at most 1 kbp, write a program that will return list containing all maximal repeats of s. **(10 Marks)**

(b) Apply the maximum base-pair method to compute the secondary structure of the RNA sequence: *GCACGAGGU* **(10 Marks)**

5. (a) Consider the following amino acid sequence. Apply Chou Fasman algorithm to predict the regions of alpha helix and beta strands in the sequence. (The propensity values of amino acids are given in Figure 1.

DMNWHIGMCRCNNTKWCQAT **(10 marks)**

Name	Abbrv	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	A	142	83	66	0.06	0.076	0.035	0.058
Arginine	R	98	93	95	0.07	0.106	0.099	0.085
Aspartic Acid	D	101	54	146	0.147	0.11	0.179	0.081
Asparagine	N	67	89	156	0.161	0.083	0.191	0.091
Cysteine	C	70	119	119	0.149	0.05	0.117	0.128
Glutamic Acid	E	151	37	74	0.056	0.06	0.077	0.064
Glutamine	Q	111	110	98	0.074	0.098	0.037	0.098
Glycine	G	57	75	156	0.102	0.085	0.19	0.152
Histidine	H	100	87	95	0.14	0.047	0.093	0.054
Isoleucine	I	108	160	47	0.043	0.034	0.013	0.056
Leucine	L	121	130	59	0.061	0.025	0.036	0.07
Lysine	K	114	74	101	0.055	0.115	0.072	0.095
Methionine	M	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	F	113	138	60	0.059	0.041	0.065	0.065
Proline	P	57	55	152	0.102	0.301	0.034	0.068
Serine	S	77	75	143	0.12	0.139	0.125	0.106
Threonine	T	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	W	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	Y	69	147	114	0.082	0.065	0.114	0.125
Valine	V	106	170	50	0.062	0.048	0.028	0.053

Figure 1: Chou-Fasman Table

(b) Consider a 6 amino acids forming a 3×2 lattice. We can have 3 unique arrangements are possible, known as π-structure, G-structure and S-structure. Which among the following sequences have unique ground states? (a) HHPHHP (b)HPPHHP **(10 marks)**

Roll No:.....

Name.....

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY KOTTAYAM
Department of Computer Science and Engineering
End-Semester Examination (Make-up)- January 2024

ISC211: Introduction to Bioinformatics

Duration: 3 hrs | Max Marks: 100
Course Instructors: Dr. Suja, Dr. Manu | Batch: CSE (2022 Admn)

Answer All Questions

1. (a) If a nucleotide sequence is given, how will you distinguish ORFs and why ORFs are important in sequence annotation?. **(5 Marks)**

   (b) Discuss the difference between local and global alignment with suitable examples. **(5 Marks)**

   (c) What is the significance of E value in a BLAST result and how it is different from that of the score? **(5 Marks)**

   (d) Explain how PAM matrix is derived? **(5 Marks)**

2. (a) Using Needleman–Wunsch algorithm, construct the partial alignment score table for the following two sequences, using the scoring parameters: match score= +5, mismatch score =-3, gap penalty =-3. What is the optimal global alignment between these sequences? Analyse the time and space complexities of the algorithm.
   ACAGTCGAACG
   CAGTCACGG **(10 Marks)**

   (b) Given a raw biological sequence, write a program to perform the following tasks:
   i. Remove the characters which are not part of DNA sequence
   ii. Compute the complement of the cleaned sequence
   iii. Compute the reverse complement **(10 Marks)**

3. Answer the following questions in the context of Phylogenetic trees.

   (a) What is the significance of Phylogenetic trees in Bioinformatics analysis? **(5 marks)**

   (b) How distance based algorithms used for construction of Phylogenetic trees? **(5 marks)**

   (c) Use UPGMA method to reconstruct a phylogenetic tree using the following distance matrix given in Table 1: **(10 Marks)**

4. Multiple sequence alignment of 5 taxas are given in Figure 1. Three possible tree alignments (out of 15) are also given in Figure 2.

   (a) Identify the informative and invariant sites from the given MSA **(5 Marks)**

Species	A	B	C	D
B	3	-	-	-
C	6	5	-	-
D	9	9	10	-
E	12	11	13	9

Table 1: Distance Matrix for question 3(c)

```
1  GAATGCTGAT
2  GGATGGTGAT
3  GGATGATGAT
4  GGATGCTGAC
5  GGATGCTGAC
```

Figure 1: MSA for 5 species

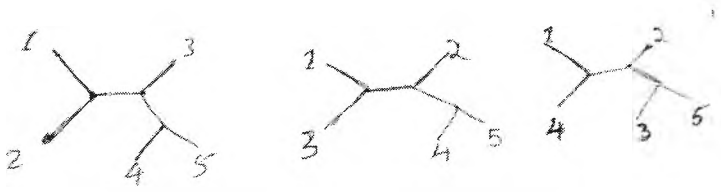

Figure 2: Three out of 15 possible unrooted trees

(b) Apply maximum parsimony algorithm to identify the best among the three given alignments. **(10 Marks)**

(c) Which among the above trees are longest? **(3 Marks)**

(d) Which among the above tress represents the optimal arrangement of species, according to Maximum parsimony algorithm? **(2 Marks)**

5. (a) Consider the following amino acid sequence. Apply Chou Fasman algorithm to predict the regions of alpha helix and beta strands in the sequence. (The propensity values of amino acids are given in Figure 3.
DMNWHIGMCRCNNTKWCQAT **(10 marks)**

Name	Abbrv	P(a)	P(b)	P(tum)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	A	142	83	66	0.06	0.076	0.035	0.058
Arginine	R	98	93	95	0.07	0.106	0.099	0.085
Aspartic Acid	D	101	54	146	0.147	0.11	0.179	0.081
Asparagine	N	67	89	156	0.161	0.083	0.191	0.091
Cysteine	C	70	119	119	0.149	0.05	0.117	0.128
Glutamic Acid	E	151	37	74	0.056	0.06	0.077	0.064
Glutamine	Q	111	110	98	0.074	0.098	0.037	0.098
Glycine	G	57	75	156	0.102	0.085	0.19	0.152
Histidine	H	100	87	95	0.14	0.047	0.093	0.054
Isoleucine	I	108	160	47	0.043	0.034	0.013	0.056
Leucine	L	121	130	59	0.061	0.025	0.036	0.07
Lysine	K	114	74	101	0.055	0.115	0.072	0.095
Methionine	M	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	F	113	138	60	0.059	0.041	0.065	0.065
Proline	P	57	55	152	0.102	0.301	0.034	0.068
Serine	S	77	75	143	0.12	0.139	0.125	0.106
Threonine	T	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	W	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	Y	69	147	114	0.082	0.065	0.114	0.125
Valine	V	106	170	50	0.062	0.048	0.028	0.053

Figure 3: Chou-Fasman Table

(b) Apply the maximum base-pair method to compute the secondary structure of the RNA sequence: *GGACGAUCA* **(10 Marks)**

Name.....

Roll No: 23BCS3

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY KOTTAYAM
Department of Computer Science and Engineering
Mid-Semester Examination- September 2024

ISC211: Introduction to Bioinformatics

Time: 1½, Hours | Max Marks: 50
Course Instructors: Athira B, Manu Madhavan | Batch: III Sem CSE

Provide precise and concise answers

1. (a) Differentiate between introns and exons **(2 marks)**
   (b) Given the following DNA sequence write the reverse complement and corresponding RNA transcription: `ATGCGTACTAGCTCGATGCTAATAGCGTAA` **(2 marks)**
   (c) Give an example for Protein structure database, and primary gene sequence database **(2 marks)**
   (d) Partial scoring matrix of following two sequences is given in Figure 1. Write any two possible optimal global alignments. Verify your answers.

   Seq-1: `GATTACA` Seq-2: `GTCGACGCA` **(4 marks)**

		G	T	C	G	A	C	G	C	A
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
G	-1	5	← 4	← 3	← 2	← 1	← 0	↖ -1	← -2	← -3
A	-2	↑ 4	↖ 5	← 4	← 3	7	← 6	← 5	← 4	← 3
T	-3	↑ 3	9	← 8	← 7	↑ 6	7	← 6	← 5	← 4
T	-4	↑ 2	↑ 8	9	← 8	← 7	↑ 6	7	← 6	← 5
A	-5	↑ 1	↑ 7	↑ 8	9	13	← 12	← 11	← 10	11
C	-6	↑ 0	↑ 6	12	← 11	↑ 12	18	← 17	← 16	← 15
A	-7	↑ -1	↑ 5	↑ 11	12	16	↑ 17	18	← 17	21

Figure 1: Partial alignment matrix

2. (a) Given two DNA sequences s and t, t is called a *motif* if t is contained as a contiguous collection of symbols in s. A *position* of nucleotide in a DNA sequence is the total number of nucleotides found to its left, including itself. Suppose a *motif* is present in s from the *position* i to j, represented as $s[i..j]$, the *location* of the motif will be its beginning position, i. There may be multiple locations if the *motif* appears more than once in the sequence. Write a *Python* **function** that will take a DNA sequence s and *motif* t and return all locations of the *motif* t in sequence s. **(5 marks)**
   (b) Given a DNA sequence, a k-mer is defiend as a substring of length k. For example, given sequence `ACTTGCT` and $k = 3$, {`ACT`, `CTT`, `TTG`, `TGC`, `GCT`} are the set of 3-mers. Write a **Python function** that take a DNA sequence and an integer k as input and return the frequency of all k-mers as a dictionary ({k-mer:frequency}).

(5 marks)

PTO

3. You are working in a bioinformatics lab analyzing two DNA sequences (`AGCTTAGCTA`, `GCTTACA`) to study their evolutionary relationship. The sequences are suspected to code for similar proteins but may have accumulated mutations over time. To compare them and determine their degree of similarity, compute their global pairwise alignment. The match and mismatch scores are derived empirically as given in Table 1. Assume gap penalty is -1.

Table 1: Scoring Matrix

	A	C	T	G
A	5	-5	-5	-1
C	-5	5	-1	-5
T	-5	-1	5	-5
G	-1	-5	-5	5

(10 marks)

4. You are part of a research team studying a new virus. The team suspects a segment of its genome may contain a key protein-coding region. Your task is to identify the Open Reading Frame (ORF) in the following DNA sequence to locate potential coding regions. Identify the ORF and corresponding translated protein sequence. Genetic code is given in Figure 2 for your reference.

Sequence: `CGATGTACGTTAGCGTAGCTAAGCTTACG` **(10 marks)**

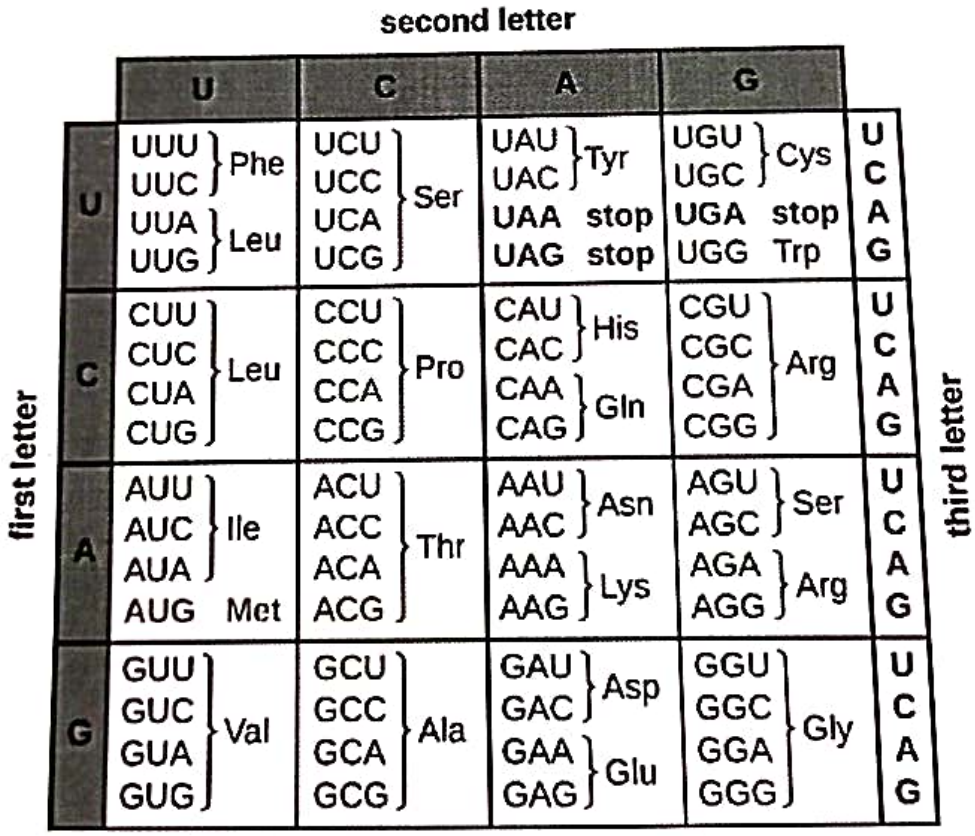

Figure 2: Genetic code

5. What is the significance of biological databases in Bioinformatics research? Discuss key types of biological databases, their contents, and applications, with suitable examples. **(10 marks)**