**Practical BioInformatics**
**BIO221**

**Mid-semester Exam**

**Total Marks 50** **Duration : 60 minutes**

**Write all the Questions.**

**Short answer questions:**

1. In the context of PAM matrices, what does it mean if two amino acids have a positive substitution score? [1]
    a. They are physically similar.
    b. They are functionally similar.
    **c. They are likely to be substituted for each other in evolution.**
    d. They are unlikely to be found in the same protein.
2. Give the two criteria for something to be called an Accepted Point Mutation [1]

**Frequency Criterion:** The mutation should occur frequently enough in a population of homologous sequences, indicating that it is a common and tolerated variation within the evolutionary context.
**Conservation Criterion:** The mutation should maintain the essential function or structure of the biomolecule (e.g., protein or RNA) to ensure its biological activity is retained despite the change.
**Selective Advantage Criterion:** The mutation may confer a selective advantage, either in terms of adaptation to the environment or improved functionality, contributing to its acceptance over evolutionary time.
**Reversibility Criterion:** The mutation should be considered reversible, implying that the mutated state can revert to the original state through subsequent evolutionary events.

**Any alternate explanation will be accepted.

3. In the Dayhoff's model, a higher PAM number means: [1]
    a. Lower rate of evolution
    b. Higher rate of evolution
    c. Lower evolutionary divergence
    **d. Higher evolutionary divergence**

Explanation: In the Dayhoff's model, PAM (Point Accepted Mutations) stands for "Point Accepted Mutation". The PAM model is used to quantify the evolutionary divergence between two homologous protein sequences.A higher PAM number indicates a higher level of evolutionary divergence. Therefore, the correct answer is: Higher evolutionary divergence

4. Which of the following is TRUE for local alignment? [1]
    a. Local alignment is used for closely related sequences

b. The first (top-left) cell always needs to start with a zero score
c. The gap penalty is always fixed as -2
**d. Local alignment is useful for finding shared motifs in unrelated sequences**

Explanation: Local alignment focuses on identifying regions of similarity between sequences, allowing for the detection of shared motifs or conserved domains even in sequences that may have overall low similarity.

5. What is a Single Nucleotide Polymorphism? How is it different from mutation? [2]

A Single Nucleotide Polymorphism (SNP) is a variation in a single nucleotide (A, T, C, or G) at a specific position in the DNA sequence among individuals within a population. SNPs are the most common type of genetic variation, and they can be used as markers for genetic studies and to understand the genetic basis of various traits or diseases.

| | |
|---|---|
| SNP is a change in the single-nucleotide of a genome. Also, it is a type of mutation. | Mutation is the variation in DNA base pairs caused due to insertion, deletion, duplication or substitution of base pairs. |
| The variation is seen only in a single nucleotide. | The variation can be due to changes in many or even a single nucleotide. |
| The SNP variation is available in a minimum of 1% of the population. | The mutation frequency is available in less than 1% of the population. |
| Example – In the sequence ATAGC, the substitution of G by C will produce ATACC. This change in a single nucleotide is termed as SNP. | The mutations are of different types. The missense mutation, silent mutation and nonsense mutation are some of them. |

6. The lack of a clear correlation between genome size and complexity of organisms is called C-value Paradox. Which of the following may explain it? [1]
**a. Genic regions are more or less common between related organisms**
b. Intergenic regions are highly divergent and variable because of non-coding sequences
c. Exons are highly divergent and variable because of non-coding sequences
d. All of the above
7. What is the role of hydrogen bonds in the DNA double helix? [1]
a. They form the backbone of the DNA molecule.
b. They connect the phosphate groups of adjacent nucleotides.
**c. They stabilize the double helical structure by connecting complementary bases.**

d. They are responsible for the replication of DNA.

8. Which of the following components is not a part of a nucleotide in DNA? [1]
   a. Phosphate group
   b. Deoxyribose sugar
   **c. Uracil base**
   d. Nitrogenous base
9. A sequence that reads the same forward and backward is called a palindromic sequence. For palindromic sequences, what is the structure of the dot plot? [1]
   **a. Two intersecting diagonal lines at the midpoint**
   b. One diagonal
   c. Two parallel diagonals
   d. No diagonals
10. True or False: Only one strand of parent DNA is replicated, the one in 5' to 3' direction, while the other is not replicated.

False.

Both strands of the parent DNA are replicated during DNA replication. However, the two strands are replicated in different directions due to the antiparallel nature of the DNA double helix.

The leading strand is replicated continuously in the 5' to 3' direction, while the lagging strand is replicated discontinuously in short fragments called Okazaki fragments.

11. Explain how the following mutation can impact protein synthesis and cellular function? Grade the impact on the scale of low, medium and high. [1+1]
    a. Point mutation (a single cytosine (C) is mistakenly replaced by thymine (T) on third base of the codon)
    b. Nonsense Mutation (a codon encoding for glutamine (CAA) is mutated into a stop codon (UAA))

A. **Impact**: This mutation can lead to the incorporation of an incorrect nucleotide into the newly synthesized DNA strand. If this mutation occurs within a coding region of a gene, it can result in a change in the DNA sequence, leading to the production of an altered mRNA transcript during transcription. This mutation may lead to a low to medium degree of impact if it is in the third position of codon as per the Wobble Hypothesis.

B. **Impact**: The premature stop codon truncates the polypeptide chain prematurely, leading to the production of a shortened protein. This truncated protein may lack functional domains or structural motifs necessary for proper protein function. Additionally, it can disrupt protein-protein interactions or enzymatic activities, ultimately impacting cellular function. This mutation may have a high degree of impact on cellular and protein functions.

**Brief Questions**

1. You are given set of sequences with the alignment score [2]
   a. ATCG Score: 3
      ATTG

b. ATCTTTA Score: 0
ATTT__A

c. ATTCTTA Score: -2
AT_TT_A

match= ?, mismatch= ?, gap extension = ? , gap open = ?

The solving for two question are taken from the reference: https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/download/lectures/PCB_Lect02_Pairwise_allign.pdf

BRIEF QUESTIONS:

(1) Match = 1

Mismatch = 0

Gap extension = -1

Gap open = -2

(Gap opening should be penalized more than extension)

(a) ATCG
ATTG

score = 1 + 1 + 0 + 1 = 3

(b) ATCTTTA
ATTT--A

score = 1 + 1 + 0 + 1 + [(-2) + (-1)] + (-1) + 1
= 0

(c) ATTCTTA
AT-TT-A

score = 1 + 1 + [(-2) + (-1)] + 0 + 1 + [(-2) + (-1)] + 1
= -2

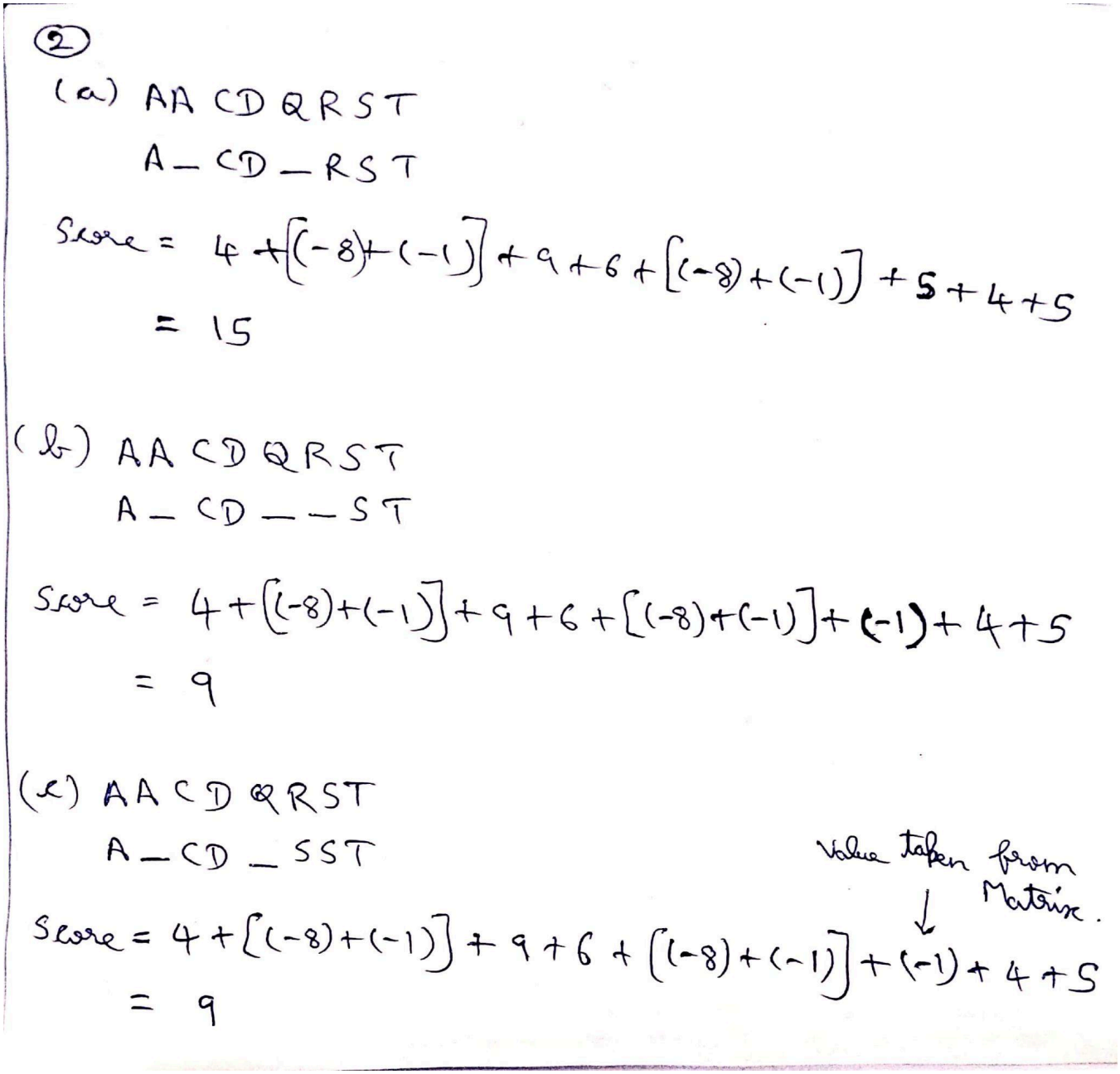
match= 1, mismatch= 0, gap extension = -1 , gap open = -2

2. You are aligning a protein sequence into a substitution matrix: [2]

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | -2 | 1 | 3 | -1 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | -2 | 0 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |

Gap opening costs -8. Gap widening (extension) costs -1. What is the score for the following sequences?

a. AACDQRST
   A–CD–RST
b. AACDQRST
   A–CD—ST
c. AACDQRST
   A–CD–SST




a. 15 b. 9 c. 9

3. BLOSUM matrix is used for local alignment instead of PAM. Reason by explaining the key differences between PAM and BLOSUM matrices. [4]

Local vs. Global Alignment:

- PAM: Primarily designed for global sequence alignments, emphasizing long evolutionary distances.
- BLOSUM: Specifically designed for local alignments, suitable for comparing sequences with shorter evolutionary distances and identifying conserved domains.

Block-Based Approach:

- PAM: Analyzes overall mutation patterns across the entire sequence, less suitable for detecting local similarities.
- BLOSUM: Utilizes conserved blocks of sequences, allowing for the detection of more recent and biologically relevant local similarities.

Sensitivity to Divergence:

- PAM: Performs better for highly divergent sequences where evolutionary changes have accumulated over a long period.
- BLOSUM: Well-suited for sequences that share recent common ancestry and have conserved regions.

Sequence Database Composition:

- PAM: Assumes a constant amino acid composition over evolutionary time.
- BLOSUM: Takes into account variations in amino acid frequencies observed in real protein databases, providing more realistic substitution matrices for local alignments.

Scoring Methodology:

- PAM: Scores are based on the probabilities of specific amino acid substitutions derived from observed mutations.
- BLOSUM: Scores are based on observed frequencies of substitutions in conserved blocks of sequences, emphasizing conserved regions.

Adaptability to Sequence Evolution:

- PAM: Less sensitive to recent changes, may not perform optimally for sequences with more recent evolutionary relationships.
- BLOSUM: More adaptable to recent evolutionary changes, suitable for local alignments where conserved regions are essential.

**Detailed questions:**

1. In the context of a gene, explain why the sense strand is also referred to as the coding strand, and the antisense strand as the template strand. Explain the importance of understanding sense and antisense strands in practical bioinformatics workflows such as designing primers. [5 + 5]

Understanding the sense and antisense strands is crucial for designing primers in PCR (Polymerase Chain Reaction) and other molecular biology techniques. Primers are short DNA sequences that serve as starting points for DNA synthesis during PCR. The primers need to anneal to the target DNA region with high specificity, and this specificity is achieved by designing primers based on the knowledge of sense and antisense strands.

**Primer Direction:**

In PCR, one primer is designed based on the sense strand, and its sequence is the same as the target DNA region. The other primer is designed based on the antisense strand, and its sequence is the reverse complement of the target DNA region.

Primers are designed to flank the region of interest, and their specificity is achieved by ensuring they anneal to the target sequence in the correct orientation.

**Avoiding Self-Annealing:**

Understanding the sense and antisense strands helps avoid self-annealing or dimer formation between the primers. Primers are designed to avoid complementarity with each other to prevent undesirable secondary structures.

2. Explain how the Basic Local Alignment Search Tool (BLAST) can be used to identify homologous DNA sequences in public databases. Describe the steps involved in performing a BLAST search and interpreting the results. [10]

Steps Involved in Performing a BLAST Search:

Access BLAST Website:
- Go to the NCBI BLAST website.

Select BLAST Program:
- Choose the appropriate BLAST program based on the nature of your query:
  - BLASTN: Compares a nucleotide query against a nucleotide database.
  - BLASTP: Compares a protein query against a protein database.

Enter Query Sequence:
- Paste or upload the DNA sequence you want to search for homologs in the "Query" box.

Choose Database:
- Select the appropriate database against which you want to search. Common choices include the entire nucleotide or protein database (nr) or specific databases.

Set Search Parameters:
- Adjust search parameters such as word size, e-value threshold, and scoring matrix based on the specificity and sensitivity required for your analysis.

Run BLAST:
- Click the "BLAST" button to submit your query.

Interpreting BLAST Results:

Alignment Summary:
- Review the alignment summary table to see top hits, including sequence descriptions, scores, and e-values.

Alignment View:
- Explore the alignment view to visually inspect the alignment of your query with homologous sequences.

E-Value:
- Focus on the e-value, which represents the expected number of random hits with a similar or better score. Lower e-values indicate more significant matches.

Identity and Similarity:
- Check for the percentage identity and similarity, providing insights into the conservation of sequences between your query and hits.

Query Coverage:

- Assess the query coverage to understand the proportion of your query that aligns with each hit.

Functional Information:

- If available, examine functional annotations of hits to understand the potential biological significance.

Selecting Hits:

- Choose hits that meet your criteria for significance and relevance to your research.

Accessing Full Sequences:

- Retrieve full sequences of hits for further analysis or download the results.

Refining Search:

- Refine search parameters and perform iterative searches to improve specificity and relevance.

3. Explain how the choice of substitution matrix (PAM or BLOSUM) can influence the outcome of a sequence alignment analysis. Provide examples of biological scenarios where each matrix would be more suitable.

The choice of substitution matrix, such as PAM or BLOSUM, can significantly influence the outcome of sequence alignment analysis:

PAM matrices: Based on evolutionary models, PAM matrices are suitable for aligning divergent sequences, such as those from distantly related species. For example, when comparing the amino acid sequences of proteins from different classes of vertebrates, PAM matrices would be appropriate due to the substantial evolutionary divergence between these species.

BLOSUM matrices: Derived from observed alignments in protein families, BLOSUM matrices are more suitable for aligning closely related sequences. For instance, when aligning protein sequences within the same species to identify conserved domains or motifs, BLOSUM matrices are preferred for their sensitivity to subtle sequence similarities.

The choice of substitution matrix should be guided by the evolutionary distance between the sequences being aligned and the specific goals of the analysis. For instance, when aligning protein sequences with a known evolutionary relationship, the appropriate substitution matrix should be selected based on the level of sequence divergence and the desired balance between sensitivity and specificity. Overall, choosing the optimal substitution matrix ensures accurate alignment and meaningful interpretation of sequence data in biological contexts.