BIF401 – BIOINFORMATICS I

MID TERM IMPORTANT QUESTIONS

1. **What is usage of sequence data?**
   Sequence data can be used to obtain:
   - Similarity of sequences
   - Evolutionary History
   - Predict the function of molecules

2. **What Multiple Sequence Alignment and Pairwise Sequence Alignment?**
   A multiple sequence alignment is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. In many cases, the input set of query sequences are assumed to have an evolutionary relationship by which they share a linkage and are descended from a common ancestor.

   Pairwise Sequence Alignment is used for prediction of probable secondary structures and fold-class; good for visualizing amphipathic helices, where present.

3. **Define Blossum Matrix.**
   BLOSUM matrices can be used to align the protein sequences. A BLOSUM matrix was first purposed in 1992 by Henikoff et al.
   BLOSUM matrices is also called the Block substitution matrix without any gap although it has mismatches in sequences.
   There are three steps to compute the BLOUSM Matrices.
   Step 1: Eliminate sequences that are identical in x% positions
   Step 2: Compute observed frequency f i, j of aligned pair Ai to Aj. Hence, f i,j becomes the probability of aligning Ai and Aj in the selected blocks.
   Step 3: Compute fi which is the frequency of observing Ai in the entire block

4. Why diagnols are selected in computing scoring matrix?

5. **How many nucleotides are present in RNA? Write their names.**
   RNA is composed of four individual nucleotides. These four nucleotides include adenine, cytosine, guanine, and uracil, which replaces thymine in DNA.. A nucleotide consists of a phosphate group, sugar, and a phosphate group.

6. **Differentiate between essential and non-essential amino acids.**
   An essential amino acid, or indispensable amino acid, is an amino acid that cannot be synthesized from scratch by the organism fast enough to supply its demand, and must therefore come from the diet.

   Non-essential means that our bodies can produce the amino acid, even if we do not get it from the food we eat. Nonessential amino acids include: alanine, arginine, asparagine, aspartic acid, cysteine, glutamic acid, glutamine, glycine, proline, serine, and tyrosine.

7. **What is the significance of the top down protein?**
   - Relatively simple sample preparation.
   - No protein digestion needed.
   - Direct detection of the molecular mass of biological proteins.
   - Mass information is retained in a great extent, such as PTMs.
   - Less time-consuming.

8. Write down the names of four scoring scheme methods in computing the position of nucleotides in Nussinov Jacobson's algorithm.

4 positions are considered to calculate the NJ scoring matrix. At every step, we can see score contributions from the 4 possible locations.

9. What are the applications of bioinformatics in genomics and proteomics?

GENOMICS

- ✓ Bioinformatics can help in assembling DNA sequencing data.
- ✓ It can help in gene finding (markers).
- ✓ Gene assembly can be performed using bioinformatics tools (nucleotide alignments)
- ✓ It can help transcribe the gene data to RNA data
- ✓ Also, databases can be generated from such data.

PROTEOMICS

- ✓ Bioinformatics can help us in decoding protein sequences.
- ✓ It can also help us in understanding protein structure.
- ✓ We can also understand post translational changes in proteins with the help of bioinformatics.
- ✓ We can better understand the protein-protein interaction in different biological reactions.
- ✓ It can also help us in generating databases of these sequences and structures.

10. Differentiate between Rooted and un-rooted phylogenetic trees and their uses.

Rooted trees show the most basal ancestor of the tree. Unrooted phylogenetic tree does not show an ancestral root. Unrooted trees represent the branching order but do not indicate the root or location of the last common ancestor. Unrooted trees show the relatedness of organisms without indicating ancestry.

In a phylogenetic tree, every node represents a species. A rooted tree is a tree in which one of the nodes is stipulated to be the root, and thus the direction of ancestral relationships is determined. An unrooted tree, as could be imagined, has no pre-determined root and therefore induces no hierarchy.

11. Define the terms: DNA replication, Translation, Transciption.

DNA Replication:

In molecular biology, DNA replication is the biological process of producing two identical replicas of DNA from one original DNA molecule. DNA replication occurs in all living organisms acting as the most essential part for biological inheritance.

Translation:

Translation is the process that takes the information passed from DNA as messenger RNA and turns it into a series of amino acids bound together with peptide bonds. It is essentially a translation from one code (nucleotide sequence) to another code (amino acid sequence).

Transcription:

Transcription, as related to genomics, is the process of making an RNA copy of a gene's DNA sequence. This copy, called messenger RNA (mRNA), carries the gene's protein information encoded in DNA.

12. Write types of methods for clustering of phylogenetic trees.

Several methods exist for constructing phylogenetic trees. Broadly, they belong to objective methods or clustering methods. We use UPGMA and Distance Methods.

13. How Needle-man wunsch algorithm finds it's optimal point and how (position) move backwards?

The traceback always begins with the last cell to be filled with the score, i.e. the bottom right cell. One moves according to the traceback value written in the cell. There are three possible moves: diagonally (toward the top-left corner of the matrix), up, or left. The traceback is completed when the first, top-left cell of the matrix is reached.

14. How we can mathematically find the distance between two phylogenetic trees?

A phylogenetic tree is a diagram that represents evolutionary relationships among organisms. In trees, two species are more related if they have a more recent common ancestor and less related if they have a less recent common ancestor. Phylogenetic trees can be drawn in various equivalent styles.

15. Define PAM and Blossum Matrix.

PAM:

Alignment scoring matrices are very useful in giving suitable scores to matches and mismatches. There are 2 types of scoring matrices i.e. PAM and BLOSUM.

PAM means "Point Accepted Mutations" • Point accepted mutation is a substitution of one amino acid by another such that the protein functions stays conserved. PAM unit is a time in which about 1% of amino acids in a sequence undergo accepted mutations.

Blosum Matrix:

BLOSUM matrices can be used to align the protein sequences. A BLOSUM matrix was first purposed in 1992 by Henikoff et al. BLOSUM matrices is also called the Block substitution matrix without any gap although it has mismatches in sequences

16. What is ensemble?

As human brain is limited to remember and store the information for long time that's why we use online database for the storage of Molecular information. ESEMBLE is genome search engine which is used to search the genome of every recorded species.

17. What is FASTA algorithm?

FASTA can search sequence databases and identify unknown sequences by comparing them to the known sequence databases. This can help obtain information on the parent organism, function and evolutionary history.

18. Enlist 4 position of Scoring sequence.

19. Enlist some step in Bottom up sequence?

- Sample containing the mixture of protein from cells and tissues is obtained.
- Enzymes such as trypsin is use to cleave the proteins.
- Enzyme cleaves the amino acids at specific sites of amino acid.
- Several peptides are formed when protein is cleaved.
- Number of peptide depends upon the number of sites where enzymes cleaved the protein. For example trypsin cleaves the protein at lysine (k)
- Mass of each peptide is measured.
- One peptide is selected at a time.
- Different enzyme is use to cleave the protein at different site.
- This process keep going until the possible number of peptides are formed or searched.
- Peptides are searched in data base and matched.

20. Explain Scoring Matrices and drive its Formula.

Scoring matrices are used to determine the relative score made by matching two characters in a sequence alignment. There are many flavors of scoring matrices for amino acid sequences, nucleotide sequences, and codon sequences, and each is derived from the alignment of "known" homologous sequences.

21. How DNA differ from RNA?

There are two differences that distinguish DNA from RNA: (a) RNA contains the sugar ribose, while DNA contains the slightly different sugar deoxyribose (a type of ribose that lacks one oxygen atom), and (b) RNA has the nucleobase uracil while DNA contains thymine. Moreover, the DNA is double stranded and RNA is single stranded.

22. What are 3 main outputs of BLAST?

Ordinary text and xml output for easy computational parsing is also available. The default layout of the NCBI BLAST result is a graphical representation of the hits found, a table of sequence identifiers of the hits together with scoring information, and alignments of the query sequence and the hits. Results are shown in HTML, plain text, and XML formats. A table lists the sequence hits found along with scores.

23. Write steps of FASTA algorithm to determine protein or nucleotides sequence.

Step 1: Local regions of identity are found.
Step 2: Rescore the local regions using PAM or BLOSUM matrix.
Step 3: Eliminate short diagonals below a cutoff score.
Step 4: Create a gapped alignment in a narrow segment and then perform Smith Waterman alignment.

24. Write four steps of FASTA input sequence.

Step 1: Specify the tool input (sequence and database).
Step 2: Entering of input sequence.
Step 3: Set up the parameters.
Step 4: Submit the query for processing.

25. What are indels?

Indels, being either insertions, or deletions, can be used as genetic markers in natural populations, especially in phylogenetic studies. OR Removal and addition of amino acids in proteins and nucleotides in DNA, RNA by using Gaps named as Indels.

26. How to identify the protein structure?

If another protein which has a similar sequence also has its structure known, the structure of an unknown protein can be predicted based on that similar protein. So, it is then possible to identify unknown protein structures by just examining the homologous protein sequences. Good sequence alignment and identity ensures that homology modeling will give accurate results

Thus, Homology modeling is used to predict structures of proteins having high sequence similarity with other proteins with known structures.

27. What is uniprot and swissprot?

UniProt is the Universal Protein resource, a central repository of protein data created by combining the Swiss-Prot, TrEMBL and PIR-PSD databases. UniProt is a freely accessible database of protein sequence and functional information, many entries being derived from genome sequencing projects.

SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.

28. Write different fields of bioinformatics by using EXPASY.

- ✓ Developed by Bioinformatics Institute (SIB)
- ✓ Website provides access to databases and tools
- ✓ Proteomics, Genomics, Phylogeny, Systems biology, Population genetics, transcriptomics etc.
- ✓ Expasy provides access to a variety of online databases and tools.
- ✓ Depending upon your requirement, you find sequence information from Expasy.

29. Differentiate between identity and similarity and write the formula for identity.

Identity means the counting number of nucleotides or amino acids which exactly match when two biological sequences are matched.

For example:
Number of match = 5
Smaller length = 5

Sequence (1) = 7
Sequence (2) = 5

Formula for Identity:
Identity = No. of Matches / smaller length × 100

Similarity means the comparison between two different sequences calculated by alignment approach. In both identity and similarity the dots are not counted.

30. How can we reduce the cost of sequencing?

One by one sequence compression is costly and time consuming process we minimize the cost with the help of algorithm.

31. What is Digital biology?

In today's world, when a biologist performs an experiment in the wet-lab, he or she in fact produces digital data which is continuously being stored on computer disks. The data may include text, numbers, symbols or images.

32. Explain mathematical relationship of cluster.

A cluster is a group of objects, numbers, data points (information), or even people that are located close together. If you plot a series of numbers on a graph and you see several of your dots gathered together, you have a cluster.

33. What is BLOSUM matrices? How we compute BLOSUM matrix? Derive its mathematical formula.

In bioinformatics, the BLOSUM matrix is a substitution matrix used for sequence alignment of proteins. BLOSUM matrices are used to score alignments between evolutionarily divergent protein sequences. They are based on local alignments.
Each value in the matrix is calculated by dividing the frequency of occurrence of the amino acid pair in the BLOCKS database, clustered at the 62% level, divided by the probability that the same two amino acids might align by chance.

There are three steps to compute the BLOUSM Matrices.
Step 1: Eliminate sequences that are identical in x% positions
Step 2: Compute observed frequency f i, j of aligned pair Ai to Aj. Hence, f i,j becomes the probability of aligning Ai and Aj in the selected blocks.
Step 3: Compute fi which is the frequency of observing Ai in the entire block

$$s_{ij} = \log_2 \frac{f_{ij}}{p_{ij}}, \quad p_{ij} = \begin{cases} f_i f_i & i = j \\ 2 f_i f_j & i \neq j. \end{cases}$$

34. What is purpose of tBLASTn and tBLASTx?

tBLASTn:
It compares a protein query sequence against a nucleotide sequence database. Nucleotide sequence dynamically translated into all reading frames.

tBLASTx:
It compares the six-frame translated proteins of a nucleotide query sequence against the six-frame translated proteins of a nucleotide sequence database.

35. Write down purine and primidine nucleotides.

Purines and Pyrimidines are nitrogenous bases that make up the two different kinds of nucleotide bases in DNA and RNA. The two-carbon nitrogen ring bases (adenine and guanine) are purines, while the onecarbon nitrogen ringbases (thymine and cytosine) are pyrimidines.

36. What is domain and domain shuffling?

Domain:

Domains are distinct functional and/or structural units in a protein. Usually they are responsible for a particular function or interaction, contributing to the overall role of a protein. Domains may exist in a variety of biological contexts, where similar domains can be found in proteins with different functions.

Domains are semi-independent functional structures in a protein. Protein may contain multiple domains. Hence, we can try to classify proteins by their domains. Locally Compact – Domains interact (H-bonds) more internally than externally. Domains have a hydrophobic core. Domains are contiguous (min. chain breaks).

Types of Domains

- ✓ Alpha Domains
- ✓ Beta Domains
- ✓ Alpha/Beta Domains
- ✓ Alpha + Beta Domains
- ✓ Alpha & Beta Multi-Domains
- ✓ Membrane & cell-surface proteins

Domain Shuffling:

Aligned portions of sequence can be considered in varying orders and this process is called as domain shuffling.

37. Enlist online databases for protein.

Both UniProt and SwissProt are the online database for proteins.

38. What are benefits of FASTA algorithm?

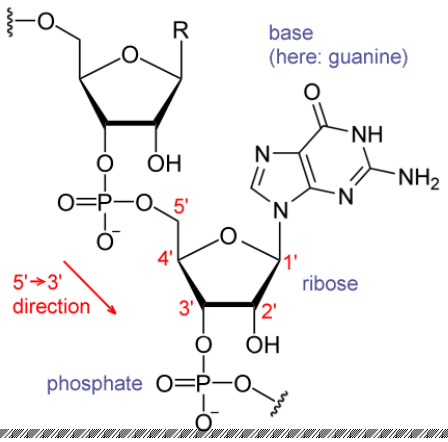
FASTA can briskly perform sequence search databases if given a query sequence. Multiple types of FASTA exist which assist in aligning DNA/RNA/Protein sequences.

39. Acronym ORF and FASTA stands for what?

Open Reading Frame (ORF)
FAST-ALL (FASTA)

40. Why RNA is less stable? Illustrate with figure.

Unlike DNA, RNA in biological cells is predominantly a single-stranded molecule. While DNA contains deoxyribose, RNA contains ribose, characterised by the presence of the 2′-hydroxyl group on the pentose ring (Figure 5). This hydroxyl group make RNA less stable than DNA because it is more susceptible to hydrolysis.



41. What are benefits of dot plot method?

Data points may be labeled if there are few of them. Dot plots are one of the simplest statistical plots, and are suitable for small to moderate sized data sets. They are useful for highlighting clusters and gaps, as well as outliers. Their other advantage is the conservation of numerical information.

42. What is Multiple Sequence Alignment?

A multiple sequence alignment is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. In many cases, the input set of query sequences are assumed to have an evolutionary relationship by which they share a linkage and are descended from a common ancestor.

43. How bulges are formed?

Bulges, are formed when a double-stranded region cannot form base pairs perfectly. Bulges can be asymmetric with varying number of base pairs on one side of the loop.

PREPARED BY: SIR MUGHAL
CONTACT: +923152113633