BIF101 – INTRODUCTION TO BIOINFORMATICS

FINAL TERM MOST IMPORTANT QUESTIONS

PREPARED BY: SHAHZAIB SHABBIR (TEAM MEMBER, BS ZOOLOGY ANNOUNCEMENT GROUP)

(ADMIN BS ZOOLOGY ANNOUNCEMENT, SHAHZEB MUGHAL)

Group link: https://chat.whatsapp.com/BjC6W0jZx8o9yaC2ISh1Gu

Q1. Nucleus structure and function.

Structure: The nucleus is a membrane-bound organelle containing chromatin (DNA and proteins) and nucleoplasm, surrounded by a double membrane called the nuclear envelope.

Function: It regulates gene expression, stores genetic information, and coordinates cell activities such as growth, metabolism, and reproduction.

Nucleus Structure in Prokaryotic and Eukaryotic Cells:

Prokaryotic Nucleus Structure:

Nucleoid Region: Prokaryotes do not have a true nucleus instead they contain a nucleoid, which is an irregularly shaped region in the cell where the circular DNA is located.

Lack of Membrane: The nucleoid is not surrounded by a membrane distinguishing prokaryotic cells from eukaryotic cells.

Eukaryotic Nucleus Structure:

Nuclear Envelope: Eukaryotic cells have a well-defined nucleus enclosed by a double membrane called the nuclear envelope which contains nuclear pores for transport.

Chromatin and Nucleoplasm: Inside the nucleus, chromatin (DNA and proteins) is organized and the nucleoplasm serves as the medium for nuclear components, including the nucleolus where ribosomal RNA is synthesized.

Q2. Features of clustalW.

ClustalW is used to align multiple nucleotide or protein sequences. It helps identify regions of similarity that may indicate functional, structural, or evolutionary relationships among the sequences.

1. Create multiple alignments
2. Optimize existing alignments
3. Profile analysis
4. Create phylogenetic trees

Q3.Types of Proteomics.

1. Expression proteomics
2. Structural proteomics
3. Functional proteomics

EXPRESSION PROTEOMICS:

Used to study the qualitative and quantitative expression of total proteins under two different conditions.

1. Normal and diseased state. E.g tumor or normal cell.
2. It studied that protein is over expressed or under expressed.
3. 2-D electrophorasis.

STRUCTURAL PROTEOMICS;

Structural proteomics helps to understand three dimensional shape and structural complexities of functional proteins.

1. It determine either by amino acid sequence in protein from a gene this process is known as homology modeling.
2. It identify all the protein present in complex system protein-protein interaction.
3. Mass spectroscopy is used for structure determination.

FUNCTIONAL PROTEOMICS;

Functional proteomics explains understanding the protein functions well as unrevealing molecular mechanisms within the cell that depend on identification of the interacting protein partners.

Q4. Define NGS.

•***Next-Generation Sequencing (NGS):*** NGS is a high-throughput sequencing technology that allows for the rapid sequencing of large amounts of DNA or RNA enabling comprehensive analysis of genomes, transcriptomes, and epigenomes.

•***Capabilities:*** It can generate millions of sequences simultaneously providing insights into genetic variations, gene expression, and complex biological processes.

Prepared by:
Shahzaib Shabbir
03308553704

•***Applications:*** NGS is widely used in genomics research, clinical diagnostics, personalized medicine, and various fields such as cancer research and microbiome studies.

Q5. Write about Simulation of the Folding Process.

Simulation of the Folding Process:

1. Build an accurate initial model (including energy and forces).
2. Accurately simulate the dynamics of the protein folding process.
3. The native structure will steadily emerge.

Q6. Principle of homology modeling.

A computational method used to predict a protein's 3D structure based on its similarity to a known structure of a homologous protein.

Q7.Define Primary key, queries and form.

Primary Key: A unique identifier for a record in a database table ensuring that each entry can be distinctly recognized.

Queries: Requests for specific data from a database allowing users to retrieve and manipulate information.

Form: A user interface element that allows data entry and interaction with a database, facilitating data input and retrieval.

Q8.Define FASTA and its types.

FASTA is a file format used to represent nucleotide or protein sequences in a plain text format where each sequence entry begins with a header line that starts with a ">" symbol.

Types of FASTA:

1. ***Fasta35:*** Scan a protein or DNA sequence library for similar sequences.
2. ***Fastx35:*** Compare a translated DNA sequence (6 ORFs) to a protein sequence database.
3. ***tfastx35:*** Compare a protein sequence to a DNA sequence database (6 ORFs).
4. ***fasty35:*** Compare a DNA sequence (6ORFs) to a protein sequence database.
5. ***fasts35:*** Compare unordered peptides to a protein sequence database.
6. ***fastm35:*** Compare ordered peptides (or short DNA sequences) to a protein (DNA) sequence database.

Q9.Write Chromosome components.

1. ***DNA:*** The primary component of chromosomes, consisting of long strands of deoxyribonucleic acid that carry genetic information.
2. ***Histones:*** Proteins that help package and organize DNA into a compact structure, forming nucleosomes that coil and fold to create the chromosome's shape.
3. ***Sister Chromatids:*** Identical copies of a chromosome formed during DNA replication, connected at the centromere.
4. ***Centromere:*** A constricted region of the chromosome that serves as the attachment point for spindle fibers during cell division.

Prepared by:
Shahzaib Shabbir
03308553704

5. ***Telomeres:*** Repetitive nucleotide sequences at the ends of chromosomes that protect them from degradation and prevent fusion with neighboring chromosomes.

Q10. Proteins as Modular Structures.

Segments of amino acid sequences can be considered as functional building blocks or modules. The modular units in proteins that confer specific properties and functions are referred to as “motifs” or “domains”.

Q11. Difference between unicellular and multicellular organisms.

Differences Between Unicellular and Multicellular Organisms

Unicellular Organisms

1. ***Cell Structure:*** Composed of a single cell that performs all life functions.
2. ***Complexity:*** Simpler in structure all processes occur within one cell.
3. ***Reproduction:*** Typically reproduce asexually, often through binary fission.
4. ***Growth and Development:*** Grow by increasing cell size.
5. ***Response to Environment:*** Respond to environmental changes as a single unit.
6. ***Lifespan:*** Often have shorter lifespans.

Multicellular Organisms

1. ***Cell Structure:*** Made up of multiple cells that work together.
2. ***Complexity:*** Exhibit greater complexity with specialized cells, tissues, and organs.
3. ***Reproduction:*** Can reproduce both sexually and asexually.
4. ***Growth and Development:*** Grow by increasing cell number and differentiation.
5. ***Response to Environment:*** Have complex responses due to specialized systems.
6. ***Lifespan:*** Can have longer lifespans due to their complex structures.

Q12.Four records of GEO.

- ***Sample (GSM):*** preparation and description of the samples.
- ***Platform (GPL):*** technology used and the features detected microarray or RNASeq.
- ***Series (GSE):*** defines a set of samples and how they are related.
- ***Datasets (GDS):*** sample data collections assembled by GEO.

Q13. How Deoxyribonucleotides join in DNA bonded prior to double helix bond.

Deoxyribonucleotides are joined together by phosphodiester bonds which form between the 5' phosphate group of one nucleotide and the 3' hydroxyl group of the next nucleotide.This bonding creates a sugar-phosphate backbone, providing structural support to the DNA strand.

Q14.What is the function of ribosomes?

Ribosomes are cellular structures that facilitate the translation of messenger RNA (mRNA) into proteins by decoding codons, which are sequences of three nucleotides that correspond to specific amino acids. They serve as the site where amino acids are linked together to form polypeptides, ultimately leading to

protein synthesis. Ribosomes play a crucial role in the overall process of gene expression, but they do not store genetic information themselves.

Q15.“Twilight zone”?

"Twilight Zone" refers to a range of sequence identity (typically around 20-30%) where reliable predictions of protein structure based on homology modeling become challenging. In this zone low identity and alignment scores indicate that the sequences may not be similar enough to confidently predict the three-dimensional structure. This phenomenon highlights the limitations of using sequence similarity alone for accurate structural predictions.

Q16.What is Modeller software and also write it uses.

Modeller is a software tool used for homology or comparative modeling of protein structures. It allows researchers to build three-dimensional models of proteins based on known structures of homologous proteins facilitating studies in structural biology, drug design, and protein function analysis. Modeller can also be used for Template recognition and initial alignment, Alignment correction, Backbone generation, Loop modeling, Side-chain modeling, Model optimization, Model validation.

Q17.How many sub-discipline of genetics also write there names.

There are four sub-discipline of genetics:

1. Transmission (Classical) Genetics
2. Population Genetics
3. Quantitative Genetics
4. Molecular Genetics

Q18.Which the two types/options of NCBI for Sequence Submission.

NCBI has two options:

- ***BANKIt***

1. For simple sequences and annotations
2. For submission through web
3. Do not require advanced tools Sequence Submission NCBI

- ***Sequin***

1. For Complex sequences and annotations
2. For off-line submissions
3. Do require advanced tools and graphical reports

Q19.Write short note on I-TASSER.

The **I-TASSER (Iterative Threading Assembly Refinement)** server is a powerful tool for automated protein structure and function prediction, utilizing a sequence-to-structure-to-function approach. It

begins with an amino acid sequence and generates 3D atomic models through multiple threading alignments and iterative structural assembly simulations.The server infers protein functions by structurally matching the generated models with known proteins, providing outputs that include full-length secondary and tertiary structures, functional annotations for ligand-binding sites, and a confidence score that estimates the accuracy of the predictions. This makes I-TASSER a valuable resource for researchers in the field of molecular biology and bioinformatics.

Q20.What is Nanopore technology? What are its applications.

Nanopore technology involves the use of nanopores tiny holes in a membrane to analyze individual molecules, such as DNA or proteins, as they pass through the pore.When a molecule translocates through the nanopore, it disrupts an ionic current, allowing for real-time detection and analysis of its size, shape, and sequence based on the characteristic changes in the current.

Applications: This technology is widely used in genomics for sequencing DNA detecting proteins, studying molecular interactions offering advantages like long read sequencing and the ability to analyze single molecules without amplification.

Q21.Write types of protein structure and how do they differ in terms of composition and function?.

Types of Protein Structures

1.***Primary Structure***:

The primary structure refers to the unique sequence of amino acids in a polypeptide chain linked by peptide bonds. This sequence determines the protein's identity and function, with a diversity represented by 20^n combinations where n is the number of amino acids.

2.***Secondary Structure:***

The secondary structure involves local folding patterns within the polypeptide chain primarily stabilized by hydrogen bonds. Common forms include α-helices and β-pleated sheets which contribute to the overall stability and shape of the protein.

3.***Tertiary Structure:***

The tertiary structure is the overall three-dimensional shape of a single polypeptide chain formed by the bending and folding of the chain due to interactions among side chains (R groups). This structure is crucial for the protein's functionality and is stabilized by various interactions, including hydrogen bonds, ionic bonds, and hydrophobic interactions.

4.***Quaternary Structure:***

The quaternary structure refers to the arrangement of multiple folded protein molecules (subunits) in a multi-subunit complex. This structure is important for the function of many proteins as the interactions between subunits can affect the protein's activity and regulation.

Q22.What are the techniques used for determining protein structures.

Techniques for Determining Protein Structures:

1.***X-Ray Crystallography:***

Prepared by:
Shahzaib Shabbir
03308553704

This technique involves crystallizing a protein and then bombarding it with X-rays. The resulting diffraction pattern is analyzed to determine the electron density and ultimately the three-dimensional structure of the protein at atomic resolution.

2.***NMR Spectroscopy:***

Nuclear Magnetic Resonance (NMR) spectroscopy is used to study proteins in solution, providing information about their structure and dynamics. By analyzing the magnetic properties of atomic nuclei, NMR can reveal distances between atoms and help construct a three-dimensional model of the protein. Both techniques are essential for understanding protein structure and function in molecular biology.

Q23.ab initio and how is it used to predict protein structures?.

Ab initio refers to computational methods that predict molecular structures and properties based solely on fundamental physical principles without relying on empirical data. In protein structure prediction ab initio methods simulate the folding and interactions of amino acids using quantum mechanics allowing for the determination of structures from amino acid sequences.

Q24.What are the steps involved in the homology modeling process for predicting the structure of a target protein?

Homology modeling of the target structure can be done as follows:

1. Template recognition and initial alignment
2. Alignment correction
3. Backbone generation
4. Loop modeling
5. Side-chain modeling
6. Model optimization
7. Model validation

Q25.Describe modification in protein lifecycle.

Proteins are synthesized by the translation of mRNAs into polypeptides on ribosomes.In most cases, the initial polypeptide-translation product undergoes some type of modification before it assumes its functional role in a living system.These changes are broadly termed “post-translational modifications” and encompass a wide variety of reversible and irreversible chemical reactions.Approximately 200 different types of post-translational modifications have been reported.

Q26.Write types of protein secondary structures.

Types of Protein Secondary Structures:

- Helices
- Beta Sheets
- Coils
- Loops

Q27.Describe Folding process.

Protein folding refers to the process through which a polypeptide chain assumes its biologically active form, achieving its native three-dimensional (3D) structure.During this process, the amino acids within the chain interact with one another, leading to the formation of a precisely defined folded protein.

Prepared by:
Shahzaib Shabbir
03308553704

Q28.Describe sequence analysis and also write its types.

Sequence analysis is a fundamental method in bioinformatics that involves examining DNA, RNA, or protein sequences using various analytical techniques to understand their features, functions, structures, or evolutionary relationships.

Types:

1. Pair-wise Sequence Alignment.
2. Multiple Sequence Alignment.

Q29.Database management system (DBMS).

The database management system (DBMS) is the software or tool that is used to manage the database and its users. A DBMS consist of different components or subsystem that we will study about later. Each subsystem or component of the DBMS performs different function(s) so a DBMS is collection of different programs but they all work jointly to manage the data stored in the database and its users. In many books and may be in this course sometimes database and database management system are used interchangeably but there is a clear difference and we should be clear about them. Sometimes another term is used, that is, the database system, again, this term has been used differently by different people however in this course we use the term database system as a combination of database and the database management system. So database is collection of data, DBMS is tool to manage this data, and both jointly are called database system.

Q30.Why Databases are important why?also write its advantages.

Traditionally computer applications are divided into commercial and scientific (or engineering) ones. Scientific applications involve more computations, that is, different type of calculations that vary from simple to very complex. Today such applications exist, like in the fields of space, nuclear, medicine that take hours or days of computations on even computers of the modern age. On the other hand, the applications that are termed as commercial or business applications do not involve much computations, rather minor computation but mainly they perform the input/output operations. That is, these applications mainly store the data in the computer storage, then access and present it to the users in different formats (also termed as data processing) for example, banks, shopping, production, utilities billing.

Advantages of Databases.

o Data Sharing: The data for different applications or subsystems is placed at the same place. This introduces the major benefit of data sharing. That is, data that is common among different applications need not to be stored repeatedly, as was the case in the file processing environment. For example, all three systems of an educational institution need to store the data about students. Now the data like registration number, name, address, father name that is common among different applications is being stored repeatedly in the file processing system environment, where as it is being stored just once in database system environment and is being shared by all applications. The interesting thing is that the individual applications do not know that the data is being shared and they do not need to. Each application gets the impression as if the data is being for stored for it. This brings the advantage of saving the storage along with others discussed later.

Prepared by:
Shahzaib Shabbir
03308553704

o Data Independence: Data and programs are independent of each other, so change is once has no or minimum effect on other. Data and its structure is stored in the database where as application programs manipulating this data are stored separately, the change in one does not unnecessarily effect other.

o Controlled Redundancy Means that we do not need to duplicate data unnecessarily; we do duplicate data in the databases, however, this duplication is deliberate and controlled.

o Better Data Integrity Very important feature; means the validity of the data being entered in the database. Since the data is being placed at a central place and being managed by the DBMS, so it provides a very conducive to check or ensure that the data being entered into the database is actually valid. Integrity of data is very important, since all the processing and the information produced in return are based on the data. Now if the data entered is not valid, how can we be sure that the processing in the database is correct and the results or the information produced is valid? The businesses make decisions on the basis of information produced from the database and the wrong information leads to wrong decisions, and business collapse. In the database system environment DBMS provides many features to ensure the data integrity, hence provides more reliable data processing environment.

Q31.Entrez.

Entrez is an integrated search engine which allows users to search and retrieve different data from the NCBI.

Q32. Difference between genomics and proteomics .

Genomics: Genomics is the study of genomes encompassing the sequencing, mapping, and analysis of the complete set of DNA within an organism, including all of its genes.

Proteomics: Proteomics is the study of the proteome focusing on the identification, quantification, and analysis of the entire set of proteins expressed by a cell, tissue, or organism, and examining their structures, functions, and interactions.

Q33. FTP.

The File Transfer Protocol (FTP) is a standard network protocol used to transfer files Via command line or application programs like FTP clients.

Q34.Z-score.

A Z-score, also known as a standard score, is a statistical measurement that describes a data point's position relative to the mean of a group of data points.

Q35.Gene-Bank.

A gene bank, also known as a genetic bank, is a facility or repository that preserves genetic material.

Q36.Force field.

Force field are used to simulate and predict the physical and chemical properties of molecules, including their structures, energies, and dynamics.

Q37.Blast.

Prepared by:
Shahzaib Shabbir
03308553704

BLAST which stands for Basic Local Alignment Search Tool is a widely-used bioinformatics program for comparing a query sequence against a database of sequences.

Q38.Types of Genomics.

1.***Structural Genomics:*** Focuses on the sequencing and analysis of the complete DNA sequence and the physical structure of entire genomes.

2.***Functional Genomics:*** Studies the roles and interactions of genes and proteins often using data from transcriptomics, proteomics, and metabolomics.

3.***Comparative Genomics:*** Compares the genomic features of different species to understand evolutionary relationships and functional biology.

4.***Epigenomics:*** Studies the complete set of epigenetic modifications on the genetic material of a cell, such as DNA methylation and histone modification, which affect gene expression without altering the DNA sequence.

5.***Metagenomics:*** Analyzes genetic material recovered directly from environmental samples, providing insights into the collective genome of microbial communities without the need for culture-based methods.

Q39.Applications of genomics and proteomics.

Applications of Genomics:

1.***Personalized Medicine:***
Genomics enables the development of personalized treatment plans based on an individual's genetic makeup, improving the effectiveness and safety of therapies.

2.***Disease Diagnosis and Prognosis:***
Identification of genetic mutations associated with diseases allows for improved diagnosis and understanding of disease progression.

3.***Drug Development:***
Genomic information can help identify new drug targets and understand variations in drug response, aiding in the creation of more effective therapeutics.

4.***Agriculture:***
Genomics assists in crop improvement through the identification of genes associated with desirable traits like disease resistance and drought tolerance.

5.***Evolutionary Biology:***
Comparative genomics studies enhance our understanding of evolutionary relationships and the genetic basis of adaptation.

6.***Forensic Science:***

Genomic techniques are used for identity verification and crime scene investigation by analyzing DNA samples.

Applications of Proteomics:

1.***Biomarker Discovery:***
Proteomics helps identify protein markers for diseases, offering insights for early detection, diagnosis, and monitoring of disease states.

2.***Understanding Disease Mechanisms:***
Analysis of protein expression and modification provides insights into the molecular pathways involved in disease processes.

3.***Drug Development and Target Identification:***
Proteomics aids in the identification of new therapeutic targets and helps understand the mechanism of action of drugs.

4.***Functional Protein Networks:***
Studying protein-protein interactions helps map cellular pathways and networks enhancing understanding of cellular functions and signaling.

5.***Agriculture:***
Proteomic analyses can improve understanding of stress responses in plants leading to the development of more resilient crop varieties.

6.***Environmental Science:***
Proteomics assists in studying the impact of environmental changes on biodiversity and ecosystem function through protein expression patterns.

Prepared by: Shahzaib Shabbir
Whatsapp no: 03308553704

Group Admin: Shahzeb Mughal
Whatsapp no:03132310940