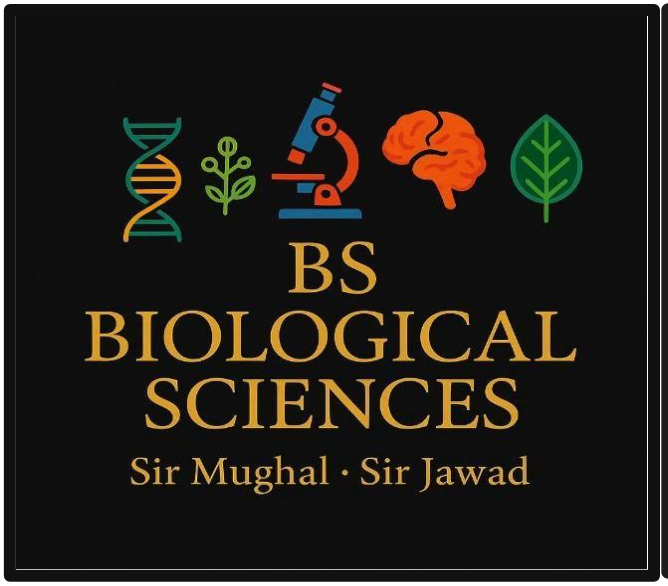

BS BIOLOGICAL SCIENCES

Admins:
Shahzeb Mughal (03132310940)
Jawad Chaudhary (03036100859)

For VU Official Updates Join:

https://chat.whatsapp.com/DWXwJpXahbxFVQutey8JXN

BIF501- Bioinformatics -II

SHORT NOTES / MIDTERM FILE

By

SHAHZAIB SHABBIR

(TEAM MEMBER BS BIOLOGICAL SCIENCES GROUP)

0330-8553704

Join Bif501 group by clicking on this link:

https://chat.whatsapp.com/BkarBswC4QDDclvBgmXvQx

SHORT NOTES / MIDTERM FILE:

Click to contact us:
Sir Mughal: https://wa.me/923132310940
Sir Jawad: https://wa.me/923036100859

Q1. What is Bioinformatics and why is it important?

Answer:

Bioinformatics is an interdisciplinary field that combines biology, computer science, mathematics, statistics, and information technology to analyze and interpret biological data. The term was introduced by Paulien Hogeweg and Ben Hesper.

Bioinformatics is important because modern biological research produces huge amounts of genomic and molecular data. Computational tools are needed to store, organize, analyze, and interpret this information efficiently. It helps researchers study genes, proteins, genomes, and biological systems, making it essential in genomics, proteomics, drug discovery, and medical research.

Q2. Differentiate between Bioinformatics and Computational Biology.

Answer:

Bioinformatics mainly focuses on the storage, management, analysis, and interpretation of biological data such as gene sequences, protein structures, and genomes using computational tools. It is often considered computational molecular biology.

Computational biology, on the other hand, is a broader field that includes all biological studies involving computation. It mainly focuses on developing theoretical algorithms and mathematical models used in biological research.

In simple words:

- ***Bioinformatics**: applying computational tools to biological data.*
- ***Computational Biology**: developing computational methods and theories for biology.*

Q3. What are the major applications of Bioinformatics?

Answer:

Bioinformatics has many important applications in biology, medicine, agriculture, and biotechnology. Some major applications are:

1. ***Drug Discovery and Development***
2. ***Personalized Medicine***
3. ***Gene Therapy***
4. ***Agriculture and Crop Improvement***

5. ***Disease Research***
6. ***Evolutionary Studies***
7. ***Environmental Applications***
8. ***Molecular medicine***
9. ***Biotechnology***
10. ***Crop improvement and Insect resistance***

Q4. What are Biological Databases and what are their main goals?

Answer:

Biological databases are organized collections of biological information such as DNA sequences, protein structures, genomes, and gene expression data. Their main goals are:

1. ***Data Storage****: storing biological information safely.*
2. ***Information Retrieval****: allowing scientists to search and access data easily.*
3. ***Knowledge Discovery*** *: helping researchers discover new biological relationships and scientific findings.*

Biological databases are classified into:

- ***Primary databases**: store original sequence data.*
- ***Secondary databases**: contain annotated and analyzed data.*
- ***Specialized databases**: focus on specific organisms or diseases.*

Q5. What are Nucleotide Sequence Databases?

Answer:

Nucleotide Sequence Databases are biological databases that store nucleotide sequences such as DNA, cDNA, and EST (Expressed Sequence Tag) sequences.

In eukaryotes, DNA contains exons and introns. Exons are transcribed into mRNA, and cDNA is produced from mRNA through reverse transcription. These cDNA sequences and ESTs are stored in nucleotide databases for research and analysis.

Major nucleotide sequence databases include:

- *NCBI*
- *EMBL*

- *DDBJ*

These databases collaborate internationally through INSDC.

Q6. What is GenBank and why is it important?

Answer:

GenBank is one of the largest nucleotide sequence databases in the world. It was first developed in 1982 at Los Alamos National Laboratory under the leadership of Walter Goad and is now maintained by NCBI.

GenBank is important because:

- *It stores huge amounts of genomic and nucleotide sequence data.*
- *Scientists worldwide use it for genetic and genomic research.*
- *It supports genome sequencing, disease research, and evolutionary studies.*
- *Its growth is exponential, with the number of bases doubling approximately every 18 months.*

GenBank also works with EMBL and DDBJ as part of the International Nucleotide Sequence Database Collaboration (INSDC).

Q7. What are Protein Databases and what type of information do they store?

Answer:

Protein databases are biological databases that store information related to proteins. These databases may contain:

- *Protein sequences*
- *Motifs (patterns of amino acids)*
- *Protein structures*
- *Structural alignments*

Protein databases help researchers study protein functions, structures, and interactions. They are very important in proteomics and structural bioinformatics.

The first biological sequences collected were protein sequences before nucleotide sequences. Early protein sequencing work was done by Frederick Sanger and Hans Tuppy in 1951.

Q8. What is PIR and how did it originate?

Answer:

PIR is a protein database resource developed from the collection of protein sequences assembled by Margaret Dayhoff and her collaborators at the National Biomedical Research Foundation (NBRF) during the 1960s.

Initially, the atlas mainly contained cytochrome protein sequences. Later, this collection became PIR, which is now a collaboration between:

- *NBRF*
- *Munich Center for Protein Sequences (MIPS)*
- *Japan International Protein Information Database (JIPID)*

PIR provides protein sequence and structural information for biological research.

Q9. What is UniProt and why is it important?

Answer:

UniProt is an international protein database created through collaboration between:

- *PIR*
- *EBI*
- *SIB*

UniProt was formed by combining:

- *Swiss-Prot*
- *PIR-PSD*
- *TrEMBL*

UniProt is important because it provides:

- *Protein sequences*
- *Protein ontology classifications*
- *Literature links*
- *Protein expression and modification data*

It is one of the most widely used resources for protein information and bioinformatics research.

Q10. What is PDB and what is its role in bioinformatics?

Answer:

Protein Data Bank (PDB) is a database that stores three-dimensional protein structures.

These structures are obtained using laboratory techniques such as:

- *X-ray crystallography*
- *Molecular biology methods*

Researchers submit experimentally determined protein structures to PDB, and scientists can use these structures to:

- *Study protein folding*
- *Compare predicted protein structures*
- *Perform structural bioinformatics research*

PDB is an important resource for understanding protein function and molecular interactions.

Q11. What is SCOP and how does it classify proteins?

Answer:

SCOP is a database that classifies proteins according to their structural features.

SCOP organizes proteins into different levels such as:

- *Class*
- *Fold*
- *Superfamily*
- *Family*
- *Domain*

Proteins are classified based on structures like:

- *Alpha helices*
- *Beta sheets*
- *Alpha/Beta structures*
- *Alpha + Beta structures*

SCOP helps scientists understand evolutionary relationships and similarities between protein structures.

Q12. What was the first free-living organism genome sequencing project?

Answer:

The first successful attempt to sequence a free-living organism was completed for Haemophilus influenzae in 1995. The project

was initiated at The Institute for Genomic Research under the leadership of Craig Venter.

The genome was sequenced using the shotgun sequencing method. The genome size was about 1.8 million base pairs, and the project took about 9 months with a cost of nearly 1 million US dollars. This success paved the way for future genome sequencing projects.

Q13. What is the Human Genome Project?

Answer:

The Human Genome Project was a scientific project started as a pilot project by the US Department of Energy (DOE) in 1986 to sequence the entire human genome.

Two major groups worked on this project:

- *NHGRI led by Francis Collins*
- *Celera Genomics led by Craig Venter*

Both groups announced completion of the human genome in 2000. The project identified approximately 3.4 billion DNA bases.

Q14. What are Genome Browsers and why are they useful?

Answer:

Genome browsers are web-based tools that allow scientists to visualize and explore genomes and their features.

One important genome browser is the UCSC Genome Browser.

Genome browsers help researchers:

- *View chromosomes and genes*
- *Zoom into specific genome regions*
- *Analyze ESTs, SNPs, and gene tracks*
- *Study genomic coordinates and annotations*

They make genome exploration easier and more interactive.

Q15. What is GEO and what type of data does it store?

Answer:

Gene Expression Omnibus (GEO) is a public database that stores gene expression data submitted by scientists worldwide.

GEO mainly stores:

- *Microarray data*

- *RNA-Seq data*
- *Gene expression profiles*

The database follows MIAME standards, which ensure proper experimental details are included.

GEO allows users to:

- *Browse and search datasets*
- *Analyze gene expression data*
- *Download raw and normalized data files*

Q16. What are the four main record types in GEO?

Answer:
GEO contains four major types of records:

1. ***Sample (GSM)***
   Stores sample details, treatments, and experimental design.
2. ***Platform (GPL)***
   Contains information about technologies such as microarray or RNA-Seq platforms.

3. ***Series (GSE)***
   A collection of related samples grouped together.

4. ***Dataset (GDS)***
   Contains processed and curated gene expression data assembled by GEO.

These records help organize and manage gene expression experiments effectively.

Q17. What is PubMed and why is it important?

Answer:
PubMed is a medical literature database developed by the National Library of Medicine.

PubMed contains millions of citations from:

- *MEDLINE*
- *Biomedical journals*
- *Life science books*

It is important because it helps researchers and students search for scientific articles, medical studies, and biomedical literature.

Q18. What is MEDLINE?

Answer:

MEDLINE is the primary database for biomedical journal articles.

It contains millions of references related to:

- *Medicine*
- *Health sciences*
- *Biomedicine*

MEDLINE uses the MeSH vocabulary system to organize medical terms and improve literature searching.

Q19. Why are sequences submitted to biological databases?

Answer:

Sequences are submitted to biological databases to share scientific data with the research community. Submission is often required by journals and funding agencies before publication.

Sequence submission helps:

- *Store biological information permanently*

- *Allow public access to scientific data*
- *Support future research and analysis*

Researchers must follow database guidelines and use proper sequence formats during submission.

Q20. What are BankIt and Sequin?

Answer:

NCBI provides two sequence submission tools:

1. ***BankIt***
    - *Used for simple sequence submissions*
    - *Web-based submission system*
    - *Suitable for small datasets*
2. ***Sequin***
    - *Used for complex sequence submissions*
    - *Supports offline submissions*
    - *Useful for large datasets and advanced annotations*

These tools help researchers submit nucleotide sequences to GenBank.

Q21. What is DNA Sequence Retrieval?

Answer:

DNA sequence retrieval is the process of searching and obtaining DNA sequence information from biological databases such as NCBI.

Researchers can retrieve:

- *Gene sequences*
- *mRNA sequences*
- *Coding regions*
- *Protein translations*
- *Genomic coordinates*

The GenBank format is commonly used for storing and displaying DNA sequence information.

Q22. What is UniProt and how is it used for protein sequence retrieval?

Answer:

UniProt is a major protein database created through collaboration between:

- *PIR*
- *EBI*
- *SIB*

UniProt allows users to retrieve:

- *Protein sequences*
- *Functional annotations*
- *Protein domains*
- *Taxonomy information*
- *Gene ontology data*

It is widely used for protein sequence analysis and bioinformatics research.

Q23. What is PDB and what information does it provide?

Answer:

Protein Data Bank (PDB) is a database that stores three-dimensional protein structures.

PDB provides:

- *Protein structural models*
- *Alpha helices and beta sheets information*
- *Structural annotations*
- *3D visualization of proteins*

Scientists use PDB to study protein folding, interactions, and structural bioinformatics.

Q22: What is FASTA sequence format?

Answer:

FASTA is a simple and widely used format for DNA and protein sequences. It starts with a ">" header line followed by the actual sequence in multiple lines.

Q23: What does the ">" symbol represent in FASTA format?

Answer:

The ">" symbol indicates the beginning of a new sequence and contains information like ID and description of the sequence.

Q24: What is GenBank format?

Answer:

GenBank format is a detailed biological sequence format that includes annotations like LOCUS, DEFINITION, ACCESSION, FEATURES, and the actual sequence under ORIGIN section.

Q25: How does a GenBank file end?

Answer:

A GenBank file ends with two forward slashes "//" which indicates the end of a sequence entry.

Q26: What is EMBL format?

Answer:

EMBL format is similar to GenBank format and is used in European databases. It contains structured information like ID, AC, DE, and SQ sections.

Q27: What is SwissProt format?

Answer:

SwissProt is a protein sequence database format that provides highly detailed and curated information about proteins including function, structure, and references.

Q28: What is XML in bioinformatics?

Answer:

XML (Extensible Markup Language) is a structured format used to store and exchange biological data in a machine-readable and human-readable form.

Q29: What is Genome Informatics?

Answer:

Genome Informatics is the use of computational tools and statistical methods to analyze genome sequences and extract biological information.

Q30: What is gene prediction in genome analysis?

Answer:

Gene prediction is the process of identifying regions of DNA that encode genes using computational methods and sequence patterns.

Q31: What are transposable elements?

Answer:

Transposable elements are DNA sequences that can move from one location to another within the genome and are also known as "jumping genes".

Q32: What are insertion sequences (IS elements)?

Answer:

IS elements are the simplest type of transposable elements found in prokaryotes. They contain only genes required for transposition.

Q33: What is the difference between introns and exons?

Answer:

Exons are coding regions that form proteins, while introns are non-coding regions removed during mRNA processing.

Q34: What is a pseudogene?

Answer:

A pseudogene is a non-functional gene formed due to mutations that prevent it from producing a functional protein.

Q35: What is comparative genomics?

Answer:

Comparative genomics is the study of comparing genomes of different organisms to find similarities and evolutionary relationships.

Q36: What are orthologs?

Answer:

Orthologs are genes in different species that evolved from a common ancestor and usually have the same function.

Q37: What are paralogs?

Answer:

Paralogs are genes within the same organism that arise from gene duplication and may evolve new functions.

Q38: What is Gene Mapping?

Answer:

Gene mapping is the process of locating genes on a chromosome and determining the relative distances between them. It helps to understand gene organization and inheritance patterns.

Q39: Difference between genetic and physical mapping?

Answer:

- *Genetic mapping is based on recombination frequency and is measured in centiMorgans (cM).*
- *Physical mapping is based on actual DNA distance and is measured in base pairs (bp).*

Q40: What is synteny?

Answer:

Synteny refers to the arrangement and order of genes on a chromosome. It helps in comparing genomes of different organisms.

Q41: What is conserved synteny?

Answer:

Conserved synteny occurs when different species have similar gene order on chromosomes due to a common ancestor. It shows evolutionary relationships.

Q42: What is genome annotation?

Answer:

Genome annotation is the process of identifying genes and assigning functions to them after genome sequencing and assembly.

Q43: What is structural annotation?

Answer:

Structural annotation identifies gene features such as promoters, exons, introns, and regulatory sequences in a genome.

Q44: What is functional annotation?

Answer:

Functional annotation assigns biological roles to genes, such as enzyme function, protein activity, gene expression, and regulation.

Q45: What is MAGPIE tool?

Answer:

MAGPIE is an automated genome analysis tool used for structural annotation. It identifies gene features and organizes genome data.

Q46: What is GENEQUIZ?

Answer:

GENEQUIZ is a tool used for functional annotation. It predicts protein function by comparing it with known protein databases.

Q47: What is EC number?

Answer:

EC number is a 4-digit system used to classify enzymes based on the reactions they catalyze. Each digit represents specific enzyme information.

Q48: What is genome sequencing?

Answer:

Genome sequencing is the process of determining the exact order of nucleotides (A, T, G, C) in DNA. It helps in studying genetic information.

Q49: What was the first sequenced genome?

Answer:

The first sequenced genome was RNA virus MS2. The first bacterial genome was Haemophilus influenzae.

Q50: What is MS2 virus?

Answer:

MS2 is a single-stranded RNA virus that infects E. coli. It contains genes for coat protein, replication, and lysis.

Q51: What is automated DNA sequencing?

Answer:

Automated DNA sequencing uses machines to rapidly sequence DNA using Sanger's method. It improved speed and accuracy of sequencing.

Q52: What is Sanger sequencing?

Answer:

Sanger sequencing is a DNA sequencing method based on chain termination. It was widely used in early genome projects.

Q53: What is gene annotation purpose?

Answer:

Gene annotation helps to identify genes, predict their function, and understand genome structure after sequencing.

Q54: What are introns and exons?

Answer:

- *Exons are coding regions that produce proteins.*
- *Introns are non-coding regions removed during RNA splicing.*

Q55: What is the importance of synteny?

Answer:

Synteny is important because it helps compare gene order between species and understand evolutionary relationships.

Q56: What are Sequence Motifs? Explain their biological importance.

Sequence motifs are short, conserved patterns in DNA that have biological significance. They are usually found in regulatory regions of genes and act as binding sites for transcription factors such as NF-κB. These motifs control gene expression by helping turn genes ON or OFF during biological processes.

Motifs are typically located upstream of genes and are important in processes like transcription regulation and RNA-level control.

Even if surrounding DNA sequences differ, these small conserved patterns remain similar across genes. This helps scientists identify functional regions in unknown DNA sequences.

Motif discovery is important in bioinformatics because it helps identify regulatory elements when the exact location or sequence is unknown. These motifs are used to understand gene regulation and disease mechanisms.

Q57: Explain Brute Force Motif Search Algorithm with formula and complexity.

Brute force motif search is a method that checks all possible combinations of l-mers from t DNA sequences. For each sequence, all possible starting positions are selected and combined to form candidate motifs. Each candidate is scored and the best one is selected.

Algorithm: BRUTEFORCEMOTIFSEARCH(DNA, t, n, l)

1. *bestScore = 0*
2. *Generate all (s1, s2, ..., st) combinations*
3. *Compute Score(s, DNA)*
4. *Update best motif if score improves*

Formula:

- *Number of combinations:*

$$(n - l + 1)^t$$

- *Time complexity:*

$$O(l \cdot n^t)$$

This algorithm is accurate but very slow due to exponential growth in search space.

Q58: Explain Simple Motif Search Algorithm and NEXTVERTEX usage.

Simple motif search improves brute force by using structured tree traversal. It explores all possible motif combinations step-by-step instead of random checking. The algorithm evaluates each full solution using a scoring function.

NEXTVERTEX is used to move through the search space systematically. It ensures that all combinations of motif positions are generated without repetition.

The algorithm updates the best motif whenever a higher score is found. It is still exhaustive but more organized than brute force.

Q59: Explain NEXTVERTEX Algorithm in detail.

NEXTVERTEX is a tree traversal algorithm used to generate all possible combinations of motif positions. It works like backtracking in a search tree.

If the current level is less than the maximum:

- *It moves deeper into the tree*
  If it reaches the last level:
- *It either increments the current value or backtracks*

This ensures that all possible combinations are visited exactly once. It is widely used in combinatorial bioinformatics problems such as motif finding and median string search.

Q60: Explain Branch and Bound Motif Search Algorithm with formula.

Branch and Bound improves brute force by skipping unpromising branches. It calculates an optimistic score for partial solutions. If the score cannot beat the current best solution, that branch is skipped using BYPASS.

Algorithm idea:

1. *Start with initial motif*

2. *Compute partial score*
3. *Estimate optimistic score*
4. *If not promising → skip branch*
5. *Otherwise continue search*

Formula:

$$optimisticScore = Score(s, i, DNA) + (t - i) \cdot l$$

This method reduces unnecessary computations and improves efficiency.

Q61: Explain Brute Force Median Search Algorithm with formula.

Brute force median search finds an l-mer that minimizes total distance between itself and all DNA sequences. It tests every possible nucleotide string from AAA… to TTT…

Algorithm:

1. *Generate all possible l-mers*
2. *Compute TOTALDISTANCE for each word*
3. *Select word with minimum distance*

Formula:

$Total\ candidates = 4^l$

$Objective:\ minimize\ TOTALDISTANCE(word, DNA)$

This method is accurate but computationally expensive.

Q62: Explain Simple Median Search Algorithm.

Simple median search uses NEXTVERTEX to generate all possible l-mers in a structured way. Each generated word is converted into a nucleotide sequence and evaluated using TOTALDISTANCE.

The algorithm keeps track of the best word found so far and updates it when a better solution appears. It ensures complete search but in an organized manner.

However, it still suffers from high computational cost when l increases.

Q63: Explain Branch and Bound Median Search Algorithm with formula.

This algorithm improves simple median search by pruning unnecessary branches. It uses prefix evaluation to estimate whether a solution can improve the best result.

Algorithm:

- *Compute prefix distance*
- *Estimate optimistic distance*
- *Skip branch if not promising*

Formula:

$$optimisticDistance = TOTALDISTANCE(prefix, DNA)$$

If:

$$optimisticDistance > bestDistance$$

→ branch is skipped (BYPASS)

This significantly reduces computation time.

Q64: Explain Greedy Motif Search Algorithm.

Greedy motif search builds motifs step by step by selecting the best possible l-mer from each DNA sequence. It does not explore all combinations but chooses locally optimal solutions.

Algorithm:

1. *Start with first sequence*
2. *Select best scoring l-mer pair*

3. *Extend selection to next sequences*

4. *Repeat until all sequences are included*

Score Function:

$$Score(s, i, DNA)$$

It is fast but may not always give the global optimal solution.

Q65: What are Genome Rearrangements? Explain with examples.

Genome rearrangements refer to changes in gene order due to evolutionary events such as inversion, duplication, deletion, or translocation. These changes affect synteny between species.

For example, human and mouse genomes share many synteny blocks, but gene order is different due to rearrangements. This shows they evolved from a common ancestor.

Genome rearrangement analysis helps understand evolution, species divergence, and genetic diseases.

Q66: Compare Brute Force, Greedy, and Dynamic Programming.

Brute force checks all possible solutions, greedy selects locally best choices, and dynamic programming breaks problems into smaller subproblems.

For example, in the Manhattan Tourist Problem, greedy fails because local choices do not guarantee optimal results. Instead, dynamic programming is used.

DP Formula:

$$s_{i,j} = max\ (s_{i-1,j}, s_{i,j-1}) + weight(i,j)$$

DP ensures optimal solution with efficient computation.

Q67: Define Sequence Similarity and Edit Distance with formula.

Sequence similarity measures how closely two DNA sequences match. Because mutations include insertions, deletions, and substitutions, direct comparison is not enough.

Edit distance measures the minimum number of operations needed to transform one sequence into another.

Formula:

$$d(v,w) = n + m - 2s(v,w)$$

Where $s(v, w)$ is the LCS length. It is widely used in genome comparison and mutation analysis.

Q68: Explain Alignment, Edit Graph, and LCS with recurrence.

Sequence alignment arranges two DNA sequences to identify matches, mismatches, and gaps. Edit graph represents alignment as a grid where moves include horizontal, vertical, and diagonal transitions.

LCS finds the longest common subsequence using dynamic programming.

Recurrence:

$$s_{i,j} = max \begin{cases} s_{i-1,j} \\ s_{i,j-1} \\ s_{i-1,j-1} + 1 if\, v_i = w_j \end{cases}$$

This method is fundamental in bioinformatics for sequence comparison and evolutionary studies.

Q69. What is a PAM matrix and why is it used?

Answer:

PAM matrix is a scoring table used in bioinformatics. It helps compare protein sequences. It shows how amino acids change

over time. Some amino acids replace others without changing function. PAM means Point Accepted Mutation. PAM1 shows 1% mutation level. It is used to study protein evolution. It helps find similarity between proteins. It is useful in sequence alignment. It is widely used in biology research.

Q70. What is PAM1 matrix?

Answer:

PAM1 matrix is built from very similar proteins. It represents small evolutionary changes. Scientists compare amino acids in aligned sequences. They calculate how often changes happen. These values become probabilities. PAM1 shows 1% mutation in proteins. It is the base matrix for all PAM models. It is used to create larger PAM matrices. It helps study protein similarity. It is important in evolution studies.

Q71. What does g(i,j) mean?

Answer:

g(i,j) means probability of amino acid change. It shows how i changes into j. It is calculated from real protein data. It is based on observed mutations. Higher values mean common substitutions. Lower values mean rare changes. It is used in

scoring matrices. It helps measure protein evolution. It is important for alignment tools. It improves biological analysis.

Q72. Write PAM scoring formula.

Answer:

PAM scoring formula is:

$$\delta(i,j) = \log (g(i,j) / (f(i) \times f(j)))$$

It compares real mutation with random chance. f(i) and f(j) are amino acid frequencies. g(i,j) is observed mutation probability. High score means strong similarity. Low score means weak similarity. It is used in protein alignment. It helps detect evolutionary relationships. It is widely used in bioinformatics.

Q73. What is PAM-n matrix?

Answer:

PAM-n matrix shows long-term evolution. It is created by multiplying PAM1 many times. PAM(n) = (PAM1)^n. Higher n means more mutations. It shows distant relationships between proteins. It is used for species comparison. It helps detect weak similarities. It models long evolution time. It is important in advanced alignment. It is used in research.

Q74. Why is PAM matrix important?

Answer:

PAM matrix helps compare protein sequences. It shows evolutionary relationships. It detects similarities in different organisms. It is used in alignment tools. It helps understand protein function. It is based on real mutation data. It improves sequence matching. It is useful in bioinformatics. It helps study evolution. It is widely used in research.

Q75. What is ORF?

Answer:

ORF means Open Reading Frame. It is a DNA sequence that can code protein. It starts with ATG codon. It ends with a stop codon. It shows possible gene regions. Long ORFs are more important. Random DNA has many stop codons. ORF helps in gene prediction. It is used in bioinformatics. It is a key concept in genetics.

Q76. Why are long ORFs important?

Answer:

Long ORFs are likely real genes. Random DNA cannot stay long without stop codons. So long ORFs are meaningful. They may

produce proteins. They are used in gene detection. However, not all ORFs are genes. Other evidence is also needed. They are first step in gene finding. They help identify coding regions. They are important in biology.

Q77. What is codon usage bias?

Answer:

Codon usage bias means preference of codons. Some codons are used more than others. Coding regions show specific patterns. Non-coding regions are different. It helps identify real genes. It improves ORF detection. It is species dependent. It is used in bioinformatics. It helps gene prediction. It is important in genetics.

Q78. What is Translation Start Site (TSS)?

Answer:

TSS is the start point of protein synthesis. It usually begins with ATG codon. ATG codes for methionine amino acid. Surrounding bases help identify it. Certain patterns appear near TSS. These patterns are used in prediction. It shows where translation starts. It is important in gene expression. It helps find gene beginnings. It is key in biology.

Q79. How is TSS predicted?

Answer:

TSS is predicted using nucleotide patterns. Scientists study bases around ATG. Some bases appear more frequently. A scoring system is used. High score means strong TSS signal. Low score means weak signal. Formula uses probability values. It helps identify start sites. It improves gene prediction. It is widely used in bioinformatics.

Q80. What is information content formula?

Answer:

Information content formula is:

$Si = \Sigma \log (Fi(X) / 0.25)$

Fi(X) is observed frequency of base. 0.25 is random probability. Higher value means strong pattern. Lower value means weak pattern. It shows conserved positions. It is used in TSS prediction. It helps detect gene start sites. It is important in bioinformatics. It improves accuracy.

Q81. What is splice junction?

Answer:

Splice junction is boundary in genes. It separates exons and introns. Donor site starts with GT. Acceptor site ends with AG. These are common biological patterns. They help remove introns. They are important in RNA processing. They help form final mRNA. They are used in gene prediction. They are key in genetics.

Q82. What is canonical splice site?

Answer:

Canonical splice site is GT–AG rule. Most introns follow this rule. GT is donor site. AG is acceptor site. About 99% introns follow it. It is very common in biology. It helps identify genes. It is used in prediction models. It is important in gene structure. It is widely studied.

Q83. How are splice sites predicted?

Answer:

Splice sites are predicted using patterns. Scientists search GT and AG signals. They also use scoring methods. Score = Σ log (Fi(X)/0.25). High score means strong site. Low score means weak site. It helps find intron boundaries. It is used in gene

annotation. It improves prediction accuracy. It is important in bioinformatics.

Q84. What are exons?

Answer:

Exons are coding parts of genes. They are used to make proteins. They remain in final mRNA. Introns are removed during splicing. Exons carry genetic information. They are very important. They are found using prediction tools. They help understand gene structure. They are key in biology. They form functional proteins.

Q85. How are exons predicted?

Answer:

Exons are predicted using multiple signals. ORF, splice sites, and coding scores are used. High scoring regions are selected. Low scoring regions are ignored. Machine learning methods are used. LDA helps classification. It separates exon and intron. It improves accuracy. It is used in gene prediction. It is important in genomics.

Q86. What is LDA?

Answer:

LDA means Linear Discriminant Analysis. It is a classification method. It separates two groups using a line. In biology, it separates exons and introns. It uses feature values for decision. It finds best boundary line. It reduces classification errors. It is used in prediction models. It improves gene finding. It is widely used.

Q87. What is genome annotation?

Answer:

Genome annotation is identifying genes in DNA. It explains genome function. It finds coding and non-coding regions. It is done after genome sequencing. It gives meaning to DNA sequence. It helps understand organisms. It is very important in biology. It is final step of genome analysis. It is used in research. It helps in medicine.

Q88. What are ab initio methods?

Answer:

Ab initio methods use only DNA sequence. They do not need external data. They detect patterns in DNA. Tools include GeneMark and AUGUSTUS. They are fast methods. But sometimes less accurate. They are used in early analysis. They

help find genes quickly. They are important in bioinformatics. They are widely used.

Q89. What are evidence-based methods?

Answer:

Evidence-based methods use real biological data. They use RNA or protein evidence. They are more accurate than ab initio. Tools include TopHat and Cufflinks. They confirm real genes. They are slower but reliable. They improve genome annotation. They are widely used in research. They help in accurate prediction. They are important.

Q90. What is pattern finding in genome?

Answer:

Pattern finding means searching DNA sequences. It finds important biological signals. It may allow mismatches and gaps. It uses scoring systems. It helps detect genes and motifs. It is used in bioinformatics tools. It improves sequence analysis. It is important in research. It helps understand DNA. It is widely used.

Good Luck ☺

Click to contact us:
Sir Mughal: https://wa.me/923132310940
Sir Jawad: https://wa.me/923036100859

BS BIOLOGICAL SCIENCES

Sir Mughal
Sir Jawad

Join Us!

Phone
0313 2310940

Phone
0303 6100859

GROUP FOR STUDENTS

- BS (Hons) Zoology
- BS (Hons) Biotechnology
- BS (Hons) Bioinformatics

Click to contact us:

Sir Mughal: https://wa.me/923132310940

Sir Jawad: https://wa.me/923036100859