Zoology Legends

BIF501

Final Term Papers Fall 2023

Muhammad Zaman 👆😍

0304-4756496

Team Zoology LEGENDS

- ORF stand for ----- **open reading frame.**
- IUB stand for---- **international union of biochemistry**
- PAM stand for ---**pointed accepted mutation**
- FTP stand for--- **file transfer protocol**.
- LINE stand for ---**long interspersed nuclear element**
- SINE stand for ---**short interspersed nuclear element**
- PIR stand for--- **protein information resource**
- HMM stand for ---**Hidden Markova model**
- GCG stand for--- **genetic computer group**
- TIGR stand for --- **the institute of genome research**
- XML stand for – **extensible markup language**
- PDGF stand for ---**platelet derived growth factor**
- NCBI stand for ---**national center for biotechnology information**
- EMBL stand for --- **European molecular biological lab**
- DDBJ stand for ---**DNA data bank of japan**
- GEO stand for--- **gene expression omnibus**
- STS stand for **sequence tag sites**
- BLOSUM stand for **block substitution matrix**
- PSSM stand for **position specific scoring matrix**

- **PAM** matrices do not perform well in case of distantly related proteins.
- **PAM** (Point accepted mutations) using substitution data from **similar proteins** then extrapolating this information to longer evolutionary distances
- Ab initio methods have **Anfinsen's** thermodynamic hypothesis at the center.
- These methods attempt to identify the structure with **minimum** free energy.
- Ab initio methods rely on computing the energies of **folded proteins.**
- The protein structures with the lowest energy are deemed as **plausible predictions**.
- BAC Abbreviation__**Bacterial Artificial Chromosome**
- GEO abbreviation__**Gene Expression Omnibus**
- DNA and histone form together____ **nucleosomes**
- Department Of Energy (**DOE**)
- National Human Genome Research Initiative (**NHGRI**)
- Human genome project begun by Department
- Of Energy (DOE) of USA in **1986**

- Association rule has two parts___ **• an antecedent (if) • a consequent (then)**
- Bioinformatics required two things **1. Smart mind 2. Internet connected computer**
- technique used for protein not present in database____
- 3D-1D Algorithm__Proposed by **Bowie et al in 1991**. It converts 3D structure into a 1-D string profile for each structure in the fold library.
- **De novo** protein sequencing algorithms- invaluable for **known, unknown** proteins
- Find a protein from a database that best matches the **experimental spectrum**.
- The technique which is used to identify different individual by DNA printing is called?__**DNA fingerprinting**
- **PCR**__Replicate a small amount of DNA to create a larger sample for analysis.
- X-RAY use techniques__ **Crystallography and**
- **NMR Spectroscopy**
- BLAST stands for? **Basic Local Alignment Search Tool**
- Fast Alignment (**FASTA**)

- Concerned with the effect of genetic factors on reactions to drugs.? **Pharmacogenetics**
- a collection of objects which are “similar” between them and are.............. to the objects belonging to other.?_**Dissimilar**
- How many Protein Structures?__**Four**
- Correct sequence? **Homology: Paralog vs. Orthology**
- Alpha helices, beta sheets and turns can be predicted using?_**Chou Fasman Algorithm**
- Alpha Helices,...............can be predicted? **Beta sheets**
- protein structures is very........... to crystallize? **Difficult**
- C-terminal ions are called? **y-ions**
- First row and first column of the matrix is filled with?_**Zero**
- .... scores can be derived from probabilistic model ?_**Substitution**
- Bowie et al in? **1991**
- IUB stand for ..... **International Union of Biochemistry**
- ADMET stand for ..... **Absorption, distribution, metabolism, excretion, and toxicity**
- Human T-cell receptor locus-Trypsinogen gene (**4 kb**).

- IF Amino acid carry 3 billions-letter sequence **500**-letter reads
- The algorithm never needs to move backwards that makes the algorithm good for processing very large files__**KMP Algorithm**
- Need for Ab Initio Modelling
  - Applicable to any sequence
  - Not very accurate biologically
  - Accuracy and applicability are limited by our understanding of the protein folding

  All of these
- Chau's fasman uses Amino acid to start process ..... Such as **4 amino** acids are needed to start an Alpha Helix and **5 amino acids** for Beta Sheet
- Genome project started in **1986**
- RNA sequence pick krna tha ..... **Ensembl**
- The sequence starts with a '**greater than'** sign (>)
- Sequence start from the amino acid residue...... **N-terminus**
- 3D ID Algorithm was present Bowie in which year ...**1991**
- European Molecular biology Lab (**EMBL** established 1980)

- The enzyme involved in DNA transcription is called **RNA polymerase**
- Hamming distance is not typically used to compare **DNA** or protein sequences.
- Gel electrophoresis may be used to measure......... of restriction fragments...? **size**
- Hamming distance is is for embryonic development? **cansus**
- technique for DNA similarity...? **DNA Finger Printing / DNA Profiling**
- **Pharmacogenetics**: Gene is a branch of pharmacology concerned with the effect of genetic factors on reaction to drugs.
- **De novo assembly**: A programming approach that does not require prior sequence alignment.
- **Bioinformatics**: Bioinformatics is a science of managing, organizing, obtaining, analyzing, and processing biological data.
- **Structure**: Structure is responsible for the function of a protein.
- **PDB**: PDB stands for Protein Data Bank.

- **True**: Secondary structures, such as alpha helices and beta sheets, are formed due to bonding between amino acids.
- Predicting protein secondary structure: **Chou-Fasman Algorithm** is first used in predicting protein secondary structure.
- **Apoptosis**: Programmed cell death is also known as apoptosis.
- AG: Needleman-Wunsch algorithm starts from "**AG**".Regions?
- **Protease**: Traceback of enzyme is used to degrade protein at certain points.
- **Requires a template structure**: This statement is not suitable for ab initio homology modeling methods. Ab initio methods do not rely on a **template** structure.
- Orthology: **Orthology** refers to genes in different species that evolved from a common ancestor and typically retain the same function.
- **True**: Chou-Fasman Algorithm helps predict secondary structures such as alpha helices and beta sheets.

- **Interdisciplinary knowledge**: Bioinformatics requires interdisciplinary knowledge in biology, computer science, statistics, and mathematics.
- **Ab initio modeling**: Ab initio modeling is used when no template is found for homology modeling.
- **False**: Not all mutations alter the functions of a protein. Some mutations may have no effect or may even be beneficial.
- **False**: Consecutive nucleotide changes can result in significant changes in the protein, depending on the amino acids encoded by the changed codons.
- Complete set of an organism's DNA: A **genome** sequence is the complete set of an organism's DNA, including all of its genes and non-coding regions.
- **Both**: De novo protein sequencing algorithms can be used for both known and unknown proteins.
- **True**: An algorithm helps predict secondary structures such as alpha helices, beta sheets, and turns.
- **Protein folds**: Stable folds in nature refer to the stable three-dimensional structures adopted by proteins.

- **Exon prediction**: Exons can be predicted within ORF from the information gathered about splice junctions.
- Alignment from the **beginning to the end**: End-to-end alignment is alignment from the beginning to the end of sequences.
- ***Salt concentration***: Salt concentration has no effect on the formation of hybrid DNA.
- **False**: The statement is incomplete. Please provide more context for clarification.
- Basic Local Alignment Search Tool: **BLAST** is a tool used for comparing biological sequences.
- **Shortest branch length**: In UPGMA, two sequences with the shortest branch length are considered.
- **Correct statement**: Evolutionary distance between them is considered when determining the clustering.
- **True**: A cluster is a collection of objects that are similar to each other and dissimilar to objects in other clusters.
- **Direct instruction**: Different learning methods do not include direct instruction.

- **All of the above**: Various factors such as dataset size, feature selection, and model complexity can affect the performance of a learner system.
- **DNA fingerprinting**: The technique to distinguish individuals based on their DNA print patterns is called DNA fingerprinting.
- **True**: NCBI provides options for sequence submission through platforms such as GenBank and the Sequence Read Archive (SRA).
- They contribute a very small proportion in yeast i.e. only **239** introns
- A neural network is a "**connectionist**" computational system
- Information is processed collectively in **parallel** throughout a network of nodes.
- Deleting **three** consecutive nucleotides results in minor changes in the protein
- An important point to note here is that 2' structures are formed due to **hydrogen bonding** between amino acids
- **Brute force** is an exhaustive search method
- **SEQUEST** Algorithm Exhaustive search approac

Current Subjective

Seven steps of homology modelling

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

Difference b/w intron and exon

Feature	Exons	Introns
Function	Protein-coding regions	Non-coding regions
Transcription	Transcribed into mRNA	Transcribed into pre-mRNA, then removed
Proportion	Smaller proportion in yeast	Larger proportion in human
Location	Interrupted by introns	Located between exons
Genome Content	Contributes to a small portion of the genome in yeast	Makes up a significant portion (95%) of the genome in humans(in Yeast 239)
Conserved Sequence	Presence of GT at 5' end and AG at 3' end (GT-------------AG)	Preserved trend throughout the genome

Describe in detail the Rationale of ab-initio modelling

Ab initio methods rely on computing the energies of folded proteins. The protein structures with the lowest energy are declared as plausible predictions

Ab initio modeling predicts a protein's 3D structure solely from its amino acid sequence, using basic physical and chemical principles:

No Similar Proteins: Used when closely related proteins aren't available for comparison.

New Discoveries: Helps predict structures when new proteins are found, expanding our knowledge.

Fundamental Principles: Relies on fundamental laws of physics and chemistry for folding predictions.

Energy Calculations: Computes energy released during folding to identify the most stable structure.

Valuable Insights: Provides crucial understanding of protein folding, enhancing our knowledge of function and interactions.

Application of drug discovery

The applications of drug discovery (DD) encompass several key areas that significantly influence the process. Here's an overview of these areas and their impacts:

Molecular Biology on Drug Discovery:

- ***Genetic Information***: Molecular biology provides insights into the genetic information related to diseases, which helps identify potential drug targets.
- ***Cloning and Gene Expression***: Techniques like cloning and gene expression enable the production of therapeutic proteins, which are crucial for drug development.

High-Throughput Screening (HTS):

- ***Pharmaceutical Industry Use***: HTS is widely employed in the pharmaceutical industry to rapidly screen large libraries of compounds for their biological activity.
- ***Automation***: Automation in HTS expedites the evaluation of numerous drug-like compounds, aiding in the identification of potential drug candidates.

Combinatorial Chemistry:

- ***Synthesis of Molecular Constructions***: Combinatorial chemistry facilitates the synthesis of millions of diverse molecular structures.
- ***Biological Activity Testing***: These synthesized compounds can be efficiently tested for their biological activity, accelerating the discovery of new drug candidates.

Procedure of RNA Sequences

- Collect biological samples containing RNA.
- Isolate RNA from samples.
- Convert RNA to cDNA (for mRNA analysis).
- Prepare RNA samples for sequencing.
- Use high-throughput sequencing to determine RNA sequences.
- Analyze sequencing data to identify and quantify RNA sequences.
- Assess the biological functions and interactions of identified RNA sequences.
- Confirm results through experimental validation techniques.

Define Target and template

Template Sequence: Known sequence of DNA is template, This is like a blueprint that we already know. It's a DNA sequence that serves as a model for processes like making copies or studying how DNA works.

Target Sequence: Unknown sequence is work as a target sequence. This is what we're interested in learning more about. It's a DNA sequence that we're trying to understand or study. We might not know much about it yet, and we want to figure out its details or how it works.

Modified Protein Identification Problem:

The Modified Protein Identification Problem (MPIP) is a computational challenge in proteomics that involves identifying a peptide sequence from a protein database that best matches an experimental spectrum, while allowing for a certain number of modifications.

Input: A database of proteins, an experimental spectrum S, a set of ion types Δ, a parent mass m, and a parameter k capping the number of modifications

Output: A protein of mass m with the best match to spectrum S that is at most k modifications away from an entry in the database. Modified Protein Identification problem P1 and P2- S1 and S2, Notion of spectral similarity, Shared peaks count, Limitations in detecting similarities by database search

FASTBLOCK in dynamic programing

```
FASTBLOCK (n, m)
if n and m are both even
    return L
else
    return W
```

What kind of gene expression in omnibus

Genes are expressed into mRNA, and whenever we talk about gene expression, we generally mean the mRNA sequences so we can normally get those mRNA from techniques like microarray and another famous technique nowadays which is being established is known as RNA seq.

And microarray data and RNA seq can be classified into Gene Expression Data which is stored in Gene Expression Databases.

Explain Primary, Secondary, Tertiary and Quaternary Structure of Protein?

Primary Structure:

The primary structure of a protein is the linear sequence of amino acids. Together, this linear sequence is referred to as a polypeptide chain. The amino acids in the primary structure are held together by covalent bonds, which are made during the process of protein synthesis (translation).

Secondary Structure:

The secondary structure of a protein refers to the local spatial arrangement of its amino acid residues into regular, recurring patterns. The two most common types of secondary structures in proteins are alpha helices and beta sheets.

Tertiary Structure:

The tertiary structure of a protein refers to the overall three-dimensional arrangement of its polypeptide chain, including the folding of secondary structures (such as alpha helices and beta sheets) and the spatial arrangement of protein domains. In the tertiary structure, amino acid side chains interact and form various bonds and interactions, such as hydrogen bonds, disulfide bonds, hydrophobic interactions, and electrostatic interactions. These interactions contribute to the stability and unique shape of the protein, which is essential for its function.

Quaternary Structure:

The quaternary structure of a protein refers to the organization of multiple protein subunits in a multi-subunit complex. This structure describes how individual protein subunits come together to form a larger functional unit. Protein complexes can range from simple dimers (two subunits) to large homo-oligomers and complexes with variable numbers of subunits. The quaternary structure is crucial for the function and regulation of many proteins, as it can influence their stability, activity, and interactions with other molecules.

Pharmacogenetics and its applications

It is the branch of pharmacology concerned with the effect of genetic factors on reactions to drugs.

It focuses on understanding how genetic factors can affect drug metabolism, efficacy, and adverse reactions.

Biotechnological science combines techniques of medicine, pharmacology & genomics which developing drug therapies to compensate for genetic differences in patients which cause varied responses to a single therapeutic regimen.

Applications:-

- Detection of genetic variability of drug effects on the genome level
- Agent selection
- Analysis of drug
- Reactions and drug toxicity on gene expression
- Development of new indications for already approved drugs
- Discovery of new drug targets
- Identification of (non) responders in clinical trials of phase I-IV
- Identification of genotype dependent adverse drug reactions
- Identification of individuals at risk for severe adverse drug effects

Dynamic programming and it's steps

Dynamic programming is used to find an optimal alignment of two sequences and its scores

It is a method by which a larger problem may be solved by first solving smaller, partial versions of the problem

Three steps in dynamic programming:

- Initialization
- Matrix filling (scoring)
- Traceback (alignment)

— **Initialization**: Create matrix with M+1 columns and N+1 rows where M and N correspond to the size of sequences to be aligned

— **Matrix filling**: Fill the matrix with highest possible score

— **Trace back:** Move from the last corner and follow the arrow

Applications of machine learning

- **Retail**: Market basket analysis, Customer relationship management (CRM)
- **Finance**: Credit scoring, fraud detection
- **Manufacturing**: Optimization, troubleshooting
- **Medicine**: Medical diagnosis
- **Telecommunications**: Quality of service optimization
- **Bioinformatics**: Motifs, alignment
- **Web mining**: Search engines
- **Genomics**

Optimal alignment and sequence

Optimal alignment refers to the process of finding the best possible alignment between two sequences, such as DNA, RNA, or protein sequences. The goal of optimal alignment is to identify the most likely evolutionary relationship between the sequences by maximizing the similarity of matching residues and minimizing the number of gaps introduced.

Sequence alignment involves arranging the sequences in a way that maximizes the number of identical or similar residues in corresponding positions. This is typically done using algorithms such as Needleman-Wunsch or Smith-Waterman for global and local sequence alignments, respectively.

Transposable elements

Transposable Elements are the elements which can transpose i.e. they can move from one place to another in the genome and then they can cause the repetitive elements (repetitive DNA within that genome)

➢ They make up a significant part of organisms' genome especially in that of the eukaryotic genome.

➢ They move within and across genomes and

➢ Causes genome expansion

Insertion Sequences (IS elements) are simple transposable elements found in prokaryotes. They are small (< 1-2 Kb) and contain inverted repeat sequences (IRs). They encode at least one gene for transposition and do not code for noticeable traits.

Transposons:

Transposons are more complex transposable elements as compared to the simple IS (Insertion Sequence) elements and they code for additional characters in addition to the gene responsible for their transposition

Chou Fasman algorithm

The Chou-Fasman algorithm is a method used in bioinformatics to predict protein secondary structure elements, such as alpha helices, beta strands, and turns. It was developed by Peter Y. Chou and Gerald D. Fasman in the late 1970s.

Chou Fasman Algorithm helps predict secondary structures such as Alpha Helices, Beta Sheets and Turns. Step by step flowchart of the entire algorithm. Beta sheets can be predicted from primary amino acid sequences

This algorithm is based on statistical analysis of amino acid occurrences in proteins.

Homology vs paralogy, Orthology

Difference	Homology	Paralogy	Orthology
Definition	Homology refers to the similarity between organisms or structures that arises from shared ancestry	Paralogy refers to the presence of multiple copies of a gene within the same genome	Relationship between genes from different species that originated from a common ancestral gene
Relationship	Similarity within or between species	Duplication within the same species	Between different species
Example	Hemoglobin genes in humans and mice	Multiple copies of a gene in humans	Insulin genes in humans and mice
Function	Often have similar functions	May diverge in function due to duplication	Tend to retain similar functions
Evolutionary Outcome	May diverge or remain similar	May develop different roles	Usually retain similar functions

Pharmacogenomics its application

Pharmacogenomics is the study of how an individual's genetic makeup influences their response to drugs.

- Improve drug safety
- Reduce ADRs (Adverse Drug Reactions)
- Tailor treatments to meet patients unique genetic predisposition
- Optimal dosing
- Improve drug discovery and Improve proof of principle for efficacy trials.

DNA & RNA Applications

DNA Applications:

- Genetic Testing
- Forensic Science
- Personalized Medicine
- Cancer Research
- Infectious Disease Diagnosis
- Biotechnology
- Drug Discovery
- Environmental Monitoring
- Agriculture and Food Safety
- Evolutionary Biology

RNA Applications:

- Gene Expression Analysis
- RNA Interference (RNAi)
- mRNA Vaccines
- RNA Sequencing
- RNA Editing
- Antisense Therapy
- Ribozyme Technology
- RNA-based Therapeutics
- RNA Nanotechnology
- Evolutionary Biology

Protein sequence

Protein sequence refers to the linear arrangement of amino acids in a protein molecule, dictated by the sequence of nucleotides in the corresponding gene.

Drugs discovery

Design & discovery of new compounds that are suitable for use as drugs

A team of workers—chemistry, biology, biochemistry, pharmacology, mathematics, medicine & computing

Tandem repeats and interspersed repeats

Tandem repeats refer to consecutive repetitions of DNA sequences occurring one after the other in the genome. They are often found clustered together.

Interspersed repeats, on the other hand, are repetitive DNA sequences that are separated by non-repetitive DNA sequences within the genome.

Local alignment output

The output of a local alignment provides a numerical score indicating the similarity between sequences, along with the alignment itself showing matched residues and gaps.

It also includes start and end positions of the aligned region, helping identify significant similarities for further analysis.

Orthologs and paralogs

Orthologs are genes found in different species that share a common ancestor and often retain similar functions.

Paralogs, on the other hand, are genes within the same species that have evolved from gene duplication events and may have diverged in function while sharing a common origin.

Types of RNA

- **mRNA - Messenger RNA**: Encodes amino acid sequence of a polypeptide.
- **tRNA - Transfer RNA**: Brings amino acids to ribosomes during translation.
- **rRNA - Ribosomal RNA**: With ribosomal proteins, makes up the ribosomes, the organelles that translate the mRNA.
- **snRNA - Small nuclear RNA**: With proteins, forms complexes that are used in RNA processing in eukaryotes. (Not found in prokaryotes.)

Haemophilus influenza: Haemophilus influenzae was the first genome that was published and the project was initiated at The Institute of Genome Research (TIGR) under the leadership of Craig Ventor.

At that time, a method which was already established known as shotgun sequencing method was being tested by this project to verify its reliability and efficiency. And by utilizing this method they sequenced the genome which was about 1.8 million base pairs (bp), it took 9-months and the cost was around 1 million US dollars and this project Paved the way for sequencing of many other organisms

Shotgun sequencing

Shotgun sequencing involves randomly breaking DNA or RNA into small fragments, sequencing them individually, and then assembling the sequences computationally to reconstruct the original genome or transcriptome. It's a high-throughput method used for sequencing large genomes or complex mixtures without prior knowledge of the target sequence.

Homology modelling last step model validation

In homology modeling, the last step is model validation, which ensures the accuracy and reliability of the generated model. This involves assessing geometric parameters, stereochemistry, and structural quality, comparing the model to experimental data or known structures. Validation ensures the model accurately represents the target protein's structure for further analysis

Difference b/w Prokaryotes and eukaryotes

- **Prokaryotes** are simple Genomes.

- They are easy models to study
- Sequencing is done on economically important organisms
- **Eukaryotes** have larger genomes
- Have tandem repeat
- Have introns in their protein-coding genes
- Heterochromatin and euchromatin region.

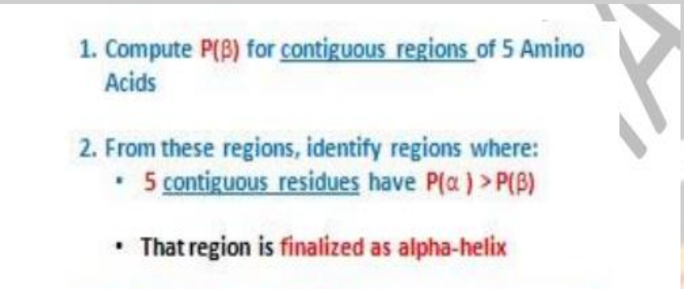
RNA splicing and its advantages

RNA splicing, a process in eukaryotic cells, removes non-coding regions (introns) from pre-mRNA transcripts and joins together coding regions (exons) to produce mature mRNA.

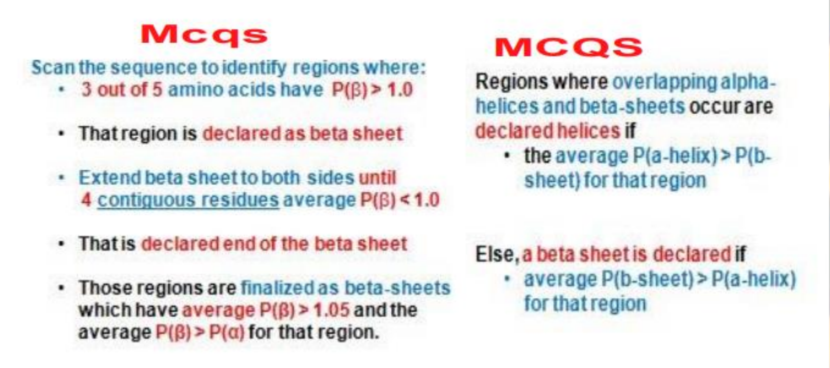
Its advantages include increased proteome diversity, regulation of gene expression, evolutionary innovation, and implications for disease. Splicing enhances protein variation, controls gene expression, drives evolution, and has disease implications.

Chou fasman ki propensity btani thi

Alpha Helices are formed from 4 contiguous amino acids having an Alpha-Helix propensity over 1.0. The Alpha-Helix stops if this propensity falls below 1.0. Alpha Helices were finalized if their propensity was higher than the propensity for Beta Sheets in regions of 5 amino acids



1. Compute P(β) for contiguous regions of 5 Amino Acids
2. From these regions, identify regions where:
   - 5 contiguous residues have P(α) > P(β)
   - That region is finalized as alpha-helix



Mcqs

Scan the sequence to identify regions where:

- 3 out of 5 amino acids have P(β) > 1.0
- That region is declared as beta sheet
- Extend beta sheet to both sides until 4 contiguous residues average P(β) < 1.0
- That is declared end of the beta sheet
- Those regions are finalized as beta-sheets which have average P(β) > 1.05 and the average P(β) > P(α) for that region.

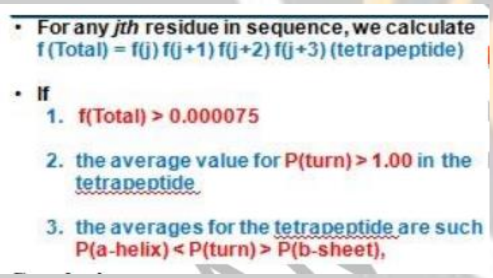
MCQS

Regions where overlapping alpha-helices and beta-sheets occur are declared helices if

- the average P(a-helix) > P(b-sheet) for that region

Else, a beta sheet is declared if

- average P(b-sheet) > P(a-helix) for that region



- For any *jth* residue in sequence, we calculate f (Total) = f(j) f(j+1) f(j+2) f(j+3) (tetrapeptide)
- If
  1. f(Total) > 0.000075
  2. the average value for P(turn) > 1.00 in the tetrapeptide
  3. the averages for the tetrapeptide are such P(a-helix) < P(turn) > P(b-sheet),

Difference b/w protein sequence and DNA sequence

Aspect	Protein Sequence	DNA Sequence
Composition	Comprised of amino acids	Comprised of nucleotides (A, T, C, G)
Function	Encodes protein structure and function	Stores genetic information for organisms
Biological Role	Acts as building blocks for proteins	Serves as the genetic blueprint
Translation	Translated from mRNA during protein synthesis	Transcribed from DNA during gene expression
Sequence Length	Typically shorter than DNA sequences	Typically longer than protein sequences
Genetic Code	Governed by the genetic code	Dictates the genetic information

Neural Network

The human brain can be described as a biological neural network an interconnected web of neurons transmitting elaborate patterns of electrical signals. A neural network is a “connectionist” computational system.

Information is processed collectively in parallel throughout a network of nodes.

Complex adaptive system.

- Learning processes in biological systems.
- Learning as an optimization process.
- Learning by modification of synaptic strength

Mass spectrometry

Mass spectrometry (MS) is an analytical technique used to identify and quantify molecules based on their mass-to-charge ratio. It involves ionizing chemical compounds to generate charged particles, which are then separated based on their mass and charge using electric and magnetic fields. The resulting spectra provide information about the molecular composition, structure, and abundance of the analyzed samples.

MS is widely used in various fields such as chemistry, biochemistry, pharmaceuticals, environmental science, and forensic analysis for molecular identification, characterization, and quantification.

Name 2 protein databases

Some widely used protein databases include:

- UniProt
- Protein Data Bank (PDB)
- NCBI Protein Database
- Ensembl Protein Database
- STRING Database
- GenBank

Threading input and output

In threading, the input is a protein sequence with an unknown 3D structure, and the output includes predicted models or alignments against known protein structures. Threading algorithms match the sequence to templates, suggesting potential structures based on similarity.

Strategy for Sequencing

The strategy involves cloning DNA into BACs, sequencing each BAC to simplify assembly. Used in the Human Genome Project, it employs mate-pair reads sequencing by Celera Genomics. Inserts of length L are sequenced from both ends, aiding in resolving repetitive regions.

How can match protein sequence

Protein sequence matching involves comparing amino acid sequences to identify similarities and infer functional relationships.

In summary, understanding protein sequence similarities and structural features is essential for deciphering protein function and biological processes.

Protein structure relationship with function

Protein structure is closely linked to function, as the specific arrangement of amino acids determines the protein's three-dimensional shape, which in turn governs its biochemical activity.

DNA profiling and it's advantage

It is a form of forensic identification that is used primarily to identify people.

Advantages

Even though all humans share 99.9% of their DNA sequences, the remaining 0.01% of sequences is unique enough to differentiate people. Because DNA is in every cell of a person’s body is present and same, any part of a human’s body including dead skin skills, hair, saliva, and more contain DNA sequences.

Ab initio method advantage and disadvantage

Advantages

- Ab Initio methods can fold any target sequence using only physical atomic properties
- Predictions are mostly accurate and correctly describe the natural folding process

Disadvantages

- Ab initio methods are the very difficult to design (energy function)
- These methods are slow due to the huge possibilities

Four type of gene expression

GEO consists of four types of records:

Sample (GSM): Stores sample information, including preparation, treatments, and experimental design.

Platform (GPL): Contains details about platforms used, indicating whether it's microarray or RNAseq data.

Series (GSE): Groups similar treatment files together, forming a series of related samples.

Datasets (GDS): Stores actual data collections assembled by GEO.

What is the role of biosafety cybernatics?

The role of biosafety cybernetics is to ensure the safe and secure management of biological information and processes within cybernetic systems. It involves risk assessment, system monitoring, compliance management, incident response, and education/training to safeguard biological assets and public health.

By looking at the PDB structure, alanine found in great number, which element will form from them?

By looking at the structures in PDB, we know that Alanine mostly found in Alpha Helices. So if we have several Alanines in the sequence, then we can anticipate that a helix may be formed by them

How to Fine tune and adjust the BLAST alignments?

To fine-tune BLAST alignments:

- Adjust E-value threshold for match significance.
- Modify word size for sensitivity.
- Change gap penalties for alignment stringency.
- Select scoring matrix for amino acid substitutions.
- Enable/disable filtering for sequence complexity.
- Customize alignment display options.
- Choose specific databases for searching.

How protein is converted into small fragments?

Proteins can be converted into smaller fragments through enzymatic digestion, chemical cleavage, ultrasonication, heat treatment, or mechanical shearing. These methods break peptide bonds or disrupt protein structures, resulting in smaller peptide fragments suitable for various applications.

Define insertions, deletions, and substitutions?

The columns of the alignment containing one space are called indels, with the columns containing a space in the top row called insertions and the columns with a space in the bottom row deletions.

Insertions: Adding nucleotides to a DNA or RNA sequence, shifting the reading frame and potentially altering protein structure.

Deletions: Removing nucleotides from a DNA or RNA sequence, also shifting the reading frame and potentially causing non-functional proteins.

Substitutions: Replacing one nucleotide with another, which may or may not change the corresponding amino acid during translation, affecting protein function.

Zoology Legends

Muhammad Zaman

03044756496

Remember me in you Precious Prayer

Jazak Allah

Team Zoology LEGENDS