Bio Unit-I

Assignment - I

1) what is Bioinformatics ? Explain the history of Bioinformatics.

⇒ Bioinformatics is new decipline which addresses the need, manages the need and a data heading past was messively generated by genomic reaserch. This decipline represent the conversion of genetics biotechnology, It and incompasses analysis as implementation of data modeling of biological phenomenon and their development of Algorithms and statistics.

History →

BI history started with an  'Gregor Mendel' is known as the father of genetics. he cross fertilizes different colours of same species of flowers. he create record of the colours of flowers that produce inheritence of trades could be more easily extend it was controlls by factors passed by generation to generation.

In 1972 all work metofirst DNA molecules using ligas in same year Stanley Cohen, annchan and narfwort Groval producing first predominont & DNA organism.

In 1988, the human genome organisimation was founded this is an international organization of scientist, involved in human genome project.

In 1993 human reasearch lab produce ~~physical~~ a human map

3) why Bioinformatics needed ? Benefits of bioinformatics.

-> Bioinformatics is crucial because it combines biology, computer science, and mathematics to analyze and interpret biological data. It helps researchers make sense of vast amounts of genetic, molecular, and clinical information. This is vital for understanding diseases, drug discovery, genetic research, and personalized medicine.

Benefits :

i) Data Analysis : It enables the efficient analysis of large-scale biological data, such as DNA sequences, protein structures.

ii) Genomic Research : Bioinformatics plays a pivotal role in deciphering genomes, identifying genes and understanding genetic variations.

iii) Drug Discovery : It aids in identifying potential drug targets, predicting drug interactions.

iv) Disease Understanding : By analyzing biological data, bioinformatics help unravel the molecular mechanisms underlying diseases, leading to better diagnosis and treatment strategies.

v) Evolutionary studies : It assists in tracing the evolutionary history of species by comparing genetic sequences, shedding light on their relationships.



4) Explain the term molecular Biology.

-> - Molecular biology is a specilized branch of biology and biochemistry which is specifically concerned study of various biological activities at molecular level.

- The history of molecular biology :
The term of molecular biology was coined by an american scientist warren in year 1938.

- The discovery of molecular biology began in ~~year 1938~~ early 1940's and its fundamental development took place in year 1953.

- Living things are made of chemicals just as non-living things are, so a molecular biologist studies how molecules interact with one another in living organisms to perform the functions of life.

- Molecular biologists conduct experiments to investigate the structure, function, processing, regulation and evolution of biological molecules and their interactions with one another - providing micro-level insights into how life works.

4/6/01

Imp.

6) what is Biological Database? explain the characteristics in detail.

⇒

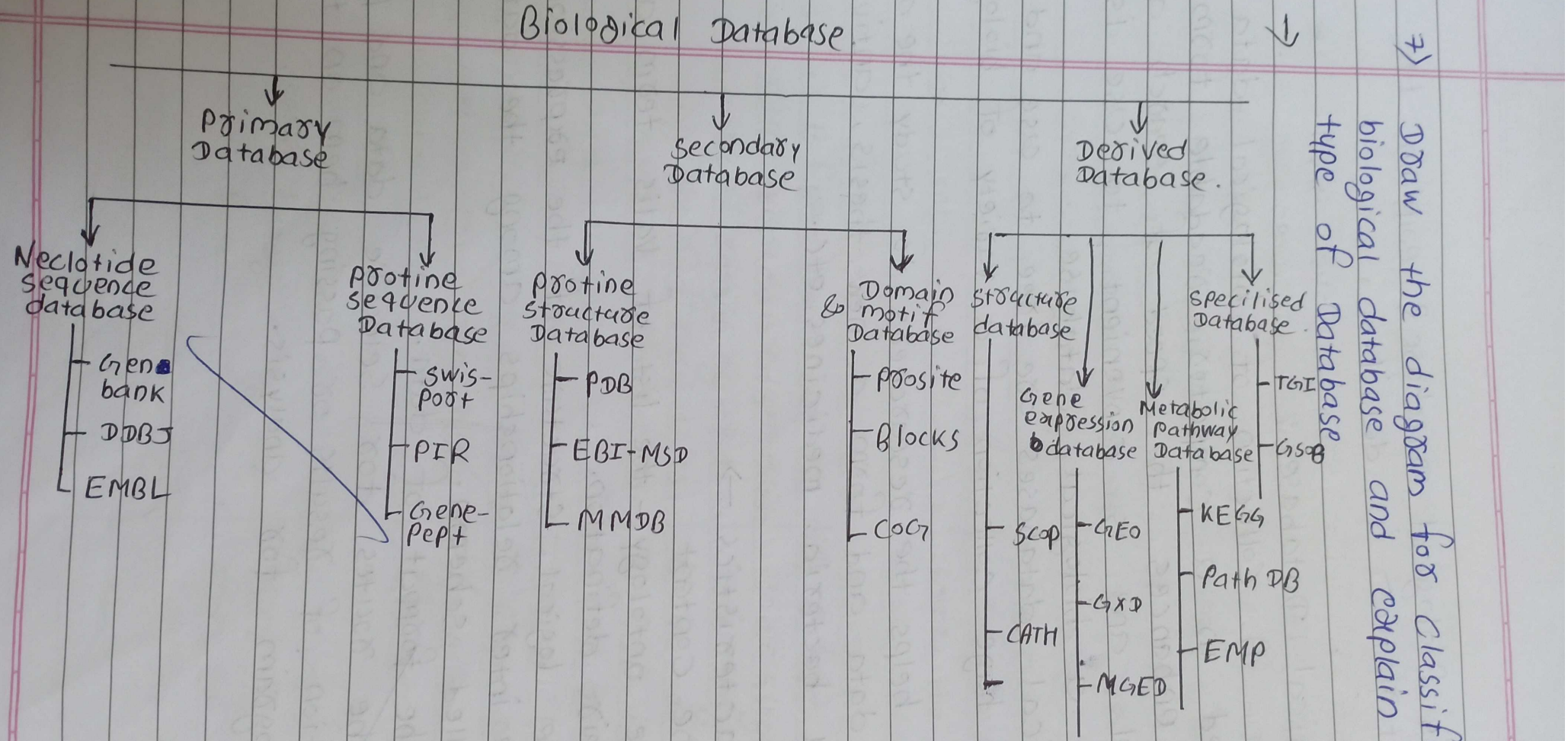
Biological Database :

- A collection of biological data arranged in a computer readable form that enhances the speed of search and retrival and is convenient to use is called a biological database.
- Biological databases are free to use and contain a huge collection of a variety of biological data.
- It helps the researchers to study the available data and form a new thesis, antivirus, helpful bacteria, medicines, etc.

Characteristics →

- The content
- The ontology - the list of valis term and their defination.
- The logical structure of the expressions of the inter relationships among the data called schema.
- The format of data.
- The routes for selective data and presentation of results or passing hem on to a program for analysis.

7) Draw the diagram for classification of biological database and explain different type of Database



10) Difference between primary and secondary database

| Primary Database | Secondary Database |
|---|---|
| i) It is also known as archival database. | i) It is also known as curated database or derived database. |
| ii) Database consisting of data derived experimentally such as nucleotide sequences and 3 dimensional structures are known as primary database. | ii) Database consisting of data derived from the analysis of primary data such as nucleotide sequences, protein structures. |
| iii) It contains the original experimental results are directly submitted into database by researches | iii) It contains results of analysis of primary database and significant data in form of conserved sequences, signature sequences |
| iv) primary database has high levels of redundancy or duplication of data. | iv) This database has low levels of redundancy of data as it is curated. |
| v) e.g. Gen bank, DBBJ, PDB | v) e.g. swiss-port. |

4/6/01

DATE PAGE NO.

Imp correct the ans.

11) What is mean by sequence database and structure database?

→ i) Sequence Database:

A sequence database contains a collection of biological sequences, such as DNA, RNA, and protein sequences, from various organisms. These sequences are typically stored in a standardized format and can be accessed based on their identifiers, accessed based on their identifiers, ~~accessed based on their identifiers~~ accession numbers, or other search criteria. Sequence databases are used for performing sequence similarity searches, where a query sequence is compared against sequences in the database to find similar or homologous sequences.

e.g. GenBank, EMBL, DDBJ.

12) Define Bioinformatics tool and explain the sequence analysis tool in Detail.

→

Bioinformatics tool are userfriendly software programs that allow researches to analyse biological data. along with construction duration of biological database, bioinformatic also consists of development of computational tool for sequence, structure and function analysis.

Sequence Analysis tool :

This tools are used for analysing nuclotides or protein sequence then also use for identifing homologous sequences and understanding the evolutionary ~~zeta~~ relationship between different organism. They include tools used for sequence aligment, sequence database searching, Motin discovering ~~the~~ and geneme assembly. some popular sequence analysis tool are BLAST, T-Coffee, MEGA, MEME.

4/6/01

14) What is mean by function analysis tool. explain GEO, Cobra tool box and Pathway tool.

→ These tools are essential for understanding the function and relationship between different protine and gense and identifying key pathways involved in diseases. It include tools that are used for profiling gene expression predicting protein-protein interaction, predicting protein sub-cellular localization and Modeling Metabolic pathways.

i) GEO :

GEO stands for Gene Expression omnibus. It's a public database maintained by the National Center for Biotechnology Information (NCBI). GEO is used for storing and sharing high-throughput gene expression and molecular abundance data. Researchers use this resource to deposit and access data from various experimental techniques like microarrays and next-generation sequencing.

ii) COBRA Toolbox :

The COBRA Toolbox (constraint-Based Reconstruction and Analysis Toolbox) is a Computational Toolkit used for function analysis of biological systems, particularly metabolic networks. It employs constraint-based modeling techniques to study how metabolic pathways function within an organism.

The COBRA Toolbox enables researchers to;

- Reconstruct Metabolic Models
- Constraint-Based Analysis
- Phenotype Prediction
- Metabolic Engineering
- Biomarker Discovery
- Drug Target Identification.

iii) Pathway tool :

Pathway tool are software tools used in bioinformatics for pathway analysis. They help researchers understand the complex network of biochemical reactions and interactions that occur within a biological system.

In functional analysis, pathway tools enable the following :

- Pathway Visualization :
- Data Integration :
- Enrichment Analysis :
- Prediction of Functional outcomes
- Comparative Analysis.
- Hypothesis Generation.

15) list out the application of Bioinformatics and explain any four.

→ Applications:

i) Genome Sequencing and Analysis:

Bioinformatics is crucial in analysing large-scale genome sequences, identifying genes, regulatory elements, and functional regions. It helps in understanding genetic variations, mutations, and their impact on diseases.

ii) Protein Structure Prediction:

Bioinformatics tools predict the 3D structure of proteins based on their amino acid sequences. this information aids in understanding protein functions, interactions, and designing drugs that target specific proteins.

iii) Phylogenetics:

Bioinformatics assists in reconstructing evolutionary relationships among species by analyzing molecular data such as DNA and protein sequences. It helps to understand the evolutionary history and relatedness between different organisms.

Bio

Unit-2. Genomics & proteomics

Date:
P. No:

Q.1. DNA.

→ - DNA is a group of molecules that is responsible for carrying and transmitting the inheriditey material or the genetic instruction from parents to offsprings.
- DNA also plays crucial roles in the production of proteins.
There are 3 types of DNA.

i) A-DNA :
It is right handed double helic similar to B-DNA form. Dehydrated DNA takes an A form protects the DNA in extreme condition such as complete dryness.

ii) B-DNA :
It is the most common DNA comformation and right handed helix. The majority of DNA has B-type conformation under normal phylogical condition.

iii) Z-DNA :
It is left handed DNA where Double helic bits to the left in a zig zag pattern.
~~This found in a zig zag pattern.~~
This found the head of start side of gene hence they need to some role in gene.

| | DNA | RNA |
|---|---|---|
| Structure | 1) DNA is double stranded molecule consisting of long chain of nucleotides. | 1) RNA is single stranded molecule consisting of short chain of nucleotides |
| Function | 2) It transmits genetic information to make other cells and new organisms. | 3) It transfers the genetic code from nucleus to ribosomes to make proteins. |
| | 3) DNA is self replicating | 3) RNA Synthesized from DNA. |
| | 4) located in nucleus of cell and in mitochondria | 4) Located in cytoplasm, nucleus and in ribosome |
| | 5) Basic Pairing : AT, GC | 5) Basic Pairing : AU, GC |
| molecules weight | 6) 2 to 6 million | 6) 25,000 to 2 million |
| | 7) DNA is more stable molecule than RNA. DNA is stable under alkaline conditions | 7) Much more reactive than DNA and is not stable in alkaline conditions. |
| | 8) DNA is vulnerable to damage by ultraviolet light | 8) RNA is much more resistant to damage from UV light than DNA. |

Date:
P. No:

Q.2 Explain regulation of gene expression.

→ Gene expression is a process by which the instruction present in DNA are converted into a functional product such as protein. During gene expression genetic code from DNA are conveited into protein with help of transcription & translation.

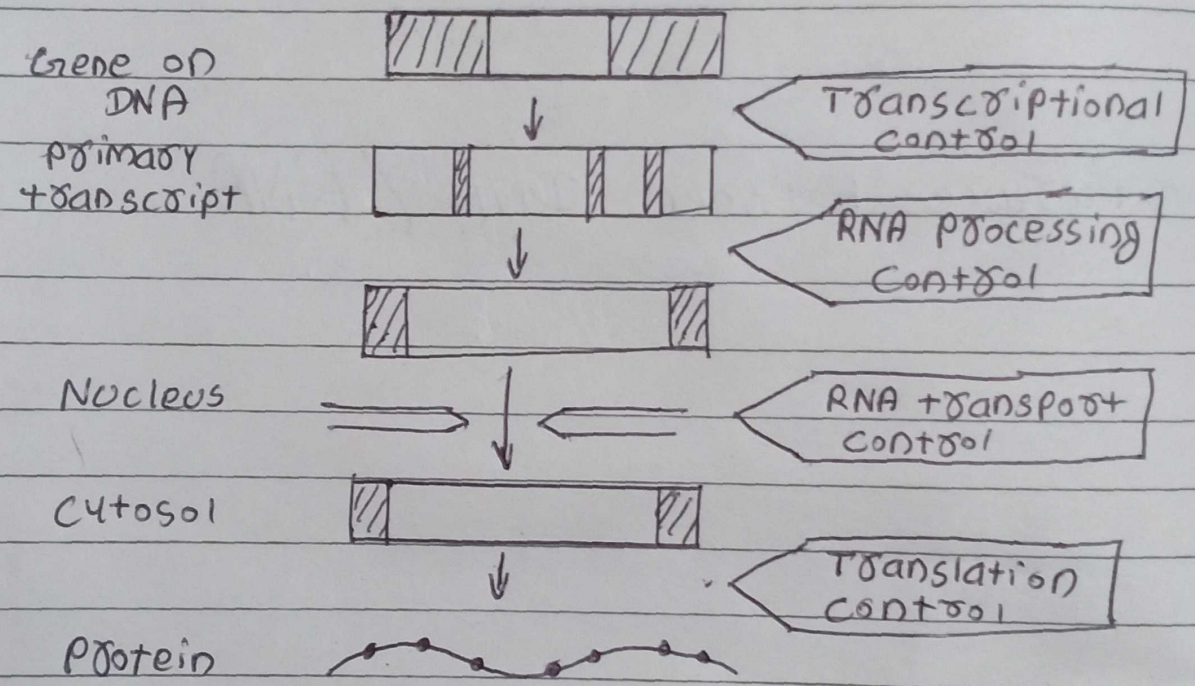



Fig. How ~~DN~~ DNA is synthesis to protein

First Transcription factors recognize specific sequences on DNA, acting like keys unlocking access for RNA Polymerase. RNA polymerase binds to promtor, starting the transcription process. RNA polymerase reads the DNA code, building a complementary RNA molecule (mRNA). The mRNA undergoes modifications before leaving the nucleus. Ribosomes in the cytoplasm read the mRNA code and assemble the corresponding protein.

3) what is DNA Sequencing ?

→ DNA sequencing is the method that determines the order of four nuclotide bases that make up DNA molecules and convey inmortant genetic information.

The human genom contain 3 billion best pairs that provides the instructions for creating and maintaining human being.

Base pair structure of the maks DNA best united for the storage of wast amount of genetic information.

Data sequencing methods:

i) Sanger sequencing method :-

This method was discover by english biochemist fedrict sanger in 1970.

N= A, T, G, C to prevent addition of another nuclotides.

ii) Next Generation sequencing method :-

It is also known as macevly parallel sequence this Technologies are largely seperated sanger sequences with advantages such as cost efficient rapidness. The NGs can determine order of millions of fragments symentionously.

NGs is short read sequencing that require construction of small fragment sequencing follow by deep sequencing, raw data preproce-ssing DNA sequence alignment, assembly, allotation and down stream analysis.

iii) Emerging third Generation Sequencing:
It is known as long read sequencing including PacBio SMART sequencing Oxford nanopore sequencing can examine billions of templets of DNA and RNA and symentionously detect variable methandation of without bias. The long bbeed method detect more variations some of which cannot be observed by short read sequencing.

4) What is mean by Genome mapping.
→ Gene mapping refers to the process of locating genes within a genome. Gene mapping refers to the techniques used to identify genes location & distance between genes. The distance between various site inside a gene can also be described through gene mapping.
Types of gene mapping:
Genetic linkage maps & physical maps are the two main categories of "Maps" used in genome mapping. Both maps consists of genetic markers & gene.
- Genetic Linkage mapping — It shows the location of each gene on a chromosome & their relative distances from one another.
- physical mapping — physical map always provide the actual DNA base pair distance between land marks. It is one genome mapping approach that has a high accuracy in determining the sequances of DNA base pair.



5) List out and explain any four implications of genomics for medical science.
→ i) Genomic medicine for personalized treatment:
Genomic allows for the tailoring of medical treatment based on an individual's unique genetic makeup. e.g. knowledge of a patient's genomic profile can guide the selection of ~~targe~~ targetcal therapies for cancer, improving treatment effectiveness.
ii) Diease Risk assessment & early detection:
- Genomic analysis enables the identification of genetic markers associated with various diseases including genetic predis possition.
- For instance, identifying specific genetic variations can help predict the risk of developing conditions like Alzheimer's or Cardiovascaular diseases, enabling proactive healthcare planning.
iii) Genomic Research for disease understanding and discovery:
— Genomic aids in researching the genetic basic of diseases, uncovering cacuative genetic mutuations or alterations.
— This understanding is crucial for developing targeted th
iv) Rare disease, Diagnosis and Research:

6) Explain the term proteomics.

→ Proteomics focuses on the study of proteins, including their identification, quantification, and the characterization and structures. This field helps in understanding proteins functions, interactions and structures.

Proteomics is the large-scale study of proteins. proteins are complex molecules that perform essential functions in living organisms, including structure, metabolism, transport, communication and defense.

Proteomics is used to study the structure, function and interactions of proteins. proteomics research use a variety of techniques to identify, characterize and quantify proteins. These technique includes.

i) Mass spectrometry.
ii) Gel electrophoresis
iii) Immunoassays.

Proteomics is a rapidly growing field with a wide range of applications. Proteomics research is being used to develop new diagonestic tests and treatments for diseases to understand basic of disease and to develop new drugs and therapies.

Applications :-
i) Cancer research
ii) Infectious disease research
iii) Neuroscience research
iv) Personalized medicine.



7) What are the applications of proteomics to the medicins.

→ Here are some important applications of the proteomics in the field of medicine :-

i) Disease, Biomark discovery.
ii) Personalized medicine
iii) Drug Targed Identification.
iv) Pharmacoproteomics.
v) Protein – protein Interaction Analysis.
vi) Analysis of Past-Translational Modifications
vii) Vaccine Development.
viii) Drug Resistance studies.
ix) Taxicology and adverse event Analysis.

8) Differentiate between genomics & proteomics.

→

| | Genomics | proteomics |
|---|---|---|
| scope and focus | study of an organism's entire genome, including genes & non-coding regions | study of an organisms entire proteone, encompassing all expressed proteins. |
| unit of study | DNA sequences | proteins. |
| Information content | Genetic code, nuclotide-sequences (A,G,T,C) | functional elements, protein presence, abundance, interactions, modifications. |
| Purpose | Understanding genetic inheritance, gene structure genotype - phenotype relationships | Understanding protein functions, cellular process-es, disease biomarkers. |
| Methodology | DNA sequencing, genome mapping, PCR | Mass spectrometry, 2D gel electrophasis, protein-protein interaction studies. |

9) Explain the term protein modeling.

→ Protein modeling, also known as protein structure modeling or protein structure prediction, is a computational approach used in bioinformatics and structural biology to predict the three dimentional structure of a protein based on its amino acid sequence.

The primary aim or goal of protein modeling is to generate a 3D model of a protein's atomic co-ordinates, which can provide insight into its structure, function, and interaction other molecules.

10) Differentiate between :-

→

| Pairwise Sequence Alignment | Multiple sequence Alignment. |
|---|---|
| i) An alignment procedure comparing two biological sequence either protein DNA or RNA | i) An alignment comparing three or more biological sequences either proteins, DNA or RNA |
| ii) Generally categorized as a global or local alignment method | ii) Generally categorized as global multiple sequence method |
| iii) Comparatively simple algorithm is used. | iii) Complex sophisticated algorithm is used |
| iv) Applications :- primary to find out conserved regions bet^n the two sequences. | iv) To detect regions of variability or consecutive in family of protein. |
| v) e.g: BLAST, LALIGN, EMBOSS Needle | e.g:- MUSCLE, T-coffee. MARRT. |



11) Explain dynamic programming method of pairwise Sequence alignment with example.

→ - Dynamic programming is used to find the optimal alignment between two proteins or nucleic acid sequences by comparing all possible pairs of characters in the sequences.

- Dynamic programming can be used to produce both global & local alignment. The global pairwise alignment algorithm using dynamic programming is based on Needleman wansh algorithm, while the dynamic programming local alignment is based on the smith waterman algorithm.

The method works in following three steps :-

1) Initialization :

The first step in global alignment dynamic programming approach is to create a matrix with M+1 colomns & N+1 rows where M & N corrosponds to the size of the sequences to be aligned.

2) Matrix fillings :

Matrix is filled with, if the characters are match then the value be 1, if characters in sequences are mismatch the value be 0 and for gap the value be -1 and for no gap value be zero.

In this way complete matrix is filled with nombers.

3) 3rd point of after question no. 16.

12) Explain BLAST Method with example.

→ BLAST stands for Basic local Alignment Search tool. It is a local alignment algorithm based tool used for aligning multiple sequences and finding similarities or dissimilarities among various species. BLAST is a heuristic method which means that is the dynamic programming algorithm that is fasted, efficient but relatively less sensitive.

BLAST first breaks the query sequence into smaller fragments, then it compares each fragment to the sequences in the database, BLAST extends the alignment to see if there is a longer match.

Use :- The scientists can use BLAST to search for similar genes in other organisms, the scientists can hypothesize that these genes are involved in the same disease pathway.

13) Explain K-tuple method with example.

→ The K-tuple method is simple but effective way to identify sequences. It works by dividing the sequences into K-tuples and identical, then the corresponding sequences are considered to be similar.

EX:-

$S_1$ - "ACGTACGT"

$S_2$ - "ACGACGT"

for, K=3



$S_1$ = ACG, CGT, GTA, TAC, ACG, CGI

$S_2$ = ACG, CGA, GAC, ACG, CGI

We can compare the K-tuple for $S_1$ & $S_2$ we find that the K-tuple "ACG" is present in both $S_1$ & $S_2$.

This indicates sequence $S_1$ & $S_2$ are similar

14) Differentiate between BLAST & K-tuple method.

→

| BLAST | K-tuple |
|---|---|
| i) Find similar sequences in a database | i) A unit used in various sequence analysis task. |
| ii) Can be used for DNA, RNA and protein sequences. | ii) A general concept application to various sequence types. |
| iii) Searches for local and global sequence similarities. | iii) Represents a fixed-length sequence segment. |
| iv) Returns a list of similar sequences from a database along with statical measures. | iv) Results in lists of matching K-tuples or sequence segment in analysed sequence. |

15) Explain the term dynamic programming method of multiple sequence alignment.

→ Dynamic programming is a fundamental technique used in bioinformatics, including MSA in the context of MSA dynamic programming is often employed to efficiently compute the optimal alignment of three or more biological sequences.

Here is how dp works in MSA:

i) Problem formalation :- Given multiple sequences, goal is to align them in a way that maximizes the similarities.

ii) Scoring system :- Define a scoring system that assign score to different possibilities such as match, mismatch, gaps, etc.

iii) Dynamic programming Table :- creat a maltidimention of table to store intermediate results & ultimately compute the optimal alignment score.

iv) Filling the table :- start filling table using a dp often based on the Needlman-Wansh or smith waterman algorithm.

v) Backtracking :- After table filled, backtrack through table from last cell to reconstruct optimal alignments.

vi) optimal Alignments :- They are set of sequences aligned in a way that maximizes the overall alignment score based on defined scoring system.



16) Explain the term clustal-w

→ clustal w is a tool for aligning multiple protein or nucleotide sequences. This alignment is achieved via three steps:-

i) Pairwise alignment.
ii) Guid tree generation.
iii) Progressive alignment.

- clustal w-MPI is a distributed & parallelism implication of cluster w.
- Clustal w uses progressive alignment methods align most similar sequences, first and work their way down to the last similar sequences until a global alignment is created.
- clustal w is a matrix based algorithm tools like T coffee and dialign are consistet base cluster w is fairly efficient algorithm competes - against other softwar

Shwetha
26/10/23

UNIT-03

Date:
P. No:

Question

① Define phylogenetic analysis

→ A phylogenetic analysis is the study of the evolutionary development of a species or a group of organism or a particular characteristic of organism.

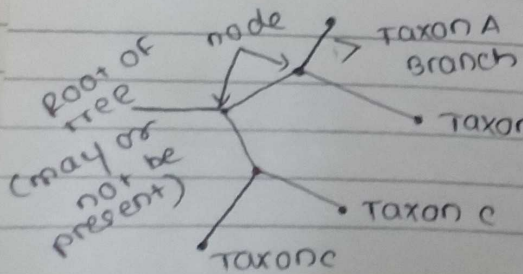
Question

② List out? and explain fundamental concept of phylogenetic tree

→ Phylogenetic tree, also known as an evolutionary tree, is a diagram that depicts the evolutionary relationship between different species or genes. It's like a family tree for the entire living world, tracing back lineages through shared ancestry.

properties

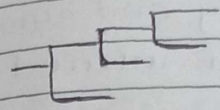
- Root, representing the oldest common ancestor of all the groups included
- Branches, representing a lineage that diverged at a specific point in time
- leaves, represent the present day species or genes being studied.



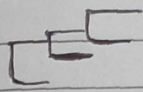

Date:
P. No:

Types.

① Rooted Tree → Have a designated root representing the oldest common ancestor. most commonly used



② unrooted Tree → do not specify a root, focusing only on the relative relationship between groups



Question

③ process of phylogenetic analysis.

→

(i) choosing the Right markers

- Homologous sequence → you need sequences from a common ancestor, like genes or protein segment
- informative sequence → The choosen markers should have enough variation to distinguish between lineages.

(ii) Gathering data → • sequencing DNA or proteins
• collecting morphological data

(iii) Alignment → multiple sequence alignment aligns the chosen sequences accounting for insertions, deletions and mutations.

Date:
P. No:

(iv) Tree building → Algorithms analyze the aligned sequences to infer evolutionary relationships.

(v) Evaluating the Tree → Bootstrapping and other statistical test assess the robustness of inferred tree.

(vi) Interpreting the Tree → branches may reflect.

(vii) visualization and analysis

question

④ Explain tree building methods

→

(i) Distance method
(ii) character method < maximum parsimony / maximum likelihood
(iii) bayesian inference.

(i) Distance method → calculate distance between taxa and build a tree minimizing total distance distance are then used to construct a tree that minimizes the total distance betwn all pairs.

(ii) character method → This method consider the presence or absence of specific characters in different taxa. goal is to find best tree

Date:
P. No:

(iii) Bayesian Interface → considers uncertainty in data and models to find the most probable tree.

question

⑤ Define paralogs & orthologs

→

(i) paralogs → paralogs are the gene that diverge within the same species. They are multiple genes (tree leafs) per species.

(ii) orthologs → are the ~~species~~ genes the diverge with species.

question

⑥ Define microarray

→ microarray is essentially a miniaturized laboratory on a chip, used to investigate thousands of biological interactions simultaneously.

Applications
- gene expression studies
- cancer research
- microbial identification

Date:
P. No:

question

(7) concept of tree evaluation

→

(i) phylogentic tree Evaluation

(ii) Decision tree Evaluation → access the predective performance of decision tree used for tasks like classifying genes. (cross validation, accuracy etc)

(iii) Expression tree Evaluation → Evaluate tree representing gene expression patterns or regulatory networks

(iv) Hierarchical clustering tree evaluation → It assess the quality of tree generated by hierarchical clustering algorithms used to group similar biological entities. (genes, proteins, samples)

(v) other tree based models < Random forest, Gradient Boosted tree.

question

(8) Microarray Techniques

→ microarray are like tiny lab-on-chips with thousands of mini detectives. They scan samples for specific target, like gene or proteins using fluorescent labels. By analyzing the glowing patterns, scientists gain insights into disease, how genes work, and evolution.

- detectives (probes) wait on the chip
- sample with labeled clues (RNA/DNA) is added
- detectives bind their matching clues, glowing brighty
- The chip is scanned, revealing the detective clue matches.

question

(9) applications of microarray

→

(i) Gene Expression profiling
(ii) mutation & SNP detection
(iii) comparative Genomics
(iv) drug development
(v) clinical diagonistics
(vi) Agriculture & plant Research
(vii) forensic science.