Bachelor of Science
(B.Sc.- CBZ)

COMPUTATIONAL BIOLOGY
(DBSZDS301T24)

Self-Learning Material
(SEM-III)

Jaipur National University
Centre for Distance and Online Education

Established by Government of Rajasthan
Approved by UGC under Sec 2(f) of UGC ACT 1956
&
NAAC A+ Accredited

TABLE OF CONTENT

EXPERT COMMITTEE

Prof. D.S. Bhatia
Former Director, Life and basic Sciences
JNU, Jaipur

Prof. Sunil Gupta
Dept. of Computer and System Sciences
JNU, Jaipur

COURSE COORDINATOR

Mr. Sumit Govil
Dept. of Life Sciences,
Jaipur National University, Jaipur

UNIT PREPARATION

Unit Writers	Assisting & Proof Reading	Unit Editor
Mr. Shish Dubey (Unit 1-2) Dept. of Computer & System Sciences, JNU, Jaipur Mr. Sumit Govil (Unit 3-4) Dept. of Life Sciences, JNU, Jaipur	Dr. Deepak Shekhawat Dept. of Computer & System Sciences, JNU, Jaipur	Mr. Ramlal Yadav Dept. of Computer & System Sciences, JNU, Jaipur

Secretarial Assistance:
Mr. Suresh Sharma

COURSE INTRODUCTION

Computational biology is an interdisciplinary study that uses mathematical, statistical, and computer science methods to analyze and interpret biological data. It makes use of computational techniques to tackle intricate biological issues, offering insights that are frequently inaccessible through conventional experimental procedures.
Key areas of the book are Genomics,Proteomics, Systems Biology, Structural Biology, Evolutionary Biology, Bioinformatics, Sequence alignment.

Through the bridging of the gap between biological research and computer approaches, computational biology enables scientists to model intricate biological systems, analyze enormous volumes of data, and uncover basic principles underlying life. Computational biology has emerged as a critical topic for expanding our knowledge of biology and medicine due to the increasing amount and complexity of biological data.

The course Computational Biology is of 3 Credits. This course is divided into 04 units and each Unit is divided into sub topics.

Course Outcomes: After completion of the course, the students will be able to

1. List out different biological database and information they provide.
2. Explain the different OMICS fields and their utility.
3. Apply the importance of various concepts and tools for biological data.
4. Analyze biological data using various software's and tools for biological data analysis.
5. Assess the biological data based on various concepts and tools.
6. Assemble the result and identify the relationship between the biological data.

Acknowledgements:
The content we have utilized is solely educational in nature. The copyright proprietors of the materials reproduced in this book have been tracked down as much as possible. The editors apologize for any violation that may have happened, and they will be happy to rectify any such material in later versions of this book.

Unit -1

Introduction to Bioinformatics

Objective:

- Gain Knowledge of various bioinformatics tools and use of bioinformatics in various field of science.
- To understand use of computational biology in the numerous fields of medicine
- Make Students aware to the application of various computational tools in bioinformatics tools and subjects.
- To integrate the fields of computational technology and scientific research to enhance research capabilities.

Definition

One field that combines biology and technology is called bioinformatics.

Both the biology sciences and engineering must be well understood.

This field is equipped with biological data and computational techniques comprises of software such as BLAST, CLUSTAL, ETC

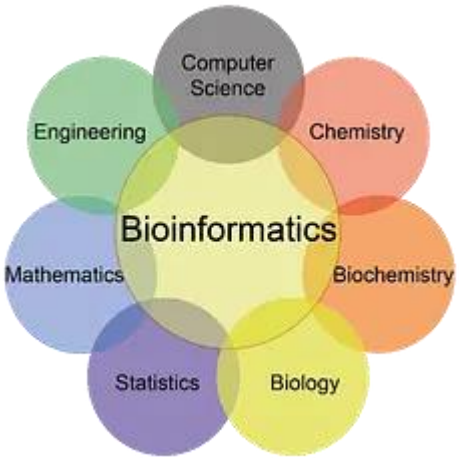

Fig 1.1 Visual depiction of Bioinformatics

Importance

Large-scale data management and suitable analysis are at the core of the vast and rapidly expanding field of bioinformatics.

The fields of 3D image processing, 3D modeling of live cells, image analysis, medication development, and many other topics are the focus of bioinformatics.

The primary goal is to reduce complexity and enhance the usability of natural processes in their entirety.

In gene therapy, bioinformatics plays a significant role.

Evolutionary notions have applied in this branch.

microbiological computation and analysis.

Understanding the structure of proteins

Storage and retrieval of biotechnological data.

Used in the discovery of novel drugs.

In agriculture to recognize crop patterns, pest control, and crop management.

Goal

The goals of objectives.

1. Bioinformatics is an organized collection of data so that scientists may access previously published data and add new items as they are developed
2. Creating resources and tools that promote data analysis is the second goal.

The creation of resources such as PSI-BLAST and FASTA requires a solid grasp of biology in addition to computational theory proficiency.

3.Utilizing these tools to evaluate the data and interpret the findings in a way that is intuitive in a biological sense is the third goal. Biological research has historically focused on researching particular systems in great detail and often compared them with a few similar ones. We are now able to conduct global assessments of all the data that is accessible in bioinformatics with the goal of emphasizing new features while discovering universal principles that apply to a variety of systems.

Scope

The field of bioinformatics encompasses a broad spectrum of biotechnological sub-disciplines that are distinguished by a profound understanding of computer science and information technology, as well as scientific ethics grounded on biological sciences.

disciplines within bioinformatics include, for instance:

Bioinformatics will grow in scope and utility. Some of the examples of many fields of bioinformatics include:

Computational biology: The application of data-driven approaches to bioinformatics complications.

Genetics: It refers to an analysis of genetic diversity and heredity in relation to inherited traits.

Genomics: It is the domain of biomolecular biology encompassing genome mapping, structure, function, and evolution.

Proteomics: investigating proteomes and their characteristics.

Metagenomics: The scientific study of genetics using environmental, biological, and sampling information.

Transcriptomics: It is the study of DNA and RNA transcriptase in its entirety.

Phylogenetics: The study of links between or within groups of organisms and their evolutionary history

Metabolomics: The investigation of metabolites and metabolic biochemistry in living organisms.

Systems biology: Mathematical designing and analysis and visualization of large sets of biodata.

Structural analysis: Modeling that determines the effects of physical loads on physical structures.

Molecular modeling: The designing and defining of molecular structures by way of computational chemistry.

Pathway analysis: A software description that defines related proteins in the metabolism of the body.

1. Genomics

- The study of an organism's complete genome, which includes all of its genes and the interactions between them and their surroundings, is known as genomics.
- The genome, which is integrated into almost all of the organism's cells, houses the whole collection of DNA for that creature.
- Structural genomics and functional genomics are the two main disciplines of genomic research they are as follows:
  - Structural Genomics:The area of genomics that studies the architecture of genome sequences is called structural genomics. Genome mapping, gene sequencing, gene feature annotation, and genome structure comparison are steps in the process of understanding the genome structure.
  - Functional Genomics: The study of gene expression and how genes operate within a genome is the focus of functional genomics. It entails applying high-throughput techniques to explore gene activities throughout the whole genome.

Few of the diverse field are as follows:

Epigenomics

- It the study of a group of molecules that bind to DNA and affect its function. The distinctions between the different cell types in the body are mostly determined by the epigenome.
- Histone modifications and DNA methylation are examples of epigenomic alterations.

Metagenomics:

- The study of genomes encompasses not just individual species but also entire biological ecosystems. Usually, it is used in relation to microbes.
- The diversity and function of microbial communities in a variety of settings, including the human gut, soil, and ocean, may be studied using metagenomics.

Pharmacogenomics:

- The branch of genomics known as pharmacogenomics studies and tailors medication selection and dosage for medical conditions by using a person's genetic information.
- It may be applied to create a more individualized prescription process and forecast medication response or toxicity.

Comparative Genomics:

- It is comparative study of the genomes of other species can provide information about their evolutionary ties, functional components, and genetic differences.
- It makes use of a number of instruments that aid in recognizing and comprehending the parallels and discrepancies among the genomes of different animals.

1.1. Method in Structural Genomics:

1.1.1. Genome Mapping

- "Genome mapping Is identifying the positions of genes, mutations, or characteristics on a chromosome
- Genome maps come in several forms, such as physical, cytologic, and genetic linkage maps.
- Genetic maps, sometimes called genetic linkage maps, are low-resolution maps that show the relative locations of genetic markers on a chromosome based on inheritance patterns.

- Regardless of inheritance patterns, physical maps are high-resolution maps that show the positions of recognized features on a genomic DNA.
- The banding patterns that are discernible on dyed chromosomes and may be directly viewed under a microscope are described by cytologic maps, commonly referred to as chromosome maps. These easily recognizable bands of bright and dark are markers on the chromosomes.

1.1.2. Sequencing of genomes

- Genome sequencing, yields the most comprehensive genome map. Sanger sequencing is a popular technique for sequencing DNA.
- It labels different length DNA chains with fluorescently tagged dideoxy nucleotides. After electrophoresis to separate the resultant fragments, the sequence is determined by examining the banding pattern on the gel. A chromatogram's peaks are assigned bases using a process known as base calling.
- There are two primary techniques for sequencing whole genomes:
  - Shotgun sequencing (DNA clones are randomly sequenced which are aligned in a sequence with the help of computational method leading to whole genome).
  - Hierarchical sequencing (Longer genomic DNA pieces are cloned onto a highly capacitate bacterial vector. The positions and sequence of the cloned fragments on the chromosome may be ascertained with the use of the physical map).

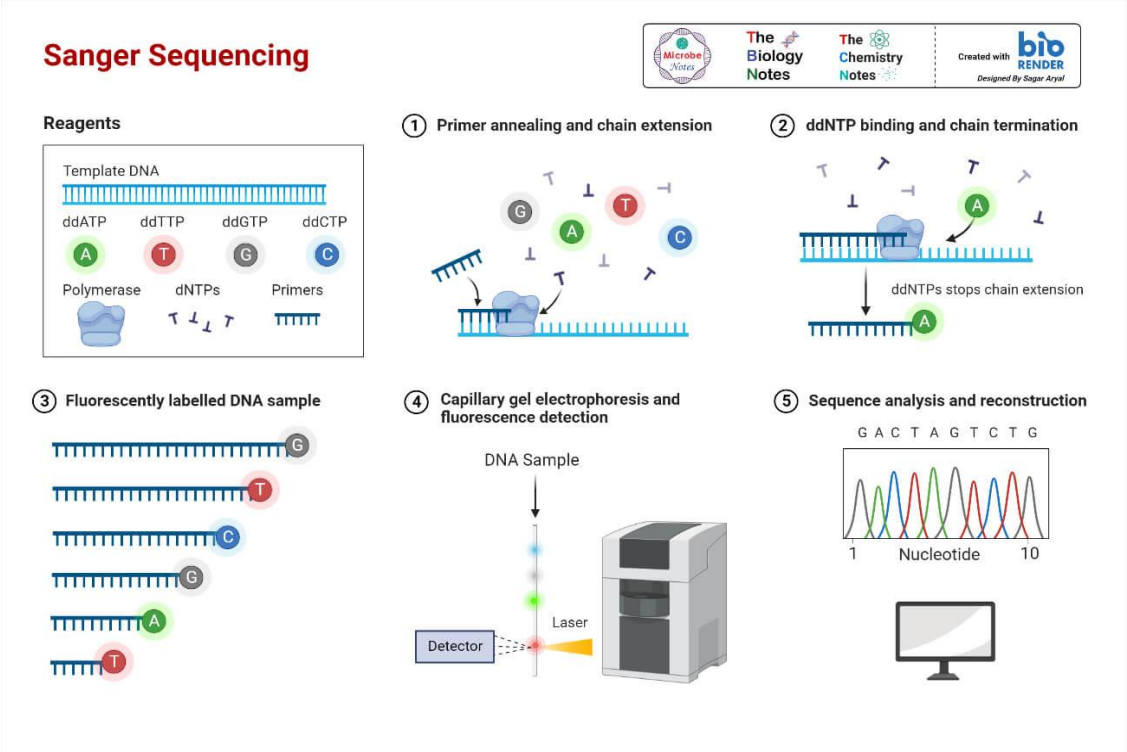

Figure 1.2 Sanger sequencing

1.1.3. Genome Sequence Assembly

- The process of piecing together a whole genome sequence from many small DNA segments, or reads, is known as genome assembly.
- Short sequence reads obtained from DNA clones are produced by DNA sequencing processes. These brief pieces must be connected while eliminating any overlaps in order to recreate the whole genome.
- Base calling, which generates base calls and gives these base calls quality scores, is the initial stage of the genome assembly process. The DNA fragments' nucleotide sequences at each site are ascertained using base calling. Each base call is given a quality score that indicates the accuracy or confidence of the base call.
- The individual sequence readings are then put together into continuous sequences, known as contigs
- Phred, Phrap, TIGR Assembler, and ARACHNE are a few frequently used tools for handling raw sequencing data and assembling contigs.

1.1.4. Genomes Annotation

- The genome sequence must be examined once it has been assembled in order to find and label important biological characteristics.
- Genomic annotation is the computational process of examining and understanding genomic sequences.
- Gene prediction and functional assignment are the two primary processes in the genome annotation process.
- Computational algorithms are employed in gene prediction to forecast the positions and configurations of genes across the genome.
- Assigning functions to the anticipated genes is the next stage. This is accomplished via homology searching, mostly with BLAST.
- The resulting annotations are then made available to the scientific community for additional study and research by being placed into open databases like as GenBank.

1.2. Methods in Functional Genomics:

1.2.1. Genetic Interactions Mapping

- The technique of genetic interaction mapping is employed to investigate the functional connections among genes.
- In order to comprehend how one gene affects the phenotype of another, this method entails perturbing genes in pairwise combinations, such as by deletion, knockdown, or overexpression.
- One such genetic connection is epistasis.
- It is a type of non-allelic gene interaction that happens when one gene's influence is obscured by another.
- The detection of gene-gene interactions can be accomplished using a variety of computational methods.
- The Bayesian Epistasis Association Mapping (BEAM), Tree-Based Epistasis Association Mapping (TEAM), Boolean Operation-based Screening and Testing (BOOST), and Two Stage-Grouped Sure Independence Screening (TS-GSIS) etc.
- Discovering new gene functions and hierarchically grouping gene products into functional complexes and pathways are two benefits of genetic interaction mapping.

1.2.2. Microarrays Technology

- A microarray is a chip that has complementary DNAs (cDNAs) or immobilized DNA oligomers arranged in a high density array.
- Every oligomer functions as a probe, capable of attaching itself to a particular complementary cDNA molecule.
- Tagged cDNA molecules are either fluorescently or radioactively tagged in a microarray and then allowed to hybridize with the oligo probes on the chip.
- The degree of mRNA abundance in the cell is shown by the quantity of fluorescence or radiolabels at each place on the microarray.
- This technique makes it possible to analyze gene expression patterns quickly and methodically, which speeds up the process of finding new gene functions and regulatory mechanisms.
- Microarrary helps in detecting changes in chromosomal copy counts, identify single nucleotide polymorphisms (SNPs) in genes, and compare patterns of gene expression between normal and sick tissues.

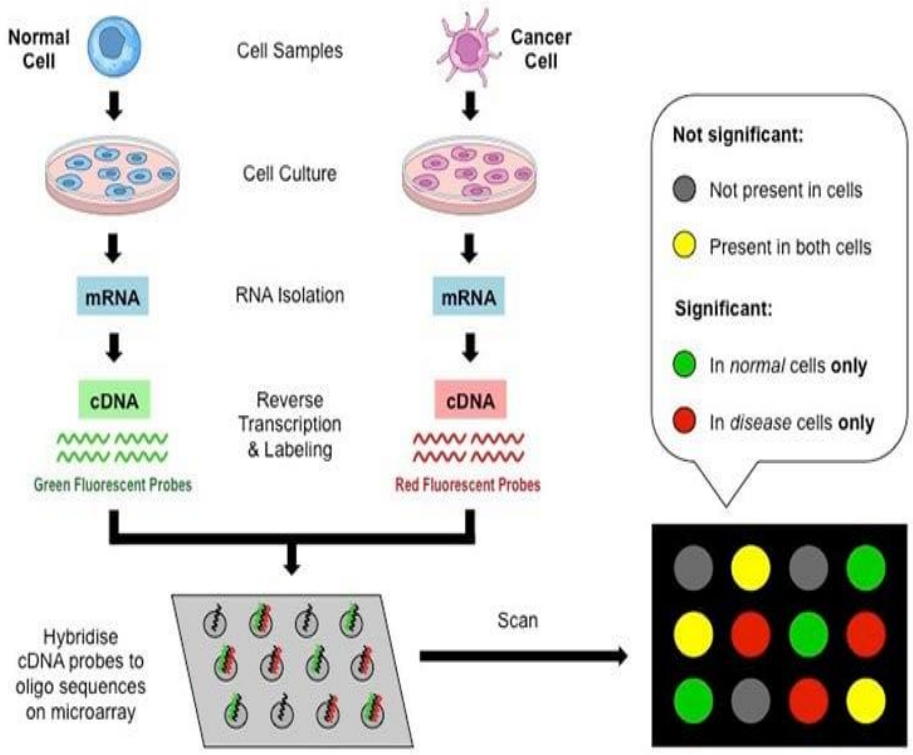

Figure 1.3 Microarray Technique

1.2.3. Serial Analysis of Gene Expression (SAGE)

- Gene expression patterns can be analyzed by a high-throughput, sequence-based technique called SAGE
- It offers details on the degrees of mRNA expression within a cell.
- SAGE involves separating individual mRNAs into short oligonucleotide sequence tags, or SAGE tags, which are then utilized as distinct identifiers for gene transcripts.
- These tags undergo cloning, sequencing, and concatenation. Gene expression analysis is carried out computationally in a serial fashion.
- The degree of gene expression is indicated by the frequency of each gene tag.
- Determining the number of tags to be sequenced for adequate coverage of the whole transcriptome may be difficult, and the sequencing process is expensive and time-consuming are the few limitation of SAGE.

Goals

The three primary goals of the bioinformatics field are as follows:

01. To efficiently arrange enormous stacks of molecular biology data

    In its most basic form, bioinformatics first classifies data so that scientists may access previously collected data.

    Furthermore, to more comprehensive inputs in a predetermined and structured manner

02. To create methods that facilitate the examination of said data

    The creation of resources and tools that promote data analysis is the second goal. Analyzing such complex data requires more than just a conventional text-based search, and technologies like PSI-BLAST and FASTA need to take into account what constitutes a physiologically meaningful match.

    The creation of such methods and tools is main goal of bioinformatics while possessing computational theoretical proficiency.

03. To appropriately and meaningfully assess the research results.

    Using the aforementioned resources to evaluate the data and analyze the findings in a way which conveys biological sense is the third goal.

We are able to conduct comprehensive evaluations of all the data that is accessible in bioinformatics with the goal of highlighting novel features while discovering universal principles that apply to a variety of systems.

1.2. Transcriptomics

The extensive investigation of the transcriptome, which consists of all RNA molecules present in an organism's genome, is known as transcriptomics.

RNA sequencing (RNA-Seq) and next-generation sequencing (NGS), continues to expand the boundaries of biology, whether it's necessary to profile genome-wide gene expression levels in a single experiment or to study only certain genes at a time.

Transcriptomics encompasses the examination of every mRNA transcript in a cell, which represents actively expressed genes at any given moment.

It is the study of RNA transcripts to learn about the production of proteins, the expression of genetic information, and the functioning of the cell.

Transcriptomics is crucial in comprehending how genetic information influences biological processes.

1.2.1 Types of Transcriptomics

2.1.1. Bulk Transcriptomics

Gene expression in large samples of millions of cells is studied using bulk transcriptomics, which yields a collective gene expression profile of the cell population as a whole.

This method aids in our comprehension of the patterns of gene expression in an extensive cell population.

It provides quantitative information on gene expression levels and may be utilized for analyzing thousands of genes at once.

2.1.2. Single-Cell Transcriptomics

Gene expression is examined at the individual cell level and is helpful in studying cellular variety.

To comprehend the many cell types and biological processes, it is crucial to be able to examine the gene expression levels of individual cells.

Cellular diversity and uncommon cell are studied using single-cell transcriptomics.

It also makes it possible for researchers to discover entirely novel types of cells.

2.1.3. Spatial Transcriptomics

Gene expression in a sample is measured using spatial transcriptomics, which maintains spatial information.

Understanding the function and structural arrangement of genes requires an examination of both their spatial distribution and expression level.

In this, spatial information is preserved while yielding important details about the composition, relationships, and activities of cells.

There are essentially two distinct types of spatial transcriptomics technologies: imaging-based and sequencing-based.

Transcriptomics is studied using both approaches while spatial information is preserved.

2.1.4. Meta Transcriptomics

The scientific investigation of RNA transcripts and gene expression across a broad range of species, including eukaryotes, plants, animals, and microbes, is known as meta transcriptomics.

The functions and interactions of organisms in their natural environments can be successfully studied with the help of this technique.

Meta transcriptomics is a widely used technique in environmental research for the study of intricate ecosystems, including soil, water, and gut microbiomes.

1.2.2 Methods Used in Transcriptomics

1.2.2.1. Expressed Sequence Tag (EST) Sequencing

- RNA transcripts are the source of short nucleotide sequences known as ESTs.
- Reverse transcription is the first step in the EST sequencing process, which turns RNA into complementary DNA (cDNA), which is then sequenced.
- Prior to the discovery of high-throughput techniques, the main sequencing method for ESTs was the Sanger sequencing method.

1.2.2.2. Serial Analysis of Gene Expression (SAGE)

- In SAGE mRNA is used to create short sequence tags which identify and quantify transcripts in biological samples.
- SAGE's primary goal is to transform mRNA into brief, distinct sequence tags that are associated with specific transcripts.
- The levels of gene expression are then determined by concatenating and sequencing these tags.
- SAGE involves isolating mRNA from the biological sample which is converted into cDNA using reverse transcription.
- Restriction enzymes cleave the cDNA into small fragments, which are then ligated to oligonucleotide adapters.
- These fragments are further cleaved to produce short tags, which are then joined together to form ditags.
- The ditags are concatenated and cloned into a vector for amplification, and sequenced.
- The resulting sequences are analyzed to identify and quantify the original mRNA transcripts.

1.2.2.3. Microarrays

- A microarray uses hybridization to quantify gene expression by placing small nucleotide oligomers, or DNA probes.
- The target samples' RNA is isolated, reverse transcribed into cDNA and tagged with fluorescent dyes.
- The microarray is hybridized with the labeled cDNA.
- A laser is used to scan the array once it has been cleaned.
- Gene expression levels are determined by measuring the intensity of light emitted by probes that hybridize to transcribed RNA.
- Smaller and more manageable data are produced using microarrays at minimal cost.
- Due to variations in laboratory protocols, analytical techniques, microarray results may vary significantly.

- As a result, RT-PCR and other techniques should be utilized to validate the results of microarray-based mRNA expression analysis.

1.2.2.4. RNA Sequencing (RNA-Seq)

- To quantify RNA transcripts, RNA-Seq combines computational techniques with high-throughput sequencing.
- The highly automated next-generation sequencing (NGS) systems constitute the foundation of this innovative technique.
- RNA-Seq enables the simultaneous sequencing of thousands of distinct RNA molecules in a single run.
- Extraction and conversion of mRNA into stable cDNA is the first step in the RNA-Seq process.
- The high-throughput techniques are used to sequence the cDNA fragments.
- Numerous transcripts, including rare and low-abundance transcripts, can be quantified using RNA-Seq.
- RNA-Seq is significantly more costly than microarrays.
- Large data sets produced by RNA-Seq need sophisticated bioinformatics expertise and powerful computer systems.

1.2.3 Applications of transcriptomics

- Applications for transcriptomics include illness profiling and diagnosis.
- Identification of regulatory elements, such as transcriptional start sites, which are crucial components in human illness.
- Transcriptomics offers valuable insights into many possible treatment drugs and helps comprehend the molecular pathways behind various illnesses.
- Transcriptomics contributes in personalized healthcare and utilized in pharmacogenomics to study how individuals respond to particular medications.
- Studying immunological responses and the interplay between the host and pathogen areas in which transcriptomics is helpful.
- Gene fusions, allele-specific expression, and Single Nucleotide Polymorphisms (SNPs) linked to illness can also be found using RNA-Seq.

- RNA-Seq is using to investigate immune-related disorders by sequencing patient derived T cell and B cell receptor repertoires.
- In order to characterize RNA expression in both the pathogen and the host concurrently throughout the infection phase, dual RNA-Seq has recently been employed.
- Transcriptomics enables the identification of the genes and mechanisms that respond to and reduce biotic and abiotic environmental stressors.
- RNA-Seq can be used to find previously unknown protein coding regions in sequenced genome.

1.3. System Biology

Systems biology is the study of how molecules, cells, organs, and organisms interact and behave as a functional unit.

The various elements of biological creatures interact in a multitude of ways.

Due to this extreme complexity. They can therefore be broadly regarded as integrated systems.

It is cooperative and integrates numerous scientific fields, including biology, computer science, engineering, bioinformatics, and physics,

Which helps to forecast how these systems will change in the future and under different circumstances.

1.3.1 Applications of System Biology

- In systems biology, experimental methods discover the elements of system their interrelationships, and the effects of perturbations on these elements.
- Systems biology is necessary to address complicated issues like cancer and other related ones.
- All impacts of potential perturbations (disturbances to the network) may be reasonably inexpensively anticipated in silico after a thorough model has been built. Forecast the result of intricate procedures, such as cancer therapy on the tumor
- Predicting system behavior in relation to medication development is necessary.
- Health care: models are required to comprehend illnesses and create treatment plans.

1.3.2 Future prospects

Systems biology models and predicts biological entities' actions at several scales by utilizing a wide range of computer approaches, including machine learning (ML) and artificial intelligence.

Convolutional neural networks (CNNs) are a type of neural network that utilized for tasks such as protein structure prediction, gene expression monitoring, and sequence alignment.

Typically, random forest is used for regression and classification tasks. The analysis of unstructured data, which uncovers underlying biological processes at the genomic level, requires the use of clustering methods.

As the discipline develops, a deeper comprehension of the processes behind these intricate illnesses will lead to novel therapeutic approaches and, ultimately, personalized medicine.

1.4. Functional Genomics

Functional genomics is an area of study that uses genome-wide methods to explain the roles and interactions of genes and proteins.

Data from several processes pertaining to DNA sequence, gene expression, and protein function are utilized in research.

These processes include transcription, both coding and noncoding, protein translation, protein–DNA, protein–RNA, and protein–protein interactions.

The interactive and dynamic networks that control gene expression, cell differentiation, and cell cycle progression are modeled using this data.

In functional genomics, genome editing is a crucial technology that allows one to alter or remove genes in cells in order to comprehend their functions in illness.

CRISPR is one of the most well-known genome editing technologies.

1.4.1 Difference Between Structural and Functional Genomics

Structural genomics: The study of genome structure and characterization is the focus of structural genomics.

It includes mapping genes and identifying other structural markers within the genome,

It also involves sequencing genomes to ascertain the whole DNA sequence of an organism's genetic material.

Functional Genomics: The goal of functional genomics is to comprehend the intricate connection between phenotype and genetics.

It entails the investigation of the roles of genes, frequently utilizing advanced techniques to examine gene expression, gene regulation, and protein interactions throughout the whole genome.

Two areas of genomics are structural and functional genomics.

Although structural genomics offers the genetic material's blueprint.

Functional genomics investigates how this blueprint translates into biological function inside an organism.

Functional genomics stresses the dynamic elements, such as gene transcription, translation, regulation, and the function of genetic sequences in cellular activities.

Whereas structural genomics is primarily concerned with the structure and organization of the genomic information.

1.5. Metabolomics

The systematic quantification of many metabolites is the focus of the developing discipline of metabolomics.

The primary goal of metabolomics is to determine which metabolites are associated with each biological trait and then to analyze the underlying processes.

Metabolomics serves as a method for characterizing observable molecular traits (also known as molecular phenotypes) connected to metabolism.

The topic of metabolomics is being explored in several fields, including environmental monitoring, microbial fermentation, and agriculture.

However, its applications are mostly focused on medical research.

1.5.1 Metabolomic methods

In metabolomics, nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) are the two dominant complimentary techniques.

The integrity of metabolites and specimens may be preserved using NMR.

Despite its potential to provide important insights into molecule structure, NMR's sensitivity is restricted.

Numerous mass spectrometry methods are utilized in metabolomics research, yielding remarkable levels of analytical sensitivity and specificity.

Lipidomics, the MS-based analysis of lipids, has proven to be a very productive field of medicine.

1.6. Molecular Phylogeny

Molecular phylogenetics is a branch of phylogeny that studies genetic, hereditary molecular distinctions, particularly in DNA sequences, to learn more about the evolutionary relationships of species.

If a character's overall similarity is measured and statistically assessed, the associations are referred to as phonetic.

The associations are said to as phylogenetic if the analysis of the characteristics reveals their evolutionary or genealogical development.

1.6.1 Types

There are three types of phylogenetic classifications:

monophyletic, polyphyletic, and paraphyletic.

Monophyletic: A monophyletic group of species consists of all of its offspring in addition to their common ancestor.

On a phylogenetic tree, a node and all of its progeny, represented by both nodes and terminal taxa, make up a monophyletic group.

Polyphyletic: When a creature is classified as polyphyletic, it means that it had at least two distinct ancestors.

Paraphyletic: When some of the progeny of the latest common ancestry are left out of a taxonomy, it is said to be paraphyletic.

1.6.2 Applications

In the domains of population genetics, genomics, and virology, among others, molecular phylogenies are today an indispensable tool.

Phylogenetics is used to:

- classify creatures,

- study reproductive biology in lower organisms,
- assess cryptic speciation in a species,
- understand the history of life
- resolve controversial historical disputes,

Phylogenetics is a trustworthy method to assess the evolutionary history current species. Scientists can gain insight into how species have developed and explain the differences and parallels among them by examining phylogenetic trees.

1.7. Applications and Limitations of Bioinformatics

Bioinformatics is an interdisciplinary scientific area that analyzes and interprets biological data by integrating computer science, information engineering, mathematics, and statistics with biology.

Methods and software tools for interpreting biological data are developed in the multidisciplinary subject of bioinformatics.

Biological databases, sequence alignment, gene and promoter prediction, molecular phylogenetics, structural bioinformatics, genomics, and proteomics are among the main fields of bioinformatics.

1.8.1 Applications of Bioinformatics

- Transcriptome analysis, which determines mRNA expression levels, is done using bioinformatics.
- In the fields of structural genomics, functional genomics, and nutritional genomics, bioinformatics is essential.
- Analyses including virtual screening, clustering, QSAR modeling, and similarity searches are all included in the field of cheminformatics analysis.
- A growing amount of bioinformatics is being used in practically every facet of medication research and discovery.
- Clinical and preclinical outcomes may be predicted, analyzed, and interpreted with great efficacy using bioinformatics techniques.
- Transcriptome analysis, which determines mRNA expression levels, is done using bioinformatics.

- The development of grain cultivars with increased resistance to iron, free aluminum, and soil alkalinity has progressed by the help of bioinformatics

1.8.2 Limitations of bioinformatics

- The quality of input data has a significant impact on the accuracy of bioinformatics analysis.
- Inadequate or inaccurate information might provide conclusions that are not accurate.
- No one algorithm works in every situation, and the algorithms used by bioinformatics tools determine their effectiveness.
- Handling genetically identifiable information presents ethical questions that need for careful consideration of data security and privacy.
- Large data volumes can put stress on the infrastructure and cause access to the data to get blocked up.
- For effective analysis, algorithms must be scalable to accommodate datasets of different sizes.
- Integration is difficult since biological data is available in many different forms (DNA sequences, protein structures, gene expression patterns, etc.).
- Large databases need a lot of work and resources to maintain and update. As databases get bigger, it gets harder to guarantee data relevance, correctness, and consistency.

Summary

This session will introduce you to the fascinating field of bioinformatics and its various objectives, including the use of computational biology in a wide range of scientific fields such as medicine, agriculture, and drug discovery. Bioinformatics is essential for large-scale data management and meaningful analysis across a range of disciplines including genomics, genetics, transcriptomics, and systems biology. It is also crucial for understanding molecular phylogeny and the intricate ways in which complex biological systems interact. Throughout this introductory session, we will explore topics such as genome sequence assembly, annotation, and the methods used in functional genomics, transcriptomics, and metabolomics. We will also delve into the world of molecular phylogeny and discuss its various applications. By the end of this

session, you will have a better understanding of how bioinformatics can revolutionize scientific research and enhance our understanding of the natural world.

Self-Assessment:

1. Explain the method used in Transcriptomics in detail with an example of each.
2. Illustrate the role of bioinformatic in molecular Phylogeny.
3. Give applications of SAGE analysis.
4. What do you mean by System Biology?

Unit-2
Biological Databases

Objectives:

- To gain the knowledge about different biological databases.
- Illstratating the analysis of biological information using various computational algorithms to assisst biological investigation.
- Explain the effects of the structure of chemical compounds in their metabolic activity.
- To understand the different metabolic activity of various dieseases by using metabolic pathway databases.

2.1 Introduction:

- A biological database is well-structured collection of long-term data that is typically paired with computer programs that allow users to update, query and retrieve specific data elements from the system.
- A single file holding several records, all of which have the same set of data, might constitute a basic database.
- The primary goal of creating a database is to arrange information into a collection of organized records that make information retrieval simple.
- A few popular databases are Gen Bank, Swiss Prot, PIR(Protein Information Resource)

2.2 Types of biological Database:

Biological databases are divided mainly in 3 categaories (See in Figure 2.1)

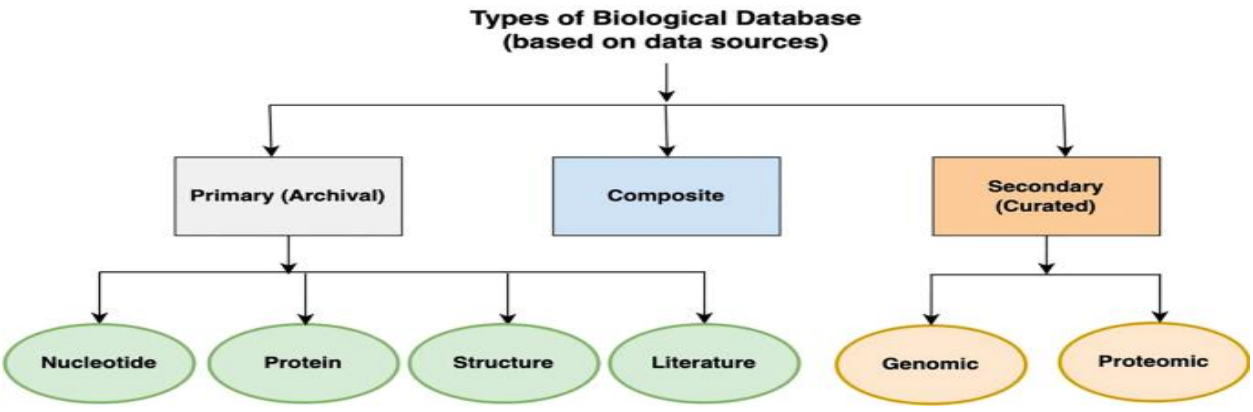

Figure 2.1: Different types of biological databases, Image sources: Springer Link

2.2.1 Primary databases:

- It can also be called an orgranize database because it holds the experiment findings that the scientists have provided, it is sometimes referred to as an archival database.
- since it archives the experimental results submitted by the scientists.
- It contain experimentally derived data.
- The information entered here is left uncurated (no changes are made to the information).
- It includes original data that was collected from the laboratory and available to regular users.
- When the data are added to the database, they are assigned accession numbers (a unique number assigned by a particular database).

Examples –

- GenBank and DDBJ are the primary databases for nucleic acids.
- SwissProt, PDB, TrEMBL, Metacyc, and PIR are examples of primary databases for proteins.
-

2.2.2 Secondary Database:

- The information kept in these kinds of databases is the outcome of analysis from the main database.
- In it, significant and informative data is stored after computational techniques are applied to the primary database.
- The information displayed here has been carefully selected and processed prior to being added to the database.
- Compared to the primary database, a secondary database has more important knowledge.
- InterPro (protein families, motifs, and domains) examples UniProt Knowledgebase (protein sequence and functional data)

2.2.3 Composite Databases:

These kinds of databases compare data first, then filter them according to predetermined standards.

They combine in accordance with predetermined criteria when initial data is received from the primary database.

It has non-redundant data and aids in quickly finding sequences.

Examples –

- Composite Databases -OWL,NRD and Swissport +TREMBL

2.3 Nucleic acid databases:

It is a type of biological database containing genetic information(DNA and RNA), that come from a variety of sources (genomes, transcriptomes, and individual genes).(See in figure 2.2)

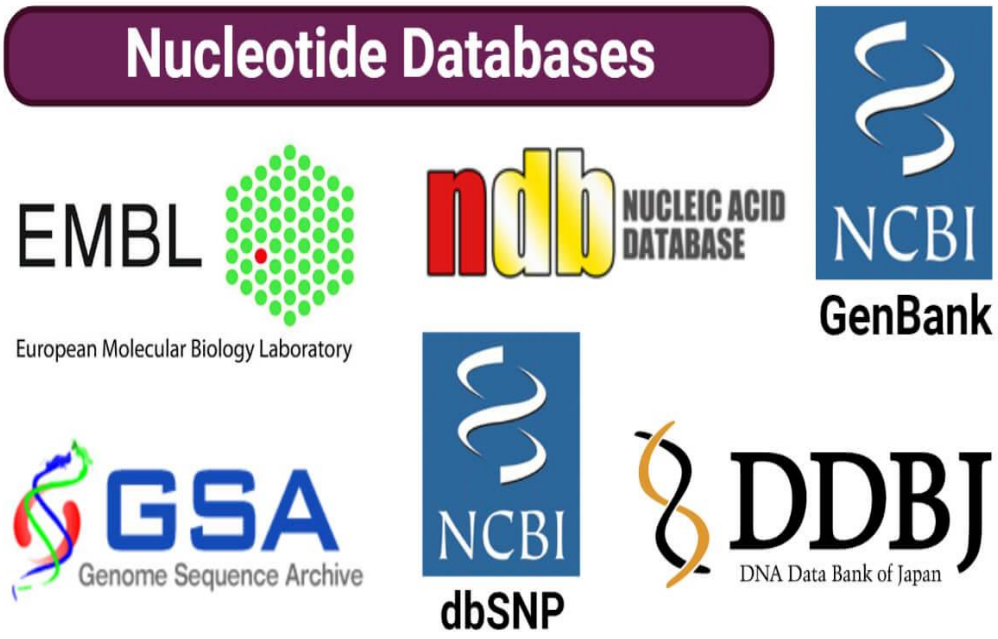

Figure2.2: Nucleotide Databases, Image Sources: Respective database websites.

Numerous nucleotide databases exist. These are a few of the nucleotide databases:

2.3.1 GenBank:

- This database has been developed and maintained at the NCBI(National Center for Biotechnology Information), as a part of International Sequence Database Collaboration (INSDC).
- Annotated nucleic acid sequence data are collected in the GenBank sequence database.

- It contains a variety of genetic information, including high-throughput raw sequence data, mRNA ,cDNA, expressed sequence tags , genomic DNA, and sequence polymorphisms.
- GeneBank files include phylogenetic categorization, accession numbers, gene names, and references.
- The database is an open access series.
- It collaborates with other sequencing databases such as DDBJ and EMBL as well as individual laboratories.
- It is a publicly accessible annotated library of all nucleotide sequences.
- The tools BankIt, Sequin, and tbl2asn can be used to submit sequences to GeneBank.

2.3.2 DDBJ (DNA Data Bank of Japan)

- It was established in 1984.
- It is one of the three biggest DNA databases in the world and joins GenBank and the EBI to establish an international DNA database.
- It is gathered and shared by DDBJ, which also manages bioinformatics tools for data submission and retrieval.
- Additionally, it creates biological data analysis tools and plans Japanese bioinformatics training courses.
- Most of the data in DDBJ comes from several sequencing centers, namely the International Consortium on Human Genome Sequencing.

2.3.3 EMBL(European Molecular Biology Lababoratory)

- It is an extensive collection of DNA and RNA sequences which is collected from patient offices, scientific publications, and submissions made directly by researchers.
- The DNA Database of Japan (DDBJ) and GeneBank (USA) have worked together to develop EMBL.
- It is established in 1980.
- It is maintained by EBI (European Bioinformatics Institute).
- It focuses on how protein and nucleotide sequences are distributed and stored.

2.3.4 NDB (Nucleic Acid Databases)

- Protein Data Bank (PDB) provides the 3D nucleic acid structures and their complexes that make up the Nucleic Acid Database (NDB).
- The database serves as a central location to store and retrieve annotations and structural data associated with nucleic acid.
- Annotations related to the structure and functioning of nucleic acids are included in NDB.
- It offers to explore the database, download information and structures, examine nucleic acids, and discover more.
- RNA and DNA oligonucleotides are included in the database. Protein-RNA and protein-DNA structures are also included.

2.4 Protein databases:

- One type of biological database is a protein database, which is a collection of data on proteins.
- It contains information about the biological role, interactions with other proteins, domain structure, amino acid sequence, and 3D structure of individual proteins.
- There are several protein databases which are available freely. Protein databases divided into different types based on the kind of data they hold.(See figure 2.3)

Figure2.3: Protein Databases., Image Source: Respective database websites.

Some of the Protein Databases are as follows:

2.4.1 PIR (Protein Information Resource)

- A popular protein sequence database that offers details on functionally annotated protein sequences is called PIR (Protein Information Resource).
- The Non-redundant Reference (NREF) sequence database, the Protein Sequence Database (PSD), and the Integrated Protein Classification (iProClass) database are the three databases that maintains by PIR .
- The iProClass database includes annotated protein sequences, classification data, and information on protein family, function, and structure.

2.4.2 SWISSPROT

- High level of annotation are available in the protein sequence database SWISS-PROT, which contains details on the function, domain structure, post-translational modifications, and variations of the protein.
- The European Bioinformatics Institute (EBI) and the Swiss Institute of Bioinformatics (SIB) worked together for Swiss-Prot.
- Three features of SWISSPROT that are different from other protein sequence databases: are :
    (i) Extensive cross-referencing and retrieval of information from related databases.
    (ii) Minimal redundancy, which guarantees that each sequence is represented only once.
    (iii) Integration with other databases.

2.4.3 TrEMBL

- It is established in 1996.
- Swiss-Prot included a computer-annotated supplement called TrEMBL. The format used for TrEMBL entries is Swiss-Prot.
- It contains all translations of nucleotide sequence entries from the EMBL that have not yet been included to Swiss-Prot.

- It also having protein sequences taken from literature and protein sequences submitted directly by the user.

2.4.4 PDB (Protein Data Bank)

- Drs. Edgar Meyer and Walter Hamilton founded the Protein Data Bank at Brookhaven National Laboratory in 1971. Members of the Research Collaboratory for Structural Bioinformatics (RCSB) took over the Protein Data Bank's maintenance in 1998.
- The Research Collaboratory for Structural Bioinformatics (RCSB) is now in under of PDB, a worldwide central repository for structural data on biological macromolecules.
- The about 8000 entries in the Protein Data Bank (PDB) at Brookhaven National Laboratory contain experimentally determined 3D structures of proteins, nucleic acids, and other biological macromolecules.
- The PDB website provides different types of services for structure submission and data searching and retrieval.
- The major of the database entries are protein structures.However, theoretical models, carbohydrates, and nucleic acids make just a tiny percentage of the database.

2.5 Metabolic pathway databases

Enzymes, biological processes, and metabolic pathways all are the part of metabolic pathway databases.

It contain different components(See in figure 2.4)

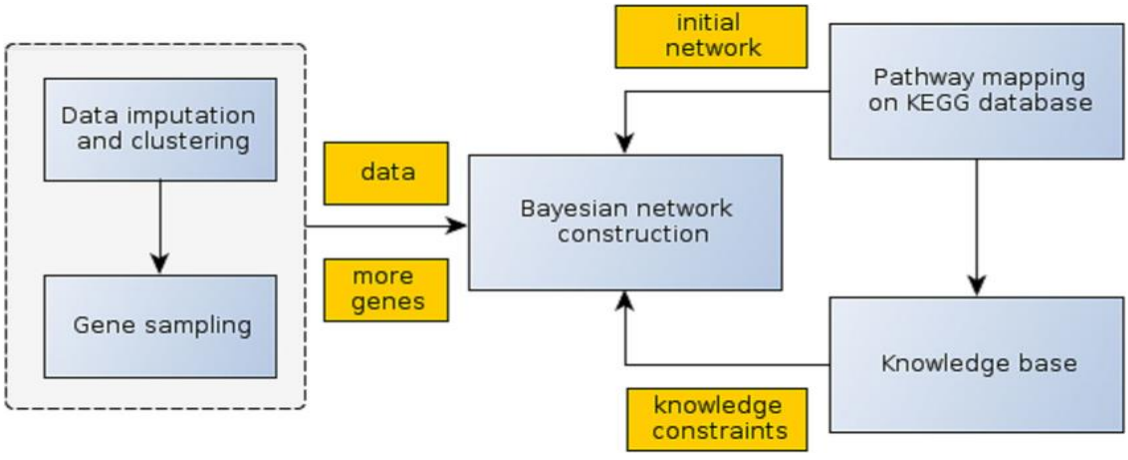

Figure 2.4The main components of the metabolic pathway construction method and their relationships. Image resource ResearchGate

Different types of Metabolic pathway databases are as follows:

2.5.1 KEGG (Kyoto Encyclopedia of Genes and Genomes):

- A huge database called KEGG that charts cellular and molecular pathways including interactions between molecules and genes.
- It is used to genrate functional maps of metabolic and regulatory pathways and is composed of pathway maps, molecule tables, gene tables, and genome maps.
- Graphical direction of maps for every known metabolic pathway from a range of species are available in the KEGG route Database.
- One of the most comprehensive and extensively used databases, KEGG (Kyoto Encyclopedia of Genes and Genomes) contains metabolic pathways (372 reference pathways) from a large number of species (>700).
- There are hyperlinks from these pathways to data on metabolites and protein_complex/enzymes. At the moment, KEGG has over 11,000 glycan structures, 7742 medications (including various salt forms and drug carriers), and more than 15,000 chemicals derived from plants, animals, and microorganisms.
- The KEGG Databases organization (See in figure 2.5)

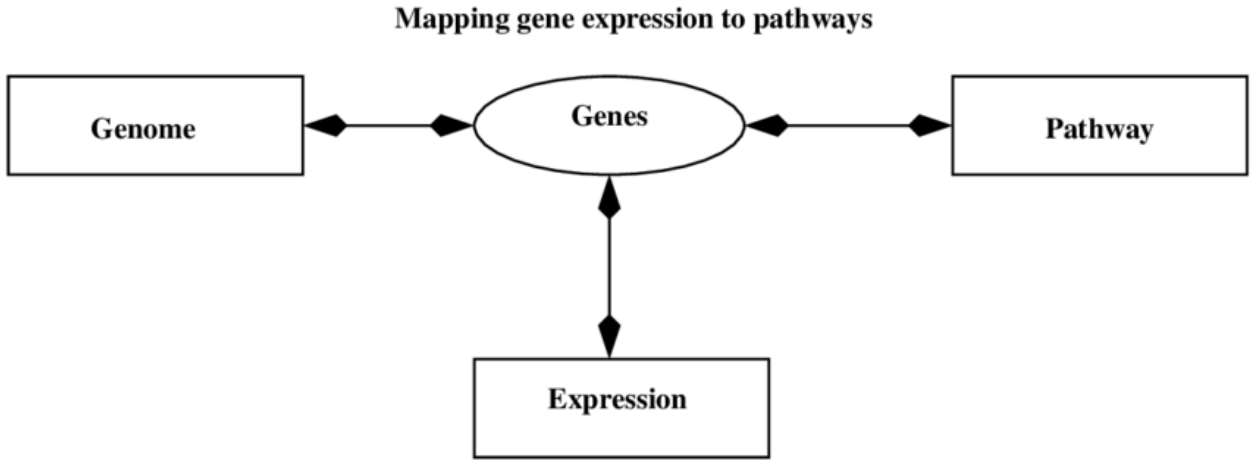

Figure 2.5 The KEGG Database organization, Image resource: ResearchGate

2.5.2 EcoCyc:

- EcoCyc is a database that contains details on the metabolic system and genome of Escherichia coli.
- EcoCyc collect data for Escherichia coli K-12 substr. MG1655 from 44,000 articles. Utilize EcoCyc to do comparative studies, review transcriptomics data, run metabolic models, and look for information on the genes, regulation, and metabolism of E. coli.
- The project's long-term goal is to provide a system-level knowledge of Escherichia coli by describing the whole molecular assets of the cell and the roles played by each of its molecular components. For scientists studying E. coli and related bacteria, EcoCyc is an electronic reference resource. Every E. Col gene product, metabolite, reaction, operon, and metabolic pathway has an information page in the database. The database also contains details on the control of gene expression, the essentiality of genes for E. coli, and the nutritional environments that either promote or inhibit E. coli growth. High-throughput data set analysis capabilities are available on the website and in the downloaded program.

2.5.3 MetaCyc

In order to predict metabolic pathways from an organism's annotated genome and create a PGDB, like AraCyc, MetaCyc is the reference library of pathways and enzymes that is used in conjunction with SRI's Pathway Tools system.

An entire reference library of enzymes and metabolic pathways from all area of life is called MetaCyc (MetaCyc.org). It is the biggest curated collection of metabolic pathways, with 2749 pathways culled from over 60 000 papers. With its evidence-based and carefully selected data, MetaCyc is an encyclopedic metabolic reference tool. In addition, hundreds of organism-specific Pathway/Genome Databases (PGDBs) are created using MetaCyc as a knowledge base and made available on BioCyc.org and other genomic portals.

2.6 Small molecule databases

Most biologists, biochemists, and bioinformaticians are still fascinated by "big" molecules like proteins and genes.However, the majority of biologists, biochemists, and bioinformaticians would prefer to overlook "small" molecules. Now it's becoming clearer than ever before that

tiny molecules like carbohydrates, lipids, and amino acids are significantly more significant in all facets of illness genesis and therapy than before thought.

Some of the small molecule databases are as follows:

2.6.1 PubChem

Developed and controlled by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM), an institution under the U.S. National Institutes of Health (NIH), PubChem is a public chemical information resource (Kim, 2016; Kim et al., 2016a; Wang et al., 2017). It collect descriptions of chemical substances and their biological activity from over 500 data sources and provides the public with free access to this information. PubChem has been an essential information resource for the biomedical research community in several fields, including cheminformatics, chemical biology, medicinal chemistry, and drug development, since its establishment in 2004 as part of the NIH Molecular Libraries Roadmap Initiatives.

Two- and three-dimensional structures, chemical and physical characteristics, bioactivity data, pharmacology, toxicity, drug target, metabolism, safety and handling, pertinent patents and scientific articles, and other sorts of chemical information are all available in PubChem. PubChem has information on a broad range of chemical entities, including siRNAs, miRNAs, carbohydrates, lipids, peptides, chemically modified macromolecules, and many more, even though the majority of its data are on tiny molecules. Many sources, including government organizations, academic institutions, pharmaceutical businesses, chemical suppliers, publications, and a variety of chemical biology websites, have contributed this data. The majority of the chemical databases covered in this research provide data to PubChem as well.

2.6.2 Drug Bank

The University of Alberta and The Metabolomics Innovation Centre in Alberta, Canada, are the creators and maintainers of the vast, publicly available DrugBank database, which is an online resource with data on medications and pharmacological targets.

A web-enabled database called DrugBank (www.drugbank.ca) offers rich molecular data regarding medications, their mechanisms, interactions, and targets. Initially presented in 2006, DrugBank has undergone continuous modifications in the last 12 years due to significant

advancements in web standards and shifting requirements for drug development and research. DrugBank 5.0, the update released this year, is the biggest addition to the database in over a decade. Existing data content has frequently increased by 100% or more since the last upgrade.
DrugBank is a well annotated resource that containingextensive drug target and drug action information together with specific drug data.

2.6.3 ZINC

It is developed by John Irwin.It is a chemical database.
ZINC is an open-access, free tool for ligand finding.More than 20 million commercially accessible molecules in physiologically suitable representations are included in the database, which may be downloaded in widely used ready-to-dock formats and subsets. Searches by structure, biological activity, physical characteristic, vendor, catalog number, name, and CAS number are also possible on the website. It is possible to build, modify, share, dock, download, and send small custom subsets to a vendor for purchase. The database is publicly accessible at zinc.docking.org and is regularly updated and reviewed to ensure a high purchase success rate.

2.6.4 CSD

CSD stands for CAMBRIDGE STRUCTURAL DATABASE.
The world's archive for organic and organometallic small-molecule crystal structures is CSD. This database, which is made available by the Cambridge Crystallographic Data Centre (CCDC), is extensively utilized in structural chemistry. The Cambridge University Department of Chemistry, which compiles and makes the CSD database accessible, founded CCDC. "Advancement and promotion of the science of chemistry and crystallography for the public benefit" is CCDC's stated mission. Comprehensive data on small-molecule crystal structures determined by X-ray crystallography or neutron diffraction methods is compiled by CCDC into a database.
The CCDC has amassed a significant amount of experience in a variety of scientific fields, such as information mining, crystal engineering, medication design, and molecular structure analysis.

In celebration of 50 years of crystal structure data exchange via the CSD, the CCDC held a celebration.

Olga Kennard started gathering crystal structures in 1965 because she thought that by pooling experimental data, new information might be discovered that would go beyond the specific findings of particular research. The CCDC is mostly responsible for compiling the more than 7,83,500 updated entries and making them accessible to all scientists worldwide. Information on chemical, experimental, physical, structural, and bibliographic aspects is included in every structure deposited in CSD.

Summary

This research aims to achieve the following objectives:

1. Gain knowledge about different biological databases.
2. Illustrate the analysis of biological information using various computational algorithms to assist biological investigation.
3. Explain the effects of the structure of chemical compounds on their metabolic activity.
4. Understand the different metabolic activity of various diseases by using metabolic pathway databases.

Biological databases are structured collections of long-term data, typically paired with computer programs that allow users to update, query, and retrieve specific data elements from the system. The primary purpose of creating a database is to organize information into a collection of records for easy information retrieval. Some popular databases include GenBank, SwissProt, and PIR (Protein Information Resource).

There are different types of biological databases:

1. Primary databases hold experiment findings and contain experimentally derived data. Examples include GenBank and DDBJ.
2. Secondary databases store information resulting from analysis of the main database. InterPro and UniProt are examples of secondary databases.
3. Composite databases compare and filter data according to predetermined standards. Examples include OWL, NRD, and Swissprot + TREMBL.

Nucleic acid databases contain genetic information such as DNA and RNA from various sources. Examples include GenBank, DDBJ, EMBL, and NDB.

Protein databases store data about the biological role, interactions, domain structure, amino acid sequence, and 3D structure of individual proteins. Examples include PIR, SWISSPROT, TrEMBL, and PDB.

Small molecule databases contain information about chemical substances and their biological activity. Examples include PubChem, DrugBank, ZINC, and CSD (Cambridge Structural Database).

This research seeks to provide an in-depth understanding of biological databases, their types, and their applications for broader knowledge and research in the field of biological information.

Self-Assessment:

1) Write down the various biological databases and their significance in storage of sensitive biological data.
2) Differntiate between the primary, secondary and composite databases.
3) Explain different types of metabolic pathways databases.
4) Define :
    i. Drug Bank
    ii. CSD

Unit - 3

Data Generations and Data Retrieval

Objectives:

- Make student aware to the biological data generation sources and their protocols.
- Explain different types sequencing methods that are used extensively in research.
- Describing and illustrating different types of biological data file formats and their applications.
- Define the various DRS(Data retrieval system) use to collect large amount of biological data from online repositories.

3.1 Generation of Data

3.1.1 Genomic Sequencing

The whole of an organism's DNA or RNA is called its genome.
All of the information needed for a creature to operate is included in its genome.
The human genome has roughly 30,000 genes and is composed of about 3 billion base pairs.

Certain machines sequence it in order to determine the etiology of a certain illness.
There is extremely little variation in the DNA that causes some disorders.

By sequencing the genome, we can identify the specific DNA changes that are causing the problem.
Genome sequencing is a flexible technology that may be applied to a wide range of organisms, including viruses, bacteria, fungi, parasites, animal vectors, and human hosts.
It is a technique used to analyze the genetic composition of a particular organism or cell type.

The discovery of harmful organisms, the tracking of disease outbreaks, and the mutations that cause medication resistance have all benefited from this knowledge.

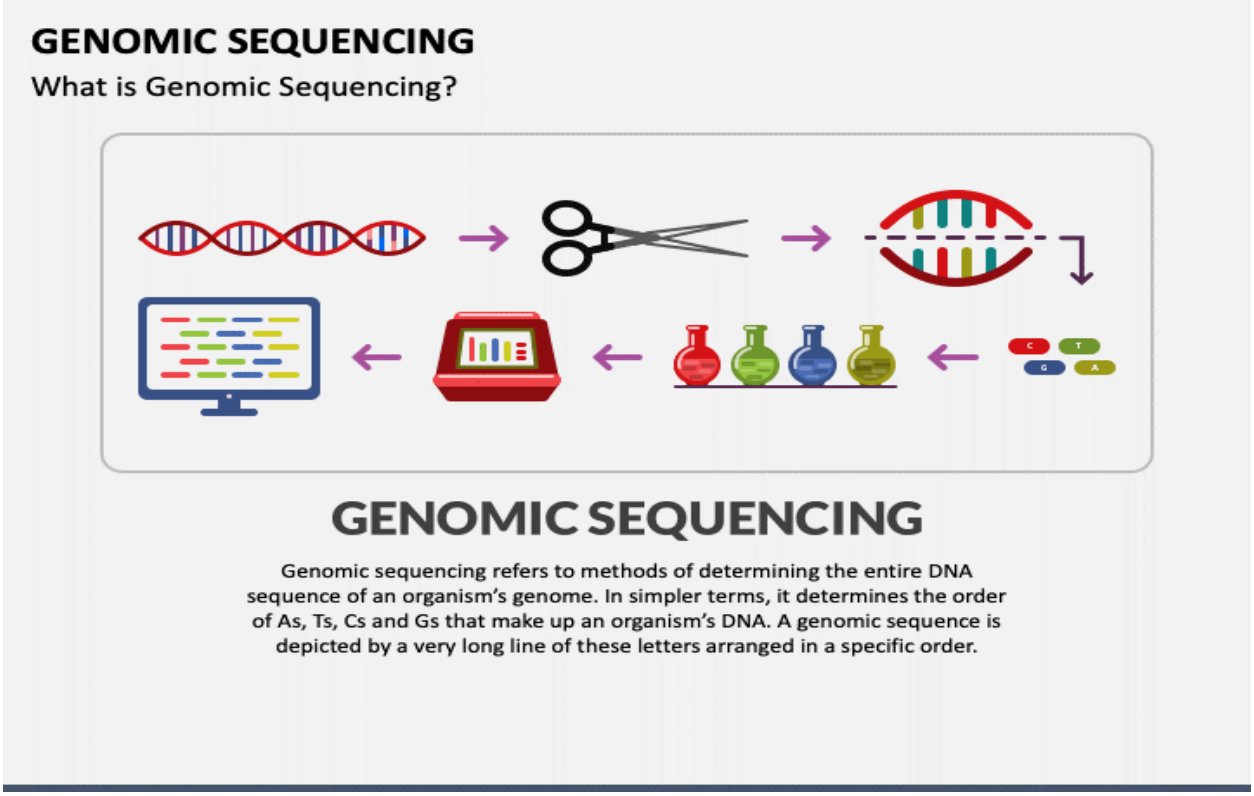

Fig 3.1 Genome Sequencing Simplified

Image Source: (1https://www.insightsonindia.com/wp-content/uploads/2024/03/genomic_.png)

decreasing sequencing costs quickly.

Genome sequencing is an effective research technique because of the modern sequencers' enhanced capacity to generate vast amounts of data.

Methods

The different strategies used for genome sequencing such as Sanger method, shotgun sequencing, and NGS.

Sanger method:

Sanger sequencing is sometimes referred to as the sequencing by synthesis method, the dideoxynucleotide technique, or the chain termination method.

It entails employing a single strand of double-stranded DNA as the sequencing template.

Dideoxy-nucleotides (dNTPs), chemically modified nucleotides, are used in his sequencing. These dNTPs are designated as ddG, ddA, ddT, and ddC for each DNA base. Dideoxynucleotides (dNTPs) are used to elongate nucleotides.

Once they are integrated into the DNA strand, they stop further elongation, indicating that the elongation process is complete.

After that, we get DNA pieces of varying sizes that are terminated by a dNTP.
The fragments are separated based on size using a gel slab, and an imaging system (UV or X-ray) can see the resulting bands that correspond to the DNA fragments.

Shotgun sequencing:

- A DNA fragment is sliced into several smaller pieces at random from multiple copies using the shotgun sequencing technique.
- The chain-sequencing approach is then used to sequence each segment.
- Next, the pieces are examined using a computer to determine where their sequences overlap.
- The complete DNA sequence may be rearranged by matching the overlapping sequences at the end of each fragment.
- The creation of pairwise-end sequencing is also formally known as double-barrel shotgun sequencing.
- Pairwise-end sequencing examines the overlap between each fragment's two ends.
- Because there is more information available, pairwise-end sequencing is more laborious than shotgun sequencing, but it is also easier to rebuild the sequence.

Next-generation sequencing:

- NGS is a collective term for a set of automated methods used for quick DNA sequencing that includes the automated sequencing methods used in labs.
- In a single day, these inexpensive, automated sequencers may produce sequences including hundreds of thousands or even millions of small fragments (25–500 base pairs).

- The tedious task of organizing all the pieces is handled by sophisticated software.
- The Sanger technique, shotgun sequencing, pairwise-end sequencing, and NGS are among the several approaches utilized for whole-genome sequencing.
- NGS has a lot of potential as a surveillance tool for hospitals. Even while technology has advanced to the point that microbiology labs might be able to employ it, the hospital still has to set up a supportive infrastructure.

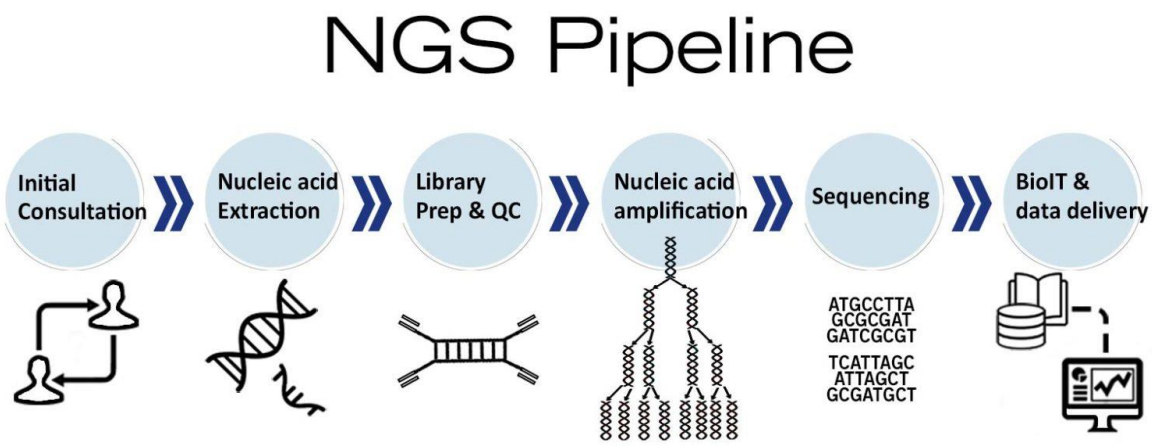

Fig 3.2 Depiction of NGS workflow used as molecular diagnostics tool.

Source : https://the-dna-universe.com/wp-content/uploads/2020/09/NGS-Pipeline_bigger-1536x638.jpg

3.1.2 Protein Sequencing

A protein's amino acid sequence may be determined by a mix of methodologies, methods, and techniques known as protein sequencing.

An amino acid's position in a protein's linear chain is known as its amino acid sequence, or primary structure.

Protein sequencing is a part of post translational modifications.

The basic procedure for primary structure determination is developed by Sanger.

The procedure consists of three conceptual parts, each of which requires several laboratory steps:

1. Get the protein ready for sequencing: Count the number of subunits, or chemically distinct polypeptide chains, that make up the protein. Break the disulfide links in the

protein.

Sort and refine the distinct components.

2. Sequence the chains of polypeptides by:
   a. Dividing the constituent subunits at certain locations to produce peptides short enough for direct sequencing.
   b. Sort and cleanse the pieces.
   c. Find out each peptide fragment's amino acid sequence.
   d. To ensure that the subunit is cleaved at distinct peptide bonds than previously, repeat Step a using a separate fragmentation technique.
3. Arrange the final structure as follows:
   a. Spread the cleavage sites across two sets of peptide fragments. In contrast, the amino acid sequence of the subunit may be determined by placing the sequences of these sets of polypeptides in the order that they appear in it.
   b. Clearly state where any disulfide linkages between and within the subunits are located.

3.1.3 Mass Spectrometry

Identifying novel substances, quantifying existing materials, and clarifying the structure and chemical characteristics of molecules are all made possible by the potent analytical method known as mass spectrometry.

One tool that assists scientists is mass spectrometry, which allows scientists to:

1. identify compounds that are present in solids, liquids, and gasses.
2. ascertain the amount of every kind of molecule.
3. ascertain the composition and arrangement of the atoms that make up a molecule.

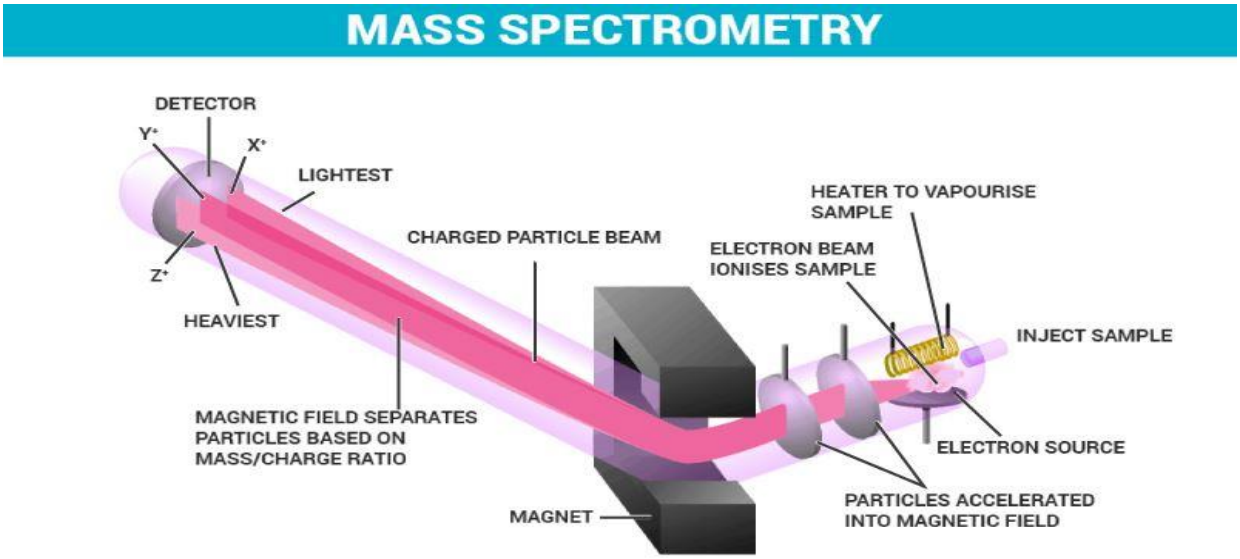

Fig 3.3 Mass Spectrometry Instrumentation

Source: https://cdn1.byjus.com/wp-content/uploads/2018/11/Mass-spectrometry-1.jpg

In the biotechnology industry, mass spectrometry is used

- to monitor fermentation processes,
- identify steroid use in athletes,
- detect heart attacks from blood tests,
- identify molecular species found in space,
- Measure the amounts of petroleum precursors in rock to find oil reserves.
- detect breathalyzer readings during surgery.
- Detect dioxins in contaminated fish
- Determine how drugs are used by the body
- Determine gene damage from environmental causes
- Sequencing proteins, nucleic acid and oligosaccharides

To begin a mass spectrometric examination of a substance, the compound must first be produced as gasphase ions, maybe by electron ionization:

$M + e^- \longrightarrow M^{\bullet +} + 2e^-$

It is typical for this molecular ion to fragment. Its odd number of electrons allows it to fracture into either a molecule and a new radical cation, or into a radical and an even number of electrons in the form of an ion.

Fragmentation can occur in each main product ion that is produced from the molecular ion, and so on. The mass-to-charge ratio of each of these ions determines how they are divided in the mass spectrometer are detected in proportion to their abundance.

The mass spectrum of the molecule is thus created. This finding is shown as an ion abundance vs mass-to-charge ratio plot.

An apparatus used to measure the masses of individual molecules that have been transformed into ions is a mass spectrometer., i.e., molecules that have been electrically charged.

3.1.4 Microarray

- DNA is connected to solid supports—typically made of silicon or glass—in an orderly, pre-planned grid pattern to create DNA microarrays.
- Every single DNA spot, referred to as a probe, represents a single gene.
- Tens of thousands of genes can have their expression analyzed at once using DNA microarrays.
- DNA microarrays can also be referred to as DNA chips, gene chips, DNA arrays, gene arrays, and biochips.
- The hybridization of nucleic acid strands is the basis for DNA microarray technology.
- Complementary nucleic acid sequences have the ability to precisely pair with one another by creating hydrogen bonds between base pairs of complementary nucleotides.
- Fluorescent dyes are used to label samples for this purpose.
- To chip, at least two samples must be hybridized.
- Hydrogen bonds are used to match complementary nucleic acid sequences on the chip between the probe and the sample.
- During the procedure' washing stage, the non-specific bonding sequences wash away and stay detached.
- When fluorescently labeled target sequences attach themselves to a probe sequence, a signal is produced.
- The signal is dependent on the hybridization circumstances (such as temperature, post-hybridization washing), but the total signal intensity is dependent on the quantity of target sample present.

This approach allows for the screening of the presence of a single genomic or cDNA sequence in up to one million sequences during a single hybridization process.

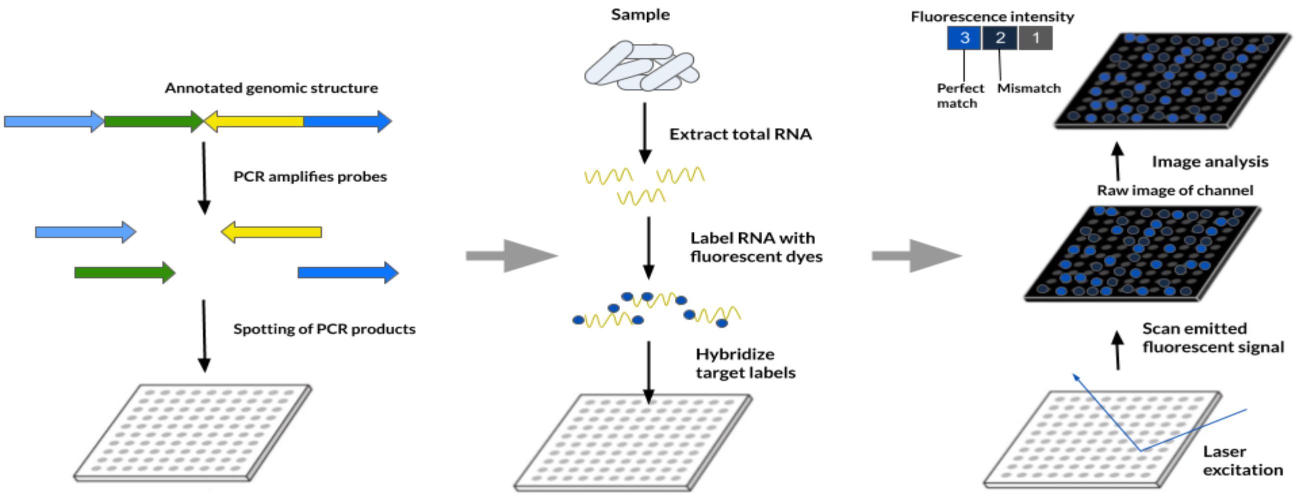

Fig 3.4 Visual Depiction of Microarray

Source:
**

The DNA microarray reaction technique includes several stages.

1. Gathering of examples

A cell or tissue from the creature we want to examine might be the sample.

To acquire data and for comparison, two types of samples—infected cells and healthy cells—are collected.

2. mRNA Isolation

With the use of a column or solvent such as phenol-chloroform, RNA is extracted from the material.

rRNA and tRNA are left behind after mRNA is isolated from the retrieved RNA.

Since mRNA has a poly-A tail, mRNA is bound using column beads that have poly-T tails.

In order to separate the mRNA from the beads, the column is washed with buffer following the extraction.

3. Production of tagged cDNA

Reverse transcription of the mRNA results in the creation of cDNA, or complementary DNA strand.

After then, various fluorescent dyes are added to both samples to create fluorescent cDNA strands. This aids in determining the cDNA sample category.

4. Hybridization

The tagged cDNAs from both samples are extensively cleaned to eliminate unbounded sequences before being inserted into the DNA microarray to allow each cDNA to hybridize to its corresponding strand.

5. Gathering and evaluating

The microarray scanner is used to capture the data.

Three components make up this scanner: a computer, a camera, and a laser. Signals are produced when the laser stimulates the cDNA's fluorescence.

The camera captures the pictures created while the laser scans the array.

After that, the computer saves the information and displays the outcomes right away. After then, the data are examined.

The characteristics of the gene at each place are determined by the variations in color intensity.

Applications of DNA Microarray

In humans, they can be used to identify the identity of the infecting organism from the expression profile or to ascertain how specific illnesses impact the expression profile, or pattern, of gene expression in different organs. Therefore, DNA microarrays offer enormous diagnostic potential in clinical medicine alone.

It may be used in a variety of disciplines, including:

Discovery of drugs

Diagnostics and genetic engineering

Alternative splicing detection

Proteomics

Functional genomics

DNA sequencing

Gene expression profiling

Toxicological research (Toxicogenomics)

3.2 Sequence Submission Tools

3.2.1 BankIt

The National Center for Biotechnology Information's (NCBI) GenBank sequence database submission tool is commonly referred to as "bankit".

Researchers can contribute their DNA or RNA sequences to the GenBank database online via BankIt.

It is an online submission mechanism offered by NCBI.

Nucleotide sequences are accessible to the general public in GenBank, an extensive database..

Researchers can submit their sequences to BankIt along with pertinent metadata. Which include facts on the organism, publication history, and experimental procedures.

The freshly produced sequence data may now be accessed and used for study and analysis by the scientific community thanks to this submission process.

Sending sequences via Bank to GenBankEnsuring data sharing and reproducibility is crucial for molecular biology and bioinformatics research.

Following submission and approval, the sequences are added to the GenBank database, enabling researchers from all over the world to search, examine, and compare them with other sequences.

BankIt should use this web-based GenBank submission facility to submit one or more of the following types of sequences:

Complete or partial genes that code for proteins found in viruses, prokaryotes, and eukaryotes (such as the gene for mitochondrial cytochrome oxidase subunit I)

Other gene types, excluding ribosomal RNA genes (rRNA) and internal transcribed spacer sequences (ITS), including those sequenced as a part of phylogenetic, population, or mutational studies/sets

Mobile components like insertion sequences and transposons

Genomic markers other than microsatellites

Viruses' whole or partial genomes, except influenza and norovirus

Phage genomes (complete or partial); unless you are also submitting the entire bacterial genome as a Complete genome submission or a Whole Genome Shotgun (WGS)

Complete or partial genomes of mitochondria, chloroplasts, or other plastids, as well as the genes derived from these organelles

Cloning and expression vectors are examples of synthetic constructions.

Third Party Annotation/Assembly sequences (TPA), let you submit your own annotation or assembly on sequences that are made publicly available (GenBank main sequences).

3.2.2 Sequin

Sequin is a stand-alone tool developed by the NCBI that allows sequences to be uploaded and updated to the GenBank, EMBL, and DDBJ databases.

Long sequences and collections of sequences (segmented entries, as well as demographic, phylogeny, and mutation studies) may be handled using Sequin.

It also offers sophisticated annotation features and permits editing and update of the sequence. Furthermore, Sequin has several integrated validation features for improved quality assurance.

It is capable of handling a broad variety of sequence lengths and complexity, including big datasets from phylogenetic or demographic studies as well as complete chromosomes.

The GenBank and Reference Sequence indexers at NCBI also employ Sequin for standard record processing prior to publication.

Sequin's modular architecture makes it easier to use, create, and execute. Since Sequin depends on several NCBI Toolkit components, it serves as a quality control measure to ensure that these features are operating as intended.

The sequence (or combination of sequences) may be an already-existing GenBank sequence record or fresh data that has not yet been given a GenBank Accession number.

In the event that Sequin is used to submit a sequence or sequences to GenBank, the scientist will first be required to provide contact details, details about additional authors, and the sequence.

Following the completion of all required fields, the sequence may be edited using Sequin's editing tools and seen in a number of ways.

3.2.3 Webin

The method used to submit nucleotide sequences and the biological annotation associated with them to EMBL-Bank is called Webin.

The web-based tools for submitting nucleotide sequences, nucleotide sequence alignments, and protein sequences to the EMBL-Bank, EMBL-Align, or SWISS-PROT databases are called Webin, Webin-Align, and SPIN, respectively.

Webin is intended to facilitate the quick submission of one or more nucleotide sequences. Submitters of bulk submissions can enter their data with little effort thanks to auto-generated templates.

These tools lead you through a series of interactive Web forms that enable submission. During this phase, all the data needed to build a database entry is gathered.

The EBI developed the EMBL-Align database in response to the growing need for alignment data storage.

Alignments are provided in alignment flat file format through the EBI FTP server and the SRS server. The specific web-based submission mechanism for uploading multiple sequence alignments to the EMBL-Align database is called Webin-Align.

It is compatible with all popular alignment formats, including Phylip, NBSF/PIR, Nexus (interleaved), Clustal, and GCG/MSF.

3.2 Sequence File Format

3.2.1 Flat File

In essence, a flat file is a simple text file without any organized relationships to other files or databases. Simple examples of this kind of file are text documents, Excel spreadsheets, binary files, and CSV (Comma-Separated Values) files.

One table makes up a flat file database, and the size of the table depends on how much data it needs to store.

A list of names, addresses, and phone numbers is a common example; this type of data structure is known as a flat file database.

Because they make data import and export across various software programs or systems easier, flat files are especially appreciated.

They are kept in plain text format, with fields inside a record being separated by a designated delimiter (such as a tab or comma) and each line denoting a separate record.

Furthermore, flat files may be easily created, edited, and customized with simple text editors, enabling users to modify the content and structure of these files to satisfy particular needs for data export or import.

3.2.2 FASTA

Base pairs or amino acids are represented using single-letter codes in the text-based FASTA format, which may be used to represent nucleotide sequences or peptide sequences.

A sequence in FASTA format is made up of lines of sequence data that follow a single line description.

The first column's greater-than (">") symbol divides the sequence data from the description line.

The FASTA defining line, which appears before the nucleotide sequence in the format, must start with a carat (">") and end with a distinct SeqID (sequence identifier).

Every nucleotide sequence needs a different SeqID, and spaces should be avoided in them.
Researcher should keep the SeqID to a maximum of 25 characters. Only letters, numerals, hyphens (-), periods (.), underscores (_),asterisks (*),colons (:) and number signs (#) are permitted in the SeqID.

The database staff will change the sequence identification with an Accession number after processing your submission.
›SeqABCD [organism=Mus musculus] [strain=C57BL/6]

After the SeqID comes the format of the source organism information, which has to be in the form of [modifier=text].

The "=" should not have spaces surrounding it. The organism's scientific name ought to be mentioned at the very least. Modifiers that are optional can be added to offer further details.

The sequence title, which will be utilized as the DEFINITION field in the flatfile, is the last optional part of the FASTA definition line.

The sequence should be briefly described in the title.
Titles for proteins and nucleotides have a recommended format. During processing, the database staff will format the specified title correctly.
›SeqABCD [organism=Mus musculus] [strain=C57BL/6] Mus musculus neuropilin 1 (Nrp1) mRNA, complete cds.

Hard returns must not appear on the FASTA defining line. Every piece of text needs to fit on a single line.

3.2.3 GCG file format

There is only one sequence in a sequence file in GCG format, which starts with annotation lines and is identified at the beginning by a line that ends in two dots ("..").

The sequence length, checksum, and sequence identification are also contained in this line. Use of this format is limited to files generated using the GCG package.

The header and sequence data are the two primary portions that make up GCG files.

Important details about the sequence, including as its name, length, source organism, and any further annotations or remarks, are included in the header section.

The actual nucleotide sequence, on the other hand, is shown as a string of characters in the sequence data portion.

When displaying a sequence, line breaks and spaces are usually used to separate individual nucleotides for ease of reading and analysis.

Included in the GCG DNA Sequence data are the.Nucleotide sequences are stored in specialized text files that end with GCG.

GCG files, which were created by the Genetics Computer Group (GCG), are often used in bioinformatics to represent and analyze DNA sequences.

Users can use a variety of software programs intended for sequence analysis to access GCG files.

3.2.4 EMBL File Format

EMBL entries (as seen below) are formatted such that both computer programs and human readers may use them. Lines make up each record in the database.

Various line formats, each with a unique purpose, are employed to document the many kinds of data that comprise the entry.

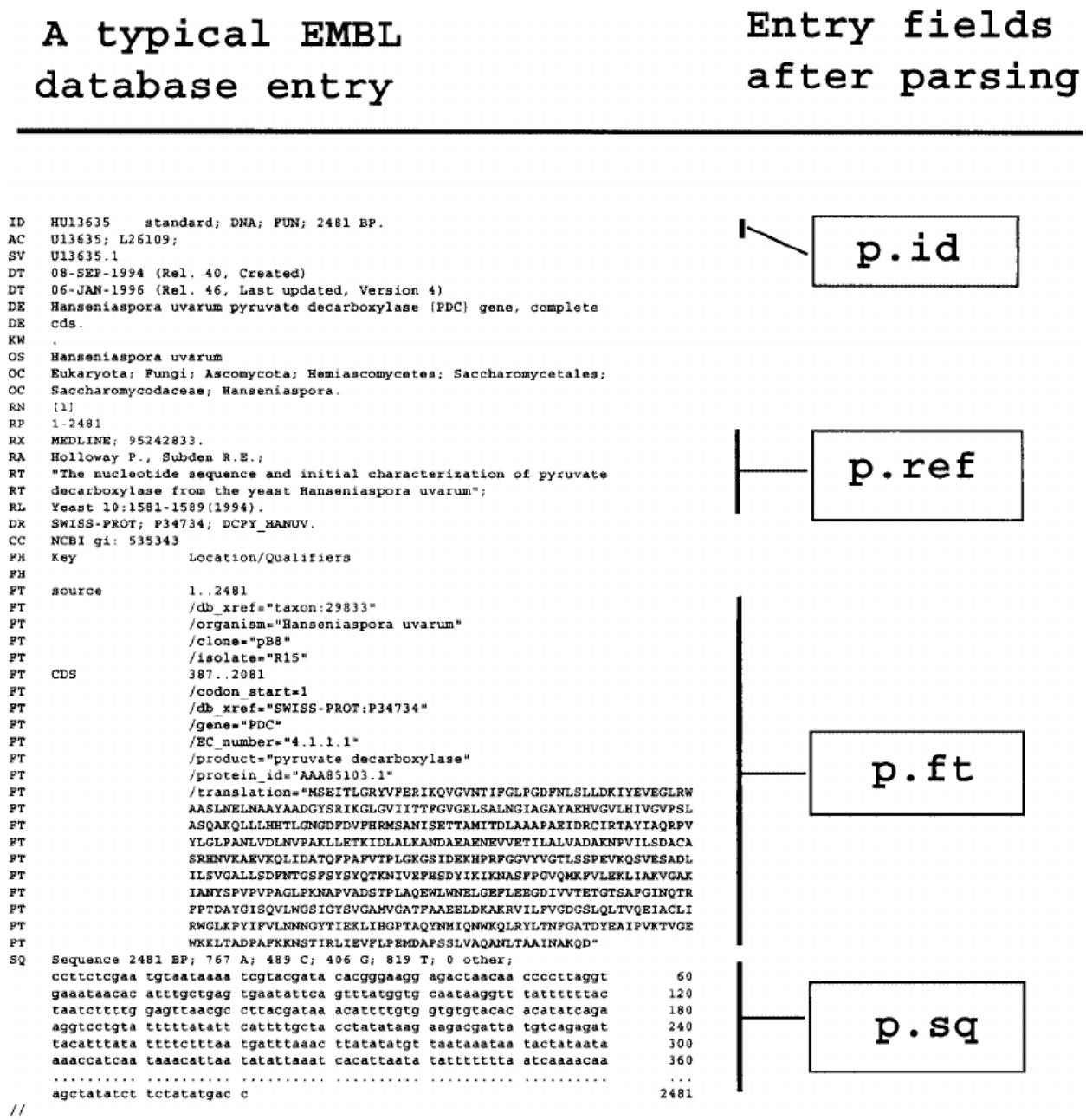

Fig 3.5 Example of an EMBL database entry.

Source: https://www.researchgate.net/profile/Christine-Gemuend/publication/12286757/figure/fig1/AS:281757795536897@1444187701775/Example-of-an-EMBL-database-entry.png

Certain line types appear more than once in a single entry, and some entries do not contain all of the line kinds.

An identification line (ID) and a terminator line (//) mark the start and finish of each entry, respectively. Refer to the EMBL user handbook for a more thorough instruction set.

The ID, SQ, and last "//" are format elements that are essential for accurate sequence analysis. The sequence data is stripped of any whitespace (spaces and tabs), return carriages, and numbers.

3.2.4 Clustal File Format

The extension ".aln" is typically used to identify files in CLUSTAL format.

The alignment software ClustalW is where the ALN format first appeared. The term "CLUSTAL" appears at the beginning of the file, followed by details about the specific clustal application that was used and its version.

Such is "CLUSTAL W multiple sequence alignment" The version of the clustal application is 2.1 and its type is "W".

The alignment is written in sixty-residue blocks. The sequence names for each block are taken from the input sequence, and at the end of each line is a count of all the residues.

Below each block of residues is the information pertaining to which residues match:

A "*" indicates that all sequences in the alignment have the same residues or nucleotides in that column.
":" indicates the presence of conserved substitutions.
"." denotes the observation of semi-conserved replacements.

- The following is a more precise description of the format:
- The words "CLUSTAL W" must be at the beginning of the file.
- One or more empty lines.
- One or more blocks of sequence data. Each block consists

One line for each sequence in the alignment. Each line consists of:

the sequence name

white space

up to 60 sequence symbols.

White space is optional and is followed by the sequences' cumulative residue count.

A line that indicates how much the alignment's columns have been conserved in this block.

One or more empty lines.

3.2.5 PHYLIP File format

The two distinct parts of the PHYLIP format are the multiple sequence alignment and a header that describes the alignment's dimensions. PHYLIP is a plain text format.

The multiple sequence alignment is stored in the PHYLIP file format. Since Joe Felsenstein's PHYLIP software first defined and employed the format, several additional bioinformatics tools have supported it.

At the head of the file is a list of all the sequences and nucleotides or amino acids that make up the alignment.

The .phy extension is used to save files in this format.

Section Header:

The alignment's dimensions are described on a single line that serves as the header. It must be the first line in the file.

Two positive numbers (n and m) separated by one or more spaces comprise the header, which is followed by optional spaces.

The number of rows, or sequences, in the alignment is indicated by the first integer (n). The length of the sequences, or the number of columns, in the alignment is indicated by the second integer (m).

1x1 is the lowest alignment dimension that is supported.

Section of Alignment:

The header is immediately followed by the alignment section. It is made up of n lines, or rows, one for each alignment sequence.

A sequence identifier (ID) and characters from the sequence are included in each row in a fixed width format.

Ten characters is the maximum length for the sequence ID. This restriction could be loosened by other bioinformatics tools to accommodate longer sequence IDs.

To meet the 10-character fixed width, IDs with less characters than 10 must have spaces added to them.

All characters, including spaces, underscores, and numerals, are acceptable within an ID, with the exception of newlines.

The original format definition of PHYLIP format defines the IUPAC nucleic acid lexicon for DNA or RNA sequences and the IUPAC protein lexicon for protein sequences, however it does not expressly limit the range of supported characters that may be used to describe a sequence.

3.2.6 Swiss-Prot File Format

The goal of the curated protein sequence database SWISS-PROT is to offer a high degree of annotations (e.g., function description, domain structure, post-translational modifications, variants, etc.), low redundancy, and high degree of database integration.

Format for records

Fig 3.6 Swiss Prot file Example

Source: https://www.ercim.eu/EU-NSF/semweb/slides/goble/img014.jpg

Every entry in the sequence is made up of lines. The different data that make up the entry are recorded using several sorts of lines, each with its unique structure.

A two-character line code that denotes the kind of data on each line appears at the beginning of the line.

3.3 Sequence Annotation

The crucial step in identifying non-protein coding regions and unique genomic elements and then connecting pertinent biological information to these elements is known as sequence annotation.

Sequence annotation is the act of adding biological information to sequences.

Sequence annotations are used to define areas or places of interest within the protein sequence. These may include binding sites, enzyme active sites, post-translational modifications, local secondary structure, or other properties expected or documented in the given references.

This also describes sequence problems between references.

In an earlier iteration of the UniProtKB entry view, sequence annotations (position-specific annotations) were located in the 'Sequence annotation (Features)' section.

The different position-specific annotations are dispersed among the pertinent parts, and the current entry view shows annotation by subject (Function, PTM & processing, etc.).

In both the text and XML formats, all position-specific annotations are still compiled into a "feature table" (FT).

The emergence of automated genome annotation technologies also serves as a catalyst for improvements in sequence and functional annotation.

With the help of pre-existing patterns and annotations found in local or public databases, these programs are skilled in annotating fresh genomes.

By enhancing the precision and effectiveness of genome annotation, the incorporation of these automated techniques has been crucial in enabling researchers to examine and comprehend the genomic sequences with increased depth and precision.

Sequence Annotation Tools:

Apollo

Pros:

- User-friendly, graphical annotation editor.
- Allows collaborative annotation.
- Supports community curation.

Cons:

- Requires some setup and configuration.
- Not recommended for very big genomes.

MAKER

Pros:

Suitable for annotating newly sequenced genomes.

Incorporates a range of resources and tools for thorough annotation.

Cons:

Steeper learning curve for beginners.

Output may require additional filtering and refinement.

These instruments have been important in helping scientists decode protein and genomic sequences.

which has made it possible for them to efficiently annotate and infer the activities of diverse genetic components.

By making use of these resources, we may greatly improve our comprehension of genetics and find the fundamental ideas that underpin life.

3.4 Data Retrieval Systems

Working with flat files, the Sequence Retrieval System is a database system. Many bioinformatics tools are also included and can be used in conjunction with database searches.

The European Bioinformatics Institute (EBI) in Hinxton, UK, developed SRS, a standardized interface to over 80 biological datasets.

The following categories of databases are represented in them:

mapping, mutations,

locus-specific mutations,

application results (such BLAST),

transcription factors,

protein 3-D structure,

metabolic pathways

sequence and sequence-related databases.

You are able to browse through them and view their contents.

A link to a page describing each database, along with the date of its most recent update, may be found on the Web page that lists all of the databases.

Before entering your query, you choose which database or databases to look through. Currently available on the WWW are more than thirty versions of SRS.

A distinct collection of databases and related analytical tools are included in each.

SRS databases are properly indexed, which cuts down on search time even if there are a lot of possible databases to look through.

Every database's data fields are dissected into their constituent parts, and a subset of words is taken out and added to an index.

Generally speaking, every field has an index. You have the choice to enter search terms in a specific field on the query form or search all fields by selecting the "All text" option.

SRS has a different query form that makes it possible to create more intricate Boolean inquiries.

ENTREZ

On the National Center for Biotechnology Information (NCBI) website, users may search several distinct health sciences databases using the Entrez Global Query Cross-Database Search System, a federated search engine or web portal.

With a single query string and user interface, Entrez Global Query is an integrated search and retrieval system that offers simultaneous access to all databases.

Entrez has the ability to quickly extract linked structures, sequences, and references.

Views of chromosomal maps and gene and protein sequences are available through the Entrez system.

All of the databases that Entrez has indexed may be searched with a single query string that allows for the use of Boolean operators and search word tags to limit certain fields in the search statement.

This yields a unified results page with links to the actual search results for each database as well as the number of hits for the search in each database.

An interface such to this is offered by Entrez for both searching and honing in on certain databases. Through a web forms interface, the user may refine their search by using the Limits function.

Entrez search through the following repositories:

- PubMed Central: free, full-text journal articles

- Site Search: FTP and NCBI webpages
- Books: online books
- Genome: whole genome sequences and mapping
- Structure: three-dimensional macromolecular structures
- Taxonomy: organisms in GenBank Taxonomy
- Online Mendelian Inheritance in Man (OMIM)
- Nucleotide: sequence database (GenBank)
- Protein: sequence database (GenPept)
- PubMed: Citations and abstracts for biomedical literature, comprising Medline
- HomoloGene: eukaryotic homology groups
- PubChem Compound: distinct chemical structures of tiny molecules
- PubChem Substance: deposited chemical substance records

For more direct access to query results, NCBI offers the Entrez Programming Utilities [4] (eUtils), in addition to the search engine forms for querying the data in Entrez.

Posting specially constructed URLs to the NCBI server and analyzing the XML return is how you access the eUtils.

All of the key databases of NCBI, such as PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, OMIM, and many more, are searchable and retrievable using Entrez, a text-based system. Entrez serves as an indexing and retrieval system, a repository for information gathered from many sources, and a framework for structuring biological data. This chapter is centered around these broad ideas.

Summary

Objectives:

• Educate students about the sources and protocols for generating biological data.
• Explain different sequencing methods used in research.
• Describe and illustrate various biological data file formats and their applications.
• Define the DRS (Data Retrieval System) used for collecting large amounts of biological data from online repositories.

Generation of Data

Genomic Sequencing

The entire DNA or RNA of an organism is called its genome, containing all the information needed for its functioning.

Sequencing the genome can identify specific DNA changes causing disorders, offering flexibility to analyze the genetic composition of various organisms.

Genome sequencing has benefited the discovery of harmful organisms, disease outbreak tracking, and identifying medication resistance mutations.

Different sequencing methods include Sanger method, shotgun sequencing, and next-generation sequencing (NGS).

Protein Sequencing

Includes a mix of methodologies and techniques to determine the amino acid sequence of proteins as a part of post-translational modifications.

Mass Spectrometry

A powerful analytical method used to identify and quantify compounds, as well as understand the molecular structure and chemical characteristics of substances.

Microarray

- DNA microarrays, also known as DNA chips, are used to analyze the expression of tens of thousands of genes simultaneously, enabling a variety of applications from gene expression profiling to toxicological research.

Sequence Submission Tools

- BankIt: Submission tool for DNA or RNA sequences to the GenBank database at NCBI.
- Sequin: Stand-alone tool for updating GenBank, EMBL, and DDBJ databases with sequencing data.
- Webin: Web-based submission method for nucleotide sequences and biological annotations to EMBL-Bank.
- Sequence File Formats
- Flat File: A simple text file without organized relationships to other files or databases, often used for data import and export across different software programs or systems.
- FASTA: A text-based format used to represent nucleotide or peptide sequences, with sequences made up of lines following a single line description.

- PHYLIP: A plain text format that stores multiple sequence alignments, with a header describing the alignment's dimensions and the alignment section containing sequence identifiers and characters.
- Swiss-Prot: The curated protein sequence database format aimed at providing high annotations, low redundancy, and integration with other databases.

These objectives and explanations provide students with a comprehensive understanding of biological data generation sources, sequencing methods, data file formats, retrieval systems, and sequence submission tools. They cover various aspects of biological data and equip students with the knowledge necessary for their research and studies.

Self-Assessment:

1. What do you mean by data generation? Explain different sources of biological data.
2. List different types of file formats which are used to store biological sequential data.
3. Explain DRS. Which software is use in DRS?
4. Give detailed explanation of NGS (Next Generation Sequencing).

Unit -4

Basic concepts of sequence Alignment and applications of Bioinformatics

Objectives:

- To help in interpreting the homology model of protein structure by sequence similarity.
- To Develop a comman understanding of the means and meathod of sequence alignment.
- To explain the purpose of scoring matrix in the evolutionary relationship and common pattern between genes.
- To apply the structural bioinformatics in drug discovery methods.

4.1 Scoring Matrices:

- Each alignment method requires a system of points for both matches and mismatches. Each place in the sequence for a specific alignment is given a number based on the match at that position.
- The best alignment among possible alignments is picked by adding the points for each location in the alignment to get a total score.
- Transferring a number for a match and a different number for a mismatch is the most smiplest method of scoring. A unitary matrix is a common term used to describe such a matrix.
- It may be sufficient to use a unitary scoring matrix for nucleic acid alignments. Since the nature of the residue directly affects protein function and, consequently, survival risks, changes in amino acid sequences are often more revealing than changes in base sequences.
- It is more likely to identify a change from a valine to an isoleucine, for example, than from a valine to an aspartate.
- The alignment systems consequently use score matrices for amino acid sequences, which either directly or indirectly carry information about the likelihood of a particular alteration. However, a matrix scoring simply for identities will usually produce the same alignment for closely related proteins.

- Several substitutes for the unitary scoring matrix have been put forward. A score matrix based on the least amount of bases required to convert a codon for one amino acid into a codon for another amino acid was one of the first ideas.
- More distant links between protein sequences have been identified by this matrix—also referred to as the minimum mutation distance matrix—than with the unitary matrix method.
- Due to the addition of information regarding the process of producing mutations from one amino acid to another, the minimal mutation distance matrix represents an improvement. It does not, however, take into account the mechanisms of selection that decide which mutations in a population will survive.
- A score matrix based on specific physiochemical or structural feature that the 220 pairs of amino acids share and do not share is another enhancement. Certain applications of this strategy are effective for sequences that have undergone significant evolutionary conservation.
- Based onMcLachlan's simple scheme, amino acids were categorized according to their shape, size, polar or non-polar characterand charge. Interconversions between identical rare amino acids, such as F and F, were assigned a score of six, while substitutions between amino acids with quite different characters, such as E and F, were given a score of zero.
- Combining data from thephysical propertiesof amino acids and genetic coding is another strategy that is comparable to McLachlan's. However, this method has issues balancing the contributions of various features to the positive selection of mutations and fails to take into account the varying rates at which distinct mutations arise.
- By selecting a scoring matrix is not the only relevant factor—a measure of homology between amino acid pairings is also significant. It's also crucial that the method used and the model used to create a particular score matrix serve as the foundation for all further outcomes.
- The statistical theory from which the most utilized scoring matrices derive.
    - You should be focused on scoring matrices for the following reasons: 1) In every investigation requiring sequence comparison, scoring matrices are present.

2) The analysis's conclusion can be significantly impacted by the matrix selection.
3) A certain theory of evolution is implicitly represented by scoring matrices.
4) Making an informed decision can be facilitated by being aware of the theories behind a certain scoring matrix.

4.2 Types of scoring matrix

Two types of scoring matrices are:

4.2.1PAM (**P**ercent **A**ccepted **M**utations)

- Evolutionary distances explained the most significant improvement over the unitary matrix. In the 1970s, **Margaret Dayhoff** developed this method. She conducted a thorough investigation of the frequency at which amino acids evolved to substitute one another.
- This study includes building phylogenetic trees for each family of proteins after manually aligning every protein in a number of protein families. This resulted in a table showing the relative rates at which different amino acids are replaced during the course of evolution.
- This table was utilized to calculate the PAM (Point Accepted Mutation) family of scoring matrices, together with the relative frequency of occurrence of amino acids in the proteins under study.
- Amino acid mutations resulting from single base alterations will dominate in the PAM series as this is based on projected mutation rates (Percent Accepted Mutations) from closely related proteins. Since the numbers are proportional to the logarithm of the "odds" that the replacement won't be a random change, it is also known as a log-odds matrix. PAM 1 represents 1% all approved alterations. Extrapolations are used to produce PAM matrices for sequences that are less comparable. Since a residue may change more than once, two sequences with this amount of mutation will have roughly 50% identity. The PAM100 matrix equates to 100 approved mutations per 100 residues. The PAM250 matrix correlates to a level of roughly 20% identical residues in a similar manner.
- In theory, the PAM matrices are better than other alignment scoring techniques. Based on observed mutations, PAM matrices are constructed from a biological perspective. As

a result, they contain data on the mechanisms that lead to mutations as well as the critical parameters for selection and population-level mutation correction. This kind of grading matrix also offers an advantage from a statistical perspective.

- PAM matrices, which are arise from data, provide precise descriptions of the changes in amino acid composition that can be predicted with a specific number of mutations. As a result, evolution rather than chance is statistically more likely to have produced the highest scoring alignment. The PAM matrices, along with the other substitution matrices that will be covered later, are typically displayed as log-odds matrices.
- As the logarithm of an odds ratio represents each score in the matrix. Rest "X" is observed to replace residue "Y" a certain number of times, divided by the number of times residue "X" would be predicted to replace residue "Y" if the replacement happened at random. This ratio is known as the **odds ratio**. Positive matrix scores, therefore, indicate a pair of residues that happen to replace each other more frequently than would be predicted by chance. Pairs of residues with negative scores in the matrix indicate that they replace one another less frequently than would be predicted by chance and provide evidence that the sequences are not identical.
- Because the matrix summarizes the documented replacements that have occurred while maintaining the structural and functional qualities of proteins, it can be used to objectively choose groups of amino acids that indicate conservative substitutions in proteins. Finally, compared to the matrices previously stated, the PAM matrix is better suited for scoring an alignment since it offers an empirical, experimental determination of conserved replacement.

4.2.2BLOSUM

- The Dayhoff model is predicated on the idea that the rates of evolution are constant throughout the entirety of the protein sequence. Since there is a clear difference in the rates of evolution between conserved and non-conserved protein domains, it seems unlikely to be the case. A novel approach has been taken to this by Henikoff and Henikoff . Specifically for more distantly related proteins, their goal was to provide a better measure of the differences between two proteins.

- They looked for sequence variations simply within the highly conserved areas of a protein family using the BLOCKS database. As a result, the BLOSUM term stands for **BLO**cks Substitution Matrix. To create a frequency chart showing how frequently certain pairs of amino acids are found together in these conserved sections, they first gather all of the sequences in the BLOCKS database. For each sequence, they then add up the amount of amino acids in each location. These frequencies can be expressed as a frequency table, and the relatedness chances are computed using a methodology akin to that of the Dayhoff matrix.
- This approach includes different evolutionary distances through a clustering procedure: two sequences that are identical for more than a threshold of places are clustered. If a sequence is identical to another sequence at the same level already in the cluster, it is added to the cluster. This resulted in an array of matrices. The matrices are known as BLOSUM matrices, and each one has an index that indicates the degree of clustering. For instance, BLOSUM62 is made up of sequence blocks that have a 62% identity level of clustering.

4.3 Methods of Alignment

- A computational method for comparing and analyzing the similarities and differences of two or more biological data sequences, such as DNA, RNA, or protein sequences, is called sequence alignment. Sequence alignment allows scientists to identify functional elements, find conserved sections, find mutations, and deduce evolutionary relationships.
- Refining sequence arrangement to maximize matches and reduce mismatches and indels (insertions and deletions) is the goal. Multiple sequence alignment expands pairwise alignment to include three or more sequences. Pairwise alignment compares two sequences. In order to evaluate the most likely evolutionary link or functional similarity between the sequences, sequence alignment algorithms award scores or penalties. This method provides insights into the structure, function, and evolution of biological sequences and is essential in areas including drug development, forensic investigation, genomics, proteomicsand evolutionary biology.
- Some of the methods are discussed below.

4.3.1 Dot Matrix Method:

- Plotting two sequences in a 2D matrix allows for a graphical approach of sequence alignment called thedot plot method or dot matrix method.
- Two sequences that need to be compared are plotted along the horizontal and vertical axis of a matrix in a dot matrix. Next, each residue in one sequence is scanned by the technique to find similarities with every residue in the other sequence.
- A dot is placed in the appropriate spot in the matrix if a residue in one sequence matches a residue in the other sequence. If not, there is a blank in the matrix position.
- The dot plot along the main diagonal of the matrix appears as a single line if the two sequences under comparison are quite similar. The dot plot, on the other hand, will have more dispersed dots with fewer diagonal lines when the sequences are less similar, showing that the sequences share less similarity.
- Repeated items in a single sequence can also be identified via dot plots. Repeats are indicated by short parallel lines above and below the main diagonal.
- Comparing two sequences are shown in Figure 4.1

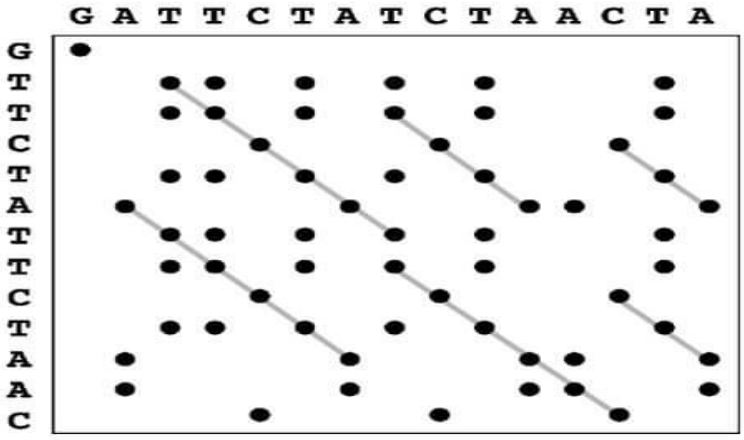

Figure 4.1: Example of comparing two sequences using dot plots. (Image resource Xiong, J., 2006).

4.3.2 Dynamic Programming Method:

- By comparing each potential pair of characters in the sequences, dynamic programming is utilized to determine the best alignment between two proteins or nucleic acid sequences.

- Both local and global alignments can be achieved through the use of dynamic programming. While dynamic programming in local pairwise alignment is based on the Smith-Waterman method, it is based on the Needleman-Wunsch algorithm in the global pairwise alignment technique.
- The following three stages are below how this method operates.
    1) **Initialization of the scoring matrix:** First, a two-dimensional matrix having the two sequences to be aligned written along the top and left sides is generated. This is the initialization of the scoring matrix. Gap penalties and a zero beginning score are added to the matrix at the upper-left corner.
    2) **Matrix filling:** The next stage is adding scores to the matrix based to a scoring matrix. Nucleotide sequence scoring matrices are easy to understand. For every match, a positive value is returned, and for every mismatch, a negative value. PAM and BLOSUM scoring matrices are utilized for amino acids.

       The method begins at the upper left corner of the matrix and moves one row at a time toward the lower right corner to calculate the alignment scores. The algorithm aligns the appropriate residues to fill each cell in the matrix with the maximum score allowed.
    3) **Traceback to determine best alignment**: The algorithm tracks back to discover the ideal alignment path once the matrix has been filled. Neighboring cells are inspected in reverse order, starting from the bottom-right corner and progressing towards the top-left corner, in order to identify the optimal path with the highest total score. The alignment path with the highest score is the best one.

4.3.3 BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool **):**

It first introduce by **Stephen Altschul et al.** in **1990**, this extensively utilized bioinformatics program evolved to become one of the most well-liked resources for sequence similarity search. Home page of BLAST looks as shown in figure 4.2

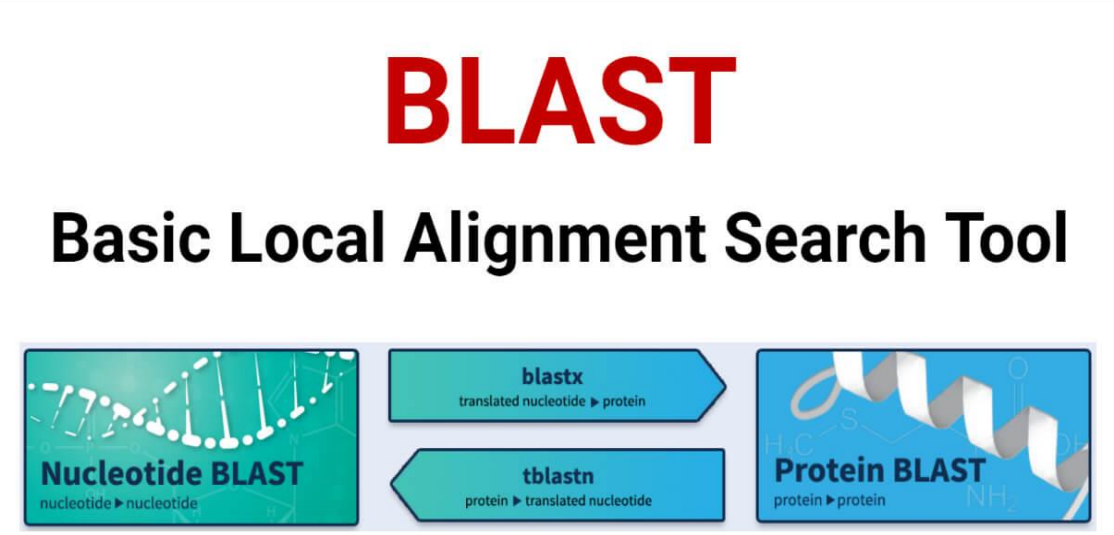

Figure 4.2 Basic Local Alignment Search Tool (BLAST). Image Source: NCBI.

A useful tool for examining biological sequencing data is BLAST. To increase its speed and accuracy, BLAST has undergone constant changes since its initial release in 1990. These days, BLAST is regarded as an essential and often utilized technique in the bioinformatics community. It has been essential to many research projects and cleared the path for the creation of other sequence comparison tools.

4.3.3.1 Types of BLAST:

Based on the type of sequence (DNA or protein) in the query and database sequences, there are five main types of BLAST are as follows:

1) **BLASTP** A protein sequence database and a protein query sequence are compared.

2) **BLASTN** A nucleotide query sequence is compared to a nucleotide sequence database.

3) **BLASTX** By translating a nucleotide query sequence into one of its six potential reading frames and matching them with protein sequences, it compares nucleotide query sequences to protein sequence databases.

4) **TBLASTX** By translating the query sequence in each of the six reading frames and aligning it with the nucleotide sequences, it compares a nucleotide query sequence to a nucleotide sequence database.

5) **TBLASTN** By translating and aligning the nucleotide sequences in each of the six reading frames with the protein sequence, it compares a protein query sequence to a nucleotide sequence database.

4.3.4 FASTA

- Among of the earliest frequently used database similarity search techniques was FASTA. FASTA, often known as FastA, is a sequence alignment program that compares input nucleotide or protein sequences to databases that already exist. It is a short form for "Fast-All." Since its initial development in **1985** by **David J. Lipman** and **William R. Pearson**, it has undergone numerous refinements and adaptations for a range of uses.
- Starting from the FASTA program, the text-based file format for representing nucleotide or protein sequences has now become the industry standard in bioinformatics. The FASTA file format is also used by many additional sequence database search tools.
- FASTA homepage looks like shown in firgure 4.3

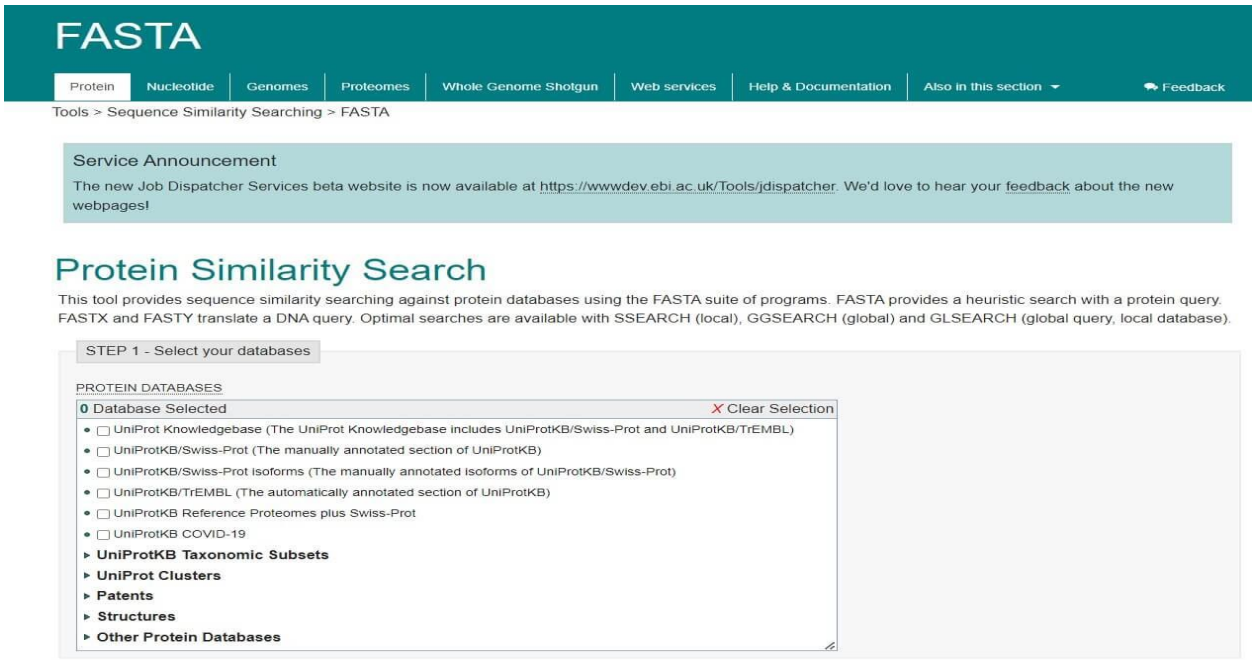

Figure4.3: FASTA Web Interface. Image Source: EMBL.

4.4 Sequence Alignment:

Among the most important steps in comparing biological sequences is thought to be sequence alignment. In order to find similarities between two or more nucleotide or amino acid sequences, sequence alignment arranges the sequences. Understanding the functional, structural, and evolutionary links between the sequences is made easier by looking at these areas of commonality.

Local alignment and Global alignment are two frequently utilized sequence alignment techniques.

Local alignment: Instead of attempting to align the entire length of the sequences, only the regions with the highest density of matches are aligned. This is useful for identifying short conserved regions in protein or nucleotide sequences.

Global alignment: By optimizing the total similarity between two sequences, global alignment aligns them along their whole length. Sequences with the same length are compared using this method. (Both are shown in figure 4.4)

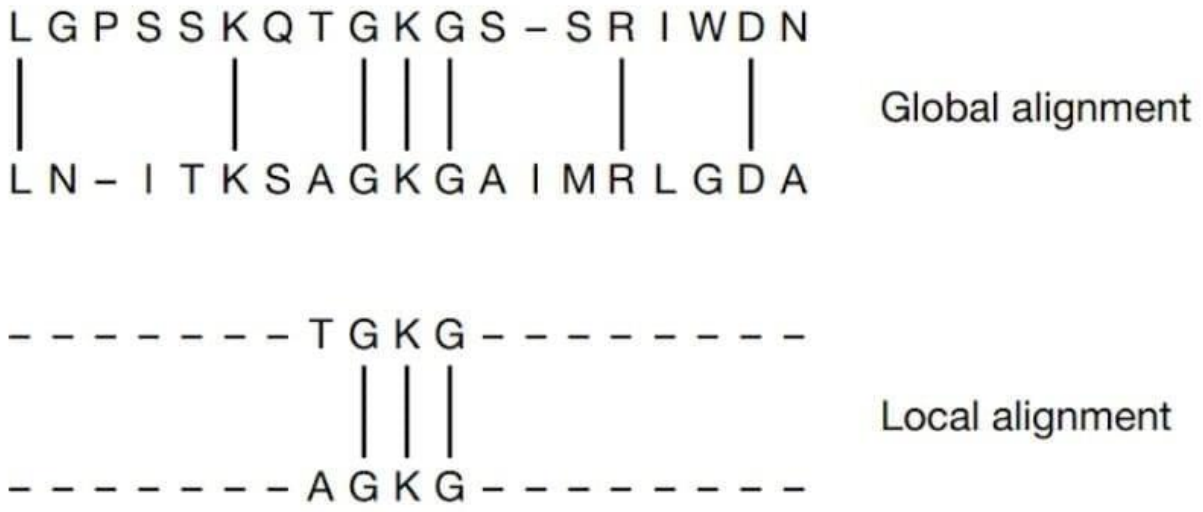

Figure4.4: Global and Local Alignment of two sequences (Image resource Mount, D. W., 2001).

4.4.1 Types of Sequence Alignment:

There are mainly two types of sequence alignment:

1) **Pairwise Alignment:**
   - Sequence alignment involves aligning two sequences to determine their ideal pairing is known as pairwise sequence alignment.
   - It depend on a scoring system that gives points for matching characters and deductions for gaps or mismatched characters.
   - The main goal of pairwise sequence alignment is to get the greatest score—a measure of how similar the two sequences are—possibly.
2) **Multiple Sequence Alignment:**
   - Various Sequence To attain ideal sequence matching, alignment entails aligning three or more biological sequences.

- For us to build phylogenetic trees and discover conserved sequence areas, which aid in our understanding of the functional and evolutionary links between various species or groups of animals, multiple sequence alignments are required.

4.5 Sequence Similarity :

- A moderate substitution is defined as the mutation of one amino acid to a residue that is like and preserves the physiochemical properties.Therefore, the total of both identical and similar matches (residues that have undergone conservative replacement) determines the percent similarity of two sequences. The criteria used to compare two amino acid residues to one another determine how similar the measurements are.
- Positives are another term for similarity on a BLAST search.

4.6 Sequence Identity :

- The total number of characters that precisely match between two distinct sequences is known as sequence identity.
- Once the alignment was obtained using whatever means, there are numerous approaches to determining the percentage identity (PID). Divide the total number of identities, for instance, by:

  1)Alignment length.

  2)Quantity of non-gap jobs.

  3)The shortest sequence's length.

  4)The sequence's arithmetic mean length.

  5)Length (alignment): columns that are aligned, even those with gaps in either series

  6)Number of places that are equivalent, omitting overhangs.

4.7 Homology of Sequence:

- Using the known structure of a homologous protein as a guide, homology modeling—also referred to as comparison modeling—predicts the 3D structure of an unknown protein.

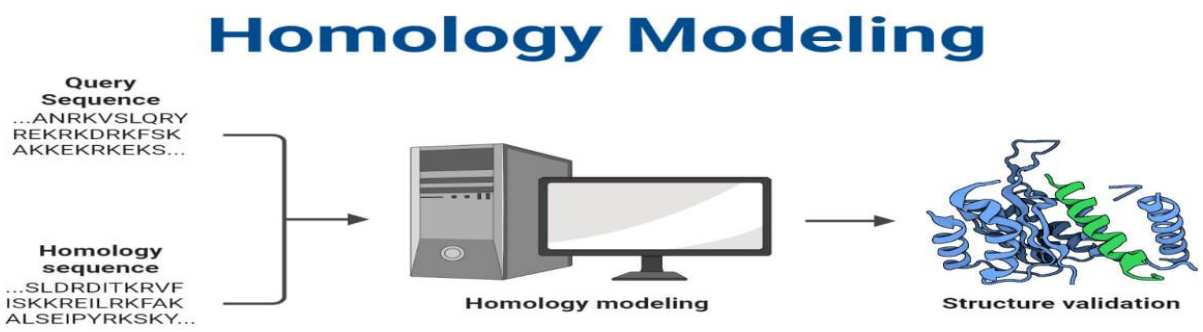

Figure 4.5 Homology Modeling(Image source Microbes Note)

- The technique is based on the observation that proteins' 3D structures are frequently more conserved than their amino acid sequences. As a result, proteins with comparable sequences are likely similar structurally. (Shown in figure 4.5)
- As there are less experimentally confirmed structures than there are protein sequences available, this approach is becoming more and more significant. The importance of homology modeling in bridging the gap between the quantity of protein sequences and experimentally established structures is becoming more and more significant.

4.8 Structural Bioinformatics

An interdisciplinary area known as structural bioinformatics uses concepts from computer science, biology, biochemistry, and mathematics to research and analyze the three-dimensional (3D) structures of biological macromolecules.

Finding the connection between the structure, function, and dynamics of biomolecules is the focus of the field.

Protein structures must be predicted, modeled, analyzed, and compared using computational techniques and algorithms.

Structural bioinformatics use several techniques, including molecular modeling, molecular dynamics simulations, and structure prediction tools, to gain understanding of the atomic-level characteristics of biomolecular structures.

Utilizing the abundance of high resolution, experimentally determined 3D structures that are accessible from public databases.

Structural bioinformatics can either directly study structures or provide model structures for molecules or their complexes that lack experimental structures.

These methods are also used in biological and pharmaceutical contexts, and we may drastically cut down on the time and expense associated with drug discovery and development campaigns by utilizing computational tools and structural bioinformatics.

A few approaches and their potential applications include:

Analysis of protein structures is made possible via structural analysis, which may also be used to forecast the consequences of mutations, for example, in protein engineering.

Docking enables the prediction and study of molecular interactions by providing 3D structures of complexes including proteins and their binding partners, such as peptides, small molecules, proteins, glycans, lipids, and nucleic acids.

Through the computation of binding free energies, which helps anticipate the strength of binding interactions between complexes, such as protein-ligand complexes, molecular dynamics simulation provides insight into the mobility of molecules and their complexes as well as the stability of these.

Virtual ligand screening makes it possible to identify molecules that behave as activators or inhibitors.

Ligand optimization is the process of making desired-property compounds better. This phase involves analysis using DSC or ITC or SPR for direct binding evaluation. The platform provides the biophysical techniques and expertise.

The following are the main objectives of structural bioinformatics:

1) Structure Prediction: When experimental structure identification is challenging or impossible, computational methods are used to predict the three-dimensional structure of biomolecules.

This includes techniques like fold identification, homology modeling, and ab initio modeling.

The process of figuring out a biomolecule's three-dimensional structure involves analyzing experimental data with methods including cryo-electron microscopy (cryo-EM), nuclear magnetic resonance (NMR) spectroscopy, and X-ray crystallography. This process includes data processing, refinement, and validation.

Structure analysis is the study and examination of the relationships, features, and structural aspects of biomolecules.

This procedure includes the identification of binding interfaces, structural motifs, functional domains, and active sites.

Investigations of interactions between proteins and ligands, proteins, and nucleic acids are also included.

determining a biomolecule's biological significance and function by examining its structural characteristics, such as ligand binding sites, active site residues, and structural resemblances to other proteins with established functions.

2) Drug Design and Discovery: The process of designing and optimizing tiny molecules for usage as possible drugs by using structural information. This involves the use of molecular dynamics, docking, and virtual screening to predict and examine interactions between drugs and target proteins.

Structural bioinformatics is essential for developing new drugs, understanding the molecular principles behind biological processes, and providing individualized treatment.

It clarifies interactions between proteins, the relationship between protein structures and functions, and the impact of genetic variations on protein architectures.

Structural bioinformatics integrates computational methods with experimental data to make advancements in biology, biochemistry, and medicine.

3) Ab Initio (De Novo) Prediction: Using no templates or known structures, the goal of ab initio approaches is to predict protein structures right away.

By using statistical potentials, energy functions, and physical principles, these methods explore a protein's conformational space and provide predictions about its three-dimensional structure.

Ab initio prediction is challenging because of the vast number of potential conformations and the computational complexity required.

In order to effectively explore the conformational environment, it usually makes use of sampling techniques and simplified representations of protein structures.

Methods Based on Templates and Fold Recognition: Fold recognition algorithms seek to detect proteins that, even in the lack of substantial sequence similarity, have folds that are similar to the target protein.

After finding an appropriate template, the target sequence is aligned with the template's structure, and the coordinates are sent to create a three-dimensional (3D) model of the target protein.

4) Hybrid Methods: To improve the accuracy and scope of protein structure prediction, hybrid methods combine many strategies, such as homology modeling and ab initio procedures.
These techniques provide more dependable models by utilizing physical principles, evolutionary knowledge, and experimental data.

5) Model Validation and Refinement: After a preliminary protein structure model is created, methods for improving the model's accuracy and quality are used.
This might involve molecular dynamics simulations, energy minimization, or optimization methods to improve atomic coordinates and remove steric conflicts.
The quality of the projected model is assessed using validation methods including Ramachandran plots, which show the sterically permitted portions of the protein backbone, and different statistical potentials, which gauge the model's dependability and quality.

4.8.1 PDB

Atomic coordinates and other data pertaining to proteins and other significant biological macromolecules are stored in the PDB collection.
Structural biologists employ techniques like cryo-electron microscopy, NMR spectroscopy, and X-ray crystallography to ascertain each atom's position in relation to the others inside the molecule.
They then deposit this data, which the wwPDB then annotates and makes publicly available in the archive.
The research taking on in labs all around the world is reflected in the ever expanding PDB.
This may make using the database for study and teaching both thrilling and difficult.
Many of the proteins and nucleic acids involved in essential biological processes have structures accessible, therefore one may search the PDB library for structures of ribosomes, oncogenes, pharmacological targets, and even whole viruses.
But since the PDB archives so many distinct structures, it might be difficult to retrieve the information you need.
For a particular molecule, you will frequently encounter incomplete structures, numerous structures, or structures that have been altered or rendered inactive from their original form.

PDB Information:

- The PDB archive's coordinate files for biological molecules contain the majority of the data. Each protein's atoms are listed in these files together with their three-dimensional (3D) spatial position.
- These files are offered in PDB, mmCIF, and XML formats.
- A typical PDB formatted file consists of a lengthy "header" portion of text that lists the sequence and a long list of the atoms and their coordinates after summarizing the protein, citation information, and the specifics of the structural solution.
- The experimental data that are utilized to calculate these atomic coordinates are also included in the archive.

Visualizing Structures:

- Although PDB files may be viewed directly in a text editor, utilizing a browser or visualization software to see them is frequently the most beneficial option.
- Users may search and browse the data under the PDB heading, including details on experimental procedures and the chemistry and biology of the protein, using online resources like those found on the RCSB PDB website.
- After locating the PDB entries that catch the spotlight, users may utilize visualization software to access the PDB file, see the protein structure on your desktop, and make personalized images of it.
- Additionally, these applications frequently come with analytical capabilities that let user find intriguing structural characteristics and analyze bond angles and distances.

4.9 Functional Genomics

- The study of how genes and intergenic areas of the genome affect various biological processes is known as functional genomics. (Shown in figure 4.6)
- In order to reduce the number of genes or areas under study to a list of potential genes or regions that require further analysis, researchers in this discipline often investigate genes or regions on a "genome-wide" scale, meaning that all or multiple genes/regions are studied simultaneously.

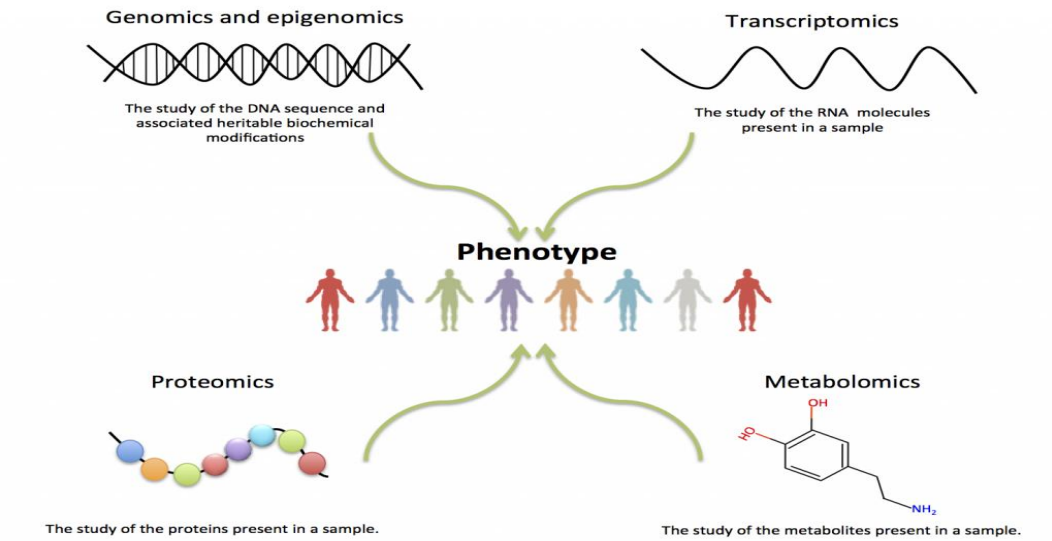

Figure 4.6 Functional genomics is the study of how the genome, transcripts (genes), proteins and metabolites work together to to produce a particular phenotype.

Image Source: (https://www.ebi.ac.uk/training/online/courses/functional-genomics-i-introduction-and-design/wp-content/uploads/sites/60/2020/05/Figure01-1024x747.png)

- The study of how genes and intergenic areas of the genome affect various biological processes is known as functional genomics.
- In order to reduce the number of genes or areas under study to a list of potential genes or regions that require further analysis, researchers in this discipline often investigate genes or regions on a "genome-wide" scale, meaning that all or multiple genes/regions are studied simultaneously.
- Functional genomics studies are distinguished by their genome-wide approach to these concerns, which often use high throughput technology instead of a more conventional gene by gene method.
- Understanding the function of genes, proteins, and eventually all of a genome's components is the aim of functional genomics.
- Functional genomics encompasses a wide range of scientific methods for studying the genes and proteins of an organism, including the analysis of the "biochemical, cellular, and/or physiological properties of each and every gene product."
- Studies of natural genetic variation spanning geography (e.g., an organism's body regions) and time (e.g., an organism's development) as well as functional disruptions (e.g., mutations) are also included in the field of genomics.

4.10 Drug Discovery Method:

Creating and discovering novel chemical compounds that can be utilized as medications to treat illnesses is the difficult and expensive process of drug design and development. Conventional drug design techniques need a lot of work and time. Contemporary drug development techniques, like in silico methods, have surfaced as a means of mitigating the shortcomings of conventional procedures by reducing the expenses, duration, and workforce involved in the drug discovery process.

The phrase "In Silico" describes the analysis and investigation of biological systems through computer-based studies.

Described as Computer-Aided Drug Design (CADD) or in silico drug design, this process uses bioinformatics tools and computational methods to find compounds that resemble drugs. Potential therapeutic candidates' biological activity is analyzed and predicted using in silico techniques, which also forecast their physicochemical characteristics.(Shown in figure 4.7)

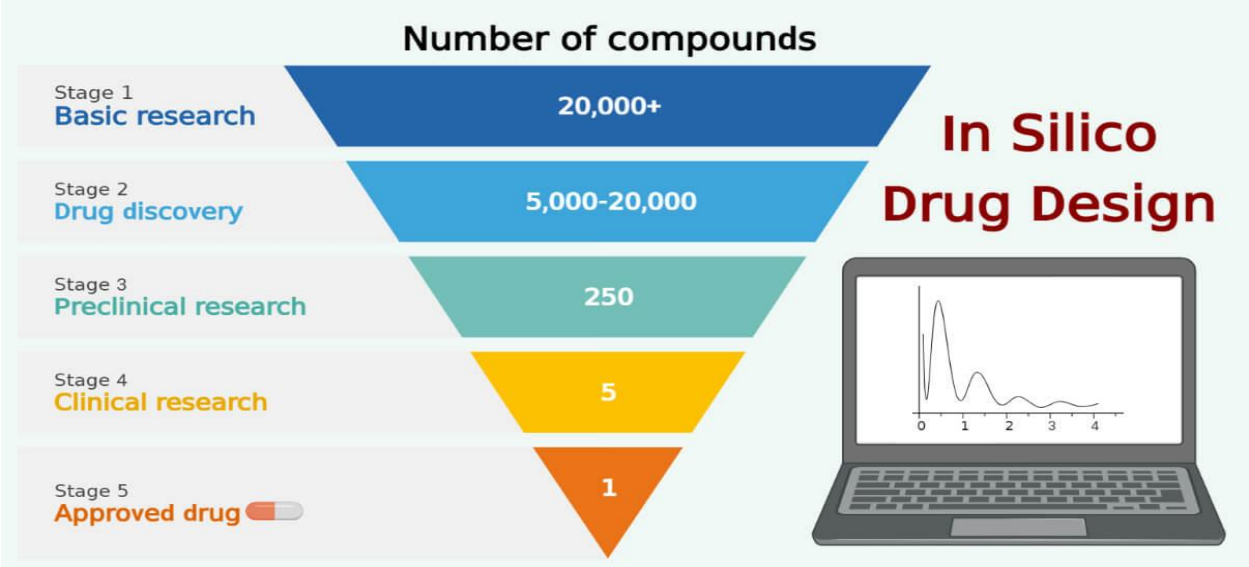

Figure 4.7 In Silico Drug Design (Image Resource : Microbe Notes)

4.10.1 Methods if In Silico Drug Design:

Different in silico methods for drug design have been created. Two distinct approaches are used: Structure-based drug design (SBDD) and Ligand-based drug design (LBDD)

In cases when the structure of the protein is unknown, LBDD might be the method of choice since it can be utilized to create molecules with characteristics akin to those of known ligands that attach to the protein.The availability of structural information about the target protein determines which approach is best. SBDD may be the better method if the protein's three-dimensional structure is known since it makes it possible to create compounds that exactly fit into the binding site of the protein.

4.10.1.1 Ligand based drug design:

- Using the known structure and functionalities of a ligand that binds to a target molecule, LBDD is a technique for creating novel medications. In order to build molecules with similar structural features for interaction with the target molecule the ligand serves as a starting point.
- There are various LBDD methodologies available; the most widely utilized ones are pharmacophore modeling, similarity searches, and quantitative structure-activity relationship (QSAR).

4.10.1.2 Structure based drug design:

- Using the 3D structure of a target protein related to a particular disease as a basis, SBDD is a drug design technique that helps researchers create new medications.
- A variety of steps are involved in the process, such as identifying a figuring out the target protein's three-dimensional structure therapeutic target and active ligands, docking small molecules into the target protein's binding cavity, synthesizing and optimizing the most promising hits, and assessing the compounds' biological characteristics.

Summary

Scoring matrices are essential in sequence comparison research, as they provide points for matches and mismatches, and the best alignment is determined by adding points for each location in the sequence. Unitary matrices are commonly used for nucleic acid alignments, as changes in amino acid sequences are more revealing than changes in base sequences. Several substitutes have been proposed, including the minimum mutation distance matrix, which

identifies more distant links between protein sequences, and the minimal mutation distance matrix, which considers the mechanisms of selection that determine which mutations will survive. Other strategies include McLachlan's simple scheme, which categorizes amino acids based on shape, size, polar or non-polar character, and charge. Combining data from physical properties and genetic coding is another strategy, but it struggles to balance the contributions of various features to the positive selection of mutations. The method and model used to create a particular score matrix serve as the foundation for all subsequent outcomes. Understanding the

statistical theory behind a scoring matrix is crucial for making informed decisions.

Margaret Dayhoff developed the Point Accepted Mutation (PAM) method in the 1970s to study the frequency of amino acid substitutions. She built phylogenetic trees for each protein family and manually aligned them, resulting in a table showing the relative rates of amino acid replacements. This table was used to calculate the PAM family of scoring matrices, which are based on projected mutation rates from closely related proteins. PAM matrices are better than other alignment scoring techniques as they contain data on mutation mechanisms and critical parameters for selection and population-level mutation correction. They provide precise descriptions of changes in amino acid composition that can be predicted with a specific number of mutations, making evolution more likely to produce the highest scoring alignment. The PAM matrix is better suited for scoring an alignment since it offers an empirical, experimental determination of conserved replacement. However, the Dayhoff model is predicated on the idea that the rates of evolution are constant throughout the protein sequence.

Self-Assessment:

1. What are scoring matrices.
2. Differentiate between PAM and BLOSUM.
3. Write down the methods alignments?
4. Explain the sequence alignment and its types.

GLOSSARY

Algorithm- A step-by-step procedure or formula for solving a problem or performing a task. In computational biology, algorithms are used to analyze biological data and model biological systems.

Alignment- The process of arranging sequences of DNA, RNA, or protein to identify regions of similarity. This is crucial for identifying functional, structural, or evolutionary relationships between the sequences.

Bioinformatics-The application of computational techniques to store, analyze, and interpret biological data. It is a key component of computational biology.

Clustering-A method of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Used in various types of biological data analysis, such as gene expression profiling.

Computational Genomics-A field of computational biology that focuses on the study of genomes using computational and statistical approaches to understand the structure, function, and evolution of genes and genomes.

Data Mining- The process of discovering patterns and knowledge from large amounts of data. In computational biology, data mining techniques are used to extract meaningful information from biological datasets.

Gene Expression- The process by which information from a gene is used to synthesize functional gene products, usually proteins.

High-Throughput Sequencing- Advanced sequencing technologies that allow for the rapid sequencing of large amounts of DNA or RNA.

Machine Learning- A subset of artificial intelligence that involves the use of algorithms and statistical models to enable computers to improve their performance on a specific task through experience.

Metagenomics- The study of genetic material recovered directly from environmental samples. Computational methods are essential for analyzing metagenomic data to understand the diversity and function of microbial communities.

Omics- A collective term for fields of study in biology ending in -omics, such as genomics, proteomics, transcriptomics, and metabolomics.

Phylogenetics- The study of the evolutionary history and relationships among individuals or groups of organisms.

Protein Structure Prediction- The prediction of the three-dimensional structure of a protein from its amino acid sequence.

Sequence Analysis- The study of the sequences of DNA, RNA, or proteins to understand their structure, function, and evolution

Single-Cell Analysis- A technique that allows the study of individual cells to understand cellular heterogeneity within tissues.

Systems Biology- An interdisciplinary field that focuses on complex interactions within biological systems, aiming to understand how these interactions give rise to the function and behavior of the system.

Transcriptomics- The study of the transcriptome, the complete set of RNA transcripts produced by the genome under specific circumstances or in a specific cell.

REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). "Basic Local Alignment Search Tool." Journal of Molecular Biology, 215(3), 403-410. doi:10.1016/S0022-2836(05)80360-2

- Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.

- Lander, E. S., Linton, L. M., Birren, B., et al. (2001). "Initial sequencing and analysis of the human genome." Nature, 409, 860-921. doi:10.1038/35057062

- Trapnell, C., Williams, B. A., Pertea, G., et al. (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nature Biotechnology, 28, 511-515. doi:10.1038/nbt.1621

- Eddy, S. R. (2011). "Accelerated Profile HMM Searches." PLoS Computational Biology, 7(10), e1002195. doi:10.1371/journal.pcbi.1002195

- Kanehisa, M., & Goto, S. (2000). "KEGG: Kyoto Encyclopedia of Genes and Genomes." Nucleic Acids Research, 28(1), 27-30. doi:10.1093/nar/28.1.27

- Consortium, T. U. (2018). "UniProt: a worldwide hub of protein knowledge." Nucleic Acids Research, 47(D1), D506-D515. doi:10.1093/nar/gky1049

- Van der Maaten, L., & Hinton, G. (2008). "Visualizing Data using t-SNE." Journal of Machine Learning Research, 9, 2579-2605.

- Shendure, J., & Ji, H. (2008). "Next-generation DNA sequencing." Nature Biotechnology, 26, 1135-1145. doi:10.1038/nbt1486

- MacArthur, D. G., Manolio, T. A., Dimmock, D. P., et al. (2014). "Guidelines for investigating causality of sequence variants in human disease." Nature, 508, 469-476. doi:10.1038/nature13127

- Koonin, E. V., & Galperin, M. Y. (2002). Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Kluwer Academic.

- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2015). "KEGG as a reference resource for gene and protein annotation." Nucleic Acids Research, 44(D1), D457-D462.
- doi:10.1093/nar/gkv1070

- Stamatakis, A. (2014). "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." Bioinformatics, 30(9), 1312-1313. doi:10.1093/bioinformatics/btu033

- Langmead, B., & Salzberg, S. L. (2012). "Fast gapped-read alignment with Bowtie 2." Nature Methods, 9, 357-359. doi:10.1038/nmeth.1923

- Krebs, J. E., Goldstein, E. S., & Kilpatrick, S. T. (2017). Lewin's GENES XII. Jones & Bartlett Learning.

- Benjamini, Y., & Hochberg, Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289-300.

- Stuart, T., Butler, A., Hoffman, P., et al. (2019). "Comprehensive Integration of Single-Cell Data." Cell, 177(7), 1888-1902.e21. doi:10.1016/j.cell.2019.05.031

- Szklarczyk, D., Gable, A. L., Lyon, D., et al. (2019). "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets." Nucleic Acids Research, 47(D1), D607-D613. doi:10.1093/nar/gky1131

- Wang, B., Mezlini, A. M., Demir, F., et al. (2014). "Similarity network fusion for aggregating data types on a genomic scale." Nature Methods, 11, 333-337. doi:10.1038/nmeth.2810

- Meyer, M., & Kircher, M. (2010). "Illumina sequencing library preparation for highly multiplexed target capture and sequencing." Cold Spring Harbor Protocols, 2010(6), pdb.prot5448.
- doi:10.1101/pdb.prot5448