MSCZO-603

M. Sc. III
Semester
BIOINFORMATICS AND BIOSTATISTICS

DEPARTMENT OF ZOOLOGY

SCHOOL OF SCIENCES

UTTARAKHAND OPEN UNIVERSITY

Bioinformatics and Biostatistics
(MSCZO-603)

DEPARTMENT OF ZOOLOGY
SCHOOL OF SCIENCES
UTTARAKHAND OPEN UNIVERSITY
Phone No. 05946-261122, 261123
Toll free No. 18001804025
Fax No. 05946-264232, E. mail info@uou.ac.in
htpp://uou.ac.in

Board of Studies and Programme Coordinator

PROGRAMME COORDINATOR

DR. PRAVESH KUMAR (ASSOCIATE PROFESSOR)
DEPARTMENT OF ZOOLOGY
SCHOOL OF SCIENCES, UTTARAKHAND OPEN UNIVERSITY
HALDWANI, NAINITAL, UTTARAKHAND

EDITOR
Dr. Shyam S. Kunjwal
Department of Zoology,
School of Sciences,
Uttarakhand Open University,
Haldwani, Nainital

Writer
Unit 1- Dr. Sunil Bhandari
Department of Zoology Govt P. G. College, Gopeshwar, Chamoli.

Unit 2,3 and 4-Dr. M. Faisal (Associate Professor)
School of Agriculture Forestry and Fisheries
Himgiri Zee University
Dehradun

Poornima Nailwal (Unit No.5, 6&7)
Assistant Professor, Department of Zoology, Uttarakhand Open University

Course Title and Code : **Bioinformatics and biostatistics (MSCZO 603)**

ISBN :

Copyright : **Uttarakhand Open University**

Edition : **2022**

Published by : **Uttarakhand Open University, Haldwani 263139**

Printed by: Shri Bagnath Paper Products, Ghatbagar Ward Near Kotwali, Bageshwar

CONTENTS

Course 1: Bioinformatics and Biostatistics

Course Code: MSC-ZO 603 **Credit: 3**

UNIT 1- BIOLOGICAL DATABASES

CONTENTS

1.1 INTRODUCTION

In biology, bioinformatics is defined as, "the use of computer to store, retrieve, analyze or predict the composition or structure of bio-molecules". Bioinformatics is the application of computational techniques and information technology to the organization and management of biological data. Classical bioinformatics deals primarily with sequence analysis. **Bioinformatics is an emerging branch of biological science that emerged as a result of the combination of biology and information technology.** It is a multidisciplinary subject where information technology is incorporated by means of various computational and analytical tools for the interpretation of biological data. In bioinformatics the term of biological databases are libraries of biological sciences, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures. Bioinformatics is subdivided into two sections, namely,

- Animal bioinformatics
- Plant bioinformatics

1.2 SCOPE AND APPLICATIONS OF BIOINFORMATICS

Bioinformatics and its application depend on taking out useful facts and figures from a collection of data reserved to be processed into useful information. Some examples of the application of bioinformatics are as follows:

1- Bioinformatics is largely used in gene therapy
2- This branch finds application in evolutionary concepts.
3- Microbial analysis and computing.
4- Understanding protein structure and modeling.
5- Storage and retrieval of biotechnological data.
6- In the finding of new drugs.

7- In agriculture to understand crop patterns, pest control, and crop management.

8- Management and analysis of a wide set of biological data.

9- It is specially used in human genome sequencing where large sets of data are being handled.

10- Bioinformatics plays a major role in the research and development of the biomedical field.

11- Bioinformatics uses computational coding for several applications that involve finding gene and protein functions and sequences, developing evolutionary relationships, and analyzing the three-dimensional shapes of proteins.

12- Research works based on genetic dieses and microbial disease entirely depend on bioinformatics, where the derived information can be vital to produce personalized medicines.

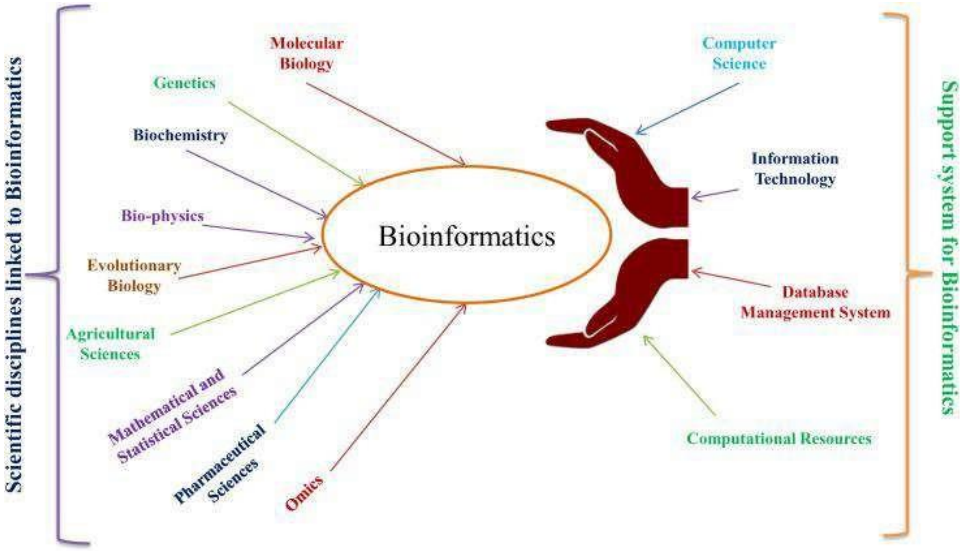


Fig.1.2 Scope of bioinformatics

Bioinformatics Subfields & Related Disciplines

The area of bioinformatics incorporates a wide range of biotechnological sub-disciplines that are highlighted by both scientific ethics based on biological sciences and deep knowledge of computer science and information technology. Bioinformatics will grow in scope and utility. Some of the examples of many fields of bioinformatics include:

a- **Computational biology:** The uses of data-based solutions to the issues in bioinformatics.

b- **Genetics:** It is the study of heredity and the gene diversity of inherited characteristics/features.

c- **Genomics:** It is the branch of bimolecular biology that works in the area of structure, function, evolution, and mapping of genomes.

d- **Proteomics:** The study of proteomes and their features.

e- **Metagenomics:** The study of genetics from the environment and living beings and samples.

f- **Transcriptomics:** It is the study of the complete **RNA** and **DNA** transcriptase.

g- **Phylogenetics:** The study of the relationships between groups of animals and humans.

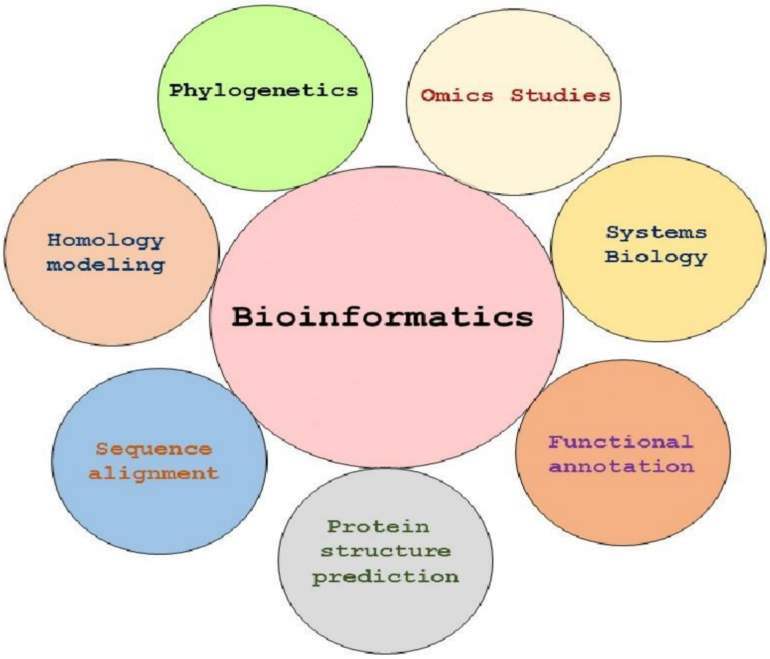
h- **Metabolomics:** The study of the **biochemistry** of metabolism and metabolites in living beings.

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

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

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

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



Biological Databases- Importance

1- One of the hallmarks of modern genomic research is the generation of enormous amounts of raw sequence data.

2- As the volume of genomic data grows, sophisticated computational methodologies are required to manage the data deluge.

3- Thus, the very first challenge in the genomics era is to store and handle the staggering volume of information through the establishment and use of computer databases.

4- A biological database is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system.

5- A simple database might be a single file containing many records, each of which includes the same set of information.

6- Databases act as a store house of information.

7- Databases are used to store and organize data in such a way that information can be retrieved easily via a variety of search criteria.

8- It allows knowledge discovery, which refers to the identification of connections between pieces of information that were not known when the information was first entered. This facilitates the discovery of new biological insights from raw data.

9- Secondary databases have become the molecular biologist's reference library over the past decade or so, providing a wealth of information on just about any gene or gene product that has been investigated by the research community.

10- It helps to solve cases where many users want to access the same entries of data.

11- Allows the indexing of data.

12- It helps to remove redundancy of data.

Example: A few popular databases are GenBank from NCBI (National Center for Biotechnology Information), SwissProt from the Swiss Institute of Bioinformatics and PIR from the Protein Information Resource.

1.3 PRIMARY, SECONDARY AND COMPOSITE DATABASE

Biological Databases are three types:

1- Primary database

a- Primary databases are also called as archival database.

b- They are populated with experimentally derived data such as nucleotide sequence, protein sequence or macromolecular structure.

c- Experimental results are submitted directly into the database by researchers, and the data are essentially archival in nature.

d- Once given a database accession number, the data in primary databases are never changed: they form part of the scientific record.

Examples-

1- ENA, GenBank and DDBJ (nucleotide sequence)

2- Array Express Archive and GEO (functional genomics data)

3- Protein Data Bank (PDB; coordinates of three-dimensional macromolecular structures)

2- Secondary database-

a- Secondary databases comprise data derived from the results of analyzing primary data.

b- Secondary databases often draw upon information from numerous sources, including other databases (primary and secondary), controlled vocabularies and the scientific literature.

c- They are highly curated, often using a complex combination of computational algorithms and manual analysis and interpretation to derive new knowledge from the public record of science.

Examples-

1- InterPro (protein families, motifs and domains)

2- UniProt Knowledgebase (sequence and functional information on proteins)

3- Ensembl (variation, function, regulation and more layered onto whole genome sequences)

Table 1- Essential aspects of primary and secondary databases

	Primary database	**Secondary database**
Synonyms	Archival database	Curated database; knowledgebase
Source of data	Direct submission of experimentally-derived data from researchers	Results of analysis, literature research and interpretation, often of data in primary databases
Examples	ENA, GenBank and DDBJ (nucleotide sequence) ArrayExpress and GEO (functional genomics data) Protein Data Bank (PDB; coordinates of three-dimensional macromolecular structures)	InterPro (protein families, motifs and domains) UniProt Knowledgebase (sequence and functional information on proteins) Ensembl (variation, function, regulation and more layered onto whole genome sequences)

3. Composite Databases:

1- The data entered in these types of databases are first compared and then filtered based on desired criteria.

2- The initial data are taken from the primary database, and then they are merged together based on certain conditions.

3- It helps in searching sequences rapidly. Composite Databases contain non-redundant data.

Examples –

Examples of Composite Databases are as follows.

a- Composite Databases - OWL,NRD and Swiss port +TREMBL

However, many data resources have both primary and secondary characteristics. For example, UniProt accepts primary sequences derived from peptide sequencing experiments. However, UniProt also infers peptide sequences from genomic information, and it provides a wealth of additional information, some derived from automated annotation (TrEMBL), and even more from careful manual analysis (SwissProt).

There are also specialized databases that cater to particular research interests. For example, Flybase, HIV sequence database, and Ribosomal Database Project are databases that specialize in a particular organism or a particular type of data.

1.3.1 NUCLEOTIDE SEQUENCES DATABASE

a- As biology has increasingly turned into a data-rich science, the need for storing and communicating large datasets has grown tremendously.

b- The obvious examples are the nucleotide sequences, the protein sequences, and the 3D structural data produced by X-ray crystallography and macromolecular NMR.

c- The biological information of nucleic acids is available as sequences while the data of proteins are available as sequences and structures. Sequences are represented in a single dimension whereas the structure contains the three-dimensional data of sequences.

d- A biological database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated.

e- The database is complemented with generalized software for processing, archiving, querying and distributing data.

f- Such databases consisting of nucleotide sequences are called nucleic acid sequence databases.

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

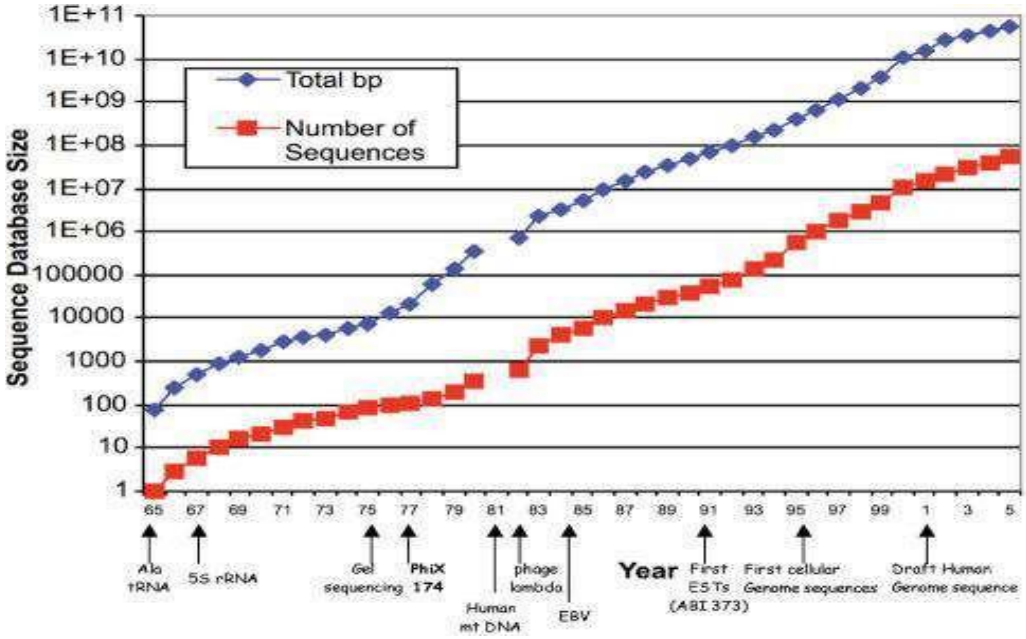


Fig-1.4Nucleotide sequence database

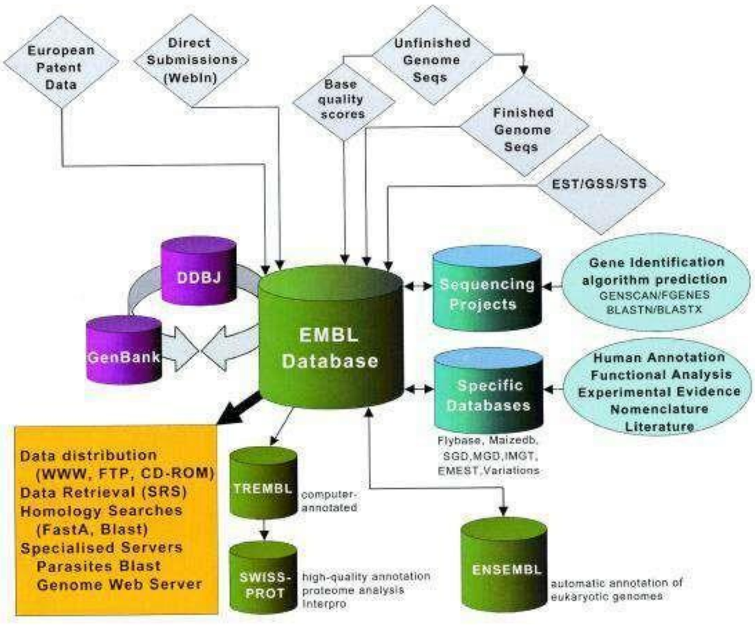

European Patent Data
Direct Submissions (WebIn)
Unfinished Genome Seqs
Base quality scores
Finished Genome Seqs
EST/GSS/STS
DDBJ
GenBank
EMBL Database
Sequencing Projects
Gene Identification algorithm prediction GENSCAN/FGENES BLASTN/BLASTX
Specific Databases
Human Annotation Functional Analysis Experimental Evidence Nomenclature Literature
Flybase, Maizedb, SGD,MGD,IMGT, EMEST,Variations
Data distribution (WWW, FTP, CD-ROM) Data Retrieval (SRS) Homology Searches (FastA, Blast) Specialised Servers Parasites Blast Genome Web Server
TREMBL
computer-annotated
SWISS-PROT
high-quality annotation proteome analysis Interpro
ENSEMBL
automatic annotation of eukaryotic genomes

Fig 1.5 Databases of nucleotide sequences

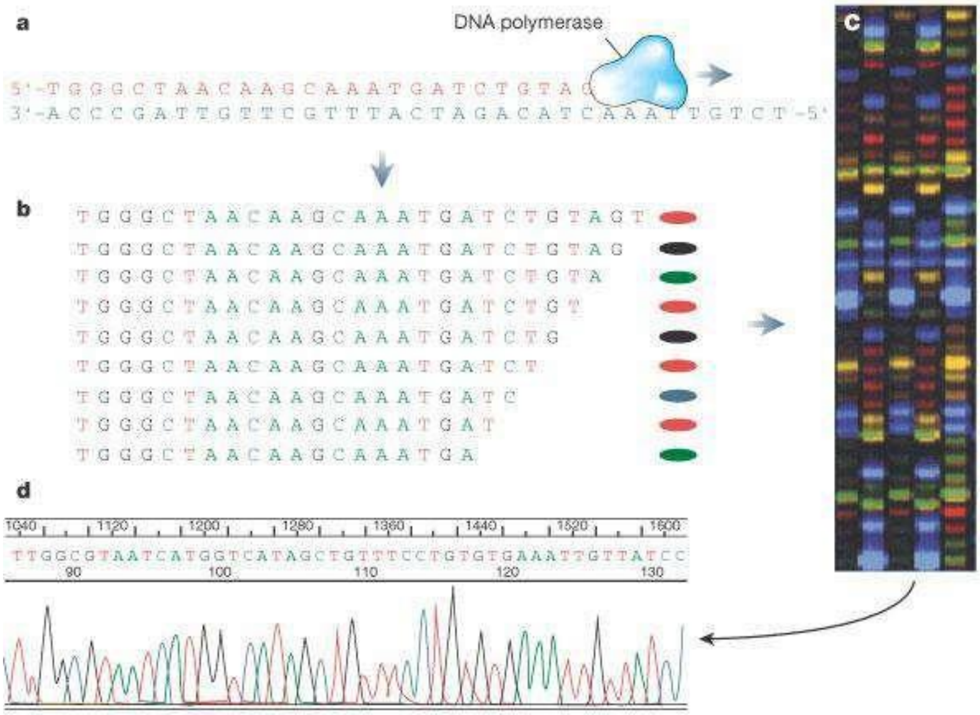

a
DNA polymerase
5'-TGGGCTAACAAGCAAATGATCTGTAG
3'-ACCCGATTGTTCGTTTACTAGACATCAAATTGTCT-5'
b
TGGGCTAACAAGCAAATGATCTGTAGT
TGGGCTAACAAGCAAATGATCTGTAG
TGGGCTAACAAGCAAATGATCTGTA
TGGGCTAACAAGCAAATGATCTGT
TGGGCTAACAAGCAAATGATCTG
TGGGCTAACAAGCAAATGATCT
TGGGCTAACAAGCAAATGATC
TGGGCTAACAAGCAAATGAT
TGGGCTAACAAGCAAATGA
c
d
1040
1120
1200
1280
1360
1440
1520
1600
TTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCC
90
100
110
120
130

Fig1.6-Human genome project-sequencing

1. Primary databases of nucleotide sequences

a- There are three chief databases that store and make available raw nucleic acid sequences to the public and researchers alike: **GenBank, EMBL, DDBJ**.

b- They are referred to as the primary nucleotide sequence databases since they are the repository of all nucleic acid sequences.

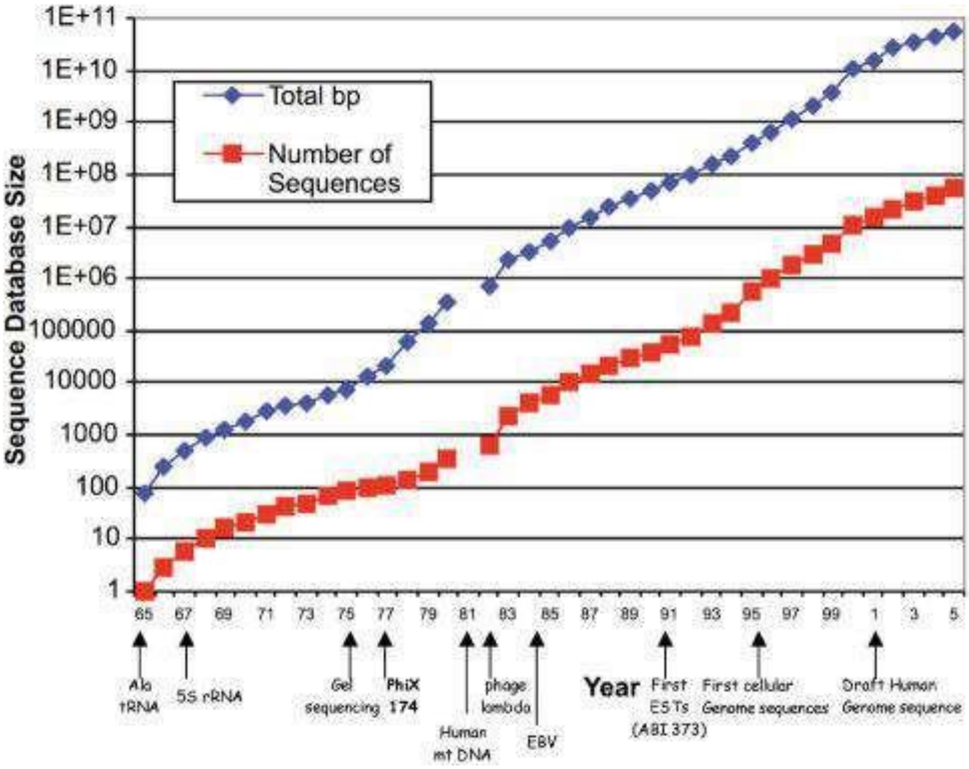
c- GenBank is physically located in the USA and is accessible through the NCBI portal over the intern.

d- EMBL (European Molecular Biology Laboratory) is in UK and DDJB (DNA databank of Japan) is in Japan.

e- All three accept nucleotide sequence submissions and then exchange new and updated data on a daily basis to achieve optimal synchronization between them.

f- These three databases are primary databases, as they house original sequence data.

g- They collaborate with Sequence Read Archive (SRA), which archives raw reads from high-throughput sequencing instruments.



a. GenBank

The GenBank sequence database is open access, annotated collection of all publicly available nucleotide sequences and their protein translations. This database is produced and maintained by the National Center for Biotechnology Information (NCBI) as part of the International Nucleotide Sequence Database Collaboration (INSDC). receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms. GenBank has become an important database for research in biological fields and has grown in recent years at an exponential rate by doubling roughly every 18 months.

b. EMBL (European Molecular Biology Laboratory)

The European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database is a comprehensive collection of primary nucleotide sequences maintained at the European Bioinformatics Institute (EBI). Data are received from genome sequencing centers, individual scientists and patent offices.

c. DDBJ (DNA databank of Japan)

It is located at the National Institute of Genetics (NIG) in the Shizuoka prefecture of Japan. It is the only nucleotide sequence data bank in Asia. Although DDBJ mainly receives its data from Japanese researchers, it can accept data from contributors from any other country.

2. Secondary databases of nucleotide sequences

a- Many of the secondary databases are simply sub-collection of sequences culled from one or the other of the primary databases such as GenBank or EMBL.

b- There is also usually a great deal of value addition in terms of annotation, software, presentation of the information and the cross-references.

c- There are other secondary databases that do not present sequences at all, but only information gathered from sequences databases.

a. Omniome Database:

Omniome Database is a comprehensive microbial resource maintained by TIGR (The Institute for Genomic Research). It has not only the sequence and annotation of each of the completed genomes, but also has associated information about the organisms (such as taxon and gram stain pattern), the structure and composition of their DNA molecules, and many other attributes of the protein sequences predicted from the DNA sequences.

It facilitates the meaningful multi-genome searches and analysis, for instance, alignment of entire genomes, and comparison of the physical proper of proteins and genes from different genomes etc.

b. FlyBase Database:

A consortium sequenced the entire genome of the fruit fly *D. melanogaster* to a high degree of completeness and quality.

c. ACeDB:

It is a repository of not only the sequence but also the genetic map as well as phenotypic information about the *C.* e*legans* nematode worm.

1.3.2 PROTEIN SEQUENCE DATABASE-

a- As biology has increasingly turned into a data-rich science, the need for storing and communicating large datasets has grown tremendously.

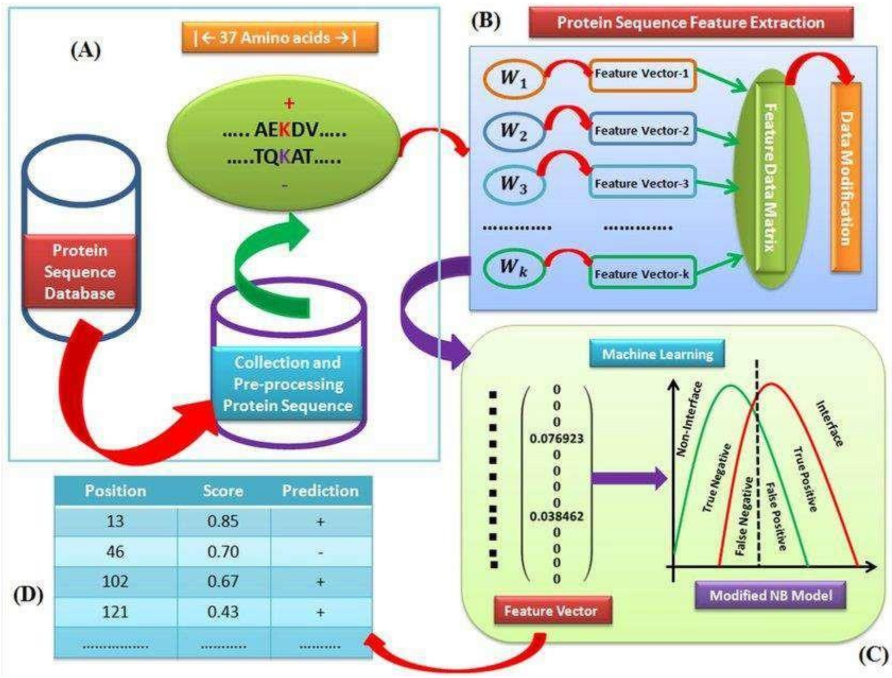
b- The obvious examples are the nucleotide sequences, the protein sequences, and the 3D structural data produced by X-ray crystallography and macromolecular NMR.

c- The biological information of proteins is available as sequences and structures. Sequences are represented in a single dimension whereas the structure contains the three-dimensional data of sequences.

d- A biological database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated.

e- A protein database is one or more datasets about proteins, which could include a protein's amino acid sequence, conformation, structure, and features such as active sites.

f- Protein databases are compiled by the translation of DNA sequences from different gene databases and include structural information. They are an important resource because proteins mediate most biological functions.



Position	Score	Prediction
13	0.85	+
46	0.70	-
102	0.67	+
121	0.43	+
....	....	....

Fig-1.8 (A) Protein sequence collection from protein database and preprocessing of the collected sequences, (B) Protein sequence feature extraction using statistical methods, (C) Fit the machine learning method modified naïve Bayes model, and (D) Protein-protein interaction sites prediction score.

Importance of Protein Databases

Huge amounts of data for protein structures, functions, and particularly sequences are being generated. Searching databases are often the first step in the study of a new protein. It has the following uses:

1. Comparison between proteins or between protein families provides information about the relationship between proteins within a genome or across different species and hence offers much more information that can be obtained by studying only an isolated protein.
2. Secondary databases derived from experimental databases are also widely available. These databases reorganize and annotate the data or provide predictions.
3. The use of multiple databases often helps researchers understand the structure and function of a protein.

Primary databases of Protein

The primary databases hold the experimentally determined protein sequences inferred from the conceptual translation of the nucleotide sequences. This, of course, is not experimentally derived information, but has arisen as a result of interpretation of the nucleotide sequence information and consequently must be treated as potentially containing misinterpreted information. There are a number of primary protein sequence databases and each requires some specific consideration.

a. Protein Information Resource (PIR) –Protein Sequence Database (PIR-PSD):

a- The PIR-PSD is a collaborative endeavor between the PIR, the MIPS (Munich Information Centre for Protein Sequences, Germany) and the JIPID (Japan International Protein Information Database, Japan).

b- The PIR-PSD is now a comprehensive, non-redundant, expertly annotated, object-relational DBMS.

c- A unique characteristic of the PIR-PSD is its classification of protein sequences based on the superfamily concept.

d- The sequence in PIR-PSD is also classified based on homology domain and sequence motifs.

e- Homology domains may correspond to evolutionary building blocks, while sequence motifs represent functional sites or conserved regions.

f- The classification approach allows a more complete understanding of sequence function-structure relationship.

b. SWISS-PROT

a- The other well known and extensively used protein database is SWISS-PROT. Like the PIR-PSD, this curated proteins sequence database also provides a high level of annotation.

b- The data in each entry can be considered separately as core data and annotation.

c- The core data consists of the sequences entered in common single letter amino acid code, and the related references and bibliography. The taxonomy of the organism from which the sequence was obtained also forms part of this core information.

d- The annotation contains information on the function or functions of the protein, post-translational modification such as phosphorylation, acetylation, etc., functional and structural domains and sites, such as calcium binding regions, ATP-binding sites, zinc fingers, etc., known secondary structural features as for examples alpha helix, beta sheet, etc., the quaternary structure of the protein, similarities to other protein if any, and diseases that may arise due to different authors publishing different sequences for the same protein, or due to mutations in different strains of an described as part of the annotation.

TrEMBL (for Translated EMBL) is a computer-annotated protein sequence database that is released as a supplement to SWISS-PROT. It contains the translation of all coding sequences present in the EMBL Nucleotide database, which have not been fully annotated. Thus it may contain the sequence of proteins that are never expressed and never actually identified in the organisms.

c. Protein Databank (PDB):

a- PDB is a primary protein structure database. It is a crystallographic database for the three-dimensional structure of large biological molecules, such as proteins.

b- In spite of the name, PDB archive the three-dimensional structures of not only proteins but also all biologically important molecules, such as nucleic acid fragments, RNA molecules, large peptides such as antibiotic gramicidin and complexes of protein and nucleic acids.

c- The database holds data derived from mainly three sources: Structure determined by X-ray crystallography, NMR experiments, and molecular modeling.

Secondary databases of Protein

The secondary databases are so termed because they contain the results of analysis of the sequences held in primary databases. Many secondary protein databases are the result of looking for features that relate different proteins. Some commonly used secondary databases of sequence and structure are as follows:

a. PROSITE:

a- A set of databases collects together patterns found in protein sequences rather than the complete sequences. PROSITE is one such pattern database.

b- The protein motif and pattern are encoded as "regular expressions".

c- The information corresponding to each entry in PROSITE is of the two forms – the patterns and the related descriptive text.

b. PRINTS:

- In the PRINTS database, the protein sequence patterns are stored as 'fingerprints'. A fingerprint is a set of motifs or patterns rather than a single one.

- The information contained in the PRINT entry may be divided into three sections. In addition to entry name, accession number and number of motifs, the first section contains cross-links to other databases that have more information about the characterized family.
- The second section provides a table showing how many of the motifs that make up the fingerprint occurs in the how many of the sequences in that family.
- The last section of the entry contains the actual fingerprints that are stored as multiple aligned sets of sequences; the alignment is made without gaps. There is, therefore, one set of aligned sequences for each motif.

c. MHCPep:

- MHCPep is a database comprising over 13000 peptide sequences known to bind the Major Histocompatibility Complex of the immune system.
- Each entry in the database contains not only the peptide sequence, which may be 8 to 10 amino acid long but in addition has information on the specific MHC molecules to which it binds, the experimental method used to assay the peptide, the degree of activity and the binding affinity observed , the source protein that, when broken down gave rise to this peptide along with other, the positions along the peptide where it anchors on the MHC molecules and references and cross-links to other information.

d. Pfam

a- Pfam contains the profiles used using Hidden Markov models.

b- HMMs build the model of the pattern as a series of the match, substitute, insert or delete states, with scores assigned for alignment to go from one state to another.

c- Each family or pattern defined in the Pfam consists of the four elements. The first is the annotation, which has the information on the source to make the entry, the method used and some numbers that serve as figures of merit.

d- The second is the seed alignment that is used to bootstrap the rest of the sequences into the multiple alignments and then the family.

e- The third is the HMM profile.

f- The fourth element is the complete alignment of all the sequences identified in that family.

1.3.3 GENE EXPRESSION DATABASE

The Gene Expression Database (GXD) is a community resource for gene expression information from the laboratory mouse. GXD stores and integrates different types of expression data and makes these data freely available in formats appropriate for comprehensive analysis.

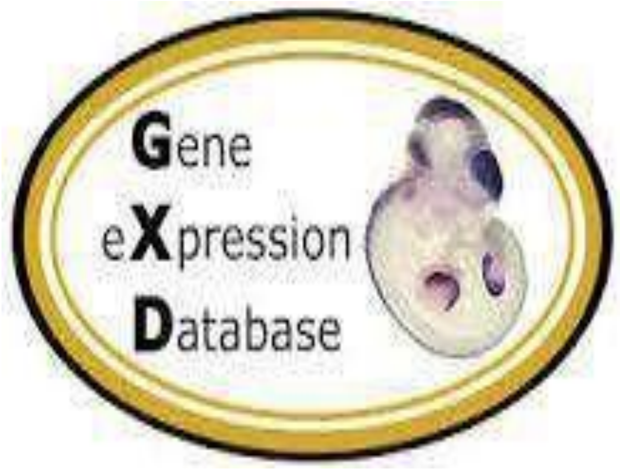


Fig.1.9 Gene expression database

Gene expression profiling is presented for developing and adult mammalian organs, tissues, anatomical compartments and cells, as well as for cultured stem, progenitor and primary cells, or cells derived via differentiation protocols. This allows for characterization of cells by their gene expression patterns. Gene expression profiles include annotations relating to developmental path-specific and enriched genes, selective gene markers in cells, and other genes whose expression has been reported in the scientific literature, high throughput experiments and public large scale datasets.

Gene expression data extracted from public large scale in situ hybridization and immunostaining databases are linked to the organs/tissues/compartments/cells

HIGH THROUGHPUT GENE EXPRESSION

DNA MICROARRAY

A DNA microarray is a collection of thousands of microscopic DNA spots attached to a solid surface. The number of genes attached depends on the array design, but generally covers all the expressed genes in the genome. RNA is extracted from cells of two populations under investigation, and reverse transcribed to cDNA. The cDNA is fluorescently labeled and applied to the microarray chip. After hybridization of the labeled cDNA to the probe, the microarray is scanned. The fluorescence intensities of the spots, which correspond to the level of gene expression in each population, are normalized and compared. The comparison is usually performed by calculating the normalized intensity fold change of one sample versus the other, with additional statistical analysis to exclude false-positive results.

RNA Sequencing

RNA sequencing (RNA-Seq) allows for quantitative determination of RNA expression levels. The method features an advantage over microarrays in that it provides coverage of the entire genome, including the various single-nucleotide polymorphisms (SNPs). In this method, RNA is extracted from cells, and the mRNA is isolated. In some cases, the mRNA is fragmented at this stage. The mRNA is then reverse transcribed into cDNA and then, if necessary, fragmented to lengths compatible with the sequencing system. Once all the fragments are sequenced, the transcripts (or reads) are assembled into genes. Although it is possible to assemble the transcriptome de novo, it is usually more efficient to align the reads to a reference genome or reference transcripts. As RNA-Seq is quantitative, a direct comparison between experiments can be made.

IN SITU HYBRIDIZATION

In situ hybridization (ISH) provides high-resolution gene expression information within the context of their natural location within an organ or organism. ISH uses a labeled cDNA fragment (i.e., probe) to locate a specific DNA segment in a portion or section of a tissue (in situ). The basic steps in ISH include cell permeabilization, hybridization of the labeled probe, and detection of the probe, thereby revealing the location of the mRNA of interest. This process can be adapted to a large scale system and the results are often shown in databases such as MGI, Gensat etc.

STRUCTURAL DATA BASE

Structural databases are essential tools for all crystallographic work and often need to be consulted at several stages of the process of producing, solving, refining and publishing the structure of a new material. Examples of such uses are:

i. Before deciding to synthesize a new compound the database could be used to check how many compounds with a particular chemical composition have been reported.
ii. After synthesizing and indexing the unit cell of a material the database can be searched to see if a material with the same or a similar unit cell is already known.
iii. If a material is found in the database with a similar unit cell to the new material then its structure may be close enough (i.e. same symmetry and similar unit cell contents) to be used as the starting model for the Rietveld refinement of the new material.
iv. To verify the results of a structure refinement the database can be consulted to find structures that have comparable bond distances, bond angles or coordination environments to the new structure.

The structures in the databases have been solved using X-ray, neutron and electron diffraction techniques on samples that are generally single crystals, but with the advances in structural solution using powder diffraction data, may be powders. There are some entries whose structures are predicted from computational modeling and some determined using NMR spectroscopy, these entries generally occur for protein samples.

1.4 SUMMARY

1. Bioinformatics is an emerging branch of biological science that emerged as a result of the combination of biology and information technology. It is a multidisciplinary subject where information technology is incorporated by means of various computational and analytical tools for the interpretation of biological data.
2. Primary databases are also called as archival database. They are populated with experimentally derived data such as nucleotide sequence, protein sequence or macromolecular structure.

3. Secondary databases comprise data derived from the results of analysing primary data. Secondary databases often draw upon information from numerous sources, including other databases (primary and secondary), controlled vocabularies and the scientific literature.

4. The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

4. A protein database is one or more datasets about proteins, which could include a protein's amino acid sequence, conformation, structure, and features such as active sites. Protein databases are compiled by the translation of DNA sequences from different gene databases and include structural information. They are an important resource because proteins mediate most biological functions.

5. The Gene Expression Database (GXD) is a community resource for gene expression information from the laboratory mouse. GXD stores and integrates different types of expression data and makes these data freely available in formats appropriate for comprehensive analysis.

1.5 TERMINAL QUESTION AND ANSWER

1. Define the Bioinformatics.

2. Write to scope and applications of bioinformatics.

3. Write to difference between primary, secondary and composite database.

4. Describe to nucleotide sequence database.

5. Comment upon protein sequence database.

6. Explain the gene expression database (GXD).

7. Write a short note on structural database.

1.7 REFERENCES

- Xiong J. (2006). Essential Bioinformatics. Texas A & M University. Cambridge University Press.
- Arthur M Lesk (2014). Introduction to bioinformatics. Oxford University Press. Oxford, United Kingdom.
- https://www.ebi.ac.uk/training/online/course/bioinformatics-terrified-2018/primary-and-secondary-databases.
- https://www.omicsonline.org/scholarly/bioinformatics-databases-journals-articles-ppts-list.php.
- https://www.ncbi.nlm.nih.gov/books/NBK44933/
- https://sta.uwi.edu/fst/dms/icgeb/documents/1910NucleotideandProteinsequencedatabasesDGL3.pdfphys.1
- https://www.nature.com/subjects/protein-databases

UNIT 2: DATABASE AND SEARCH TOOL

CONTENTS

2.1 OBJECTIVES

After studying this module, you shall be able to:

- Determine orthologs and paralogs for a protein of interest, assign putative function.
- A new bacterial genome is sequenced, how many genes have related genes in other species.
- Determine if a genome contains specific types of proteins.
- Determine the identity of a DNA or protein sequence.
- What is the identity of a clinical pathogen?
- Determine if particular variant has been described before.
- Many pathogens, especially viruses, mutate rapidly. We should like to know if we have a new strain.

2.2 INTRODUCTION

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data. Bioinformatics has been used for in silico analyses of biological queries using mathematical and statistical techniques. Bioinformatics derives knowledge from computer analysis of biological data. These can consist of the information stored in the genetic code, but also experimental results from various sources, patient statistics, and scientific literature. Research in bioinformatics includes method development for storage, retrieval, and analysis of the data. Bioinformatics is a rapidly developing branch of biology and is highly interdisciplinary, using techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, and linguistics. It has many practical applications in different areas of biology and medicine.

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

2.3 COMPUTATIONAL TOOLS AND BIOLOGICAL DATABASES

Computational Biology: The development and application of data-analytical and theoretical Methods, mathematical modeling and computational simulation techniques to the study of Biological, behavioral, and social systems.

2.3.1 NATIONAL CENTRE FOR BIOTECHNOLOGY INFORMATION (NCBI)

The National Center for Biotechnology Information (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health (NIH). The NCBI is located in Bethesda, Maryland and was founded in 1988 through legislation sponsored by Senator Claude Pepper.

The NCBI houses a series of databases relevant to biotechnology and biomedicine and is an important resource for bioinformatics tools and services. Major databases include Gene Bank for DNA sequences and PubMed, a bibliographic database for the biomedical literature. Other databases include the NCBI Epigenomics database. All these databases are available online through the Entrez search engine. NCBI was directed by David Lipman, one of the original authors of the BLAST sequence alignment program and a widely respected figure in bioinformatics. He also led an intramural research program, including groups led by Stephen Altschul (another BLAST co-author), David Landsman, Eugene Koonin, John Wilbur, Teresa Przytycka, and Zhiyong Lu. David Lipman stood down from his post in May 2017.

Gene Bank

NCBI has had responsibility for making available the GenBank DNA sequence database since 1992.Gene Bank coordinates with individual laboratories and other sequence databases such as those of the European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ).

Since 1992, NCBI has grown to provide other databases in addition to Gene Bank. NCBI provides Gene, Online Mendelian Inheritance in Man, the Molecular Modeling Database (3D protein structures), dbSNP (a database of single-nucleotide polymorphisms), the Reference Sequence Collection, a map of the human genome, and a taxonomy browser, and coordinates with the National Cancer Institute to provide the Cancer Genome Anatomy Project. The NCBI assigns a unique identifier (taxonomy ID number) to each species of organism. The NCBI has software tools that are available by WWW browsing or by FTP. For example, BLAST is a sequence similarity searching program. BLAST can do sequence comparisons against the Gene Bank DNA database in less than 15 seconds.

NCBI Bookshelf

The "NCBI Bookshelf is a collection of freely accessible, downloadable, on-line versions of selected biomedical books. The Bookshelf covers a wide range of topics including molecular biology, biochemistry, cell biology, genetics, and microbiology, disease states from a molecular and cellular point of view, research methods, and virology. Some of the books are online versions of previously published books, while others, such as Coffee Break, are written and edited by NCBI staff. The Bookshelf is a complement to the Entrez PubMed repository of peer-reviewed publication abstracts in that Bookshelf contents provide established perspectives on evolving areas of study and a context in which many disparate individual pieces of reported research can be organized.

Basic Local Alignment Search Tool (BLAST)

BLAST is an algorithm used for calculating sequence similarity between biological sequences such as nucleotide sequences of DNA and amino acid sequences of proteins. BLAST is a powerful tool for finding sequences similar to the query sequence within the same organism or in different organisms. It searches the query sequence on NCBI databases and servers and posts the results back to the person's browser in chosen format. Input sequences to the BLAST are mostly in FASTA or Gene bank format while output could be delivered in variety of formats such as HTML, XML formatting and plain text. HTML is the default output format for NCBI's Web-page. Results for NCBI-BLAST are presented in graphical format with all the hits found, a table with sequence identifiers for the hits having scoring related data, along with the alignments for the sequence of interest and the hits received with analogous BLAST scores for these

ENTREZ

The Entrez Global Query Cross-Database Search System is used at NCBI for all the major databases such as Nucleotide and Protein Sequences, Protein Structures, PubMed, Taxonomy, Complete Genomes, OMIM, and several others. Entrez is both indexing and retrieval system having data from various sources for biomedical research. NCBI distributed the first version of Entrez in 1991, composed of nucleotide sequences from PDB and Gene Bank, protein sequences from SWISS-PROT, translated Gene Bank, PIR, and PRF, PDB and associated abstracts and citations from PubMed. Entrez is specially designed to integrate the data from several different sources, databases and formats into a uniform information model and retrieval system which can efficiently retrieve that relevant references, sequences and structures.

GENE

Gene has been implemented at NCBI to characterize and organize the information about genes. It serves as a major node in the nexus of genomic map, expression, sequence, protein function, structure and homology data. A unique Gene ID is assigned to each gene record that can be followed through revision cycles. Gene records for known or predicted genes are established here and are demarcated by map positions or nucleotide sequence. Gene has several advantages over its predecessor, Locus Link, including, better integration with other databases in NCBI, broader taxonomic scope, and enhanced options for query and retrieval provided by Entrez system.

PROTEIN

Protein database maintains the text record for individual protein sequences, derived from many different resources such as NCBI Reference Sequence (RefSeq) project, Gene Bank, PDB and UniProtKB/SWISS-Prot. Protein records are present in different formats including FASTA and XML and are linked to other NCBI resources. Protein provides the relevant data to the users

Such as genes, DNA/RNA sequences, biological pathways, expression and variation data and literature. It also provides the pre-determined sets of similar and identical proteins for each sequence as computed by the BLAST. The Structure database of NCBI contains 3D coordinate sets for experimentally-determined structures in PDB that are imported by NCBI. The Conserved Domain database (CDD) of protein contains sequence profiles that characterize highly conserved domains within protein sequences. It also has records from external resources like SMART and Pfam. There is another database in protein known as Protein Clusters database which contains sets of proteins sequences that are clustered according to the maximum alignments between the individual sequences as calculated by BLAST.

Pubchem database

PubChem database of NCBI is a public resource for molecules and their activities against biological assays. PubChem is searchable and accessible by Entrez information retrieval system.

2.3.2 EUROPEAN BIOINFORMATICS INSTITUTE (EBI)

The European Bioinformatics Institute (EMBL-EBI) is an Intergovernmental Organization (IGO) which, as part of the European Molecular Biology Laboratory (EMBL) family, focuses on research and services in bioinformatics. It is located on the Well come Genome Campus in Hinxton near Cambridge, and employs over 600 full-time equivalent (FTE) staff. Institute leaders such as Rolf Apweiler, Alex Bateman, Ewan Birney, and Guy Cochrane, an adviser on the National Genomics Data Center Scientific Advisory Board, serve as part of the international research network of the BIG Data Center at the Beijing Institute of Genomics.

Additionally, the EMBL-EBI hosts training programs that teach scientists the fundamentals of the work with biological data and promote the plethora of bioinformatics' tools available for their research, both EMBL-EBI and non-EMBL-EBI-based.

BIOINFORMATIC SERVICES

One of the roles of the EMBL-EBI is to index and maintain biological data in a set of databases, including Ensembl (housing whole genome sequence data), UniProt (protein sequence and annotation database) and Protein Data Bank (protein and nucleic acid tertiary structure database). A variety of online services and tools is provided, such as Basic Local Alignment Search Tool (BLAST) or Clustal Omega sequence alignment tool, enabling further data analysis.

BLAST

BLAST is an algorithm for the comparison of bio macromolecule primary structure, most often nucleotide sequence of DNA/RNA and amino acid sequence of proteins, stored in the bioinformatics' databases, with the query sequence. The algorithm utilizes scoring of the available sequences against the query by a scoring matrix such as BLOSUM 62. The highest scoring sequences represent the closest relatives of the query, in terms of functional and evolutionary similarity.

The database search by BLAST requires input data to be in a correct format (e.g. FASTA, GenBank, PIR or EMBL format). Users may also designate the specific databases to be searched, select scoring matrices to be used and other parameters prior to the tool run. The best hits in the BLAST results are ordered according to their calculated E value (the probability of the presence of a similarly or higher-scoring hit in the database by chance).

CLUSTAL OMEGA

Clustal Omega is a multiple sequence alignment (MSA) tool that enables to find an optimal alignment of at least three and maximum of 4000 input DNA and protein sequences. Clustal Omega algorithm employs two profile Hidden Markov models (HMMs) to derive the final alignment of the sequences. The output of the Clustal Omega may be visualized in a guide tree (the phylogenetic relationship of the best-pairing sequences) or ordered by the mutual sequence similarity between the queries. The main advantage of Clustal Omega over other MSA tools (Muscle, ProbCons) is its efficiency, while maintaining a significant accuracy of the results.

ENSEMBL

Based at the EMBL-EBI, the Ensembl is a database organized around genomic data, maintained by the Ensembl Project. Tasked with the continuous annotation of the genomes of model organisms, Ensembl provides researchers a comprehensive resource of relevant biological information about each specific genome. The annotation of the stored reference genomes is automatic and sequence-based. Ensembl encompasses a publicly available genome database which can be accessed via a web browser. The stored data can be interacted with using a graphical UI, which supports the display of data in multiple resolution levels from karyotype, through individual genes, to nucleotide sequence.

Originally centered on vertebrate animals as its main field of interest, since 2009 Ensembl provides annotated data regarding the genomes of plants, fungi, invertebrates, bacteria and other species, in the sister project Ensembl Genomes. As of 2020, the various Ensembl project databases together house over 50,000 reference genomes.

PDB

PDB is a database of three dimensional structures of biological macromolecules, such as proteins and nucleic acids. The data are typically obtained by X-ray crystallography or NMR spectroscopy, and submitted manually by structural biologists worldwide through PDB member organizations –PDBe, RCSB, PDBj and BMRB. The database can be accessed through the webpages of its members, including PDBe (housed at the EMBL-EBI). As a member of the wwPDB consortium, PDBe aids in the joint mission of archiving and maintenance of macromolecular structure data.

UniProt

UniProt is an online repository of protein sequence and annotation data, distributed in UniProt Knowledgebase (UniProt KB), UniProt Reference Clusters (UniRef) and UniProt Archive (UniParc) databases. Originally conceived as the individual ventures of EMBL-EBI, Swiss Institute of Bioinformatics (SIB) (together maintaining Swiss-Prot and TrEMBL) and Protein Information Resource (PIR) (housing Protein Sequence Database), the increase in the global protein data generation led to their collaboration in the creation of UniProt in 2002.

The protein entries stored in UniProt are cataloged by a unique UniProt identifier. The annotation data collected for the each entry are organized in logical sections (e.g. protein function, structure, expression, sequence or relevant publications), allowing a coordinated overview about the protein of interest. Links to external databases and original sources of data are also provided. In addition to standard search by the protein name/identifier, UniProt webpage houses tools for BLAST searching, sequence alignment or searching for proteins containing specific peptides.

2.3.3 EMBL NUCLEOTIDE SEQUENCE DATABASE

The European Bioinformatics Institute (EBI) is an Outstation of the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. The EBI is located in the grounds of the Wellcome Trust Genome Campus near Cambridge, UK, next to the Sanger Centre and the UK Human Genome Mapping Project Resource Centre.

The main missions of the Service Programme of the EBI centre on building, maintaining and providing biological databases and information services to support data deposition and exploitation. In this respect a number of databases are operated, namely the EMBL Nucleotide Sequence Database (EMBL-Bank), the Protein Databases (SWISS-PROT and TrEMBL), the Macromolecular Structure Database (MSD) and Array Express for gene expression data plus several other databases many of which are produced in collaboration with external groups.

The EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/) is the European member of the tri-partide International Nucleotide Sequence Database Collaboration DDBJ/EMBL/GenBank. Main data sources are large-scale genome sequencing centers, individual scientists and the European Patent Office (EPO). Direct submissions to EMBL-Bank are complemented by daily data exchange with collaborating databases DDBJ (Japan) and GenBank (USA).

The EMBL database is growing rapidly as a result of major genome sequencing efforts. Within a 12 month period the database size has increased from about 6.7 million entries comprising 8255 million nucleotides (Release 63, June 2000) to over 12 million entries and 12 820 million nucleotides (Release 67, June 2001). During the same period the number of organisms represented in the database has risen by >30% to over 75,000 species.

Databases at EBI

The following section will deal with selected databases of EBI-EMBL:

Nucleotide databases

a. European Nucleotide Archive (ENA): ENA receives nucleotide data from a variety of sources, including small scale sequencing studies, sequencing centers and the INSDC (i.e.Genbank and DDBJ). In order to better manage the sequencing resources, ENA has been divided in several sub-databases such as

- ❖ **ENA-Genome** - for genome sequencing data
- ❖ **Sequence Read Archive (SRA)** –for Next Generation Sequencing (NGS) data
- ❖ **EMBL-Bank**- for assembled and annotated sequence data (note that submission of nucleotide data should be done at Genbank, EBI or DDBJ and not to all of these, as the data submitted in one of the database is automatically replicated or sent to the other two).

b. DGva: Database of Genomic Variant Archive (DGVa) is a publicly accessible database that stores information about genomic structural variants having role in causing diseases. Such variant may be in the form of

- Size ranging from few nucleotides to several Kilobase or even Megabases,
- Structural, i.e. insertions, deletions, translocations, and

- Copy number variants (CNV)

The DGVa is analogous to the dbVar database of NCBI. The data at DGVa can be accessed via the ensemble (www.ensembl.org) portal.

C.EGA: The European Genome Phenome Archive (EGA) stores data from studies that are carried out with an objective to understand the linkages between genotype and phenotype, especially from biomedical research. This database is analogous to the dbGaPdatabase at NCBI. Such data may have been generated from Genome wide association studies (GWAS). As the studies and datasets generally deals with disorders such as cancer, coronary artery defects, hypertension, Rheumatoid arthritis and diabetes, strict control during submission and public access is implemented on ethical grounds (as it contains information about patients and subjects taking part in the study) to prevent misuse or data.

D. ENA- Genome: This database contains the completed genome sequence data from a variety of organisms such as:

- Archaea and archeal virus
- Bacteria
- Eukaryotes
- Organelles
- Phages
- Plasmid
- Viroids

EMBL-EBI developed the ENSEMBL genomes tool to browse, analyses and visualize the genome sequencing data. Currently, there are close to 350 completed genome sequences available for browsing, analysis and downloading. The sequence analysis tools at ENSEMBL genome server provides tools for analysis at all levels of genome organization, such as whole genome, chromosome, genome segment, gene and transcript level. The genome visualization and analysis tool at ENSEMBL genome also provides links to molecular function, gene ontology, protein summary and structure tables.

e. Several other databases such as Immuno Polymorphism database (IPD) (such as IMGT/HLA, IMGT/LIGM, IPD-MHC, IPD-KIR e etc), Meta genomics and Patent data resources are also part of the nucleotide resources at EBI-EMBL. IMGT/HLA database is the nucleotide sequence database for human major histo-compatibility complex HLA. This database is a part of the International Immuno Genetics Project (IMGT) and the data has been subdivided into the following five classes of alleles of HLA (http://www.ebi.ac.uk/ipd/imgt/hla/stats.html):

- HLA Class I alleles (6725)
- HLA Class II alleles (1771)
- HLA alleles (8496)
- Other non-HLA alleles (148)
- Confidential alleles (8)

Alignment tools built into the database allows users to perform analysis and detect polymorphism at HLA loci. IMGT/LIGM similarly is a database for Immuglobulins and T-Cell receptors IPD- MHC contains sequences for Major histocompatibility factors for a large number of species

IPD-HPA is the database for human platelet antigens IPD-KIR is the database for human Killer cell Immunoglobulin like Receptors and contains information about 614 EBI-Metagenomic contains sequence information from microflora samples that have been collected from various environments. Some such examples include core gut microflora, aquatic microflora from Antarctica, glaciers, ocean samples, meat samples and so on. The metagenome sequences are analyzed to reveal the frequency of predicted CDS (coding DNA sequence), their GO (genome Ontology) annotation, putative proteins with biochemical, cellular and molecular functions.

2.3.4 DNA DATA BANK OF JAPAN (DDBJ)

Databases are like information banks which are used for storing and retrieval of sequence information. DNA Databank of Japan (DDBJ; http://www.ddbj.nig.ac.jp) is one of three nucleotide database which together with National Center for Biotechnology Information (NCBI) and European Bioinformatics Institute (EMBL), form a consortium known as International Nucleotide Sequence Database Collaboration (INSDC). DDBJ is the only nucleotide sequence databank of Asian origin and mainly collects sequences from Japanese researches. It is a primary nucleotide database; it collects data directly from the researchers. On accepting a nucleotide sequence, DDBJ issues an accession number to the submitter which has an international recognition. From July 2011 and June 2012, DDBJ had collected and released 15243000 entries/

12270462217 bases (Ogasawara et al., 2012). During 2009-10 DDBJ contributed 25.4% of the entries and 21.5% of the bases added to INSDC (Kaminuma et al., 2011).

History

DDBJ was established in the year 1986 at the National Institute of Genetics (NIG), Japan with support from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT). Later on for its efficient functioning, Center for Information Biology (CIB) was established at NIG in 1995. In 2004, NIG was made a member of Research Organization of Information and Systems.

The functioning and maintenance of DDBJ is monitored by an international advisory committee consisting of 9 members from Japan, Europe and USA. The committee reviews the functioning of DDBJ and reports the progress of DDBJ in database issue of Nucleic Acid Research Journal every year. Since its inception there has been a tremendous increase in the number of sequences submitted to DDBJ.

Roles of DDBJ

As a member of INSDC, primary objective of DDBJ is to collect sequence data from researchers all over the world and to issue a unique accession number for each entry. The data collected from the submitters is made publically available and anyone can access the data through data retrieval tools available at DDBJ. Everyday data submitted at either DDBJ or EMBL or NCBI is exchanged, therefore at any given time these three databases contain same data.

Activities of DDBJ

Collection of sequences

The sequences collected from the submitters are stored in the form of an entry in the database. Each entry consists of a nucleotide sequence, author information, reference, organism from which the sequence is determined, properties of the sequence etc.

Tools for data retrieval

Retrieval of data is as important as submission and one of the main objectives of any database is to provide the users with the required information. Any database contains enormous amount of information and retrieving the required information is also a tricky task which depends on right use of search strings. DDBJ hosts a number of tools for data retrieval like get entry (database retrieval by unique identifiers) and All-round Retrieval of Sequence and Annotation (ARSA). Unique identifiers required for retrieval through get entry can be accession number, gene name etc. Following are the steps along with snapshots showing data retrieval from DDBJ using get entry.

1. Open the homepage of DDBJ (http://www.ddbj.nig.ac.jp).

2. Click on the Search/Analysis link on the menu bar

(http://www.ddbj.nig.ac.jp/searches-e.html)

3. Click on get entry link (http://getentry.ddbj.nig.ac.jp/top-e.html)

4. Type in the accession number in the search box and click on search.

5. Desired sequence will be retrieved.

2.3.5 SWISS-PROT

Biological database can be defined as biological information stored in an electronic format and can be easily accessed throughout the world. These databases can be classified into various categories depending upon data type, data source, maintainer status etc. A variety of databases contain nucleotide and/or protein sequences data that are pertinent to a specific gene. Protein databases are specific to protein sequences. There are three important publicly accessible protein databases: Protein Information Resource (PIR), Swiss-Prot and Protein Data Bank (PDB). Whereas PIR and Swiss-Prot contain protein sequences, PDB is a structural database of biomolecules.PIR is considered as a primary database whereas Swiss-Prot falls into secondary database category. The aim of this chapter is to explain Swiss-Prot database and strategies to retrieve information from this database. Some of the tools and databases that are linked to each entry will also be discussed briefly.

HISTORY

Swiss-Prot is an annotated protein sequence database which was formulated and managed by Amos Bairoch in 1986. It was established collaboratively by the Department of Medical Biochemistry at the University of Geneva and European Molecular Biology Laboratory (EMBL).

Later it shifted to European Bioinformatics Institute (EBI) in 1994 and finally in April 1998, it became a part of Swiss Institute of Bioinformatics (SIB) (Bairoch and Apweiler, 1998). In 1996, TrEMBL was added as an automatically annotated supplement to Swiss-Prot database (Bairoch and Apweiler, 1996). Since 2002, it is maintained by the UniProt consortium and information about a protein sequence can be accessed via the UniProt website (http://www.uniprot.org/) (Apweiler et al., 2004). The Universal Protein Resource (UniProt) is the most widespread protein sequence catalog comprising of EBI, SIB and PIR (UniProt Consortium, 2009).

FEATURES

Swiss-Prot database is characterized for its high quality annotation which comes at a price of lower coverage. It provides information about the function of protein, its domain structure, post translational modifications (PTM) etc. In other words, it imparts whole information about a specific protein. Swiss-Prot database is curated to make it non- redundant. Therefore, this database contains only one entry per protein. As a result, the size of Swiss-Prot is very less as compared to DNA sequence databases. Figure 1 shows the development of the size of this database. The high quality annotation and minimum redundancy distinguish Swiss-Prot from other protein sequence databases.

There are four main features of Swiss-Prot:

1. High Quality Annotation: It is achieved through manually creating the protein sequence entries. It is processed through 6 stages:

a. Sequence curation: In this step, identical sequences are extracted through blast search and then the sequences from the related gene and same organism are incorporated into a single entry. It makes sure that the sequence is complete, correct and ready for further curation steps.

b. Sequence Analysis: It is performed by using various sequence analysis tools. Computer-predictions are manually reviewed and important results are selected for integration.

c. Literature curation: In this step, important publications related to the sequence are retrieved from literature databases. The whole text of each article is scanned manually and relevant information is gathered and supplemented to the entry.

d. Family based curation: Putative homologs are determined by Reciprocal Blast searches and phylogenetic resources which are further evaluated, curated, annotated and propagated across homologous proteins to ensure data consistency.

e. Evidence attribution: All information incorporated to the sequence entry during manual annotation is linked to the original source so that users can trace back the origin of data and evaluate it.

f. Quality assurance, integration and update: Each completely annotated entry undergoes quality assurance before integration into Swiss-Prot and is updated as new data become available.

2. Minimum redundancy: During manual annotation, all entries belonging to identical gene andfrom similar organism are merged into a single entry containing complete information. This results in minimal redundancy.

3. Integration with other Databases: Swiss-Prot is presently cross- referenced to more than 50 specialized databases. This extensive interlinking allows Swiss Prot to play a major role as a connecting link between various biological databases.

4. Documentation: Swiss-Prot Database contains a large number of index files and specialized documentation files. 'Documentation file' section provides an updated descriptive list of all document files.

2.4 SUMMARY

Databases are a source of vast amount of information generated from various sequencing projects. There are numerous kinds of databases available on web, but for protein sequence analysis, PIR, Swiss-Prot and PDB are the most relevant.

- The Swiss Institute of Bioinformatics collaborated with European Molecular Biology Laboratory to provide a high quality annotated database of protein sequences termed

Swiss-Prot. The latter is manually curated which makes it non-redundant and is directly linked to specialized databases.

- In 2002, SIB, EBI and PIR agreed to amalgamate their resources and resulted in the embodiment of UniProt consortium. Now, any information about a protein can be achieved via UniProtKB/Swiss-Prot database.

TrEMBL is a computationally generated annotation which is unreviewed and is not of higher quality. It is accessed through UniProt/TrEMBL database.

2.5 REFERENCES

- Apweiler R, Bairoch A, Wu CH (2004) "Protein sequence databases". Current Opinion in Chemical Biology 8(1): 76–80.
- Bairoch A, Apweiler R (1996) "The SWISS-PROT protein sequence data bank and its new
- Supplement TREMBL". Nucleic Acids Research 24 (1): 21–25.
- Bairoch A, Apweiler R (1998) The SWISS-PROT protein sequence data bank and its Supplement TrEMBL in 1998. Nucleic Acids Research 26(1): 38-42.
- UniProt Consortium (2009) The Universal Protein resource (UniProt). Nucleic Acids Research 37: D169-D174.
- "Background | European Bioinformatics Institute". Ebi.ac.uk. 16 May 2018. Retrieved 29 October 2019.
- "Jobs at EMBL-EBI". Retrieved 20 June 2016.
- "Scientific report" (PDF). www.embl.de. 2017. Retrieved 29 October 2019.
- BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences. (2018). Annual Report, p. 6. Retrieved 26 March 2020.
- "NCBI BLAST at EMBL-EBI". www.ebi.ac.uk. Retrieved 3 November 2021.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (October 1990). "Basic local alignment search tool". Journal of Molecular Biology. 215 (3): 403 410. Doi: 10.1016/S0022-2836(05)80360-2. PMID 2231712.
- Wheeler D, Bhagwat M (2007). BLAST QuickStart. Humana Press. PMID 17993672.
- "Clustal Omega at EMBL-EBI". www.ebi.ac.uk. Retrieved 3 November 2021.

- "Clustal Omega Documentation at EMBL-EBI". www.ebi.ac.uk. Retrieved 3 November 2021.
- Sievers F, Higgins DG (January 2018). "Clustal Omega for making accurate alignments of many protein sequences". Protein Science. 27 (1): 135–145. doi:10.1002/pro.3290. PMC 5734385. PMID 28884485.
- "Ensembl homepage". www.ensembl.org. Retrieved 3 November 2021.
- Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. (January 2021). "Ensembl 2021". Nucleic Acids Research. 49 (D1): D884 –D891. doi:10.1093/nar/gkaa942. PMC 7778975. PMID 33137190.
- "About the Ensembl Project". www.ensembl.org. Retrieved 3 November 2021.
- "Protein Data Bank: the single global archive for 3D macromolecular structure data". Nucleic Acids Research. 47 (D1): D520–D528. January 2019. doi:10.1093/nar/gky949. PMC 6324056. PMID 30357364.
- "About PDBe". www.ebi.ac.uk. Retrieved 3 November 2021.
- "About UniProt". www.uniprot.org. Retrieved 3 November 2021.
- "UniProt: the universal protein knowledgebase in 2021". Nucleic Acids Research. 49 (D1): D480–D489. January 2021. doi:10.1093/nar/gkaa1100. PMC 7778908. PMID 33237286.

2.6 TERMINAL QUESTIONS AND ANSWERS

1. The single point database search and retrieval system of NCBI is termed as ----- --.

2. What are the major domains under which NCBI databases and tools are organized?

3. Are the domains of NCBI standalone or interlinked?

4. What are the categories of databases at EBI?

5. What is ENA? What are the various sub-databases of ENA?

6. Which database stores genomic structural variation information? What is the comparable?

database at NCBI?

7. Write short note on various activities of DDBJ?

8. What are the different data submission tools available at DDBJ?

9. What is the difference between DSA and DTA?

UNIT 3: SEQUENCE ALIGNMENT AND DATABASE SEARCHING

CONTENTS

3.1 OBJECTIVES

After studying this module, you shall be able to:

- Determine the evolutionary basis of sequence alignment.
- Know how to do Database similarity searching.
- Know about using the Sequence Similarity search tools: BLAST and FASTA.
- Get the Concept of Alignment.
- Know about Multiple Sequence Alignment (MSA).
- Describe "Percent Accepted Mutation" (PAM).
- Use the Blocks of Amino Acid and Substitution Matrix (BLOSUM)

3.2 INTRODUCTION

One of the central themes in bioinformatics is the concept of "similarity" and "relatedness" which in turn is based on evolutionary relationship or ancestry. We use such themes of "similarity/relatedness" in a variety of applications such as

- Gene and genetic element finding
- Molecular evolution or phylogeny
- Comparative genomics
- Structure prediction through homology modeling, and several others

The principle on which all these are based is sequence similarity that can be deduced via Sequence Alignment. We often can deduce relationship among objects by identifying similar features or characters. Alignment also attempts to identify similarity between two or multiple sequences by applying a similar logic, except that several events (such as types, frequency and Occurrence of mutation) that may have led to similarity or dissimilarity are also taken into account. Before we delve into the principles of sequence alignment, it may be useful to refresh some of the concepts of mutation and evolution and keep them in mind while understanding alignment.

a. Mutations occur at the level of DNA

b. Mutations can survive or are accepted if they are potentially non-harmful (Selectively neutral) or confer some selective advantage to the organism and population. A mutation that is harmful, has a negative impact and may be lethal will be lost from the population.

c. Small mutations such as single-base changes include transitions and transversions, and insertion and deletion of bases.

d. Transitions are more frequently encountered than transversions.

e. Non-coding DNA can accumulate mutations or changes at a higher rate than coding regions (because of the subsequent consequences on the encoded proteins).

f. Due to degeneracy of codons and Wobble bases, all mutations at DNA level do not have an impact at the protein level and are thus deemed to be silent.

g. Any change in DNA sequence that does not alter protein sequence is termed as synonymous; and a change in DNA that leads to incorporation of an alternate amino acid is termed non-synonymous.

h. Within proteins, replacement rate of one amino acid with another is rarely observed within domains or functional units

i. Amino acids belonging to similar chemical or physical properties are more likely to replace one another

j. Rate of evolution among DNA is higher than proteins; or in other words, proteins are more conserved than DNA sequences

As alignment aims to find matches between similar residues, concepts of evolutionary biology are widely used.

3.3 THE EVOLUTIONARY BASIS OF SEQUENCE ALIGNMENT

The course of evolution proceeds in small incremental stages i.e. instead of large scale disruptions that span entire genomes, evolution favors small variations spread throughout the genome. Off-course it is difficult to actually define the physical boundaries of what constitutes “large” or “small”! For the sake of simplicity, let us limit our definition of “small” to single base or amino acids, and “large” being several Kilo bases or even Mega bases in dimensions. As majority of the changes are small, it is possible for us to detect similar regions with the genome through alignment. We also presume that regions that share considerable levels of similarity as measured through alignment must have shared ancestry or have common evolutionary history. Such regions are termed as homologous sequences. Homology can be further sub-divided into orthology and paralogy which are shared evolutionary history either by speciation or through duplication. A note of caution: Two sequences can also share high similarity without sharing recent ancestry. Such sequences are termed as xenologs and are generally acquired through horizontal gene transfer. An alignment attempts to create a matrix of rows and columns where each row denotes a sequence and each column is occupied by similar characters derived from each sequences or a gap. Pair wise alignment attempts to align two sequence at-a-time, whereas multiple sequence alignment (MSA) attempts to align more than two sequences. If there are several sequences are derived from organisms having a common shared ancestry or evolutionary history, we expect that these sequences will exhibit similarity but will not be exactly identical i.e. we expect to find similar characters or residues and also some differences. The differences or dissimilarities encountered are a result of mutational events; more the time since common ancestry, more the number or accumulated mutation and therefore more the number of dissimilar residues. The number of changes is therefore directly proportional to evolutionary time.

Therefore alignment tools will try to generate the matrix such that there are more identical and/or similar residues. It may be worthwhile to point out in case a mutational event or events lead to deletion of the nucleotides, “gaps” are introduced while performing the alignment to “mimic” the event and “achieve” an alignment with maximal identity. Therefore sequence alignment is a combination of correctly identifying and placing similar and dissimilar residues in columns.

Given the complexity involved because of length and types of changes observed in sequences, it is impossible to derive alignment manually and we have to rely on various algorithms and software for an automated alignment process.

Pair wise alignment employs two distinct strategies for alignment or similarity searching; termed as “local” and “global”. Local alignment attempts to generate an optimal alignment of similar and dissimilar residues over a short block of sequences with maximal identity; whereas global alignment tries to identify an average identity over the entire length of the sequences; local alignment algorithm was developed by Temple Smith and Michael Waterman (1981), whereas

Saul Needleman and Christian Wunsch developed the algorithm for global alignment. Sequence alignments employ matrices to find an optimal alignment and the following section will introduce you to some of the matrices commonly used.

3.4 DATABASE SIMILARITY SEARCHING

3.4.1 SEQUENCE SIMILARITY SEARCH TOOLS: BLAST AND FASTA

Introduction to Sequence Analysis

Analysis of the sequence data is one of the major challenges of computation biology and is the first step towards understanding molecular basis of development and adaptation. Several types of analysis can be performed that range from

DNA Sequence analysis

□ Sequence similarity searches

□ Prediction of genes and other genetic elements

□ Evolutionary tendencies and trends

□ Functional information

RNA analysis

□ Expression analysis

□ Structure

□ Functional information

Protein level

□ Domain finding

□ Structure prediction

- Evolution

- Function

Genome level

- Comparative genomics

- Genome organization and re-organization

- Genome annotation

Similarity search with Nucleotide queries

DNA sequence analysis constitutes one of the major applications in bioinformatics. Some of the basic objectives of performing sequence analyses are

- Sequence retrieval

- Finding similar sequence through similarity searching

- Phylogenetic or evolutionary analysis

- Finding homology relationships (orthologus and paralogous nature)

- Discovering new genes and genetic elements

- Exploring importance of residues (nucleotides and amino acids) that are important for structure and function

Central to the process of searching for similar sequences from database and retrieval are concepts of homology that are derived from evolutionary relationships. DNA data can be used to retrieve similar sequences that have diverged upto 600 million years ago! Sequences can be retrieved from NCBI database by using the identity of the sequence in the form of accession number and/or using the "Identity" of the sequence as "query" to search against the entire database or by selecting a specific database. Sequence can be retrieved as **FASTA** formatted sequence or in Genbank format. **FASTA formatted files** are simple text files of nucleotide or protein sequence where a single definition beginning with a “greater than (>)” sign is placed at the beginning of the the sequence. This is one of formats that are recognized by almost all sequence analysis Software’s. A single file can contain several **FASTA** formatted sequence that can then be used for analysis such as in multiple sequence alignment.

Genbank formatted files contains detailed annotation and the associated sequence Sequence similarity search is performed using a suite of tools called **“BLAST” i.e. Basic Local Alignment Search Tool**. Two distinct types of sequence similarity searches can be performed – Local and Global. Needleman and Wunsch developed the GLOBAL alignment algorithm (1970) whereas Michael Waterman and Temple Smith co-developed the Smith- Waterman sequence alignment algorithm for LOCAL alignment (1981). Global alignment attempts to find an “optimal or average” similarity via alignment over the entire length between the user provided “query” and “subject” sequences that are part of the database. Local alignment, in contrast attempts to find “local” regions of high similarity between query and subject sequences. Sequence similarity searches, performed via alignment are a measure of relatedness i.e. sequences that are evolutionary closely related will align over larger distances; in other words similarity is a function of evolutionary relatedness. Similarity searches carried out against subject sequences in the database are based on pairwise alignment, i.e. between two sequences at-a-time. One of the two sequences is always the “query” sequence, whereas the subject sequences retrieved from the database changes. Similarity being a function of evolutionary relationship can also be extended for employing sequence alignments to evaluate molecular phylogeny via multiple sequence alignment.

3.5 CONCEPT OF ALIGNMENT

3.5.1 MULTIPLE SEQUENCE ALIGNMENT (MSA)

Once we have retrieved a number of sequences using BLAST (see earlier chapters on this),

Several questions that can be raised are:

a. how are all these sequences related?

b. what is the level of similarity or divergence between all these?

c. what residues are conserved across all the sequences and thereby may be of

functional importance?

d. what is the evolutionary relationship between these sequences?

e. are there motifs present in these sequences?

f. are there polymorphic sites?

Multiple sequence alignment can be employed to answer these questions. Multiple sequences

alignment can be viewed as a "reiterative pairwise alignment" i.e. all the sequences are aligned with each other in a pairwise manner so as to arrive at an output that attempts to align such that all the similar or identical residues from the various sequences appear in the same column. Gaps are introduced to arrive at an optimal multiple alignments.

There are five different methods to perform Multiple Sequence Alignment with some representative software in parentheses:

a. Exact method

b. Progressive method (CLUSTAL)

c. Iterative method (MUSCLE)

d. Consistency-based method (MAFFT)

e. Structure-based method (Expresso)

The most common tool for multiple sequence alignment is Clustal that can be either be used as a web-based service or the software can be downloaded from http://www.clustal.org/. It employs progressive alignment as to perform a MSA. Clustal first creates a global pairwise alignment for all sequence pairs with alignment/similarity scores and then starts the MSA with the two sequences with highest score and progressively adds more and more sequences to complete the alignment. Along with the MSA, Clustal also generates a tree depicting the "phylogenetic" relationship among the sequences analysed.

Clustal can be downloaded from www.clustal.org and installed on any computer. Also prepare a text file containing FASTA formatted sequences for alignment. Such sequences could have been identified through BLASTN or BLASTP.

The sequences have to be loaded onto ClustalX, and then aligned using "perform complete alignment" command. Once a complete alignment has been performed, the resultant alignment can be viewed in Clustal itself or the alignment file can be viewed using any text editor such as "notepad" or "WordPad". The Tree can be viewed using several software's, including Tree View that can be downloaded free-of-cost from http://taxonomy.zoology.gla.ac.uk/rod/treeview.html.

3.5.2 PERCENT ACCEPTED MUTATION (PAM)

PAM or Percent Accepted Mutations also sometimes expanded as **Point Accepted Mutations** was developed by Margaret Dayhoff and was published in 1978. She and her co-workers analyzed 71 families belonging to 34 super-families of evolutionarily related proteins by comparing their sequences over their entire lengths (diverged over various time scales) and observed 1572 changes. These changes were tabulated or used to create a matrix of 20x20 (number of amino acids). The observation that one amino acid could be substituted by another amino acid prompted the authors to term these as Point Accepted Mutation to signify that the substituted amino acid is accepted by natural selection. This is deemed to be an outcome of two distinct processes:

a. occurrence of a mutation in DNA leading to a change in amino acid

b. acceptance of the newly incorporated amino acid through natural selection

Some of the protein families that were analysed by Dayhoff are (Pevsner 2009):

a. Immunoglobulin (Ig) Kappa chain C region

b. Casein

c. Epidermal growth factors

d. Serum albumin

e. Alpha chain of hemoglobin

f. Myoglobin

g. Trypsin

h. Nerve growth factor

i. Insulin

j. Cytochrome C

k. Glutamate dehydrogenase

l. Histone H3 and H4

The analysis of such a superfamilies allowed Dahyhoff to study the replacement or substitution frequencies over a large phylogenetic distance and examine the rates at which substitutions occur.

PAM1 for example is derived from proteins that are nearly 1% diverged i.e. 1 accepted change per 100 residues; whereas PAM250 matrix is derived from proteins that exhibit 250 changes over 100 residues. PAM1 and PAM250 therefore represent two different classes of substitution probabilities—PAM1 for closely related proteins, and PAM250 for diverged proteins.

Such probabilities were computed for all the amino acid pairs for number of PAM values, such as PAM10, PAM30, PAM70 and so on. Increasing number of PAM indicates larger evolutionary time scale.

These substitution probability values or odds ratio values are then converted into PAM matrices by taking 10 times the base 10 logarithmic of the odd ratio to obtain the Log-odds scoring matrix or PAM matrix. So similar substitution odds when considered in combination with evolutionary time scale or divergence time has different scores. For example, the log- odds score of Alanine remaining conserved as Alanine in PAM 10 and PAM received scores of 7 and 2, respectively.

3.5.4 BLOCKS OF AMINO ACID AND SUBSTITUTION MATRIX (BLOSUM)

BLOSUM matrix: One of the drawbacks associated with the PAM matrix was the fact that the PAM250 matrix was generated by multiplying PAM1 matrix to itself 250 times and thus although it is meant for long evolutionary scale, it is only an approximation derived by compounding values obtained over short evolutionary time scale. Additionally, an unintended major drawback was the fact that in 1978 and earlier, very few protein sequences were available and therefore, PAM matrices were based on a very small set of proteins of similar nature and thus may not represent the entire spectrum of amino acid changes or substitution. Steven Henikoff and Jorga G Henikoff in 1992 published a new and updated amino acid substitution matrix termed as BLOSUM matrix. The matrix was derived from analysis of BLOCKS database of protein domains (blocks.fhcrc.org/) which contains ungapped multiple aligned regions of domains or most conserved segments of proteins. Henikoff and Henikoff (1992) used nearly 2000 ungapped aligned sequences from more than 500 groups of related proteins to devise the BLOCKS SUBSTITUTION MATRIX or BLOSUM matrix. The BLOSUM matrix was found to be more accurate and closer to observed changes than PAM matrix because a. The use of large and evolutionary diverse dataset meant more realistic estimation of substitution probabilities, and b. Probabilities were based on conserved domains that are under greater selection pressure and thereby reflecting true estimation of substitution rates. Like PAM matrices, Henikoff and Henikoff also computed the log-odd ratio of substitution probabilities for varying evolutionary time scale to generate several matrices such as BLOSUM 45, BLOSUM 50, BLOSUM 62, BLOSUM 80, and BLOSUM 90. The scheme of numbering in BLOSUM reflects the "proportion" of conserved residues; higher the number, higher is the conserved residues and thereby more suited for analysis of closely related protein sequences. BLOSUM62 is now default matrix used by several sequence similarity tools including BLAST.

A comparison of PAM and BLOSUM reveals that PAM is based on global alignment of proteins whereas BLOSUM is based on local alignment of conserved domains. The numbering in scheme in PAM and BLOSUM is also opposite, lower number in PAM means less divergence and in BLOSUM means less conservation; higher numbers in PAM denotes high divergence whereas in BLOSUM means high level of conservation. PAM matrix is preferred for global alignment of proteins whereas BLOSUM matrices are preferred for local alignment.

3.6 SUMMARY

The concept of sequence alignment that estimates similarity or relatedness is based on the fundamental principles of evolution. Before attempting to perform sequence alignment, it is imperative to understand that mutations occur at the level of DNA with non-coding regions accumulating mutations at a higher rate than coding regions, and that not all mutation lead to an alteration in the amino acid sequence. Given this background, the information content of DNA and proteins are thus variable. Orthologous sequences are likely to share more similarity compared to paralogs because of their evolutionary history. Sequence similarity and alignment can be performed either in a pairwise manner or using multiple sequence alignment. The objective of alignment is to create a matrix with rows and columns; the rows represent the taxonomic units with an objective of placing similar or identical sequence data in a single column. While creating the alignments, unitary matrix is used to compute the mutational substitution rates for DNA whereas PAM and BLOSSUM matrices are employed to compute Mutational probability Indices in case of proteins. Tools such as BLASTN and BLASTP are used for pairwise sequence alignment, and CLUSTAL, MUSCLE, MAAFT and Expresso are employed for multiple sequence alignment. The output generated upon Multiple sequence alignment can be further viewed either as an alignment file (using any text viewer such as Wordpad or notepad) and as tree file using TreeView.

3.7 TERMINAL QUESTIONS AND ANSWERS

1. What is the principle of similarity searching?

2. What are the objectives of analysis of sequence data?

3. Define the following:

    a. Accession number

    b. Query

    c. Subject

4. What is the relationship between alignment and similarity?

5. What are applications of sequence alignment?

6. Which of the following accumulate mutation at higher rate: Non-coding or coding DNA?

REFERENCES

- Altschul S.F, Gish W, Miller W, Myers E W and Lipman D J. Basic Local Alignment Search Tool. J. Mol. Biol. 215, 403-410 (1990)
- Bioinformatics and Functional Genomics: 2nd Edition, Jonathon Pevsner (2009), WileyBlackwell
- Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. A model of evolutionary change in proteins. In "Atlas of Protein Sequence and Structure" 5(3) M.O. Dayhoff (ed.), 345 -352 (1978)
- Henikoff, S. and Henikoff, J. Amino acid substitution matrices from protein blocks Proc.Natl. Acad. Sci. USA. 89(biochemistry): 10915 - 10919 (1992).
- Robert C Edgar (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics: 5:113 doi:10.1186/1471-2105-5-113
- Katoh, Misawa, Kuma, Miyata (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059-3066)
- Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. (2007). Clustal W and Clusta

UNIT 4: COMPUTATIONAL TOOLS FOR DNA SEQUENCE ANALYSIS

CONTENTS

4.1 OBJECTIVES

After studying this module, you shall be able to:

1. Determine how to submit data and how to retrieve data.

2. Determine the relationship between sequence and biological functions.

3. Define Molecular Phylogeny and its uses.
4. Determine the Consistency of Molecular Phylogenetic Prediction.
5. Know the applications of bioinformatics.

4.2 INTRODUCTION

The National Center for Biotechnology Information was established in 1988 at the National Institute of Health (NIH) as part of the National Library of Medicine (NLM) and is located at Bethedsa, Maryland, USA. This association of NCBI with NIH and NLM is reflected in its web-address (www.ncbi.nlm.nih.gov). NCBI was set up to collate information, create databases and conduct research in the field of molecular biology especially for biomedical data, and develop computational tools. Since then, the database and computational tools have expanded to include diverse organisms including plants so as to encompass not only data from biomedical field but also include agriculture, food and other plant derived resources. NCBI has now emerged as the primary source of free public-access data encompassing a wide range of disciplines ranging from literature, sequence information, expression profile data, protein sequence and structure, chemical structure and bioassays, taxonomy; in addition, NCBI has developed a variety of analysis tools that are available for free download and use.

The diverse activities of NCBI can be broadly categorized into:

a. research at molecular level using mathematical and computational tools on fundamental problems in biology

b. formulating uniform standards for generation and deposition of computational data, nomenclature or annotation of biological material and information; and facilitating exchange of such standards

c. developing and distributing databases and software

d. developing and maintaining collaborations with academia, industry and other governmental agencies at national and international level through visitors program

e. fostering scientific communication through sponsoring and organizing meetings, workshops and lectures

f. supporting training program on basic and applied aspects of computational biology

The resources at NCBI are categorized into major groups and following are some of the broad sets of various databases and tools developed, curated and hosted at NCBI:

Submissions:

Genbank: BankIt

Genbank: Barcode

Genbank: Sequin

GEO Web deposit

NIH Manuscript submission (NIHMS)

SNP submission

PUBChem Deposition gateway

BioProject Submission

Databases:

Literature (PubMed, PubMed Central; NCBI Bookshelf):

Entrez and Entrez Programming utilities:

DNA and RNA (Refseq, nucleotide, EST, GSS, WGS, PopSet, trace archive, SRA): Proteins (Reference sequences, GenPept, UniProt/SwissProt, PRF, PDB, Protein clusters, GEO, Structure, UniGene, CDD): Genomes (Map Viewer, Genome workbench, Plant Genome Central, Genome

Reference Consortium, Epigenomics, Genomics Structural variation): Maps Taxonomy PubChem Substance OMIM

Tools:

Data mining

Sequence analysis (Vector Screen, BLAST, CDART)

Electronic PCR (forward and Reverse)

GEO-BLAST

Genetic codes

ORF finder

Splign

3-D structure viewer (Cn3D)

3-D structure and similarity searching

1000 Genome Browser

Others:

FTP downloads sites:

Collaborative cancer research:

Entrez is the single point database search and **retrieval system** that allows a user to perform the search and retrieve action against "all" or a "specific" database in an interlinked manner.

4.3 & 4.4 DATABASE SUBMISSION & DATA RETRIEVAL

NCBI relies on submission of accurately annotated and curated data submitted by the research community. The data can be grouped into two major types - sequence and non-sequence. The diverse types and categories of data hosted at NCBI require that these are deposited into one of the many databases in an appropriate format with annotations. The following section will introduce you to the several forms of biological data and the submission gateways at NCBI.

Submission of sequence data:

The field of computational biology has experienced tremendous / exponential growth on account of deluge of nucleotide sequence data. This in turn has been helped by the advancements in automated sequencing capabilities, vastly improved chemistry of sequencing and greatly reduced cost.

The sequence data generated under a variety of research objectives or goals such as sequencing from "traditional" studies, whole genome sequencing programs (small genomes, large complex genomes), population genetics studies, sequence variation or barcoding projects, and other sequencing projects need to be deposited into one of the following databases:

i. Genbank

ii. Sequence Read Archives (SRA)

iii. dbSNP

iv. dbVar

v. GEO

Sequence submission tools:

Sequence types

Traditional sequence information may constitute of

i. Single or multiple sequences for different genes or loci

ii. Multiple sequences for same gene or loci but derived from several individuals, varieties, species or other taxonomic units

iii. Barcode sequences such as those originating from Cytochrome Oxidase I loci of mitochondria

The sequence data can be submitted using one the several following tools:

Using BankIt: It is a web-based tool that is preferred for submission of single or a small set of sequences, and has relatively simple features for annotation. The submitter has to register at NCBI and after filling the requisite details can deposit the sequence/s using the web-based tool.

Using Sequin: Sequin is the preferred standalone option for submission if the sequences to be submitted

- need advanced network accessible analytical tools,
- are complex and require detailed annotation,
- The ability to launch graphical viewing and editing option, and
- Pre-submission work can be done offline i.e even in the absence of network
- an option to update and edit the sequences and features in future

The sequence file prepared using Sequin Program allows users to view the sequence and its associated features in Genbank and Graphical view, in addition to several other formats.

Barcode submission tool: "DNA Barcode" can be defined as short nucleotide sequence from a standard, well characterized genetic loci such as mitochondrial cytochrome oxidase, chloroplastic maturase K, tRNA Lysine (trnK), large subunit of RUBISCO (rbcL), and ITS region from nuclear rDNA. The inherent sequence conservation and variability in such loci aids researchers in species identification. As per the policy of NCBI, all sequences originating from Mitochondrial oxidase loci can be submitted through the DNA-barcode submission tool, while the rest of the sequences need to be submitted through the BankIt tool.

Batch submission: Sequences that have been generated through high-throughput sequencing projects such as single pass sequencing of cDNAs (EST), genomic survey sequences (GSS) and genomic mapping projects (STS) are to be submitted through either ftp (file transfer protocol) or via e-mail after annotation.

Submission of Genome sequences: Submission of genome sequences require that the project is first registered with NCBI for allotment of project ID. Small genomes such as chloroplast, mitochondria, plasmids, phages and viral do not require registration and can be submitted using Sequin tool. Large prokaryotic genome sequences have to be formatted as a FASTA file followed by adding annotation features. Annotation requires that the following information must be provided along with coordinates or positions in the genome:

i. Genes

ii. Coding region of known proteins and protein identity

iii. Structural RNAs i.e. tRNA and rRNA

In addition to these mandatory annotation features, optional features can also be submitted such as information about other non-coding RNAs, transposons etc.

The sequence can typically be prepared for submission through Sequin or using tbl2asn and then submitted via FTP.

Eukaryotic genomes need FASTA formatted sequence files to be annotated with the following features mandatorily:

- ➢ Genes
- ➢ Coding regions of known proteins along with protein names and ID
- ➢ mRNA features
- ➢ Transcript ID

The annotation can be prepared using Sequin and submitted using Genome Submission tool. NCBI also accepts deposition of information for sequence based reagents such a primer pairs, siRNA, probe-sequences into Probe database. Such information must be accompanied by probe unique identifier, name and probe type. In addition, optional information on target can also be provided.

Submission of High throughput sequences derived from transcript survey sequence assemblies and metagenomic studies are to be deposited to transcriptome shotgun assembly (TSA) archive and metagenome archive, respectively.

Submission of Non-Sequence data

Non-sequence data comprises of information being generated through microarray, Human clinical studies, chemical substances, structure and bioassay data, manuscripts etc. The Gene Expression Omnibus (GEO) database is meant for deposition and cataloguing of a variety of functional genomics and quantitative data generated via high throughput technologies. Some such data types include:

- Expression analysis performed via Microarray (oligonucleotide, cDNA) RT-PCR (real-time reverse transcriptase PCR) SAGE (Serial Analysis of Gene Expression)
- High throughput sequence submissions
- mRNA sequencing
- small RNA sequencing
- ChIP-sequencing
- Methyl sequencing
- Digital gene expression

GEO accepts microarray data that have been generated using microarray chips manufactured by Affymetrix, Agilent, Nimblegen, Illumina and also custom made by users. The user needs to provide information about

i. Array or sequencer

ii. Array template or array design i.e. the identity of the spots

iii. Description of biological sample

iv. Protocol

v. Hybridization result or sequence count

vi. Raw and/or processed data file of intensity values or sequence counts

The experiments must have perfomed following Minimum Information About Microarray Experiment (MIAME) guidelines.MIAME guidelines include-(http://www.ncbi.nlm.nih.gov/geo/info/MIAME.html):

- Raw data for hybridization
- Processed data for hybridization
- Sample annotation
- Experimental deisgn
- Array annotation
- Laboratory and data processing protocols

GEO also accepts quantitative data generated through

- RT-PCR experiments
- ChIP-Chip (Chromatin Immuno-precipitation-chip)
- ArrayCGH (array Comparative Genomic Hybridization)
- SNP array (Single nucleotide Polymorphism)
- SAGE (serial Analysis of Gene Expression) and
- Protein Array

For RT-PCR experiments, the following information should be supplied

- Studies that have been performed with atleast 50 genes (sometimes data for 20 genes can also be accepted)
- Protocol
- Sample information
- Non-normalized data and normalized data
- Fold-change data

dbGaP (database of Genotype and Phenotype) collates data originating from several studies that have analysed the relation between genotype and phenotype, including Genomewide association studies (GWAS), medical sequencing, molecular diagnostic studies and also from other genetic studies that are non-clinical in nature. As some of this data may have confidentiality and ethical aspects including identity of participants in clinical studies, a data access committee (DAC) and data use certification (DUC) regarding ethical treatment, biosafety approval, and confidentiality is required.

As most of the GWAS deal with understanding the relationship between genetic factors and health, a Submission Certificate must be obtained prior to submission of data to dbGaP confirming the ethical nature and confidentiality of the study.

The submitter must deposit de-identified subject identity and consent for each subject that have participated in the study, genetic data such as sequence and/or array information, and phenotype data. Phenotype data may consist of body site, Histological type etc.

PubChem database accepts information about chemical substance, structure and bio-activity and for ease of usage has been further sub-divided into PcSubstance, PcCompound and PcBioAssay database. Before submission, a user has to register at PubChem with an option to open a test or a deposition account. A test account allows users to first validate all steps of submission and format of submission without actually depositing and releasing the data; a deposition account on the other hand will allow the user to deposit and release the information into the database.

In order to successfully deposit into PubChem database, the user should provide the following information:

- Biological properties
- Chemical reactions
- Imaging agents (in case of Bioassay)
- Metabolic pathway
- Physical properties
- Protein 3-D structure
- Toxicological information

Original and novel findings that have been peer reviewed by subject experts and accepted for publication in a research journal can be deposited in the PubMed Central database using NIH manuscript submission system (NIHMS).

Each data (sequence, literature, microarray, structure, genome sequence, primer etc.) that is deposited in the NCBI is allotted a unique identifier in the form of an accession number.

4.5 RELATIONSHIP BETWEEN SEQUENCE AND BIOLOGICAL FUNCTIONS

What is the relationship between the sequence similarity and structure similarity in biological proteins?

Proteins with high sequence identity and high structural similarity tend to possess functional similarity and evolutionary relationships, yet examples of proteins deviating from this general relationship of sequence/structure/function homology are well-recognized.

What is the relationship between DNA sequence and protein structure?

DNA sequence provides the code for the amino acid sequence. The amino acid sequence determines the structure of the protein, which affects the function of the protein.

4.6 MOLECULAR PHYLOGENY

Introduction

Mutation is the basis of evolution driven by the process of selection. All life forms are expected to be part of a tree of life, which should be able to explain their origin and evolution. Practically, this may not happen due to extinction of species and further complications arising from ways by which organisms can acquire genes (e.g. lateral transfer of genes). Phylogenetics exploits available comparative information to generate trees, which can explain evolution. Traditionally morphological features were used to compare data and generate trees. More recently molecular sequences are used for comparisons among species, helping in defining species, families and other taxa, hence named as "Molecular Phylogeny".

How to generate trees

Trees are generated by comparing traits among organisms. For classical phylogeny these traits are morphological traits but for molecular phylogeny we can use DNA, RNA or protein sequence data. As a general rule DNA has more phylogenetic information as compared to proteins. Proteins are derived through triplet code, in which third bases follow the "wobble hypothesis" leading to loss of phylogenetic information. DNA sequences comprise coding and non-coding regions that have differing rates of evolution. The rate of evolution also depends on the type of organism.

Comparison of sequences can only be done after aligning them. Without alignment it is very difficult to decide which nucleotide/amino acid should be compared with which one (homology). Proteins show two types of changes- synonymous and non-synonymous. A synonymous change does not result in change in the coded amino acid.

Positive and negative selection

Traditionally, any change which is favored by natural selection is called positive selection. It is favored by natural selection because it helps in the survival of organism. Similarly, any trait which is not favored by natural selection is normally eliminated and is called negative selection.

Similar kind of selection also operates for molecular sequences. It is common among genes to go through duplication. A duplicated copy of gene is free to undergo mutation and create variation. This variation goes through positive/negative selection and often leads to neo-functionalization, leading to new genes with new functions.

Understanding Trees

Cladograms vs Phylograms

Trees fall under two categories–Cladogram and Phylogram. Cladogram just provide the information about relationship between different organisms while phylograms also provide a measure of the amount of evolutionary change, as seen in the branch-lengths. Due to this fact, branch length has no meaning in cladograms while it has meaning in phylograms.

Rooted vs Unrooted trees

The root in a tree denotes the ultimate common ancestor and provides direction in time. At times, it is not possible to have this information hence there are both types of algorithms available- those we do apply a common ancestor hypothesis and those we does not. A common way to decide the root of tree is by using an outgroup. An outgroup is a taxon from a group closely related to the ingroup, which includes the taxa under study.
Another way to identify the root is to use midpoint as the rooting point for the longest branch.

Tree Terminology

Trees can be described based on branches and nodes. Terminal branches represent Operational Taxonomic Unites (OTU's). When two branches are connected, it results in internal nodes. When two terminal branches are directly connected to each other, they are called sister branches.

If two lineages (branches) originate from one internal node, it is called bifurcation or dichotomy. If there are more than two branches are coming out of one internal node, this is called as polytomy and tree is said to be multifurcating.

Methods of Phylogenetic reconstruction

Various methods have been proposed to build a phylogenetic tree. We will only consider three here: distance based method (UPGMA and NJ), maximum parsimony (MP) and maximum likelihood (ML).

Distance Method

Distance based methods start with calculating pairwise distances between sequences based on pairwise alignment. These distances form a distance matrix which is used to generate the tree. Commonly known methods to generate the tree from this matrix are Unweighted Pair Group Method using Arithmetic mean (UPGMA) and Neighbor Joining (NJ). Distance based methods are fast but overlook substantial amount of information in a multiple sequence alignment. Distance is calculated as dissimilarity between the sequences of each pair of taxa.

UPGMA distance based method

It is no longer a popular method and distance based tree now use NJ as a method of choice. In UPGMA is a progressive clustering method. All the sequences are first considered in calculating the matrix. Now closest taxa are considered as a group. Again matrix is calculated considering this group as a node, subsequent to which taxa with minimum distance are considered as a group. Now matrix is calculated again and so on...continue till only two groups are formed and connect them also. UPGMA assumes that rate of nucleotide or amino acid substitution is constant due to which branch length reflects actual dates of divergence. This assumption is often not true hence can produce an inaccurate tree. Midpoint rooting is applied in this method.

Neighbor joining method

It allows different rates of evolution in different branches of tree. It starts with connecting OTU's with minimum distance and the node thus created is used for subsequent calculation. The tree is not rooted because it does not assume a constant rate of evolution but can be rooted using an out group.

Corrections: Observed distances are not always a good measure of evolutionary distance. Because they do not take into account hidden changes due to multiple hits. Due to this reason converting a measure of distance to a measure of evolution requires correction. Two such common corrections are Jukes–Cantor and Kimura-2 parameter models. The Jukes-Cantor one parameter model considers that each nucleotide is free to convert to others with equal rates for transition and transversion hence any nucleotide has equaled chance to covert to other three. It also assumes that four bases are present in equal frequencies.

Usually, transition rate is higher than transversion rate. Kimura two parameter models adjust pairwise distances taking into account the transition transversion ratio. Various other models have been developed that are more sophisticated.

Maximum Parsimony

Parsimony based method work on the principle of choosing the most parsimonious tree. The maximum parsimony works on the idea of minimizing the number of evolutionary changes. It works as follows:

- Identify informative sites in a dataset. Sites which represent alternative possibilities for OTU's are considered informative.
- Construct trees. All possible trees are constructed and evaluated. Score is based on number of evolutionary changes required to generate the particular tree.

- The trees with minimum score are retained. It is possible to retain more than one tree if they have equal minimum score.

Statistical methods of phylogeny

Distance and Maximum parsimony method are often criticized for lack of a statistical approach. Both these methods do have criteria to select trees but are unable to calculate the probability of one tree being the true tree over the other. Various methods have been proposed to overcome this drawback. Two such methods are provided by likelihood and Bayesian approaches.

In simplistic terms, likelihood can be considered as the probability assigned to each dataset (observed characters such as nucleotides) generated for a particular hypothesis (tree and model of evolution). In a way this is similar to maximum parsimony because each tree is assigned a score, but this score is a likelihood score based on statistical analysis. The best tree is the one, which has highest probability for a particular model of how changes occur. Both maximum parsimony and maximum likelihood are computationally exhaustive exercise and hence are slow. A detailed discussion about likelihood can be found in referenced text books.

Another statistical method for phylogeny is Bayesian method. In maximum likelihood we calculate the probability of observing data for a given hypothesis, in Bayesian method, probability is calculated for a particular hypothesis.

4.7 CONSISTENCY OF MOLECULAR PHYLOGENETIC PREDICTION

What is phylogenetic consistency?

A phylogenetic method is a consistent estimator of phylogeny if and only if it is guaranteed to give the correct tree, given that sufficient (possibly infinite) inde- pendent data are examined.

How do you calculate the consistency index of a phylogeny?

It is calculated by dividing the minimum possible number of steps by the observed number of steps. If the minimum number of steps is the same as the observed number of steps, then the character will have a CI of 1.0.

What makes molecular phylogeny inconsistent?

Possible reasons for these inconsistencies are: disparities in evolutionary rates among lineages. uneven taxonomic sampling. single explosive radiation of major eukaryotic taxa.

4.8 APPLICATIONS OF BIOINFORMATICS

Introduction

Bioinformatics is the application of IT to address a biological data. Bioinformatics helps us in understanding biological processes and involves development and application of computational techniques to analyse and interpret a biological problem. Major research efforts in the area of bioinformatics and computational biology include sequence alignment, genome annotation, prediction of protein structure and drug discovery.

Bioinformatics and Genomics

Genome and Genome Sequencing Projects

The word "genome" was coined by Hans Winkler from the German “Genome” in as early as 1926. The total DNA present in a given cell is called genome. In most cells, the genome is packed into two sets of chromosomes, one set from maternal and another one set from paternal inheritance. These chromosomes are composed of 3 billion base pairs of DNA. The four nucleotides (letters) that make up DNA are A, T, G, and C. Just like the alphabets in a sentence in a book make words to tell a story, same do letters of the four bases – A, T, G, C in our genomes.

Genomics is the study of the genomes that make up the genetic material of organism. Genome studies include sequencing of the complete DNA sequence in a genome and also include gene annotation for understanding the structural and functional aspects of the genome. Genes are the parts of your genome that carry instructions to make the molecules, such as proteins that are responsible for both structural and functional aspects of our cells. The first organism that was completely sequenced was *Haemophilus influenzae* in 1995 that led to sequencing of many more organisms from both prokaryotic and eukaryotic world.

The Human Genome Project (HGP)

The Human Genome Project (HGP) was global effort undertaken by the U.S. Department of Energy and the National Institutes of Health with a primary goal of determining the complete genome sequence in a human cell. It also aimed at identifying and mapping the genes and the non-genes regions in the human genome.

Some key findings of the draft (2001) and complete (2004) human genome sequences included-

1. Total number of genes in a human genome was estimated to be around 20,500.
2. Gene expression studies helped us in understanding some diseases and disorders in man.
3. Identification of primate specific genes in the human genome.
4. Identification of some vertebrate specific protein families.
5. The role of junk DNA was being elucidated.
6. It is estimated that only 483 targets in the human body accounted for all the pharmaceutical drugs in the global market.

How was the whole genome sequenced?

The human genome was sequenced by two different methods Hierarchical Genome Shotgun (HGS) Sequencing and Whole Genome Sequencing (WGS).

Why do we want to determine the sequence of DNA of an organism?

1. Genome variation among individuals in a population can lead to new ways to diagnose, treat, and someday prevent the thousands of disorders that affect mankind.

2. Genome studies help us to provide insights into understanding human disease biology.

3. Studies on nonhuman organisms' DNA sequences can contribute to solving challenges in health care and agriculture. Understanding the sequence of genomes can provide insights in the identification of unique and critical genes involved in the pathogenesis of microorganisms that invade us and can help identifying novel drug targets to offer new therapeutic interventions. Increasing knowledge about genomes of plants can reduce costs in agriculture, for example, by reducing the need for pesticides or by identification of factors for development of plants under stress.

4. HGP studies also included application of research on the ethical, legal and social implications (ELSI) of the genomic research for individuals and communities.

Where are the genome data stored?

The genome sequence and the genes mapped are stored in databases available freely in the Internet. The National Centre for Biotechnology Information (NCBI) is a repository of the gene/protein sequences and stores in databases like GenBank. This large volume of biological data is then analyzed using computer programs specially written to study the structural and functional aspects of genome.

Prediction Methods

Computational approaches for prediction of genes is one of the major areas of research in bioinformatics. Finding genes by the traditional molecular biology techniques becomes time consuming process. Two classes of prediction methods for identifying genes from non-genes in a genome are generally adopted: similarity or homology based searches and ab initio prediction.

Gene discovery in prokaryotic genomes becomes less time consuming as compared to prediction of protein coding regions in higher eukaryotic organisms due to the absence of intervening sequences called introns.

Comparative Genomics and Functional Genomics

Comparative genomics is the analysis and comparison of genomes from two or more different organisms. Comparative genomics is studied to gain a better understanding of how a species has evolved and to study phylogenetic relationships among different organisms.

One of the most widely used sequence similarity tool made available in the public domain is Basic Local Alignment Search Tool (BLAST). BLAST is a set of programs designed to perform sequence alignment on a pair of sequences (both nucleotide and protein sequence).

Functional genomics attempts to study gene functions and interactions. Functional genomics seeks to address questions about the function of DNA at the levels of genes, RNA at the levels of transcription and proteins at the structural and functional levels.

Pharmacogenomics

Pharmacogenomics analyzes how the genetic constitution affects a person's response to drugs and help us in the creation of personalized medicine to create and design drugs based on an individual's unique genetic makeup. Pharmacogenomics is used for the development of drugs to treat a wide range of health problems including diabetes, cancer, cardiovascular disorders, HIV, and tuberculosis.

Next Generation Sequencing

The advancement of the field of molecular biology has been principally due to the capability to sequence DNA. Over the past eight years, massively parallel sequencing platforms have transformed the field by reducing the sequencing cost by more than two folds. Previously, Sanger sequencing ('first-generation' sequencing technology) has been the sole conventional technique used to sequence genomes of several organisms. In contrast, NGS platforms rely on high-throughput massively parallel sequencing involving unison sequencing of millions of DNA fragments from a single sample. The former facilitates the sequencing of an entire genome in less than a day. The speed, accessibility and the cost of newer sequencing technologies have accelerated the present -day biomedical research.

These technologies reveal large scale applications outspreading even genomic sequencing. The most regularly used NGS platforms in research and diagnostic labs today have been-the Life Technologies Ion Torrent Personal Genome Machine (PGM), the IlluminaMiSeq, and the Roche 454 Genome Sequencer. NGS platforms rapidly generate sequencing read data on the gigabase scale. So the NGS data analysis poses the major challenge as it can be time-consuming and require advanced skill to extract the maximum accurate information from sequence data. A massive computational effort is needed along with in-depth biological knowledge to interpret enormous NGS data.

Bioinformatics and Protein Structure Prediction

Proteins are linear polymer of amino acids joined by peptide bonds. Every protein adopts a unique three-dimensional structure to form a native state. It is this native 3D structure that confers the protein to carry out its biological activity. Proteins play key roles in almost all the biological process in a cell. Proteins are important for the maintenance and structural integrity of cell.

Levels of protein architecture

There are four levels of protein structure. The primary structure of a protein is the arrangement of linear sequence of amino acids. The patterns of local conformation within the polypeptide are referred to as as secondary structure. The two most common types of secondary structure occurring in proteins are α-helices and β-sheets. These secondary structures are connected by loop regions. The tertiary structure represents the overall three dimensional structure of these elements and the protein folds into its native state. The quaternary structure includes the structure of a multimeric protein and interaction of its subunits. Figure illustrates the hierarchy in protein structure.

Explosion in the growth of Biological Sequence and Structure Data

Experimental determination of the tertiary structure of proteins involves the use of X-ray crystallography and NMR. In addition, computational techniques are exploited for the structural prediction of native structures of proteins. There has been an exponential growth of both the biological sequence and structure data, mainly due to the genome sequencing projects underway in different countries around the world. As of Oct 2013, there are 94,540 structures in the protein data bank (RCSB-PDB).

Computational approaches to protein structure prediction

There are three different methods of protein 3D structure prediction using computational approaches

1. Comparative Protein Modeling or Homology Modeling

Homology modeling predicts the structure of a protein based on the assumption that homologous proteins share very similar structure, as during the course of evolution, structures are more conserved than amino acid sequences. So a model is generated

based on the good alignment between query sequence and the template. In general we can predict a model when sequence identity is more than 30%. Highly homologous sequences will generate a more accurate model.

2. Protein Threading

If two sequences show no detectable sequence similarity, threading or fold recognition is employed to model a protein. Threading predicts the structure for a protein by matching its sequence to each member of a library of known folds and seeing if there is a statistically significant fit with any of them.

3. Ab initio method

Ab initio protein modeling is a database independent approach based exploring the physical properties of amino acids rather than previously solved structure. Ab-initio modeling takes into consideration that a protein native structure has minimum global free energy.

4.9 SUMMARY

The National Center for Biotechnology Information (NCBI), established in 1988 has emerged as one of the largest repositories of biological data and related literature from a diverse range of organisms. This enormous amount of information available at NCBI relies on individual researchers and consortia (i.e. a collaborative effort by a group of individuals/institutes) for submission of several forms of datasets. The various forms of data have been further categorized as sequence based and non-sequence based such as microarray, phenotype, chemical structures etc. The different forms of datasets can be submitted using the appropriate submission tools;

tools such as BankIt, Sequin and Barcode are meant for submission of nucleotide data. Whereas Sequin is a stand-alone sequence submission tool, BankIt and Barcode are web-based; In addition web based tools can also be used for deposition of Whole /Complete Genome, trace files and Short Read archive nucleotide data; Gene expression data, RT-PCR data, mRNA sequence data, Chromatin-Immuno-precipitation (ChIP) and methylation sequencing data are submitted using the GEO web-deposit gateway. Along with the datasets, the depositor must also submit experimental details and designs. Upon acceptance of datasets (nucleotide, gene expression, chemical structures etc) after strict quality control and verification, the team at NCBI assigns a unique number or Identifier termed as Accession number. The datasets are organized and catalogued in the most appropriate database and can be accessed using keywords of the accession number.

Molecular Phylogeny is to study evolutionary relationships based on molecular sequence data. Different methods have been proposed for studying phylogeny. Earlier methods were distance based and considered constant evolutionary rates. These methods used more exhaustive and computationally exhaustive methods like maximum parsimony. These methods are now being supplemented or replaced with more sophisticated statistical methods like maximum likelihood and Bayesian method. The benefits and pitfalls of these methods are still debated and their applicability may depend upon the situation. A basic understanding of these methods is a must for effective use of them for reconstructing phylogeny.

1. Bioinformatics is application of IT to address biological problems.

2. Bioinformatics and its related fields like Genomics, Proteomics, Transcriptomics, Metabolomics and Systems biology finds useful applications in agriculture, health sector and environmental issues.

3. The three major thrust areas of research include genome and transcriptome and proteome analysis, protein structure prediction and computer aided drug design.

4. Many softwares/tools are being developed and are available freely over the internet to locate genes in a genome and predict structures of protein.

5. Bioinformatics and computational biology help in reducing the cost and time for designing new drugs and are nowadays routinely now used in pharmaceutical companies.

4.10 TERMINAL QUESTIONS AND ANSWERS

1. What tools would you use to submit sequences?

2. Prepare a list of databases dealing with literature and their characteristic features.

3. Sequence data are deposited in which databases?

4. Compare the features of two sequence submission tools, BankIT and Sequin.

5. How will you differentiate a dendrogram from a cladogram?

6. What is the difference between a distance based method (NJ) and maximum parsimony (MP) methods?

7. What is the difference between UPGMA and NJ method?

8. Differentiate between maximum parsimony (MP) and maximum likelihood method (ML).

9. Discuss the major research areas in the field of bioinformatics

10. What is the difference between the gene organization in prokaryotes and eukaryotes?

11. Differentiate between comparative genomics and functional genomics

12. What is pharmacogenomics?

REFERENCES

- Rastogi SC, Mendiratta N, Rastogi P (2011) Bioinformatics: Concepts, Skills & Applications. CBS Publishers & Distributors Pvt. Ltd. ISBN: 81-239-1482-2.
- Mount DM (2004) Bioinformatics: Sequence and Genome Analysis 2. Cold Spring Harbor Laboratory Press. ISBN: 0-87969-712-1.

3. Ghosh Z, Mallick B (2012) Bioinformatics: Principles and Applications. Oxford University Press. ISBN-13: 978-0-19-569230-3.

4. Campbell AM, Heyer LJ (2006) Discovering Genomics, Proteomics, and Bioinformatics. CSHL Press. ISBN: 0-8053-8219-4.

5. Young DC (2009) Computational Drug Design. John Wiley & Sons, Inc. ISBN: 978-0-470-12685-1.

6. Tramontano A (2006) Protein structure Prediction: Concepts and Applications. WILEY-VCH. ISBN-13: 978-3-527-31167-5.

7. Hiroaki Kitano (2001) Foundation of Systems Biology. MIT Press. ISBN: 0-262-11266-3.

8. Bioinformatics and Functional Genomics: 2nd Edition, Jonathon Pevsner (2009), Wiley Blackwell

UNIT – 5: INTRODUCTION TO BIOSTATISTICS

Contents

5.1 OBJECTIVES

To study

- Statistical symbol
- Scope & applications
- Collection, organization and representation of data
- Importance of statistics in biological research

5.2 INTRODUCTION

Statistics is the science of figures which deals with collection, analysis and interpretation of data. Data is obtained by conducting a survey or an experiment study. The use of statistics in biology is known as Biostatistics or biometry.

Purpose and scope of statistics: The purpose of statistics is not only to collect numerical data but is to provide a methodology for handling, analysing and drawing valid inferences from the data. It has wide application in almost all sciences–social as well as physical such as biology, psychology, education, economics, planning, business management, mathematics etc.

SOME IMPORTANT STATISTICAL TERMS AND NOTATIONS

While studying various aspects of problems of statistics one has to come across several statistical terms. Few important statistical terms are given below:

1. Population: The popular idea of population is universe. But statistician's idea of population is quite different from the popular idea. Biometric study regards the population of some limited region as its universe. The population in a statisticalinvestigation refers to any well-defined group of individuals or of observations of a particular type. In short one can say that a group of study element is called population. For example all fishes of one species present in a particular pond could be a population. All patients of a hospital suffering from AIDS may be considered as population while few patients are used as study elements.

2.Sample: In case of large population, it becomes practically impossible to collect data from all the members. In order to study the Haemoglobin percentage (Hb %) of patients of a hospital, it will be more convenient and quicker to collect data from few patients. Here patient taken for study are sample.

Sample may be defined as fraction of a population drawn by using a suitable method so that it can be regarded as representative of the entire population.

3. Variable: In everyday life, we come across living beings and phenomena, which vary in a number of ways, even though they belong to the same general category or type. Measurement of characteristics is called variable.

Animals of some species may differ in their length, weight, age, sex, Hb %, YO2 intake, fecundity (Rate of reproduction), RBCs count, habits, personality traits etc. The above mentioned characteristics on which individuals differ among themselves are called a variable'. Variables may be of two types:

(a) Quantitative variable: Whenever the measurement of characteristics is possible on a scale in some appropriate units, it is called a quantitative variable. Examples of quantitative variables are measurement of length, weight, age, intellectual ability etc. Quantitative variables can be further sub –divided into two types:

(i) Discrete or discontinuous variable and, (ii) Continuous variable.

(i) Discrete or discontinuous variable is one where the values of the variables differ from one another by definite amounts, i.e., these vary only by finite 'jumps' or 'breaks'. For example the number of persons in a family or number of fish in a pond

(ii) Continuous variable can assume all values within a certain interval and as such are divisible into smaller and smaller fractional units. Thus values of a continuous variable have no 'breaks' or 'jumps'. Measurement of length, weight, Hb %, VO2 consumption, intelligence quotient (I.Q.) etc. is some examples of a continuous variable.

(b) Qualitative variable: It is unmeasurable variable and is unexpressible in magnitudes. But it can be expressed in quality. These qualities are called attributes. Colour of flower or animal, wrinkled seeds or smooth seeds etc., are examples of a qualities are called attributes. Colour of flower or animal, wrinkled seeds or smooth seeds etc. are examples of a qualitative variable.

4. Parameter: The numerical quantities which characterise a population (in respect of any variable) are called parameters of the population. For example, if the characteristic is length and a measurement of length is variable then the mean length can be regarded as a parameter. Usually all the important characteristics of a population can be specified in terms of a few parameters.

5. Statistics: Description of the properties of a population in terms of its parameters can be done with the help of statistical methods.

The term statistics is used to denote summary value of any quantity that is calculated from sample data. A statistics that serves as an estimate of the parameter, population mean

6. Observation: Measurement of an event is only possible by observation. For example Hb % in any animal is an event while 14 g/100c.c, is a measurement and these are observing experiments.

7. Data: A set of values recorded on one or more observational unit is called data. First step of statistical study is the collection of data. In scientific research work data is collected only from personal experimental study. Data collected by personal investigation is called primary data.

5.3 STATISTICAL SYMBOLS

Some of the statistical symbols which are useful to biostatistics students are:

f: Frequency of the variate.

$\bar{X}$: Arithmetic Mean of a given set of values or of a distribution

Me: Median of a given set of values or of a distribution

Mo: Mode of a given set of values or of a distribution

σ:Standard Deviation of a given set of values or of a distribution

2 σ: Variance of a given set of values or of a distribution

δ: Mean deviation

x: Deviation

c: Correction

Σ: Stands for summation of observations, sum of all the values of a given set

df: Degree of freedom

O: Observed number

E: expected number

P: Probability

%: Per cent

w: Assumed mean

i: Length of class interval

Q: Quartile deviation

5.4 SCOPE & APPLICATIONS

The scope of statistics is not only to collect numerical data but is to provide a methodology for handling, analysing and drawing valid inferences from the data. It has wide application in almost all sciences -social as well as physical such as biology, psychology, education, economics, planning, business management, mathematics etc.

Applications of Biostatistics

In Anatomy and Physiology

- To define what is normal or healthy in a population.
- To find the limits of normality in variables such as weight and pulse rate etc. in a population.
- To find the correlation between two variables such as height and weight (weight increases with increase in height).

In Pharmacology

- To find the action of drug on human– A drug is given to humans to check whether the changes produced are due to the drug or by chance.
- To compare the action of two different drugs or two successive dosages of the same drug.

- To find the relative potency of a new drug with respect to a standard drug.

In Medicine

- To compare the efficacy of a particular drug, operation or line of treatment- for this, the percentage cured, relieved or died in the experiment and control groups, is compared and difference due to chance or otherwise is found by applying statistical techniques.
- To find correlation between two attributes such as cancer and smoking or filariasis and social class.
- To identify signs and symptoms of a disease or syndrome.
- Cough in typhoid is found by chance and fever is found in almost every case.
- To test usefulness of vaccines in the field- Percentage of attacks or deaths among the vaccinated subjects is compared with that among the unvaccinated ones to find whether the difference observed is statistically significant.

Clinical medicine

- Documentation of medical history of diseases.
- Planning and conduct of clinical studies.
- Evaluating the merits of different procedures.
- In providing methods for definition of 'normal' and 'abnormal'.

Preventive medicine:

- To provide the magnitude of any health problem in the community.
- To find out the basic factors underlying the ill- health.
- To evaluate the health programs which was introduced in the community (success/failure)?
- To introduce and promote health legislation.

In Health Planning and Evaluation:

- The methods used in dealing with statistics in the fields of medicine, biology and public health for planning, conducting and analyzing data.
- In carrying out a valid and reliable health situation analysis, including in proper summarization and interpretation of data.

- In proper evaluation of the achievements and failures of a health programs.

In Biotechnology

- Study of genetic modification of plants, and animals to gene therapy,
- Medicine and drug manufacturing,
- Reproductive therapy
- Energy production
- In all these cases, research is carried out and testing whether or not it has the desired performance.

In Community Medicine and Public Health:

- To evaluate the efficacy of vaccines.
- In epidemiological studies-the role of causative factors is statistically tested.
- To test whether the difference between two populations is real or a chance occurrence.
- To study the correlation between attributes in the same population.
- To measure the morbidity and mortality.
- To evaluate achievements of public health programs.
- To fix priorities in public health programs.
- To help promote health legislation and create administrative standards for oral health.
- It helps in compilation of data, drawing conclusions and making recommendations.
- To test the usefulness of vaccines in the field–the percentage of attacks or deaths among the vaccinated subjects is compared with that among the non-vaccinated ones to find whether the difference is observed as statistically significant.
- In epidemiological studies –the role of causative factors is statistically tested. The deficiency of iodine as an important cause of goitre in a community is confirmed only after comparing the incidence of goitre cases before and after giving iodized salt.

In Genetics

- Statistical and probabilistic methods are now central to many aspects of analysis of questions is human genetics.

- The find an extensive applications of statistical methods in human genetics is * Human Genome Project * Linkage Analysis * Sequencing.
- Analysis of DNA, RNA, protein, low- molecular-weight metabolites, as well as access to bioinformatics databases.

In Dental Science

- To find the statistical difference between means of two groups. Ex: Mean plaque scores of two groups.
- To assess the state of oral health in the community and to determine the availability and utilization of dental care facilities.
- To indicate the basic factors underlying the state of oral health by diagnosing the community and find solutions to such problems.
- To determine success or failure of specific oral health care programs or to evaluate theprogram action.
- To promote oral health legislation and in creating administrative standards for oral health care delivery.

In Environmental Science

- Baseline studies to document the present state of an environment to provide background in case of unknown changes in the future.
- Targeted studies to describe the likely impact of changes being planned or of accidental occurrences.
- Regular monitoring to attempt to detect changes in the environment.

5.5 COLLECTION, ORGANIZATION AND REPRESENTATION OF DATA

Collection of data:

Statistical data is a set of facts expressed in quantitative form. The data can be obtained through primary sources or secondary source. Data obtained by the investigator from personal experimental study is called primary data.

If the data is obtained from secondary sources such as, journals, magazines, paper, etc. it is known as secondary data. In scientific work only primary data are used.

Primary Data Collection Methods:

Primary data obtained directly from the first-hand source through experiments, surveys or observations. The primary data collection method is further classified into two types:

- Quantitative Data Collection Methods
- Qualitative Data Collection Methods

Quantitative Data Collection Methods

This method is based on mathematical calculations using mean, median or mode measures,close-ended questions, correlation and regression method.

- It is cheaper than qualitative data collection method.
- It can be applied in a short duration of time.

Qualitative Data Collection Methods:

It does not involve any mathematical calculations. This method is closely associated with elements that are not quantifiable. This qualitative data collection method includes interviews, questionnaires, observations, case studies, etc. There are several methods to collect this type of data. They are

Observation Method

Observation method is used when the study related to behavioral science. This method is planned systematically. It is subject to many controls and checks.

The different types of observations are:

- Structured and unstructured observation
- Controlled and uncontrolled observation
- Participant, non-participant and disguised observation

Interview Method

The method of collecting data verbally, it consists of:

- **Personal Interview** – In this method, an interviewer is required to ask questions face to face to the other person. The personal interview can be structured or unstructured, direct investigation, focused conversation, etc.
- **Telephonic Interview** –In this method, an interviewer obtains information by contacting people on the telephone to ask the questions or views, verbally.

Questionnaire Method

In this method, the set of questions are mailed to the respondent. They should read, reply and subsequently return the questionnaire. The questions are printed in the definite order on the form. A good survey should have the following features:

- Short and simple
- Should follow a logical sequence
- Provide adequate space for answers
- Avoid technical terms
- Should have good physical appearance such as colour, quality of the paper to attract the attention of the respondent

Schedules

This method is slightly different from the questionnaire method. The enumerators are specially appointed for the purpose of filling the schedules. It explains the aims and objects of the investigation and may remove misunderstandings, if any have come up. Enumerators should be trained to perform their job with hard work and patience.

Secondary data collection method:

Secondary data is collected by someone other than the actual user. It means that the information is already available, and someone analyses it. The secondary data comprised of magazines, newspapers, books, journals, etc.

It can be either published data or unpublished data.

Published data include:

- Government publications

- Public records
- Historical and statistical documents
- Business documents
- Technical and trade journals

Unpublished data include:

- Diaries
- Letters
- Unpublished biographies, etc.

Presentation of data:

Data obtained by the researcher can be displayed in tabular form, diagrams and through charts. Display of data in tabular form, diagrams and through charts. Display of data in tabular form is called classification of data and through charts is known as charting of data.

Process to arrange and present primary data in a systematic way is called classification of data. Data may be grouped or classified in following various ways:

(i) **Geographical;** i.e., according to area or region. If we take into account production of fish or lac or silk state wise, this would be called geographical classification.

(ii) **Chronological;** i.e., according to occurrence of an event in time.

Egg production of a poultry farm for five years are given below which is an example of chronological classification:

Year	Egg Production
95-96	1590
96-97	1672
97-98	1882
98-99	1961
99-2000	2233

(iii) **Qualitative;** i.e., according to attributes or quality. For example, if a species of fish in a pond is to be classified in respect to one attribute say sex, we can classify them into two groups. One is of males and other is of females.

When the classification is done with respect to one attribute, which is simple or dichotomous in nature, two classes are formed, one possessing the attribute and the other not possessing the attribute. This type of qualitative classification is called simple or dichotomous classification.

When we classify fishes simultaneously with respect to two attributes, *i.e,* sex and infected condition, then fishes are first classified with respect to 'sex' into 'males' and 'females'. Each of these classes may then be further sub-divided into 'infected' and 'uninfected'. Thus the attribute sex and condition infection in fishes are classified into four classes, namely– (a) Male uninfected, (b) Male infected, (c) Female uninfected, (d) Female infected. The classification, where two or more attributes are considered and several classes are formed is called a manifold classification.

(IV) Quantitative; i.e., according to magnitudes. For example, the thickness of a plant may be classified according to their growth rate. Quantitative data may be of two types:

(a) Continuous data: It covers all values of a variable. Hb % of a person can be expressed in any values such as 13 mg/100 c.c., 13.1 mg/100c.c. and so on. Water percentage in the body of a species may be 65 %, 65.1 %, 65.2 %, 65.3 % and so on.

(b) Discrete data: The term discrete data is limited to discontinuous numerical values of a variable. It can be done only in whole number. For example number of persons in a family or number of books in a library can be said only in whole number. One can't say that there are 4 ½ (Four and half) persons in my family or there are 500 ½ books in this library.

Preparation of frequency distribution table:

Quantitative data is grouped or classified and presented in the form of a frequency distribution table. The frequency distribution table presents the quantitative data very concisely indicating the number of repetition of observations. It records how frequently a variable occurs in a group study.

Following raw data is obtained in an investigation. 100 pea plants bore pods ranging from 15 to 41 in a garden of pea plants.

Raw Data Table A:

33, 31, 28, 15, 17, 17, 16, 18, 16, 18, 20, 22, 24, 25, 31, 27, 30, 29, 33, 28, 20, 22, 23, 25, 41, 39, 30, 36, 37, 27, 33, 28, 31, 29, 32, 31, 29, 34, 19, 22, 25, 40, 19, 21, 24, 30, 26, 37, 27, 28, 32, 32, 31, 29, 34, 21, 23, 25, 40, 26, 38, 27, 26, 33, 28, 34, 29, 30, 30, 35, 29, 23, 29, 26, 38, 27, 32, 28, 34, 35, 29, 30, 33, 32, 35, 29, 24, 26, 38, 27, 36, 28, 34, 29, 35, 30, 33, 32, 36, 37.

Raw Data Table B:

15, 16, 17, 17, 18, 18, 19, 19, 20, 20, 21, 21, 22, 22, 22, 23, 23, 23, 24, 24, 24, 25, 25, 25, 26, 26, 26, 26, 27, 27, 27, 27, 27, 27, 28, 28, 28, 28, 28, 28, 28, 29, 29, 29, 29, 29, 29, 29, 29, 29, 30, 30, 30, 30, 30, 30, 30, 31, 31, 31, 31, 31, 32, 32, 32, 32, 32, 32, 33, 33, 33, 33, 33, 33, 34, 34, 34, 34, 34, 35, 35, 35, 35, 35, 36, 36, 36, 37, 37, 37, 38, 38, 38, 39, 39, 40, 40, 41.

Our first step in the preparation of frequency distribution table is to arrange them in ascending order of magnitude. The data is then said to be in array. The above raw data table A is arranged in ascending order of magnitude as shown in raw data table B.

Steps for the preparation of a discrete frequency distribution table may be taken as follows:

A table of two columns is prepared. First column contains variables and second column contains repetition number of variable i.e. frequency of variables.

In above data variable 15 is obtained only once. Therefore frequency 1 is mentioned against variable 15. Variable 16 is obtained twice; therefore, frequency 2 is mentioned against this variable. In the same fashion frequencies of all variables of above data are mentioned and a frequency distribution table 1.1 is obtained.

Table 1

No. of Pods **(Variables)**	**No. of Plants** **(Frequency)**	**No of Pods** **(Variables)**	**No. of Plants** **(Frequency)**

15	1	29	9
16	2	30	7
17	2	31	5
18	2	32	6
19	2	33	6
20	2	34	5
21	2	35	4
22	3	36	3
23	3	37	3
24	3	38	3
25	4	39	2
26	5	40	2
27	6	41	1
28	7		

For convenience discrete frequency table may be prepared with the help of tally mark. Following steps have to be taken to prepare discrete frequency table using tally mark:

A table of three columns is prepared. In first column variables are mentioned. In second column repetition (frequency) of each variable is denoted by tally mark. In third column, total of tally mark, of each variable is written which is of course the frequency of variable.

If variable appears only once then tally mark I is mentioned, for second repetition II, for third III but for fifth a cut of fourth IV is mentioned.

Following simple frequency table 1.2 is prepared using raw data B in array with the help of tally mark.

Table 1.2

No. of Pods or Variable	Tally mark	Repetition number of Plants of frequency
15	I	1
16	II	2
17	II	2
18	II	2
19	II	2
20	II	2
21	II	2
22	III	3
23	III	3
24	III	3
25	IIII	4
26	~~IIII~~	5
27	~~IIII~~ I	6
28	~~IIII~~ II	7
29	~~IIII~~ IIII	9
30	~~IIII~~ II	7
31	~~IIII~~	5
32	~~IIII~~ I	6
33	~~IIII~~ I	6
34	~~IIII~~	5

35	IIII	4
36	III	3
37	III	3
38	III	3
39	II	2
40	II	2
41	I	1

Preparation of frequency distribution table in class – interval:

What is class interval and how it is prepared?

To make data comprehensible one should classify or group identical values of the variables into ordered class intervals.

To illustrate, the construction of a frequency distribution table in class interval, consider the raw data B, which represents the pods per plant in a garden.

Here we first decide about the number of classes into which data are to be grouped. Ordinarily, the number of classes should be between 5 and 20, but this may be done arbitrarily. The number of classes depends on the number of observations– with larger number of observations – with larger number of observations one can have more classes.

The width or range of class is usually called class-interval and is denoted by h. The width of class-interval must be of uniform size.

After deciding about class-interval we calculate range (The highest score H minus lowest score L or length of class interval) (H-L). From Raw data B, Range of score R = 41-15 = 26 (Range is denoted by R).

Now following formula may be applied to get the approximate number of classes which should expect to group the given observations.

Number of classes k = Range of scores / Class interval = R/h.

Mid-point of class interval: Class mid-point is the sum of highest and lowest limits of class-interval divided by two. Thus the mid-point falls in the middle of upper and lower level of class-interval.

Class mid-point = Highest limit of C.I. + Lowest limit of C.I. / 2

For example mid – point of a class interval 10-20 may be as follows:

Mid-point of C.I. = 20 + 10/ 2 = 30/2 =15.

5.6 IMPORTANCE OF STATISTICS IN BIOLOGICAL RESEARCH

Biostatistics is the application of statistics in different fields of biology:

- Biostatistics includes the design of biological experiments, especially in medicine, pharmacy, agriculture, forestry, environmental science, fishery etc., the collection, summarization, and analysis of data from those experiments; and executes interpretation and inference from the results.
- A major branch of this is medical biostatistics, which is exclusively concerned with health and medical sciences. In modern world, the scope of biostatistics is increasing rapidly. Almost all educational programmes in biostatistics are at postgraduate level. They are most often found in schools of public health, affiliated with schools of medicine, forestry, or agriculture, or as a focus of application in departments of statistics. In larger universities where both a statistics and a biostatistics department exist, the degree of integration between the two departments may range from the bare minimum to very close collaboration.

5.7 SUMMARY

Biostatistics is the application of statistical principles to questions and problems in medicine, public health or biology. In other circumstances in would be important to make comparisons among groups of subjects in order to determine whether certain behaviour's (e.g., smoking, exercise, etc.) are associated with a greater risk of certain health outcomes. It would, of course, be impossible to answer all such questions by collecting information (data) from all subjects in the populations of interest. A more realistic approach is to study samples or subsets of a population. The discipline of biostatistics provides tools and techniques for collecting data and then summarizing, analysing, and interpreting it. Consequently, in biostatistics one analyses samples in order to make inferences about the population. This module introduces fundamental concepts and definitions for biostatistics. A survey research can be objectivist or subjectivist in nature. An objectivist approach is a more rigid and scientific approach. In this the hypothesis is tested using publicly standard procedure. There is little or no latitude available to deviate from the stated procedures or questions. Data Analysis and Data Presentation have a practical implementation in every field. The transformed raw data assists in obtaining useful information. The presentation is the key to success. Once the information is obtained the user transforms the data into a pictorial presentation so as to be able to acquire a better response and outcome.

5.8 TERMINAL QUESTION AND ANSWERS

Question1. Define, explain and mention uses of biometry.

Question2. What do you mean by data, population, sample, variable, parameter, class interval, frequency distribution, cumulative frequency distribution, primary data, and secondary data?

Question3. The lowest and highest levels of few class intervals are given below. Mention length and mid-points of each class interval:

40-50, 32-44, 20-32, 10-22, 20-30, 30-39, 40-49, 40-52, 53-64.

Question4. Differentiate between primary data and secondary data.

Question5. Write short note on following:

A) Class interval

B) Interview method

C) Schedule

D) Qualitative data

E) Quantitative data

REFERENCES

- Mahajan BK 2002 (Methods in Biostatistics) (6th edition)
- Zaman SM, HK Rahim and M Howlader 1982. (Simple Lessons from Biometry), BRRI
- Research methodology methods and techniques, C. R, Kothari, New Age International Limited Publisher,

UNIT 6: MEASURES OF CENTRAL TENDENCY AND VARIABILITY

CONTENTS

6.1 OBJECTIVES

To study measures of central tendency and variability

- Mean, median and mode
- Mean deviation
- Standard deviation and error
- Variance and coefficient of variation
- Chi square test
- Student t Test

6.2 INTRODUCTION

Central tendency may be considered as a synonym of average. Average is a general term which describes the general value of series, around which all other observations are dispersed.

There are two types of central tendency.

1. Mathematical average

2. Average of positions

1. **Mathematical average:** Average represented mathematically is called mathematical average. There are three types of mathematical average:

- Arithmetic mean
- Geometric mean
- Harmonic mean

2. **Average of positions:** Average exhibited by position is called average of positions. There are two types of average positions.

- Median
- Mode

Median indicates the average position of a series. In a series all observations are arranged in ascending or descending order and the middle observation is called the median.

Mode is that value which is repeated maximum times in a series. In other words we can say that the mode is that value which has the maximum frequency.

Standard deviation is the most important and widely used measure of dispersion. It is denoted by a Greek letter sigma. The standard deviation defined as 'the square root of the arithmetic mean of the squared deviations of measurements from their mean. It has accordingly often been called the root mean square deviation.

Variance may be defined as "Square of sum of deviation divided by number of observations" or the square of standard deviation is termed as variance.

The Chi- square test was developed by Prof. A. R. Fisher in 1870. Karl Pearson improved Fisher's chi-square test in its modern form in 1900. Chi-square is derived from the Greek letter (chi χ) and pronounced as 'kye'.

Student's t-test is used not only to test the significance of difference between two means but also to test the significance of product moment correlation, point- biserial correlation, rank difference correlations etc.

Student's t test is also known as t-ratio because it is the ratio of difference between two means and standard error of difference between two means.

6.3 MEAN, MEDIAN AND MODE

In statistics, the mean is one of the measures of central tendency, apart from the mode and median. Mean is nothing but the average of the given set of observations. It denotes the equal distribution of values for a given data set.

Mean is the simple mathematical average of a set of two or more numbers. The mean for a given set of numbers can be computed in more than one way, which uses the sum of the numbers in the series.

Mean obtained arithmetically is called the arithmetic mean. Arithmetic means can be obtained both from grouped and ungrouped data.

Ungrouped data: Arithmetic mean is obtained by summing up all the observations and dividing it by total number of observations.

Arithmetic mean = Sum of all the observation ÷ Total number of observation

$$\bar{X} = (x_1 + x_2 + x_3 + \ldots. + x_n) / n$$

To calculate the arithmetic mean of a set of data we must first add up (sum) all of the data values (x) and then divide the result by the number of values (n). Since ∑ is the symbol used to indicate that values are to be summed (see Sigma Notation) we obtain the following formula for the mean ($\bar{x}$):

$\bar{X} = \sum x/n$

Here, $\bar{X}$ **= Mean**

∑ = Sigma

X = observations

N = Number of observations

Example: In a class there are 20 students and they have secured a percentage of 88, 82, 88, 85, 84, 80, 81, 82, 83, 85, 84, 74, 75, 76, 89, 90, 89, 80, 82, and 83.

Find the mean percentage obtained by the class.

Solution:

Mean = Total of percentage obtained by 20 students in class/Total number of students

= [88 + 82 + 88 + 85 + 84 + 80 + 81 + 82 + 83 + 85 + 84 + 74 + 75 + 76 + 89 + 90 + 89 + 80 + 82 + 83]/20

= 1660/20

= 83

Hence, the mean percentage of each student in the class is 83%.

Grouped data: When data is presented in frequency distribution, it can be obtained by two obtained.

Merits mean: Arithmetic mean is the most important measures of central tendency because

- It covers all the observations.
- It can be calculated easily and it expresses a simple relation between the whole and the parts
- It does not get affected by the fluctuations of sampling.

Demerits of mean:

- By observing data on graphs mean cannot be observed.
- Mean obtained by calculation may not be represented by any series.

Geometric mean: This is the central tendency of a set of data following a geometric progression. In case of 'n' number of the items the geometric mean is defined as the nth root of the product of 'n' items of series. Geometric mean is denoted by GM.

GM = $^n\sqrt{}$ ($a_1 \times a_2 \times ... \times a_n$)

Harmonic mean: Harmonic mean is the reciprocal of arithmetic mean of given observation.

= $1 / 1/n\ (1/x_1 + 1/x_2 + 1/x_3 + ….. +1/x_n)$

Average of position:

Median indicates the average position of a series. In a series all observations are arranged in ascending or descending order and the middle observation is called the median. The median is most suitable for expressing qualitative data such as colour, health, intelligence etc. Median is calculated differently for ungrouped and grouped data. Ungrouped data: Median of ungrouped data is calculated by two different methods: When scores are in odd number, formula to obtain median is as follows:

Median = $(\frac{n+1}{2})^{\text{th item}}$ When the data is continuous and in the form of a frequency distribution, the median is calculated through the following sequence of steps.

Step 1: Find the total number of observations (n).
Step 2: Define the class size (h), and divide the data into different classes.
Step 3: Calculate the cumulative frequency of each class.
Step 4: Identify the class in which the median falls. (Median Class is the class where n/2 lies.)
Step 5: Find the lower limit of the median class (l), and the cumulative frequency of the class preceding the median class (c).

Merits of Median:

- ✓ Median is a better indicator average than mean when one or more of the lowest or the highest observations are wide apart or not so evenly distributed.
- ✓ It can be calculated easily and can be exactly located.
- ✓ The value of the median is not influenced by abnormally large or small values or the change of any one value of the series.
- ✓ It can also be used in qualitative measures.

Demerits of Median:

- ✓ Arithmetic expression of median is not possible.
- ✓ To obtain median data must be kept in ascending and descending order.
- ✓ It gives equal importance to all series.

MODE:

Mode is that value which is repeated maximum times in a series. In other words we can say that the mode is that value which has the maximum frequency. Mode can be obtained by two methods: Determination of mode at a glance: The value which is repeated maximum times in a series is considered as mode. The value occurring most frequently in a set of observations is its mode. In other words, the mode of data is the observation having the highest frequency in a set of data. There is a possibility that more than one observation has the same frequency, i.e. a data set could have more than one mode. In such a case, the set of data is said to be multimodal.

In the case of grouped frequency distribution, calculation of mode just by looking into the frequency is not possible. To determine the mode of data in such cases we calculate the modal class. Mode lies inside the modal class. The mode of data is given by the formula:

$$Mode = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

Where,

l = lower limit of the modal class

h = size of the class interval

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class

Let us take an example to understand this clearly.

Let us learn here how to find the mode of a given data with the help of examples.

Example 1: Find the mode of the given data set: 3, 3, 6, 9, 15, 15, 15, 27, 27, 37, and 48.

Solution: In the following list of numbers,

3, 3, 6, 9, 15, 15, 15, 27, 27, 37, 48

15 is the mode since it is appearing more number of times in the set compared to other numbers.

Example 2: Find the mode of 4, 4, 4, 9, 15, 15, 15, 27, 37, 48 data sets.

Solution: Given: 4, 4, 4, 9, 15, 15, 15, 27, 37, 48 is the data set.

As we know, a data set or set of values can have more than one mode if more than one value occurs with equal frequency and number of time compared to the other values in the set.

Hence, here both the number 4 and 15 are modes of the set.

Example 3: Find the mode of 3, 6, 9, 16, 27, 37, and 48.

Solution: If no value or number in a data set appears more than once, then the set has no mode.

Hence, for set 3, 6, 9, 16, 27, 37, 48, there is no mode available.

Example 4: **In a class of 30 students marks obtained by students in mathematics out of 50 is tabulated as below. Calculate the mode of data given.**

Marks Obtained	Number of Student
10-20	5
20-30	12
30-40	8
40-50	5

Solution:

The maximum class frequency is 12 and the class interval corresponding to this frequency is 20 – 30. Thus, the modal class is 20 – 30.

Lower limit of the modal class (l) = 20

Size of the class interval (h) = 10

Frequency of the modal class (f_1) = 12

Frequency of the class preceding the modal class (f_0) = 5

Frequency of the class succeeding the modal class (f_2) = 8

Substituting these values in the formula we get;

$$Mode = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h = 20 + \left(\frac{12 - 5}{2 \times 12 - 5 - 8}\right) \times 10 = 26.364$$

6.4 MEAN DEVIATION

Mean deviation may be defined as the mean of all the deviations, in a given set of data obtained from an average. All the deviations are treated as positive. Mean deviation is treated as positive.

Example: Hb% of 10 patients of a ward of a hospital were obtained as 5, 7, 8, 10, 14, 12, 13, 5, 8, 8. Compute the mean deviation.

Calculation: Following 4 steps have to be taken to calculate mean deviation of ungrouped data.

1) Find out the mean of the series.

2) Find out the mean score and mean.

3) Sum up all deviations. All deviations are taken as positive.

4) Divide sum of all deviations by total number of observations.

Step 1: Mean= Sum of all observation/ Total Number of observations

= 5+7+8+10+14+12+13+5+8+8/10

= 90/10 = 9

Step 2: Following table is prepared to obtain deviations between each score and mean.

Score (X)	**Score-Mean(X-$\bar{X}$)**	**Deviation (x)**
5	5-9	-4
7	7-9	-2
8	8-9	-1
10	10-9	-1
14	14-9	5
12	12-9	3
13	13-9	4
5	5-9	-4
8	8-9	-1
8	8-9	-1

Step 3: Sum up all deviations regardless of sign.

= 4+2+1+1+5+3+4+4+1+1

=26

Step 4: MD= 26/10= 2.6

Grouped data: Following formula is used to obtain mean deviation from grouped data:

MD = f.x/f

Here, fx = Sum of multiplication of each frequency and each score.

F= Sum of all frequency.

Merits and demerits of mean deviation:

Merits: It is easy to calculate.

Demerits: It is less reliable because positive and negative signs are ignored.

6.5 STANDARD DEVIATION AND STANDARD ERROR

Standard deviation is the most important and widely used measure of dispersion. It is denoted by a Greek letter sigma. The standard deviation may be defined as 'the square root of the arithmetic mean of the squared deviations of measurements from their mean. It has accordingly often been called the root mean square deviation. Standard deviation is calculated differently in ungrouped and grouped data.

Ungrouped data: Following formula is used where deviation is obtained from mean.

Note: Standard deviation is computed by using N-1 in the denominator of the above formula in place of N if size of the sample is less than 30. If size of sample is more than 30 then previous formula i.e.

σ = √Σx2/N is used.

The above formula calls for following six steps in computation in fixed order:

Step 1. Find mean of the series.

Step 2. Find deviation of each score from the mean.

Step 3. Square each deviation, finding x_2.

Step 4. Sum the squared deviations, finding Σx_2.

Step 5. Divide this sum by N or N-1, finding

$\Sigma x_2/N$ or $\Sigma x_2/N-1$.

Step 6. Extract the square root of the result of step 5. This is standard deviation.

Solved Example Haemoglobin percent g/100mL of *Heteropneustes fossilis* was recorded as 23, 22, 20, 24, 16, 17, 18, 19 and 21. Compute the standard deviation (σ) by indirect method.

Calculation. Following table 4.7 having four columns are prepared on the basis of above observations. As per above steps one has to find mean of the series. Here ΣX = 180 and number of observations

N = 9.

Mean = 180/9 =20.

Observation X	**Observation- Mean** **$X-\bar{X}$**	**Deviation x**	**$(\text{Deviation})^2$ x^2**
16 17 18 19 20 21 22 23 24	16-20 17-20 18-20 19-20 20-20 21-20 22-20 23-20 24-20	-4 -3 -2 -1 0 +1 +2 +3 +4	16 9 4 1 0 1 4 9 16
ΣX = 180			$\Sigma x^2 = 60$

Here the size of the sample is less than 30. Therefore the following formula is applicable.

$\sigma = \sqrt{\Sigma x^2/N-1}$

On putting the values in the above formula-

$= \surd 60/9-1$

$= \surd 60/8$

$= 2.75$

Standard deviation from grouped data:

Following formula is used to obtain standard deviation by long method from grouped data:

$\sigma = \surd \mathbf{\Sigma f.x^2 / \Sigma f}$

The above formula calls for following steps in computation in fixed order.

Step 1- Find the midpoint of each class interval.

Step 2- Find mean value of the series using formula $\Sigma f.\ X/f$

Step 3- Find each deviation from the mean.

Step 4- Square each deviation, finding x^2.

Step 5- Multiply each squared deviation with corresponding frequency, finding $f.x^2$.

Step 6- Sum the squared deviations multiplied by frequency, finding $\Sigma f.x2$.

Step 7- Divide the sum of step 6 i.e. $\Sigma f.x2$ by $\Sigma f.x^2$ by Σf, finding $\Sigma f.x^2/\Sigma f$

Step 8- Extract the square root of the result of step 7.

Solved example- Weight of testis of 50 frogs is given below with their frequency. Find the standard deviation.

C.F	Mid-Point X	Frequency f	f.X	Deviation x	Deviation squared x^2	$f.x^2$
2-2.9	2.45	6	14.7	-2.14	4.5796	27.47
3-3.9	3.45	13	44.85	-1.14	1.2996	16.88
4-4.9	4.45	11	48.95	-0.14	0.0196	0.21
5-5.9	5.45	8	43.6	+0.86	0.7396	5.91
6-6.9	6.45	12	77.4	+1.86	3.4596	41.4

		$\Sigma f = 50$	$\Sigma f.X =$ 229.5		Σx^2 =10.088	$\Sigma f.X^2 =$ 91.87

Deviation x of each score (from mid-point) obtained from this actual mean using formula X- $\bar{X}$. For instance deviation

$x = X - \bar{X}$

$= 2.45 - 4.59 = -2.14$

Compute the obtained value in following formula

$\sigma =\surd\mathbf{\Sigma f.X^{2/}\Sigma f}$

$= \surd 91.87/50$

$= \surd 1.05$

$= 1.02$

Merits and demerits of standard deviation:

Merits:

(i) It summarizes the deviation of a large distribution from mean in one figure used as a unit of variation.

(ii) It indicates whether the variation of difference of an individual from the mean is real or by chance.

(iii) It helps in calculating the standard error.

(iv) It helps in finding the suitable size of sample for valid conclusions.

Demerits: It gives weightage to only extreme values. The process of squaring deviations and then taking square root involves lengthy calculation.

6.6 VARIANCE AND COEFFICIENT OF VARIATION

V stands for variance and it has algebraic interrelationship with standard deviation. The square of standard deviation is called variance. This may be demonstrated symbolically as follows:

$$\sigma\ \mathbf{2 = V}$$

Variance may be defined as "Square of sum of deviation divided by number of observations" or the square of standard deviation is termed as variance.

Suppose that we have a sample of one case with only one score. There is no possible basis for individual difference in such a sample; therefore there is no variance and variability. Consider a second individual with his score in the same test or experiment. We now have one difference. Consider a third case and we then have two additional differences, three altogether. There are as many differences as there are possible pairs of individuals. We could compute all these inter pair differences and could average them to get a single, representative value. We could also square them and then average them. It is most easy to find a mean of all scores and to use that value as a common reference point.

Each difference then becomes a deviation from that reference point and there are only as many deviations as there are only as many deviations as there are individuals. Either the variance or the S.D. is a single representative value for all the individual differences when taken from a common reference point.

Example: Hb% of 10 patients in a ward was recorded as 7, 8, 9, 10, 11, 12, 13, 14.5, 15 and 15.5g/100mL. Find out the variance of the data.

Following table was prepared from the above ungrouped data.

Hb % X	**Deviation X- $\bar{X}$=x**	**Standard deviation x^2**
7	7-11.5= -4.5	20.25
8	8-11.5= -3.5	12.25
9	9-11.5=-2.5	6.25
10	10-11.5=-1.5	2.25
11	11-11.5=-0.5	0.25
12	12-11.5=0.5	0.25
13	13-11.5=1.5	2.25
14.5	14.5-11.5=2.5	9.0
15	15-11.5=3.5	12.0
15.5	15.5-11.5=4	16.0

Σ X= 115.0		Σ x^2=80.75

Σ X̄= Σ X/N

Here, Σ X = 115, N = 10, X bar = 115/10 = 11.5

Variance or V = Σ χx2 / N = 80.75/10 = 8.075 **Ans.**

Measurement of relative dispersion (Coefficient of variation) Measures of dispersion gives us an idea about the extent to which variations are scattered around their central value. Therefore, two distributions having the same central values can be compared directly with the help of various measures of dispersion. If for example, an analysis of seed number per unit in two batches of 10 fruits in a garden, batch 1 has a mean score of 70 and standard deviation of 1.25 and batch II have a mean score 80 with standard deviation of 2, 4 then it is clear that batch I having a lesser value of S.D. are more consistent in producing seed number than batch II.

On the other hand we have a situation when two or more distributions having unequal means or different units of measurements are to be compared in respect of their variability. For making such a comparison we use a statistic called relative dispersion or coefficient of variation (c.v.). Formula for coefficient of variation is as follows:

c.v. = 100 × Standard deviation / Mean

Example: Mean values of Hb % of 20 males and 20 females were calculated as 13.5 and 14 mg/100mL. SD of males as 3 and 4 respectively. Find coefficients of variation of both male and female. Mention which sex is more variable and which more consistent

Group of sex	**Mean**	**S.D.**
Males	13.5	3
Females	14	4

For males, c.v = 100 × 3/13.5 = 22.22%

For females c.v. = 100 × 4/ 15 = 28.57%

Deductions: Females are variable than males in respect of Hb %. In other words, contrary to females, males are more consistent in Hb %.

Merits and demerits of variance:

1) It is easy to calculate.

2) It indicates the variability clearly.

But the use of cv creates two difficulties:

a) The unit of expression of variance is not the same as that of the observations, because variance indicates squared deviations. For example if 'x' values are obtained in cm variance will be in square cm.

b) Variance is usually a large number compared to the values of observations. Therefore variance is now seldom used to express the variability.

6.7 CHI-SQUARE TEST (χ2)

By significance of statistics we mean non-chance difference between obtained scores on the basis of sample and scores based on some hypothesis. If the observed difference is significant then we say that the observed difference is not influenced by chance defying null hypothesis. On the other hand if the observed difference is obtained by chance. Here we shall deal with only one method of test of significance known as "standard error" and will learn about the use of chi square test.

Standard error and student t tests are parametric tests and are applied to only quantitative data. In biological experiments a non-parametric test is very commonly called chi-square test. It is applied only for qualitative data such as colour, health, intelligence, cure response of drugs etc. which do not have numerical values.

The Chi- square test was developed by Prof. A. R. Fisher in 1870. Karl Pearson improved Fisher's chi-square test in its modern form in 1900. Chi-square is derived from the Greek letter (χ) and pronounced as 'kye'.

Definition: Chi-square test is the test of significance of overall deviation square in the observed and expected frequencies divided by expected frequencies.

General computing formula for chi-square:

Chi- square (χ 2)= Σ{(O-E)^2/E} or Σ{(f0-f3)2/fe}

Here, fo or ΣO = Observed Frequency

fe or ΣE = Expected frequency

Common applications of chi- square (χ 2) test:

1.As an alternate test to find significance of difference in two or more than two proportions. Chi – square test is very useful test which is applied to find significance in the same type of data with two more advantages:

a) To compare the values of two binomial samples even if they are small such as oxygen consumption in 5 control and 5 thyroxin injected fishes of the same species.

b) To compare the frequencies of two multinomial samples such as oxygen consumption in control and T4 injected groups of fishes weighing.

2. As a test of association between two events in binomial or multinomial samples. Chi- square measures the probability of association between two discrete attributes. Two events can be studied for their association such as iron intake and Hb%, season and fecundity, T4 injection and oxygen consumption, nutrition and intelligence, weight and diabetes etc. There are two possibilities, either they influence or they do not.

χ 2 is very useful tool in ascertaining mendelian ratio.

Association table: Table is prepared by enumeration of qualitative data. Since one wants to know the association between two sets of events, the table is also called association table.

Four fold or 2 × 2 contingency table: When there are only two samples, each divided into two classes, it is called fourfold, four cell of 2 × 2 contingency table.

Table: Outcome of treatment with drug and placebo.

Groups	**Outcome or Result** **Died**	**Survived**	**Total**

A (Control on placebo)	10	25	35
B (Experiment on drug)	5	60	65
Total	15	85	100

Multifold Association Table: The association of two sets of events having more than two classes will be larger than a fourfold or four cell contingency table.

Table: Social class and *Wuchereria* Positivity:

Social Class	**Outcome or Result**			
	Number +ve	Number -ve	Total	Percentage +ve
I.	4	76	80	5
II.	20	180	200	10
III.	60	440	500	12
IV.	144	576	720	20
	228	1272	1500	47

3. As a test of goodness of fit. Chi square test is also applied as a test of "goodness of fit". Goodness of fit reveals the closeness of observed frequency with those of the expected. Thus it helps to answer whether something (physical or chemical factors) did or did not have an effect. If observed and expected frequency are in complete agreement with each other than the chi-square will be zero. But it rarely happens in biological experiments. There is always some degree of deviation.

Prerequisite of $\chi2$ test. There are three basic prerequisites of $\chi2$ test such (i) Sample must be random. (ii) Data should be qualitative. (iii) Observed frequency should not be less than five.

Method to draw inferences: If the calculated value of χ 2, then observed value of χ 2 is more than the tabulated value of χ 2 is more than the two variables are dependent on each other and value is insignificant.

The quantity in the denominator which is one less than the independent number of observations in a sample is called degree of freedom. If there are 2 classes (For example control and T4 injected, male and female) the degree of freedom would be 2 - 1 = 1. If there are three classes then d.f = 3 - 1 = 2, in case of 4 classes d.f. = 4 - 1 = 3 and so on.

If the χ 2 value obtained in more than two pairs of data then d.f. = (2-1) × (2 - 1) = 1.

Calculation of χ 2 test. The calculation of χ 2 is easy and same for each case. If fo is the observed frequency of a particular category of a variable and fe is the expected frequency of some hypothesis, then χ 2 is calculated by following formula

$$\chi\ \mathbf{2 = \{(fo\text{-}fe) / fe\}.}$$

Following steps have to be taken to obtain χ 2 values:

Make a contingency table and note the observed frequency (fo or O) in each class of one event, row wise i.e. horizontally and then the numbers in each group of the other event, column wise i.e. vertically.

Determine the expected frequency (fe or E) in each group of the sample on the assumption of null hypothesis (Ho) i.e. no difference in the proportion of the group from that of the population.

Find the difference between the observed and expected frequency in each cell (fo-fe) or (O-E).

Calculated χ 2 value applying formula:

χ 2 = {(fo-fe) 2/ fe}

Calculated χ 2 value is compared with tabulated value χ 2 at desired degree of freedom under different probabilities 0.5, 0.1, 0.05, 0.01, 0.001 etc.

If calculated χ 2 value is higher than tabulated value then it is considered as significant. But if the calculated value is less than the table value then it is considered insignificant.

Example: In a monohybrid cross between tall (TT) and dwarf (tt) 1574 tall and 554 dwarf were obtained. Suggest if a ratio of 3 : 1 is suitable or not.

Calculation: Total number = 1574 Tall + 554 Dwarf

= 2128.

Therefore expected 3 : 1 will be 2128 × ¾ : 2128 × ¼.

Therefore expected 3:1 will be 2128×2128×1/4

Expected ratio = 1596 : 532

Observed ratio = 1574 : 554.

Putting the values in the formula:

χ 2 = {(fo-fe)2/fe}

= (1574-1596)2/1596 + (554 - 532)2/532

= (-22)2/1596 + (22)2/532 = 484/1596 + 484/532

= 0.303 + 0.909

= 1.212. Ans

Here, d.f. = 2 - 1 = 1

Significance, At 5 % level, at 1 degree of freedom the table value of χ 2 is 1.212.

This shows that the two series of frequencies, observed and expected, are not in agreement with the theoretical ratio of 3 : 1.

Example- In a cross between black male and gray female Drosophila the offspring obtained were 25 black and 35 gray. Calculate the χ 2 and give your inference on the ratio of black and gray offsprings. Expected number is calculated from the fact that gray body color is dominant and the expected ratio of this nature is 1 : 1 [Total number of offspring are 60].

Calculation. Following table is prepared to obtain required values.

Table:

BLACK	**GRAY**
Observed number 25 Expected number 30 (O-E) 25-30 = -5 (O-E)2 = (25-30)2 = 25	35 30 35-30 = +5 (35-30)2 = 25

Putting values in the formula:

χ 2 = Σ {(fo-fe)2 / fe}

= 25/30 + 25/30

= $\frac{5}{6} + \frac{5}{6}$ = 10/6 = 1.66.

The table value of χ 2 at d.f. 1 p 0.05 is 3.84. The obtained value is 1.6. The obtained value is less and therefore it is not in agreement with the theoretical ratio of 1 : 1.

6.8 STUDENT T TEST

W. S. Gossett (1908) applied a statistical tool called t test, to test the significance of the difference between two means. The pen name of Mr. Gossett was Student and therefore t-test is known as 'student t-test'. Later on R.A. Fisher developed the t-test and explained it in various ways. Student's t-test is used not only to test the significance of difference between two means but also to test the significance of product moment correlation, point- biserial correlation, rank difference correlations etc.

Student's t test is also known as t-ratio because it is the ratio of difference between two means and standard error of difference between two means.

Following formula is used to obtain t-ratio.

t = Difference between two means / SE of diff. between two means

Here, $\bar{X}_1$= Mean of one variable

$\bar{X}_2$= Mean of second variable

SED = Standard error of difference between two means.

To determine the significance: Probability of occurrence of any calculated value of't' table corresponding to the degree of freedom derived from the number of conservations in the sample under study. If the calculated value of't' exceeds the value given in the t table at different levels (0.01, 0.05 etc.), it is said to be significant. But if the calculated value of't' is less than the table value then the difference between two means is insignificant.

Applications of t-test: The significance of difference between two means is obtained differently in uncorrelated or unpaired t-test and paired t-test.

Unpaired or uncorrelated t-test: Unpaired t-test is applied to unpaired data of independent observations made on individuals of two different or separate groups or samples drawn from two populations.

[Note: Standard error of the difference between two uncorrelated means is calculated differently during t-test]

Following steps have to be taken to test the significance of difference between two uncorrelated means:

Find the observed difference between means of two samples (X_1-X_2).
The standard error of the difference between uncorrelated sample means is obtained with the help of following formula;

$$\text{SED} = \sqrt{}\text{SE } X_1^2 \text{bar} + \text{SE } X^{22} \text{ bar}$$

Here, SED = Standard error of the difference between the two sample means.
SE $\bar{X}$ $_1$bar = Standard error of the first mean.
SE $\bar{X}_2$ bar = Standard error of the second mean.

We find SE_M of each mean with the help of following formula

$$\mathbf{SE\sigma = \sigma/\sqrt{}N} \text{ or } \mathbf{\sigma/\sqrt{}N\text{-}1}$$

To obtain SE**σ**one have to obtain the value of combined (σ). Following formula is used to obtain

Combined $\sigma = \sqrt{\Sigma x_1^2 + \Sigma x^2_2 / N_1 - N_2}$

Example: Body length of 10 fishes of a species of fish was obtained from two ponds (population) of Gaya town. They were measured as:

Pond A: 20cm, 24cm, 20cm, 28cm, 22cm, 20cm, 24cm, 32cm, 24cm and 26 cm.

Pond B. 12 cm, 10cm, 8cm, 10cm, 6cm, 4cm, 14cm, 20cm, 10cm, and 6cm.

Calculate the mean difference in total body length between the two pounds of fish is significant or not.

Calculation: Following steps have to be taken to obtain t ratio:

- Make a table of six columns.
- First column having length of sample A; Second column for $X_1 - X_2 = x_1$; Third column $(\bar{X}_1 - \bar{X}_1)^2$ or x_1^2 Fourth column having length of sample B; Fifth column for $(\bar{X}_2 - \bar{X}_2) = x_2$; Mean value; Sixth column for $(\bar{X}_2 - \bar{X}_2)^2$ or x_2^2.

6.9 SUMMARY

Central tendency may be considered as a synonym of average. Average is a general term which describes the general value of series, around which all other observations are dispersed.

The Chi- square test was developed by Prof. A. R. Fisher in 1870. Karl Pearson improved Fisher's chi-square test in its modern form in 1900. Chi-square is derived from the Greek letter (χ) and pronounced as 'kye'.

Variance may be defined as "Square of sum of deviation divided by number of observations" or the square of standard deviation is termed as variance.

6.10 TERMINAL QUESTIONS AND ANSWERS

Question1. Three groups with 20 patients in each were administered analgesics A, B and C. Relief was noted in 20, 10 and 6 cases respectively. Is this difference due to the drug or by chance?

Question2. Define and mention formula of Chi square test ($\chi2$)? What do you mean by goodness of fit?

Question3. What do you mean by measures of variability? Name measures of variability of individual variations.

Question4. Define and explain mean, median and mode. Mention formula to find out mean, median and mode both for ungrouped and grouped data.

Question5. Calculate mean and median from the following ungrouped data:

I. 10, 8, 20, 22, 39, and 18.

II. 19, 21, 17, 16, 19, 21, 23, and 23.

III. 17, 19, 13, 17, 13, 11, and 21.

IV. 15.8, 13.3, 15.2, 13.3, 17.8, 18.1, and 18.9.

Question6. Compute N (No of observation) when mean = 5 and sum of means = 30.

Question7. Calculate mean, median and mode from the data given in following table

Table A

Class interval	Frequency

16-20	4
21-25	4
26-30	9
31-35	7
36-40	13
41-45	3
46-50	3
51-55	2
56-60	2
61-65	3

Question 8: What do you mean by student's t test? Mention formula to obtain't' ratio.

REFERENCES

- Mahajan BK 2002 (Methods in Biostatistics) (6th edition)
- Zaman SM, HK Rahim and M Howlader 1982. (Simple Lessons from Biometry), BRRI
- Research methodology methods and techniques, C. R, Kothari, New Age International Limited Publisher,

UNIT- 7 CORRELATIONS AND REGRESSION

Contents

7.1 OBJECTIVES

To study

- ✓ Types of correlation.
- ✓ Regression analysis.

7.2 INTRODUCTION

Correlation means association of two or more facts. In statistics correlation may be defined as 'the tendency of simultaneous variation between two variables'. The distribution involving two variables is called bivariate distribution and the distribution involving more than two variables is called multivariate distribution. In statistics we study the degree of correlation between two or more variables. Sometimes two variables are measured in the same individual such as length and weight, oxygen consumption and body weight, Body weight and Hb% etc. At other times the same variable is measured in two or more related groups such as tallness in parents and offspring, intelligent quotient (IQ) in brothers and in corresponding sisters (siblings) and so on.

F. Galton coined the term regression in 1885 to explain the data obtained during the study of inheritance. Galton observed the height of offspring's of a few generations of generations of a family and came to the conclusion that the height of offspring's tend to remain in middle position. "The tendency to remain towards central position was called regression by Galton."

We have studied that in order to draw a relationship; observations of two variables are plotted in the form of dots in a scatter diagram. A straight line is drawn which will approach as close as possible to all these points in the graph. The statistical analysis employed to find out the exact position of the straight line is known as the linear regression analysis. The main objective of regression analysis is to predict the value of one variable using the known value of the other. The existence of a relationship between the independent variable X and the dependent variable Y can be expressed in a mathematical form known as the linear regression analysis.

The main objective of regression analysis is to predict the value of one variable using the known value of the other. The existence of a relationship between the independent variable X and dependent variable Y can be expressed in a mathematical form known as the regression equation. The equation expressed by a straight line is called the linear regression equation.

The degree of association is measured by a correlation coefficient, denoted by 'r'. It is sometimes called Pearson's correlation coefficient after its originator and is a measure of linear association. If a curved line is needed to express the relationship, other and more complicated measures of the correlation must be used.

The correlation coefficient is measured on a scale that varies from + 1 through 0 to– 1. Complete correlation between two variables is expressed by either + 1 or -1. When one variable increases as the other increases the correlation is positive; when one decreases as the other increases it is negative. Complete absence of correlation is represented by 0. Figure 11.1 gives some graphical representations of correlation.

7.3 TYPES OF CORRELATION

Types of correlation: There may exist five kinds of correlation between two variables depending on its extent and direction. Each type may be shown both mathematically and graphically:

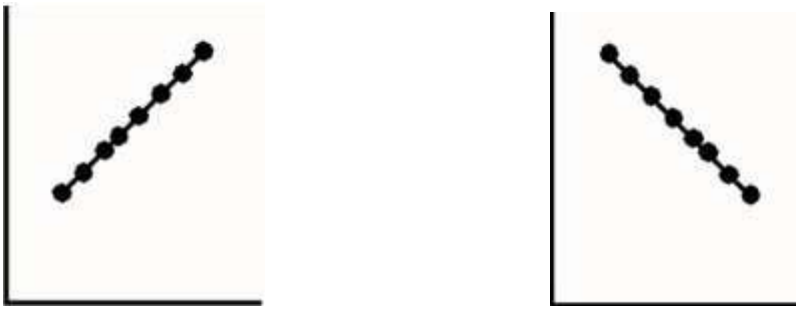
I. **Perfect Positive Correlation:** The two variables denoted by letter X (Body length) and Y (Body weight) are directly proportional and fully correlated with each other. Both variables rise or fall in the same proportion. Examples of perfect or total correlation is very rare in nature but some approaching to that extent are there such as day length and temperature; rain and humidity; body weight and height; age and height; age and weight etc, upto certain age. The imaginary mean line rising from the lower ends of both X and Y axes forms a straight line. When the scatter diagram is drawn all the points fall around the mean line.

II. **Moderately positive correlation:** The two variables denoted by X (Age of husband) and Y (Age of wife) are partially positively correlated. Values of correlation coefficient (r) lie between 0 and +1, i.e., $0<r<1$. Other examples of positive correlation may be infant mortality rate and overcrowding, tallness of plants and the quantity of manure used, nutrition and death rate in pregnancy etc.In such moderately positive correlation, the scatter will be there around an imaginary mean line, rising from the lower ends of both X and Y variables.

III. **Perfect negative correlation:** The variables denoted by letter X (Temperature) and Y (Lipid content of body of a species of fish) are inversely proportional to each other, i.e., when one (X) rises the other (Y) falls in the same proportion. The correlation coefficient (r)= -1 to 0. Examples of perfect negative correlation are also very rare in nature but some approaching to that extent is there such as temperature and lipid content of the body, RBCs number and Hb%, T4 injection and oxygen.

IV. **Moderately negative correlation:** The two variables denoted by X (Economic condition of States) and Y (case of tuberculosis). In this case values of correlation coefficient lie between -1 and 0 such as income and infant mortality rate, age and vitality in adults etc. In such moderately negative correlation, the scatter will be there around an imaginary mean line rising from the extreme values of the variable.

V. **Absolutely no correlation.** In this case the value of correlation coefficient (r) is zero, indicating that no linear relationship exists between the two variables. There is no imaginary mean line indicating a trend of correlation. X is completely independent of Y such as Hb% and body weight; Body weight and IQ etc. In absolutely no correlation X variable is completely independent of Y variable. In this case points are so scattered that no imaginary line can be drawn.



(a) Perfect Positive Correlation(r= +1) (b) Perfect Negative correlation(r=-1)

7.3.1 SIMPLE CORRELATION

In bivariate distribution, the correlation may be positive or negative, and linear or curvilinear.

Two variables co-varying in the same direction are positively correlated. For example, we expect a positive correlation between height and weight of a group of individuals.

Correlation between the two variables in opposite directions is negatively correlated. The increase in one variable results in a decrease in the other. For example, an increase in the number of caterpillars results in a corresponding decrease in leaves of plants.

The correlation of two variables which can be expressed by a straight line is called linear correlation. In perfect linear correlation the amount of change in one variable bears a constant ratio to the amount of change in the other. For example, the length of five fishes of a species and their snout length is measured in cm. The measurement is given below:

Body length: 8, 9, 11, 12, 13 X variable

Snout length: 1, 2. 4, 5, 6 Y variables

The above correlation indicates that each individual scored 1 cm more on test Y. This means that the correlation between the above two variables is expressible in the form Y=X+1, which is an expression representing a straight line, i.e. a perfect positive linear relationship, in which correlation between X and Y will be +1. (Though it rarely happens in biological experiments)

The correlation of cores of two variables based on some quality shown by a curve line on a graph is called curvilinear correlation.

Correlation coefficient: Numerical expression of correlation is called mathematical correlation. In other words, correlation of two variables by mathematical method is obtained by correlation coefficient. In biological experiments use of correlation coefficient is very significant.

According to J. P. Guilford 'A coefficient of correlation is a single number that tells us to what extent two or more things are related and to what extent variations in one go with variations in other." The correlation coefficient is expressed by a letter 'r'.

7.3.2 METHODS OF STUDYING CORRELATION

Methods of studying correlation:

There are three methods of studying correlation between two variables in the case of ungrouped data:

1. Scatter diagram method.

2. Pearson's product moment method and

1. Scatter diagram method: Scatter diagram or dot diagram is a graphic device for drawing certain conclusions about the correlation between two variables. In preparing a scatter diagram, the observed pairs of observations are plotted by dots on a graph paper by taking the measurements on variable X along the horizontal axis and that on variable Y along the vertical axis. The placement of these dots on the graph reveals the change in the variable as to whether they change in the same or in opposite directions. Scatter diagram showing various degrees of correlation.

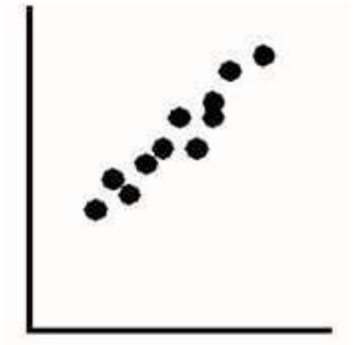
I. **Perfect Positive Correlation:** The two variables denoted by letter X (Body length) and Y (Body weight) are directly proportional and fully correlated with each other. Both variables rise or fall in the same proportion. Examples of perfect or total correlation is very rare in nature but some approaching to that extent are there such as day length and temperature; rain and humidity; body weight and height; age and height; age and weight etc, upto certain age. The imaginary mean line rising from the lower ends of both X and Y axes forms a straight line. When the scatter diagram is drawn all the points fall around the mean line.

II. **Moderately positive correlation:** The two variables denoted by X (Age of husband) and Y (Age of wife) are partially positively correlated. Values of correlation coefficient (r) lie between 0 and +1, i.e., $0<r<1$. Other examples of positive correlation may be infant mortality rate and overcrowding, tallness of plants and the quantity of manure used, nutrition and death rate in pregnancy etc.In such moderately positive correlation, the scatter will be there around an imaginary mean line, rising from the lower ends of both X and Y variables.

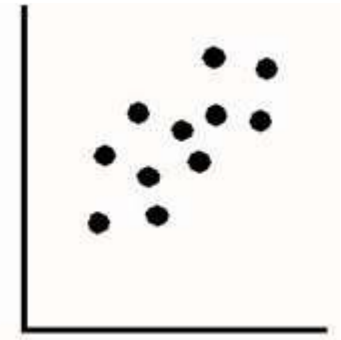
III. **Perfect negative correlation:** The variables denoted by letter X (Temperature) and Y (Lipid content of body of a species of fish) are inversely proportional to each other, i.e., when one (X) rises the other (Y) falls in the same proportion. The correlation coefficient (r)= -1 to 0. Examples of perfect negative correlation are also very rare in nature but some approaching to that extent are there such as temperature and lipid content of the body, RBCs number and Hb%, T4 injection and oxygen.

IV. **Moderately negative correlation:** The two variables denoted by X (Economic condition of States) and Y (case of tuberculosis). In this case values of correlation coefficient lie between -1 and 0 such as income and infant mortality rate, age and vitality in adults etc. In such moderately negative correlation, the scatter will be there around an imaginary mean line rising from the extreme values of the variable.

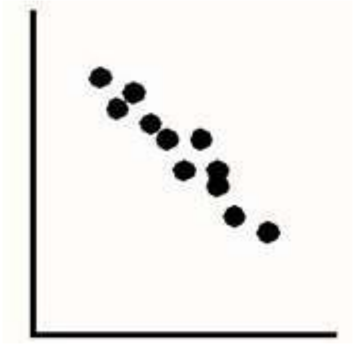
V. **Absolutely no correlation.** In this case the value of correlation coefficient (r) is zero, indicating that no linear relationship exists between the two variables. There is no imaginary mean line indicating a trend of correlation. X is completely independent of Y such as Hb% and body weight; Body weight and IQ etc. In absolutely no correlation X variable is completely independent of Y variable. In this case points are so scattered that no imaginary line can be drawn.



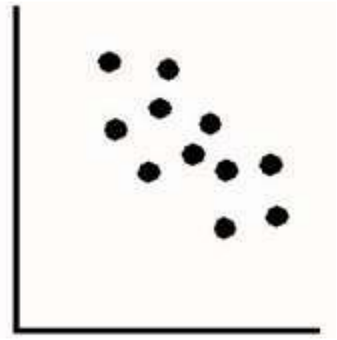
(a) Strong Positive Correlation



(b) Weak Positive Correlation



(c) Strong Negative Correlation



(d) Weak Negative Correlation

2. Pearson's product moment method

It is also known as Pearson's coefficient of correlation. It is one of the most widely used algebraic methods of finding correlation between two variables. The coefficient of correlation (r) gives an idea about the degree of linear relationship between two variables. Formula to obtain coefficient of correlation (r) is used as follows:

$$\mathbf{r = \Sigma x \times y / \sqrt{}\Sigma x^2 \times y^2}$$

Where X is the independent variable normally represented by the abscissa and Y is the dependent variable represented by the ordinate. x and y are the deviations from the respective means (as used for other purposes ~~d~~etermination of variance and standard deviation).

In language we can say that ‘r’ can be calculated by dividing the sum of products of deviation from their respective means by the square root of the products of the sums of squares of deviations from the respective means of two variables.

Here, r = correlation coefficient

x = deviation of X variable

y = deviation of Y variable

$\sum xy$ = Sum of multiplication of deviation x and y

Pearson’s product moment method:

The above formula is applied when ‘r’ is obtained by applying actual mean.

Example: The length and weight of 7 groups of fishes of a species is given below. Find out the correlation coefficient of the two variables.

Length of body 11.7 cm, 13.9 cm, 15.5 cm, 17.8 cm, 18.5 cm, 19.2 cm, 21 cm.

Weight of the body 7.10 g, 12.42 g, 15.35 g, 23.20 g, 28.45 g, 32.25 g and 39.84 g.

S. No.	Length X	Weight Y	x	y	x2	y2	x.y
1. 2. 3. 4. 5. 6. 7.	11.7 13.9 15.5 17.8 18.5 19.2 21	7.10 12.42 15.35 23.20 28.45 32.45 39.84	-5.1 -2.9 -1.3 +1 +1.7 +2.4 +4.2	-15.58 -10.236 -7.33 +5.5 +5.8 9.6 +17.19	26.01 8.41 1.69 1.0 2.89 5.76 17.64	241.8 104.6 53.2 .30 33.64 92.16 295.49	79.3 23.36 9.49 0.55 9.86 28.8 72.2
N = 7	ΣX =	ΣY =			Σx^2 = 66.64	Σy^2 = 821.19	$\Sigma x \times y$ = 223.56

	117.6	158.81					

Calculation: For calculation of correlation coefficient from above data (ungrouped series) a table is prepared with help of following steps;

1) Make a table of 8 columns.

2) Mention a serial number in column 1, value of X in column 2 and value of Y in column 3.

3) Find out the actual mean of X and Y with the help of a formula.

4) Find the deviation of all scores of X and Y. Formula to find out deviation from actual mean (Score - Mean). Put all values of deviations in column 4 against their scores for variable X and in column 5 for variable Y.

5) Find the square of x and y and put them in column 6 and 7 respectively.

Put the multiplication value of x and y of each score in the last column i.e. 8th column. All values of **x×y** are summed and given in the last column.

X1 bar of X = 117.6/7 = 16.8. Mean length of fish.

X2 bar of Y = 158.81/7 = 22.68 Mean of weight of fish

$r = \mathbf{\Sigma} x \times y / \sqrt{\mathbf{\Sigma} x^2 \times y^2}$

$\mathbf{= 223.56 / \sqrt{66.66} \times 821.19}$

$= 223.56 / \sqrt{54724.101}$

$\mathbf{= 223.56 / 233.93}$

= 0.96.

Inference: The calculated value of correlation coefficient (r) is 0.96. One has to see the significance of 'r' at 0.05 and 0.01 level. First of all we find out df. Here df= N-2 i.e 7-2=5.

Here df=5 and calculated value of r = 0.96.

On verification of correlation table we observe that the value of 'r; at d.f. 5 is 0.755 at 0.05 level. Calculated value of 'r' is 0.96. Since the calculated value is higher, therefore it is clear that 'r' is significant at 0.05 levels at df 5. Now we can safely say that both variables i.e. length and weight of the body is in complete +ve correlation.

7.4 REGRESSION ANALYSIS

1) The first aim or regression analysis is to predict the value of one character of variable (variable say Y) from the known value of the other character or variable (variable say X). The formal variable Y to be predicted is called dependent variable and the latter known variable X is called the independent variable. This is done by the regression line and by finding another constant called regression coefficient. Regression line explains the mean relationship between X and Y variables.

2) To find out the measures of error, present during the use of a regression line for prediction, is another aim of regression analysis.

For this standard error of estimate is calculated.

Linear regression: Linear regression between values of two variables are possible only when one unit change in the independent variable (X) influences change in the definite quantity in the dependent variable (Y). This change may be on the positive or negative side beyond the mean.

The lines of the best fit passing through the middle of points on the plotted graph are drawn. These lines are called regression lines. The two regression lines are drawn one is X, Y and the other is Y on X indicating conditions of moderately +^{ve} and moderately–^{ve} correlation respectively. The two regression lines interest at the point where perpendiculars drawn from the means of X and Y variables meet.

When there is perfect correlation (r = +1 or -1) the two regression lines will coincide or become one straight line. Though perfect correlation is not possible in biological experiments. When the correlation is not possible in biological experiments. When the correlation is partial, the lines will be separate and diverge forming an acute angle at the meeting point of perpendiculars drawn from the means of two variables. Lesser the correlation, greater will be the divergence of angle.

Steepness of the lines indicates the extent of correlation. Close to the correlation greater is the steepness of regression lines X on Y and Y on X.

Composition of regression lines is based on least square assumptions. The general condition for regression analysis is based on lines of the best fit. It is called least squared error.

Regression equation: The existence of a relationship between the independent variable X and the dependent variable Y can be expressed in a mathematical form is known as the regression equation. These equations represent the regression lines.

Regression equation of Y on X indicates the changes in the values of X for changes given in Y. Likewise regression equation of X and Y indicates the changes in the values of Y for changes given in X.

Regression equation of X on Y:

$$X = a + b\,y$$

Regression equation of Y on X:

$$Y = a + b\,x$$

In both equations x and y are values of variables whereas a and b are constant. Constant a is intercepted i.e. it is that point where the regression line touches Y axis. In other words, the distance between the touching point of the regression line on the Y axis from the point of origin is a. If correlation is + ve regression lines touch the Y axis above of origin and in case of $-^{ve}$ regression line touches Y axis below point of origin.

$\mathbf{x = a + b\times y}$

$\sum x = n.a + b\times\sum y$

$n\times a = \sum x - b\sum y$

Both sides are divided by n.

$na / n = \sum x/n - b \sum y/n$

$a = \bar{X}b\ \bar{y}$

Likewise, y = a + b x

Or a = y + b $\bar{X}$

$\bar{X}$ is the mean of x series and y is the mean of y series. Constant b exhibits the slope of the line. It is the value of angle made by the regression line and its horizontal line (X-axis). In other words b is gradient or slope. It means for the measurement of any distance on X axis:

Change in values of Y axis/ Distance on axis. In the given graph the position of a and b has been made clear from the equation y = a + bx.

This clears that determination of any special straight lines depends on the value of a and b and the best least square line can be obtained only when the real value of a and b is determined. Values of a and b can be obtained by following two normal equations.

In Y = a + bx, value of a and b can be obtained by following equation

$\sum \mathbf{Y} = \mathbf{n} \times \mathbf{a} + \mathbf{b} \times \sum \mathbf{x}$

$\sum \mathbf{XY} = \mathbf{a} \sum \mathbf{x} + \mathbf{b} \sum \mathbf{y}^2$

7.4.1 USES OF REGRESSION ANALYSIS

The regression analysis attempts to accomplish the following:

1. Regression analysis provides estimates of values of the dependent variable from values of the independent variable. The device used to accomplish this estimation procedure is the regression line. The regression line describes the average relationship existing between X and Y variables, i.e., it displays mean values of X for given values of Y. The equation of this line, known as the regression equation, provides estimates of the dependent variable when values of the independent variable are inserted into the equation.

2. A second goal of regression analysis is to obtain a measure of the error involved in using the regression line as a basis for estimation. For this purpose the standard error of estimate is calculated. This is a corresponding value estimated from the regression line. If the line fits the

data closely, that is, if there is little scatter of the observations around the regression line, good estimates can be made of the Y variable. On the other hand, there is a great deal of scatter of the observations around the fitted regression line; the line will not produce accurate estimates of the dependent variable.

3. With the help of regression coefficients we can calculate the correlation coefficient. The square of correlation of coefficient (r), called coefficient of determination, measures the degree of association or correlation that exists between the two variables. It assumes the proportion of variance in the dependent variable that has been accounted for by the regression equation. In general, the greater the value of r^2, the better is the fit and the more useful the regression equation as a predictive device.

7.5 SUMMARY

Correlation means association of two or more facts. In statistics correlation may be defined as ‘the tendency of simultaneous variation between two variables’. The distribution involving two variables is called bivariate distribution and the distribution involving more than two variables is called multivariate distribution. In statistics we study the degree of correlation between two or more variables. Sometimes two variables are measured in the same individual such as length and weight, oxygen consumption and body weight, Body weight and Hb% etc.

Regression analysis is a branch of statistical theory that is widely used in almost all the scientific disciplines. In economics it is the basic technique for measuring or estimating the relationship among economic variables that constitute the essence of economic theory and economic life. The uses of regression are not confined to economics and business fields only. Its applications are extended to almost all the natural, physical, and social sciences.

7.6 TERMINAL QUESTIONS AND ANSWERS

Question 1: Define and explain correlation and correlation coefficient with examples.

Question 2: What is regression? Differentiate between correlation and regression. Explain the method of least square to estimate the regression coefficient in a linear regression of Y on X.

Question 3: What is the purpose of regression analysis? What do you mean by linear regression? Explain regression equation.

Question 4: The body length and girth of 7 groups of a species of fish in cm is as follows. Find the regression equation.

Body length-X	**Girth of body-Y**
13.9	4.2
15.7	4.7
15.8	4.7
17.5	5.2
18.1	5.4
19.9	6.0
22.0	6.5

Question 5: Deviations taken from the mean of X and Y (two variables) is given below. Find ‘r’ by Pearson’s product moment method and explain their significance.

X- 4, -3, -2, -1, 0, 1, 2, 3, 4

Y- 3, -3, -4, 0, 4, 4, 1, 2, -2, -1

7.7REFERENCES

- Mahajan BK 2002 (Methods in Biostatistics) (6th edition)
- Zaman SM, HK Rahim and M Howlader 1982. (Simple Lessons from Biometry), BRRI
- Research methodology methods and techniques, C. R, Kothari, New Age International Limited Publisher,

UTTARAKHAND OPEN UNIVERSITY

Teenpani Bypass Road, Behind Transport Nagar,
Haldwani- 263139, Nainital (Uttarakhand)
Phone: 05946-261122, 261123; Fax No. 05946-264232
Website: www.uou.ac.in; e-mail: info@uou.ac.in
Toll Free No.: 1800 180 4025

MSCZO-603-1(003443)