ZO (N) 220&ZO (N)-220 L

B.Sc. 3rd Semester

Bioinformatics, Biostatistics & Instrumentation Techniques

DEPARTMENT OF ZOOLOGY
SCHOOL OF SCIENCES
UTTARAKHAND OPEN UNIVERSITY

ZO(N)-220 & ZO(N)-220L

Bioinformatics,Biostatistics&Instrumentation Techniques

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

Board of Studies

Dr. Neera Kapoor
Professor & Head
Department of Zoology
School of Sciences
IGNOU Maidan Garhi, New Delhi.

Dr. O. P. Gusain
Department of Zoology
HNB Garhwal (Central University)
Srinagar (Garhwal)
Uttarakhand.

Dr. Pravesh Kumar Sehgal
Associate Professor
Department of Zoology
Uttarakhand Open University
Haldwani, Nainital.

Dr. Mukta Joshi
Assistant Professor
Department of Zoology
Uttarakhand Open University
Haldwani, Nainital.

Dr. A. K. Dobriyal
Professor & Head
Department of Zoology
BGR Campus Pauri
HNB Srinagar Garhwal.

Dr. Shyam S. Kunjwal
Assistant Professor
Department of Zoology
Uttarakhand Open University
Haldwani, Nainital.

Dr. Jaya Upreti
Assistant Professor
Department of Zoology,
Uttarakhand Open University
Haldwani, Nainital.

Poornima Nailwal
Assistant Professor
Department of Zoology
Uttarakhand Open University
Haldwani, Nainital.

Programme Coordinator

Dr. Pravesh Kumar Sehgal (Associate Professor)
Department of Zoology
School of Sciences, Uttarakhand Open University
Haldwani, Nainital.

Unit writing and Editing

Editor

Dr.Shyam Kunjwal
Department of Zoology
Uttarakhand Open University
Unit No: 1, 2,3,4,5,6,7,8,9,10

Dr.Anju Thapliyal
BGR Campus Pauri
HNB Garhwal University
&
Dr.Mukta Joshi
Department of Zoology
Uttarakhand Open University
Unit No: 11,12,13,14

RIVISION/RE-EDITING/CONTENT-EDITING/LANGUAGE EDITING

Dr.Shyam Kunjwal
Assistant Professor
Department of Zoology
Uttarakhand Open University

Writer

Dr.sunil Bhandari, Unit 1

Department of Zoology, Govt.PG College, Gopeshwar

Dr.M.Faisal, Unit 2, 3, 4:

School of Agriculture Forestry and Fisheries, Himgiri Zee University Dehradun

Dr.Lakhan Singh Unit 5, 6 &7
HNB Garhwal University,
Uttarakhand

Dr.Harish Chandra Unit: 8, 9& 10
Assistant Professor, HAPPRC
HNB Garhwal University,
Uttarakhand

Dr.N.C.Khanduri Unit: 11, 12 & 13
Assistant Professor
Govt. PG College Agustyamuni,
Rudraprayag, Uttarakhand

Dr.Shyam S.Kunjwal Unit: 14
Assistant Professor, Department of Zoology
Uttarakhand Open University

Course Title and Code: Bioinformatics, Biostatistics & Instrumentation Techniques ZO (N)-220 Lab Work ZO (N)-220L

ISBN	**: 978-93-85740-54-1**
Copyright	**: Uttarakhand Open University**
Edition	**2024**
Published By	**: Uttarakhand Open University, Haldwani, Nainital- 263139**

Contents

Bioinformatics, Biostatistics &
Instrumentation Techniquesand
Lab Work

Course code: ZO (N)-220 & ZO (N)-220 L **Credit: 3+1**

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

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.

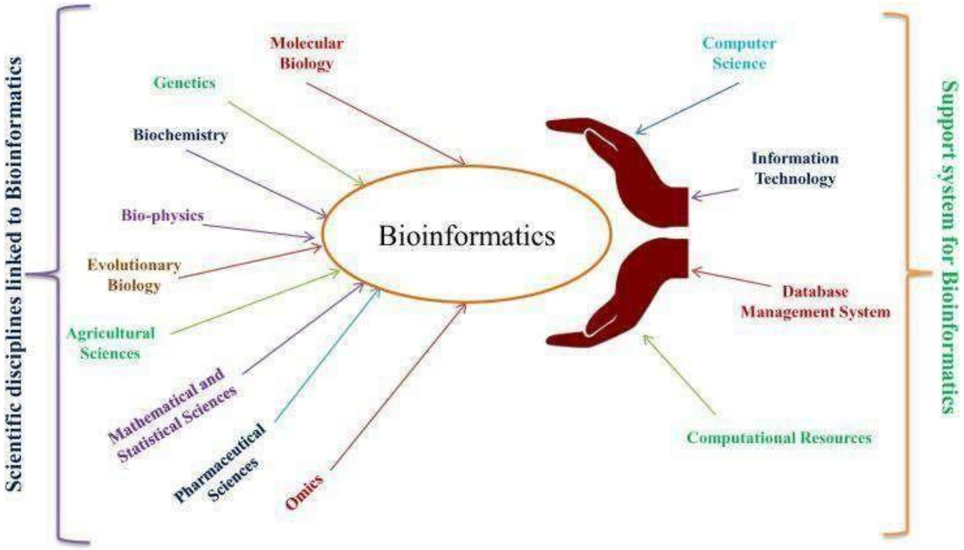


Fig.1.2 Scope of bioinformatics

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.

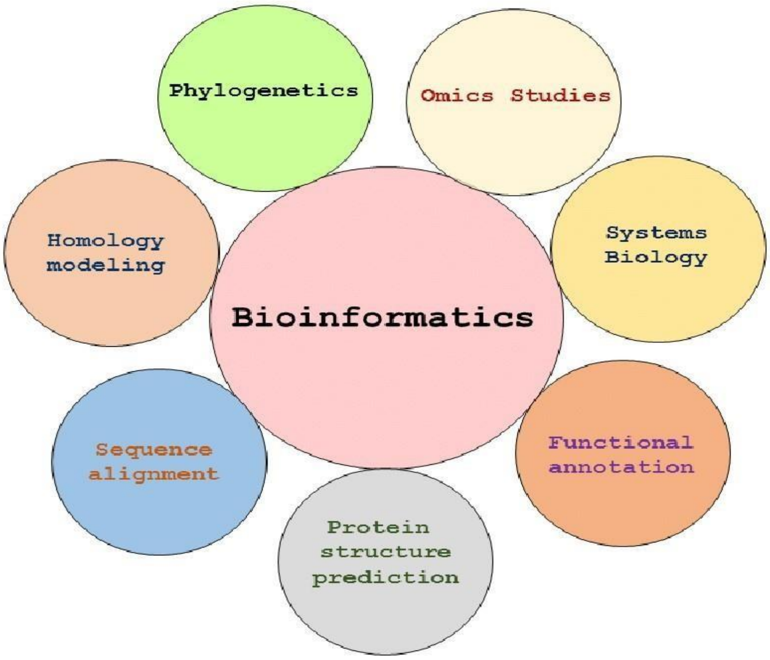
h- **Metabolomics:** The study of the **biochemistry** of metabolism and metabolites in living beings.

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.

1.2 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.

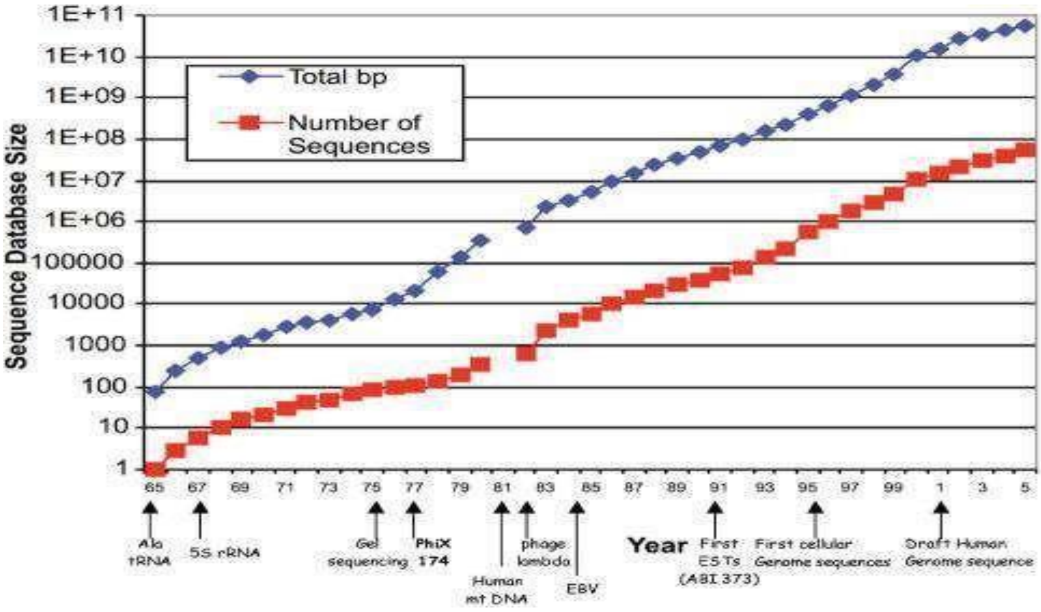
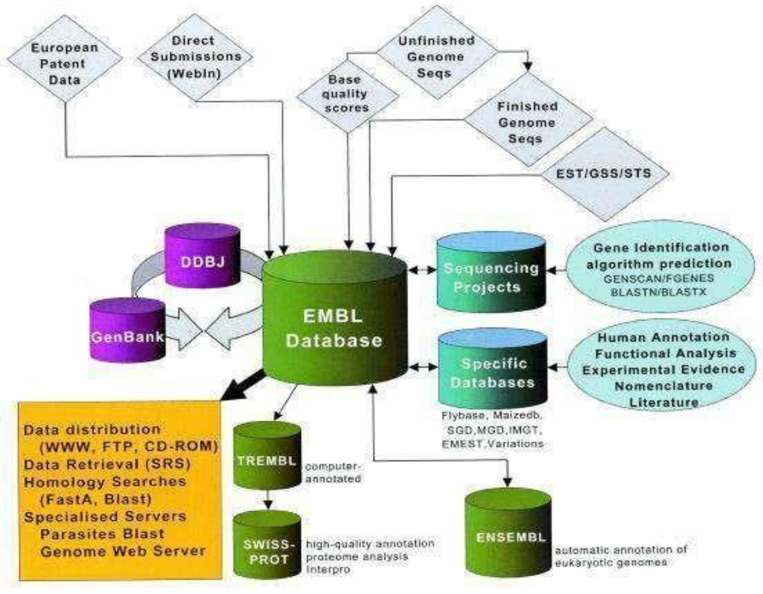


Fig-1.4Nucleotide sequence database



European Patent Data
Direct Submissions (WebIn)
Base quality scores
Unfinished Genome Seqs
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

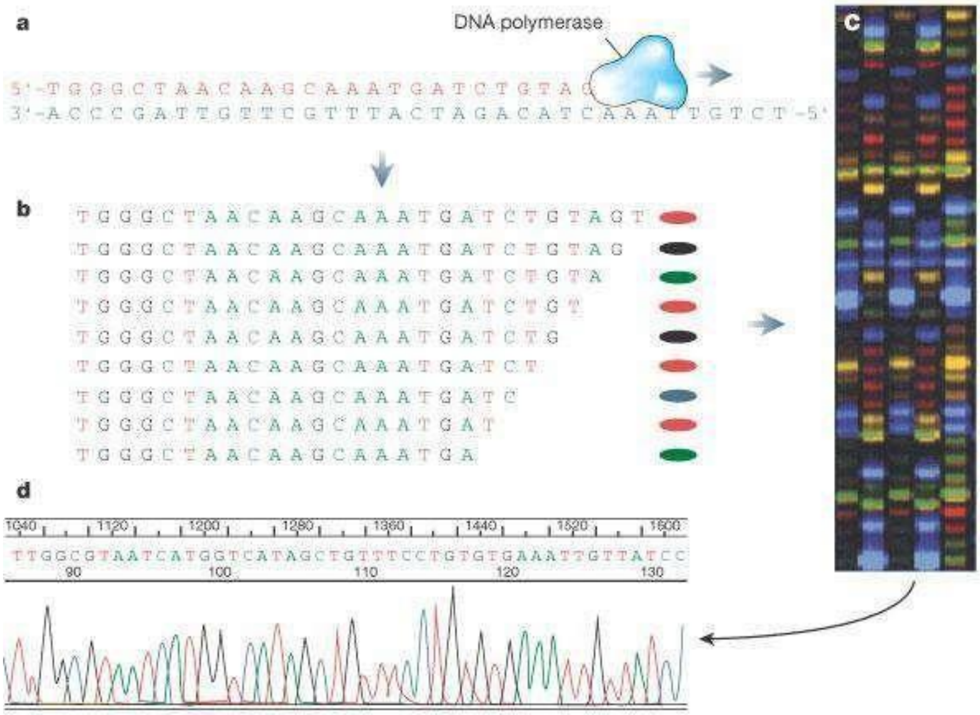

a
DNA polymerase
5'-TGGGCTAACAAGCAAATGATCTGTAG
3'-ACCCGATTGTTCGTTTACTAGACATCAAATTGTCT-5'
b
TGGGCTAACAAGCAAATGATCTGTAGT
TGGGCTAACAAGCAAATGATCTGTAG
TGGGCTAACAAGCAAATGATCTGTA
TGGGCTAACAAGCAAATGATCTGT
TGGGCTAACAAGCAAATGATCTG
TGGGCTAACAAGCAAATGATCT
TGGGCTAACAAGCAAATGATC
TGGGCTAACAAGCAAATGAT
TGGGCTAACAAGCAAATGA
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. They collaborate with Sequence Read Archive (SRA).

Gen Bank

The Gen Bank 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.

a. 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.

b. 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.

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.

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.

a. FlyBase Database:

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

b. 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.

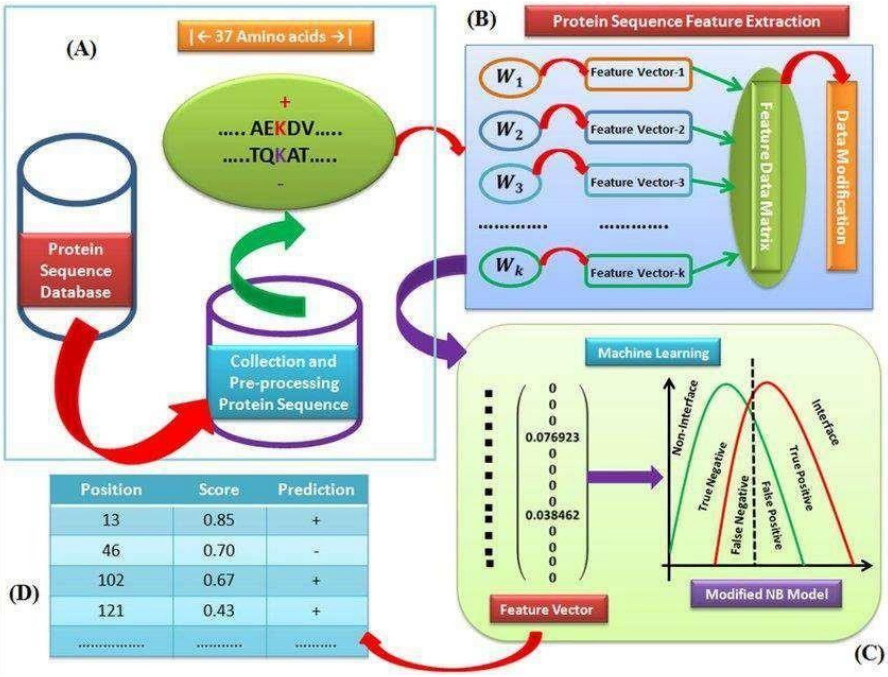
b- The obvious examples are the nucleotide sequences, the protein sequences, and the 3D structural data produced by X-ray crystallography and macromolecular NMR.

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 data bases 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 super family 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.

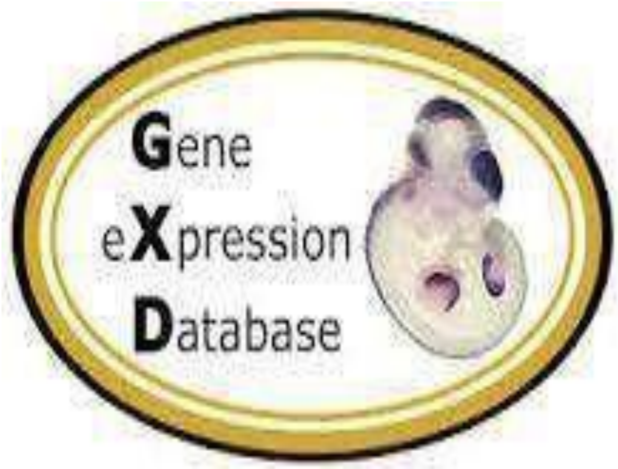


Fig.1.9 Gene expression database

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.

1.3 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.

5. 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.

6. 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.

1.4 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

1.5REFERENCES

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 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 algorithmutilizes scoringof 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 ~~P~~DBe, 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.

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. O ff-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 forstructure 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.

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.

4.2 & 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 d eluge 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 longestbranch.

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 from 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 Illumina MiSeq, and the Roche 454 Genome Sequence. NGS platforms rapidly generate sequencing read data on the giga base 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 distances 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

1. Rastogi SC, Mendiratta N, Rastogi P (2011) Bioinformatics: Concepts, Skills &Applications. CBS Publishers & Distributors Pvt. Ltd. ISBN: 81-239-1482-2.

2. Mount DM (2004) Bioinformatics: Sequence and Genome Analysis 2. Cold SpringHarbor 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 DATA COLLECTION AND PRESENTATION

CONTENTS

5.1 OBJECTIVES

Following are the objectives of this chapter:

1. To know the definitions of Statistics and Biostatistics.
2. To know about some statistical symbols.
3. To know the scope and applications of Biostatistics.
4. To know about data, its collection and collection techniques.
5. To know about organization and representation of data by many graphical techniques such as histogram, pie chart, frequency polygon etc.

5.2 INTRODUCTION

We welcome the reader who wishes to learn biostatistics. In this chapter we introduce you to the subject. First of all we define statistics and biostatistics and then examples are given where bio- statistical techniques are useful. These examples show that biostatistics has an importance in advancing our biological knowledge; biostatistics helps to evaluate many life-and-death issues in medicine.

We advise you to read the examples carefully and then think yourself, “What can be inferred from the information presented?” What would you do with the data after they are collected? How can it be presented and what you can get from it? We want you to realize that biostatistics is a tool that can be used to benefit you and society.

There is no royal road to biostatistics. You need to be involved. You need to work hard. You need to think. If you analyze the actual data, the result will be a powerful tool that has immediate practical uses. Our main purpose is to develop thought patterns in your mind that are useful in evaluating information in all areas of your life.

5.3 DEFINITIONS OF STATISTICS AND BIOSTATISTICS

Much of the joy and pain in life arises in situations that involve considerable uncertainty. Here we are giving two situations which show that the study of statistics and biostatistics is necessary.

1. Parents of a child with a genetic defect consider whether or not they should have another child. They will base their decision on the chance that the next child will have the same defect.

2. To choose the best therapy, a physician must compare the diagnosis or future course, of a patient under several therapies. A therapy may be a success, a failure, or somewhere in between; the evaluation of the chance of each occurrence necessarily enters into the decision.

5.3.1 DEFINITION OF STATISTICS

Statistics is the science which deals with the collection, classifying, presenting, comparing and interpreting numerical data collected to throw light on any sphere of enquiry- Lovitt.

The science of statistics is a most useful servant, but only of great value to those who understand its proper use- W.I.King. Statistics provides tools and techniques for research workers- A.M. Mood. Planning is the order of the day and without statistics planning is inconceivable- L.H.C. Tippet.

Statistics may be defined as a science of numerical information which employs the process of measurement and collection, classification, analysis, decision making and communication of results in a manner understandable and verifiable by other- Cecil H. Meyers

5.3.2 DEFINITION OF BIOSTATISTICS

Biostatistics is the application of statistics methods applied to biological areas. Biological laboratory experiments, medical research (including clinical research), and health services research all use statistical methods. Many other biological disciplines rely on statistical methodology.

There are three reasons for focusing on biostatistics:

1. Some statistical methods are used more deeply in biostatistics than in other fields. For example, a general statistical textbook would not discuss the life-table method of analyzing survival data of importance in many bio-statistical applications. The topics in this book are adapted to the applications in mind.

2. Examples are drawn from the biological, medical, and health care areas; this helps you maintain motivation. It also helps you in understanding how to apply statistical methods.

3. A third reason for a book on biostatistics is to teach the material to the audience of health professionals. In this case, the interaction between students and teacher, but especially among the students themselves, is of great value in learning and applying the subject matter.

5.4 STATISTICAL SYMBOL

5.4.1 STATISTICAL SYMBOL

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

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

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

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

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

Σ : Sum of all the values of a given set

Q.D.: Quartile deviation of a given set of values or of a distribution

M.D.: Mean deviation of a given set of values or of a distribution

5.4.2 SCOPE OF BIOSTATISTICS

Biostatistics is the application of statistics in different fields of biology. The science of 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 execute interpretation and inference from the results. A major branch of this is medical biostatistics, which is exclusively concerned with health and medical sciences.

In current world, the scope of biostatistics is increasing rapidly. If we discuss about biostatistics, we see that 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. In general, the difference between a statistics program and a biostatistics program is twofold: (i) statistics departments will often host theoretical/methodological research which are less common in biostatistics programs and (ii) statistics departments have lines of research that may include biomedical applications but also other areas such as industry (quality control), business and economics and biological areas other than medicine

There is a special need of the subject bio statistics because it related with such areas as medical , pharmacy, forestry, agriculture, etc, which are very necessary for the betterment of society.

5.4.3 APPLICATION OF BIOSTATISTICS

The importance and application of statistics in the field of biology is increasing day

by day. Why it is so? The reason is that in biology the interplay of casual and response variables follow the laws that are not in the classic mold of 19th century physical science. In that century, biologists such as Robert Mayer, Helmholtz, and others in trying to show that biological process were nothing but physicochemical phenomena, helped create the impression that the experimental methods and natural philosophy that had led to such dramatic progress in the physical sciences should be imitated fully in biology.

Many biologists even to this day have retained the tradition of strictly mechanistic and deterministic concepts of thinking, while physicists, as their science became more refined and came to deal with ever more elementary particles, began to resort to statistical approaches. In biology most phenomena are affected by many casual factors, uncontrollable in their variation and often unidentifiable. Statistics is needed to measure such variable phenomena with a predictable error and to ascertain the reality of minute but important differences.

A Biostatistics centre could jointly organize working groups, the seminar series, computing infrastructure and possibly consulting and clinical trials coordinating centre cervices. The main objective of the centre would be to estimate, collaborate on, and circulate results of research in a particular subspecialty in the following reasons:

1. Statistical methods for longitudinal studies;
2. Statistical genetics;
3. Foundations of inference;
4. Bayesian biostatistics
5. Biostatistician practice and education.

The most critical short term problem in the field of biostatistics is the information system. We need to incorporate modern, web-based technologies into the everyday workings of the department of biostatistics. We need reliable and accessible systems that are competitive with those available to departments of statistics and biostatistics. We likely build collaborations with computer science students.

5.5 DATA AND ITS TYPES

5.5.1 DATA

The information collected from census or surveys or from other sources is called raw data. The word data means information. The adjective raw attached to data indicates that the information collected cannot be used directly. It has to be converted into more suitable form before it begins to make sense to be utilized gainfully. Raw data is like raw rice. Raw rice has to be cooked properly and tastefully before it is eaten and digested. Similarly, raw data has to be converted into proper form such as tabulation, frequency distribution form, etc, before any inference is drawn from it.

There are two ways of statistical data;

1. Primary data 2. Secondary data

Primary data: It is the data collected by some person or organization for his own use from any primary source. For example the data of census report collected by the centre government, the data collected by any agency for its own purpose, the gadget of India, etc.

Secondary data: It is the data collected by some other person or organization for their own use but the investigator also gets it for his own use. For example the data collected by any medical agency can be used by some other medical institute students.

In other words, primary data are those data which are collected by you to meet your own specific purpose where as the secondary data are those data which are collected by somebody else. A data can be primary for one purpose and secondary for the other.

5.5.2 TYPES OF DATA

The primary data are of the following types:

5.5.2.1 Nominal Data

In the study of biostatistics, we meet many different types of numerical data. The different types have varying degrees of structure in the relationships among possible values. One of the simplest types of data is nominal data, in which the values fall into unordered categories or classes. In a certain study, for instance, males might be assigned the value 1 and females the value 0. Numbers are used mainly for the sake of convenience; numerical values allow us to use computers to perform complex analysis of the data. Nominal data that take on one of two distinct values-such as male and female are said to be dichotomous or binary, depending on whether the Greek or the Latin root for two is preferred. However, not all nominal data need be dichotomous. Often there are three or more possible categories into which the observations can fall. For example, persons may be grouped according to their blood type, such that 1 represents type O, 2 is type A, 3 is type B, and 4 is type AB.

5.5.2.2 Ordinal Data

When the order among categories becomes important, the observations are referred to as ordinal data. For example, injuries may be classified according to their level of severity, so that 1 represents a fatal injury, 2 is severe, 3 is moderate, and 4 is minor. Here a natural order exists among the groupings; a smaller number represents a more serious injury. A second example of ordinal data is Eastern Cooperative Oncology Group's classification of patient performance status.

Status 0: Patient fully active, able to carry on all predisease performance without restriction.

Status 1: Patient restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature.

Status 2: Patient ambulatory and capable of all self-care but unable to carry out any work activities; up and about more than 50% of waking hours.

Status 3: Patient capable of only limited self-care; confined to bed or chair more than 50% of

Waking hours.

Status 4: Patient completely disabled; not capable of any self-care; totally confined to bed.

5.5.2.3 Ranked Data

In some situations, we have a group of observations that are first arranged from highest to lowest according to magnitude and then assigned numbers that correspond to each observation's place in the sequence. This type of data is known as ranked data. As an example, consider all possible causes of death in the India. We could make a list of all of those causes, along with the number of lives that each one claimed in. If the causes are ordered from the one that resulted in the greatest number of deaths to the one that caused the smallest and then assigned consecutive integers, the data are said to have been ranked.

5.5.2.4 Discrete Data

For discrete data, both ordering and magnitude are important. In this case, the numbers represent actual measurable quantities rather than mere labels. In addition, discrete data are restricted to taking on only specified values-often integers or counts-that differ by fixed amounts; no intermediate values are possible. Examples of discrete data include the number of motor vehicle accidents in Dehradun in a particular month, the number of times a woman has given birth, the number of new cases of tuberculosis reported in the India during a one-year period, and the number of beds available in a particular hospital. Note that for discrete data a natural order exists among the possible values. If we are interested in the number of times a woman has given birth, for instance, a larger number indicates that a woman has had more children. Furthermore, the difference between one and two births is the same as the difference between four and five births. Finally, the number of births is restricted to the nonnegative integers; a woman cannot give birth 3.4 times because it is meaningful to measure the distance between possible data.

5.5.2.5 Continuous Data

Data that represent measurable quantities but are not restricted to taking on certain specified values (such as integers) are known as continuous data. In this case, the difference between any two possible data values can be arbitrarily small. Examples of continuous data include time, the serum cholesterol level of a patient, the concentration of a pollutant, and temperature. In all instances, fractional values are possible. Since we are able to measure the distance between two observations in a meaningful way, arithmetic operations can be applied. The only limiting factor for a continuous observation is the degree of accuracy with which it can be measured; consequently, we often see time rounded off to the nearest second and weight to the nearest pound or gram. The more accurate our measuring instruments, however, the greater the amount of detail that can be achieved in our recorded data.

In a study of the effects of maternal smoking on newborns, for example, we might first record the birth weights of a large number of infants and then categorize the infants into three groups: those who weight less than 1500 grams, those who weight between 1500 and 2500 grams, and those who weight more than 2500 grams. Although we have the actual measures of birth weight, we are not concerned with whether a particular child weighs 1560 grams or 1580 grams; we are only interested in the number of infants who fall into each category. From prior experience, we may not expect substantial differences among children within the very low birth weight, low birth weight, and normal birth weight groupings. Furthermore, ordinal data are often easier to handle than continuous data and thus simplify the analysis. There is a consequent loss of detail in the information about the infants, however. In general, the degree of precision required in a given set of data depends on the questions that are being studied.

As we progressed, the nature of the relationship between possible data values became increasingly complex. Distinctions must be made among the various types of data because different techniques are used to analyze them. It does not make sense to speak of an average blood type of 1.8; it does make sense, however, to refer to an average temperature of 4.55°C.

5.6 DATA COLLECTION AND RELATED TERMS

5.6.1 POPULATION, SAMPLE, SAMPLING UNIT AND SAMPLING FRAME

5.6.1.1 Population

Population is a collection of units or objects of which some property is defined for every unit or object. Population may consist of finite or infinite number of units. Population is also called universe by a number of statisticians and scientists.

The inhabitants of a region, number of wheat fields in a state or district, fruit plants in a city, number of students in a institution, insects in a field, persons suffering from any particular disease, workers in a institution, total no of person in city, total households, total no of students in any university, are a few examples of finite populations. All real numbers, all stars in the sky are examples of infinite populations. Generally, the population has a large number of animates and inanimates. Moreover, the units or subjects constituting the population may vary from survey to survey in the same region of activity depending upon the aims and objective of the survey.

In brief, one should very well keep in mind that statistical population is not only the human population which is usually considered in literary sense. It is generally a group or collection of items specified by certain characteristics or defined under certain restrictions.

5.6.1.2 Sample

A sample is the portion of the population that is examined to make inferences about the population or a part or fraction of population, which represent it, is known as sample. Sample consists of a few items of the population. In principle a sample should be such that it is a true representative of the population.

If poll stars are trying to take the pulse of the nation prior to an election their target population consists of those who will go to the polls and vote, whereas those whose opinions they actually obtain constitute a sample of that population.

5.6.1.3 Sampling Unit

The constituents of a population which are the individuals to be sampled from the population and cannot be further subdivided for the purpose of sampling at a time are called sampling units. For example to know the average income per family, the head of the family is a sampling unit. To know the average marks in a paper of a class a single student is a unit.

5.6.1.4 Sampling Frame

For accepting any sampling procedure it is necessary to have a list or a map identifying each sampling unit by a number. Such a list or map is called sampling frame. For example, a list of students of a class, a list of patients of a particular disease a list of workers in a factory, a list of voters, a list of staff members of a college, a list of villages in a district, etc., are a few examples of sampling frame.

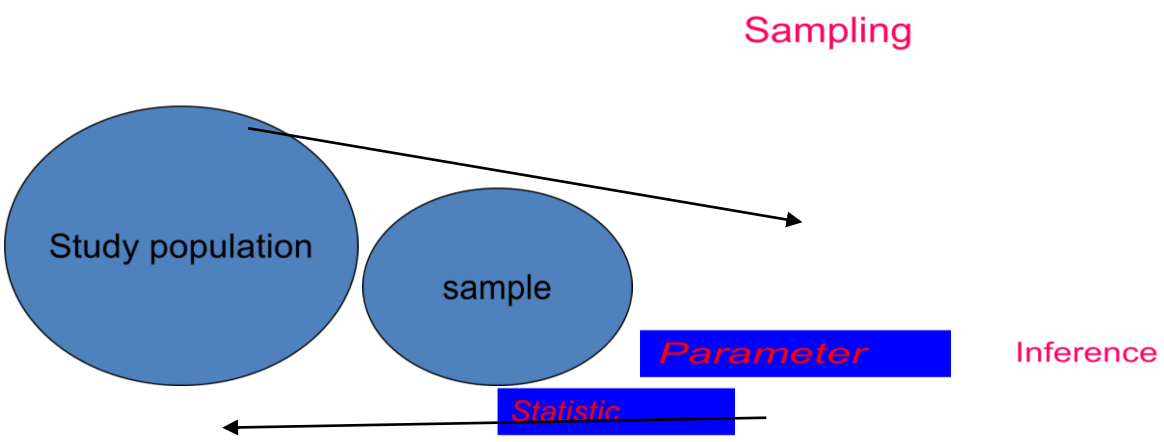


Figure- 1.1 Population v/s Sample

5.6.2 PRINCIPLES STEPS IN A SURVEY

The main steps in the planning and execution of a survey are as follows:

1. Objective of the survey. The first step is to define clearly the objective of the survey. It is generally found that even the sponsoring agency is not quite clear in its thinking as to what it wants and how it is going to use the results. The sponsors of the survey should take care that the objectives should be fulfilled with the available resources presented in the form of time, money and manpower.

2. Defining the population to be sampled. The population from which the sample is to be taken should be defined clearly. For example in sampling of farms clear-cut rules must be framed to define a farm in respect of shape, size, etc., keeping in mind the border- line cases so as to enable the investigating person to decide whether to include or not a particular farm in the population.

3. The frame and sampling unit. The sampling units must cover the entire population and they must be distinct, obvious and non-overlapping in the sense that every element of the population belongs to one and only one sampling unit. For example, in socio-economic survey for selecting people in a town, the sampling unit might be an individual person, a family, a household or a block in a locality.

In order to cover the population decided upon, there should be a list, map or some other acceptable material, called the frame, which serves as a guide to the to the population to be covered. The construction of the frame is one of a main problem since it is the frame which determines the structure of the sample survey. If the frame is not up-to-date, it should be brought up-to-date before using it.

4. Data to be collected. The data should be collected keeping in view the objective of the survey. The tendency should not be to collect too many data some of which are never subsequently examined and analyzed. A practical method is to chalk out an outline of the tables that the survey should produce. This would help in eliminating the collection of irrelevant information and ensure that no necessary data are omitted.

5. The questionnaire or schedule. Having decided about the type of the data to be collected, the next important step of the sample selection is the construction of the questionnaire (to be filled by the respondent) or schedule of enquiry (to be filled by the interviewer) which requires skill, special technique as well as familiarity with the subject matter. The questions should be clear, brief, non- offending, polite in tone, clear-cut and to the point so that not much scope of guessing is left on the part of the respondent or interviewer. Suitable and detailed instructions for filling up the questionnaire or schedule should also be prepared.

6. Method of collecting information. The two methods commonly used to collect the sample data are:

(i) Interview method. In this method, the investigator goes from house to house and interviews the individuals personally. He asks the questions one by one and fills up the schedule on the basis of the information gained from the individuals.

(ii) Mailed questionnaire method. In this method the questionnaire is mailed to the individuals who are required to fill it up and returns it duly completed.

Whether the data should be collected by interview method or mail questionnaire method or by physical observation has to be decided keeping in view the cost, time, accuracy and money.

7. non – respondents. Due to practical difficulty the data cannot be collected for all the sampled units. For example the selected respondent may not be available at his place when the investigator goes to him or he may refuse to give certain information. This is called non- response problem. Such cases of non-response should be handled with caution in order to draw unbiased and valid conclusions.

8. Selection of proper sampling design. The size of the sample (n), the procedure of selection and the estimation of parameters along with their margins of uncertainty are some of the important problems that should be tackled carefully.

A number of designs for the selection of a sample are available and a good selection will guarantee good and reliable estimates but the relative time and money factors should also be considered for adopting any sampling design.

9. Organization of field work. It is very essential that the investigator should be trained in locating the sample units, recording the measurements, the methods of collection of required data before starting the field work. The success of a survey to a great extent depends upon the reliable field work. It is very necessary to make provisions for adequate supervisory staff for inspection after field work.

10. Pretest. From practical point of view a small pre-test has been found very useful. It always helps to decide upon effecting methods of asking questions and results in the improvement of the questionnaire. In case of large scale surveys it provides the better idea about the cost and time factor.

11. Summary and analysis of the data. The analysis of the data may be classified into the following steps:

(i) Scrutiny and editing of the data. An initial quality check should be done by the supervisory staff when the investigators are in the field. This will help in amending recording errors or in eliminating data that are inconsistent.

(ii) Tabulation of data. Before carrying out the tabulation of the data, we should decide the procedure for tabulation of the data which are incomplete due to non-response. The method of tabulation, hand or machine should be depending upon the size of the data.

(iii) Statistical analysis. A properly scrutinized, edited and tabulated data now prepared for the statistical analysis. There are different methods of estimation; therefore a suitable formula should be used for the estimation of the parameters.

(iv) Information for future surveys. Any completed survey helps in providing a note of caution for designing future surveys. The information in the form of the data means, standard deviation, the nature of the variability and the cost, time, etc., are important which are helpful for future surveys. Any completed sample survey is a lesson for future surveys in recognizing and rectifying the mistakes committed in the post survey.

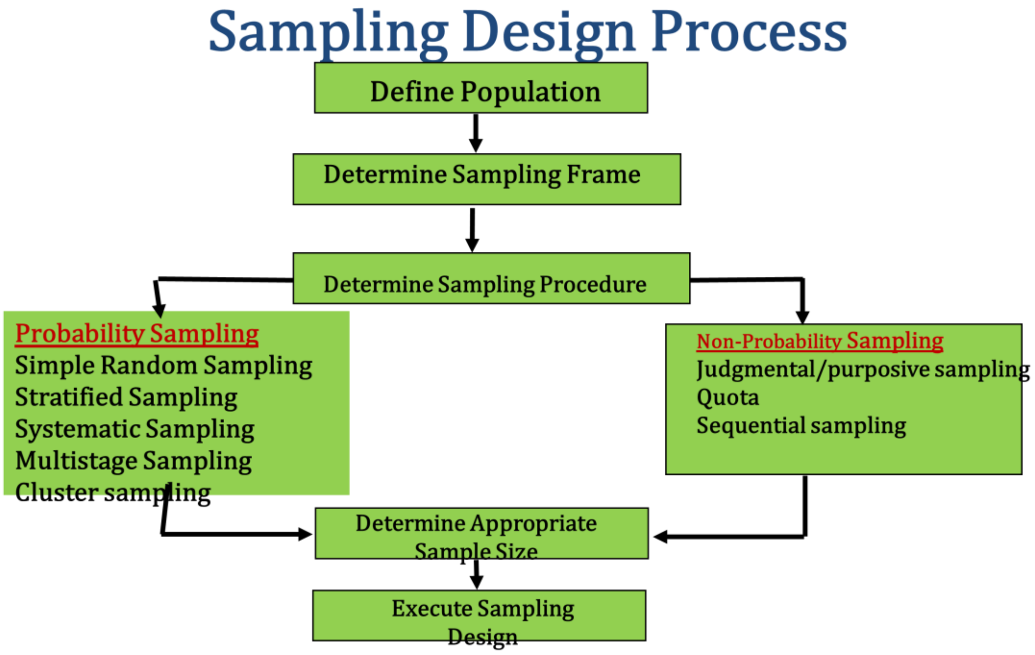


Figure-1.2 Sampling Design Process

5.6.3 SAMPLING AND NON- SAMPLING ERRORS

The errors in the collection, processing and analysis of the data are of two types:

1. Sampling error   2. Non-sampling error

5.6.3.1 Sampling errors

Sampling errors arise in the collection of a sample and the reason is because only a small part of the population is used for getting the population parameter estimates. Therefore these are absent in the complete enumeration. The main reasons of these errors are:

1. Faulty selection of the sample;
2. Substitution of the existing unit;

3. Faulty demarcation of the sampling units;
4. Constant error due to improper choice of the statistics for estimating the population parameters.

5.6.3.2 Non- Sampling errors

The errors due to the inductive process of inferring about the population on the basis of a sample, the non-sampling errors arise at the stages of observation, ascertainment and processing the data and so are present in the complete enumeration and sample survey both. The reasons of these errors are:

1. Faulty planning or definitions
2. Response errors
3. Non-response errors
4. Errors in coverage
5. Compiling errors
6. Publication errors

5.7 TYPES OF SAMPLING SCHEMES

The technique of selecting a sample is of fundamental importance in the theory of sampling and usually depends upon the nature of the data and type of enquiry. The procedure of selecting a sample may be classified into three forms:

1. Subjective or judgment sampling
2. Probability or random sampling
3. Mixed sampling

5.7.1 SUBJECTIVE OR PURPOSIVE OR JUDGMENT SAMPLING

In this scheme of sampling the sample is selected with some definite purpose in mind of the selector and so the selection of the sampling units depends completely on the decision of the selector. This sampling suffers from the drawback of favoritism depending the beliefs and prejudices of the selector and so does not provide a true representative sample of the population. For example if the selector wants to give the picture that the standard of living has increased in Dehradun city, he may take individuals in the sample from the posh colonies of the city and ignore the colonies where low income group and middle class group live. Another example, suppose a sample of TB patients has to be drawn. Since, it is not possible to ascertain a population of TB patients, the persons turning up to TB sanitorium and having TB are selected in the sample.

This sampling method is seldom used and cannot be recommended for general use. However, if the selector is experienced and skilled and this sampling is carefully applied, then judgment sampling may provide useful results. This sampling is used in the selection of- national players of a national team and opinion surveys.

5.7.2 PROBABILITY SAMPLING

Probability sampling is the scientific method of selecting samples according to some laws of chance in which each unit of the population has some pre-assigned probability of being selected in the sample. The different types of probability sampling are:

- **(i)** Where each unit has an equal chance of selection.
- **(ii)** Where sampling units have different chances of selection.
- **(iii)** Where chances of selection of unit is proportional to the sample size.

Some techniques which are commonly used in sampling are as follows:

5.7.2.1 Simple random sampling

This is the basic and most commonly used method of sampling. In this method each unit of the population has an equal chance of selection in the sample.

In this method, an equal probability is attached to each unit of the population at the first draw. It also indicates an equal probability of selection for the remaining units at the subsequent draws.

For example, to draw a simple random sample from an outdoor patient register of the department of obstetrics and gynecology, each entry would need to be numbered subsequently. If you want to draw a sample of size 700 out of 3500, a list of 700 random numbers between 1 and 3500 would need to be prepared using one of the known procedures (described later). The 700 entries made in the register corresponding to 700 random numbers present in the prepared list would make up the required sample.

There are two ways in simple random sampling, if the unit drawn is replaced back before the next unit is drawing, the technique is called simple random sampling with replacement and if the drawing units are not replaced back and the next draws are done without selected units, the technique is called simple random sampling without replacement.

5.7.2.1. Selection of a simple random sample

Mainly two approaches are use to draw a simple random sample:

(a) Lottery system method
(b) Mechanical randomization or random numbers method

(a) Lottery system. This is the simplest method of selecting a random sample. The process is given below:

Suppose for a survey we want to select (n) students out of a class of (N) students. We assign the numbers 1 to N; one number to each student and we write these numbers on (N) identical chits which are same in size shape and color. These chits are put in a bag and thoroughly shuffled and then (n) chits are drawn one by one. The (n) students corresponding to the numbers on these chits will make the required sample.

This method is quite independent of the properties of the population. Generally, in place of chits, cards are used. This is one of the most reliable methods of selecting a random sample.

(b) Mechanical randomization or random numbers method. The lottery method is time consuming, if the population is large. In random numbers methods, a randomly generated numbers' table known as random number table is used to draw the required sample. There are many tables of this types prepared by many professors and scientists. These tables are so constructed that each of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9 appears same number of times independently of each other. If we want to select a sample from a population of size N ($\leq$ 99) then the numbers can be combined two by two and we will get pairs from 00 to 99. Similarly if N ($\leq$ 999) or N ($\leq$ 9999), and so on, then we combine the numbers three by three for N ($\leq$ 999) and four by four for N ($\leq$ 9999), and so on. Since each of the digits 0, 1, 2,…..,9 occurs equal number of times independently of each other, so does each of the pairs 00 to 99 or triplets 000 to 999, or quadruplets 0000 to 9999, and so on.

The steps of drawing the random sample are as follows:

(i) Identify the N units in the population with numbers from 1 to N.

(ii) Select at random any page of the random number table and pick up the numbers row wise or column wise or diagonal wise at random.

(iii) The population units corresponding to these selected numbers form the required sample.

Some commonly used random number tables are:

1. Tippet's (1927) random number table. (Tracts for computers No. 15, Cambridge University Press).

2. Fisher and Yates (1938) Tables (in Statistical tables for Biological, Agricultural and Medical Research)
3. Kendal and Babington Smith's (1939) random number tables (Tracts for computers, No. 24, Cambridge, University, Press).
4. Rand Corporation (1955) random number table. (Free Press, Illinois).

Example1. Draw a random sample (without replacement) of size 15 from a population of size 500.

Solution First of all we identify 500 units in the population with numbers from 1 to 500. Then we select at random any page of the random number table discussed above row wise or column wise or diagonally, we select on by one three digited numbers, discarding the numbers over 500, until 15 numbers below 500 are obtained. Since we are using simple random sampling technique, the numbers selected previously will also be discarded. The 15 numbers finally, so selected will constitute the required sample. The following is an extract from the first set of 40 four-digited numbers in Tippet's random number tables:

2952	6641	3992	9792	7669	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2370	7483	3408	2762	3563	1089	6913	7691
0560	5246	0112	6107	6008	8126	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

The table for the selection of 15 units of the sample is as follows:

Table -1

S. No.	Random No.	Unit selected/ not selected
1.	295	Unit selected. Since, $295 \le 500$
2.	266	Unit selected. Since, $266 \le 500$
3.	413	Unit selected. Since, $413 \le 500$
4.	992	Discard random number. Since, $992 \ge 500$
5.	979	Discard random number. Since, $979 \ge 500$
6.	279	Unit selected. Since, $279 \le 500$

7.	695	Discard random number. Since, $695 \geq 500$
8.	911	Unit not selected. Since, $911 \geq 500$
9.	317	Unit selected. Since, $317 \leq 500$
10.	056	Unit selected. Since, $056 \leq 500$
11.	244	Unit selected. Since,$244 \leq 500$
12.	167	Unit selected. Since, $167 \leq 500$
13.	952	Discard random number. Since, $952 \geq 500$
14.	415	Unit selected. Since, $415 \leq 500$
15.	451	Unit selected. Since, $451 \leq 500$
16.	396	Unit selected. Since, $396 \leq 500$
17.	720	Discard random number. Since, $720 \geq 500$
18.	353	Unit selected. Since, $353 \leq 500$
19.	561	Discard random number. Since, $561 \geq 500$
20.	300	Unit selected. Since, $300 \leq 500$
21.	269	Unit selected. Since, $269 \leq 500$

Starting with first number and moving row-wise, the units in the population with the numbers:

295, 266, 413, 279, 317, 056, 244, 167, 415, 451, 396, 353, 300, 269, will be the 15 units selected in the required sample.

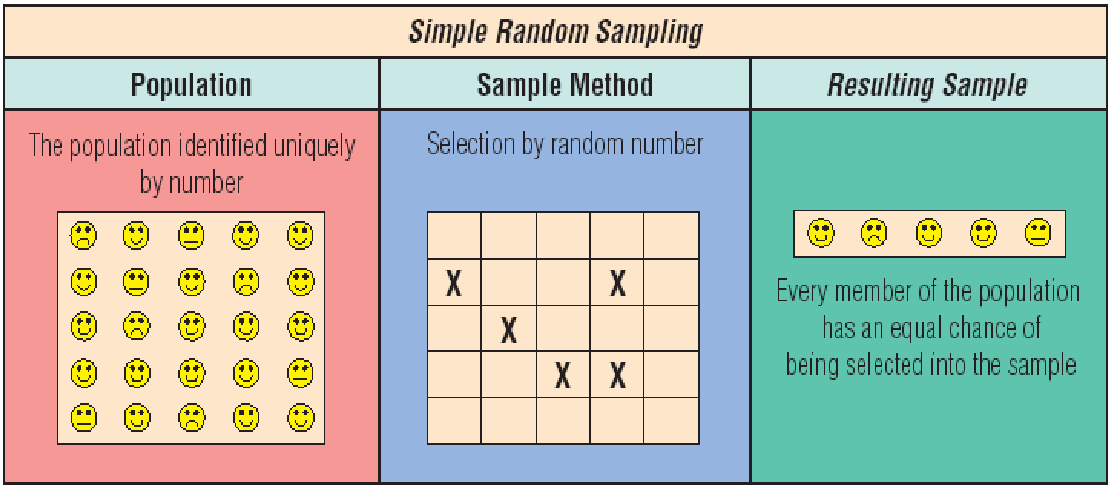

Simple Random Sampling
Population
Sample Method
Resulting Sample
The population identified uniquely by number
Selection by random number
X
X
X
X
X
Every member of the population has an equal chance of being selected into the sample

Figure-1.3Simple random sampling

5.7.2.2 Stratified random sampling

Stratified random sampling comes under the category of restricted sampling. When the population is heterogeneous with respect to some major characteristics, applying simple random sampling directly is not suitable. In such a situation, first of all the population is divided into homogeneous groups under certain criteria. These groups are termed as strata or stratum. From each stratum independent, independent samples are selected using any of the known sampling methods. If the sample selection is carried out using simple random sampling in each stratum, the sample design is called stratified random sampling.

Information about each individual sampling unit is rarely available. Hence, the strata are formed on some broad basis such as localities in a city, districts in a state, etc. If the population is heterogeneous, stratified random sampling is more efficient. This is because a large sample is necessary to get an estimate of a characteristic with the same precision, if we ignore stratification. To be more specific, if every person in the population has the same hemoglobin level, and then a sample of even one individual would be enough to get a precise estimate of the average hemoglobin level. Let us clarify it further: in stratified random sampling, sapling units within each stratum have similar characteristic (e.g., hemoglobin levels) but different from those in other strata (e.g., disease status). In such a case, only a small sample from each stratum may provide a precise estimate of the hemoglobin level for that stratum. The estimates obtained for each stratum may be combined to get a precise estimate of hemoglobin levels for the population. A simple random sampling approach to the entire population without stratification would require comparatively large sample size than the total of stratum-specific samples to obtain an estimate of hemoglobin level with the same level of precision.

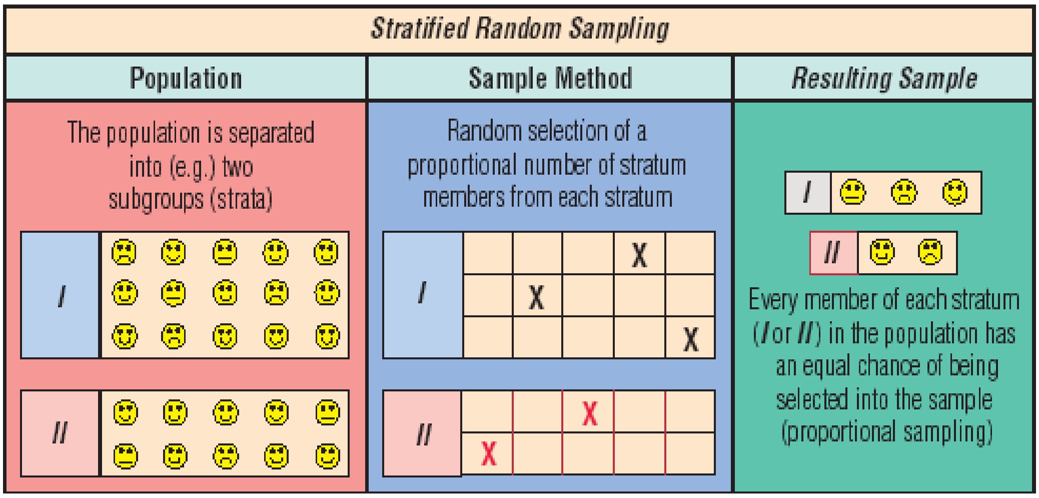

Stratified Random Sampling
Population
Sample Method
Resulting Sample
The population is separated into (e.g.) two subgroups (strata)
I
II
Random selection of a proportional number of stratum members from each stratum
X
I
II
X
Every member of each stratum (I or II) in the population has an equal chance of being selected into the sample (proportional sampling)

Figure-1.4 Stratified random Sampling

5.7.3 MIXED SAMPLING

If the samples are selected partly according to some laws of chance and partly according to a fixed sampling rule, i.e., there is no involvement of probabilities, they are termed as mixed samples and the technique of selecting such samples is known as mixed sampling.

5.8 ORGANIZATION AND REPRESENTATION OF DATA

5.8.1 ORGANIZATION

Data collected as such do not give any meaning. This is divided according to its type explained in 1.5 and then it is consolidated by way of tabulation. Rearrangements and grouping according to requirement and standards are done, thus summarizing tables from data tabulation, which give meaning to the information collected. Data are tabulated by (1) manual procedure (2) Mechanical procedure (3) Computer feeding. IN preparation of tables following principles are followed:

(i) A rough draft of the table should be prepared first. Before drawing out the final table, rough draft should be examined carefully.

(ii) Headings of the rows and columns should be brief and clear.

(iii) Title, note, row and column are made specific, connoting meaning or expressions.

(iv) Numbers of class intervals are decided as per aims of study which should not be too small or too big.

(v) Symbols used, should be explained.

(vi) Tabulated data should specify the units of their measurements.

(vii) The sources from which data are obtained should be given.

5.8.2 REPRESENTATION OF DATA

Tabulated data will give some information and also allow for further analysis. The columns and rows in a table make eye strain and there are chances of poor visual impression of data presented in a tabular form. Now the well tabulated data can be represented in the form of picture, diagram or figure which will help in good comparison through good visual impression. The representation of quantitative data through charts and diagrams is known as graphical representation of statistical data. A picture is said to be more effective than words for describing a particular thing or phenomenon. Main objective of diagram is to help the eye to grasp series of numbers and to grasp the meaning of series of data and also to assist the intelligence.

There are various types of graphs in the form of charts and diagrams. Some of them are:

5.8.2.1 Bar diagram

The simplest type of graph that can be used to represent the categorical data is the bar diagram. It is also called a columnar diagram. The bar diagrams are drawn through columns of equal width. In this diagram we show the category of the variable on the X-axis and the frequencies on the Y-axis on a graph paper. A bar of each category is of the variable is drawn and the height of the bar represents the frequency of that category. Since the data is of qualitative nature or quantitative data of discrete type, bars should not be next to each other and there should be an equal gap between two successive bars. Following rules were observed while constructing a bar diagram:

(a) The width of all the bars or columns is similar.

(b) All the bars should are placed on equal intervals/distance.

The following types of bar graphs are possible:

(a) Simple bar graph

(b) Double bar graph

(c) Multiple bar graphs

We will illustrate each of these graphs by the following illustrations:

(a) Simple Bar Diagram

A simple bar diagram is constructed for an immediate comparison. It is advisable to arrange the given data set in an ascending or descending order and plot the data variables accordingly. In Base hospital has been found patients in OPD in particular disease as below in year 2012.

Month:	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct	Nov	Dec
Patients:	285	315	250	289	386	410	452	620	421	186	450	500

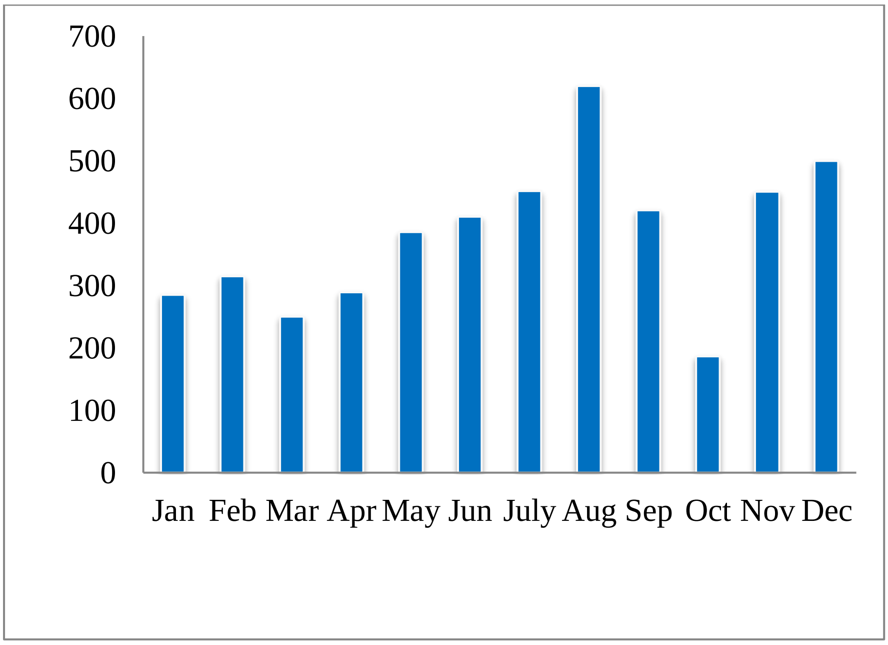


Figure -1.5 simple bar diagram

(b) Double Bar Diagram

When two components are grouped in one set of variable or different variables of one component are put together, their representation is made by a double bar diagram. In this method, different variables are shown in a single bar with different rectangles. From above example, patients were divided in two categories as male and female and the data is given below:

Month:	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Male	100	250	150	189	270	200	350	275	215	86	300	200

Female: 185 115 100 100 116 210 102 345 206 100 150 300

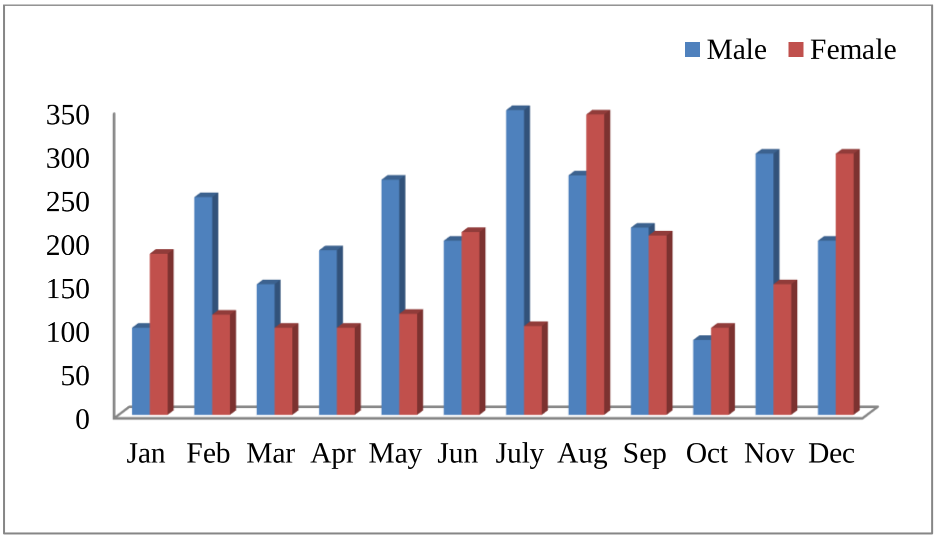


Figure -1. 6 Double Bar Diagram

(c) Multiple Bar Diagram

Multiple bar diagram shows that the proportion of subgroup between two or more categories are represented with a bar giving proportion to each of them within the bar. It is also advisable to make one bar as 100% and each subcategory is given proportion within the graph.

5.8.2.2 Pie Chart

Pie diagram is another graphical method of the representation of categorical data. Pie is a mathematical constant defined as the ratio of the circumference of a circle to the diameter and is equal to 22/7. It is drawn to depict the total value of the given attribute using a circle. In the pie chart, a circle (total 360^{o}) is divided into sectors with areas proportional to the frequencies or the relative frequencies of the categories of a variable. Dividing the circle into corresponding degrees of angle then represent the sub– sets of the data. Hence, it is also called as Divided Circle Diagram.

Example 2. A household with a monthly salary of Rs. 7200 plans his budget for a month as given below:

Item	Food	Rent	Education	Savings	Misc.	Total
Amount (Rs.)	3000	800	1200	1500	700	7200

Make a pie chart for this data.

Solution. First of all we find the angles of each sector as follows:

Total of data corresponds to 360°. Let x° = the angle at the centre for item A, then for the data given in above example to draw pie graph, we find the angles of each category.

Calculation of Angles

For Food:

Angle at centre = $\frac{f}{\sum f}\times 360^{o} = \frac{3000}{7200}\times 360^{o} = 150^{o}$. Here f= Frequency of food and $\sum f$ = Total frequency

For Rent:

Angle at centre = $\frac{f}{\sum f}\times 360^{o} = \frac{800}{7200}\times 360^{o} = 40^{o}$

Similarly, we can calculate the remaining angles, and the total of angles column should always come to 360^{o}.

Table-2

Item	Amount (Rs.)	Angle
Food (A)	300	150
Rent (B)	800	40
Education (C)	1200	60

Savings (D)	1500	75
Miscellaneous	700	35
Total	7200	360

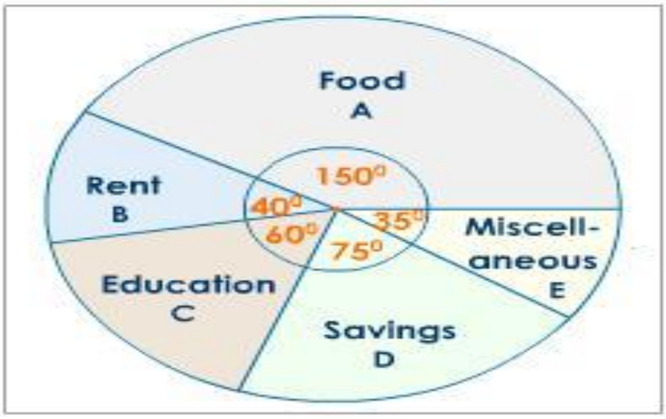


Figure-1.7Pie chart

5.8.2.3 Histogram

A two dimensional frequency density diagram is called a histogram. A histogram is a diagram which represents the class interval and frequency in the form of a rectangle. There will be as many adjoining rectangles as there are class intervals. There are two types of histograms-

(1) Histogram with equal class intervals

(2) Histogram with unequal class intervals

To draw a histogram, you should follow the steps as stated below:

1. Class intervals must be exclusive. If the intervals are in inclusive form, convert them to the exclusive form.
2. Draw rectangles with class intervals as bases and the corresponding frequencies as heights.
3. If the intervals are equal, then the height of each rectangle is proportional to the corresponding class frequency.
4. If the intervals are unequal, then the area of each rectangle is proportional to the corresponding class frequency density.

Example 3. Draw a histogram for the following data showing the class interval and their corresponding frequencies.

Class interval	0-5	5-10	10-15	15-20	20-25
Frequency	4	10	18	8	6

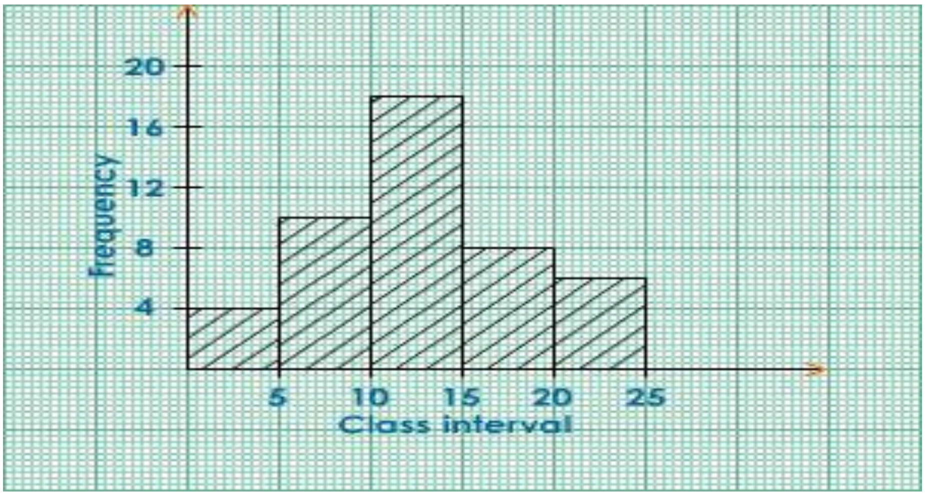


Figure-1.8 Histogram

Example 4. Following is the distribution of shops according to the number of wage - earners employed at a shopping complex.

Table-3 showing the distribution of wage earners

Number of wage earners	No. of shops	Frequency density
Under 5	18	3.6
5 – 10	27	5.4
10 – 20	24	2.4
20 – 30	20	2.0
30 – 50	16	0.8

Illustrate the above table by a histogram, showing clearly how you deal with the unequal class intervals.

Solution. When the class intervals are unequal, we construct each rectangle with the class intervals as base and frequency density as height.

Frequency density = Frequency/ Class width

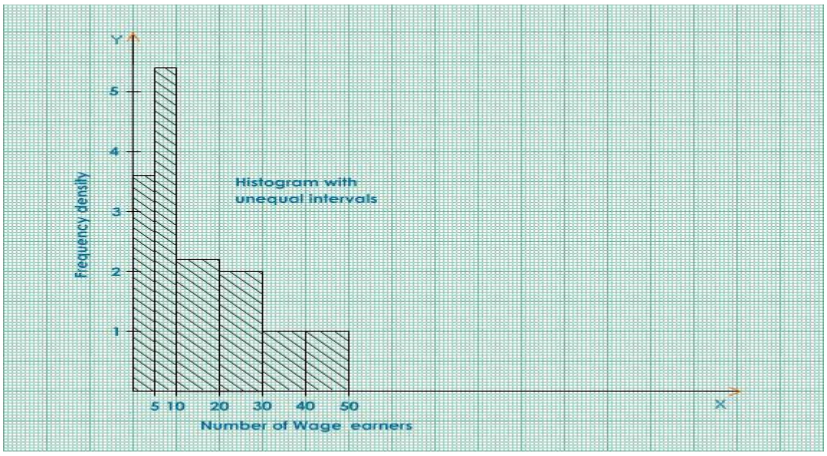


Figure- 1.9

5.8.2.4 Frequency Polygon and Frequency Curve

In a frequency distribution, the mid-value of each class is obtained. Then on the graph paper, the frequency is plotted against the corresponding mid-value. These points are joined by straight lines. These straight lines may be extended in both directions to meet the X - axis
to form a polygon. If these points are joined by a free hand smooth curve then it is called
Frequency curve.

Example 5. The growth rate of different crops like rice, wheat, birth rates, death rates and life expectancy are given in the following table. Make a frequency polygon from it.

Table-4 Showing class interval and frequency

Class interval	Mid Marks	Frequency
40 – 44	42	3
45 – 49	47	10
50 – 54	52	12
55 – 59	57	15

60 – 64	62	7
65 – 69	67	5

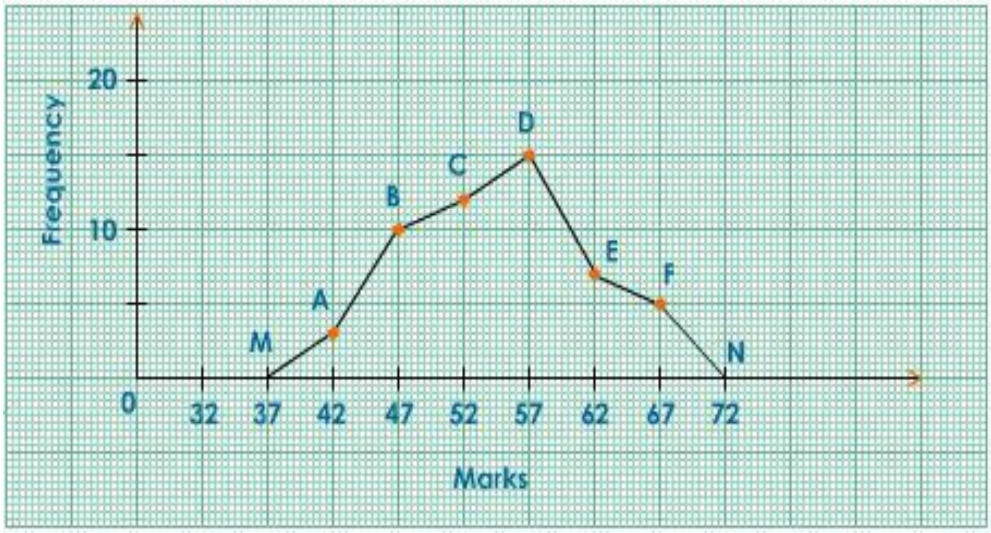


Figure-1.10

5.8.2.5 Pictograms

Pictograph is the use of pictures or images to present data. They will give the quick idea for the frequency of the characteristics and fraction also marks on pictures, e.g., bus for transport, man for cases, cot for hospital beds, etc. It is widely used by government and private organizations. The chief advantage of this method is its attraction.

5.8.2.6 Line chart

It is most widely used in medical science. It shows the trend of times. Data having some order as age – wise incidence of a disease can be represented by a line chart. It is drawn by taking one variable on the horizontal X-axis and the other variable on the vertical Y-axis. This graph shows the effect of one variable on the other variable, e.g., age specific incidence of cancer among males of Delhi.

Cumulative frequency curve

If we plot the less than cumulative frequencies rather than frequencies against the upper limits of the classes, the curve obtained on joining these points by free hand curve is called less than cumulative frequency curve or ogive or less than ogive and If we plot the more than cumulative frequencies rather than frequencies against the lower limits of the classes, the curve obtained on joining these points by free hand curve is called more than cumulative frequency curve. The advantage of this curve is that it enables us to answer the queries related to the frequency distribution of the variable.

5.8.2.8 Scatter diagram

It is the simplest way of the representation of bivariate data. Thus for the bivariate distribution (x, y) ; if the values of the Variable X and Y be plotted as x along X-axis and the y along the Y-axis respectively in the x y plane, the diagram of dots so obtained is called scatter diagram.

5.9 SUMMARY

From the study of this chapter the students came to know the definitions of statistics and biostatistics, the scope and applications of biostatistics. The students studied and learnt about data. What is data? What are the types of data? The classification of different types of data provides knowledge to treat different types of data. We learn from the study of this chapter the different steps necessary for adopting any sampling procedure and the two types of error involved in the collection of sample and complete census. We learn definitions of some terms related to sampling as population, sample, sampling frame and sampling unit. Also we learn many sampling techniques for collecting sample data. In the last of this chapter we study and learn the organization and many representation techniques of data through graphs, pictures and chart.

5.10 GLOSSARY

Data: Collected and recorded information which is either in numerical form otherwise it has no meaning.

Qualitative data: The data collected on some quality characteristic.

Quantitative data: The data collected on quantitative variable.

Variable: Any attribute, any phenomenon or any event that can have different values.

Tabulation: It is the process to [put the data into different groups or classes.

Sampling Frame: It is the list or map of population units on the basis the selection of a sample is carried out.

Srswr: Simple random sampling a technique where each unit of the population has equal chance of selection in the sample and when sampling is done with replacement, i.e., selected unit is replaced before the next unit selection is made.

Srswor: It is also simple random sampling technique but without replacement, i.e., the unit selected is not replaced back.

Stratified random sampling: It is the technique of random sample when the population has heterogeneity by dividing it into homogeneous groups.

Bar diagram: Diagrammatic representation technique of frequency data for nominal classes by bars in which the length of bars is proportional to the class frequencies.

Cumulative frequency curve: The curve in which the cumulative frequencies rather than frequencies are plotted against the class intervals. These are of two types less than and more than types and their intersection represent median.

Histogram: It is a diagrammatic representation of the frequency distribution of a quantitative data with areas of the rectangles proportional to the class frequency.

5.11 SELF ASSESMENT QUESTIONS

Question 1. For the given data draw a bar chart.

Year	1978	1979	1980	1981	1982	1983
Rice (in tons)	4500	5700	6100	6500	4300	7800

Question 2. For the data give below make a pie chart.

Blood group	A	B	AB	O	Total
Frequency	15	25	20	30	90

Question 3. Make a histogram and frequency polygon for the data given below.

Profit per shop	0-100	100-200	200-300	300-400	400-500	500-600
No. of shops	12	18	27	20	17	6

Question 4. What are the different types of data? Explain in brief. Also explain the different sampling schemes of data collection.

Question 5. The data collected directly by the investigator or his/her team is called

(a) Secondary data (b)Primary data

(c) Population data (d) Sample data

Question 6. The data collected from already published material is called

(a) Secondary data (b)Primary data

(c) Population data (d) Sample data

Question 7. In bar diagram the base line is:

(a) Horizontal (b) Vertical

(c) False base line (d) any of the above

Question 8. Less than and more than ogives intersects at;

(a) mean (b) median

(c) mode (d) origin

Question 9. In cases of frequency distribution with classes of unequal width, the heights of bars in a histogram are proportional to:

(a) class frequency (b) class intervals

(c) frequencies in percentage (d) frequency densities

Question 10. Sampling frame is a term used for:

(a) a list of random numbers (b) a list of voters

(c) a list of sampling units of the population (d) none of the above

Question 11. A selection procedure of a sample having no involvement of probability is:

(a) Purposive sampling (b) Judgment sampling

(c) Subjective sampling (d) all the above

Answers: 5- (b), 6- (a), 7- (b), 8- (b), 9- (d), 10- (c), 11- (d).

5.12 REFERENCES

Agarwal, B.L., Programmed Statistics, New Age International (P) Ltd, New Delhi, (2003).

Arora, P.N. and Malhan, P.K., Biostatistics, Himalaya Publishing House (P) Ltd. Mumbai, (1996).

Belle, G., Fisher, L., Heagerty, P.G. and Lumly, T.,Biostatistics- A Methodology for the Health Science, John Willey & Sons, inc. Pub., Hoboken, New Jersey, (2004).

Gupta, S.C. and Kapoor, V.K., Fundamentals of Mathematical Statistics, Sutan Chand & sons, Daryaganj, New Delhi, (1970).

Gupta, S.C. and Kapoor, V.K., Fundamentals of Applied Statistics, Sutan Chand & sons, Daryaganj, New Delhi, (1976).

Pagaro, M. and Gauvreau, K., Principles of Biostatistics, Duxbury Thomson Learning, U.S.A. (2000).

Prabhakara, G.N., Biostatistics, Jaypee Brothers Medical Publishers (P) Ltd, New Delhi, (2006).

Singh, D. and Chaudhary, F.S., Sample Survey Designs, New Age International (P) Ltd, New Delhi, (1986).

Sudaram, K.R., Dwivedi, S.N., Sreenivas, V., Medical Statistics, Wolters Kluwer (P) Ltd, New Delhi. (2010).

5.13 TERMINAL QUESTIONS

1. What is sampling frame?

2. What do you know about histogram?

3. Explain stratified random sampling

4. What is the difference between simple random sampling with and without replacement?

5. What are different types of graphs or charts for data representation? Explain any two of them.

6. Explain the difference between sample and population.

7. What is the need of biostatistics?

Unit-6– MEASURES OF CENTRAL TENDENCY

CONTENT

6.1 OBJECTIVES

From the study of this chapter the students will be able:

1. To know about the measures of central tendency- mean, median and mode.
2. To know the merits and demerits and uses of these measures.
3. To know about different methods of measuring mean, median and mode.
4. To know the situations where which measure is better to use?
5. To know the advantages of short cut methods of computing mean.

6.2 INTRODUCTION

In the previous chapter, we discussed data collection, data organization and data representation techniques. The data representation techniques such as frequency histograms and frequency polygons, introduced the concept of the shape of distributions of data. For example, a frequency polygon illustrated the distribution of body mass index data. We expend chapter 1 on these concepts by defining measures of central tendency.

Measures of central tendency as the name suggests are numerical measurement of the central part of the distribution. Measures of central tendency are also called averages or measures of location because they show the location of the centre of the distribution from which the data were sampled. According to Professor Bowley, averages are, “statistical constants which enable us to comprehend in a single effort the significance of the whole.” In other words, these are numbers that tell us where the majority of values in the distribution are located. For example the average marks in a distribution of marks of all the students of a class. The averages which are commonly used in biostatistics are as follows:

1. Mean or arithmetic mean 2. Median 3. Mode

6.3 MEAN

Mean or arithmetic mean of a series of data is the ratio of the sum of the observations to the number of observations. If $x_1, x_2, ...x_n$ are the observations of a series then their arithmetic mean is given by

$$\bar{x} = \frac{x_1 + x_2 + ...x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (1)$$

And if the corresponding frequencies, $f_1, f_2, ...f_n$ of the variables $x_1, x_2, ...x_n$ are given, then the arithmetic mean is defined as ratio in which the numerator is the sum of products of the variables with their frequencies and denominator is the sum of the frequencies.

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + ...f_n x_n}{\sum f_i} = \frac{\sum_{i=1}^{n} f_i x_i}{N} \qquad (2)$$

where, N = $\sum f_i$ = sum of frequencies.

6.3.1 MEAN OF INDIVIDUAL ITEMS

Mean of individual items is given by the ratio of the sum of items to the number of items as given in formula (1).

Example 1. Find the arithmetic mean of triglycerides present 10 patients in their blood samples in a hospitalas:

25, 30, 21, 55, 47, 10, 15, 17, 45, 35

Solution. Let $\bar{x}$ be the average triglyceride value and since these are individual items, their mean can be computed by formula

$$\bar{x} = \frac{x_1 + x_2 + ...x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$
$$= \frac{25 + 30 + 21 + 55 + 47 + 10 + 15 + 17 + 45 + 35}{10} = \frac{300}{10} = 30$$

6.3.2 MEAN IN DISCRETE FREQUENCY DISTRIBUTION

If $x_1, x_2, ...x_n$ are the observations in a discrete distribution according to some characteristic and $f_1, f_2, ...f_n$ be their corresponding frequencies then the arithmetic mean is given by the formula (2). The computation procedure for mean can be easily understood with the help of the example given below.

Example 2. The distribution of marks of 50 students of B.Sc. class in a botany semester examination is given below. Find the average of marks.

Marks (x)	12	23	25	35	45	15	40
Frequency(f)	3	10	12	10	2	8	5

Solution. Since this is a discrete distribution so the average of marks is given by the formula (2). For the computation of average marks we prepare the following table:

Table1. For calculation of Mean in Discrete Distribution

Marks (x)	Frequency (f)	f x
12	3	36
23	10	230
25	12	300
35	10	350
45	2	90
15	8	120
40	5	200
Total	$\sum f = 50$	$\sum fx = 1326$

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + ...f_nx_n}{\sum f_i} = \frac{\sum_{i=1}^{n} f_ix_i}{N} = \frac{1326}{50} = 26.52$$

This is clear that it is not necessary that average will be a number presenting in the data and also it is not an integer value while the marks in integers.

6.3.3 MEAN IN CONTINUOUS DISTRIBUTION

In case of continuous distribution, there are given class intervals and their corresponding frequencies. First of all we find the mid values of these classes and treat them as the variable values. Now we apply the formula (2) for the calculation of arithmetic mean. The procedure will be clear from the following example.

Example 3. For the data given in the below table on systolic BP of 68 patients, calculate the arithmetic mean.

Table 2.

Systolic BP (mmHg)	Frequency (f)	Systolic BP (mmHg)	Frequency (f)
90-100	3	140-150	11
100-110	5	150-160	9
110-120	7	160-170	6
120-130	10	170-180	2
130-140	15		

Solution. For the calculation of mean we prepare the following table:

Table3. For calculation of Mean in Continuous Distribution

Systolic BP (mmHg)	Frequency (f)	Mid Value (x)	Fx
90-100	3	95	285
100-110	5	105	525
110-120	7	115	805
120-130	10	125	1250

130-140	15	135	2025
140-150	11	145	1595
150-160	9	155	1395
160-170	6	165	990
170-180	2	175	350
Total	$\sum f = 68$		$\sum fx = 9220$

$$\bar{x} == \frac{\sum_{i=1}^{n} f_i x_i}{N} = \frac{9220}{68} = 135.6 mmHg$$

6.3.4 SHORT-CUT METHOD FOR MEAN

For the computation of mean short –cut method is applied when the variable values and their frequencies are large. To make the computations easy we take a middle value in the given values of x as assumed mean and subtract this assumed mean from all the values of x. This assumed mean is also called provisional mean. Then the formula for the calculation of arithmetic mean is given by as follows:

$$\bar{x} = A + \frac{\sum d}{n} \qquad (3)$$

Where A= assumed mean, n= number of observations in the given data

d= x- A= deviation of all the variate values from assumed mean A.

Steps of computation for short- cut method:

Step 1. Take any observation (generally, middle value if we arrange the values in ascending or descending order of magnitude) of the individual series as assumed mean A.

Step 2. Find the deviation of the values of variate x from assumed mean A, i.e., calculate the differences d= x- A

Step 3. Find the sum of d and use above formula (3), we find the value of mean.

If the frequencies corresponding to the variate values are given, then we use the formula for mean as follows:

$$\bar{x} = A + \frac{\sum fd}{N} \qquad (4)$$

Where, N = $\sum f$ = sum of frequencies. Here we find the product of f and d.

If the data is continuous, we find the mid values as x and then d= x- A. Now apply the above formula (4). The procedure will be clear from the examples as give below.

Example 4. The marks of the 7 students of a class in a test are as given below:

12, 15, 22, 25, 35, 40, 45

Find the mean by short-cut method.

Solution. Let us take assumed mean A=25. Now we prepare the table for the computation of mean as given below:

Table-4 Mean for individual data by short cut method

X	d = x- 25
12	-13
15	-10
22	-3
25	0
35	10
40	15
45	20
Total	$\sum d = 19$

Arithmetic mean $\bar{x} = A + \frac{\sum d}{n} = 25 + \frac{19}{7} = 25 + 2.71 = 27.71$

Thus the average of marks of the given 7 students of the class is 27.71

Example 5. Ten patients were examined for uric acid test. The operation was performed 1050 times and the frequencies so obtained for different number of patients (x) are shown in the table given below. Compute the arithmetic mean by short- cut method.

x:	0	1	2	3	4	5	6	7	8	9	10
f:	2	8	43	133	207	260	213	120	54	9	1

Solution. Let 5 be the assumed mean. Now we prepare the table for the calculation of mean.

Table-5 Mean for discrete grouped data by short cut method

X	Frequency (f)	d = x- 5	fd
0	2	-5	-10
1	8	-4	-32
2	43	-3	-129
3	133	-2	-266
4	207	-1	--207
5	260	0	0
6	213	1	213
7	120	2	240
8	54	3	162
9	9	4	36
10	1	5	5
Total	$\sum f = 1050$		$\sum fd = 12$

Arithmetic mean $\bar{x} = A + \frac{\sum fd}{N} = 5 + \frac{12}{1050} = 5 + 0.0114 = 5.0114cm$

Thus the average for uric acid is 5.0114.

6.3.5 STEP DEVIATION METHOD OF MEAN

It can be used in grouped data. When all the classes are of equal width (say h), in continuous data and the values of x are at equal interval in discrete grouped data then the we may simplify the calculations by taking d= (x- A)/ h in short-cut method. Now the formula for the calculation of mean becomes.

$$\bar{x} = A + \frac{\sum fd}{N} \times h$$

Here, the symbols have the same meaning as in short-cut method above and h is the gap between the two values of x or class interval.

Example 6. Find the mean by step deviation method for the data of blood pressure of 68 patients as given in the following table.

BP(mmHg) (x)	90	100	110	120	130	140	150	160	170
Frequency (f)	3	5	7	10	15	11	9	6	2

Solution. We take assumed mean A= 130 and here interval between any two values of x is 10, i.e., h= 10. Now prepare the table for the computation of mean.

Table6. Step Deviation Method of Mean in discrete grouped data

BP (mmHg)	Frequency (f)	$d = \frac{x-130}{10}$	fd
90	3	-4	-12
100	5	-3	-15
110	7	-2	-14
120	10	-1	-10
130	15	0	0

140	11	1	11
150	9	2	18
160	6	3	18
170	2	4	8
Total	$\sum f = 68$		$\sum fd = 4$

Arithmetic mean $\bar{x} = A + \frac{\sum fd}{N} \times h = 130 + \frac{4}{68} \times 10 = 130 + 0.588 = 130.588 mmHg$

Thus the average for BP is 130.588mmHg.

Example7. For example 3, calculate arithmetic mean by step deviation method.

Solution. For the calculation of mean the table is given below:

Table-7. Mean by Step Deviation Method in Continuous Data

Systolic BP (mmHg)	Frequency (f)	Mid Value (x)	$d = \frac{x-135}{10}$	fd
90-100	3	95	-4	-12
100-110	5	105	-3	-15
110-120	7	115	-2	-14
120-130	10	125	-1	-10
130-140	15	135	0	0
140-150	11	145	1	11
150-160	9	155	2	18
160-170	6	165	3	18
170-180	2	175	4	8
Total	$\sum f = 68$			$\sum fd = 4$

Arithmetic mean $\bar{x} = A + \frac{\sum fd}{N} \times h = 135 + \frac{4}{68} \times 10 = 135 + 0.588 = 135.588 mmHg$

Thus the average for uric acid is 135.588mmHg.

Example 8. In a study on patients of typhoid fever the following data are obtained. Find the arithmetic mean.

Age in years	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89
No. of cases	1	0	1	10	17	38	9	3

Solution. This is inclusive type data; first of all we convert it to exclusive type data. The procedure for converting inclusive type data to exclusive type data is as follows:

We see that the upper limit of the first class is 19 and the lower limit of the second class is 20 and their difference is 20-19=1. Now subtract half of the difference, i.e., 0.5 from the upper limit and 0.5 to the lower limit. Also we see that this difference is the same for each of the class. So the new classes are as 9.5-19.5, 19.5-29.5 and so on.

Now for the calculation of mean any method discussed above can be used. Here we apply step deviation method.

Table-8. Mean by Step Deviation Method in Inclusive Data

Age	Frequency (f)	Mid Value (x)	$d = \frac{x - 44.5}{10}$	fd
9.5-19.5	1	14.5	-3	-3
19.5-29.5	0	24.5	-2	0
29.5-39.5	1	34.5	-1	-1
39.5-49.5	10	44.5	0	0
49.5-59.5	17	54.5	1	17

59.5-69.5	38	64.5	2	76
69.5-79.5	9	74.5	3	27
79.5-89.5	3	84.5	4	12
Total	$\sum f = 79$			$\sum fd = 128$

Arithmetic mean $\bar{x} = A + \frac{\sum fd}{N} \times h = 44.5 + \frac{128}{79} \times 10 = 44.5 + 16.2 = 60.7$

6.3.6 WEIGHTED MEAN

In computation of arithmetic mean some items are more important than the others, in such cases the weight age should be given to the items according to their importance. For example if we want to have an idea of the change in cost of living of group of people of a certain locality, then the simple mean of the prices of the commodities consumed by them will not do, since all the commodities are not equally important, e.g., wheat, rice and pulses are more important than cigarettes, tea, confectionery, etc.

If $x_1, x_2, ...x_n$ are the variate values of a distribution and $w_1, w_2, ...w_n$ be their corresponding weights then weighted mean is give by:

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$$

Example 9. The following table gives the platelets count (in lakh/cmm) from the analysis of the blood samples on five different days in a pathology laboratory. Find the average platelets count per patient.

Day	1	2	3	4	5
Platelates count (in lakh/cmm) (w)	0.50	0.75	1.00	1.50	2.00
No. of patients (x)	65	80	95	90	70

Solution. The table for the calculation of weighted mean is given by:

Table 9. Table for Weighted Mean

Platelets count (x)	No. of patients (w)	Wx
0.50	65	32.5
0.75	80	60.0
1.00	95	95.0
1.50	90	135
2.00	70	140
Total	$\sum w = 400$	$\sum wx = 462.5$

$$x_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{462.5}{400} = 1.156$$

Thus, the average platelets per patient are 1.156 lakh/cmm.

6.3.7 COMBINED MEAN

If $\bar{x}_1, \bar{x}_2, ...\bar{x}_m$ are the means of m series of sizes $n_1, n_2, ...n_m$ respectively, then their combined arithmetic meaning $\bar{x}$ is given by:

$$\bar{x} = \frac{\sum n_i \bar{x}_i}{\sum n_i}; i = 1,2,......m$$

Example 10. There are 40 male and 10 female employees in a firm. The mean salary of male employees is Rs.520 and that of female employees Rs. 420. Find the combined average salary of all the employees.

Solution. Here, $n_1 = 40, n_2 = 10, \bar{x}_1 = 520, \bar{x}_2 = 420$

Combined mean $\bar{x} = \frac{\sum n_i \bar{x}_i}{\sum n_i} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{520 \times 40 + 420 \times 10}{40 + 10} = \frac{25000}{50} = 500$

Hence, the average salary of all the employees is Rs. 500.

6.3.8 CORRECTED MEAN

Some times there are problems of such type that we used wrong digits while the actual digits were different, then we replace the wrong digits with the correct digits and now we can get the correct mean. The procedure will be clear from the example given below.

Example 11. A student calculates the mean of 20 observations as 25.2. Later on he found that he misread one observation 34 in place of 43, find the correct mean.

Solution. We know the mean of individual series is given by:

$$\bar{x} = \frac{\sum x}{n} or \sum x = n\bar{x} = 20 \times 25.2 = 504$$

But he misread 43 as 34. So the correct total of x= 504-34+43=513.

So correct mean=513/ 20 =25.65

6.3.9 MERITS, DEMERIRS AND USES OF MEAN

Merits:

1. Mean is rigidly defined.
2. It can be calculated easily by a non mathematical person also.
3. It is based upon all the observations.
4. Among all the averages, it is affected least by fluctuations of sampling.
5. It is the best measure to compare two or more series.
6. It is easily understandable.
7. It is the most widely used method of central tendency.

Demerits:

1. It is affected much by extreme values.
2. It cannot be calculated in case of open end classes.
3. It cannot be calculated in case of qualitative data such as intelligence, beauty, etc.

`4. In extremely asymmetrical distribution, mean is not a suitable measure of central tendency.

5. It cannot be calculated if any observation is missing.

6. It may lead to wrong conclusions if the details of the data are not given. For example the marks of two students in three successive tests are respectively 30, 40, 50 and 50, 40, 30. We see that average score of both the students is same, we can say that both students are of same level while first is improving and the second is deteriorating.

Uses of Mean:

1. It is very much used in practical situations.

2. A common man uses it for computing his monthly budget.

3. It is very much used in sampling and inference.

4. A businessman uses it for computing per unit profit, output per person, average expenditure and average profit per week or per month, etc.

6.*4 MEDIAN*

Median of a distribution is the middle most value of the variable if the values of the variable are arranged in ascending or descending order of their magnitude. The median divides the observations of the variable in such a way that half of the observations of the variable lie above the median and half below this. Median is thus called a positional average because it locates at the middle of the observations. But if the number of observations is even then after arrangement there will be two middle values and the median will be the average of these two middle values.

6.4.1 MEDIAN IN INDIVIDUAL SERIES

Arrange the data observations, (say n) in ascending or descending order of magnitude. Now there can be two cases:

Case 1. If n is odd then middle most, i.e., (n+1/2)th term value is the median.

Case2. If n is even then there are two middle terms (n/2)th and (n+1/2)th, then median is given by:

$$M_e = \frac{\left[\frac{n}{2}th + \left(\frac{n}{2}+1\right)th\right]term}{2}$$

Example 12. The marks of 9 children in a test exam are: 12, 23, 34, 11, 14, 15, 13, 16, 45.

Find the median of the marks.

Solution. Arrange the given observations in ascending order of magnitude, we get

`11, 12, 13, 14, 15, 16, 23, 34, 45

Here the number of observations n =9, i.e., odd.

So the median is the (9+1)/2 th, i.e., 5th term value, i.e., 15.

Example 13. The number of blood LDL (in mg/dl) present the blood samples of 12 patients are: 5, 19, 42, 11, 50, 30, 21, 0, 22, 52, 36, 27

Find the median of the data.

Solution. On arranging the given observations in ascending order of magnitude, we get,

0, 5, 11, 19, 21, 22, 27, 30, 36, 42, 50, 52

Here number of observations = 12, i.e., even. So median is given by

$$M_e = \frac{\left[\frac{n}{2}th + \left(\frac{n}{2}+1\right)th\right]term}{2} = \frac{\left[\frac{12}{2}th + \left(\frac{12}{2}+1\right)th\right]term}{2}$$

$$= \frac{(6th+7th)term}{2} = \frac{22+27}{2} = \frac{49}{2} = 24.5$$

So median is 24.5 mg/dl which does not belong to the data. So in case of even number of observations median is not present in the data observations.

6.4.2 MEDIAN IN DISCRETE FREQUENCY DISTRIBUTION

If $x_1, x_2, ...x_n$ are the observations in a discrete distribution according to some characteristic and $f_1, f_2, ...f_n$ be their corresponding frequencies then for the calculation of median we calculate the cumulative frequencies. The median is calculated with the help of the following steps.

Working steps for median

Step1. Arrange the given values in ascending order of magnitude.

Step2. Find the total of frequencies, called cumulative frequency and denoted by c.f. Step3. Find $\frac{N}{2}$, where N=$\sum f$.

Step4. Find cumulative frequency just greater than $\frac{N}{2}$. The value of x corresponding to this cumulative frequency is the required median.

Example14. Find the median for the following data.

X	21	15	17	9	5	7	8	10
F	2	5	3	4	5	1	6	12

Solution. For calculating the median we arrange the values of x in ascending order and then prepare the cumulative frequency table as follows:

Table10. Median in Discrete Distribution

x	F	c.f.
5	5	5
7	1	6
8	6	12
9	4	16
10	12	28
15	5	33
17	3	36
21	2	38

$$N = \sum f = 38$$

Here N/2 = 38/2= 19 and cumulative frequency just greater than 19 is 28. The value of x corresponding to cumulative frequency 28 is 10. So the median of the given data is 10.

6.4.3 MEDIAN IN CONTINUOUS FREQUENCY DISTRIBUTION

When the data is in class interval form, the class corresponding to c.f. just greater than N/2 is called the median class and the median is computed by the following formula:

$$M_e = L + \frac{\left(\frac{N}{2} - C\right)}{f} \times h$$

Where L= lower limit of the median class

C= cumulative frequency just before the median class

N= Total of frequency

f = frequency of the median class

h = magnitude of the median class

Example15. The following table gives the distribution of weights of 100 persons. Find the median of this data.

Weight	40-45	45-50	50-55	55-60	60-65	65-70	70-75	75-80	80-85	85-90
Frequency	1	3	6	10	15	25	15	10	11	4

Solution. For computing the median we prepare the following table:

Table11. Median in Continuous Distribution

Weight (in kg) (x)	Frequency (f)	Cumulative Frequency (c.f.)
40-45	1	1
45-50	3	4
50-55	6	10
55-60	10	20
60-65	15	35

65-70	25	60
70-75	15	75
75-80	10	85
80-85	11	96
85-90	4	100
Total	$N=\sum f=100$	

Here N/2 = 10/2= 50 and cumulative frequency just greater than 50 is 60. The class corresponding to cumulative frequency 60 is 65-70. So class 65-70 is the median class. Now median is given by:

$$M_e = L + \frac{\left(\frac{N}{2} - C\right)}{f} \times h$$

here L= lower limit of the median class= 65

C= cumulative frequency just before the median class=35

N= Total of frequency =100

f = frequency of the median class=25

h = magnitude of the median class=5

$$M_e = 65 + \frac{(50-35)}{25} \times 5 = 65 + \frac{15}{25} \times 5 = 65 + 3 = 68$$

So median of the given data is 68kg

6.4.4 MERITS, DEMERIRS AND USES OF MEDIAN

Merits:

1. Median is rigidly defined.

2. It is not affected at all from extreme values.

3. It is easy to understand and to calculate.

4. In case of individual series data it can be located merely by inspection.

5. It can be calculated in case of open end classes.

6. Its graphical representation is also possible.

7. It can be computed even if the classes are of unequal interval.

8. In case of qualitative data, e.g., beauty, honesty, intelligence, etc. it is the best measure of central tendency.

Demerit:

1. It is not amenable to algebraic treatment.

2. It is a positional average and is based only on the middle term. It does not use all the observations of the data.

3. In case of irregular distribution, it is not a good measure.

4. In case of even number of observations it cannot be determined exactly, it can be estimated only by the average of the two middle terms.

5. In comparison to mean it is affected much by fluctuations of sampling.

Uses:

1. It is a good measure if numerical measurements are not possible.

2. In case of qualitative data where the observations cannot be determined quantitatively, it is the only average.

3. It is generally used in studying the average intelligence or average honesty of a group of people.

6.5 MODE

Mode is the most frequent item of the series, i.e., in a given set of observations a item or observation which is repeated maximum number of times an all other observations cluster around this, is called mode. For example, the average height of an Indian male is 5 feet 6 inch; the average size of the shoes of an Indian male is number 7, etc. Mode is also known as norm.

6.5.1 TYPES OF MODE OF A DISTRIBUTION

Unimodal: If the data of a distribution has only one mode then the distribution is called unimodal.

Bimodal: If we find that there are two items in a distribution which have the same number of repetitions, then these two items are the modes and the distribution is called bimodal.

Trimodal: Similarly, in a distribution, if there are three such items that they have the same frequency then these three items are called the modes of the distribution and the distribution is called trimodal.

Ill- defined mode: If there exists more than one mode in a distribution, then mode is called ill-defined.

6.5.2 MODE IN IDDIVIDUAL SERIES

In case individual series mode is the most frequent observation. It is clear from the following example.

Example 16. Find the mode of the series given below:

2, 3, 4, 7, 9, 3, 2, 1, 5, 3, 6, 3, 8, 3

Solution. In the given series the observation 3 is repeated maximum number of times (5) so the mode of the given series is 3.

6.5.3 MODE IN DISCRETE FREQUENCY DISTRIBUTION

In case of discrete frequency distribution, mode is the value of the variable which has the maximum frequency. Consider the following example:

Example 17. Find the mode of the following frequency distribution:

Variable (x)	2	5	7	9	11	25	35	43	52
Frequency (f)	1	3	4	8	25	12	11	10	8

Solution: Here we see that in the given distribution, the variable 1 has the maximum frequency 25. So the mode of this distribution is 11.

6.5.4 GROUPING METHOD OF MODE

When the distribution is irregular, the frequencies are increasing and decreasing in

An irregular pattern or the difference between the maximum frequency and the frequency succeeding or proceeding to it is small and the observations are concentrated on either side, in such a situation mode cannot be determined merely by inspection. In such a case, we apply the grouping method for the computation of mode. The procedure of grouping method will be clear from the following example.

Example 18. Find the mode of the following distribution.

Variable (x)	2	3	4	5	6	7	8	9	10	11	12	13
Frequency (f)	1	3	4	5	7	10	11	10	9	14	7	5

Solution. Here we see that initially the frequencies are increasing from 1 to 11 and then decreasing but the frequency 14 of the variable value 11 is again increasing and then decreasing up to frequency 5. This distribution shows an irregular pattern. So for the calculation of mode we apply the grouping method of mode. For this we prepare a table and the procedure of preparing the table is explained below the table.

Table12. Table for grouping the frequencies

Variable (x)	Frequency(f)					
	Column (i)	Column (ii)	Column (iii)	Column (iv)	Column (v)	Column (vi)
2	1	}4				
3	3	}9	}7	}8		
4	4	}17	}12		}12	
5	5	}21	}21	}22		}16
6	7	}23	}19		}28	

7	10	}12	}21	}30		}31
8	11				}33	
9	10			}26		}30
10	9					
11	14					
12	7					
13	5					

Prepare a table from the frequencies of the distribution. In column (i), we have the original frequencies. Mark bold type the maximum frequency in this column. Column (ii) is prepared by adding the frequencies two by two as 1+3 = 4; 4+5 = 9 and so on. Mark bold type the maximum frequency in this column also. Column (iii) is prepared by adding the frequencies two by two leaving the first frequency. Column (iv) is prepared by adding the frequencies three by three. Column (v) is prepared by adding the frequencies three by three leaving the first frequency and column (VI) is prepared by adding the frequency three by three leaving the first two frequencies. In each column make bold type the maximum frequency. The table is given above:

Now to find the mode we prepare the following analysis table:

Table13. ANALYSIS TABLE

Column number (1)	Maximum frequency (2)	Value(s) of x related to the maximum frequency (3)
i	14	11
ii	23	10, 11
iii	21, 21	7, 8, 11, 12
iv	30	8, 9, 10
v	33	9, 10, 11
vi	31	7, 8, 9

In the analysis table column number (1) shows the columns serially from the above table 12, column number (2) shows the maximum frequency from the same table 12 and column number (3) shows the value of x related to the maximum frequency or the values of x which contributes in the maximum frequency. Finally, in column number (3) of the analysis table we see that the value 11 is repeated maximum number of times. So 11 is the mode of the above distribution.

6.5.5 MODE IN CONTINUOUS FREQUENCY DISTRIBUTION

In case of grouped continuous frequency distribution the maximum frequency shows that the related class is the modal class and for the computation of mode we use the following formula:

$$M_o = L + \frac{(f_1 - f_0)}{(2f_1 - f_0 - f_2)} \times h$$

Where L= lower limit of the modal class

h= magnitude of the modal class

f_1 = frequency of the modal class

f_2 = frequency of the class succeeding the modal class

f_0 = frequency of the class preceding the modal class

For a moderately asymmetrical distribution the mode can be calculated by a formula given by Karl Pearson as follows:

Mode = 3 Median – 2 Mean

Example 19. Following table shows the blood pressure and the frequency related to it. Find the mode of this distribution.

Table 14.

C.I.	Frequency	C.I.	frequency
70-80	2	110-120	32
80-90	4	120-130	28

90-100	14	130-140	12
100-110	35	140-150	5

Solution. From the table it is clear that maximum frequency is 35 and the related class is the 100-110. So 100-110 is the modal class. Now to compute the mode we use the following formula:

$$M_o = L + \frac{(f_1 - f_0)}{(2f_1 - f_0 - f_2)} \times h$$

Here L= lower limit of the modal class= 100

h= magnitude of the modal class= 10

f_1 = frequency of the modal class= 35

f_2 = frequency of the class succeeding the modal class= 32

f_0 = frequency of the class preceding the modal class= 14

$$\therefore \qquad M_o = 100 + \frac{(35-14)}{(70-14-32)} \times 10 = 100 + \frac{210}{24} = 100 + 8.75 = 108.75$$

So mode of the given distribution is 108.75.

6.5.6 MERITS, DEMERITS AND USES OF MODE

Merits:

1. Mode is easy to understand and to calculate.

2. It is not affected by extreme values.

3. It can be determined graphically.

4. In some cases it can be located by inspection only.

5. It can be computed for the distributions of unequal class intervals provided the modal class; the class preceding the modal class and succeeding the modal class are of equal width.

6. It represents the most frequent value of the distribution, practically it is very useful.

Demerits:

1. It is not based upon all the observations.

2. It is not subjected to algebraic treatments, i.e., we cannot compute the combined mode if we have the modes of the two series.

3. In some cases mode is ill defined. In some cases it is not possible to find a clear mode. Some series have two modes and some more than two modes.

4. As compared to mean, mode is affected much by fluctuations of sampling; it is an unstable measure of central tendency.

5. If the modal class or the class preceding or succeeding the modal class are of unequal width, it cannot be determined.

6. There are different formulas for the calculation of mode.

Uses:

1. It is used to find the ideal size; it is very useful in business forecasting.

2. It is very useful in ready-made market, e.g., shoes, shirts, jeans etc.

3. It is very useful in commercial management.

6.6 SUMMARY

The study of this chapter provides us the knowledge of central tendency and measures of central tendency. From the study of this chapter we came to know the definitions of the measures of central tendency as mean, median and mode. We studied and learnt different methods of computing mean. We learnt about weighted mean and combined mean. We learnt how we can calculate the mean, median and mode in case of individual series, in case of discrete distribution and continuous distribution. We studied the grouping method of mode. We studied the merits, demerits and uses of mean, median and mode also. From the study of these methods, merits, demerits and uses we came to know the situations where which method is suitable and also which measure is suitable for the particular situation? Over all we learnt a lot about measures of central tendency.

6.7 GLOSSARY

Measures of central tendency: These are mean, median and mode and they provide us the most important knowledge about the distribution. All these give the study of central part of the distribution.

Inclusive class: A class in which its upper and lower both limits are included is called inclusive class.

Exclusive class: A class in which only one limit (generally lower) is included and other limit (upper) is not included in the class is called exclusive class.

Arithmetic Mean: Arithmetic means or simply mean is ratio of sum of values to the number of values.

Median: On arranging the values according to magnitude, the middle most value of the arranged data is the median.

Mode: The most frequent item in the series is called mode.

Multimodal Distribution: If a distribution has more than one mode, it is called multimodal distribution.

Ill-defined Mode: If in a distribution there exists more than one mode, mode is called ill- defined mode.

Cumulative frequency Table: A table got on adding the successive frequencies is called cumulative frequency table.

6.8 SELF ASSESMENT QUESTIONS

Question 1. Find the arithmetic mean of the following series:

2, 13, 4, 12, 13, 5, 16, 17, 11, 22, 32, 42, 44

Question2. Find the median and mode of the following series:

12, 32, 2, 3, 15, 16, 12, 13, 23, 43, 35, 12

Question 3. Find the mean of the following distribution:

x:	12	22	13	14	15	17	24	25
f:	2	1	3	10	8	5	3	2

Question4. Find the median for the given data.

I.Q.:	92	65	77	72	110	89	112	98
f:	3	5	4	10	6	2	1	5

Question5. Find the mean, median and mode for the data given in the following table:

I.Q.	Frequency	I.Q.	Frequency
90-100	11	130-140	43
100-110	27	140-150	28
110-120	36	150-160	16
120-130	38	160-170	1

Question6. Which of the following is not a measure of central tendency?

a. Mean b. variance c. median d. none of these

Question7. If a constant 10 is added in each observation of a set of data, the effect on mean is:

a. Mean is decreased by 10. b. Mean is increased by 10

c. Mean is multiplied by 10 d. Mean is not affected

Question8. The best measure of central value in case of qualitative data is:

a. Mean b. mode c. Median d. None of these

Question9. The measure which is affected much by extreme observations is:

a. Mean b. mode c. median d. Variance

Question10. If we add 3 in each observation, the effect on median is:

a. Median is decreased by 9 b. Median is increased by 3

c. median is not affected d. Median is multiplied by 3

Question11. The most frequent item in the series is called:

a. median b. mean c. mode d. none of these

Question12. The average of 3 numbers is 2 and the average another 4 numbers is 5. What is the combined average?

a. 7 b. 10 c. 3 d. 3.71

Question13. For a group of 100 students, the mean score in a test was found to be 40. Later on it it was found that a value 45 was misread as 54. The correct mean will be:

a. 40.50 b. 39.85 c. 39.80 d. 39.91

Question14. Arrangement of observations according to magnitude is carried out in:

a. mean b. mode c. median d. none of these

Question15. Grouping method is used for:

a. mean b. mode c. median d. all of these

Question16. Which measure of central tendency is most widely used?

a. mean b. mode c. median d. all of these

Question17. For asymmetrical distribution, the relation between the measures of central values is:

a. mean= mode- 2 median b. mode= 2 median- 3 mean

c. mode= 3 median- 2 mean d. mode = 3 mean- 2 median

Question18. Which measure is affected least by fluctuation of sampling is:

a. mode b. mean c. median d. none of these

Question19. The measure which cannot be calculated with open end classes is:

a. mean b. mode c. median d, all of theses

Answers: 1. 17.92, 2. 14, 12, 3. 16.24, 4. 77, 5. 126.4, 126.84, 122.86, 6. b, 7. b, 8. c, 9. a, 10. b, 11. c, 12. d, 13. d, 14. c, 15. b, 16. a, 17. c, 18. b, 19. a.

6.9 REFERENCES

Agarwal, B.L., Programmed Statistics, New Age International (P) Ltd, New Delhi, (2003).

Arora, P.N. and Malhan, P.K., Biostatistics, Himalaya Publishing House (P) Ltd. Mumbai, (1996).

Belle, G., Fisher, L., Heagerty, P.G. and Lumly, T.,Biostatistics- A Methodology for the Health Science, John Willey & Sons, inc. Pub., Hoboken, New Jersey, (2004).

Gupta, S.C. and Kapoor, V.K., Fundamentals of Mathematical Statistics, Sutan Chand & sons, Daryaganj, New Delhi, (1970).

Pagaro, M. and Gauvreau, K., Principles of Biostatistics, Duxbury Thomson Learning, U.S.A. (2000).

Prabhakara, G.N., Biostatistics, Jaypee Brothers Medical Publishers (P) Ltd, New Delhi, (2006).

Sudaram, K.R., Dwivedi, S.N., Sreenivas, V., Medical Statistics, Wolters Kluwer (P) Ltd, New Delhi. (2010).

6.10 TERMINAL QUESTIONS

Question1. Define arithmetic mean and write down its merits and demerits

Question2. Explain step deviation method of mean by taking an example.

Question3. Which measure is the best in qualitative data? Give its merits and demerits.

Question4. Which measure cannot be calculated with open end classes and why?

Question5. Which measure is best for a industry and why?

Unit-7 MEASURES OF VARIABILITY/ DISPERSION

CONTENTS

7.1 OBJECTIVES

From the study of this chapter the students will be able:

1. To know about the measures of variability such as range, interquartile range, mean deviation and standard deviation.
2. To know the advantages and disadvantages of these measures.
3. To know about the differences of these measurement.
4. To know the situations where which measure is better to use?
5. To know why the coefficient of variation is the best to comparing two or more series.

7.2 INTRODUCTION

In the previous chapter we study various measures of central tendency and the methods of measuring them. The measures of central tendencies give us only an idea of the central part of the distribution. Although these measures provide us important information about the population, i.e., they are necessary but not sufficient for explaining a distribution properly. These measures do not provide the information about the variability of the observations. Variability is an important factor in nature. There is slight variability in heights of a region of persons but there is more variation from the heights of the persons of other region. To understand a distribution more clearly we should study about variability.

Let us consider the example of distribution of marks 5 students of three sections of a class. The marks of section A are 5, 5, 5, 5, 5; the marks of section B students 3, 4, 5, 6, 7 ; and the marks of section C students are 0, 1, 5, 9, 10. We see that the total score of all the three section students is 25 and the average score of all the three section students is also same 5. But we see that there is no variability in section A marks; it is a constant series. There is less variability in marks of section B; but there is more variability in section C students marks. It is clear from the example that the distribution of marks of the three sections is quite different while the average is same.

The entire concept of statistics is based on variation. If there is no variation and all individuals are alike, there is no need for collecting any sample data because one individual can provide all the information of the complete data which we want to study on the basis of the complete data set. Bur variation in any phenomenon is inherent and statistical techniques help us to explain this variation. Thus we have to know dispersion. Dispersion means scatteredness and this tells us about the homogeneity and heterogeneity of the distribution. To know a distribution more clearly we study the variability or scatteredness.

7.3MEASURES OF VARIABILITY

Commonly used measures for variability are as follows:

1. Range
2. The interquartile range
3. Mean deviation
4. Standard deviation

7.3.1 RANGE

The simplest measurement of variability is range. It is defined as the difference between the extreme observations of the distribution, i.e., the difference between the largest and the smallest observation of the distribution. This uses the two extreme observations and does not provide the information on the middle values of the distribution. It is simple measurement of variability but its results are misleading if the two extreme observations are unusual.

There are four situations of occurrence in range value.

i. There may be no variability at all (see table A Below).

ii. There may be less variation between the extreme observations (see table B below) .

iii. There may be very much variation between the extreme observations (see table C below).

iv. None of the observation is representative of the mean (see table D below).

Table1. Situations in range value

A	B	C	D
7	6	2	49
7	6	5	0
7	8	6	0

7	7	7	0
7	8	8	0
7	6	9	0
7	8	12	0
Mean= 7	Mean= 7	Mean= 7	Mean= 7
Range= 5-5	Range= 6-8	Range= 2-12	Range= 0-49
Variation= 0	Variation= 2	Variation= 10	Variation= 49

7.3.1. Merits and Demerits of Range

Merits:

1. It is simple understand and to calculate
2. Unit of measurement is same as the unit of variable under study.
3. It does not require mathematical calculation.

Demerits:

1. It uses only two extreme observations and does not use all other information so there is no importance for collection of all other data observations. A good measure should use all the data observations for providing better estimate of variability.
2. It is not an ideal measure.
3. It cannot be calculated if the distribution contains open-end classes.
4. It is not amenable to further mathematical treatment.

7.3.2 INTERQUARTILE RANGE

It provides more knowledge about the distribution as it includes middle 50% observations of the distribution. It is defined as the difference between the first and third quartile of the distribution. The third or upper quartile (Q_3) divides the distribution in such a way that 75% observations lie below it and 25% observations of the distribution above it. It's just opposite first quartile or lower quartile (Q_1) divides the distribution in such a way that 25% observations of the distribution lie below it and 75% observations above it. Thus, interquartile range gives us information about 50% central observations of the distribution.

If interquartile range is high, it means that the middle 50% observations are a long distance and if it is low, it means that middle 50% observations are closely near to each other. Quartiles divide a series into four equal parts.

There are different methods for calculating interquartile range, depending upon the type of distribution. For calculating interquartile range we have to calculate the two quartiles.

7.3.2.1 Interquartile Range for Individual Series

Let the variable under study X takes the values $x_1, x_2, ...x_n$, then for the calculation of interquartile range we apply the following steps;

Step1. Arrange the given observations in ascending or descending order of their magnitude.

Step2. Now calculate first quartile (Q_1) and third quartile (Q_3) by the method:

Q_1 = Value of $\left(\frac{n+1}{4}\right)$th term in the arranged series.

Q_3 = Value of $3\left(\frac{n+1}{4}\right)$th term in the arranged series.

Step 3. The value of $Q_3 - Q_1$ is the interquartile range.

Example1. Find the interquartile range for the following series:

2, 12, 23, 4, 34, 25, 21, 17, 14

Solution. Arrange the given observations in ascending order of magnitude, we get,

2, 4, 12, 14, 17, 21, 23, 25, 34

Here n= 9

$\therefore$ Q_1 = Value of $\left(\frac{n+1}{4}\right)$th term in the arranged series.

= Value of $\left(\frac{9+1}{4}\right)th = 2.5$ th term in the arranged series.

= Value of 2nd term + ½ (3rd term - 2nd term)

= 4+ ½ (12 – 4) = 4+ 8/2 = 4+4 = 8

Q_3= Value of $3\left(\frac{n+1}{4}\right)$th term in the arranged series.

= Value of 3(2.5)th term = value of 7.5th term

= Value of 7th term + value of ½ (8th term – 7th term)

= 23+ ½ (25 – 23) = 23+ 1 = 24

Interquartile range = $Q_3 - Q_1$ = 24 – 8 =16

This shows that 50% observations of the given series lie between 8 and 24.

7.3.2.2 Interquartile Range for Discrete Distribution

If $x_1, x_2, ...x_n$ are the observations in a discrete distribution according to some characteristic and $f_1, f_2, ... f_n$ be their corresponding frequencies then for the calculation of quartiles, we calculate the cumulative frequencies. The interquartile range is calculated with the help of the following steps.

Working steps for interquartile range

Step1. Arrange the given values in ascending order of magnitude.

Step2. Find the total of frequencies, called cumulative frequency (denoted by c.f)

.Step3. Find $\frac{N}{4}$ for first quartile, where N=$\sum f$.

Step4. Find cumulative frequency just greater than $\frac{N}{4}$. The value of x, corresponding this cumulative frequency is the value of first quartile.

Step5. Find $3\left(\frac{N}{4}\right)$ for third quartile.

Step6. Find cumulative frequency just greater than$3\frac{N}{4}$. The value of x, corresponding to this cumulative frequency is the value of third quartile.

Step7. For interquartile range, find $Q_3 - Q_1$.

Example2. Find the interquartile range for the following data.

X	21	15	17	9	5	7	8	10
F	2	5	3	4	5	1	6	12

Solution. For calculating the interquartile range, first of all we calculate first and third quartile and for this, we arrange the values of x in ascending order and then prepare the cumulative frequency table as follows:

Table2. Quartiles in Discrete Distribution

x	F	c.f.
5	5	5
7	1	6
8	6	12
9	4	16
10	12	28
15	5	33
17	3	36
21	2	38

$$N = \sum f = 38$$

Here $\frac{N}{4} = \frac{38}{4} = 9.5$. The cumulative frequency just greater than 9.5 is 12 and the value of x, corresponding to cumulative frequency 12 is 8. So the first quartile is 8.

Here $3\frac{N}{4}=3\times\frac{38}{4}=28.5$. The cumulative frequency just greater than 28.5 is 33 and the value of x, corresponding to cumulative frequency 33 is 15. So the third quartile is 15.

Interquartile range = $Q_3 - Q_1$ = 15 – 8 =7.

This shows that 50% observations of the given series lie between 8 and 15.

7.3.2.3 Interquartile Range for Continuous Distribution

When the data is in class interval form, the class corresponding to c.f. just greater than $\frac{N}{4}$ is the first quartile class and the value of first quartile is given by:

$$Q_1 = L + \frac{\left(\frac{N}{4} - C\right)}{f} \times h$$

Where L= lower limit of the first quartile class

C= cumulative frequency just before the first quartile class

N= Total of frequency

f = frequency of the first quartile class

h = magnitude of the first quartile class

When the data is in class interval form, the class corresponding to c.f. just greater than $3\frac{N}{4}$ is third quartile class and the value of third quartile is given by:

$$Q_3 = L + \frac{\left(3\frac{N}{4} - C\right)}{f} \times h$$

Where L= lower limit of the third quartile class

C= cumulative frequency just before the third quartile class

N= Total of frequency

f = frequency of the third quartile class h = magnitude of the third quartile class

Example3. The following table gives the distribution of weights of 100 persons. Find the interquartile range for this data.

Weight	40-45	45-50	50-55	55-60	60-65	65-70	70-75	75-80	80-85	85-90
Frequency	1	3	6	10	15	25	15	10	11	4

Solution. For calculating the interquartile range, first of all we calculate first and third quartile and for this we prepare the cumulative frequency table as follows:

Table3. Quartiles in Continuous Distribution

Weight (in kg) (x)	Frequency (f)	Cumulative Frequency (c.f.)
40-45	1	1
45-50	3	4
50-55	6	10
55-60	10	20
60-65	15	35
65-70	25	60
70-75	15	75
75-80	10	85
80-85	11	96
85-90	4	100
Total	$N = \sum f = 100$	

Here $\frac{N}{4}=\frac{100}{4}=25$. The cumulative frequency just greater than 25 is 35 and the corresponding class is 60 – 65. So 60 – 65 class is the first quartile class and the first quartile is given by:

$$Q_1 = L + \frac{\left(\frac{N}{4} - C\right)}{f} \times h$$

Where L= lower limit of the first quartile class = 60

C= cumulative frequency just before the first quartile class = 20

N= Total of frequency = 100

f = frequency of the first quartile class = 15

h = magnitude of the first quartile class = 5

$$\therefore \quad Q_1 = L + \frac{\left(\frac{N}{4} - C\right)}{f} \times h = 60 + \frac{(25-20)}{15} \times 5 = 60 + 1.66 = 61.66$$

Here $3\frac{N}{4}=3\frac{100}{4}=75$. The cumulative frequency just greater than 75 is 85 and the corresponding class is 75 – 80. So 75 – 80 class is the third quartile class and the third quartile is given by:

$$Q_3 = L + \frac{\left(3\frac{N}{4} - C\right)}{f} \times h$$

Where L= lower limit of the third quartile class = 75

C= cumulative frequency just before the third quartile class = 75

N= Total of frequency =100

f = frequency of the third quartile class = 10

h = magnitude of the third quartile class = 5

$$\therefore \quad Q_3 = L + \frac{\left(3\frac{N}{4} - C\right)}{f} \times h = 75 + \frac{(75-75)}{10} \times 5 = 75$$

So the third quartile is 75.

Interquartile range = $Q_3 - Q_1$ = 75 – 61.66 =13.34.

This shows that 50% observations of the given series lie between 61.33 and 75.

7.3.2.4 Merits and Demerits of Interquartile Range

Merits:

1. It is better measure than range.
2. It provides information on 50% observations of the distribution.
3. It is not affected from extreme observations
4. It can be calculated in case of open- end classes.
5. Unit of measurement is same as the unit of variable under study.

Demerits:

1. It provides information only on middle 50% observations and does not give information on rest of the 50% observations. A good measure should use all the data observations for providing better estimate of variability.
2. It is not an ideal measure.
3. Its value is quite variable from sample to sample.
4. It is not amenable to further mathematical treatment.

7.3.3 MEAN DEVIATION

Arithmetic mean, the measure of central tendency plays an important role in the measure of variability also. A measure of variability in the data can be obtained as the average deviation of observations from the arithmetic mean. This is called mean deviation. A measure of variability can be obtained by taking the deviations of the observations from their mean, ignoring the signs. This deviation can be taken from any mean but usually arithmetic mean, median and mode is used for this. Mean deviation is least when deviation is taken from median.

Let us consider the marks of 5 students of a class. The marks are 6, 7, 9, 10, and 18. The arithmetic mean of the marks is 10. The deviations of marks from their arithmetic mean are – 4, – 3, –1, 0, 8. We see that the sum of deviations is zero. This is the property of arithmetic mean, i.e., the sum of deviations taken from mean is zero. So in calculating mean deviation come across two problems:

1. Which average should be taken for deviation?

The answer to this problem is that any of the averages – mean, median or mode can be used for deviation. But generally the arithmetic mean is used.

2. What should be the sign of the deviations?

The answer to this problem is that we should use the positive sign for the deviations, i.e., the absolute deviations should be taken. This is done because the sum of deviations taken from mean is zero. This is clear from the example of marks of 5 students taken in the above paragraph.

7.3.3.1 Mean Deviation for Individual Series

For an individual series if $x_1, x_2, ...x_n$ are the values of the variable x, then the mean deviation is given by the following formula:

$$\text{Mean Deviation (M.D.)} = \frac{1}{n}\sum |x_i - A|$$

When A is taken as arithmetic mean, it is called mean deviation about arithmetic mean. When A is taken as median, it is called mean deviation about median and when A is taken as mode, it is called mean deviation about mode.

Example4. Find the mean deviation about mean for the following series:

2, 4, 5, 7, 10, 13, 9, 15, 25

Solution. For finding the mean deviation about mean first of all we find the arithmetic mean of the given series.

$$\bar{x} = \frac{\sum x}{n} = \frac{90}{9} = 10$$

Now we form the following table:

Table4. Mean deviation in Discrete Series

X	$\lvert x - \bar{x} \rvert$

2	8
4	6
5	5
7	3
10	0
13	3
9	1
15	5
25	15
$\sum x = 90$	$\sum \lvert x-\bar{x}\rvert$ =46

Mean Deviation (M.D.) = $\frac{1}{n}\sum \lvert x-\bar{x}\rvert = \frac{1}{9}\times 46 = 5.11$

Hence the mean deviation of the given series is 5.11

7.3.3.2 Mean Deviation for Discrete Distribution

If $x_1, x_2, ...x_n$ are the observations in a discrete distribution according to some characteristic and $f_1, f_2, ...f_n$ be their corresponding frequencies then the mean deviation about mean is given by the following formula:

Mean Deviation (M.D.) = $\frac{1}{N}\sum f\lvert x-\bar{x}\rvert$, where N = $\sum f$ = Total frequency

Example5. Ten patients were examined for uric acid test. The operation was performed 1050 times and the frequencies so obtained for different number of patients (x) are shown in the table given below.

x:	0	1	2	3	4	5	6	7	8	9	10
f:	2	8	43	133	207	260	213	120	54	9	1

Compute the Mean deviation about arithmetic mean.

Solution. Let 5 be the assumed mean. Now we prepare the following table for the calculation of mean deviation about mean.

Table5. Mean Deviation for Discrete Distribution

X	Frequency (f)	d = x- 5	Fd	$\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert$
0	2	-5	-10	5.01	10.02
1	8	-4	-32	4.01	32.08
2	43	-3	-129	3.01	129.43
3	133	-2	-266	2.01	267.33
4	207	-1	--207	1.01	209.07
5	260	0	0	0.01	2.6
6	213	1	213	0.99	210.87
7	120	2	240	1.99	238.8
8	54	3	162	2.99	161.46
9	9	4	36	3.99	35.91
10	1	5	5	4.99	4.99
	$\sum f=1050$		$\sum fd=12$		$\sum f\lvert x-\bar{x}\rvert=1302.56$

Arithmetic mean $\bar{x}=A+\frac{\sum fd}{N}=5+\frac{12}{1050}=5+0.0114=5.01cm$

Now, Mean deviation (M.D.) = $\frac{1}{N}\sum f\lvert x-\bar{x}\rvert=\frac{1}{1050}\times 1302.56=1.24cm$

7.3.3.3 Mean Deviation for Continuous Distribution

In case of continuous distribution, there are given class intervals and their corresponding frequencies. First of all we find the mid values of these classes and treat them as the variable values. Then we calculate mean and then mean deviation about mean. The procedure will be clear from the following example.

Example 6. For the data given in the below table on systolic BP of 68 patients, calculate the mean deviation about arithmetic mean.

Systolic BP (mmHg)	Frequency (f)	Systolic BP (mmHg)	Frequency (f)
90-100	3	140-150	11
100-110	5	150-160	9
110-120	7	160-170	6
120-130	10	170-180	2
130-140	15		

Solution. For the calculation of mean deviation we prepare the following table:

Table6. For calculation of Mean in Continuous Distribution

Systolic BP (mmHg)	Frequency (f)	Mid Value (x)	Fx	$\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert$
90-100	3	95	285	40.6	121.8
100-110	5	105	525	30.6	153
110-120	7	115	805	20.6	144.2
120-130	10	125	1250	10.6	106
130-140	15	135	2025	0.6	9.0
140-150	11	145	1595	9.4	103.4
150-160	9	155	1395	19.4	174.6
160-170	6	165	990	29.4	176.4
170-180	2	175	350	39.4	78.8
Total	$\sum f = 68$		$\sum fx = 9220$		$\Sigma f\lvert x-\bar{x}\rvert = 866.6$

$$\bar{x} == \frac{\sum_{i=1}^{n} f_i x_i}{N} = \frac{9220}{68} = 135.6mmHg$$

Mean deviation about mean (M.D.) is given by

$$\frac{1}{N}\sum f\left|x-\bar{x}\right| = \frac{1}{68}\times 866.6 = 12.74mmHg$$

Hence the mean deviation about mean of the systolic BP is 12.74mmHg.

7.3.3.4 Merits, Demerits and uses of Mean Deviation

Merits:

1. Mean deviation is easy to understand and to calculate.
2. Mean deviation is least when taken from median.
3. Mean deviation is less affected from extreme values as compared to range and standard deviation.
4. Mean deviation is better measure of variability than range and interquartile range as it uses all the observations of the data.

Demerits:

1. It is rarely used in social sciences.

2. Its results are not accurate because it is least when taken from median but median itself is not a good measure of central value when the variations in the series are large.

3. It is not suitable to further algebraic treatment as the negative signs of the deviations are also taken as positive.

Uses:

It is used in studying the economic problems.

7.3.4 STANDARD DEVIATION

For describing the scatteredness of the data values the best measure of variability is the standard deviation.

It is denoted by σ. If standard deviation in a data is small, it means there is high degree of homogeneity in the data values and vice versa if the value of standard deviation is large, it means there is a large heterogeneity in the data values.

It is defined as the positive square root of the arithmetic mean of the deviations of values when the deviations are taken from their arithmetic mean.

7.3.4.1 Standard Deviation for Individual Series

Let the variable under study X takes the n values $x_1, x_2, ...x_n$, their standard deviation is given by the following formula:

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \quad \text{or} \quad \sigma = \sqrt{\frac{\sum d^2}{n}} \quad \text{where} \quad d = x_i - \bar{x}$$

The steps of the procedure are as follows:

Step1. Compute the arithmetic mean of the given series.

Step2. Compute the deviations of the series values from the mean, i.e., compute $d = x_i - \bar{x}$

Step3. Compute the square of the values got in step 2, i.e., compute $d^2 = (x_i - \bar{x})^2$.

Step4. Find the sum of values got in step 3 and divide it by the number of values, i.e., compute $\frac{\sum d^2}{n} = \frac{\sum (x_i - \bar{x})^2}{n}$.

Step5. Take the square of the value got in step 4. This is the required value of the standard deviation.

The procedure will be clear from the example given below:

Example7. Compute the standard deviation of the following series:

12, 15, 17, 21, 28, 27,

Solution. For the computation of standard deviation we prepare the following table:

Table7. S.D. for Individual Series

x	$d = (x_i - \bar{x})$	$d^2 = (x_i - \bar{x})^2$
12	-8	64

15	-5	25
17	-3	9
21	1	1
28	8	64
27	7	49
$\sum x = 120$		$\sum d^2 = \sum(x_i - \bar{x})^2 = 212$

Arithmetic mean $\bar{x} = \frac{\sum x}{n} = \frac{120}{6} = 20$

Standard deviation $\sigma = \sqrt{\frac{\sum d^2}{n}} = \sqrt{\frac{212}{6}} = \sqrt{35.33} = 5.94$

Hence the standard deviation of the given series is 5.94

7.3.4.2 Short-Cut Method of Standard Deviation

This method is applied when mean is in fractional form because in that case the deviations and their squares make the calculations difficult. So in this case we take the deviations of the values from an assumed mean.

Let d = x – A, here A is the assumed mean, then in this case the formula for standard deviation is given by as :

$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$ Where n is the number of observations.

We follow the following steps for the commutation of S.D. in this case:

Step1. Take any value of the series as assumed mean A.

Step2. Compute the deviations of the series values from the assumed mean, i.e.,

Compute $d = x_i - A$

Step3. Find the total of step 2 values, i.e., find total of d, i.e., Σ d.

Step4. Divide the value of step 3 by number of values 'n' and find its square, i.e,

$$\left(\frac{\sum d}{n}\right)^2.$$

Step5. Compute the square of the values got in step 2, i.e., compute $d^2 = (x_i - A)^2$.

Step5. Find the sum of values got in step 5 and divide it by the number of values, i.e.,

$$\text{Compute} \frac{\sum d^2}{n} = \frac{\sum (x_i - \bar{x})^2}{n}.$$

Step6. Subtract the value of step 4 from value of step 5 and then take its square root.

This is the required value of the standard deviation.

The procedure will be clear from the example given below:

Example8. Find the standard deviation in the above example 7 by short- cot method.

Solution. Let us take 21 as assumed mean A. Now we prepare the following table for the computation of standard deviation.

Table8. For Short –Cut Method of S.D.

x	$d = (x-21)$	d^2
12	-9	81
15	-6	36
17	-4	16
21	0	0
28	7	49
27	6	36
Total	$\sum d = -6$	$\sum d^2 = 218$

$$\therefore \qquad \sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} = \sqrt{\frac{218}{6} - 1} = \sqrt{35.33} = 5.94$$

7.3.4.3 Standard Deviation in Discrete Frequency Distribution

If $x_1, x_2, ...x_n$ are the observations in a discrete distribution according to some characteristic and $f_1, f_2, ...f_n$ be their corresponding frequencies then standard deviation can be calculate with the help of these methods:

1. Actual mean method
2. Assumed mean method
3. Step deviation method

The procedures of the three methods will be clear with the help of the examples.

1. Actual mean method:

For this we use the following formula:

$$\sigma = \sqrt{\frac{\sum f(x - \bar{x})^2}{N}}$$

Where $N = \sum f$ =Total frequency

Example9. Calculate the standard deviation of the distribution of marks of the B.Sc. botany class students. The data is given below:

x: 12, 18, 17, 15, 20, 25, 32, 42

f: 2, 1, 3, 5, 1 2 10 1

Solution. For the calculation of standard deviation we prepare the following table:

Table9. S.D. for Discrete Frequency Distribution by Actual Mean

X	f	f x	$d = (x - \bar{x})$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$

12	2	24	-12	144	288
18	1	18	-6	36	36
17	3	51	-7	49	147
15	5	75	-9	81	405
20	1	20	-4	16	16
25	2	50	1	1	2
32	10	320	8	64	640
42	1	42	18	324	324
	$N = 25$	$\sum fx = 600$			$\Sigma f\left(x-\bar{x}\right)^2 = 1858$

Arithmetic mean $\bar{x} = \dfrac{\sum fx}{\sum f} = \dfrac{600}{25} = 24$

Standard deviation $\sigma = \sqrt{\dfrac{\sum f\left(x-\bar{x}\right)^2}{N}} = \sqrt{\dfrac{1858}{25}} == \sqrt{74.32} = 8.62$

Hence the standard deviation of the given distribution is 8.62

2. Assumed Mean Method:

In this method we take a middle vale of x as the assumed mean A and the apply the following formula:

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$; where symbols are in their usual meaning.

For the procedure of this method we take a example below:

Example10. For the above example 9 apply assumed mean method for computing the standard deviation.

Solution. Let us assume A= 20. Now for the calculation of standard deviation we prepare the following table:

Table10. S.D. for Discrete Frequency Distribution by Assumed Mean

X	f	$d=(x-A)$	fd	fd^2
12	2	-8	-16	128
18	1	-2	-2	4
17	3	-3	-9	27
15	5	-5	-25	125
20	1	0	0	0
25	2	5	10	50
32	10	12	120	1440
42	1	22	22	484
	$N=25$		$\Sigma fd = 100$	$\Sigma fd^2 = 2258$

Standard Deviation $\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{2258}{25} - \left(\frac{100}{25}\right)^2}$

$$= \sqrt{90.32 - 16} = \sqrt{73.68} = 8.62$$

Hence the standard deviation of the given distribution is 8.62

3. Step Deviation Method:

This method is applied when the values have some common interval (say h), we divide the deviations by this common interval and apply the following formula:

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times h$$

The procedure of the method will be clear from the example given below:

Example11. Daily high blood pressure of a patient on 100 days is given below:

Bp (mmHg):	102	106	110	114	118	122	126
No. of days:	3	9	25	35	17	10	1

Calculate the standard deviation of the above data.

Solution. Let us take the assumed mean A= 114. Here common interval h= 4. Now we prepare the following table for the calculation of standard deviation.

Table11. S.D. for Discrete Frequency Distribution by Step Deviation

BP (mmHg)	f	$d = \frac{(x-114)}{4}$	fd	fd^2
102	3	-3	-9	27
106	9	-2	-18	36
110	25	-1	-25	25
114	35	0	0	0
118	17	1	17	17
122	10	2	20	40
126	1	3	3	9
Total	$N = 100$		$\Sigma fd = 100$	$\Sigma fd^2 = 2258$

$$\text{S.D.= } \sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times h = \sqrt{\frac{154}{100} - \left(\frac{-12}{100}\right)^2} \times 4$$

$= \sqrt{1.54 - .0144} \times 4 = 1.235 \times 4 = 4.94$ MmHg

7.3.4.4 Standard Deviation in Continuous Frequency Distribution

In case of continuous distribution we find the mid values of classes and treated them as the variable values x. In this case we can apply all the three methods discussed in previous section. But generally step deviation is applied. The formula is the same as in case of discrete distribution discussed. The procedure is described in the example given below.

Example12. Calculate the standard deviation for the following table giving the age distribution of 542 persons of a city.

Age in years:	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80	80 – 90
No. of members:	3	61	132	153	140	51	2

Solution. For the calculation of standard deviation, let us take $d = (x - 55)/10$. Here we let assumed mean A = 55 and common interval (h) = 10. Now we prepare the following table:

Table12. S.D. in Continuous Distribution by Step Deviation Method

Age group	Mid-value (x)	Frequency (f)	$d = \frac{(x-55)}{10}$	fd	fd^2
20-30	25	3	-3	-9	27
30-40	35	61	-2	-122	244
40-50	45	132	-1	-132	132
50-60	55	153	0	0	0
60-70	65	140	1	140	140
70-80	75	51	2	102	204
80-90	85	2	3	6	18
Total		N=542		$\sum fd = -15$	$\sum fd^2 = 765$

$$\text{S.D.} = \sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times h$$

$$= \sqrt{\frac{765}{542} - \left(\frac{-15}{542}\right)^2} \times 10 = \sqrt{1.334} \times 10 = 11.55$$

Hence the standard deviation of age of the given distribution is 11.55 years.

7.3.4. Merits, Demerits and Uses of Standard Deviation

Merits:

1. It is rigidly defined.
2. It uses all the observations of the data in calculation.
3. It is used in correlation.
4. It is affected least by fluctuation of sampling.
5. It is suitable for further mathematical treatments.
6. It is the best measure of variability.

Demerits:

1. Its calculation is difficult in comparison to other measures of dispersion.
2. It is sensitive to extreme values.
3. It is not easily understandable for a common person.

Uses:

1. It is best measure of comparison of variability.
2. It is used in partitioning between groups and within groups in analysis of variance and design of variance.
3. It is used with mean in normal distribution for finding the areas.
4. It shows best dispersion of values from the mean.
5. It is very much used in medical field.

7.4 VARIANCE AND COEFFICIENT OF VARIATION

7.4.1 VARIANCE

It is just the square of the standard deviation. It is denoted by σ^2 . In other words variance is the arithmetic mean of the squares of the deviations, when deviations are taken from their arithmetic mean.

7.4.2 COEFFICIENT OF VARIATION

It is the best measure of the comparison of variability of the two series or populations. The units of measurement of the two populations may be different. This comparison is possible because it is a unit free m easure. It is presented in percentage and

is expressed as:

Coefficient of variation (C.V.) = $\frac{\sigma}{\bar{x}}\times 100$; where notations have their usual meaning.

A series having lesser c.v. is called more consistent or more homogeneous, i.e., the values of the series are closer to the mean of the series and if the c.v. of a series is larger, it is called more variable or in other words more heterogeneous series, i.e., the values of the series far apart from the mean of the series.

Example13. Calculate the coefficient of variation of the distribution of marks of the B.Sc. botany class students. Given the following information:

Average marks	Standard deviation of marks
$\bar{x} = 24$	$\sigma = 6$

Solution. Coefficient of variation (C.V.) = $\frac{\sigma}{\bar{x}}\times 100 = \frac{6}{24}\times 100 = 25\%$

Hence the C.V. of the marks is 25%.

Example14. The following data shows the mean and standard deviation on systolic BP and weight of 10 persons as:

BP		Weight	
Mean	S.D.	Mean	S.D.
120	15	60	4.5

Compare the two characteristics.

Solution. For comparison of the two characteristics we find the C.V. of these characteristics.

$$\text{C.V for BP} = \frac{\sigma}{\bar{x}} \times 100 = \frac{15}{120} \times 100 = 12.5\%$$

$$\text{C.V for Weight} = \frac{\sigma}{\bar{x}} \times 100 = \frac{4.5}{60} \times 100 = 7.5\%$$

We see that the coefficient of variation of BP is more than the coefficient of variation of weight so BP is more variable than the weight of the given persons.

7.5 SUMMARY

The study of this chapter provides us the knowledge of dispersion or variability and measures of variability. From the study of this chapter we came to know the definitions of the measures of variability - range, interquartile range, mean deviation and standard deviation. We learnt the different method for calculating interquartile range, mean deviation and standard deviation for individual series, in case of discrete distribution and continuous distribution. We studied the merits, demerits and uses of these measures. From the study of these methods, merits, demerits and uses we came to know the situations where which method is suitable and also which measure is suitable for the particular situation? Coefficient of variation is the best measure for comparison on the basis of variability as it uses all the values and is a unit free measure.

7.6 GLOSSARY

Range: It is the difference between the largest and the smallest observation of the data.

Interquartile Range: It is based on the first and third quartile of the distribution and it express the limit for the middle 50% observations of the data.

Dispersion: It shows the spread of the data values around the central value of the distribution.

Mean Deviation: It is the average of the absolute deviations of the data values. Generally deviations are taken from mean.

Variance: It is the arithmetic mean of squares of the deviations of the values when deviations are taken from their mean.

Standard Deviation: It is the positive square root of the variance.

Coefficient of Variation: It is the best measure of comparison on the basis of variability of two or more series. It can also be used if the units of measurements are different because it is a unit free measure.

7.7 SELF ASSESMENT QUESTIONS

Question1. Find the range of the following series:

1, 23, 32, 24, 45, 42, 35, 37

Question2. Find interquartile range for the given series:

2, 4, 3, 5, 7, 10, 12, 13, 15, 14, 9, 1, 8, 6, 11, 16

Question3. Find the standard deviation of erythrocyte sedimentation rate (ESR) of the data 3, 4, 5, 4, 2, 4, 5, 3 found in 8 normal persons.

Question4. Calculate the mean and standard deviation of the following data on the length of fishes (in cm) .

Length:	10	20	30	40	50	60	70
No. of Fishes:	4	6	10	15	20	15	10

Question5. A firm has selected a random sample of 100 from its production line and has obtained the data shown in the table below:

Class interval	Frequency	Class interval	Frequency
130- 134	3	150-154	19
135- 139	12	155-159	12
140- 144	21	160-164	5
145-149	28		

Question6. Compute the standard deviation for data of hypo B.P. (in mmHg).

B.P. :	50	60	70	80	90	100	110	120

No. of Patients; 14 40 54 46 26 12 6 2

Question7. The C.V. of two series of observations are 20% and 25% with S.D. 5 and 8 respectively. What are their arithmetic means?

Question8. Which of the following is not a measure of dispersion?

a. Range b. Mean deviation c. median d. standard deviation

/Question9. The easiest measure of dispersion is:

a. Mean deviation b. S.D. c. interquartile range d. range

Question10. The range of the series 2, 3, 7, 9, 12, 17, 21 is:

a. 2 b. 21 c. 19 d. None of these

Question11. Which measure of dispersion is the best?

a. Range b. mean deviation c. C.V. d. S.D.

Question12. For comparison of two different series, the best measure of dispersion is:

a. Range b. S.D. c. mean deviation d. C.V.

Question13. Mean deviation is least when taken about:

a. Median b. mean c. zero d. mode

Question14. The mean of a series is 10 and C.V. 40%. Its S.D is:

a. 4 b. 16 c. 40 none of these

Question15. If all the values of a series are 4, the S.D. of this series is:

a. Zero b. 16 c. 4 d. 2

Answers: 1. 44, 2. 8, 3. 0.97, 4. 14.79, 5. 7.21, 6. 15mmHg, 7. 25, 32, 8. c, 9. d. 10. c, 11. d, 12. d, 13. a, 14. a, 15. a.

7.8 REFERENCES

Agarwal, B.L., Programmed Statistics, New Age International (P) Ltd, New Delhi, (2003).

Arora, P.N. and Malhan, P.K., Biostatistics, Himalaya Publishing House (P) Ltd. Mumbai, (1996).

Belle, G., Fisher, L., Heagerty, P.G. and Lumly, T.,Biostatistics- A Methodology for the Health Science, John Willey & Sons, inc. Pub., Hoboken, New Jersey, (2004).

Gupta, S.C. and Kapoor, V.K., Fundamentals of Mathematical Statistics, Sutan Chand & sons, Daryaganj, New Delhi, (1970).

Pagaro, M. and Gauvreau, K., Principles of Biostatistics, Duxbury Thomson Learning, U.S.A. (2000).

Prabhakara, G.N., Biostatistics, Jaypee Brothers Medical Publishers (P) Ltd, New Delhi, (2006).

Sudaram, K.R., Dwivedi, S.N., Sreenivas, V., Medical Statistics, Wolters Kluwer (P) Ltd, New Delhi. (2010).

7.9 TERMINAL QUESTIONS

Question 1. Define mean deviation and write down its merits and demerits

Question 2. Explain standard deviation and write down its merits and demerits.

Question 3. Which measure of dispersion is the best and why?

Question 4. Which measure is used if measurement units are different?

UNIT8: PRINCIPLES AND USES OF ANALYTICAL INSTRUMENTS

CONTENT

8.1- OBJECTIVES

In this chapter you will learn about the basic or common instrument used in the laboratory such as pH me UV-visible spectrophotometer, Centrifuges (clinical, High-speed and ultra- centrifuge), Geiger Muller and scintillation counters

8.2. INTRODUCTION

In biological science there is basic requirement of having knowledge of instrument. Instrument play important role in laboratories such as soil analysis lab, diagnostic lab, molecular biology lab, microbiology lab, chemistry lab, physics lab and diagnostic lab. We cannot imagine the laboratories without basic instrument such as pH meter, centrifuge and spectrophotometer. These instruments are very necessary for analysis. Instrument not only speedup the analysis but also have accuracy, specificity and sensitivity. pH meter is one of the important instruments used in basic laboratory to measure pH of water, soil and food etc. Centrifuge is used to separate the mixture of two liquid depending on the density of liquid. UV spectrophotometer is generally used in analytical chemistry for the quantitative determination of different analyses, such as transition metal ions, highly conjugated organic compounds, and biological macromolecules. G-M counter and Scintillation counter is used for the detection of radioactivity.

8.3. PRINCIPLES AND USES OF ANALYTICAL INSTRUMENTS

8.3.1 pH METER

pH meter is generally used to determine the pH of soil, water and culture medium used for the cultivation of fungi and bacteria. There is presence of electrode which very sensitive to detect the change in H ion concentration. The electric circuit measure the electromotive force developed across the electrode pair. pH is abbreviation of “Pondus Hydrogenii”and it was proposed by Sorenson in 1909 in order to express the small concentration of hydrogen ions. pH is a unit of measure that describe acidity and alkalinity of solution. It is measured on a scale 0 to 14 and defined as the negative logarithm of hydrogen ion activity.

$pH= -\log [H^+]$

The pH value of a substance is directly related to the ratio of the hydrogen ion and hydroxyl ion concentrations. If the H+ concentration is higher than OH- the material is acidic. If the OH^- concentration is higher than H^+ the material is basic.

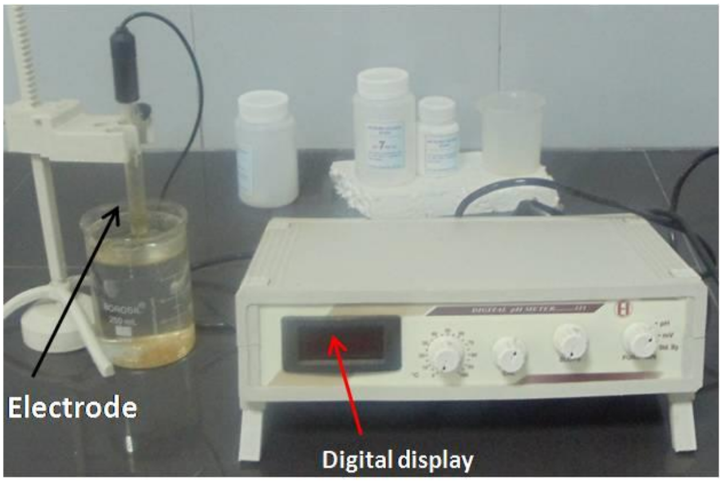


Figure 4.1. Digital pH meter

Principle of pH meter

The pH electrode consist of the pH-sensitive electrode which is a thin glass membrane whose outside surface contacts the solution to be tested. The inside surface of the glass membrane is exposed to a constant concentration of hydrogen ions (0.1 M HCl) (Fig.8).

Inside the glass electrode assembly, a silver wire, coated with silver chloride and immersed in the HCl solution known as Ag/AgCl electrode. This electrode carries current through the half-cell reaction. The potential between the electrode and the solution depends on the chloride ion concentration, but, since this is constant (0.1 M), the electrode potential is also constant. To complete electrical circuit reference electrode is needed. So, Ag/AgCl electrode is immersed in an 0.1 M KCl solution which makes contact with the sample through a porous fiber which allows a small flow of ions back and forth to conduct the current. The potential created at this junction between the KCl solution and the test solution is nearly zero and nearly unaffected by anything in the solution, including hydrogen ions. A voltmeter in the probe measures the difference between the voltages of the two electrodes. The meter then translates the voltage difference into pH and displays it on the screen.

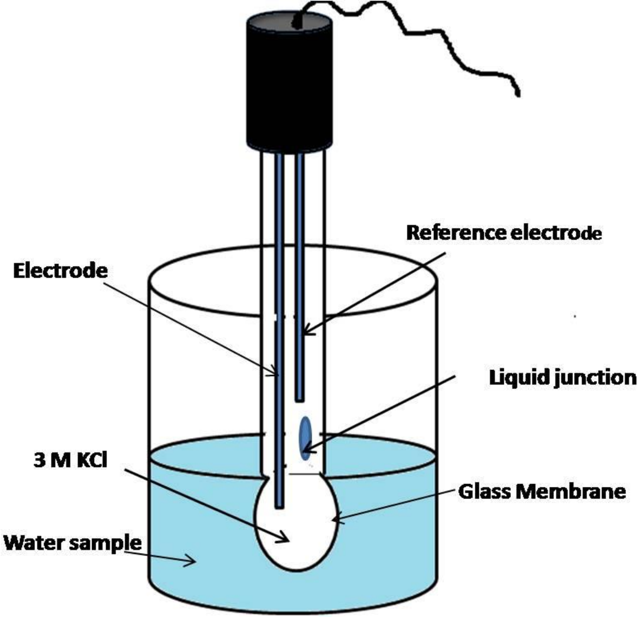


Figure 4.2: Working of pH meter

Calibration of pH meter

1. Switch on the pH meter and move the knob from stand by to pH.
2. Rinse the electrode with double distilled water and wipe out the tip gently with tissue paper,
3. Place the electrode into the solution of pH 7.0 and adjust the pH by adjustment given in pH meter.
4. After adjusting again wash the tip with distilled water and wipe out with tissue paper.
5. Dip the electrode in to solution of pH 4.0 (if your solution is acidic) and pH 9.0 (when your solution is basic)
6. Turn on the knob to slope position till it shows 4.0 or 9.0.
7. After reading the pH of sample. Again wash the electrode with distilled water.
8. Immerse the electrode in to the distilled till used for the next time.

8.3.2- UV-visible spectrophotometer

Spectroscopy is the measurement of (spectrum of light) electromagnetic radiation, absorbed, scattered, or emitted by atoms, molecules, or other chemical species. Each chemical species has unique energy states; spectroscopy can be used to identify the interacting species separately. Moreover; interaction of light with different compounds offers several possibilities of qualitative and quantitative measurements. Electrons in an atom move around the nucleus in orbitals and possess characteristic energy also known as *ground state*.

Energy transfer to these electrons energizes them substantially to let them leave their orbital to jump over the next level of energy orbital or level; this is called as *excited state*. Generally, on ceasing the energy supply, electrons emit the absorbed energy in the form of radiation giving rise to the *atomic spectrum* or simply *"line spectrum"*, which is represented as graph of the amount of energy absorbed or emitted by a system against wavelength or similar electromagnetic parameters. Visible light forms a part of the electromagnetic spectrum with γ rays at one end having wavelength of the order of 10^{-14} m, and radio waves at the other end having wavelength 3 × 10^{3} m or greater. As atoms, each of the electrons in a molecule usually occupies the available lowest energy level (ground state). Electrons in a molecule change their energy level only after the absorption or emission of the distinct quanta (particular energy radiation) of radiation by the molecule. Depending upon the absorption and emission of energy quanta, there occur two viz., absorption and emission spectra respectively. For molecules, each ground and excited state is subdivided into a number of vibrational and rotational energy sublevels, molecular spectra is therefore seen as "*band spectra*".

Types of spectra

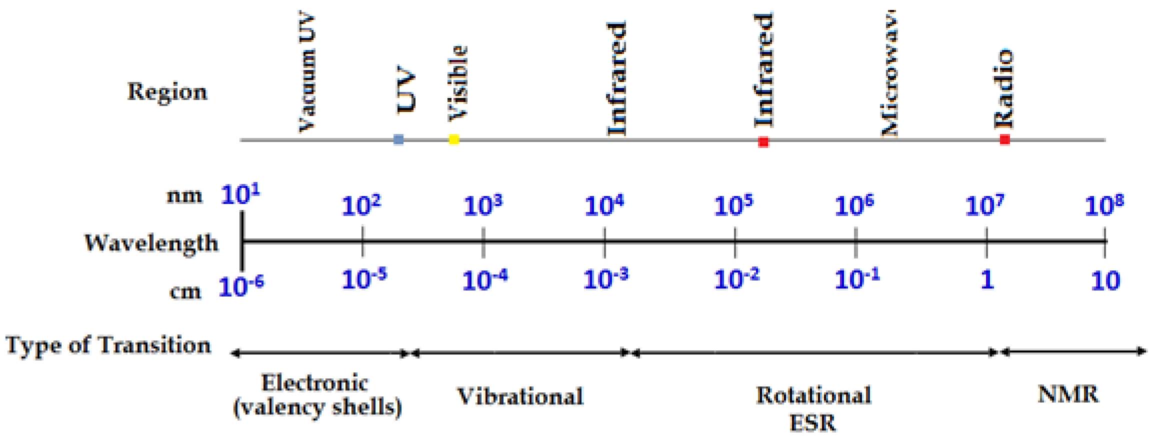


Figure 4.3. Types of spectra and their wavelengths

Electronic Spectra

Electronic spectra arise due to the outer electrons of atoms changing between major electronic energy levels. Such spectra occur in the visible and ultraviolet regions and are usually accompanied by changes in the rotational and vibrational energy levels. These spectra are used routinely in biochemistry. Fluorescence spectra may also arise owing to these transitions.

Vibration – Rotation Spectra

Vibration spectra caused by changes in the vibration energy levels. They occur in the near infra-red region and may be accompanied by changes in the rotational energy levels. Such spectra are sometimes used in studies of the detailed structure of biological macromolecules in non-aqueous environments.

Electron Spin Resonance (ESR) Spectra and Nuclear Magnetic Spectra

These spectra arise due to changes in the directions of the spins of electrons and nuclei respectively in a magnetic field. These two types of spectra are valuable for studying the structure of biological macromolecules.

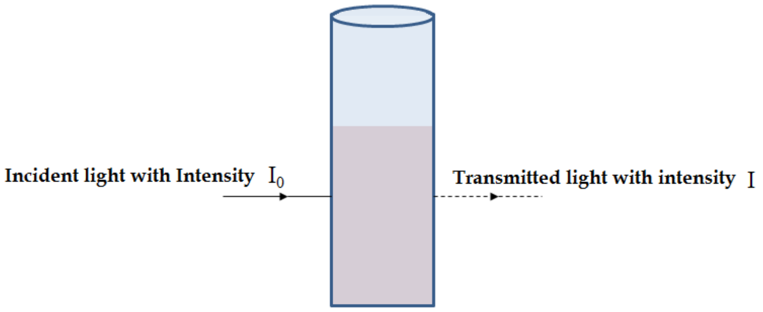
Molecular Band Spectra

Molecular band spectra may be resolved into a number of very close line spectra, corresponding to the vibrational and rotational energies of the electrons only at extremely high resolution.

3.0.1 Absorption

Photons of UV and Visible light may sometimes impart their energy to materials by interaction with individual atoms or molecules. Energy is imparted to the atoms or molecules, causing the excitation of valence electrons. Molecules with excited electronic states represent an unstable state will relax by allowing their electrons to fall to the ground state as soon as possible (10^{-16} sec.).

How much radiation is absorbed?


Incident light with Intensity I_0
Transmitted light with intensity I

The absorbance (A) is defined as the $\log_{10}$ of the ratio of the incident to transmitted light intensities:

$A = log_{10}\left(\frac{I_0}{I}\right)$ - The intensity is measured in the unit of power; therefore the absorbance is a unit less quantity. The intensity of light can be detected by photo-multiplier tubes. The collision of a photon of appropriate energy with the suitable molecule results in absorption of light.

Beer-Lambert Law or Law of absorption

The Beer-Lambert law relates the absorption of most molecular species to the concentration (C_n), the path length (l) and the molar absorptivity (ε).

$A = \varepsilon C_n l$

The wavelength of maximum absorption is known as λ max and is usually used as the wavelength for molar absorptivity (ε). ε is also sometimes known as extinction coefficient and is measured in $dm^3mol^{-1}cm^{-1}$. The Beer-Lambert law is sometimes expressed in terms of transmittance T as:

T = 1/A

Pathlength of 1 cm is normally chosen to simplify the calculation of the absorbance or molar absorptivity. All modern spectrophotometers are designed to comply the Beer-Lambert law. A plot of absorbance versus wavelength is known as UV-Visible Spectrum and is measured by a UV-Visible Spectrophotometer.

3.0.2. Spectrometers /Spectrophotometer

A spectrometer or spectrophotometer is a device which measures the absorbance by molecules in a given sample. Spectrometers are a monochromator equipped with a photo-transducer. Single and double beam spectrophotometers are readily available in the market whereas double beam spectrophotometers are commonly used in scientific laboratories because of the ease of function and precision in readings of absorbance.

A

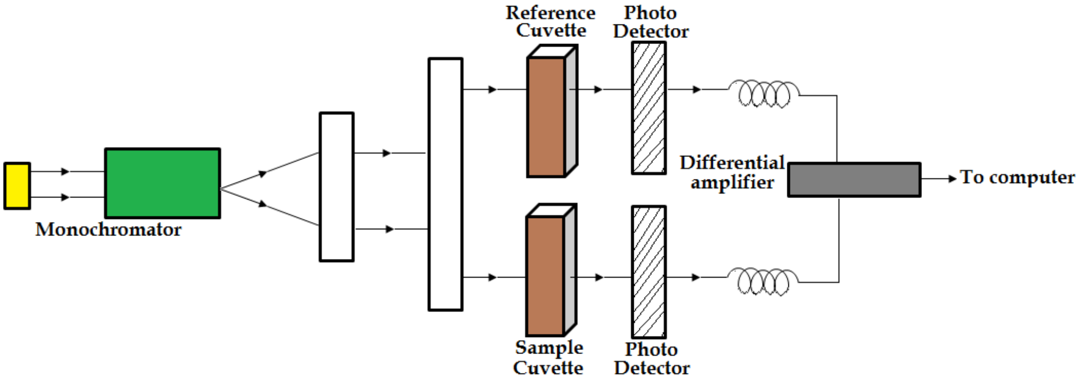


Figure 4.4. Schematic of Double Beam Spectrophotometer

spectrophotometer comprised of several devices (figure 3.1) such as monochromatic for monochromatic light of particular wavelength, grits, photo-detectors, and a processor that send signals to the interfaced computer. To understand and to define a spectrophotometer, it is recommended to study each and every component in details.

Light Source

The irradiation of a sample for UV and Visible spectroscopy requires a light source with constant output intensity. Tungsten (W) filament lamp are used for UV and visible light (320 – 2500) nm while hydrogen and deuterium lamps UV radiation. The light source should also be sufficiently intense so as to allow sufficient transmitted radiation to be detected when the absorption falls within a range of 0 -2. In more advanced spectrophotometers, tungsten-halogen lamps are frequently used. Such lamps contain small quantity of iodine and the lamp is enclosed in *quartz* housing. The iodine (a halogen) raises the temperature of lamp to 3500K, which permits the intensity of the output radiation down to ~ 190 nm (UV range). Quartz is permeable for UV radiation whereas glass blocks the UV radiation.

Hydrogen and deuterium lamps produce high intensity UV radiation. Electrical excitation of hydrogen atoms at low pressure may produce two hydrogen of either low energy with the high energy photon or higher energy with low energy photons. As a result of this unequal uniform output the deuterium or hydrogen lamps give rise to the outputs of radiation over a wavelength range of 160 – 375 nm.

Monochromator

A monochromator is the most commonly used device to select a wavelength of light for the irradiation of a sample. Monochromators use a series of lenses, mirrors, slit and windows together with either prisms and /or diffraction gratings to isolate a narrow band of wavelengths. It is impossible to select one wavelength range. The wavelength of the radiation source will tend to follow a *Gaussian* distribution around a mean value of wavelength known as the nominal wavelength. The effective *bandwidth* is defined as the wavelength range which corresponds to the half peak height width of

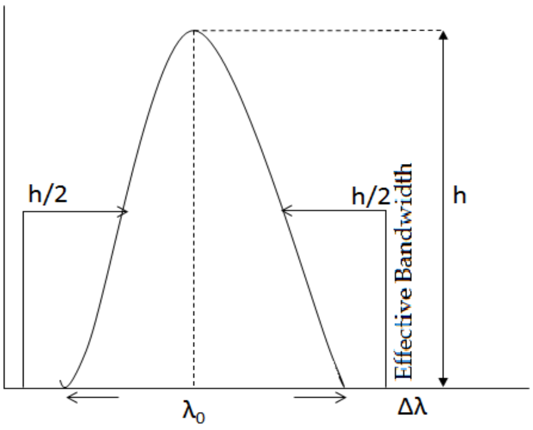


Figure 4.5 Effective Bandwidth

The wavelength distribution profile. There are two common types of monochromators viz., Monochromators based on refracting prisms and Monochromators based on diffraction grating are used in the modern age spectrophotometers. A monochromator with a bandwidth of less than ~ 0.5 nm is perfectly acceptable.

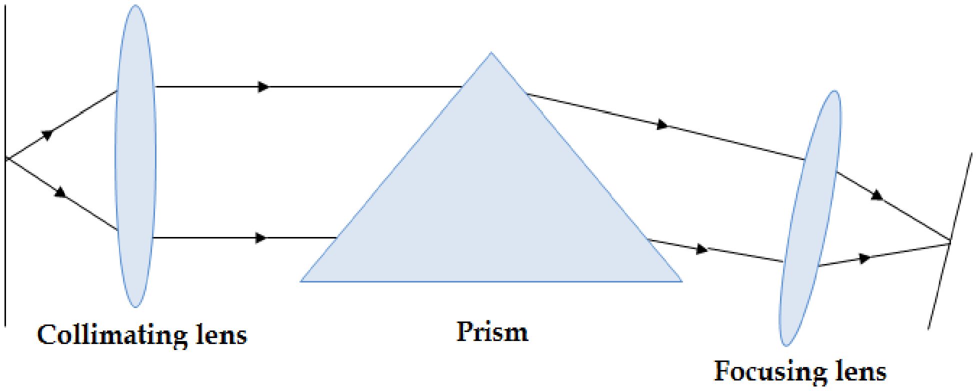


Figure 4.6 Simplified Schematic of a Prism Monochromator

In refracting prism monochromators, the white light enters via a slit passing through a collimator into the monochromator. A collimator is equipped with collimating lens, produce a parallel beam of radiation as shown in Figure 6. The light then passes through a refraction prism that disperses the light into its component wavelengths. The light is focused by another lens on the focal point where another slit is placed. By rotating prisms, radiation of different frequency is selected.

Diffraction grating based monochromator

White light passes through an entrance slit and is focused on a diffraction grating through a concave mirror. The diffraction grating disperses the light into its component wavelengths and reflects the light onto a second concave mirror. The grating can be rotated by a stepper motor. Light is then focused on the exit slit by this second concave mirror. By rotating the grit, light of different wavelengths can be selected.

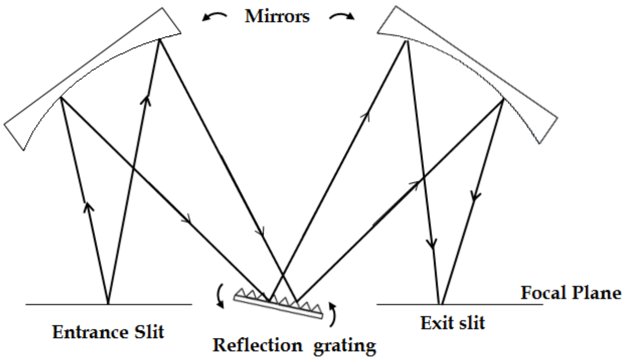


Figure4. 7. Optics of Diffraction Based Monochromator

Light Detector

The absorbance of an analyte is performed by a photon tube or light detector which measures the intensity of transmitted light. A photo-multiplier tube or photon tube works on the principle of photo-emission or electronic transition involving the counting of photons enabling the monitoring of the intensity of light. Figure 3.5 show a photon tube where by the photons enter into the tube to emit electrons from the large cathode that are captured by a wire anode placed in front of the cathode. The entire circuit is encased in quartz housing under high vacuum. The circuit gets completed and the current generated is directly

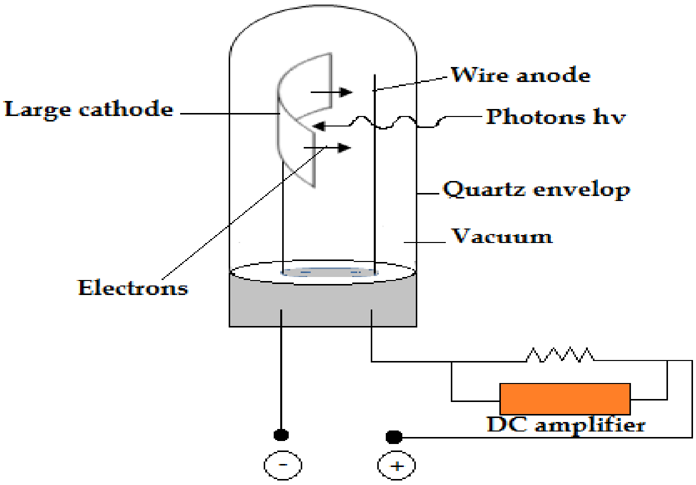


Figure4. 8. A photo-multiplier tube

Proportional to the intensity of light entering into the photon tube. Photon tube operates as photo-multiplier tube whereby a cascade effect is produced, which enables the collection of $10^6 – 10^7$ electrons per photon entering into the tube.

8.4 CENTRIFUGE

Centrifugation is based on the principle of centrifugal force in which liquid are subjected to high speed to separate solid from liquid or liquid from liquid depending on the density. In centrifugation heavy particle settled down and light particle will rises to the top. The substance which is settled down is called “Pellet” and the remaining fluid or overlying fluid is called “Supernatant”. Centrifugation process separates two substance of different density.

Basic principle of centrifugation

When the tube is filled with fluid and allowed to spin, as the rotor spin the apparent centrifugal force act on the sample of fluid and the analyte inside the fluid is pushed both radially outwards to the side of centrifuge tube.

Relative centrifugal force $F = m\ \omega^2 r$

Where,

F= Relative centrifugal force

m= Mass of particle

r= radius of the circular motion of the centrifuge (unit meter).

ω= angular velocity of the centrifuge (unit radian per sec)

The rate of sedimentation depends upon the applied centrifugal field (G), which depend upon the radical distance of the particle from the rotation axis and square of the angular velocity.

$G = \omega^2 r$

Where angular velocity (ω) = $\frac{2\pi}{60}$ revolution/min (one revolution is equal to 2 radian)

Putting the value in G , therefore

$$G = \frac{2\pi\,(\text{revolution/min})^2 \times r}{(60)^2}$$

$$G = \frac{4\pi r}{3600}$$

The relative centrifugal force (RCF) in g units (g=9.8065 M/sec^2 or 980 cm /sec ^{2})

$$\text{RCF} = \frac{4\pi r}{3600 \, x \, 980} = 1.11 \times 10^{-5}\ (\text{revolution min}^{-1})^2\ r$$

The rate of settling of particle depends upon (a) the relative centrifugal force (b) the density of the specimen, (c) the density of the fluid, (d) frictional forces, and (e) the size and shape of the specimen Consider if the suspending particle is spherical, then the rate of settling of spherical particle can be expressed by following equation.

$$R = \frac{2r^2\, g\, (dp-df)}{9n}$$

Where,

R= Rate of settling
r= Radius of the particle
g= acceleration due to gravity
dp= Density of the particle
df=Density of the fluid
n=viscosity of the fluid

The rate of sedimentation depend on the applied centrifugal field i.e. RCF in g units (980 cm/sec^2).
Types of rotor
The centrifuge is equipped with two types of rotor

1. Fixed angle rotor: In this type of rotor, the centrifuge tube containing samples are placed in the shield in the rotor at fixed pre-set angle. The rotor holds the centrifuge at fixed inclination i.e. 35

degree to the vertical. The solute during centrifugation forced against the side of the tube resulting in faster separation of solute from the suspension. The disadvantage of this rotor is that there are chances of abrasion due to striking the particle to the wall of centrifuge tube. Another disadvantage of this rotor is there is formation of smear like sedimentation than clear pellet formation. This type of rotor has short run time (Fig 9).

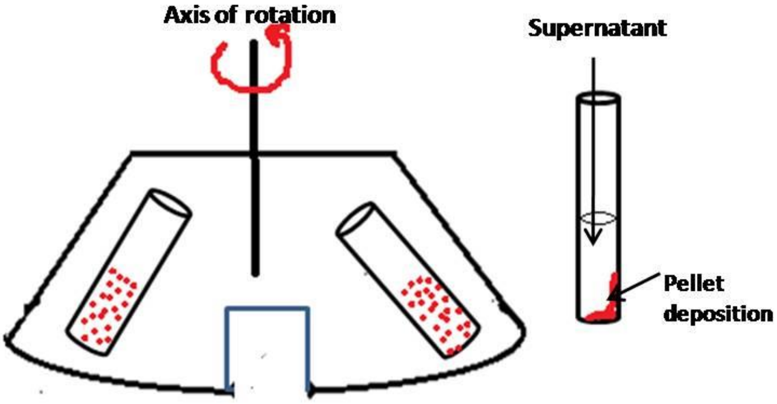


Figure 4.9: Fixed angle rotor and formation of pellet on side wall

2. Swinging bucket rotor: In this type of rotor the sample centrifuge tube are placed vertically and when the machine is started the bottom of the tube swing outward (horizontally) as the shaft rotates. This rotor has advantage than fixed rotor having clear pellet at the bottom of the tube (Fig.10).

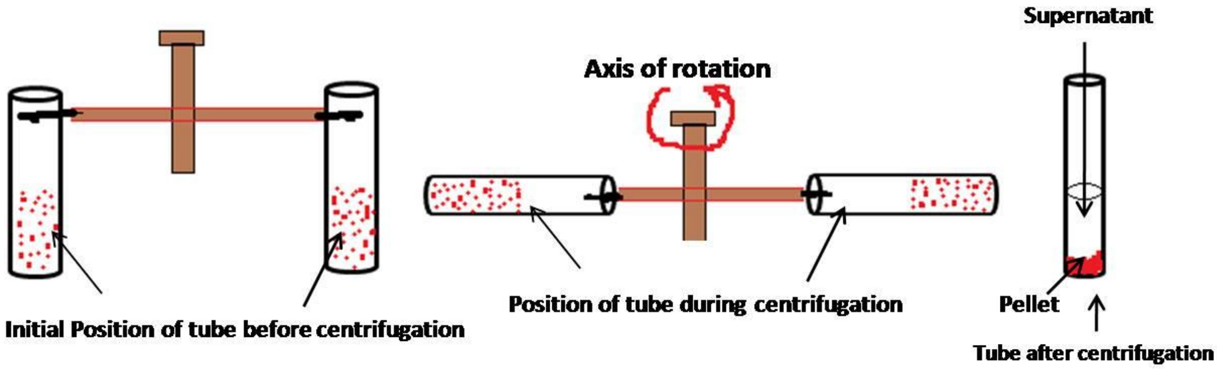

Supernatant
Axis of rotation
Initial Position of tube before centrifugation
Position of tube during centrifugation
Pellet
Tube after centrifugation

Figure 4.10: Working of swing bucket rotor before and after centrifugation

3. Vertical Rotor: In this type of rotor angle of placement of rotor is fixed but not at the slanting position it is vertical i.e. perpendicular to centrifugal field and the rotation axis as shown in Fig. 4.11. Band separate across the diameter of the tube rather than down the length of tube.

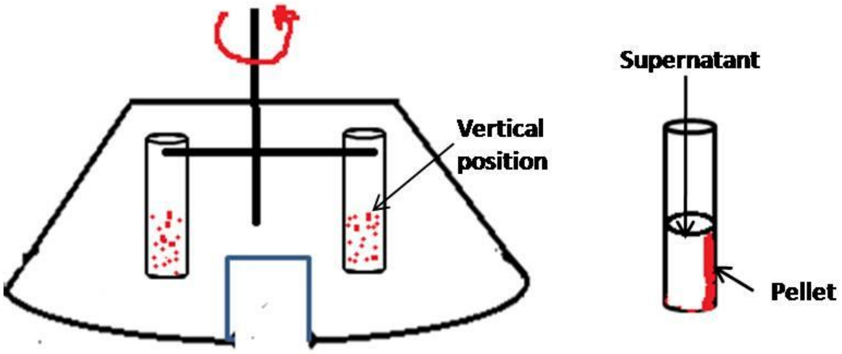


Figure 4.11: Working of vertical rotor before and after centrifugation

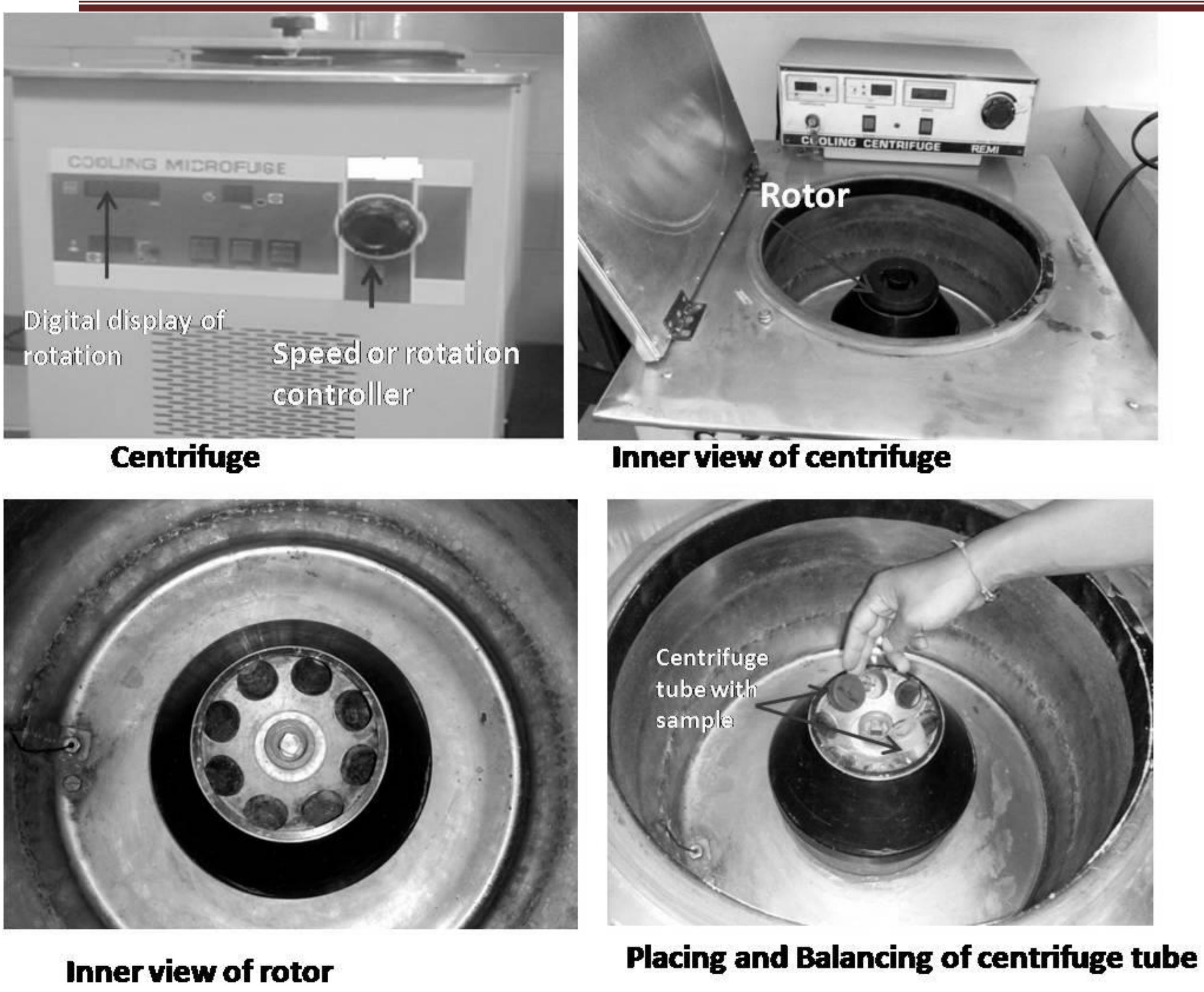
Types of centrifuge

Bench centrifuge: The most common type of centrifuge which generally used for common purposes such as separation of serum, plasma from blood sample required for serological reaction. It is not used for the fine substance such as organelle etc. The maximum speed of this type of centrifuge is 4000-5000 rpm and they operate is at ambient temperature. Now a day's small microfuge are available which can easily put into refrigerator to keep temperature cool for sedimentation to prevent denaturation heat sensitive substance such as Protein. .

Refrigerated centrifuge (Large capacity):

Refrigerated centrifuge have inbuilt cooling system for rotor such as compressor. It has speed of 6000 rpm with relative centrifugal field 6500g. Refrigerated centrifuge is used for the heat sensitive substance. This type of centrifuge can be run with fixed type rotor or swinging bucket type rotor. The balancing and placing of sample is very important in centrifuge. Always keep balance by placing centrifuge tube opposite to each other so that the load is distributed equally around the axis of rotor as shown in Fig. The sample tubes or centrifuge tube of 10, 50 and 100 ml can be used for centrifugation. The machine comes with changeable rotor according to the requirement of researcher (Fig. 4.12).

High Speed refrigerated Centrifuge: As name indicates this type of centrifuge work on high speed i.e. 25000 rpm generating RCF of about 60000 g. It is also having flexibility of using both types of rotor i.e. fixed rotor or swinging bucket type rotor. It is used for separating bacteria, cell organelles and precipitated proteins. It is not used for the sedimentation of viruses, and small organelles such as ribosome, for this purpose ultracentrifuge is required having maximum speed.


COOLING MICROFUGE
Digital display of rotation
Speed or rotation controller
COOLING CENTRIFUGE
REMI
Rotor
Centrifuge tube with sample

Centrifuge

Inner view of centrifuge

Inner view of rotor

Placing and Balancing of centrifuge tube

Figure 4.12: Working and parts of Refrigerated centrifuge

Ultracentrifuge: Centrifuge which is used to sediment viruses and small organelles like ribosome have speed of 80000 rpm with a relative centrifugal field of upto 600000 g. it was developed by Svedberg in 1929 and demonstrated the subunit of proteins. Ultracentrifuges are of two types:

a. Preparatory ultracentrifuge:. Preparatory centrifuge is generally used for the centrifugation or separation of cellular organelles such as mitochondria, ribosome, microsome and viruses. It is also used for the gradient separation of solution containing increasing concentration of dense substance. For e.g. sucrose solution is used for the separation of cell organelles. Caesium salt is used for the separation of nucleic acid. In this centrifugation the sample is run at high speed and then rotor is allowed to come to smooth stop and gradient is taken out to isolate the separated component.

a. Analytical centrifuge (AUC): A typical analytical centrifuge generate centrifugal field of 2500000 g. Analytical centrifuge is used for the determination of the purity of macromolecule, determination of relative molecular mass of solute in their native state, for the examination changes in molecular mass of super molecular complex. It consists of optical detecting system to monitor the material during sedimentation for concentration distribution in the sample at any time during centrifugation using ultra violet light absorption and optical refractive index sensitive system. Measurement of sample concentration at wavelengths from 200 to 800 nm detection of macromolecules containing strong chromophores. It is used to analyse protein, polysaccharide, nucleic acid, drug, ligands, gases, organelles and viruses.

When the machine switched on or rotor moves the image of analyte (for e.g cells or protein) are projected by an optical system on to film or computer. The concentration of the solution at various points in the cell is determined by absorption of light of the appropriate wavelength and them measuring the degree of blackening of photographic film. This will facilitate to observe the separation of sample concentration versus the axis of rotation due to applied centrifugal force. AUC are generally used for the two types of experiment

a. Sedimentation equilibrium experiment has aim to determine the total time course of sedimentation and report on molar mass and size distribution of dissolved macromolecule

b. Sedimentation velocity experiment is dealing with the final steady state of the experiment where sedimentation is balanced by diffusion opposing the concentration gradient

8.4.1- GEIGER MULLER COUNTERS (GM COUNTER)

The GM counter was named after its inventor Hans Geiger and Walther Muller. This instrument was first developed by Geiger in 1908 with Ernest Rutherford and name Geiger counter. At that time this device has limited used i.e. it can detect only alpha particle. Later on the research scholar of Geiger, Walther Muller improved this device for detecting more types of ionization radiation. Now a day this instrument is known as Geiger- Muller Counter which is used as particle detector that measure ionization radiation i.e. beta particle and alpha particles except gamma particle because it does not ionize the gas.

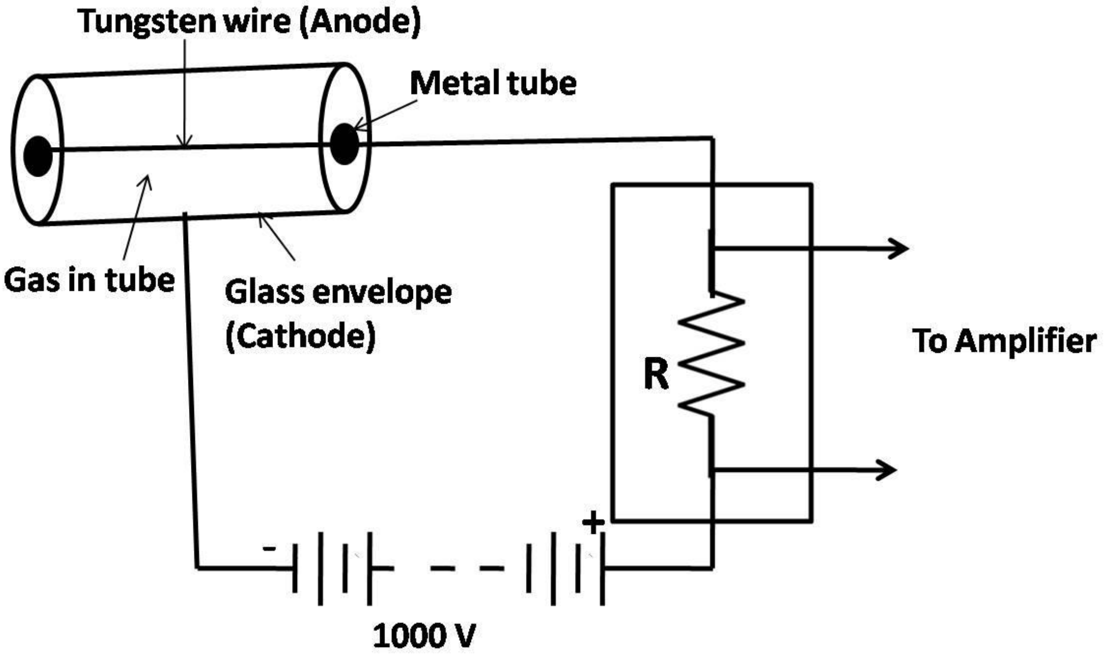


Figure 4.13. Working principle of G-M counter

GM counter consist of a metal tube covered with glass which acts as cathode and along the axis there is thin wire made up of tungsten which acts as anode. The tube is filled with inert gas such as Argon, Helium, Neon with halogen added at low pressure. There is window at one end where mica sheet is fitted, from this window radiation enter into the tube. A high potential difference of about 1000 V is applied between the electrodes through a high resistance R of about 100 mega ohm. (Fig.4.13)

Working of G-M counter

It is based on the principle of ionization as the charged particle passing through the gas medium present in the tube ionizes the atoms of gas by energy transfer.

High potential difference i.e. 1000 V is applied across the anode and cathode of the tube so that a high radial electric field nears the central wire is obtained (Fig.13). Due to the formation of high electrical field electrons generated by ionizing collisions between a high-speed particles entering the tube and the inert gas atoms are accelerated towards the anode wire by the strong electric field and acquire within a very short distance a high speed of their own. Because of this speed, they too can ionize other atoms and free more electrons. The object of the counter is to produce a single pulse for each particle entering the tube. This can only be achieved if spurious pulses due to secondary electrons released from the cathode surface by the bombardment of ions are completely suppressed so that the tube can recover as quickly as possible to be in a state when it is able to record the next entering particle. A quenching gas (halogen or organic vapours) introduced into the tube is to serve this purpose. The idea is to allow the inert gas ions on their way to the cathode to collide with the heavy molecules thereby transfer

Their charges to the molecules and become neutralized - a process known as quenching. The molecular ions thus produced move slowly to the cathode and on reaching there, capture electrons from the cathode surface to become neutral molecules. The multiplication of charges repeats itself in rapid succession producing within a very short interval of time an avalanche of electrons. The electron avalanche is concentrated near the central wire while the positive ions, being much heavier, drift slowly toward the cathode. The number of electrons striking the wire can be measured to detect the presence of radioactive emission.

8.4.2- SCINTILLATION COUNTERS

Scintillation counter is important instrument generally used for the detection and measurement of radiation. The scintillation counter in which radioactive materials exposed to atoms within the detector that temporarily absorb the radiated energy. These excited atoms return to their unexcited state and emit photons that are detected by the scintillation counter.

Working of Scintillation Counter

In scintillation counter there is lining of phosphor in one end of the photomultiplier tube. Its inner surface is coated with photoemission called photocathode. It acts as negative terminal. There is presence of several numbers of electrodes called dynodes which are arranged in the tube at increasing positive potential. When charge particles reach and strike the photocathode in photomultiplier tube releasing an electron. These electrons accelerate toward the first dynode and strike it. More numbers of secondary electron are emitted which accelerates towards the second dynode so on. Finally the chain continues multiplying the affect of the first charged dynode. Due to this process there is release of a voltage pulse across the external resistance. This voltage pulse is amplified and electronic counter (Fig.4.14).

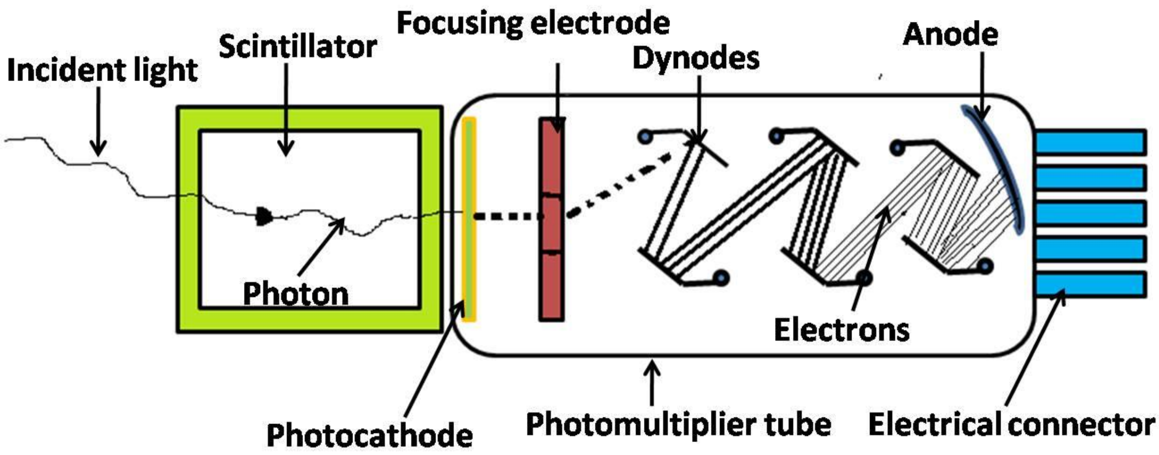


Fig.4.14 Scintillation Counter

Scintillation counter consist of two parts

1. Scintillator: Scintillation is a material may be solid, liquid (organic or inorganic) or gaseous which give luminescence or exhibit scintillation when struck by ionization radiation. The NaI crystal is the most widely used scintillation material which under UV light glows blue.
2. Photo multilplier Tube: PMT is generally used to creates strong electrical output from a weak signal developed by striking of photon on photocathode. This electron accelerates toward the first dynode and strikes it. More numbers of secondary electron are emitted which accelerates towards the second dynode so on. Lastly electron hits the anode and creates current which flow to ground through a resistor where it creates a voltage drop that can be counted.

8.5. SUMMARY

Modern laboratories of biological sciences should equipped with basic instrument like pH meter, UV-Vis spectrophotometer, centrifuge etc. pH meter is generally used to determine the pH of soil, water and culture medium used for the cultivation of fungi and bacteria. There is presence of electrode which very sensitive to detect the change in H ion concentration. The electric circuit measure the electromotive force developed across the electrode pair. Spectroscopy is the measurement of (spectrum of light) electromagnetic radiation, absorbed, scattered, or emitted by atoms, molecules, or other chemical species.

Each chemical species has unique energy states; spectroscopy can be used to identify the interacting species separately. Centrifugation is based on the principle of centrifugal force in which liquid are subjected to high speed to separate solid from liquid or liquid from liquid depending on the density. In centrifugation heavy particle settled down and light particle will rises to the top. Geiger- Muller Counter which is used as particle detector that measure ionization radiation i.e. beta particle and alpha particles except gamma particle because it does not ionize the gas. Scintillation counter is important instrument generally used for the detection and measurement of radiation. The scintillation counter in which radioactive materials exposed to atoms within the detector that temporarily absorb the radiated energy.

8.6. GLOSSARY

Bandwidth: The frequency span where constant amplitude input will produce a meter reading within a specified limit (usually 3db). In controllers, the region around the set point where control occurs.

Controller: A device capable of receiving a signal from a process and regulating an input to that process in order to maintain a selected operating condition

Electromotive force (emf): An electrical potential difference which produces or tends to produce an electric current.

Double Beam: In a double beam spectrophotometer the beam from the light source is split in two. One beam illuminates the reference cell holder and the other illuminates the sample.

8.7. SELF ASSESSMENT QUESTIONS AND POSSIBLE ANSWERS

Multiple Choice Questions:

1. Which filament lamp is used for UV and visible light:

(a) Tungten (b) Argon

(c) Platinum (d) Lithium

2. G-M Counter is used to measure the

(a) RBC (b) WBC

(c) Intensity of the radioactive radiation (d) intensity of visible light

3. Centrifugation is based on the principle of

(a). Gravitational force (b) Centrifugal force

(c) Frictional force (d) Vander wall force

4 pH scales ranges from

(a) 0 to 6 (b) 0 to 7

(c) 0 to 10 (d) 0 to 14

8.8. REFERENCES

1. Bair, E.J. (1962). Introduction to chemical Instrumentation, McGraw hill, New York

2. R.C.Dubey and D.K. maheshwari.(2012). Text book of Microbiology. S. Chand & company.

8.9 TERMINAL QUESTIONS

Q1. Write principle, procedure and working of pH meter.

Q2. Draw labeled diagram of GM counter.

Q3. Write different types of centrifuge used in laboratory

Q4. Write short note on

- **a.** Bench centrifuge
- **b.** Cooling centrifuge
- **c.** Ultra centrifuge

Q5. Describe the working and principle of Scintillation counter?

Q6. Explain Beer-Lambert law?

Q7. Write principle and working of UV-Vis spectrophotometer

UNIT 9: MICROTOMY AND MICROSCOPY

CONTENT

9.1OBJECTIVES

In this chapter, students will able to learn how tissue or cells are fixed and stain to study the histopathology. The chapter also deals with the different types of microscope used for the study of the microscopic world.

9.2 INTRODUCTION

At the beginning of science, there was no sophisticated and advanced instrument available to explore the world of an invisible entity. It was made possible by the development of a microscope in which the microscopic cell, bacteria, fungi, and other microscopic things was seen and detailed structure studied. As the science progress, there was the development of Microbiology has been progressed and now we studied the ultramicroscopic microorganism such as viruses and bacteriophages. Due to the development of the microscope it is possible that we can differentiate between different types of cells. The importance of the microscope in the histopathological investigation cannot be ignored. The abnormal changes in the cell and tissue of human being can be easily seen with the help of the microscope. There are different types of microscope has been developed these days according to the need of researchers and scientists i.e. Compound microscope, Stereo microscope, phase contrast microscope, Bright field microscope, Dark field microscope, Transmission electron microscope, Scanning electron microscope.

The morphological changes in the cell or tissues can be visualized through microscope when it is correctly processed and stained. This chapter focuses on different types of microscope used in the laboratories for various investigations.

9.3MICROTOMY

The trimmed block of paraffin is then fixed on a block holder. This block holder is then secured to the microtome and oriented appropriately with respect to the knife. With each revolution of the microtome handle, the specimen moves through the blade and a section of desired thickness is produced. Each successive section adheres to the preceding are forming a continuous ribbon. The ribbon is cut into small piece and transferred onto clean albumenized slides. The ribbon is flattened by putting a few drops of water. These floating sections are stretched using a hot plate or warming table.

Staining

Next, the paraffin is removed with xylene or another appropriate solvent and the specimen is rehydrated. It is then stained, dehydrated, cleared (made transparent) with xylene, covered with a suitable mounting medium, and topped with a cover-slip. Various stains are available to the histologist. Hematoxylin and eosin (H&E) is a frequently used combination of stains. Haematoxylin imparts a purple colour to substances, but must be linked to a metallic salt called a mordant before it can function effectively. This combination, called a lake, carries positive charges and behaves as a basic (cationic) stain. The lake combines electro statically with negatively charged radicals such as phosphate groups of nucleoproteins. Substances that become colored by a basic stain are said to be basophilic. Methylene blue, toluidine blue, and basic fuchsine are basic stains. Unlike haematoxylin, these stains have molecules that carry a positive charge of their own and do not require a mordant. Acidic (anionic) stains carry a negative charge and colour cell or tissue components that bear positive charges. Eosin is an acid stain. It imparts orange or red colorto acidophilic substances. Other commonly used acid stains are orange G, phyloxine, and aniline blue.

In addition to the widely used H & E staining procedure, numerous other stain combinations and techniques are available. Some are especially useful for identifying certain tissue elements. For example, trichrome procedures such as Mallory's and Masson's specifically stain collagenous fibers within connective tissue. Orcein and Weigert'sresorcinfuchsin are stains used to color elastic fibers, providing a means of distinguishing them from other fibrous elements. Reticular fibers and nervous tissue components such as neurons, myelin, and cells of the neuroglia can be stained by procedures employing the use of silver. There are also special histochemical and immune histochemical procedures that make possible the localization of various carbohydrates, lipids, and proteins found in tissue. Stains such as Wright's and-Giemsa's (Romanovsky stains) are available for differentiating the various cells found in blood and bone marrow.

Fig 9.1Microtomy

Fixation

As discussed in the preceding chapter, fixation of a sample is the first and most important step in microscopical examination. After being removed from an animal/plant, a tissue or organ is cut into pieces. These are placed in a fixative for suitable time at room temperature or at 4^{o}C in a refrigerator. Commonly used fixatives are 4-10% formalin, aqueous or alcoholic Bouin, acetic alcohol formalin, Comoy's fluid. The purpose of fixation is to preserve normal morphology of the tissue and prepare it for further processing.

Dehydration

After fixation, the next step is to remove the fixative. The aqueous fixative may be removed by washing the fixed tissue in running water. Alcoholic fixative is removed by washing the tissue in 70% alcohol several times.

Removal of fixative is followed by dehydration. Dehydration is performed by transferring the sample through a series of alcohols of increasing concentrations upto 100% alcohol.

Clearing

Dehydration is followed by clearing. Several clearing agents such as xylene, toluene and benzene can be used. These agents are miscible with alcohol as well as paraffin wax. This intermediate step is essential before infiltrating the dehydrated tissue with paraffin because alcohol and paraffin do not mix.

Embedding

Cleared tissue is embedded in a suitable embedding medium. Paraffin wax is the most preferred medium. Several plastics are also available nowadays and they offer better sections. Paraffin embedding is carried out in an oven at a temperature just above the melting point of the paraffin. When infiltration is complete, the specimen is transferred to an embedding mold of fresh paraffin which is allowed to harden. Then the mold is removed and excess paraffin is trimmed away.

9.4. TYPES OF MICROSCOPES

The type of microscope was designed as per the requirement and nature of the specimen. So, there exist huge variations in microscopes required as per the desired magnification and visualization. However; some most common types of microscopes include compound (having a combination of lenses) light microscopes, bright and dark field, phase contrast, fluorescence, and electron microscopes. The modern optical science of microscopy is based on the designs of German physicist Ernst Abbé. The applicability of various microscopes depends upon the most important feature, the resolution power.

RESOLUTION OF MICROSCOPES

To ascertain the distinguished minimum distance between two closely located objects with the help of a lens is called resolution which can be measured using the Abbé equation:

$$d = \frac{0.5\,\lambda}{nSin\theta}$$

Where, λ is the wavelength of the light used to illuminate the sample and nSinθ is the numerical aperture (NA). Angular aperture (θ) as in figure 1 is ½ the angle of the cone of light (illuminating a specimen) entering the lens of an objective piece of the microscope.

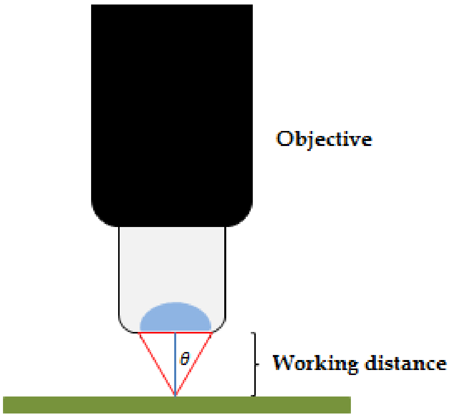


Figure 9.2Angular aperture θ of the cone formed on lens by specimen

The numerical aperture is nSinθ. NA depends upon the working distance which influences the angular aperture. For higher values of NA, working distance should be less for achieving higher angle of cone resulting in the higher resolution. It should be remembered that angle of the cone depends on the refractive index (n) of the medium and the optical material of lens. The refractive index of air is 1.0 and the maximum value of angle is 90^0. Sin90 is 1 therefore; lenses working in air cannot have a numerical aperture greater than 1.0. To achieve higher resolution greater angle of cone is required for which *immersion oil* (colorless) having a refractive index

Equal to that of medium can be used. It also helps to collect the otherwise dispersing light (from the edges of specimen) on the objective lens.

LIGHT MICROSCOPE

In the modern age of advanced scientific endeavor, sophisticated microbial examinations are being deliberated. All the modern age microscopes are compound and light microscope is the simplest compound microscope, which is very well understood by students of science. Several variants of light microscopes viz.Bright field, dark field, phase contrast, fluorescence and many others are available for microscopic examination. A simple compound light microscope has two sets of lenses i.e., an objective which form magnified image, which is further enlarged by another lens. All the light microscopes have a source of white light.

BRIGHT FIELD MICROSCOPES

This is an ordinary microscope, commonly used by students of preliminary science. The bright field microscope forms a dark image against a bright field. Two types of lens sets viz., objective (commonly 4 in number of different magnifying power) and eyepiece (ocular) are used (figure 2). The total magnification power of microscope using particular objective is calculated by multiplying the power of objective to the power of eyepiece for e.g., 45X objective and 10X ocular yield 450X magnification.

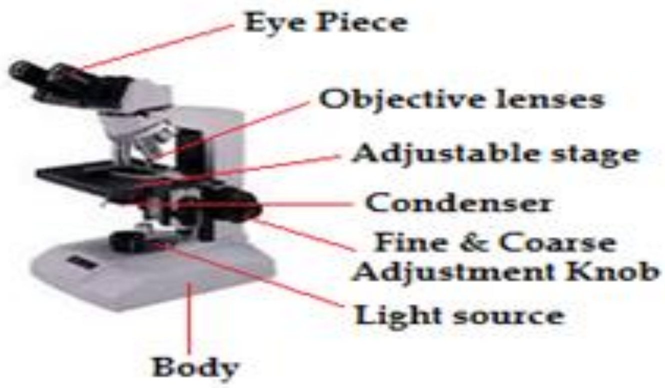


Figure 5.3. Bright Field

The objective lens forms an enlarged real image which is further magnified by eyepiece and the image is virtual formed 25 cm beyond the stage. Sub-stage condenser is also used some times for the purpose to increase the resolution. The specimen can be brought in focus of objective lens by fine and coarse adjustment knob (figure 2). It should be remembered that a microscopes are par focal i.e., even if the objective is changed the specimen remain in focus. To visualize by bright field microscopes, staining of the specimen is often required and it is very difficult to observe living cells through this microscope.

DARK FIELD MICROSCOPE

Dark field microscopes are used to magnify the images of living and unstained living microorganisms. In fact this is also technically and structurally similar as bright field microscopy, only difference is the addition of a dark field stopper between light source and condenser (figure 3). This light stopper (figure 4) let the light to pass from the open space and eventually condensers form a hollow cone of light and focused on the specimen to illuminate. The light that is reflected and refracted by the specimen forms an image, rest of the part around the Specimen appears black. Since the background of the specimen image is dark, the microscopy is called dark field.

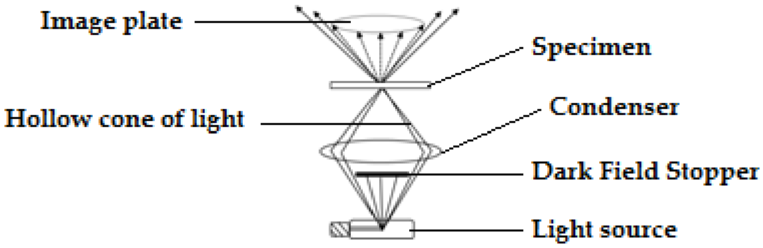


Figure 95.3: Optics of a Dark Field microscope

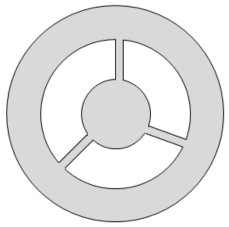


Figure 9.4. Dark Field

PHASE-CONTRAST MICROSCOPE

Phase-contrast microscopes produce images as a result of slight differences in refractive index and cellular density. The images so formed let the observer to have idea of the thickness of the organism under study.

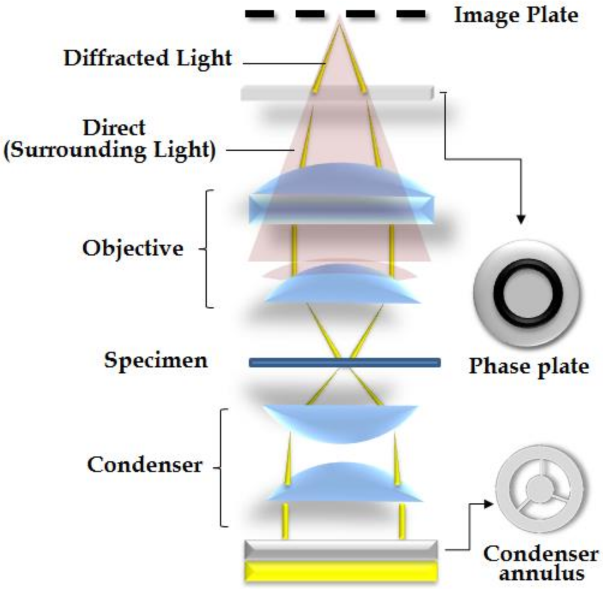


Figure 9. 5 Optics of a phase-contrast microscope

Like dark field, phase-contrast microscopes also have an annular stop disk with a slight difference of thin transparent ring which produces a hollow cone of light. As the cone passes through a living organism (bacterial cell or any other living cell) some light rays are bent because of the density variations and refractive index. The light wavelength gets retarded by ¼. This deviation focuses to form an image of specimen. The un-deviated light strike a phase

ring in a phase plate (figure 5), a disk placed in the objective. The deviated light (from the specimen) miss the phase plate and pass through the rest of the plate. The deviated and un-deviated lights have ½ and ¼ wavelengths of the light from the source. The un-deviated passing through the plate cancels each other during the formation of image.The un-deviated light render a bright background and deviated light form dark image giving a sharp contrast resulting in the peculiar image of specimen. This is dark-phase-contrast microscope.

FLUORESCENCE MICROSCOPE

Usually, light source is required to illuminate the specimen, which is the basis of image formation by microscopes. However; some objects emit light giving a base for fluorescence microscopy. In fact, some molecules reach to an excited state by absorbing radiant energy and re-emit the same energy in the form of light of lower wavelength than what was absorbed earlier, this is called fluorescence.

In fluorescence microscope, usually specimens are stained with certain dye molecules called *fluorochromes*such as acridine orange, Lucifer yellow, ethidium bromide, TOTO – 1(vital stain), DAPI (diamidino – 2 – phenylindole a DNA specific stain),which are excited by intense light produced by *mercuryvaporarc* lamp.

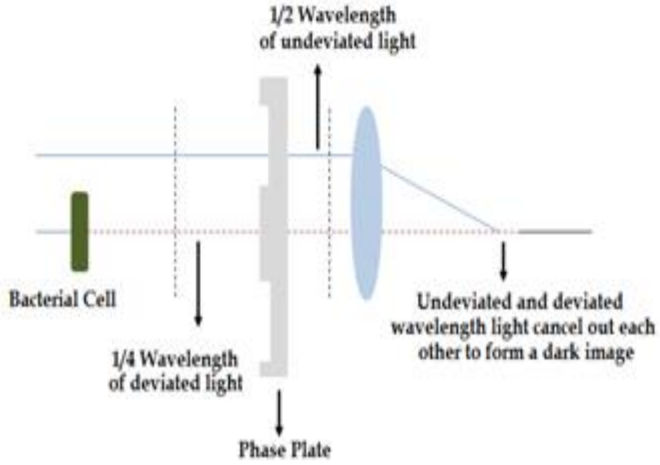


Figure 5.6. Wavelength deviations through phase plate in a phase contrast microscope

Since the light intensity is very high hence an infrared filter is usually used to avoid likely damages to the microscope itself. Like dark field microscopes, a dark field condenser renders a dark background to the image so that fluorescence coming out from the specimen could be seen. Light of specific wavelength excite the stained specimen, the image is formed from the light emitted. A barrier filter is also placed to remove the ultraviolet light that could otherwise damage the viewer’s eye.

ELECTRON MICROSCOPY

In practical naked eyes can resolve only up to 10^{-3} m (thickness of hairs) while light microscopes can resolve up to 10^{-7}m, although the objects having lesser size exist and cannot be visualized by light microscopy. In the year 1931 German physicist Ernst Ruska and electrical engineering Max Knoll constructed the first prototype of electron microscope. Modern age electron microscopes have magnification power 10,000,000X and resolution up to 50 Pico-meters because electron beam having wavelength about 100,000 times shorter than visible light is used to illuminate the specimen. Two basic type viz., transmission and scanning electron microscopes are commonly used these days.

TRANSMISSION ELECTRON MICROSCOPE (TEM)

In the transmission electron microscopes, an electron beam is generated by a heated tungsten filament positioned in the electron gun. This electron beam is focused on the specimen, which scatters the electron as per the density. Since the electrons cannot pass through the glass lens therefore magnetic lenses are used to focus. The entire set of magnetic lenses and specimen are worked under high vacuum as the air inside the column may get charged and moisture along with other molecules may scatter the electron beam. The electrons passing through the specimen are scattered, the dense part of the specimen scatter more as compared to the thin region. Thus the

image formed by the scattered electrons will be dark indicating the density and the brighter region indicate the electron transparent regions (thin regions).

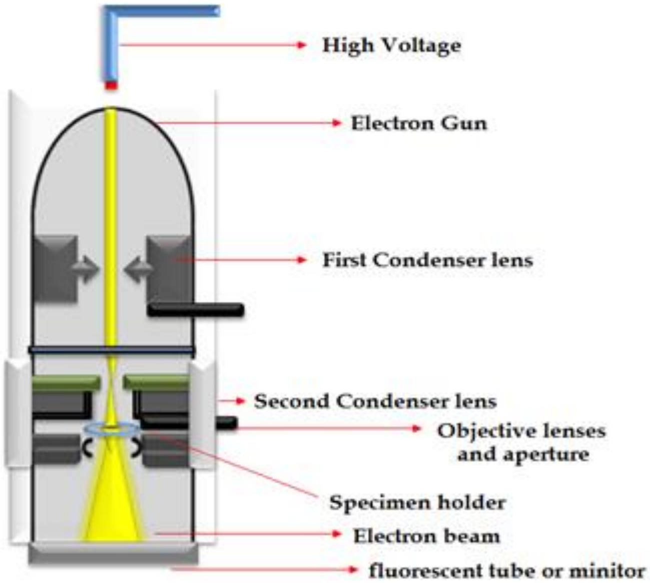


Fig 9.7 Transmission electron microscopes

The scattered electrons form enlarged image on the fluorescent screen and on the photographic plates, which can be used for reporting purposes.

Scanning Electron Microscope (SEM)

Scanning electron microscopes are the recent advancement in electron microscopy facilitating ease in the microbiological and other life science related examinations. Principally same as TEM except having advancements such as specimen preparation is easy, more well defined images, 3D direct viewing and photography and ease of working. Images in SEM are formed by the secondary electron produced as a result of shower of primary condensed electron beam under high vacuum. Figure 7 present the optics of a SEM. An Electron beam is produced by an electron gun so positioned that primary electrons are focused on the shielded specimen (metal coated). As soon as electron beam strikes the surface secondary electrons are produced (more from elevated and fewer from depressed region of the specimen surface), which are detected by a detector as lighter (indicating elevation because of more secondary electron) and darker (indicating depression because of few secondary electron production) regions. The electronic signals are sent to the photomultiplier which is further amplified. A cathode ray tube receives signals from the photo-multiplier and forms a picture. The image can be seen just as a computer monitor or television and can be photographed.

SPECIMEN PREPARATION FOR ELECTRON MICROSCOPY

TEM

Since electrons are rapidly absorbed and scattered by solid matter, therefore, extremely finely cut slices of 20 to 100 nm irrespective of material can only be observed by TEM. The slicing of the object intended to be observed through TEM need technical practice whereby specimen are fixed with the fixer chemical for e.g., osmium tetra oxide followed by rigorous dehydration with organic solvents such as ethanol, acetone etc. Usually, paraffin or any other liquid epoxy plastic are used to soak the specimen and then solidified to form a block. The thin slices are usually cut using ultra microtome’s. It should be remembered that prior to use electron microscopes, complete dehydration of the specimen is a must.

SEM

The specimen preparation for SEM is easy however; it involves some of the basic requirement as that of TEM for e.g., dehydration and preservation. In scanning electron microscopes the entire procedure occurs at high vacuum condition, this may also damage the specimen and since secondary electrons are required for image formation, specimen are shielded with a thin layer of metal which prevent the damage and also helps emission of secondary electron from the specimen surface.

9.5 SUMMARY

The secret of the microscopic world cannot be deciphered until the microscope has not been discovered. The microscopy is a very useful instrument for the study of materials and can be used to gain valuable information about a large variety of specimens such as bacteria, fungi, protozoa etc. Some knowledge of the material and the information that is required is essential to determine the best techniques to employ when preparing and examining specimens. In Microscopy sample preparation is the most important of microscopy, as this determines the quality of the images produced. Disadvantage of the optical microscope is its resolution. This limitation is overcome by a scanning electron microscope (SEM).In addition, for 'transparent' specimens; in particular those required for motility, polarized light microscopy is best method for this type sample.

9.6. GLOSSARY

Analyzer: A analyzer is used with a polarizer to provide polarizing light.

Achromatic Lens: A lens that helps to correct the misalignment of light that occurs when it is refracted through a prism or lens.

Abbe Condenser: A lens that is specially designed to mount under the stage and which typically moves in a vertical direction.

Eyepiece: The eyepiece is the lens nearest to your eye.

Compound Microscope:Originally used to describe a microscope with more than one objective lens, a compound microscope is now generally understood to be a high power microscope with multiple, selectable objective lens of varied magnifications.

Condenser:A lens that concentrates the light on a specimen and increases the resolution. Found in or below the stage on compound microscopes.

Cover Slip:A thin, square piece of glass or plastic placed over the specimen on a microscope slide. It flattens out liquid samples and helps single plane focusing.

Electron Microscope:A type of microscope that uses electrons rather than light to create an image of the target.

Darkfield Microscopy: A technique used to enhance the contrast in unstained specimens. It works on the principle of illuminating the sample with light that will not be collected by the objective lens, so not form part of the image.

Monocular Head: A *microscope* head having a single eyepiece lens.

Nosepiece: The part of the *microscope* that holds the objective lenses also called a revolving nosepiece or turret.

Numerical Aperture (N.A.): This is a number that expresses the ability of a lens to resolve fine detail in an object being observed.

9.7. SELF ASSESSMENT QUESTIONS AND POSSIBLE ANSWERS

Multiple Choice Questions:

1. Numerical aperature is:

(a) nSin$\boldsymbol{\theta}$ (b) nTan$\boldsymbol{\theta}$

(c) nSec$\boldsymbol{\theta}$ (d) nCos$\boldsymbol{\theta}$

2. In fluorescence microscope, usually specimens are stained with certain dye molecules

(a) Simple stain (b) Crystal voilet

(c) Alberts Stain (d) Acridine orange

3. Transmission Electron Microscope magnify upto

(a). 100X (b) 1000X

(c) 400,000X (d) 400X

4 3D image of specimen was obtained with

(a) Compound Microscope (b) SEM

(c) TEM (d) Phase contrast Microscope

5. Image can be seen in SEM

(a) Fluorescent Microscope (b) Anode

(c) As Graph (d) Phosphorescent screen

9.8. REFERENCES

1.. R.C.Dubey and D.K. maheshwari.(2012). Text book of Microbiology. S. Chand & company.

2. P.K.Bajpai. (2006). Biological instrumentation and Methodology. S.Chand and Company.

3. Bruce Alberts. (2014). Molecular Biology of the Cells.W. W. Norton & Company; Sixth edition

9.9 TERMINAL QUESTIONS

Q1. Write principle, procedure, and working of Light Microscope.

Q2. Draw labeled diagram of a compound microscope.

Q3. Write principle and working of Fluorescent Microscope

Q4. Write short note on

- **d.** Resolution
- **e.** Numerical apertures
- **f.** Magnification

Q5. Describe the working and principle of the Transmission Electron Microscope?

Q6. Write principle and working of Scanning Electron Microscope?

UNIT 10: SEPARATION TECHNIQUES AND CRYOPRESERVATION

CONTENT

10.1 OBJECTIVE

In this chapter you will learn different techniques which are related to purification and separation of desired molecules and develop a basic understanding of Electrophoresis theory and also become familiar with the procedure involved in agrose gel electrophoresis to visualize DNA.

This chapter also deals with the preservation of cells, organelle, tissue, organs and microbes to very low temperature i.e. cryopreservation.

10.2. *INTRODUCTION*

As the biological sciences progress there is need of some sophisticated, accurate and sensitive method used for the purification and separation of desired component from the complex or mixture of two or more chemicals. The chromatography techniques have two parts i.e. mobile phase and stationary phase. In a chromatography sample mixture is placed in liquid or gas called mobile phase and the mobile phase carries the sample through a solid support called the stationary phase. In case of Thin Layer Chromatography (TLC) the stationary phase is silica coated on glass plate. These techniques have disadvantage such as it does not have long stationary phase so the length of separation is limited and another problem with this method it operates in open system so the chromatogram may influenced by temperature and humidity. Now days, there is some modern methods which is based on the principle of TLC such as HPLC, GC, HPTLC and Electrophoresis. HPLC is one of the important instruments used in the pharmaceutical industry for the analysis of drug, plant metabolite, and toxins and in medical science it is also use for the detection of hazardous chemicals in blood. Electrophoresis is most commonly used technique in molecular biology for the analysis of nucleic acid and protein. It is based on method of separating molecules on the basis of charge and molecular weight when applied electrical field. On the basis of molecular weight of nucleic acid we can predict the nucleic acid of unknown microorganism.

10.3 CHROMATOGRAPHY

10.3.1 CHROMATOGRAPHY

Chromatography is another chemical procedure to find out and to separate the contents of a mixture of two or more chemicals. The literal meaning of chromatography is the study of colors first developed and described by Russian scientist Mikhail Tswett in 1900 which was further developed by Martin and Synge in early 1950s. In fact the chromatography is a separation technique based on the “partition or distribution coefficient or more appropriately partition constant or partition ratio (K_d) that describe the way in which a compound distribute itself between two immiscible phases".

K_d = concentration of compound in phase A / concentration of compound in phase B

The term effective distribution coefficient is defined as the total amount, as distinct from the concentration, of substance present in one phase divided by the total amount present in the other phase. It is in fact the distribution coefficient multiplied by the ratio of the volumes of the two phases. If the K_d of a compound between two phases e.g., A & B; is 1 and if the compound distribution is of 10 cm^3 and 1 cm^3 respectively, the concentration is the two phases will be same however; the total amount of the compound in phase A will be 10 times the amount in phase B.

There are two basic and essential phases' viz., stationary and mobile in chromatography. Stationary phase remain stable and do not move at all and provide base matrix for eg., solid, gel, liquid solid / liquid mixtures etc. on which compounds are separated. Mobile phase are either liquid or gaseous which flows over or through the stationary phase. If base matrix support is polar (e.g., paper, silica etc.) it is forward phase or if the matrix is non-polar (C-18) it is reverse phase chromatography. The choice of stationary and mobile phase is made so as to have different distribution coefficient for the mixture sample. This is also achieved by setting up following:

1) Adsorption equilibrium between stationary and mobile phase
2) Partition equilibrium
3) Ion exchange equilibrium
4) An equilibrium between a liquid phases trapped inside the pores of a stationary porous structure and a mobile liquid phase as in permeation or molecular exclusion chromatography.
5) Equilibrium between a stationary immobilized ligand and a mobile liquid phase.

Chromatography may be preparative (separating the components of mixture or in advanced terms "purification") or analytical (measuring the relative proportion of analyte in a mixture).Furthermore, there exist two basic types of chromatographic viz., column chromatography and planar chromatographic techniques.

10.3.2. COLUMN CHROMATOGRAPHY

Column chromatography is a separation technique in which the stationary bed is within a tube (glass, plastic or metal). The stationary phase comprised of various polar or non-polar porous materials. The filled column is packed by a liquid mobile phase carrying the mixture to be separated through the column.

There are two types of column used: 1) packed column which is characterized by the complete filling of the volume of column (figure 1), 2) open tubular column characterized as stationary phase concentrated along the column wall giving an unrestricted path to the mobile phase.

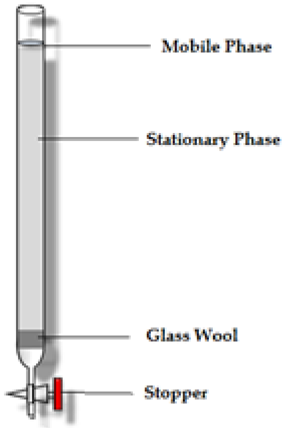
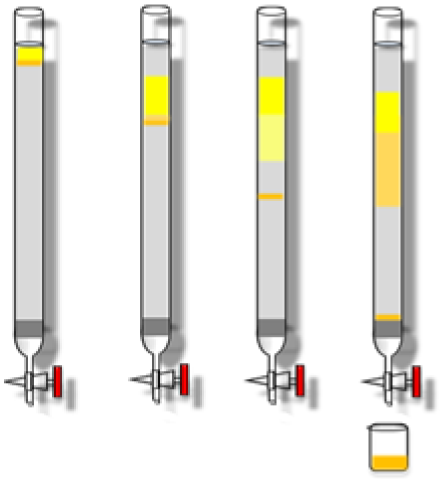


Fig: Packet Column

The separation is done in the column shown in the figure normally to separate the components of the mixture as for e.g., in the plant extract and not to quantify them. The packed column is gently handled following by the repeated washing with the solvent used as mobile phase. Sample is poured in to the column carefully and packed with the liquid mobile phase as in figure 2. The components of sample mixtures immediately get separated depending upon their partition coefficient (stationary to mobile) in the column forming band of different color having negligible differences therefore extra care is needed to separate each component. Separated components are called as *eluent*. The washing of column with the solvent to remove the components is called *elution*.



GAS CHROMATOGRAPHY

Gas chromatography is a column chromatography where the mobile phase is a carrier gas, usually inert gasses such as helium and/or nitrogen. As shown in figure 3, a GC typically consists of a gas tank providing inert carrier gas, the flow of which is controlled by a flow controller connected with the column oven. The column oven is a chamber is a metal casing wherein a column made of glass or metal is connected at one side with the carrier gas tubing and sample injection chamber and to the detector at the other end.

The column oven provides temperature for the fugitive diffusion of gas inside the column containing microscopic layer either of polymer or liquid as solid bed. The separated compounds are detected by the detector. Detectors may be of various types whereas; flame ionization and electron capture detectors are commonly used in most of the researches of life sciences and environmental sciences.

COLUMN OF GC

The heart of a gas chromatography is the column performance of which sets limit to the separation attainable and determines time of analysis. Two main type of columns viz., 1) packed column, first introduced by Martin in 1952 and 2) capillary packed column introduced by Halasz in 1968 are used these days. Pressure factors such as column inlet pressure, required speed of sample introduced, required sample size (detection limit) are largely determined by properties of the selected type of column.

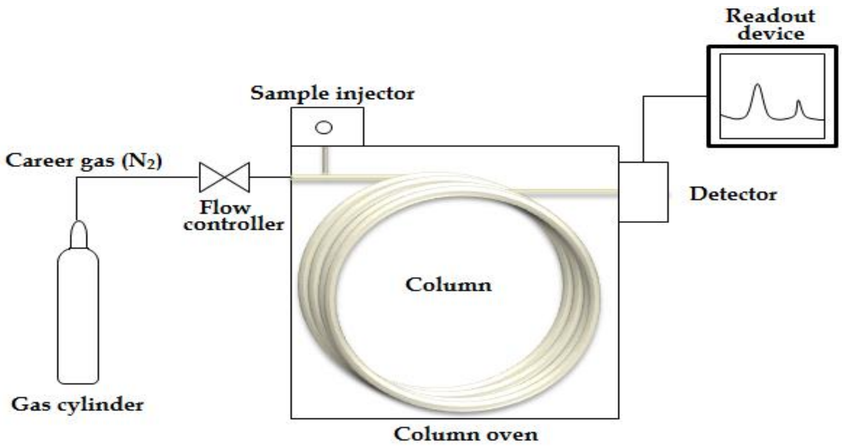


Figure 10.1. Schematic of a gas chromatograph

Factors determining gas chromatographic performance are directly linked to the column efficiency. These factors are:

(a) **Analysis time:** The residence time of the peak maxima in a column, the retention time t_R can be calculated by

$$\boldsymbol{t_R = t_0\,(r + k)}$$

Where t_0 is gas hold up time, r is the retention time of an unreacted component such as air, k is the capacity ratio. k can be derived from the partition coefficient K_d by multiplying the volumetric ration of stationary to mobile phase in the column.

$$k = K_d\,(\frac{V_{liquid}}{V_{gas}})$$

(b) **Efficiency:** Efficiency **is** expressed as the plates (n), the degree of band broadening a solute undergoes in a given column. This can be derived from the chromatogram:

$$n = \frac{t_{R^2}}{\sigma 2}$$

Where σ is the standard deviation (time units) is equal to half width at 0.607 of the height of a Gaussian peak.

The plate number (n) or the peak (bands) is related to the length of column and height H equivalent of a theoretical plate. H = L/n

H is further dependent on the linear velocity (u) of the carrier gas as in Figure 4.

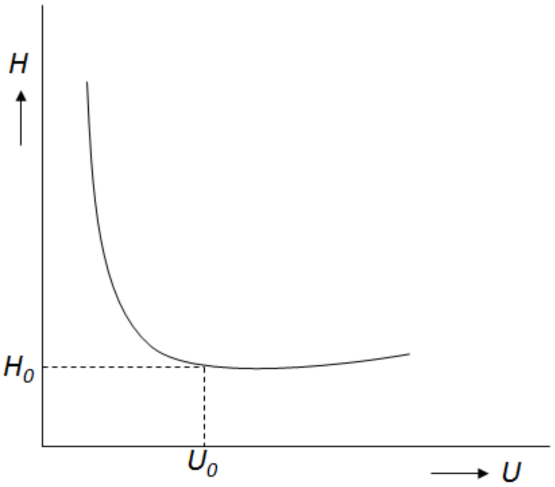


Figure 10.2 Dependency of Plate Height (H) and linear velocity (U) of carrier gas

(c) Resolution: (R) express the degree of separation of two components leaving the column shortly after each other

$$R = \frac{\left(t_{R_2} - t_{R_2}\right)}{\sigma_1}$$

Taking the values of t_R:

$$R = \frac{(\propto\ -1)k}{(k + I)}\ \sqrt{3}$$

Where $\propto$ is relative of two components

The separation is almost complete at R = 6.0 and just acceptable for R = 4.0.

(d) Speed of sample: The capacity of the column, which determines the most practical parameter, the sample size. Quantity of sample can be introduced in a GC either as a narrow band of high concentration or as a plug of correspondingly lower concentration.

If the sample is fed into the column as a band of high concentration the K_d will remain concentration dependent and each part of solute band will move at different speed resulting in the asymmetric peaks. If the sample is fed as plug of lower concentration, it goes under dilution with carrier gas as soon as it enters the column. This causes additional symmetric band broadening. The variance $\sigma\ ^2$(sec^2) of an eluting peak may be considered to be composed of the variance of the sample at the inlet $\sigma_1{}^2$and the variance $\sigma_0{}^2$due to column processes.

$$\sigma^2 = \sigma_1^2 + \sigma_0^2$$

Best resolution can be obtained at $\sigma_1{}^2 = 0$ (infinitely narrow sample band).

Carrier gas selection

The selection of the carrier gas depends upon the type of analyte, column temperature and detectors. For e.g., helium is used in discharge ionization detectors (DID), nitrogen is used in electron capture detector (ECD) and in flame ionization detector (FID) (in FID hydrogen is used as fuel gas). In general helium, nitrogen, argon, hydrogen and air are used as carrier gases in chromatography.

Detectors

All though advancements have been made in the detector technology and currently number of detectors viz., ECD, FID, DID (as above), PID (photo-ionization detector), PDD (pulse discharge ionization detector), MSD (mass selective detector) etc. are available for the analysis of variety of environmental and medical samples. However; ECD and FID are the most employed detectors worldwide discussed here in details.

Electron Capture Detector (ECD)

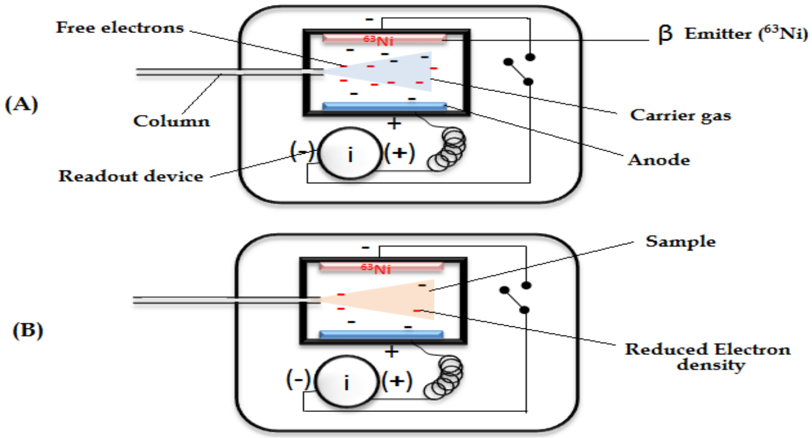


Figure 10.3. Schematic of Electron capture detector (A) high electron density in carrier gas (B) reduced electron density in sample + carrier gas

Electron capture detector (ECD) is a device used to analyze halogenated, nitriles and nitro compounds and is particularly important for the environmental, pharmaceutical, and forensic studies. ECD was first invented by James Lovelock in 1957. The functioning of ECD is based on a radioactive beta particle (electron) emitter ^{63}Ni (~ 10 millicurie or 370 mBq) which remains in a metal foil holding inside a chamber. ^{63}Ni emits electron that move towards anode so as to make the circuit complete resulting in to the generation of a current. Nitrogen is commonly used for analysis by ECD because of its low excitation energy enabling more electron density in the detector chamber generating current of high magnitude.

When the sample analyte enters along with the carrier gas, the molecules of analyte absorb / capture some of the electrons and reduce the current. while a background signal of current generated by the carrier gas is also distinctly visualize in a chromatogram. *The reduction in the current due to the analyte is directly proportional to the analyte concentration.*

The flame ionization detector (FID) is the industry standard method of measuring hydrocarbons (HC) concentrations. The sample gas is introduced in to a hydrogen flame inside the flame chamber. Any hydrocarbon in the sample will produce ions when they are burnt. Ions are detected using a metal collector (high voltage ion collector) which is based on a high DC voltage. The current across this collector is thus proportional to the rate of ionization which in turn depends upon the concentration of HC in the sample gas. Remember FID use hydrogen as fuel gas (Fig.6). If hydrogen flow is on and column is connected to the detector inlet fitting, hydrogen gas can flow in to the column oven and create an explosion hazard. Hydrogen flow inside the FID should be standardized and must not exceed 30 mL/min, while the carrier gas (N_2) flow should be 25 mL/min.

The analytical efficiency of a GC depends upon the appropriate column choice. There is variety of basic types of column (as discussed earlier in this section) present in the market out of which wall coasted open tubular (WCOT) and porous layer open tubular (PLOT) columns are widely used. PLOT columns are of three viz., 1) molecular sieve; 2) Divinylbenzene (DVB) and 3) Alumina (Al_2O_3) types are quite common for wide range of hydrocarbon analysis. The accuracy in the sample analysis by a GC has been developed to an advanced level, and gas chromatography mass spectrophotometer (GCMS) is one of them. The detection limit increases to ppt (parts per trillion) level. Since GCMS is more sophisticated and an expensive device, extra care is required while analyzing the samples.

HIGH PERFORMANCE LIQUID CHROMATOGRAPHY (HPLC)

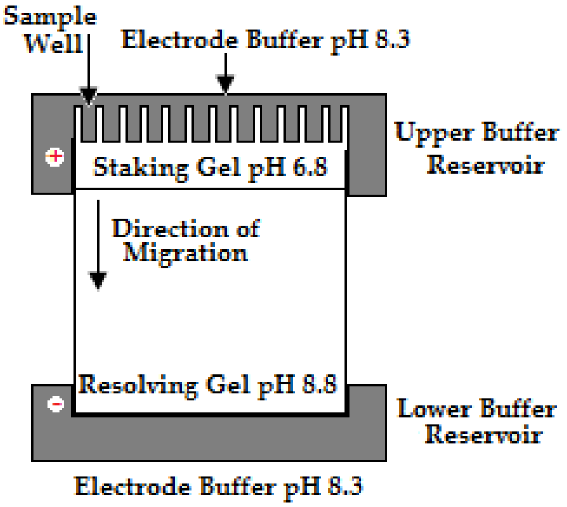
Gas chromatography required the volatilization of the sample, however; HPLC is the method to identify and quantify the analyte that cannot be converted into the gas phase. The unique retention time of the analyte in to the column and its characteristic peak as calibrated with standard is the principle of HPLC analysis. The time for a substance to pass through the column, termed the retention time which is related to the identity of the compound. Quantitative information is obtained from the area or height of the peak produced by the detector.

By the choice of appropriate solute (mobile phase) usually methanol and acetonitrile and the column (packed solid material) efficient separation can be achieved. In an HPLC the sample (20µl) is injected which is then carried by the solutes. The separation of analyte depends upon the interaction of the analyte and mobile phase with the stationary phase. High pressure are required to force a liquid through a tightly packed column filled with small particle material and the availability of high pressure solvent delivery systems is directly responsible for the "high performance".

The detectors used for the identification and quantification in HPLC are complex, in fact no single universal detector system has yet been developed. More than one detector having unique detection property can be used in a single HPLC in series. The most commonly employed detectors include *bulk property detector* (compare over all change in physical property of the mobile phase with and without an eluting solute); *solute property detectors* (respond to a physical property of the solute that is not exhibited by the pure mobile phase. These detectors are 100 times sensitive and detect sample on a nano-gram level; and *ultra-violet photometers* (most commonly used detector).

10.4. ELECTROPHORESIS

10.4.1 ELECTROPHORESIS



Electrophoresis is an analytical method frequently used in molecular biology and life sciences. It is applied for the separation and characterization of proteins, nucleic acids and subcellular-sized particles like viruses and small organelles. The working principle of electrophoresis is that the charged particles migrate towards their corresponding electrode in an applied electrical field. If conducted in solution, samples are separated according to their *surface net charge density*. The most frequent applications, however, use gels (polyacrylamide, agarose) as a support medium. The presence of such a matrix adds a sieving effect so that particles can be characterized by both charge and size. Protein electrophoresis is often performed in the presence of a charged detergent like sodium dodecyl sulfate (SDS) which usually equalizes the surface charge and, therefore, allows for the determination of protein sizes on a single gel. Additives are not necessary for nucleic acids which have a similar surface charge irrespective of their size. The resistance entrusted by the electrophoretic units also play significant role in producing migration related variations. The resistance of an electrophoresis unit depends on its size, gel thickness, amount of buffer, buffer conductivity and temperature. This resistance will normally decrease in time due to a slowly increasing temperature. Electrophoresis units which have a resistance below the minimum load resistance of a power supply will trigger an alarm.

For a comprehensive study and to achieve better separations, one should have knowledge of basics of electrophoresis. This section deals with such information.

Isoelectric point (pI)

The isoelectric point is the pH at which a particular molecule or surface carries no net electrical charge. Biological amphoteric molecules such as proteins contain both acidic and basic functional groups. Amino acids which make up proteins may be positive, negative, neutral or polar in nature, and together give a protein its overall charge. At a pH below their pI, proteins carry a net positive charge; above their pI they carry a net negative charge. Proteins can thus be separated according to their isoelectric point (overall charge) on a polyacrylamide gel using a technique called isoelectric focusing.

Isoelectric focusing (IEF)

Isoelectric focusing is an electrophoretic method in which proteins are separated on the basis of their pIs. It makes use of the property of proteins that their net charges are determined by the pH of their local environments. Proteins are positively charged in solutions at pH values below their pI and negatively charged above their isoelectric points. Thus at pH values below the pI of a particular protein, it will migrate towards the cathode during electrophoresis. While moving towards the electrode (cathode or anode) proteins lose some protons and their net charge drops and pH changes, this retard the protein movement, finally it stops to move as the pH become equals to the pI. Protein diffuses in to a pH higher than its pI, the protein will become negatively charged and will be driven towards the anode. In this way, proteins condense or focus, into sharp bands in the pH gradient at their individual pI values. Narrow pH range and high applied voltage gives excellent resolution in IEF.

Buffers of Electrophoresis

A buffer should be chosen with a pKa that is very close to the desired pH, preferably within a half point. The buffer will have the greatest capacity both to absorb and/or release protons with the acid and the base form well represented in solution. It should be noted that pKa is not constant for all conditions but is a function of the total ionic strength and the temperature, so the stoichiometry should be modeled after actual running conditions. In general, TAE (tris acetic acid with EDTA) and TBE (tris borate with EDTA) buffers are used in electrophoresis technique. TAE buffer provides optimal resolution of fragments >4 kb in length, while for 0.1 to 3 kb fragments, TBE buffer should be selected. TBE has both a higher buffering capacity and lower conductivity than TAE and therefore should be used for high-voltage electrophoresis. Additionally, TBE buffer generates less heat than TAE at an equivalent voltage and does not allow a significant pH drift.

Gel Concentration

Agarose is a polysaccharide extracted from seaweed. It is typically used at concentrations of 0.5 to 3%. The higher the agarose concentration the "stiffer" the gel. Agarose gels are extremely easy to prepare. It is also non-toxic.Agarose gels have a large range of separation, but relatively low resolving power.Polyacrylamide is a cross-linked polymer of acrylamide. The length of the polymer chains is dictated by the concentration of acrylamide used, which is typically between 3.5 and 20%. Polyacrylamide gels are significantly more complex to prepare than agarose gels. Because oxygen inhibits the polymerization process, they must be poured between glass plates. Polyacrylamide gels have a rather small range of separation, but very high resolving power.

10.5. CRYOPRESERVATION

Cryopreservation or Liquid Nitrogen

Cryopreservation is the method in which very low temperature is used to preserve living cell, organs, microorganism, sperm etc for longer period. Polge *et al* in 1949 was credited for cryopreservation of the first mammalian cell i.e. spermatozoa.   This technique is devised to keep the cells genetically stable, viable and metabolic inert. The disadvantage of low temperature preserving cell, tissue, organ and microorganism was formation of ice crystal which eventually disrupts the cell membrane resulting the death of cell. The main objective of cryopreservation is to replace water with other material that will not form ice crystals to overcome the freezing there is addition of antifreeze agent such as Glycerol and DMSO (Dimethyl Sulphoxide).

Cryopreservation media generally consists of a base medium, protein source, and a cryopreservative. The cry0preservative both protects the cells from mechanical and physical stress and reduces the water content within the cells, thus minimizing the formation of cell-lysing ice crystals. The protein source, often fetal bovine serum (FBS), also protects the cells from the stress associated with the freeze-thaw process. Cells are frozen slowly, at 1C/minute, using programmable coolers or by techniques outlined below. Generally, the optimum cell density to freeze per 1mL of cell suspension depends on the type of cell. Mammalian cells are usually frozen between 10^6 cells/mL to 10^7 cells/mL. The cryopreservation media may differ slightly for adherent and suspension cell types.

The following is procedure for cryopreservation of cells

1. Expand culture to allow for adequate cell density for the desired volume to freeze. The cell culture to be cryopreserved must be in the log phase of the growth cycle (approximately 2-4 days after subculturing). Determine the cell count for number of viable cells and the total cell concentration.
2. Centrifuge cells at approximately 200 to 400 x g for 10 minutes, allowing the cells to form a pellet. During centrifugation, determine the amount of freezing media to prepare. For example, if using 1 mL cryovials, divide the total cell concentration by the desired cell density. *Example: A $4x10^7$ cell suspension will yield a total of ten 1 mL alliquots at $4x10^6$ cells per alliquot.Prepare 10 mL of freezing medium to easily suspend the pellet at the correct cell density.*
3. Prepare the necessary volume of cryopreservation media (determined above) using the following guidelines:
4. Resuspend cell pellet(s) using the cryopreservation media, triturating to ensure a single-cell suspension with as few cell clumps as possible.
5. Dispense into the desired number of vials for cryopreservation.
6. Immediately transfer the vials to a freezer with a minimum temperature of -20 °C for one hour.
7. Transfer the vials to a -80°C freezer for 24 hours. Alternatively, a dry ice/methanol slurry using or other insulated box with a cover or lid may be used if a -80°C freezer is not available.
8. After 24 hours at -80°C, cells may be transferred to a liquid nitrogen storage (-196 °C)

10.6. SUMMARY

As the technology and techniques are developing day by day, which made the analysis and detection faster than earlier conventional method used. This chapter deals with some modern technique which is now used in research laboratories. Chromatography is chemical procedure to find out and to separate the contents of a mixture of two or more chemicals. Column chromatography is a separation technique in which the stationary bed is within a tube (glass, plastic or metal). The stationary phase comprised of various polar or non-polar porous materials. The filled column is packed by a liquid mobile phase carrying the mixture to be separated through the column. Gas chromatography is a column chromatography where the mobile phase is a carrier gas, usually inert gasses such as helium and/or nitrogen. As shown in figure 4.2, a GC typically consists of a gas tank providing inert carrier gas, the flow of which is controlled by a flow controller connected with the column oven. Gas chromatography required the volatilization of the sample, however; HPLC is the method to identify and quantify the analyte that cannot be converted into the gas phase. The unique retention time of the analyte in to the column and its characteristic peak as calibrated with standard is the principle of HPLC analysis. Electrophoresis is an analytical method frequently used in molecular biology and life sciences. It is applied for the separation and characterization of proteins, nucleic acids and subcellular-sized particles like viruses and small organelles. Cryopreservation is the method in which very low temperature is used to preserve living cell, organs, microorganism, sperm etc. for longer period.

10.7. GLOSSARY

Absorption: In chromatography, absorption signifies the process by which a solute partitions into a liquid-like stationary phase.

Mobile Phase: The eluate moving through the column. In gas chromatography (GC) this will be a gas, and in liquid chromatography (LC) a liquid.

Stationary Phase: The substance that remains in one place in the column. In GC this will be a liquid of high-viscosity, which clings to the inner walls of the column; in LC it will be some sort of packing, either solid or gel-based.

Eluate: The mobile phase exiting a column.

Eluent: The mobile phase entering a column**.**

Elution: The passage of the mobile phase through the column to transport solutes.

Flow Rate: The amount of mobile phase that has passed through the column per unit time. The units are millilitres per second (mL/sec) or, more commonly, millilitres per minute (mL/min).

Isoelectric focusing (IEF): Electrophoresis technique that separates proteins according to their isoelectric point (pI)

Isoelectric point (pI): pH value at which a molecule carries no electrical charge, or at which the negative and positive charges are equal.

PAGE: Polyacrylamide gel electrophoresis, a common method of separating proteins

Polyacrylamide gel electrophoresis (PAGE): Electrophoresis technique that uses polyacrylamide as the separation medium

Rf value: Relative distance a protein has traveled compared to the distance traveled by the ion front. This value is used to compare proteins in different lanes and even in different gels. It can be used with standards to generate standard curves, from which the molecular weight or isoelectric point of an unknown may be determined.

Adsorption: The process of retention in which the interactions between the solute and the surface of an adsorbent dominate.

Retention Time: The elapsed time between sample injection and the appearance of the chromatographic peak apex

Isocratic: Chromatographic conditions in which a constant composition eluent is used.

Ion Chromatography: An ion-exchange technique in which low concentrations of organic and inorganic anions or cations are determined using ion-exchangers of low ion-exchange capacity with dilute buffers.

Degassing: The practice of removing dissolved gases from the eluent. It can be achieved by helium sparging, applying vacuum to the eluent, ultrasonification or heating.

Isoelectric Point: The pH point at which a molecule no longer has a net charge.

Effective distribution coefficient: It is defined as the total amount, as distinct from the concentration, of substance present in one phase divided by the total amount present in the other phase.

SDS-PAGE: Separation of molecules by molecular weight in a polyacrylamide gel matrix in the presence of a denaturing detergent, sodium dodecyl sulfate (SDS). SDS denatures polypeptides and binds to proteins at a constant charge-to-mass-ratio. In a sieving polyacrylamide gel, the rate at which the resulting SDS-coated proteins migrate in the gel is relative only to their size and not to their charge or shape.

Running buffer: Buffer that provides the ions for the electrical current in an electrophoresis run. It may also contain denaturing agents. The running buffer provides the trailing ions in discontinuous electrophoresis.

Sample buffer: Buffer in which a sample is suspended prior to loading onto a gel. SDS-PAGE sample buffer typically contains denaturing agents (including reducing agents and SDS), tracking dye, and glycerol.

10.8. SELF ASSESMENT QUESTION

Multiple Choice Questions:

1. The Stationary phase in Reverse phase chromatography is made of

A. Polar B. Non Polar

C. Both D. None of the above

2. TLC is based on the principle of

A. Electrical mobility B. Partition chromatography

C. Ion exchange D. Adsorption

3. In isocratic method in HPLC the composition of solvent is

A. Remains Variable B. Remains constant

C. Both A and B D. None of the above

4. In gradient method in HPLC the composition of solvent is

A. Remains Variable B. Remains constant

C. Both A and B D. None of the above

5. HPLC stands for

A. High Performance Liquid Chromatography B. High Peak Liquid Chromatography

.C.High Pressure Liquid Chromatography D. All of the above

6. Antifreeze agent is

A. Glycerol B. DMSO

C.Both A and B D. Methanol

10.9. REFERENCES

1. Harold M. McNair, James M. Miller.2009 Basic Gas Chromatography, 2nd Edition, John Wiley.

2. Elsa Lundanes, Leon Reubsaet, Tyge Greibrokk.2013. Chromatography: Basic Principles, Sample Preparations and Related Methods. John Wiley

3. G. Lunn and N. Schmuff, John Wiley & Sons, 1997, HPLC Methods for Pharmaceutical Analysis.

10.10. TERMINAL QUESTIONS

Q1. Describe the principle and procedure of Thin Layer Chromatography.

Q.2. Write short notes on following

a. Flame Ionisation detectoe

b. Electron capture detector

Q.3. Write the principle, procedure and working of HPLC with suitable diagram.

Q.4. Write the principle, procedure and working of Gas Chromatography with suitable diagram.

Q.5. Write the principle, procedure and working of Electrophoresis.

Q.6. Write short notes on:

A. Isoelectric point
B. Isoelectric focusing (IEF)
C. Cryopreservation

Unit 11 Biotechniques (exercise based on chart/ picture or sample instrument)

CONTENT

11.1 Determination of pH using pH meter.

11.2 Demonstration of functioning of spectrophotometer.

11.3 Demonstration of use of bright field, phase contrast, dark field, fluorescence, confocal and electron microscopes (on photograph basis).

EXPERIMENT NO-01

OBJECTIVE: To demonstrate the pH using by pH meter.

INTRODUCTION

A **pH Meter** is a scientific instrument that measures the hydrogen-ion concentration (or pH) in a solution, indicating its acidity or alkalinity.

The pH meter measures the difference in electrical potential between a pH electrode and a reference electrode. It usually has a glass electrode plus a calomel reference electrode, or a combination electrode.

In addition to measuring the pH of liquids, a special probe is sometimes used to measure the pH of a **pH Meter** is a scientific instrument that measures the hydrogen-ion concentration (or pH) in a solution, indicating its acidity or alkalinity.

The pH meter measures the difference in electrical potential between a pH electrode and a reference electrode. It usually has a glass electrode plus a calomel reference electrode, or a combination electrode.

In addition to measuring the pH of liquids, a special probe is sometimes used to measure the pH of semi-solid substances.

Uses

- Knowledge of pH to greater or lesser accuracy is useful or critical in a great many situations, including of course chemical laboratory work.
- pH meters of various types and quality can be used for soil measurements in agriculture; water quality for water supply systems, swimming pools, etc.; brewing, industrially or domestically; healthcare, to ensure that solutions are safe when applied to patients or lethal as sterilants and disinfectants; and many other applications.

Circuit and operation

- Simple potentiometric pH meters simply measure the voltage between two electrodes and display the result converted into the corresponding pH value.
- They comprise a simple electronic amplifier and a pair of probes, or a combination probe, and some form of display calibrated in pH.
- The probe is the key part: it is a rod-like structure usually made of glass, with a bulb containing the sensor at the bottom.
- Frequent calibration with solutions of known pH, perhaps before each use, ensures the best accuracy. To measure the pH of a solution, the probe is dipped into it.

Probe care and cleaning

- Probes need to be kept clean of contamination as far as possible, and not touched by hand.
- Probes are best kept moist with a medium appropriate for the particular probe (distilled water, which can encourage diffusion out of the electrode, is undesirable) when not in use.
- If the bulb becomes contaminated with use it can be cleaned in the manner recommended by the manufacturer; a quick rinse in distilled water immediately after use, blotted (not wiped) off may be sufficient.
- One maker of laboratory-grade equipment gives different cleaning instructions for general cleaning (15' soak in a solution of bleach and detergent), salt (hydrochloric acid solution followed by sodium hydroxide and water), grease (detergent or methanol), clogged reference junction (KCl solution), protein deposits (pepsin and HCl, 1% solution), and air bubbles.

Calibration and use

- For very precise work the pH meter should be calibrated before each measurement. For normal use calibration should be performed at the beginning of each day.
- The reason for this is that the glass electrode does not give a reproducible e.m.f. over longer periods of time
- Calibration should be performed with at least two standard buffer solutions that span the range of pH values to be measured.
- For general purposes buffers at pH 4.00 and pH 10.00 are acceptable.
- The pH meter has one control (calibrate) to set the meter reading equal to the value of the first standard buffer and a second control which is used to adjust the meter reading to the value of the second buffer.
- A third control allows the temperature to be set.
- Standard buffer sachets, which can be obtained from a variety of suppliers, usually state how the buffer value changes with temperature.
- For more precise measurements, a three buffer solution calibration is preferred. As pH 7 is essentially, a "zero point" calibration (akin to zeroing or taring a scale or balance), calibrating at pH 7 first, calibrating at the pH closest to the point of interest (e.g. either 4 or 10) second and checking the third point will provide a more linear accuracy to what is essentially a non-linear problem.
- Some meters will allow a three-point calibration and that is the preferred scheme for the most accurate work. Higher quality meters will have a provision to account for temperature coefficient correction, and high-end pH probes have temperature probes built in.
- The calibration process correlates the voltage produced by the probe (approximately 0.06 volts per pH unit) with the pH scale.
- After each single measurement, the probe is rinsed with distilled water or deionized water to remove any traces of the solution being measured, blotted with a scientific wipe to absorb any remaining water which could dilute the sample and thus alter the reading, and then quickly immersed in a solution suitable for storage of the particular probe type.

TYPES OF PH METERS

A simple pH meter

pH meters range from simple and inexpensive pen-like devices to complex and expensive laboratory instruments with computer interfaces and several inputs for indicator and temperature measurements to be entered to adjust for the variation in pH caused by temperature. Specialty meters and probes are available for use in special applications, harsh environments, etc.

There are also holographic pH sensors, which allow pH measurement calorimetrically.

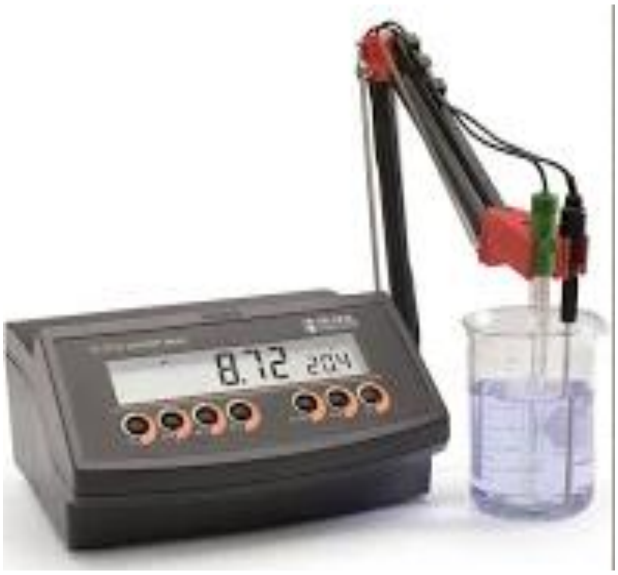


Figure 11.1 a simple pH meter

Figure 11.2 a Digital pH meter

HISTORY

The concept of pH was defined in 1909 by S. P. L. Sorensen, and electrodes were used for pH measurement in the 1920s.

In October 1934 Arnold Orville Beckman registered the first patent for a complete chemical instrument for the measurement of pH, U.S. Patent No. 2,058,761, for his "acidometer", later renamed the pH meter.

Beckman developed the prototype as an assistant professor of chemistry at the California Institute of Technology, when asked to devise a quick and accurate method for measuring the acidity of lemon juice for the California Fruit Growers Exchange (Sunkist). On April 8, 1935, Beckman's renamed National Technical Laboratories focused on the making of scientific instruments, with the Arthur H.

Thomas Company as a distributor for its pH meter.131–135. In its first full year of sales, 1936, the company sold 444 pH meters for $60,000 in sales. In years to come, it would bring in millions.

Radiometer in Denmark was founded in 1935, and began marketing a pH meter for medical use around 1936, but "the development of automatic pH-meters for industrial purposes was neglected.

Instead American instrument makers successfully developed industrial pH-meters with a wide variety of applications, such as in breweries, paper works, alum works, and water treatment systems. In 2004

The Beckman pH meter was designated an ACS National Historic Chemical Landmark in recognition of its significance as the first commercially successful electronic pH meter.

In the 1970s Jenco Electronics of Taiwan designed and manufactured the first portable digital pH meter. This meter was sold under Cole-Parmer's label.

Building a pH meter

A basic pH meter essentially measures the potential difference between two electrodes and displays the result calibrated in pH; the electronic circuit is very simple and easily built with a few cheap standard electronic components, plus the specialized pH probe.

1. SCOPE

- This test method is the procedure for determining the pH of water or soil samples by use of a pH meter. However, if the pH of any industrial by-product material (e.g.: cinders, flyash, etc.) is required, the procedure under 3.B. will be followed.

2. APPARATUS AND MATERIALS

- A 0.1 pt. (50 mL), wide-mouth glass beaker with a watch glass for cover. If lightweight material is to be tested, it may be necessary to increase beaker size up to a maximum of 0.5 pt. (250 mL).
- A pH meter, suitable for laboratory or field analysis, with either one or two electrodes.
- Standard buffer solutions of known pH values - standards to be used are pH of 4.0, 7.0, and 10.0.
- Distilled water
- A teaspoon or small scoop
- A thermometer capable of reading 77±18°F (25±10°C) to the nearest 0.1°C.

- A ¼ in. (6.3 mm) sieve conforming to the requirements of AASHTO Designation M-92-91 and a pan.
- A glass stirring rod

- A scale, minimum capacity of 1.1 lb. (500 g). It shall be accurate to 0.1% and be readable to 0.1 g.

3. ***PROCEDURE***

A. Water pH Determination

- Stir the water sample vigorously using a clean glass stirring rod.
- Pour a 40 mL ± 5 mL sample into the glass beaker using the watch glass for a cover.
- Let the sample stand for a minimum of one hour to allow the temperature to stabilize, stirring it occasionally while waiting. Measure the temperature of the sample and adjust the temperature controller of the pH meter to that of the sample temperature. This adjustment should be done just prior to testing.
- On meters with an automatic temperature control, follow the manufacturer's instructions.
- Standardize the pH meter by means of the standard solutions provided. Temperature and adjustments must be performed as stated.
- Immerse the electrode(s) of the pH meter into the water sample and turn the beaker slightly to obtain good contact between the water and the electrode(s).
- The electrode(s) require immersion 30 seconds or longer in the sample before reading to allow the meter to stabilize. If the meter has an auto read system, it will automatically signal when stabilized.
- Read and record the pH value to the nearest tenth of a whole number. If the pH meter reads to the hundredth place, a round off rule will apply as follows: If the hundredth place digit is less than 5, leave the tenth place digit as is. If it is greater than 5, round the tenth place digit up one unit. If the hundredth place digit equals 5, round the tenth place digit to the nearest even number.
- Rinse the electrode(s) well with distilled water, then dab lightly with tissues to remove any film formed on the electrode(s). Caution: Do not wipe the electrodes, as this may result in polarization of the electrode and consequent slow response.

B. SOIL pH DETERMINATION

The material must be separated on the ¼ in. (6.3 mm) sieve. Only the minus ¼ in. (6.3 mm) material is to be used for testing.

- Weigh and place 30±0.1 g of soil into the glass beaker.
- Add 30±0.1 g of distilled water to the soil sample. Stir to obtain soil slurry and then cover with watch glass.
- The sample must stand for a minimum of one hour, stirring every 10 to 15 minutes. This is to allow the pH of the soil slurry to stabilize.
- After one hour, the temperature of the sample should be stabilized. Measure the temperature of the sample and adjust the temperature controller of the pH meter to that of the sample temperature. This adjustment should be done just prior to testing. On meters with an automatic temperature control, follow the manufacturer's instructions.
- Standardize the pH meter by means of the standard solutions provided. Temperature and adjustments must be performed as stated.
- Immediately before immersing the electrode(s) into the sample, stir the sample well with a glass rod. Place the electrode(s) into the soil slurry solution and gently turn beaker to make good contact between the solution and the electrode(s). DO NOT place electrode(s) into the soil; only into the soil slurry solution.
- The electrode(s) require immersion 30 seconds or longer in the sample before reading to allow the meter to stabilize. If the meter has an auto read system, it will automatically signal when stabilized.
- Read and record the pH value to the nearest tenth of a whole number. If the pH meter reads to the hundredth place, a round off rule will apply as follows: If the hundredth place digit is less than 5, leave the tenth place digit as is. If it is greater than 5, round the tenth place digit up one unit. If the hundredth place digit equals 5, round the tenth place digit to the nearest even number.

- Rinse the electrode(s) well with distilled water, then dab lightly with tissues to remove any film formed on the electrode(s). Caution: Do not wipe the electrodes as this may result in polarization of the electrode(s) and consequent slow response.

NOTE 1 - To standardize the pH meter, use the 7.0 pH buffer standard solutions plus the other standard solution which is nearest the estimated pH value of the sample to be tested. If the manufacturer's instructions indicate a method other than that noted above, then those instructions must be followed.

NOTE 2 - When immersing electrode(s) into the glass beaker, care should be taken not to hit the bottom or side, causing damage to electrode(s).

NOTE 3 - If polarization does occur, as indicated by a slow response, rinse the electrode(s) and dab lightly again.

4. *PRECAUTIONS*

- Periodically check for damage to electrode(s).
- Electrode tips should be kept moist during storage. Follow manufacturer's instructions.

EXPERIMENT NO-02

***OBJECTIVE*:** To demonstrate the functioning of Spectrophotometer.

INTRODUCTION

In addition to the light microscope, the spectrophotometer is usually found in the biological laboratory. This instrument allows the investigator to identify compounds of biological interest and quantify them. Identification is determined by producing an absorption spectrum of a compound. Quantifying the amount of the material present, usually in solution, is done:

(a) directly if the substance is a strong absorber of a measurable wavelength (□) of light; (b) indirectly by chemically modifying the compound so that it is a strong absorber at a measurable wavelength of light; or

(c) Indirectly by stoichiometrically coupling its reaction to the formation of another light absorbing compound.

MATERIALS

2,6-dichlorophenol indophenols, graduated cylinders, flasks, stir plates, stir bars, micropipettes, spectrophotometer tubes, rulers, spectrophotometer.

A BEER-LAMBERT LAW

Spectrophotometry, or spectrophotometric analysis, refers to the quantitative determination of the radiant energy ratio of incident to transmitted light beams at a given wavelength.

Spectrophotometers are instruments that allow for the determination of this ratio. Spectrophotometers are designed to make these measurements over a given range of wavelengths of the electromagnetic spectrum.

Usually, in the biological laboratory a spectrophotometer will allow for measurements in the UV and visible wavelengths.

If $\mathbf{P_i}$ is the intensity of the incident beam of light and $\mathbf{P_t}$ is the intensity of the transmitted light, then, by definition, the ratio P_t/P_i equals the **transmittance**, **T**, and log P_i/P_t equals the **absorbance, A**. Thus:

$A = -\log T$ . . . . (1)

A fundamental law of spectrophotometry is the Beer Lambert law, or simply Beer's Law that states that the amount of radiant energy absorbed (log P_i/P_t) by a compound in solution is proportional to its concentration and the length of the path that the light beam passes through the solution:

$\log P_i / P_t = A = acl$ . . . . (2)

Where:

c = concentration

l = path length

a = proportionality constant called **absorptivity**.

The absorptivity at a given wavelength is dependent upon the chemical structure of the compound, which determines the probability that the wavelength of light will be absorbed. The absorptivity constant is sometimes referred to as the **extinction coefficient.**

If the concentration is expressed in molarity, then **a** is called molar absorptivity or **molar extinction coefficient (e)**. Remember that **a** or **e** is dependent upon the wavelength. Conventionally, the path length used is 1 cm, and thus, the units for **e** are cm^{-1} M^{-1}. Note that **A**, therefore, is a unit less value, which it should be, based on its definition above (i.e., a ratio).

A plot of the absorbance of a substance in solution at a given wavelength, as a function of the molar amount of the substance present, is called a calibration plot or **standard curve.** If you examine equation #2, you will see that the slope of a standard curve is **e**. Having determined **e**, and knowing l (the path length is 1 cm), the concentration of the substance in any other similar sample can be determined using equation #2, Beer's Law.

An **absorbance spectrum** of a given compound is a graphical representation of the absorbance at each wavelength over a given wavelength range (usually UV - visible), that is, **A** vs. 8. Chemically different compounds have their own distinctive absorbance spectrum. You will be determining an absorbance spectrum for the electron acceptor 2, 6-dichlorophenol inophenol (DCPIP) and will also plot absorbance spectra in a future lab exercise on plant pigments.

B. The Instrument

There are four essential parts to a spectrophotometer:

1. Light source
2. Monochromator or filter

3. Sample cell with holder

4. Detector

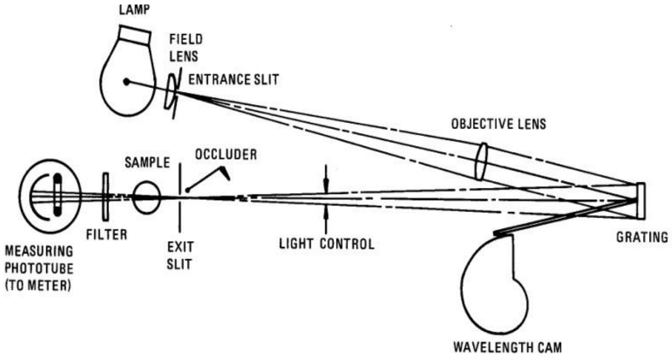


Figure 11.3.Essential parts of the spectrophotometer

The detector converts the radiant energy of the transmitted light into an electrical signal, the strength of which is proportional to the intensity of light transmitted (actually the number of photons striking its surface). A phototube is often used for this purpose. The electronics of the spectrophotometer convert this signal into a reading of transmittance or absorbance.

- Light sources do not emit the same light intensity over the entire spectrum; indeed, each source has its own inherent **emission spectrum.**
- Therefore, when determining an absorbance spectrum, it is important to adjust for the difference in light intensity at each wavelength of incident light.
- This can be done manually (using a “*ZERO CONTROL*” knob) or automatically, depending on the sophistication of the instrument.

C. Directions for Spectronic 20 Genesys ™ Spectrophotometer

1. Choose **A** from the **A/T/C** button to select for absorbance mode.

2. Press nm to select for correct wavelength.

3. Insert your blank into the cell holder and close the sample door.

4. Press the **0 ABS/100 %T** button to zero the absorbance.

5. Remove blank and insert your sample into the cell holder. Close door, and read absorbance from the LED display.

D. Preparing a standard curve and calculating the molar extinction coefficient

1. Prepare 100 ml of 1 mM 2, 6-dichlorophenol indophenol (DCPIP) in distilled water (DW) using the glassware provided. Make sure all the DCPIP goes into solution. *(The molecular weight of DCPIP is 290.08. g mole^{-1})*

2. Dilute the solution 1:10 by adding 10 ml of 1 mM DCPIP to 90 ml of DW to form a 100 M stock solution of DCPIP.

3. Prepare a serial dilution series of different concentrations of DCPIP in the spectrophotometer tubes provided. Take care to label each tube with a marker pen to indicate the appropriate concentration. Remember to use a new pipette tip for each manipulation.

4. Check that the dilution curve is correct by eye. Tube #9 should be the deepest blue in color.

5. Adjust setting to 500 nm. Zero the spectrophotometer using tube #1.

6. Remove tube 1 and place tube 9 in the spectrophotometer and record the absorbance value.

7. Repeat steps 5 and 6 at 10-nm increments from 500-700 nm. Zero at each wavelength before taking a measurement. Record values in Table. Plot these data (absorbance: y-axis; : x-axis) on the graph paper provided and determine the absorbance maximum, the wavelength at which DCPIP reaches its highest absorbance.

8. Set the wavelength to the absorbance maximum and readjust the zero control at this wavelength.

9. Determine the absorbance of tubes #1-9 at this wavelength. Record values in Table and construct a standard curve on the graph paper provided.

10. Determine the molar extinction coefficient for the dye from your standard curve. Enter this value below, for it will be used in a later laboratory exercise.

Micromolar Extinction Coefficient DCPIP @ _______ nm = ________

Wavelength (nm)	Absorbance
500	
510	
520	
530	
540	
550	
560	
570	
580	
590	
600	
610	
620	
630	
640	
650	
660	
670	

680	
690	
700	

Table: 1 Absorbance of Tube 9 as a Function of Wavelength

Tube	**Absorbance (at $_{max}$)**	**ml of stock solution**	**ml distilled water**	**Final concentration of DCPIP (µM)**
1		0.0	5.0	0
2		0.05	4.95	1.0
3		0.25	4.75	5.0
4		0.5	4.50	10.0
5		1.0	4.0	20.0
6		1.5	3.5	30.0
7		2.5	2.5	50.0
8		4.0	1.0	80.0
9		5.0	0.0	100.0

Table 11.1 Absorbance at $_{max}$ as a Function of DCPIP Concentration

11. You are provided with an additional sample containing an unknown concentration of DCPIP. Using your standard curve from part 9, determine the concentration of your unknown.

Unknown ____________ Unknown absorbance _______________

Unknown Concentration ____________

12. knowing the molar extinction coefficient for DCPIP, Determine the concentration of your unknown Using Beer's law. Recall that A = acl

Calculated "unknown" concentration using Beer's law ______________

- In chemistry, spectrophotometry is the quantitative measurement of the reflection or transmission properties of a material as a function of wavelength.
- It is more specific than the general term electromagnetic spectroscopy in that spectrophotometry deals with visible light, near-ultraviolet, and near-infrared, but does not cover time-resolved spectroscopic techniques.

- Spectrophotometry uses photometers that can measure a light beam's intensity as a function of its color (wavelength) known as spectrophotometers.
- Important features of spectrophotometers are spectral bandwidth (the range of colors it can transmit through the test sample), the percentage of sample-transmission, the logarithmic range of sample-absorption, and sometimes a percentage of reflectance measurement.
- A spectrophotometer is commonly used for the measurement of transmittance or reflectance of solutions, transparent or opaque solids, such as polished glass, or gases.
- However they can also be designed to measure the diffusivity on any of the listed light ranges that usually cover around 200 nm - 2500 nm using different controls and calibrations.
- Within these ranges of light, calibrations are needed on the machine using standards that vary in type depending on the wavelength of the **photometric determination.**

An example of an experiment in which spectrophotometry is used is the determination of the equilibrium constant of a solution. A certain chemical reaction within a solution may occur in a forward and reverse direction where reactants form products and products break down into reactants. At some point, this chemical reaction will reach a point of balance called an equilibrium point. In order to determine the respective concentrations of reactants and products at this point, the light transmittance of the solution can be tested using spectrophotometry. The amount of light that passes through the solution is indicative of the concentration of certain chemicals that do not allow light to pass through.

- The use of spectrophotometers spans various scientific fields, such as physics, materials science, chemistry, biochemistry, and molecular biology.

- They are widely used in many industries including semiconductors, laser and optical manufacturing, printing and forensic examination, and as well in laboratories for the study of chemical substances.
- Ultimately, a spectrophotometer is able to determine, depending on the control or calibration, what substances are present in a target and exactly how much through calculations of observed wavelengths.

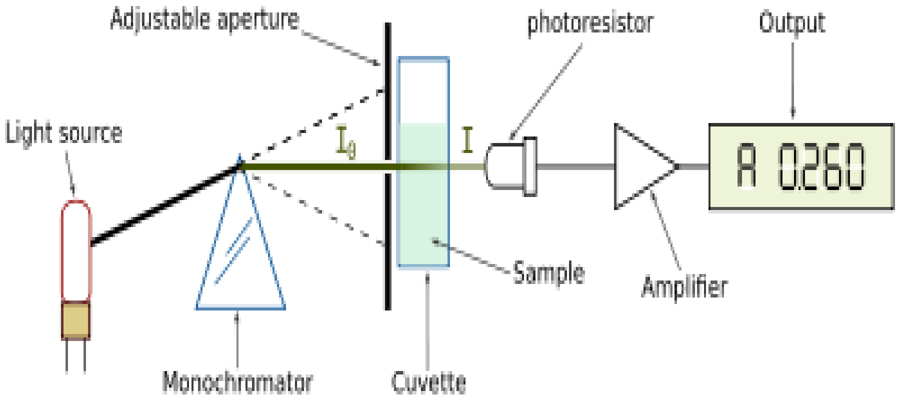


Figure 11.4 Single beam spectrophotometer

There are two major classes of devices: single beam and double beam.

- A double beam spectrophotometer compares the light intensity between two light paths, one path containing a reference sample and the other the test sample.
- A single-beam spectrophotometer measures the relative light intensity of the beam before and after a test sample is inserted.
- Although comparison measurements from double-beam instruments are easier and more stable, single-beam instruments can have a larger dynamic range and are optically simpler and more compact.

- Additionally, some specialized instruments, such as spectrophotometers built onto microscopes or telescopes, are single-beam instruments due to practicality.

Historically, spectrophotometers use a monochromator containing a diffraction grating to produce the analytical spectrum.

The grating can either be movable or fixed. If a single detector, such as a photomultiplier tube or photodiode is used, the grating can be scanned stepwise so that the detector can measure the light intensity at each wavelength (which will correspond to each "step"). Arrays of detectors, such as charge coupled devices (CCD) or photodiode arrays (PDA) can also be used.

In such systems, the grating is fixed and the intensity of each wavelength of light is measured by a different detector in the array. Additionally, most modern mid-infrared spectrophotometers use a Fourier transform technique to acquire the spectral information. The technique is called Fourier transform infrared spectroscopy.

When making transmission measurements, the spectrophotometer quantitatively compares the fraction of light that passes through a reference solution and a test solution, then electronically compares the intensities of the two signals and computes the percentage of transmission of the sample compared to the reference standard. For reflectance measurements, the spectrophotometer quantitatively compares the fraction of light that reflects from the reference and test samples.

Light from the source lamp is passed through a monochromator, which diffracts the light into a "rainbow" of wavelengths and outputs narrow bandwidths of this diffracted spectrum through a mechanical slit on the output side of the monochromator. These bandwidths are transmitted through the test sample.

Then the photon flux density (watts per metre squared usually) of the transmitted or reflected light is measured with a photodiode, charge coupled device or other light sensor. The transmittance or reflectance value for each wavelength of the test sample is then compared with the transmission or reflectance values from the reference sample. Most instruments will apply a logarithmic function to the linear transmittance ratio to calculate the 'absorbency' of the sample, a value which is proportional to the 'concentration' of the chemical being measured.

In short, the sequence of events in a modern spectrophotometer is as follows:

1. The light source is shone into a monochromator, diffracted into a rainbow, and split into two beams. It is then scanned through the sample and the reference solutions.
2. Fractions of the incident wavelengths are transmitted through, or reflected from, the sample and the reference.
3. The resultant light strikes the photodetector device, which compares the relative intensity of the two beams.
4. Electronic circuits convert the relative currents into linear transmission percentages and/or absorbance/concentration values.

Many older spectrophotometers must be calibrated by a procedure known as "zeroing", to balance the null current output of the two beams at the detector. The transmission of a reference substance is set as a baseline (datum) value, so the transmissions of all other substances are recorded relative to the initial "zeroed" substance. The spectrophotometer then converts the transmission ratio into 'absorbency', the concentration of specific components of the test sample relative to the initial substance.

APPLICATIONS IN BIOCHEMISTRY

Spectrophotometry is an important technique used in many biochemical experiments that involve DNA, RNA, and protein isolation, enzyme kinetics and biochemical analyses.

A brief explanation of the procedure of spectrophotometry includes comparing the absorbency of a blank sample that does not contain a colored compound to a sample that contains a colored compound.

The spectrophotometer is used to measure colored compounds in the visible region of light (between 350 nm and 800 nm), thus it can be used to find more information about the substance being studied.

In biochemical experiments, a chemical and/or physical property is chosen and the procedure that is used is specific to that property in order to derive more information about the sample, such as the quantity, purity, enzyme activity, etc.

Spectrophotometry is also a helpful procedure for protein purification and can also be used as a method to create optical assays of a compound. Because a spectrophotometer measures the wavelength of a compound through its color, a dye binding substance can be added so that it can undergo a color change and be measured.

Spectrophotometers have been developed and improved over decades and have been widely used among chemists. It is considered to be a highly accurate instrument that is also very sensitive and therefore extremely precise, especially in determining color change.

This method is also convenient for use in laboratory experiments because it is an inexpensive and relatively simple process.

UV-visible spectrophotometry

The most common spectrophotometers are used in the UV and visible regions of the spectrum, and some of these instruments also operate into the near-infrared region as well.

Visible region 400–700 nm spectrophotometry is used extensively in colorimetry science. It is a known fact that it operates best at the range of 0.2-0.8 O.D.

Ink manufacturers, printing companies, textiles vendors, and many more, need the data provided through colorimetry. They take readings in the region of every 5–20 nanometers along the visible region, and produce a spectral reflectance curve or a data stream for alternative presentations.

These curves can be used to test a new batch of colorant to check if it makes a match to specifications, e.g., ISO printing standards.

Traditional visible region spectrophotometers cannot detect if a colorant or the base material has fluorescence. This can make it difficult to manage color issues if for example one or more of the printing inks is fluorescent.

Where a colorant contains fluorescence, a bi-spectral fluorescent spectrophotometer is used. There are two major setups for visual spectrum spectrophotometers, d/8 (spherical) and 0/45.

The names are due to the geometry of the light source, observer and interior of the measurement chamber. Scientists use this instrument to measure the amount of compounds in a sample.

If the compound is more concentrated more light will be absorbed by the sample; within small ranges, the Beer-Lambert law holds and the absorbance between samples vary with concentration linearly.

In the case of printing measurements two alternative settings are commonly used- without/with UV filter to control better the effect of UV brighteners within the paper stock.

Samples are usually prepared in cuvettes; depending on the region of interest, they may be constructed of glass, plastic (visible spectrum region of interest), or quartz (Far UV spectrum region of interest).

APPLICATIONS

- Estimating dissolved organic carbon concentration
- Specific Ultraviolet Absorption for metric of aromaticity
- Bial's Test for concentration of pentoses

Infrared Spectrophotometry

Spectrophotometers designed for the infrared region are quite different because of the technical requirements of measurement in that region.

One major factor is the type of photo sensors that are available for different spectral regions, but infrared measurement is also challenging because virtually everything emits IR light as thermal radiation, especially at wavelengths beyond about 5 μm.

Another complication is that quite a few materials such as glass and plastic absorb infrared light, making it incompatible as an optical medium. Ideal optical materials are salts, which do not absorb strongly.

Samples for IR spectrophotometry may be smeared between two discs of potassium bromide or ground with potassium bromide and pressed into a pellet. Where aqueous solutions are to be measured, insoluble silver chloride is used to construct the cell.

Spectroradiometers

Spectroradiometers, which operate almost like the visible region spectrophotometers, are designed to measure the spectral density of illuminants.

Applications may include evaluation and categorization of lighting for sales by the manufacturer, or for the customers to confirm the lamp they decided to purchase is within their specifications.

COMPONENTS

1. The light source shines onto or through the sample.
2. The sample transmits or reflects light.
3. The detector detects how much light was reflected from or transmitted through the sample.
4. The detector then converts how much light the sample transmitted or reflected into a number.

EXPERIMENT NO - 03

OBJECTIVE: To demonstrate of use of bright-field microscopy

INTRODUCTION

Bright-field microscopy is the simplest of all the optical microscopy illumination techniques. Sample illumination is transmitted (i.e., illuminated from below and observed from above) white light and contrast in the sample is caused by absorbance of some of the transmitted light in dense areas of the sample.

Bright-field microscopy is the simplest of a range of techniques used for illumination of samples in light microscopes and its simplicity makes it a popular technique. The typical appearance of a bright-field microscopy image is a dark sample on a bright background, hence the name.

Light path

The light path of a bright-field microscope is extremely simple; no additional components are required beyond the normal light microscope setup. The light path therefore consists of:

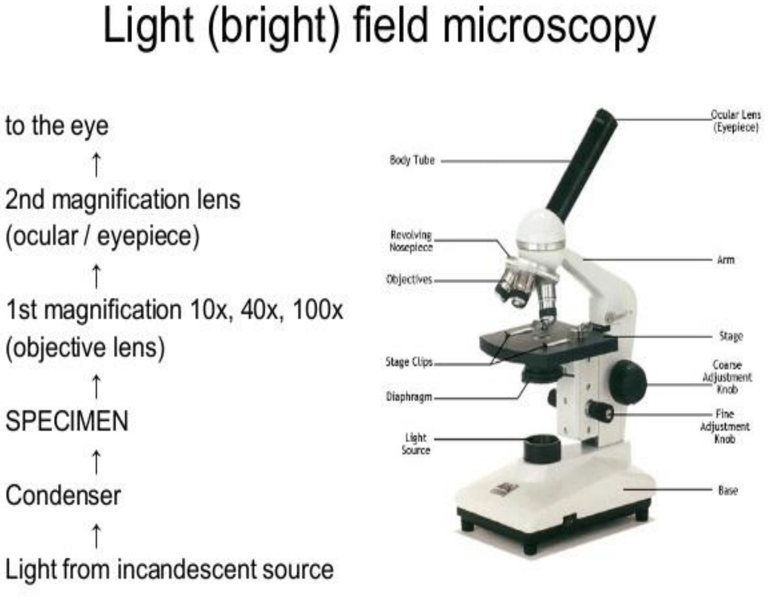
- A transillumination light source, commonly a halogen lamp in the microscope stand;
- A condenser lens which focuses light from the light source onto the sample; and
- Objective lens which collects light from the sample and magnifies the image.
- Oculars and/or a camera to view the sample image

Bright field microscopy may use critical or Köhler illumination to illuminate the sample.

Performance

Bright-field microscopy typically has low contrast with most biological samples as few absorb light to a great extent. Staining is often required to increase contrast, which prevents use on live cells in many situations. Bright field illumination is useful for samples which have an intrinsic colour, for example chloroplasts in plant cells.Comparison of transillumination techniques used to generate contrast in a sample of tissue paper. 1.559 μm/pixel. Bright field illumination, sample contrast comes from absorbance of light in the sample.

Cross-polarized light illumination, sample contrast comes from the rotation of polarized light through the sample. Dark field illumination, sample contrast comes from light scattered by the sample.



Light (bright) field microscopy

to the eye

↑

2nd magnification lens
(ocular / eyepiece)

↑

1st magnification 10x, 40x, 100x
(objective lens)

↑

SPECIMEN

↑

Condenser

↑

Light from incandescent source

Figure 11.5 Light field microscopy

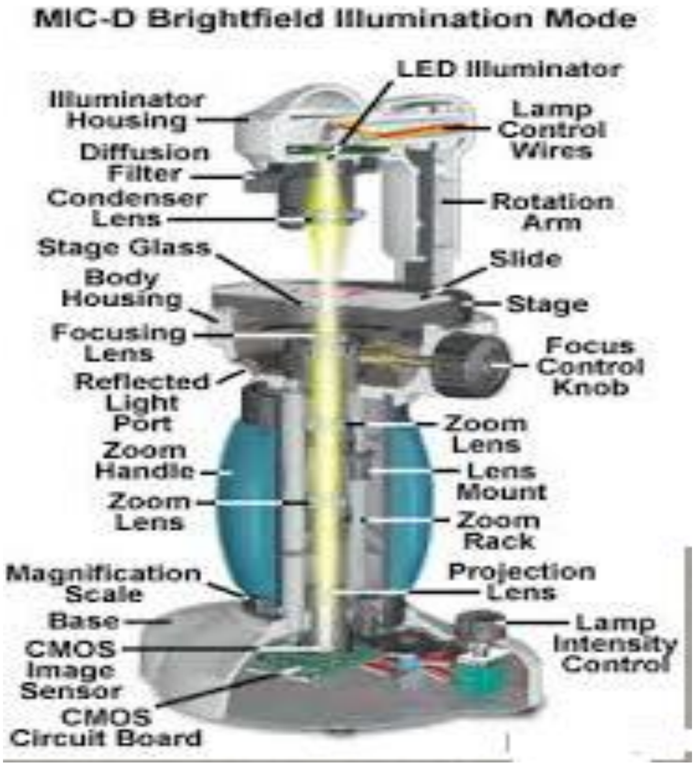


Figure 11.6 Bright field Illumination Mode

Phase contrast illumination, sample contrast comes from interference of different path lengths of light through the sample. Bright-field microscopy is a standard light microscopy technique, and therefore magnification is limited by the resolving power possible with the wavelength of visible light.

ADVANTAGES

- Simplicity of setup with only basic equipment required.

Limitations

- Very low contrast of most biological samples.
- Low apparent optical resolution due to the blur of out of focus material.

- Samples that are naturally colorless and transparent cannot be seen well, e.g. many types of mammalian cells.
- These samples often have to be stained before viewing. Samples that do have their own colour can be seen without preparation, e.g. the observation of cytoplasmic streaming in Chara cells.

Enhancements

- Reducing or increasing the amount of the light source via the iris diaphragm.
- Use of an oil immersion objective lens and special immersion oil placed on a glass cover over the specimen. Immersion oil has the same refraction as glass and improves the resolution of the observed specimen.
- Use of sample staining methods for use in microbiology, such as simple stains (Methylene blue, Safranin, Crystal violet) and differential stains (Negative stains, flagellar stains, endospore stains).
- Use of a colored (usually blue) or polarizing filter on the light source to highlight features not visible under white light. The use of filters is especially useful with mineral samples

EXPERIMENT NO – 3.1

OBJECTIVE: To demonstrate of use of Phase contrast microscopy

INTRODUCTION

Phase contrast microscopy, first described in 1934 by Dutch physicist Frits Zernike, is a contrast-enhancing optical technique that can be utilized to produce high-contrast images of transparent specimens, such as living cells (usually in culture), microorganisms, thin tissue slices, lithographic patterns, fibers, latex dispersions, glass fragments, and subcellular particles (including nuclei and other organelles).

In effect, the phase contrast technique employs an optical mechanism to translate minute variations in phase into corresponding changes in amplitude, which can be visualized as differences in image contrast.

One of the major advantages of phase contrast microscopy is that living cells can be examined in their natural state without previously being killed, fixed, and stained.

As a result, the dynamics of ongoing biological processes can be observed and recorded in high contrast with sharp clarity of minute specimen detail.

In **f**igure **1** is a cut-away diagram of a modern upright phase contrast microscope, including a schematic illustration of the phase contrast optical train.

Partially coherent illumination produced by the tungsten-halogen lamp is directed through a collector lens and focused on a specialized annulus (labeled **condenser annulus**) positioned in the sub stage condenser front focal plane.

Wave fronts passing through the annulus illuminate the specimen and either passes through undeviated or is diffracted and retarded in phase by structures and phase gradients present in the specimen.

Undeviated and diffracted light collected by the objective is segregated at the rear focal plane by a **phase plate** and focused at the intermediate image plane to form the final phase contrast image observed in the eyepieces.

Prior to the invention of phase contrast techniques, transmitted bright field illumination was one of the most commonly utilized observation modes in optical microscopy, especially for fixed, stained specimens or other types of samples having high natural absorption of visible light.

Collectively, specimens readily imaged with bright field illumination are termed **amplitude objects** (or specimens) because the amplitude or intensity of the illuminating wave fronts is reduced when light passes through the specimen.

The addition of phase contrast optical accessories to a standard bright field microscope can be employed as a technique to render a contrast-enhancing effect in transparent specimens that is reminiscent of optical staining.

Light waves that are diffracted and shifted in phase by the specimen (termed a **phase object**) can be transformed by phase contrast into amplitude differences that are observable in the eyepieces.

Large, extended specimens are also easily visualized with phase contrast optics due to diffraction and scattering phenomena that occur at the edges of these objects.

The performance of modern phase contrast microscopes is so refined that it enables specimens containing very small internal structures, or even just a few protein molecules, to be detected when the technology is coupled to electronic enhancement and post-acquisition image processing.

In figure is a comparison of living cells in culture imaged in both bright field and phase contrast illumination.

The cells are human glial brain tissue grown in monolayer culture bathed with a nutrient medium containing amino acids, vitamins, mineral salts, and fetal calf serum.

In bright field illumination, the cells appear semi-transparent with only highly refractive regions, such as the membrane, nucleus, and unattached cells (rounded or spherical), being visible.

When observed using phase contrast optical accessories, the same field of view reveals significantly more structural detail. Cellular attachments become discernable, as does much of the internal structure. In addition, the contrast range is dramatically improved.

Interaction of Light Waves with Phase Specimens

An incident wave front present in an illuminating beam of light becomes divided into two components upon passing through a phase specimen.

The primary component is an undeviated (or undiffracted; **zeroth-order**) planar wave front, commonly referred to as the **surround** (**S**) wave, which passes through and around the specimen, but does not interact with it.

In addition, a deviated or **diffracted** spherical wave front (**D**-wave) is also produced, which becomes scattered over a wide arc (in many directions) that passes through the full aperture of the objective.

After leaving the specimen plane, surround and diffracted light waves enter the objective front lens element and are subsequently focused at the intermediate image plane where they combine through

interference to produce a resultant **particle** wave (often referred to as a **P**-wave). The mathematical relationship between the various light waves generated in phase contrast microscopy can be described simply as:

Formula 1 - Relationship between Various Light Waves Generated in Phase Contrast Microscopy P = S + D

Detection of the specimen image depends on the relative intensity differences, and therefore on the amplitudes, of the particle and surround (**P** and **S**) waves.

If the amplitudes of the particle and surround waves are significantly different in the intermediate image plane, then the specimen acquires a considerable amount of contrast and is easily visualized in the microscope eyepieces.

Otherwise, the specimen remains transparent and appears as it would under ordinary bright field conditions (in the absence of phase contrast or other contrast-enhancing techniques).

In terms of optical path variations between the specimen and its surrounding medium, the portion of the incident light wave front that traverses the specimen (**D**-wave), but does not pass through the surrounding medium (**S**-wave), is slightly retarded.

For arguments in phase contrast microscopy, the role of the specimen in altering the optical path length (in effect, the relative phase shift) of waves passing through is of paramount importance.

In classical optics, the optical path length (**OPL**) through an object or space is the product of the refractive index (**n**) and the thickness (**t**) of the object or intervening medium as described by the relationship:

Formula 2 - Optical Path Length

Optical Path Length (OPL) = $n \times t$

When light passes from one medium into another, the velocity is altered proportionally to the refractive index differences between the two media.

Thus, when a coherent light wave emitted by the focused microscope filament passes through a phase specimen having a specific thickness (**t**) and refractive index (**n**), the wave is either increased or decreased in velocity.

If the refractive index of the specimen is greater than that of the surrounding medium, the wave is reduced in velocity while passing through the specimen and is subsequently retarded in relative phase when it emerges from the specimen.

In contrast, when the refractive index of the surrounding medium exceeds that of the specimen, the wave is advanced in phase upon exiting the specimen. The difference in location of an emergent wave front between the specimen and surrounding medium is termed the **phase shift** (δ) and is defined in radians as:

Formula 3 - Phase Shift $\delta = 2\pi\Delta/\lambda$

In the equation above, the term **D** is referred to as the **optical path difference**, which is similar to the optical path length:

Formula 4 - Optical Path Difference

Optical Path Difference (OPD) = $\Delta = (n_2 - n_1) \times t$.

Where **n(2)** is the refractive index of the specimen and **n(1)** is the refractive index of the surrounding medium.

The optical path difference results from the product of two terms: the thickness of the specimen, and its difference in refractive index with the surrounding medium.

In many cases, the optical path difference can be quite large even though the thickness of the specimen is small.

On the other hand, when the refractive index of the specimen equals that of the surrounding medium, the optical path difference is zero regardless of whether the specimen thickness is large or small.

Wave Interactions in Phase Contrast Microscopy

Phase relationships between the surround, diffracted, and particle (**S**, **D**, and **P**) waves in the region of the specimen at the image plane for bright field microscopy (in the absence of phase contrast optical accessories) are presented in figure.

The surround and particle waves, whose relative amplitudes determine the amount of specimen contrast, are illustrated as red and green lines (respectively).

The wave produced by diffraction from the specimen, which is never directly observed, is depicted as a blue wave of lower amplitude.

The surround and diffracted waves recombine through interference to generate the resultant particle wave in the image plane of the microscope.

The amplitude of each wave illustrated in figure represents the sum of the electric vectors of the individual component waves.

Interpretation of Phase Contrast Images

Images produced by phase contrast microscopy are relatively simple to interpret when the specimen is thin and distributed evenly on the substrate (as is the case with living cells grown in monolayer tissue culture).

When thin specimens are examined using positive phase contrast optics, which is the traditional form produced by most manufacturers, they appear darker than the surrounding medium when the refractive index of the specimen exceeds that of the medium.

Phase contrast optics differentially enhance the contrast near the edges surrounding extended specimens, such as the boundary between a cellular membrane and the bathing nutrient medium, and produce overall high-contrast images that can be roughly interpreted as density maps.

Because the amplitude and intensity of a specimen image in phase contrast is related to refractive index and optical path length, image density can be utilized as a gauge for approximating relationships between various structures.

In effect, a series of internal cellular organelles having increasing density, such as vacuoles, cytoplasm, the inter phase nucleus, and the nucleolus (or mitotic chromosomes), are typically visualized as progressively darker objects relative to a fixed reference, such as the background.

It should also be noted that numerous optical artifacts are present in all phase contrast images, and large extended specimens often present significant fluctuations in contrast and image intensity.

Symmetry can also be an important factor in determining how both large and small specimens appear in the phase contrast microscope.

Sensible interpretation of phase contrast images requires careful scrutiny and examination to ensure that artifacts are not incorrectly assigned to important structural features.

For example, some internal cellular organelles and components often have a lower refractive index than that of the surrounding cytoplasm, while others have a higher refractive index.

Because of the varying refractive indices exhibited by these numerous intracellular structures, the interior of living cells, when viewed in a positive phase contrast microscope, can reveal an array of intensities ranging from very bright to extremely dark.

For example, pinocytotic vesicles, lipid droplets, and air vacuoles present in plants and single cell protozoan's have a lower refractive index than the cytoplasm, and thus appear brighter than other components. In contrast, as discussed above, organelles that have high refractive indices (nuclei, ribosome, mitochondria, and the nucleolus) appear dark in the microscope. If the phase retardation introduced by the specimen is large enough (a phase shift of the diffracted wave by approximately a half-wavelength), interference between the diffracted waves and the surround waves becomes constructive, rendering these specimens brighter than the surrounding background.

In order to avoid confusion regarding bright and dark contrast in phase contrast images, the optical path differences occurring within the specimen preparation should be carefully considered.

As discussed above, the optical path difference is derived from the product of the refractive index and the specimen (object) thickness, and is related to the relative phase shift between the specimen and background (diffracted and surround) waves.

It is impossible to distinguish between high and low refractive index components in a phase contrast image without information pertaining to the relative thickness of the components.

For example, a small specimen having a high refractive index can display an identical optical path difference to a larger specimen having a lower refractive index. The two specimens will have approximately the same intensity when viewed through a phase contrast optical system.

In many biological experiments, conditions that produce a shrinking or swelling of cells or organelles can result in significant contrast variations. The external medium can also be replaced with another having either a higher or lower refractive index to generate changes in specimen image contrast.

In fact, the effect on image contrast of refractive index variations in the surrounding medium forms the basis.

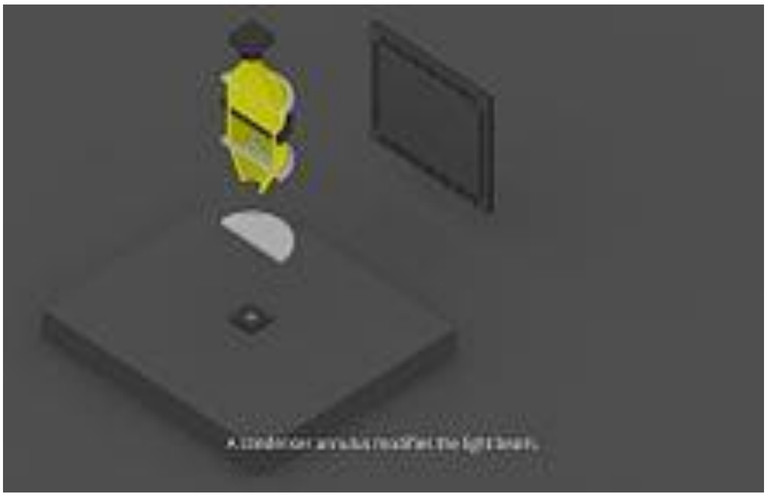


Fig 11.8 *Dark field and phase contrast microcopies operating principle*

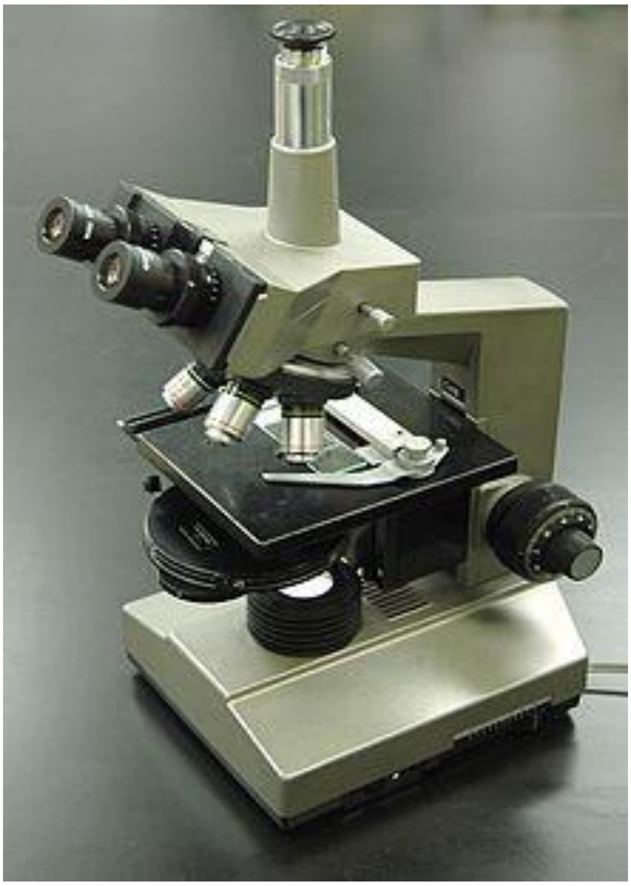


Fig.11.9 A phase-contrast microscope

WORKING PRINCIPLE

The basic principle to making phase changes visible in phase-contrast microscopy is to separate the illuminating (background) light from the specimen-scattered light (which makes up the foreground details) and to manipulate these differently.

The ring-shaped illuminating light (green) that passes the condenser annulus is focused on the specimen by the condenser. Some of the illuminating light is scattered by the specimen (yellow).

The remaining light is unaffected by the specimen and forms the background light (red). When observing an unstained biological specimen, the scattered light is weak and typically phase-shifted by −90° (due to both the typical thickness of specimens and the refractive index difference between biological tissue and the surrounding medium) relative to the background light.

This leads to the foreground (blue vector) and background (red vector) having nearly the same intensity, resulting in low image contrast.

In a phase-contrast microscope, image contrast is increased in two ways: by generating constructive interference between scattered and background light rays in regions of the field of view that contain the specimen, and by reducing the amount of background light that reaches the image plane.

First, the background light is phase-shifted by −90° by passing it through a phase-shift ring, which eliminates the phase difference between the background and the scattered light rays.

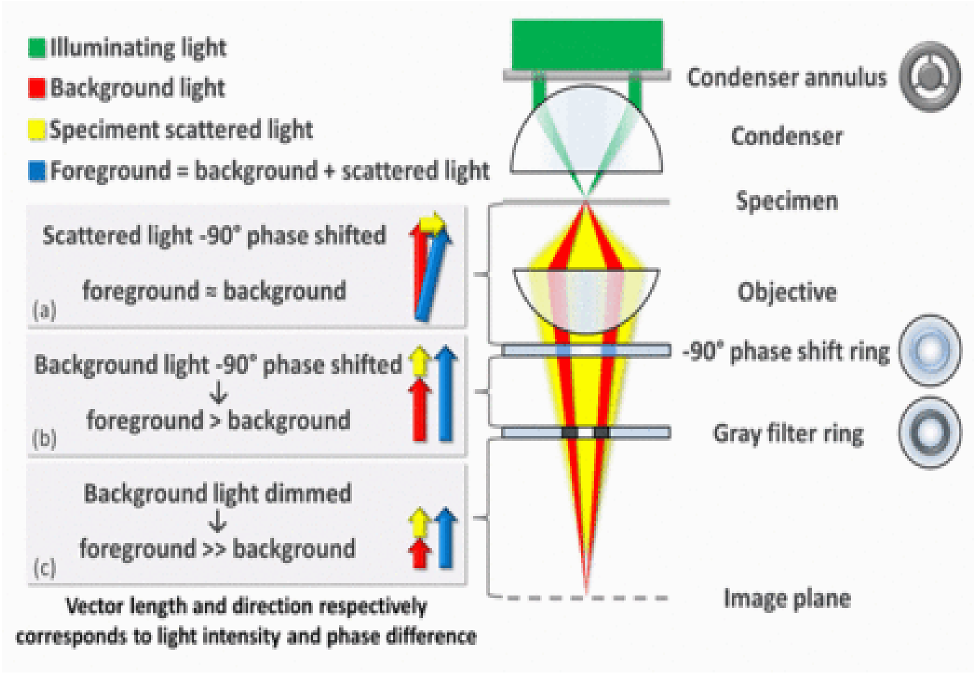


Fig 6.10 Operating principle dark field and phase contrast microcopies

EXPERIMENT NO- 3.2

OBJECTIVE: To demonstrate of use of dark field microscopy

INTRODUCTION

Have you ever heard of a dark field microscope? While such a name may sound like a sci-fi gadget used to measure black holes, in reality it's just a handy tool used to view certain types of translucent samples.

The average microscope user may not know about the concept of dark field microscopy, yet it can shed new light on the old way of viewing specimens.

Most people who have survived a biology class know what a light field microscope is. This type of scope uses bright field illumination, meaning it floods the specimen with white light from the condenser without any interference.

Thus the specimen shows up as a dark image on a light background (or white field if you will).

This type of unit works best with specimens that have natural color pigments. The samples need to be thick enough to absorb the incoming light; so staining is usually paired with this type of microscope.

Yet what if the specimen is light colored or translucent, like the plankton on the right? It certainly won't stand out against a strong white background. Additionally, some specimens are just too thin.

They cannot absorb any of the light that passes through them, so they appear invisible to the user. This is where the concept of dark field illumination comes in!

Rather than using direct light from the condenser, one uses an opaque disk to block the light into just a few scattered beams.

Now the background is dark, and the sample reflects the light of the beams only. This results in a light colored specimen against a dark background (dark field), perfect for viewing clear or translucent details.

On a grand scale, the same thing happens every day when you look up at the sky. Do the stars disappear when it's light out? Of course not! They're still there, their brilliance blotted out by the mid-day sun.

If you're still having a hard time visualizing this concept, think of a dusty room with the light on and the door open. You may feel the dust affecting your breathing, but you probably won't see it flying through the air.

Now turn off the light and close the door to just a sliver, while leaving the light on in the adjacent room. If you look at that sliver of light coming through the door, you'll see all sorts of dust motes suspended in it. You're employing a similar principle when you use dark field illumination!

Use

Dark field microscopes are used in a number of different ways to view a variety of specimens that are hard to see in a light field unit.

Live bacteria, for example, are best viewed with this type of microscope, as these organisms are very transparent when unstained.

There are multitudes of other ways to use dark field illumination, often when the specimen is clear or translucent. Some examples:

- Living or lightly stained transparent specimens
- Single-celled organisms
- Live blood samples
- Aquatic environment samples (from seawater to pond water)
- Living bacteria
- Hay or soil samples
- Pollen samples
- Certain molecules such as caffeine crystals (right)

Dark field microscopy makes many invisible specimens appear visible. Most of the time the specimens invisible to bright field illumination are living, so you can see how important it is to bring them into view!

ADVANTAGES AND DISADVANTAGES

No one system is perfect and dark field microscopy may or may not appeal to you depending on your needs. Some advantages of using a dark field microscope are:

- Extremely simple to use
- Inexpensive to set up (instructions on how to make your own dark field microscope are below)
- Very effective in showing the details of live and unstained samples

Some of the disadvantages are:

- Limited colors (certain colors will appear, but they're less accurate and most images will be just black and white)
- Images can be difficult to interpret to those unfamiliar with dark field microscopy
- Although surface details can be very apparent, the internal details of a specimen often don't stand out as much with a dark field setup.

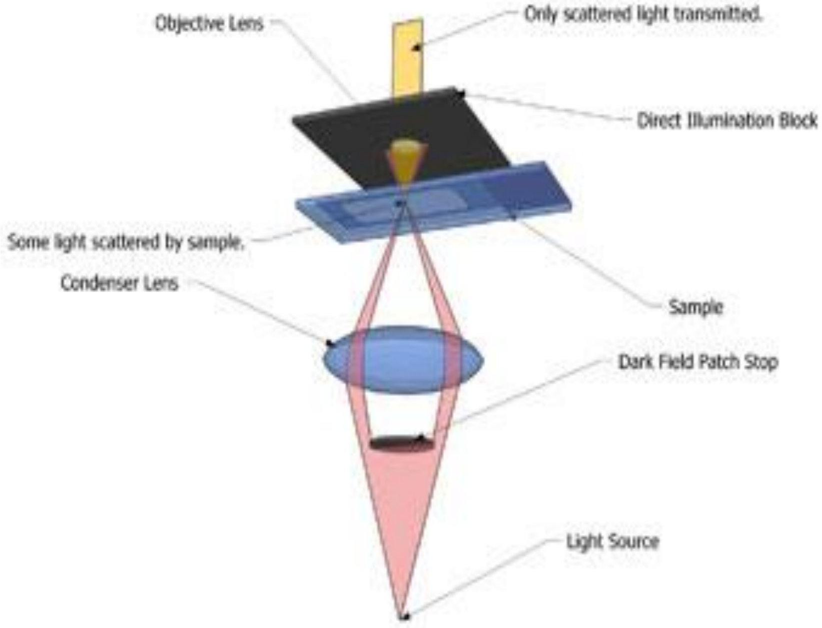

Objective Lens
Only scattered light transmitted.
Direct Illumination Block
Some light scattered by sample.
Condenser Lens
Sample
Dark Field Patch Stop
Light Source

Fig 11.11Contrasting examples of dark field illumination

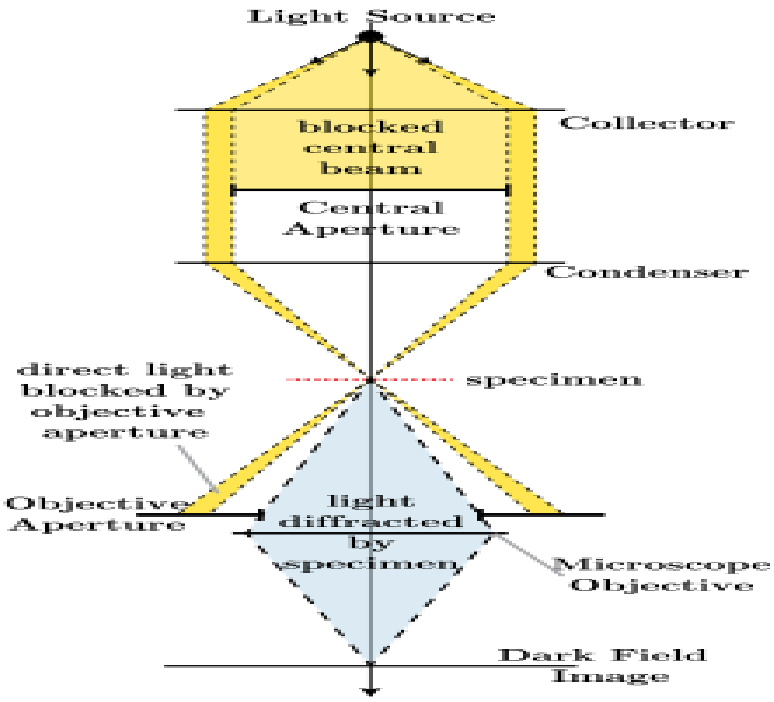


Fig 11.12 Diagram illustrating the light path through a dark field microscope

Digital dark field analysis

This a mathematical technique intermediate between direct and reciprocal (Fourier-transform) space for exploring images with well-defined periodicities, like electron microscope lattice-fringe images.

As with analog dark field imaging in a transmission electron microscope,it allows one to "light up" those objects in the field of view where periodicities of interest reside.

Unlike analog dark field imaging it may also allow one to map the Fourier-phase of periodicities, and hence phase-gradients which provide quantitative information on vector lattice-strain.

Experiment No- 3.3

OBJECTIVE: To demonstrate of use of Fluorescent microscopy

INTRODUCTION

The absorption and subsequent re-radiation of light by organic and inorganic specimens is typically the result of well-established physical phenomena described as being either **fluorescence** or **phosphorescence**.

The emission of light through the fluorescence process is nearly simultaneous with the absorption of the excitation light due to a relatively short time delay between photon absorption and emission, ranging usually less than a microsecond in duration.

When emission persists longer after the excitation light has been extinguished, the phenomenon is referred to as phosphorescence.

A fluorescence microscope is much the same as a conventional light microscope with added features to enhance its capabilities.

- The conventional microscope uses visible light (400-700 nanometers) to illuminate and produce a magnified image of a sample.
- A fluorescence microscope, on the other hand, uses a much higher intensity light source which excites a fluorescent species in a sample of interest. This fluorescent species in turn emits a lower energy light of a longer wavelength that produces the magnified image instead of the original light source.

Fluorescent microscopy is often used to image specific features of small specimens such as microbes. It is also used to visually enhance 3-D features at small scales.

This can be accomplished by attaching fluorescent tags to anti-bodies that in turn attach to targeted features, or by staining in a less specific manner.

When the reflected light and background fluorescence is filtered in this type of microscopy the targeted parts of a given sample can be imaged.

This gives an investigator the ability to visualize desired organelles or unique surface features of a sample of interest. Confocal fluorescent microscopy is most often used to accentuate the 3-D nature of samples.

This is achieved by using powerful light sources, such as lasers, that can be focused to a pinpoint. This focusing is done repeatedly throughout one level of a specimen after another.

Most often an image reconstruction program pieces the multi level image data together into a 3-D reconstruction of the targeted sample.

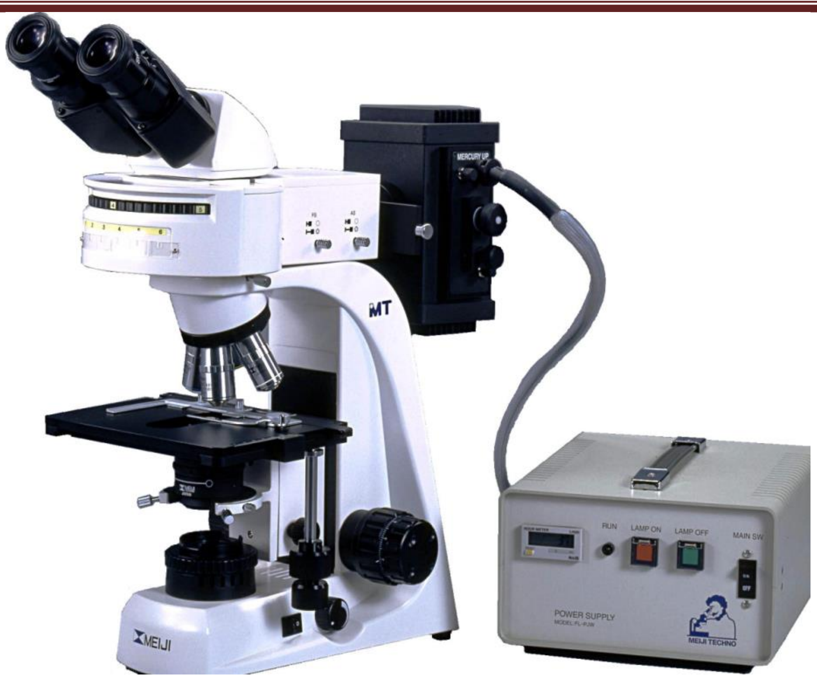


Fig 11.13Fluorescent Microscopy

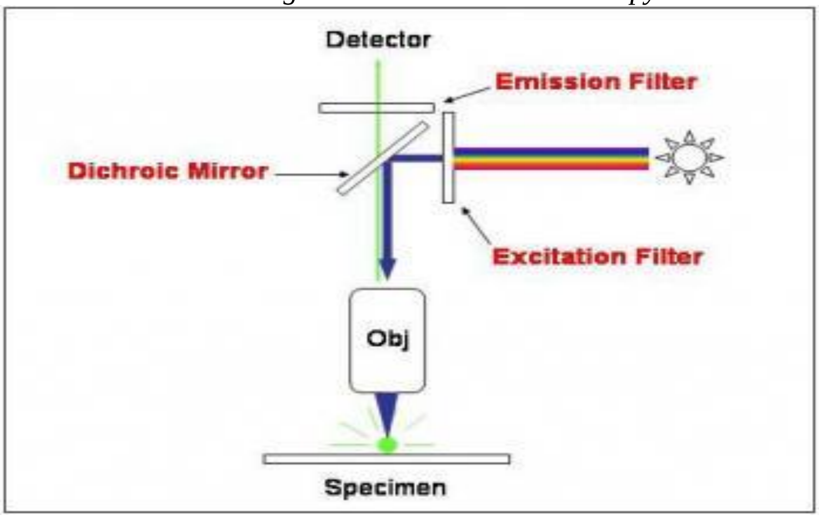


Figure 11.14 showing the filters and mirror in a fluorescent microscope

How does Fluorescent Microscopy Work

In most cases the sample of interest is labeled with a fluorescent substance known as a fluorophore and then illuminated through the lens with the higher energy source.

The illumination light is absorbed by the fluorophores (now attached to the sample) and causes them to emit a longer lower energy wavelength light.

This fluorescent light can be separated from the surrounding radiation with filters designed for that specific wavelength allowing the viewer to see only that which is fluorescing.

The basic task of the fluorescence microscope is to let excitation light radiate the specimen and then sort out the much weaker emitted light from the image.

First, the microscope has a filter that only lets through radiation with the specific wavelength that matches your fluorescing material.

The radiation collides with the atoms in your specimen and electrons are excited to a higher energy level.

When they relax to a lower level, they emit light. To become detectable (visible to the human eye) the fluorescence emitted from the sample is separated from the much brighter excitation light in a second filter.

This works because the emitted light is of lower energy and has a longer wavelength than the light that is used for illumination.

Most of the fluorescence microscopes used in biology today is epi-fluorescence microscopes, meaning that both the excitation and the observation of the fluorescence occur above the sample.

Most use a Xenon or Mercury arc-discharge lamp for the more intense light source.

APPLICATIONS

The refinement of epi-fluorescent microscopes and advent of more powerful focused light sources, such as lasers, has led to more technically advanced scopes such as the confocal laser scanning microscopes and total internal reflection fluorescence microscopes (TIRF).

CLSM's are invaluable tools for producing high resolution 3-D images of subsurfaces in specimens such as microbes.

Their advantage is that they are able to produce sharp images of thick samples at various depths by taking images point by point and reconstructing them with a computer rather than viewing whole images through an eyepiece.

These microscopes are often used for -

- Imaging structural components of small specimens, such as cells
- Conducting viability studies on cell populations (are they alive or dead?)
- Imaging the genetic material within a cell (DNA and RNA)
- Viewing specific cells within a larger population with techniques such as FISH

Fluorescence Light Sources

An unfortunate consequence of low emission levels in most fluorescence microscopy applications is that the number of photons that reach the eye or camera detector is also very low.

In most cases, the collection efficiency of optical microscopes is less than 30 percent and the concentration of many fluorophores in the optical path ranges in the micromolar or nanomolar regions.

In order to generate sufficient excitation light intensity to produce detectable emission, powerful compact light sources, such as high-energy short arc-discharge lamps, are necessary.

The most common lamps are mercury burners, ranging in wattage from 50 to 200 Watts, and the xenon burners that range from 75 to 150 Watts.

These light sources are usually powered by an external direct current supply, furnishing enough start-up power to ignite the burner through ionization of the gaseous vapor and to keep it burning with a minimum of flicker.

The microscope arc-discharge lamp external power supply is usually equipped with a timer to track the number of hours the burner has been in operation.

Arc lamps lose efficiency and are more likely to shatter if used beyond their rated lifetime (200-300 hours).

The mercury burners do not provide even intensity across the spectrum from ultraviolet to infrared, and much of the intensity of the lamp is expended in the near ultraviolet.

Prominent peaks of intensity occur at 313, 334, 365, 406, 435, 546, and 578 nanometers. At other wavelengths in the visible light region, the intensity is steady although not nearly so bright (but still useable in most applications).

In considering illumination efficiency, mere lamp wattage is not the prime consideration. Instead, the critical parameter is the mean luminance must be considered, taking into account the source brightness, arc geometry, and the angular spread of emission.

Table 1 - Luminous Density of Selected Light Sources

Lamp	**Current (Amperes)**	**Luminous Flux (Lumens)**	**Mean Luminous Density (cd/mm2)**	**Arc Size (H x W) (Millimeters)**
Mercury Arc (100 Watt)	5	2200	1700	0.25 x 0.25
Xenon Arc (75 Watt)	5.4	850	400	0.25 x 0.50
Xenon Arc (500 Watt)	30	9000	3500	0.30 x 0.30
Tungsten Halogen	8	2800	45	4.2 x 2.3

CONCLUSIONS

The modern fluorescence microscope combines the power of high performance optical components with computerized control of the instrument and digital image acquisition to achieve a level of sophistication that far exceeds that of simple observation by the human eye.

Microscopy now depends heavily on electronic imaging to rapidly acquire information at low light levels or at visually undetectable wavelengths.

These technical improvements are not mere window dressing, but are essential components of the light microscope as a system.

The era when optical microscopy was purely a descriptive instrument or an intellectual toy is past.

At present, optical image formation is only the first step toward data analysis. The microscope accomplishes this first step in conjunction with electronic detectors, image processors, and display devices that can be viewed as extensions of the imaging system.

Computerized control of focus, stage position, optical components, shutters, filters, and detectors is in widespread use and enables experimental manipulations that were not humanly possible with mechanical microscopes.

The increasing application of electro-optics in fluorescence microscopy has led to the development of optical tweezers capable of manipulating sub-cellular structures or particles, the imaging of single molecules, and a wide range of sophisticated spectroscopic applications.

EXPERIMENT NO- 3.4

OBJECTIVE: To demonstrate of use of confocal microscopy

INTRODUCTION

This microscopy facility provides users with the opportunity to prepare and image their own samples.

The EM/Confocal Specialist will provide technical support to new users, faculty, graduate students, and staff, for specimen preparation and will provide comprehensive training to operate the microscopes.

Confocal microscopy, most frequently **confocal laser scanning microscopy** (**CLSM**), is an optical imaging technique for increasing optical resolution and contrast of a micrograph by means of adding a spatial pinhole placed at the confocal plane of the lens to eliminate out-of-focus light.

It enables the reconstruction of three-dimensional structures from the obtained images by collecting sets of images at different depths (a process known as optical sectioning) within a thick object.

This technique has gained popularity in the scientific and industrial communities and typical applications are in life sciences, semiconductor inspection and materials science.

A conventional microscope "sees" as far into the specimen as the light can penetrate, while a confocal microscope only "sees" images one depth level at a time. In effect, the CLSM achieves a controlled and highly limited depth of focus.

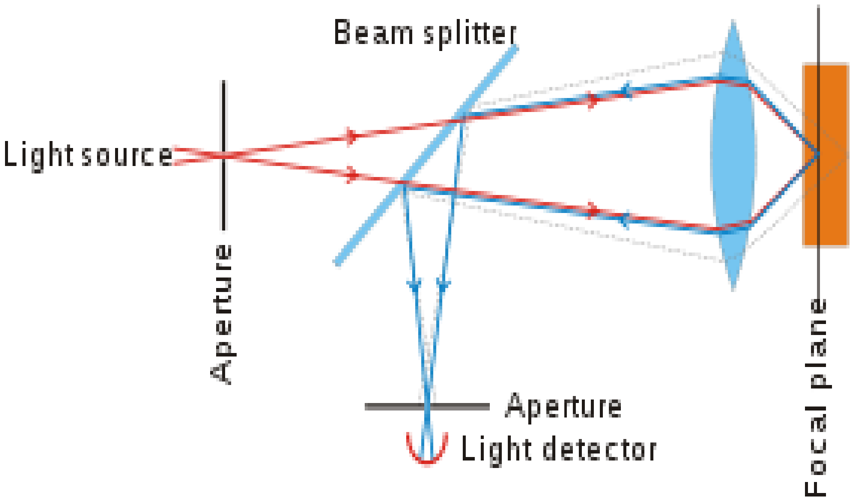


Fig 11.15Principle of confocal microscopy

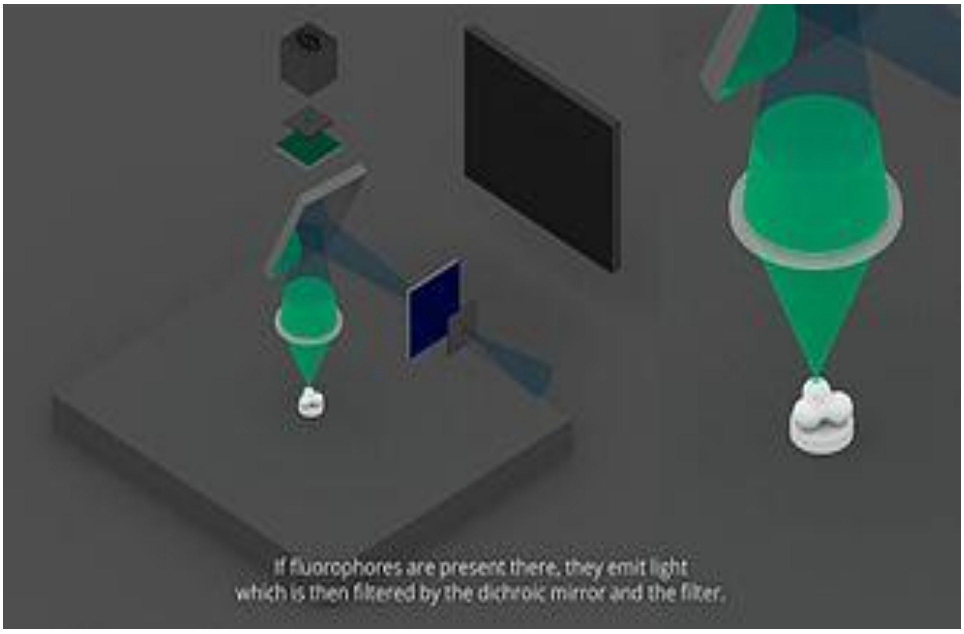
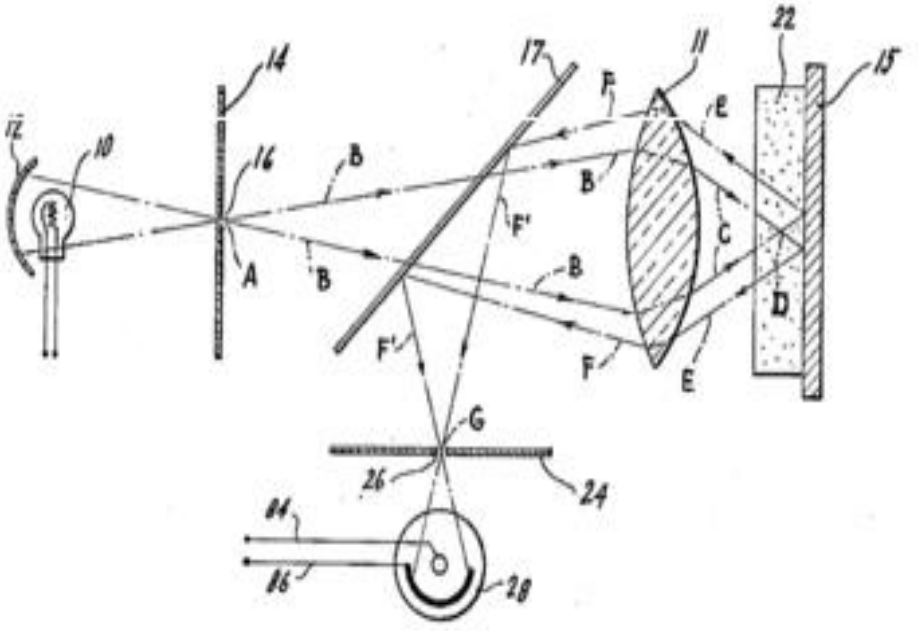

If fluorophores are present there, they emit light
which is then filtered by the dichroic mirror and the filter.

Fig 11.16 Fluorescent and confocal microscopes operating principle



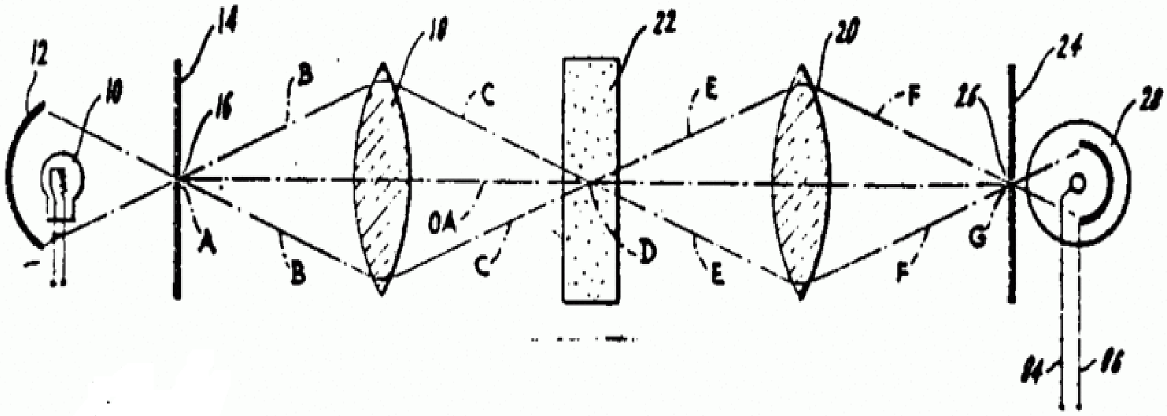


Fig 11.17Confocal point sensor principle

EXPERIMENT NO- 3.5

OBJECTIVE: To demonstrate of use of Electron microscopy

INTRODUCTION

- An **electron microscope** is a microscope that uses a beam of accelerated electrons as a source of illumination.
- As the wavelength of an electron can be up to 100,000 times shorter than that of visible light photons, electron microscopes have a higher resolving power than light microscopes and can reveal the structure of smaller objects.
- A transmission electron microscope can achieve better than 50 pm resolution and magnifications of up to about 10,000,000 xs whereas most light microscopes are limited by diffraction to about 200 nm resolution and useful magnifications below 2000x.
- Transmission electron microscopes use electrostatic and electromagnetic lenses to control the electron beam and focus it to form an image.
- These electron optical lenses are analogous to the glass lenses of an optical light microscope.
- Electron microscopes are used to investigate the ultra structure of a wide range of biological and inorganic specimens including microorganisms, cells, large molecules, biopsy samples, metals, and crystals.
- Industrially, electron microscopes are often used for quality control and failure analysis.
- Modern electron microscopes produce electron micrographs using specialized digital cameras and frame grabbers to capture the image.

The first electromagnetic lens was developed in 1926 by Hans Busch. According to Dennis Gabor, the physicist Leó Szilárd tried in 1928 to convince Busch to build an electron microscope, for which he had filed a patent.

German physicist Ernst Ruska and the electrical engineer Max Knoll constructed the prototype electron microscope in 1931, capable of four-hundred-power magnification; the apparatus was

the first demonstration of the principles of electron microscopy. Two years later, in 1933, Ruska built an electron microscope that exceeded the resolution attainable with an optical (light) microscope.[Moreover, Reinhold Rudenberg, the scientific director of Siemens-Schuckertwerke, obtained the patent for the electron microscope in May 1931.

In 1932, Ernst Lubcke of Siemens & Halske built and obtained images from a prototype electron microscope, applying concepts described in the Rudenberg patent applications. Five years later (1937), the firm financed the work of Ernst Ruska and Bodo von Borries, and employed Helmut Ruska (Ernst's brother) to develop applications for the microscope, especially with biological specimens. Also in 1937, Manfred von Ardenne pioneered the scanning electron microscope. The first commercial electron microscope was produced in 1938 by Siemens.

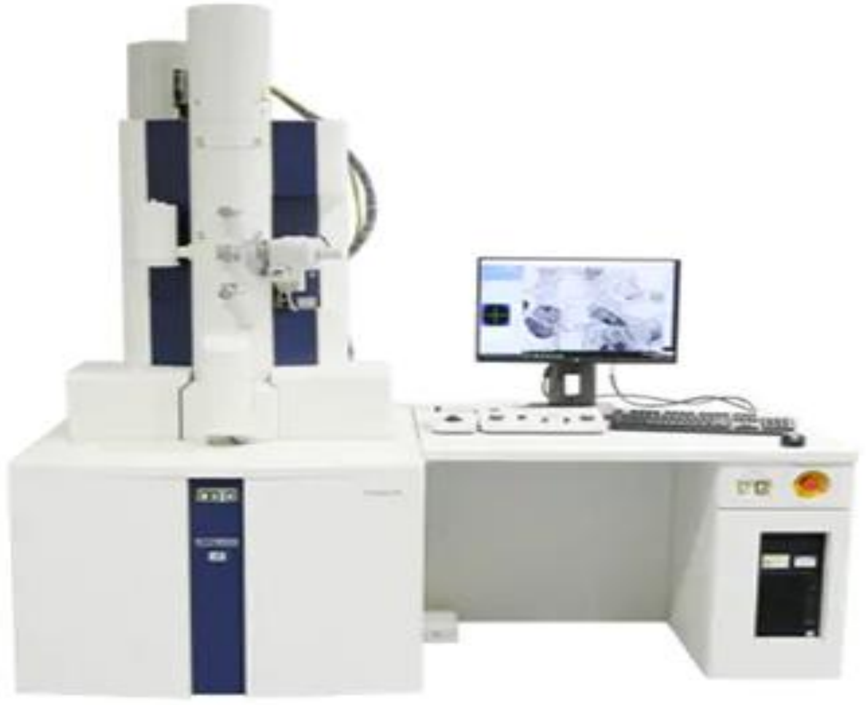


Fig 11.18A modern transmission electron microscope

Types

Transmission electron microscope (TEM)

The original form of electron microscope, the transmission electron microscope (TEM) uses a high voltage electron beam to illuminate the specimen and create an image.

The electron beam is produced by an electron gun, commonly fitted with a tungsten filament cathode as the electron source.

The electron beam is accelerated by an anode typically at +100 keV (40 to 400 keV) with respect to the cathode, focused by electrostatic and electromagnetic lenses, and transmitted through the specimen that is in part transparent to electrons and in part scatters them out of the beam.

When it emerges from the specimen, the electron beam carries information about the structure of the specimen that is magnified by the objective lens system of the microscope.

The spatial variation in this information (the "image") may be viewed by projecting the magnified electron image onto a fluorescent viewing screen coated with a phosphor or scintillator material such as zinc sulfide.

Alternatively, the image can be photographically recorded by exposing a photographic film or plate directly to the electron beam, or a high-resolution phosphor may be coupled by means of a lens optical system or a fibre optic light-guide to the sensor of a digital camera.

The image detected by the digital camera may be displayed on a monitor or computer.

The resolution of TEMs is limited primarily by spherical aberration, but new generations of aberration correctors have been able to partially overcome spherical aberration to increase resolution.

Hardware correction of spherical aberration for the high-resolution transmission electron microscopy (HRTEM) has allowed the production of images with resolution below 0.5 angstrom (50 picometres) and magnifications above 50 million times.

The ability to determine the positions of atoms within materials has made the HRTEM an important tool for nano-technologies research and development.

Transmission electron microscopes are often used in electron diffraction mode.

The advantages of electron diffraction over X-ray crystallography are that the specimen need not be a single crystal or even a polycrystalline powder, and also that the Fourier transform reconstruction of the object's magnified structure occurs physically and thus avoids the need for solving the phase problem faced by the X-ray crystallographers after obtaining their X-ray diffraction patterns of a single crystal or polycrystalline powder.

The major disadvantage of the transmission electron microscope is the need for extremely thin sections of the specimens, typically about 100 nanometers.

Biological specimens are typically required to be chemically fixed, dehydrated and embedded in a polymer resin to stabilize them sufficiently to allow ultrathin sectioning.

Sections of biological specimens, organic polymers and similar materials may require special treatment with heavy atom labels in order to achieve the required image contrast.

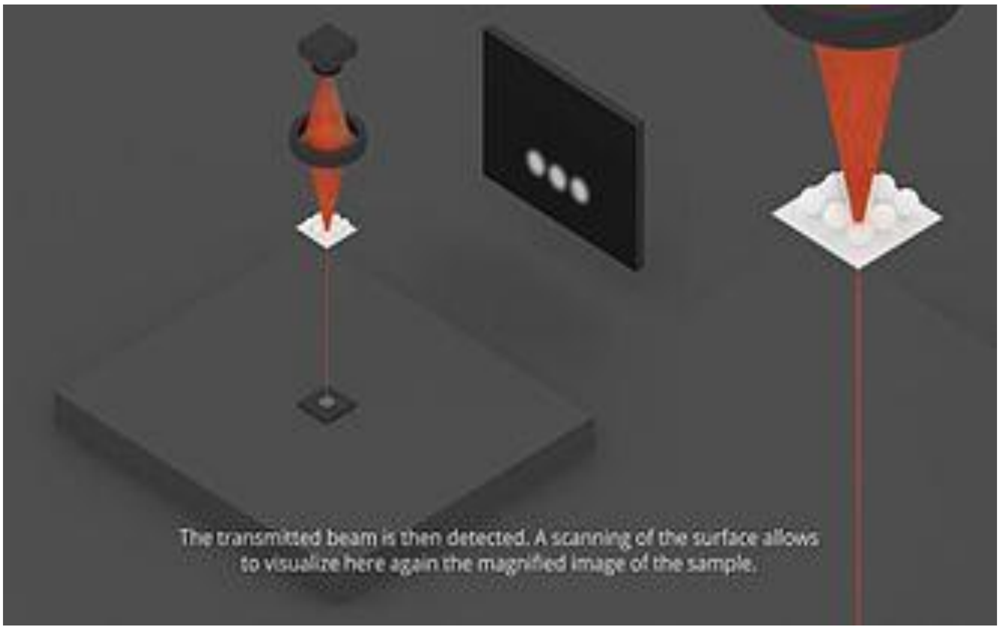

The transmitted beam is then detected. A scanning of the surface allows to visualize here again the magnified image of the sample.

Fig 11.20Operating principle of a Transmission Electron Microscope

Scanning electron microscope (SEM)

The SEM produces images by probing the specimen with a focused electron beam that is scanned across a rectangular area of the specimen (raster scanning).

When the electron beam interacts with the specimen, it loses energy by a variety of mechanisms.

The lost energy is converted into alternative forms such as heat, emission of low-energy secondary electrons and high-energy backscattered electrons, light emission (cathodoluminescence) or X-ray emission, all of which provide signals carrying information about the properties of the specimen surface, such as its topography and composition.

The image displayed by an SEM maps the varying intensity of any of these signals into the image in a position corresponding to the position of the beam on the specimen when the signal was generated.

In the SEM image of an ant shown below and to the right, the image was constructed from signals produced by a secondary electron detector, the normal or conventional imaging mode in most SEMs.

Generally, the image resolution of an SEM is at least an order of magnitude poorer than that of a TEM.

However, because the SEM image relies on surface processes rather than transmission, it is able to image bulk samples up to many centimeters in size and (depending on instrument design and settings) has a great depth of field, and so can produce images that are good representations of the three-dimensional shape of the sample.

Another advantage of SEM is its variety called environmental scanning electron microscope (ESEM) can produce images of sufficient quality and resolution with the samples being wet or contained in low vacuum or gas.

This greatly facilitates imaging biological samples that are unstable in the high vacuum of conventional electron microscopes.

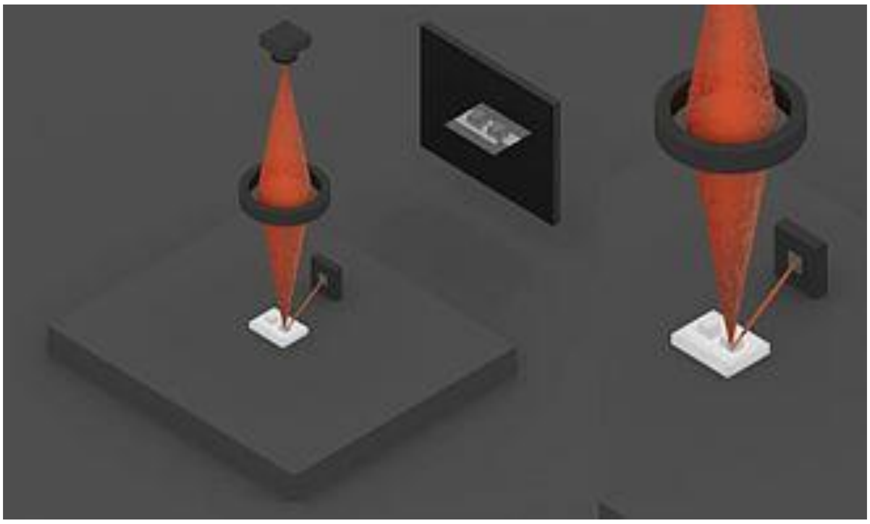


Fig 11.21Operating principle of a Scanning Electron Microscope

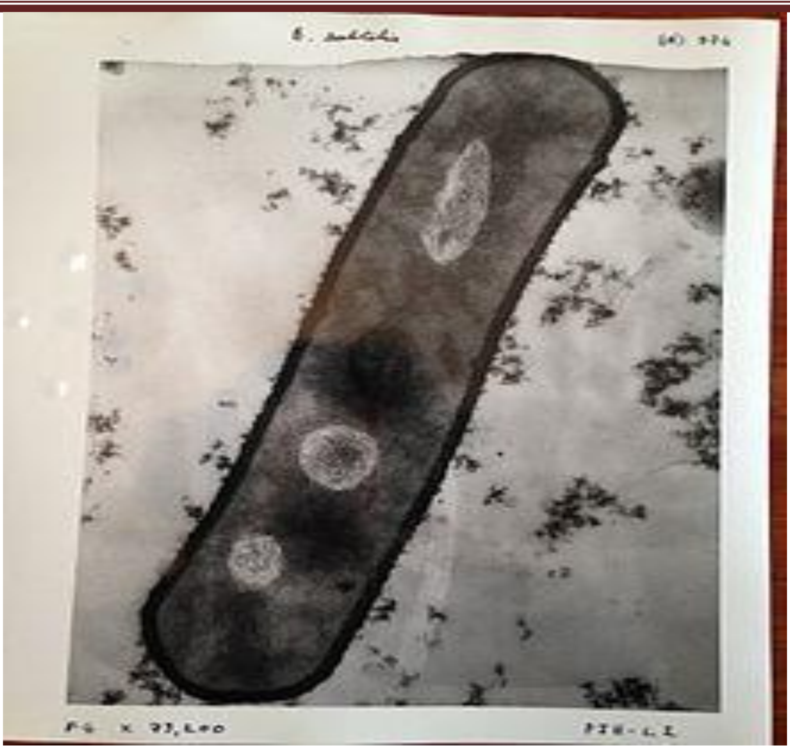


Fig 11.22Image of bacillus subtilis taken with a 1960s electron microscope

SUMMARY

1. A scientific instrument (pH meter, Spectrophotometer, Bright Field, Phase Contrast, Dark Field, Fluroscence, Cofocal and Electron Microscope) are an instrument used for laboratory purposes.

2. Most are measuring instruments. They may be specifically designed, constructed and refined for the purpose. Over time, instruments have become more accurate and precise.

3. Scientific instruments are part of laboratory equipment, but are considered more sophisticated and more specialized than other measuring instruments as scales, rulers, chronometers, thermometers or even waveform generators.

4. They are increasingly based upon the integration of computers to improve and simplify control, enhance and extend instrumental functions, conditions, parameter adjustments and data sampling, collection, resolution, analysis (both during and post-process), storage and retrieval.

GLOSSARY

1**. pH Meter**: Instrument that measures the hydrogen-ion concentration (or pH) in a solution, indicating its **acidity** or **alkalinity**. The pH meter measures the difference in electrical potential between a pH electrode and a reference electrode.

2. **Cuvet(te):** A clear, rectangular vessel of glass or plastic used to hold solutions for spectrometry. Don't place its ribbed sides in the light path of the spectrometer!

3**. Absorbance and 100% Line**: Absorbance is the unit utilized for measuring the amount of IR radiation absorbed by a material.

4. **Wavelength and Wavenumber**. Wavelength is the interval between two adjacent crests or troughs of a light wave. Wave number is 1/wavelength and is expressed in cm^{-1}. It is widely utilized as the X axis unit in infrared spectra.

5. **Zero Path Difference (ZPD), or Zero Optical Path Difference (ZOPD):** In an interferometer, the mirror displacement upon which the optical path difference for the two beams is zero. The detector signal is often much larger at ZPD, ZOPD, and is called the center burst.

6. **TEM:** Transmission Electron Microscopy.

7. **SEM:** Scanning Electron Microscope.

8. **Dark-field microscope**. A **microscope** in which an object is illuminated only from the sides so that it appears bright against a **dark** background.

9. **Absorbance (Optical Density)** - The quantity of light absorbed by a chemical or biological substance as measured in a spectrophotometer or similar device. Units of absorbance are equivalent to the logarithm of the reciprocal transmittance (the ratio of the transmitted light intensity to incident light intensity).

10. **Bioluminescence** - A biochemical oxidative process that results in the release of energy as emitted light. Firefly luminescence, which requires the enzyme luciferase to catalyze a reaction between the substrate luciferin and molecular oxygen (in the presence of adenosine triphosphate), is a commonly employed example of bioluminescence. The phenomenon occurs in a wide variety of marine organisms and insects.

SELF ASSESSMENT QUESTIONS

1. How often should I calibrate my pH sensor?
2. How can I calibrate the spectrophotometer?
3. What are the TEM and SEM?
4. What is optical density?
5. What are zero path differences?

REFERENCES

1. Advanced Light Microscopy vol. 1 Principles and Basic Properties by Maksymilian Pluta, Elsevier (1988)
2. Introduction to Light Microscopy by S. Bradbury, B. Bracegirdle, BIOS Scientific Publishers (1998)
3. Microbiology: Principles and Explorations by Jacquelyn G. Black, John Wiley & Sons, Inc. (2005).

4. Douglas B. Murphy - Department of Cell Biology and Anatomy and Microscope Facility, Johns Hopkins University School of Medicine, 725 N. Wolfe Street, 107 WBSB, Baltimore, Maryland 21205.
5. Jump up ^ "The phase contrast microscope". Nobel Media AB.
6. Jump up ^ "Phase Contrast". Leica Science Lab.
7. Jump up ^ Frits Zernike (1942). "Phase contrast, a new method for the microscopic observation of transparent objects part I". Physica. 9 (7): 686–698. Bibcode:1942Phy.....9..686Z. Doi:10.1016/S0031-8914(42)80035-X.
8. Confocal Microscopy (3rd Ed.). Berlin: Springer. ISBN 0-387-25921-X.
9. Jump up ^ Vincze L (2005). "Confocal X-ray Fluorescence Imaging and XRF Tomography for Three Dimensional Trace Element Microanalysis". Microscopy and Microanalysis. 11 (Supplement 2). doi:10.1017/S1431927605503167.

UNIT 12 BIOTECHNOLOGY EXERCISE (PART- I)

Content

12.1 Study of the principles and applications of the following equipment

12.1.1 Laminar flow

12.1.2 Autoclave

12.1.3 Elisa reader

12.1.4 PCR machine

12.1.5 Refrigerated centrifuge

12.1.6 Transilluminator

12.2 Double helical DNA modal

12.3 Chromatography or Thin Layer Chromatography (TLC)

12.4 Recombinant DNA techniques

Experiment No-(01)

OBJECTIVE**:** Study of the principal and application of Laminar flow

INTRODUCTION

In fluid dynamics, **laminar flow** (or streamline flow) occurs when a fluid flows in parallel layers, with no disruption between the layers.

At low velocities, the fluid tends to flow without lateral mixing, and adjacent layers slide past one another like playing cards.

There are no cross-currents perpendicular to the direction of flow, nor eddies or swirls of fluids.

In laminar flow, the motion of the particles of the fluid is very orderly with particles close to a solid surface moving in straight lines parallel to that surface.

Laminar flow is a flow regime characterized by high momentum diffusion and low momentum convection.

When a fluid is flowing through a closed channel such as a pipe or between two flat plates, either of two types of flow may occur depending on the velocity and viscosity of the fluid: laminar flow or turbulent flow.

Laminar flow tends to occur at lower velocities, below a threshold at which it becomes turbulent.

Turbulent flow is a less orderly flow regime that is characterised by eddies or small packets of fluid particles which result in lateral mixing. In non-scientific terms, laminar flow is smooth while turbulent flow is rough.

Laminar flow barriers

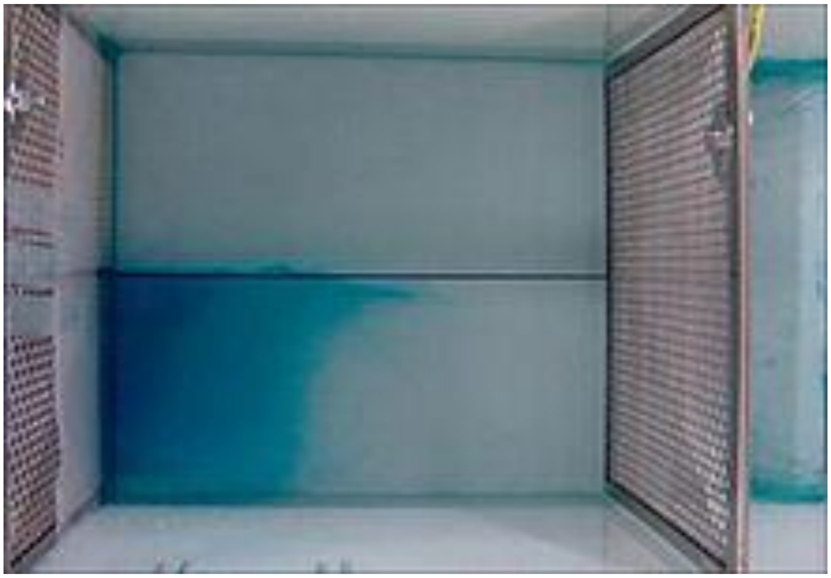


Fig 12.1Experimental chamber for studying chemo taxis in response to laminar flow

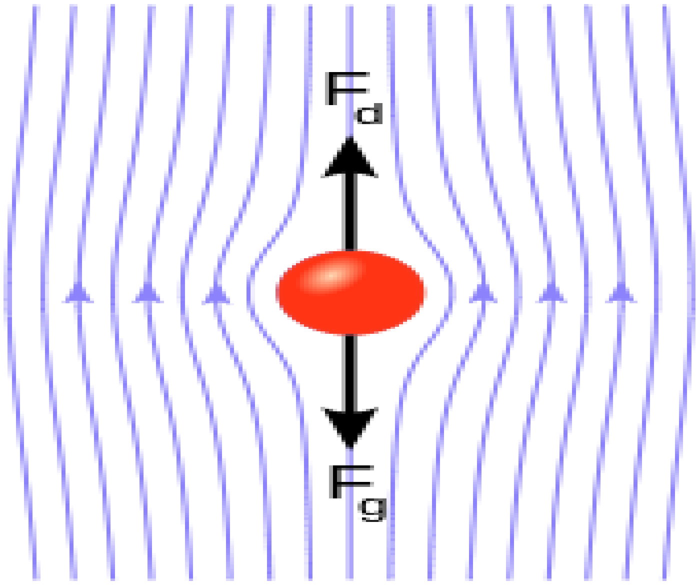


Fig 12.2 Reynolds Number

Relationship with the Reynolds number

The common example is flow through a pipe, where the Reynolds number is defined as:

$$\mathbf{Re} = \frac{\rho \mathbf{v} D_{\mathrm{H}}}{\mu} = \frac{\mathbf{v} D_{\mathrm{H}}}{\nu} = \frac{\mathbf{Q} D_{\mathrm{H}}}{\nu A}$$

Where:

- D_H is the hydraulic diameter of the pipe; its characteristic travelled length, *L*, (m).
- **Q** is the volumetric flow rate (m^3/s).
- *A* is the pipe's cross-sectional area (m^2).
- **v** is the mean velocity of the fluid (SI units: m/s).
- μ is the dynamic viscosity of the fluid ($Pa{\cdot}s = N{\cdot}s/m^2 = kg/(m{\cdot}s)$).
- ν is the kinematic viscosity of the fluid, $\nu = \mu/\boxed{\rho}$ (m^2/s).
- ρ is the density of the fluid (kg/m^3).

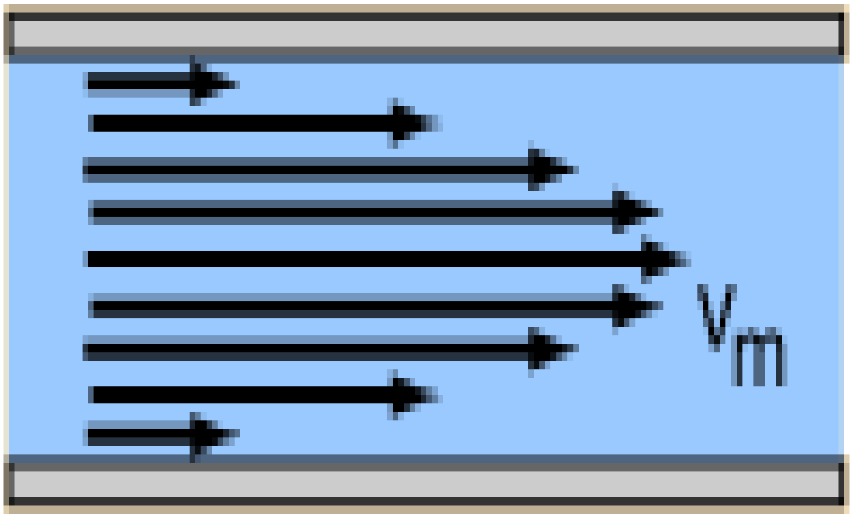
For such systems, laminar flow occurs when the Reynolds number is below a critical value of approximately 2,040; through the transition range is typically between 1,800 and 2,100.

For fluid systems occurring on external surfaces, such as flow past objects suspended in the fluid, other definitions for Reynolds numbers can be used to predict the type of flow around the object. The particle Reynolds number Re_p would be used for particle suspended in flowing fluids, for example.

As with flow in pipes, laminar flow typically occurs with lower Reynolds numbers, while turbulent flow and related phenomena, such as vortex shedding, occur with higher Reynolds numbers.

Examples:

In the case of a moving plate in a liquid, it is found that there is a layer (lamina) that moves with the plate, and a layer next to any stationary plate that is stationary.



EXPERIMENT NO- 2

***OBJECTIVE*:** Study of the principal and application of Autoclave

INTRODUCTION

An **autoclave** is a pressure chamber used to carry out industrial processes requiring elevated temperature and pressure different from ambient air pressure.

Autoclaves are used in medical applications to perform sterilization and in the chemical industry to cure coatings and vulcanize rubber and for hydrothermal synthesis.

They are also used in industrial applications, especially regarding composites, see autoclave (industrial).

Many autoclaves are used to sterilize equipment and supplies by subjecting them to high-pressure saturated steam at 121 °C (249 °F) for around 15–20 minutes depending on the size of the load and the contents.

The autoclave was invented by Charles Chamberland in 1879, although a precursor known as the steam digester was created by Denis Papin in 1679.

The name comes from Greek auto-, ultimately meaning self and Latin clavis meaning key, thus a self-locking device.

WHAT IS AN AUTOCLAVE?

The **autoclave** carries out that exact function of sterilizing materials.

It is a machine that uses pressure and steam to reach and maintain a temperature that is too high for any microorganisms or their spores to live.

Microorganisms are what most people commonly refer to as germs. These are the bacteria, viruses, fungi, parasites, etc. that are able to cause infections in our bodies.

Spores are the environment-resistant form of the microorganisms. Even though they are able to withstand harsher conditions, they still can be killed if extreme conditions are maintained for an extended period of time.

HOW IT WORKS

Autoclaves are pressure cookers very similar to the ones that you see in the stores. If you have used, or are familiar with pressure cookers, then you know that foods cook a lot faster in a pressure cooker than they do in a regular pot or in the oven. This is due to the intense heat and pressure that is applied to the food.

The same mechanism works against living microorganisms. Once an autoclave is started, steam is pushed into the chamber that contains the items that are being sterilized.

As the steam goes in, the pressure and temperature within the chamber is increased. Most autoclaves are set to increase steam pressure until a temperature of at least 121 degrees Celsius is reached (about 250 degrees Fahrenheit).

This temperature and pressure will remain at this level for at least 15 minutes. This is a high enough temperature for a long enough period of time to kill any and all microorganisms and their spores.

Uses

Sterilization autoclaves are widely used in microbiology, medicine, podiatry, tattooing, body piercing, veterinary medicine, mycology, funeral homes, dentistry, and prosthetics fabrication. They vary in size and function depending on the media to be sterilized.

Typical loads include laboratory glassware, other equipment and waste, surgical instruments, and medical waste.

Air removal

It is very important to ensure that all of the trapped air is removed from the autoclave before activation, as trapped air is a very poor medium for achieving sterility.
Steam at 134 °C can achieve in three minutes the same sterility that hot air at 160 °C can take two hours to achieve. Methods of air removal include:

Downward displacement (or **gravity-type**):

As steam enters the chamber, it fills the upper areas first as it is less dense than air. This process compresses the air to the bottom, forcing it out through a drain which often contains a temperature sensor.

Only when air evacuation is complete does the discharge stop. Flow is usually controlled by a steam trap or a solenoid valve, but bleed holes are sometimes used, often in conjunction with a solenoid valve.

As the steam and air mix, it is also possible to force out the mixture from locations in the chamber other than the bottom.

Steam pulsing

Pressurized and then depressurized to near atmospheric Air dilution by using a series of steam pulses, in which the chamber is alternately pressure.

Vacuum pumps

A vacuum pump sucks air or air/steam mixtures from the chamber.

Super atmospheric cycles

Achieved with a vacuum pump. It starts with a vacuum followed by a steam pulse followed by a vacuum followed by a steam pulse.

The number of pulses depends on the particular autoclave and cycle chosen.

Sub-atmospheric cycles

Similar to the super-atmospheric cycles, but chamber pressure never exceeds atmospheric pressure until they pressurize up to the sterilizing temperature.

A medical autoclave is a device that uses steam to sterilize equipment and other objects. This means that all bacteria, viruses, fungi, and spores are inactivated.

However, prions, such as those associated with Creutzfeldt–Jakob disease, may not be destroyed by autoclaving at the typical 134 °C for three minutes or 121 °C for 15 minutes.

Although that a wide range species of archaea, including *Geogemma barosii*, can survive at temperatures above 121 °C, no archaea are known to be infectious or pose a health risk to humans; in fact their biochemistry is so vastly different from our own and their multiplication rate is far too slow for microbiologists to worry about them.

Autoclaves are found in many medical settings, laboratories, and other places that need to ensure the sterility of an object.

Many procedures today employ single-use items rather than sterilizable, reusable items.

This first happened with hypodermic needles, but today many surgical instruments (such as forceps, needle holders, and scalpel handles) are commonly single-use rather than reusable items.

Autoclaves are of particular importance in poorer countries due to the much greater amount of equipment that is re-used.

Providing stove-top or solar autoclaves to rural medical centres has been the subject of several proposed medical aid missions.

Because damp heat is used, heat-labile products (such as some plastics) cannot be sterilized this way or they will melt.

Paper and other products that may be damaged by steam must also be sterilized another way.

In all autoclaves, items should always be separated to allow the steam to penetrate the load evenly.

Autoclaving is often used to sterilize medical waste prior to disposal in the standard municipal solid waste stream.

This application has become more common as an alternative to incineration due to environmental and health concerns raised because of the combustion by-products emitted by incinerators, especially from the small units which were commonly operated at individual hospitals.

Incineration or a similar thermal oxidation process is still generally mandated for pathological waste and other very toxic and/or infectious medical waste.

In dentistry, autoclaves provide sterilization of dental instruments according to health technical memorandum 01-05 (HTM01-05).

According to HTM01-05, instruments can be kept, once sterilized using a vacuum autoclave for up to 12 months using sealed pouches.

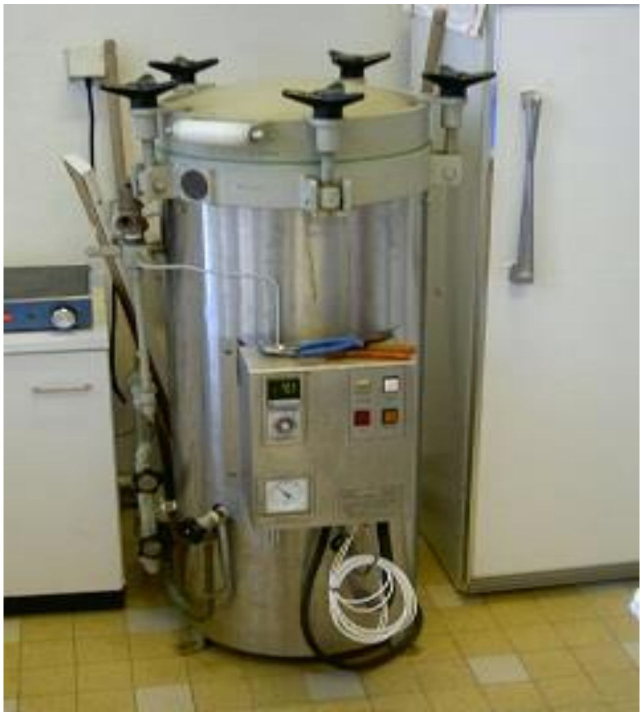
.


Fig 12.3 This is an autoclave that is used in the medical facility

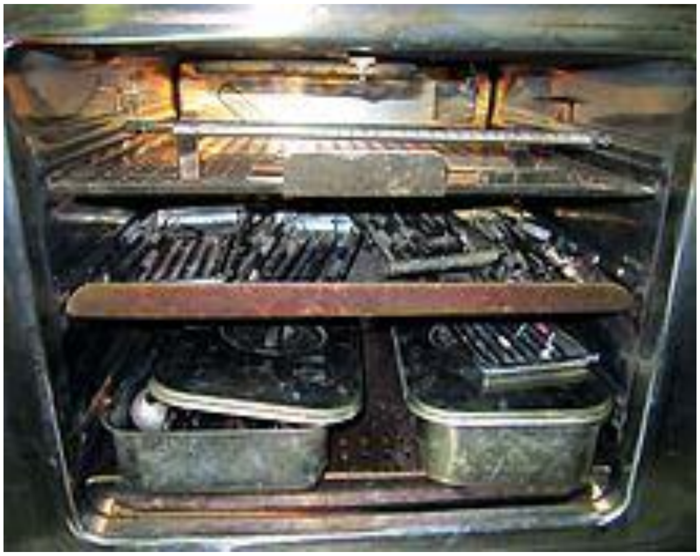


Fig 12.4Other Autoclave with dental equipment in an autoclave to be sterilized for 2 hours at 150 to 180 degrees Celsius

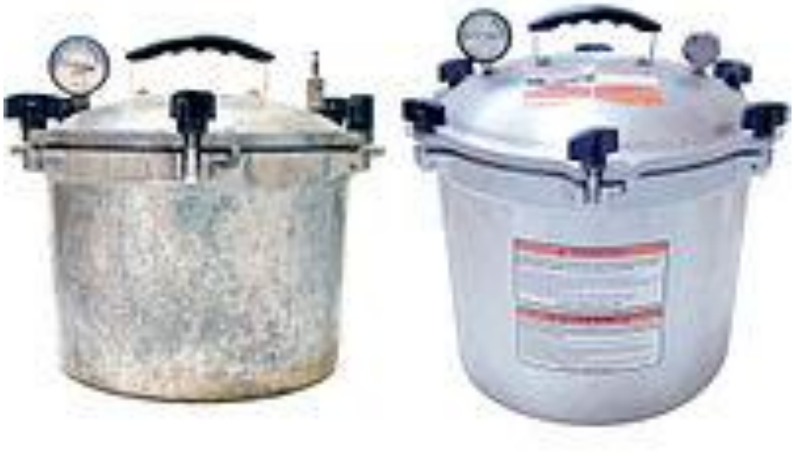


Fig 12.5 Stovetop autoclaves—the simplest of autoclaves

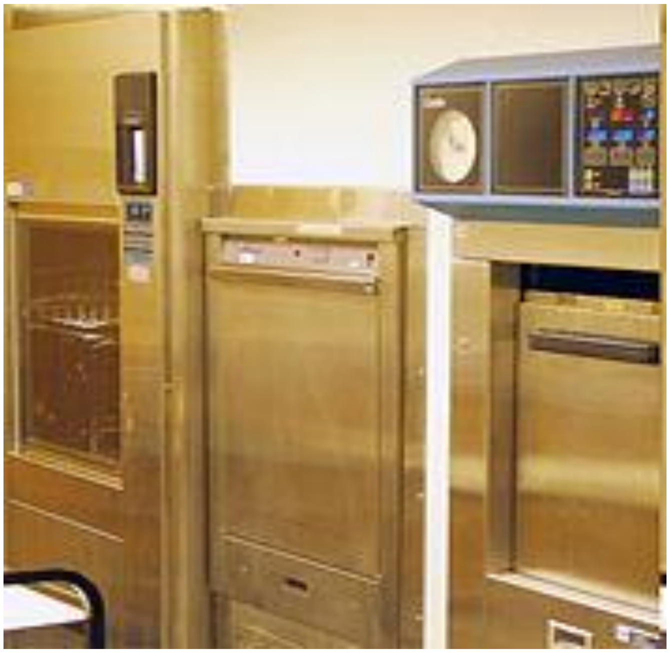


Fig 12.6The machine on the right is an autoclave used for processing substantial quantities of laboratory equipment prior to reuse, and infectious material prior to disposal. (The machines on the left and in the middle are washing machines.)

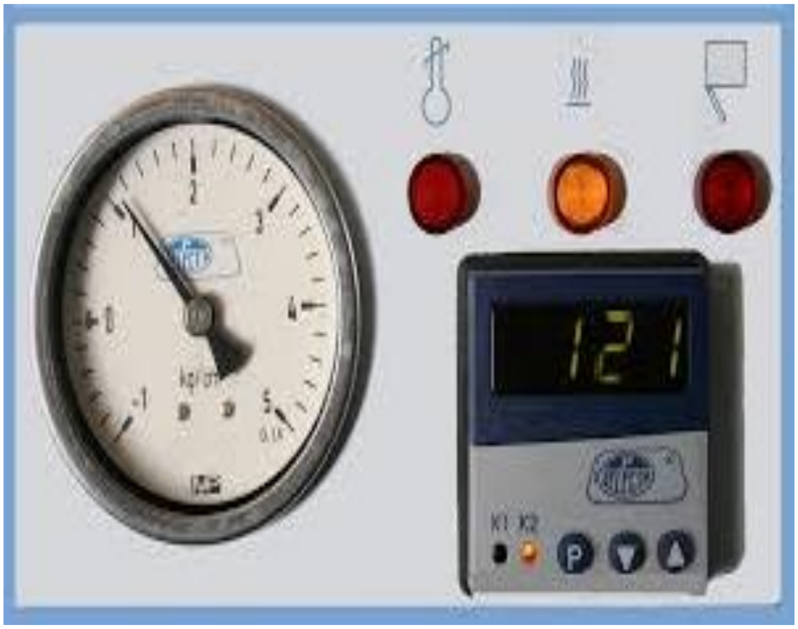


Fig 12.7 The Autoclave must reach 121 degrees Celsius

The high temperatures cause the internal parts of the microorganisms to essentially cook. Once the internal parts cannot function in the microorganisms, they will die. The steam and pressure are released and brought down to normal room temperature and pressure after the 15 or more minutes of running. The items that were autoclaved will remain sterile until they are contaminated by new microorganisms.

EXPERIMENT NO- 3

OBJECTIVE**:** Study of the principal and application of Elisa Reader

INTRODUCTION

ELISA stands for enzyme linked immune-sorbent assay. In short, it is an antibody test or a test for immune response to things attacking the body such as virus, bacteria and allergens. The test is done in an ELISA plate, also known as a 96-well plate or micro plate. The ELISA reader reads the plate.

WHAT AN ELISA READER DOES

An ELISA reader measures and quantities the color differences in the 12 wells of the plate.

ELISA readers or micro plate readers do spectrophotometer; they emit light at one wave length, and measure the amount of light absorbed and reflected by an object such as a protein. A spectrophotometer measures ultraviolet and visible light.

Additionally, ELISA plate readers can also measure fluorescence and luminescence. Chemical dyes fluoresce or emit one color or wavelength when exposed to light. The amount of reflection, absorption and the color identify, and measure the amount of a substance.

PURPOSE OF AN ELISA READER

ELISA readers were designed for measuring antibody tests. They worked so well, the machine has been adapted to other purposes. Researchers use them for protein and enzyme assays. They are also used for HIV detection and quantization of nucleic acids.

ADVANTAGES OF ELISA READER

Spectrophotometers require more sample per measurement .

To use a spectrophotometer or ELISA plate reader, the molecule must be dissolved in solution.

A spectrophotometer requires between 400 micro-liters and four milliliters, depending on the manufacturer and model. An ELISA plate reader needs about two to 100 micro-liters; ELISA plate readers use much less of a sample to get a result.

ELISA plate readers measure more samples in a shorter period of time. A spectrophotometer measures one to six samples at a time. Typically, an ELISA plate measures 96 wells in an equivalent amount of time.

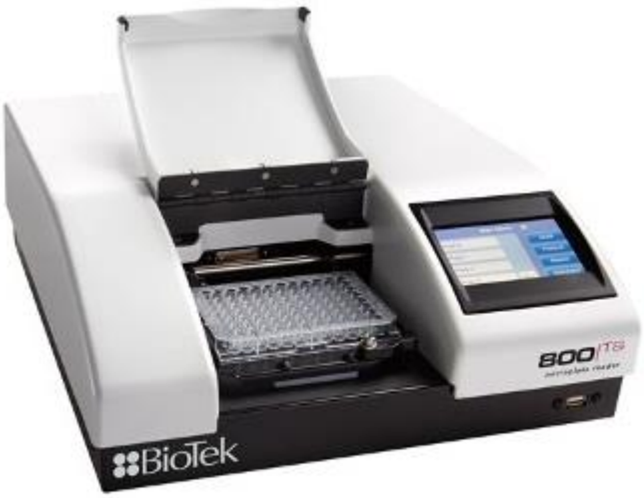


Fig 12.8 Demonstration of ELISA reader

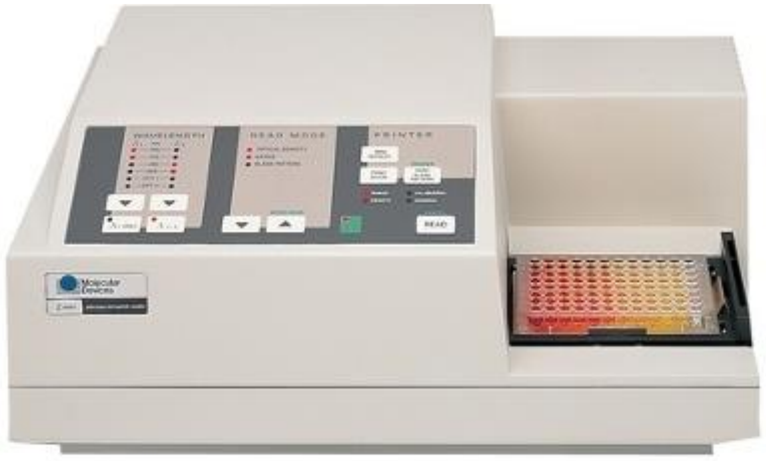


Fig 12.9 ELISA plate readers

EXPERIMENT NO- 4

***OBJECTIVE*: ** Study of the principal and application of PCR Machine

INTRODUCTION

PCR (polymerase chain reaction) machine, also known as a thermal cycler, is a DNA amplifier that regulates temperature and amplifies segments of DNA via the polymerase chain reaction**.** PCR machines are also sometimes used to facilitate reactions regarding enzyme digestion or rapid diagnostics.

PCR requires four main components. The first is the DNA sample containing the section, or sections, for copying.

Secondly, PCR requires a primer. Primers are short segments of DNA a scientist creates to match the DNA sample.

The next requirement of PCR is DNA polymerase, an enzyme that copies DNA. Human DNA polymerase denatures, or breaks down, at PCR temperatures, so researchers often use DNA polymerase from a heat-tolerant bacterium.

Finally, PCR requires nucleotides: adenine, guanine, cytosine and thymine. These are the base pairs that provide the coding elements of DNA.

Researchers first heat the PCR mixture to a temperature that denatures the DNA double helix.

A cooling step follows, which allows the primers to bind to the sample DNA. Another healing cycle follows during which DNA polymerase elongates the DNA strand.

Repeating these steps multiple times creates many new copies of the original DNA sample in as few as three hours.

PCR is useful in the diagnosis of viral diseases and some forms of cancer. It is also a common tool in forensic science for replicating small samples from crime scenes.

PCR is used for research when it is necessary to make a large amount of a single gene, such as for genetic engineering or cloning. PCR is also used to test whether or not a particular gene product is present in a sample.

In forensics, PCR technology is used to carry out DNA fingerprinting to analyze crime scene DNA evidence. PCR can also be used in medical settings to carry out tissue-typing for organ transplants.

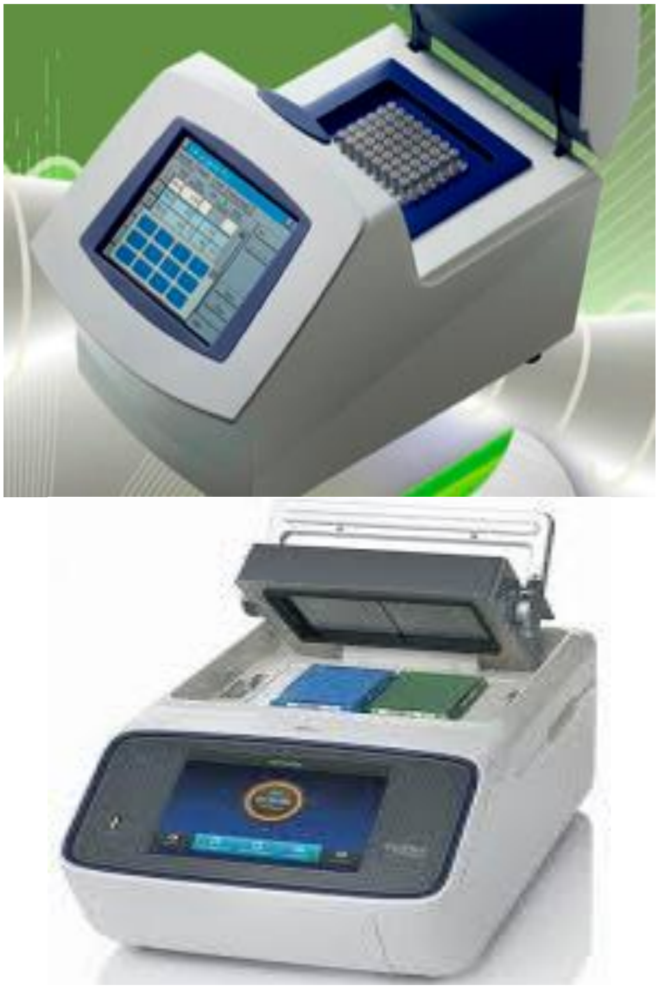
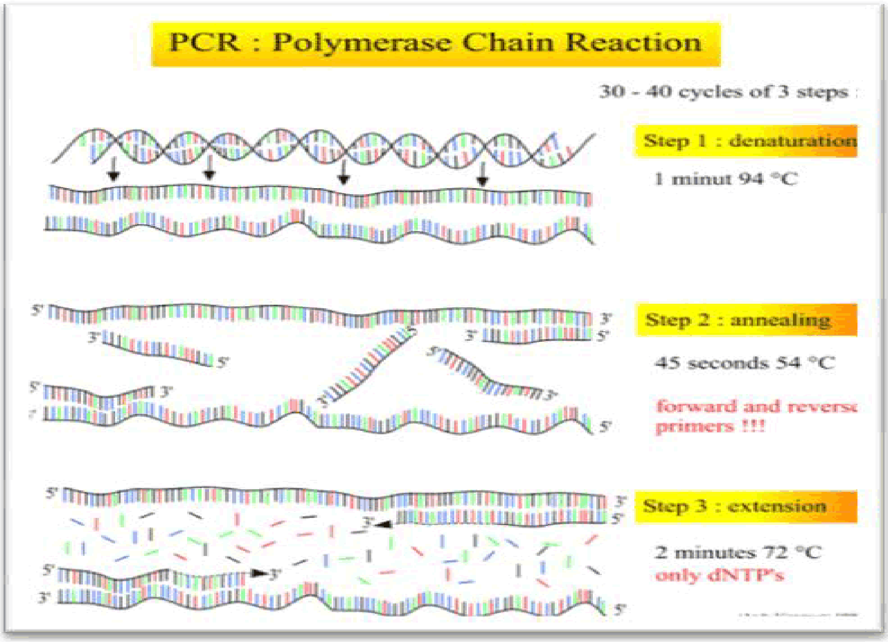


Fig 12.10 PCR Machine

Fig 12.11The purpose of PCR is to amplify small amounts of a DNA sequence of interest so it can be analyzed separately. PCR can be used to make a large amount of a specific piece of DNA or to test a DNA sample for that sequence



Experiment No- 5

***OBJECTIVE*:** Study of the principal and application Refrigerated Centrifuge

INTRODUCTION

Refrigerated centrifuge works on the concept of sedimentation principle by holding up the sample tubes with a capacity of 2ml, 10ml and 50ml in rotation around a fixed axis.

In this, the centripetal force causes the denser substances to separate out along the radial direction in the bottom of the centrifuge tube.

The rate of the centrifugation is calculated by the acceleration applied to the sample and it is typically measured in revolution per minute (RPM) or relative centrifugal force (RCF).

The particle's settling velocity during centrifugation depends on the function of their size and shape, centrifugal acceleration, the volume fraction of solids present, the density difference between the particle and the liquid, and the viscosity.

This equipment is extensively used in chemistry, biology, and biochemistry for isolating and separating suspensions.

It additionally provides the cooling mechanism to maintain the uniform temperature throughout the operation of the sample.

Fig.12.12 Refrigerated centrifuge

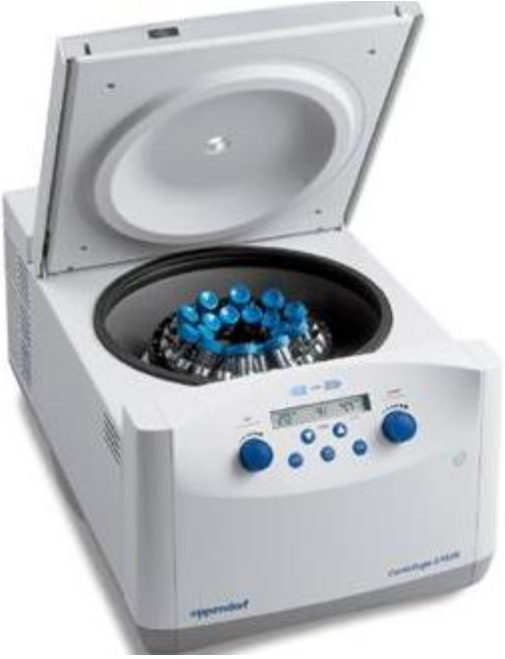


Fig 12.13 Compact and quiet 4 x 85 ml refrigerated centrifuge for low-speed clinical and research applications

FUNCTIONS

- Fast Cool function eliminates long waits for set temperatures
- Easy-to-use interface with digital display improves repeatability
- Standby cooling maintains temperature when chamber is not in use
- -9 to 40°C temperature range

Centrifuge is ideal for low-speed spinning (up to 3,000 x g) of tubes ranging in volume from 1.1 to 85 ml.

The space-saving footprint and innovative design accommodates 30 x 15 ml round-bottom tubes or 20 x 15 ml conical tubes.

Automatic, motorized locking lid makes loading and closing effortless. A variety of rotors can be used for optimal versatility in high-capacity or sensitive applications.

For delicate blood and urine samples, users can deactivate the electronic brake.

Time and speed are easily adjusted and digitally displayed for accuracy and all settings are adjustable during a run.

Two preset buttons store routine runs for added convenience. Settings can also be locked to prevent unintentional adjustment.

Timer is programmable up to 99 minutes or continuous, and an audible alert signals the end of a run. Measures 15"W x 10.2"H x 22.8"D (38 x 26 x 58 cm). Weight 79 lbs./36 kg.

Centrifuge 5702R (refrigerated) with 4x100mL swing-bucket rotor (A-4-38) including buckets. Max. RCF 3,000xg (4400rpm),

Temperature range -9 - 40°C. Compact footprint, Whisper quiet operation, SOFT brake option, 120V/50-60Hz.

Centrifuge 5702R (refrigerated) with 4x100mL swing-bucket rotor (A-4-38) including buckets. Max. RCF 3,000xg (4400rpm),

EXPERIMENT NO- 6

***OBJECTIVE*:** Study of the principal and application of Tran illuminator

INTRODUCTION

Tran illuminators are used in molecular biology labs to view DNA (or RNA) that has been separated by electrophoresis through an agarose gel.

During or immediately after electrophoresis, the agarose gel is stained with a fluorescent dye which binds to nucleic acid.

Exposing the stained gel to a UVB light source causes the DNA/dye to fluoresce and become visible.

This technique is used wherever the researcher needs to be able to view their sample, for example sizing a PCR product, purifying DNA segment after a restriction enzyme digest, quantifying DNA or verifying RNA integrity after extraction.

It describes how to make a UVB (310nm) transilluminator with a 7 x 7 cm window for viewing ethidium bromide (or SYBR-Safe) stained DNA mini-gels.

Once all of the materials are collected, the actual assembly time is approx. 1-2 hours. Some soldering is required.

For the UV transilluminator enclosure and lid. Laser cut the parts from the material listed in the design file.

If you do not have access to a laser cutter, you can send the files to any laser cutting service such as Pololu. Materials for laser cutting can be found at any supplier of acrylic materials (McMaster-Carr, US Plastics etc) except for the solacryl (UV-transmissive) which can be bought from Loop Acrylics.

Tap holes in the following parts:

- 5-40: Two holes in the enclosure side with the cutout for the power switch
- 8-32: Four holes in the solacryl cover
- 8-32: Four holes in the 0.25" clear lid side part for mounting the hinges
- 8-32: Two holes in the enclosure bottom

Safety Notes:

1. Because ethidium bromide is a toxic chemical with strict safety protocols, it is only recommended that you use this dye in a lab with established handling, storage and waste disposal procedures in place**.** Other users are strongly recommended to use SYBR-Safe instead, which can be handled and disposed of more safely.
2. The transilluminator does come with a safety lid for viewing the gel. However, when the lid is not in place, safety glasses mustbe worn when operating the UVB bulb.
3. If you prefer to avoid UVB altogether, we can recommend the blue light LED transilluminators such as the one described in this instructable instead.

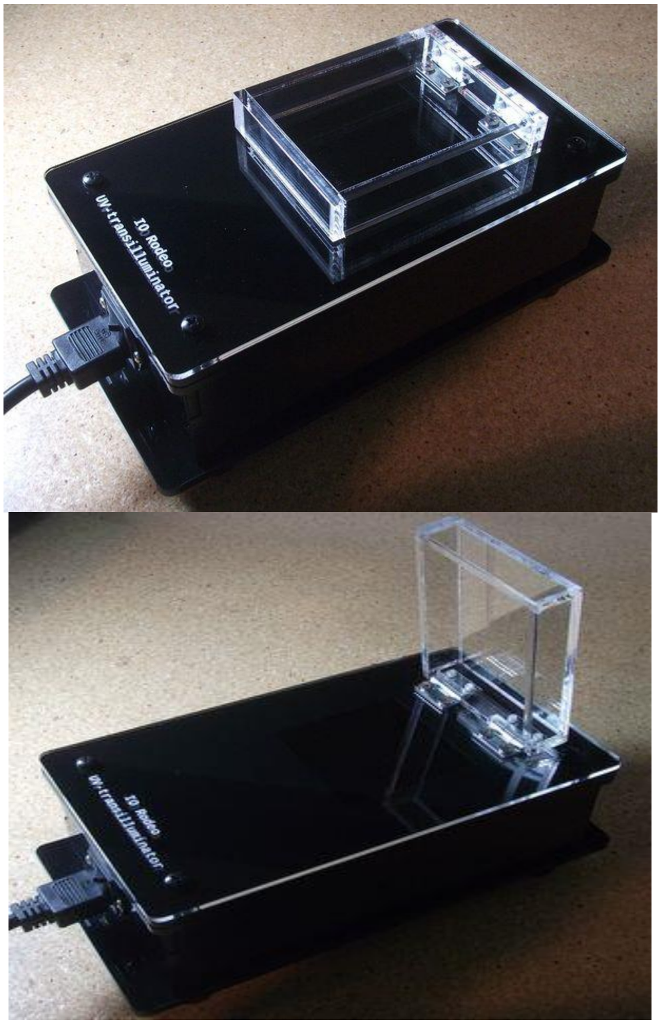


Fig12.14 ***UV***-*TransiIlluminator*

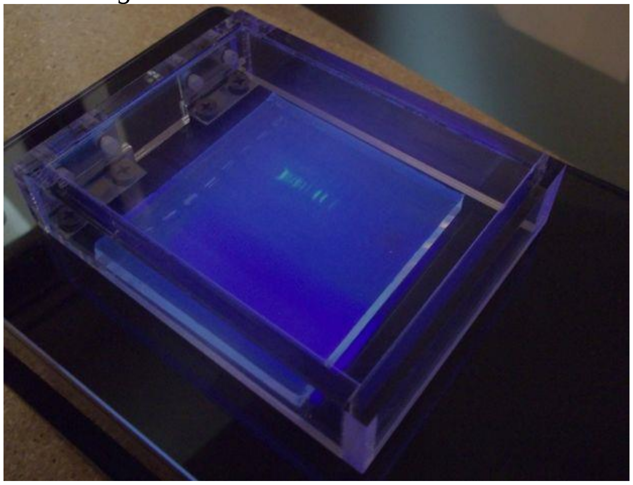


Fig 12.15   UV- Tran illuminator

EXPERIMENT NO- 07

***OBJECTIVE*:** Study of the Double helical DNA model

INTRODUCTION

Many people believe that American biologist James Watson and English physicist Francis Crick discovered DNA in the 1950s.

In reality, this is not the case. Rather, DNA was first identified in the late 1860s by Swiss chemist Friedrich Miescher.

Then, in the decades following Miescher's discovery, other scientists--notably, Phoebus Levene and Erwin Chargaff--carried out a series of research efforts that revealed additional details about the DNA molecule, including its primary chemical components and the ways in which they joined with one another.

Without the scientific foundation provided by these pioneers, Watson and Crick may never have reached their groundbreaking conclusion of 1953: that the DNA molecule exists in the form of a three-dimensional double helix.

Watson and Crick Propose the Double Helix

Chargaff's realization that A = T and C = G, combined with some crucially important X-ray crystallography work by English researchers Rosalind Franklin and Maurice Wilkins, contributed to Watson and Crick's derivation of the three-dimensional, double-helical model for the structure of DNA.

Watson and Crick's discovery was also made possible by recent advances in model building, or the assembly of possible three-dimensional structures based upon known molecular distances and bond angles, a technique advanced by American biochemist Linus Pauling.

In fact, Watson and Crick were worried that they would be "scooped" by Pauling, who proposed a different model for the three-dimensional structure of DNA just months before they did. In the end, however, Pauling's prediction was incorrect.

Using cardboard cutouts representing the individual chemical components of the four bases and other nucleotide subunits, Watson and Crick shifted molecules around on their desktops, as though putting together a puzzle.

They were misled for a while by an erroneous understanding of how the different elements in thymine and guanine (specifically, the carbon, nitrogen, hydrogen, and oxygen rings) were configured.

Only upon the suggestion of American scientist Jerry Donohue did Watson decide to make new cardboard cutouts of the two bases, to see if perhaps a different atomic configuration would make a difference. It did. Not only did the complementary bases now fit together perfectly (i.e., A with T and C with G), with each pair held together by hydrogen bonds, but the structure also reflected Chargaff's rule.

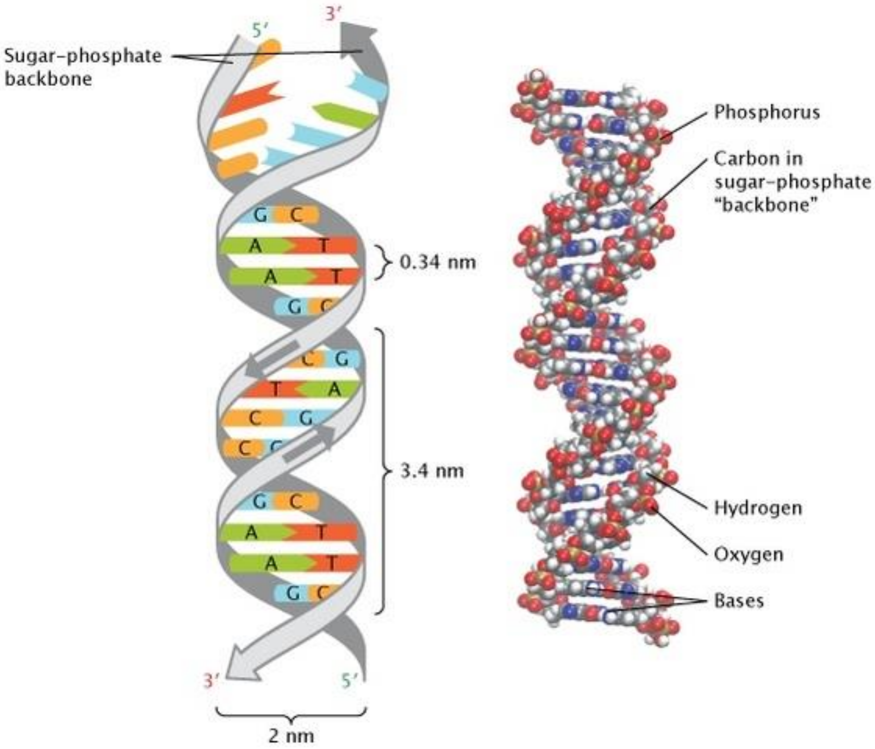


Fig 12.16 The double-helical structure of DNA,The 3- dimensional double helix structure of DNA, correctly elucidated by James Watson and Francis Crick. Complementary bases are held together as a pair by hydrogen bonds.

Although scientists have made some minor changes to the Watson and Crick model, or have elaborated upon it, since its inception in 1953, the model's four major features remain the same yet today. These features are as follows:

- DNA is a double-stranded helix, with the two strands connected by hydrogen bonds.
- A base is always paired with Ts, and Cs are always paired with Gs, which is consistent with and accounts for Chargaff's rule.
- Most DNA double helices are right-handed; that is, if you were to hold your right hand out, with your thumb pointed up and your fingers curled around your thumb, your thumb would represent the axis of the helix and your fingers would represent the sugar-phosphate backbone. Only one type of DNA, called Z-DNA, is left-handed.
- The DNA double helix is anti-parallel, which means that the 5' end of one strand is paired with the 3' end of its complementary strand (and vice versa).

- As shown in Figure 4, nucleotides are linked to each other by their phosphate groups, which bind the 3' end of one sugar to the 5' end of the next sugar.
- Not only are the DNA base pairs connected via hydrogen bonding, but the outer edges of the nitrogen-containing bases are exposed and available for potential hydrogen bonding as well.
- These hydrogen bonds provide easy access to the DNA for other molecules, including the proteins that play vital roles in the replication and expression of DNA.

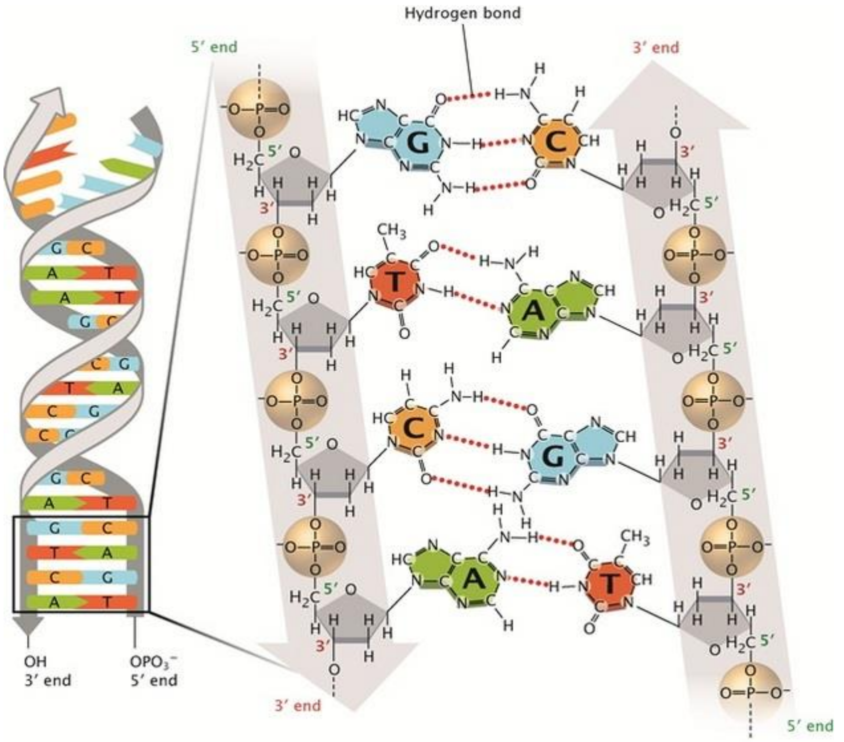


Fig 12.17 Base pairing in DNA

Two hydrogen bonds connect T to A; three hydrogen bonds connect G to C. The sugar-phosphate backbones (grey) run anti-parallel to each other, so that the 3' and 5' ends of the two strands are aligned.

One of the ways that scientists have elaborated on Watson and Crick's model is through the identification of three different conformations of the DNA double helix.

In other words, the precise geometries and dimensions of the double helix can vary. The most common conformation in most living cells (which is the one depicted in most diagrams of the double helix, and the one proposed by Watson and Crick) is known as B-DNA.

There are also two other conformations: A-DNA, a shorter and wider form that has been found in dehydrated samples of DNA and rarely under normal physiological circumstances; and Z-DNA, a left-handed conformation. Z-DNA is a transient form of DNA, only occasionally existing in response to certain types of biological activity.

Z-DNA was first discovered in 1979, but its existence was largely ignored until recently. Scientists have since discovered that certain proteins bind very strongly to Z-DNA, suggesting that Z-DNA plays an important biological role in protection against viral disease (Rich & Zhang, 2003).

Three different conformations of the DNA double helix

(A) A-DNA is a short, wide, right-handed helix. (B) B-DNA, the structure proposed by Watson and Crick, is the most common conformation in most living cells. (C) Z-DNA, unlike A- and B-DNA, is a left-handed helix.

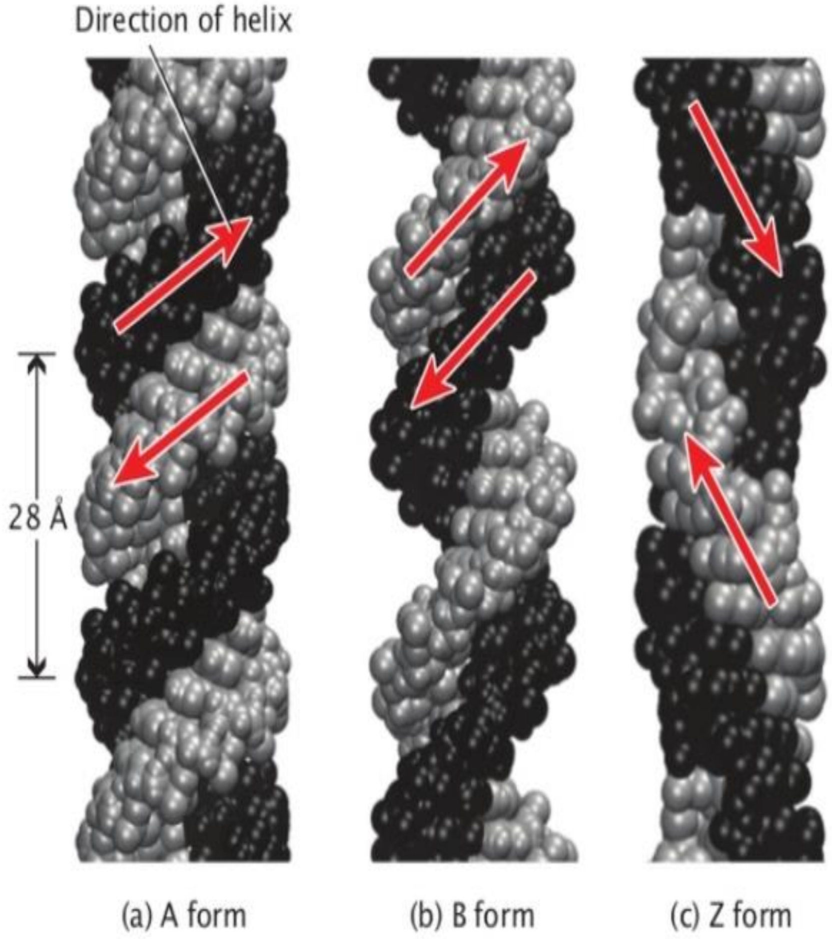
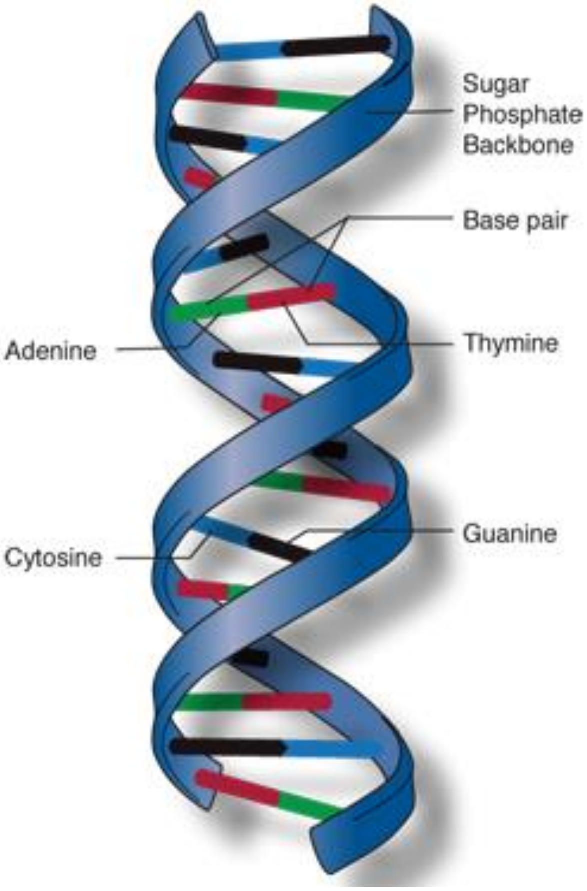


Fig.12.18 Three different conformations of the DNA double helix


Sugar
Phosphate
Backbone
Base pair
Adenine
Thymine
Cytosine
Guanine

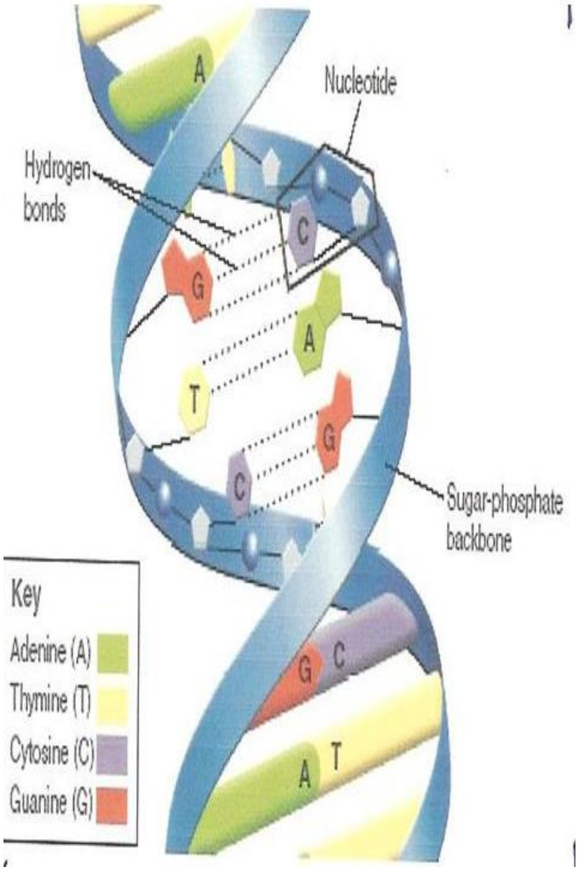


Fig.7.19 double Helical structure of DNA

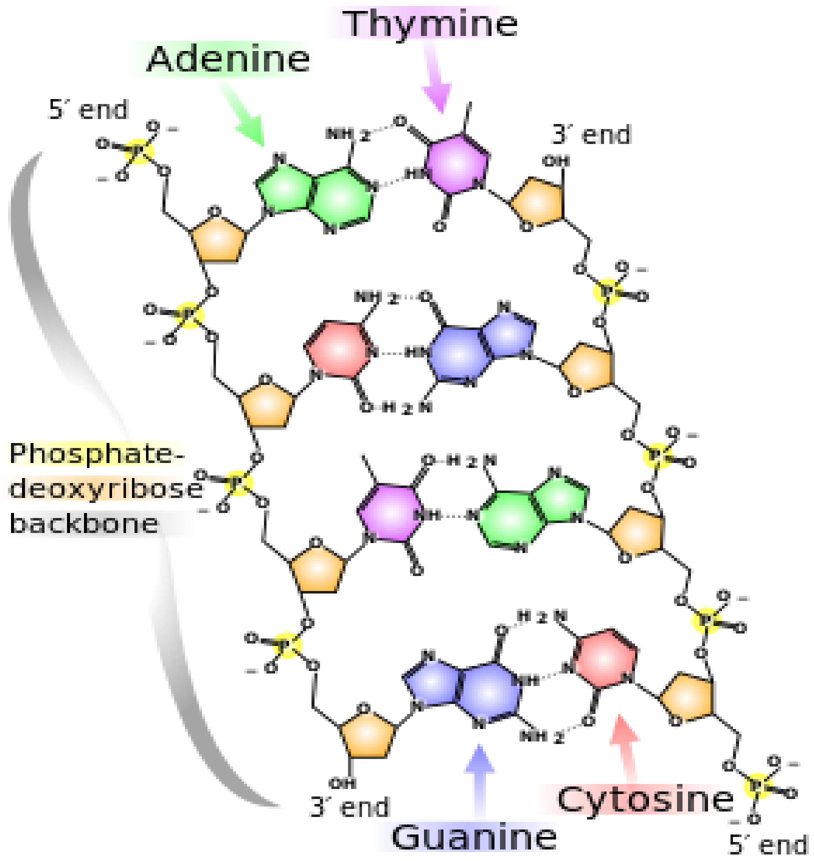


Fig.12.20 Chemical structure of DNA; hydrogen bonds shown as dotted lines

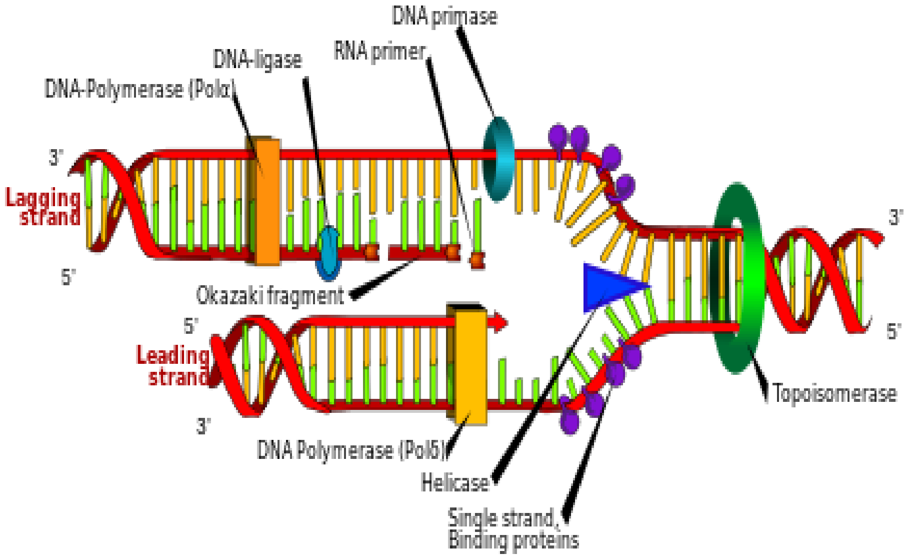


Fig.12.21 DNA replication

(The double helix is unwound by a helicase and to `poisomerase. Next, one DNA polymerase produces the leading strand copy. Another DNA polymerase binds to the lagging strand. This enzyme makes discontinuous segments (called Okazaki fragments) before DNA ligase joins them together.)

EXPERIMENT NO- 03

OBJECTIVE: Study of the Chromatography or Thin layer Chromatography (TLC)

INTRODUCTION

Thin-layer chromatography is performed on a sheet of glass, plastic, or aluminium foil, which is coated with a thin layer of adsorbent material, usually silica gel, aluminium oxide (alumina), or cellulose.

This layer of adsorbent is known as the stationary phase.

After the sample has been applied on the plate, a solvent or solvent mixture (known as the mobile phase) is drawn up the plate via capillary action. Because different analytes ascend the TLC plate at different rates, separation is achieved.

The mobile phase has different properties from the stationary phase.

For example, with silica gel, a very polar substance, non-polar mobile phases such as heptanes are used.

The mobile phase may be a mixture, allowing chemists to fine-tune the bulk properties of the mobile phase.

After the experiment, the spots are visualized. Often this can be done simply by projecting ultraviolet light onto the sheet; the sheets are treated with a phosphor, and dark spots appear on the sheet where compounds absorb the light impinging on a certain area.

Chemical processes can also be used to visualize spots; anisaldehyde, for example, forms colored adducts with many compounds, and sulfuric acid will char most organic compounds, leaving a dark spot on the sheet.

To quantify the results, the distance traveled by the substance being considered is divided by the total distance traveled by the mobile phase. (The mobile phase must not be allowed to reach the end of the stationary phase.)

This ratio is called the retention factor or R_f. In general, a substance whose structure resembles the stationary phase will have low R_f, while one that has a similar structure to the mobile phase will have high retention factor.

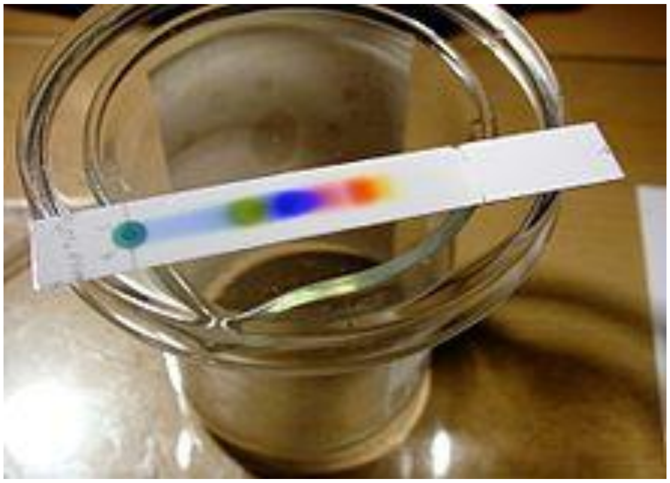
Retention factors are characteristic, but will change depending on the exact condition of the mobile and stationary phase. For this reason, chemists usually apply a sample of a known compound to the sheet before running the experiment.

Thin-layer chromatography can be used to monitor the progress of a reaction, identify compounds present in a given mixture, and determine the purity of a substance.

Specific examples of these applications include: analyzing ceramides and fatty acids, detection of pesticides or insecticides in food and water, analyzing the dye composition of fibers in forensics, assaying the radiochemical purity of radiopharmaceuticals, or identification of medicinal plants and their constituents

A number of enhancements can be made to the original method to automate the different steps, to increase the resolution achieved with TLC and to allow more accurate quantitative analysis.

This method is referred to as HPTLC, or "high-performance TLC". HPTLC typically uses thinner layers of stationary phase and smaller sample volumes, thus reducing the loss of resolution due to diffusion.



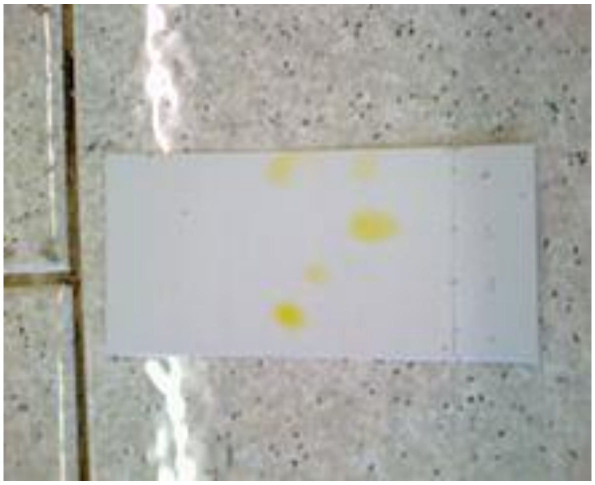


Fig.12.22 Separation of black ink on a TLC plate

TLC of three standards (ortho-, meta- and para-isomers) and a sample

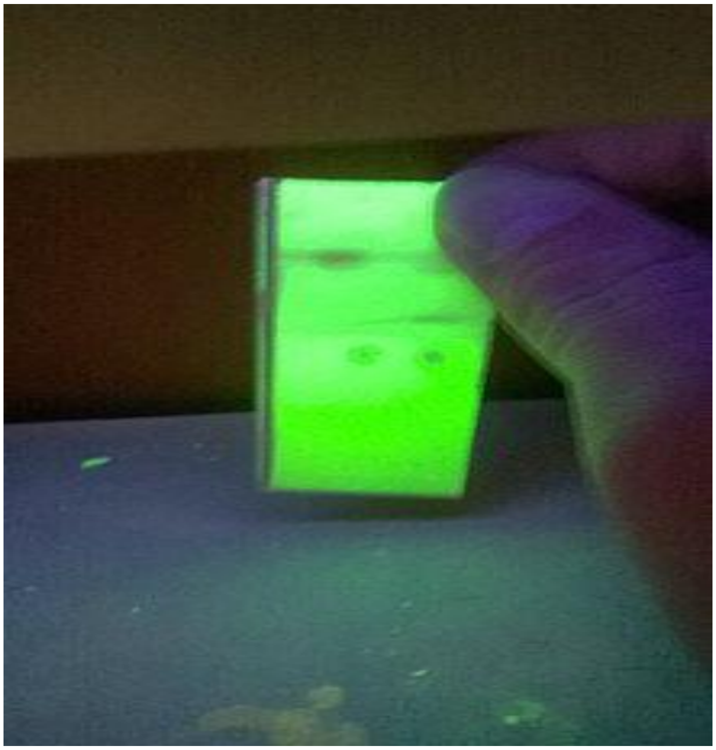


Fig.12.23 Fluorescent TLC plate under UV light

PLATE PREPARATION

TLC plates are usually commercially available, with standard particle size ranges to improve reproducibility.

They are prepared by mixing the adsorbent, such as silica gel, with a small amount of inert binder like calcium sulfate (gypsum) and water.

This mixture is spread as thick slurry on an uncreative carrier sheet, usually glass, thick aluminum foil, or plastic.

The resultant plate is dried and activated by heating in an oven for thirty minutes at 110 °C. The thickness of the absorbent layer is typically around 0.1 – 0.25 mm for analytical purposes and around 0.5 – 2.0 mm for preparative TLC.

TECHNIQUE

The process is similar to paper chromatography with the advantage of faster runs, better separations, and the choice between different stationary phases.

Because of its simplicity and speed TLC is often used for monitoring chemical reactions and for the qualitative analysis of reaction products.

Plates can be labeled before or after the chromatography process using a pencil or other implement that will not interfere or react with the process.

To run a thin layer chromatography plate, the following procedure is carried out:

Using a capillary, a small spot of solution containing the sample is applied to a plate, about 1.5 centimeters from the bottom edge.

The solvent is allowed to completely evaporate off to prevent it from interfering with sample's interactions with the mobile phase in the next step. If a non-volatile solvent was used to apply the sample, the plate needs to be dried in a vacuum chamber.

This step is often repeated to ensure there is enough analyte at the starting spot on the plate to obtain a visible result. Different samples can be placed in a row of spots the same distance from the bottom edge, each of which will move in its own adjacent lane from its own starting point.

- A small amount of an appropriate solvent (eluent) is poured into a glass beaker or any other suitable transparent container (separation chamber) to a depth of less than 1 centimeter.
- A strip of filter paper (aka "wick") is put into the chamber so that its bottom touches the solvent and the paper lies on the chamber wall and reaches almost to the top of the container. The container is closed with a cover glass or any other lid and is left for a few

minutes to let the solvent vapors ascend the filter paper and saturate the air in the chamber. (Failure to saturate the chamber will result in poor separation and non-reproducible results).

- The TLC plate is then placed in the chamber so that the spot(s) of the sample do not touch the surface of the eluent in the chamber, and the lid is closed. The solvent moves up the plate by capillary action, meets the sample mixture and carries it up the plate (elutes the sample).
- The plate should be removed from the chamber before the solvent front reaches the top of the stationary phase (continuation of the elution will give a misleading result) and dried.
- Without delay, the solvent front, the furthest extent of solvent up the plate, is marked.
- The plate is visualized. As some plates are pre-coated with a phosphor such as zinc sulfide, allowing many compounds to be visualized by using ultraviolet light; dark spots appear where the compounds block the UV light from striking the plate. Alternatively, plates can be sprayed or immersed in chemicals after elution. Various visualising agents react with the spots to produce visible results.

SEPARATION PROCESS AND PRINCIPLE

Different compounds in the sample mixture travel at different rates due to the differences in their attraction to the stationary phase and because of differences in solubility in the solvent.

By changing the solvent, or perhaps using a mixture, the separation of components (measured by the Rf value) can be adjusted. Also, the separation achieved with a TLC plate can be used to estimate the separation of a flash chromatography column. (A compound elutes from a column when the amount of solvent collected is equal to 1/Rf.) Chemists often use TLC to develop a protocol for separation by chromatography and they use TLC to determine which fractions contain the desired compounds.

Analysis

As the chemicals being separated may be colorless, several methods exist to visualize the spots:

- Fluorescent analytes like quinine may be detected under black light (366 nm)

- Often a small amount of a fluorescent compound, usually manganese-activated zinc silicate, is added to the adsorbent that allows the visualization of spots under UV-C light (254 nm). The adsorbent layer will thus fluoresce light-green by itself, but spots of analyte quench this fluorescence.
- Iodine vapors are a general unspecific color reagent
- Specific color reagents into which the TLC plate is dipped or which are sprayed onto the plate exist.
    - Potassium permanganate - oxidation
    - Bromine
- In the case of lipids, the chromatogram may be transferred to a PVDF membrane and then subjected to further analysis, for example mass spectrometry, a technique known as Far-Eastern blotting.

Once visible, the R_f value, or retardation factor, of each spot can be determined by dividing the distance the product traveled by the distance the solvent front traveled using the initial spotting site as reference. These values depend on the solvent used and the type of TLC plate and are not physical constants.

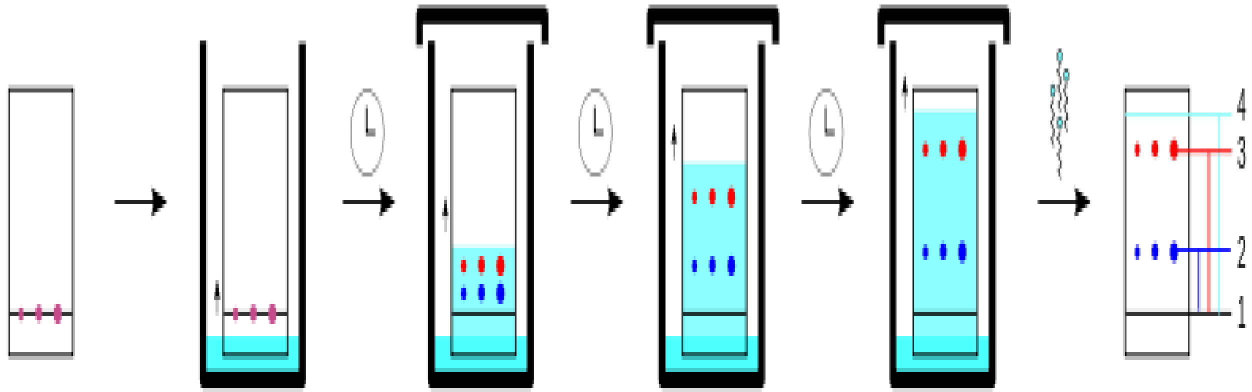


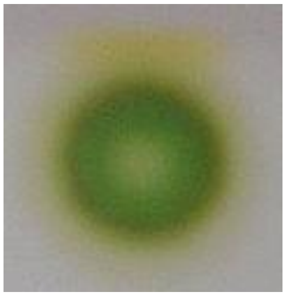
Fig.12.24 Development of a TLC plate, a purple spot separates into a red and blue spot

ISOLATION

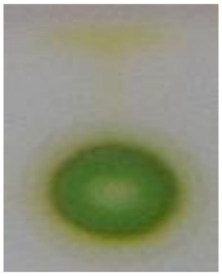
Since different compounds will travel a different distance in the stationary phase, chromatography can in effect be used as an isolation technique.

The separated compounds each occupying a specific area on the plate, they can be scraped away, put in another solvent to separate them from the stationary phase and used for further analysis.

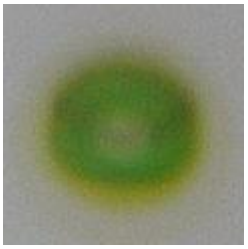
As an example, in the chromatography of an extract of green leaves (for example spinach) in 7 stages of development, Carotene elutes quickly and is only visible until step 2. Chlorophyll A and B are halfway in the final step and lutein the first compound staining yellow. Once the chromatography is over, the carotene can be removed from the plate, put back into a solvent and ran into a spectrophotometer to characterize its wavelength absorption.

- 

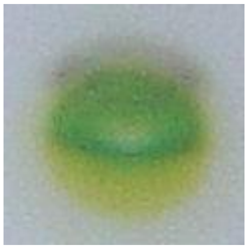
Step 1

- 

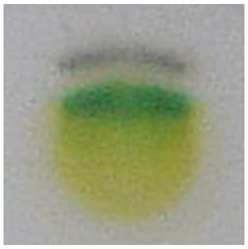
Step 2

- 

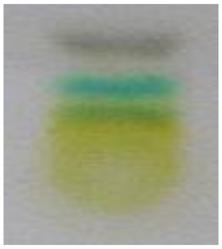
Step 3

- 

Step 4

- 

Step 5

- 

Step 6

-

Step 7

Fig.12.25 Isolation

EXPERIMENT NO- 04

***OBJECTIVE*:** Study of the Recombinant DNA techniques

INTRODUCTION

Recombinant DNA (rDNA) molecules are DNA molecules formed by laboratory methods of genetic recombination (such as molecular cloning) to bring together genetic material from multiple sources, creating sequences that would not otherwise be found in the genome.

Recombinant DNA is possible because DNA molecules from all organisms share the same chemical structure. They differ only in the nucleotide sequence within that identical overall structure.

Recombinant DNA is the general name for a piece of DNA that has been created by the combination of at least two strands.

Recombinant DNA molecules are sometimes called chimeric DNA, because they can be made of material from two different species, like the mythical chimera. R-DNA technology uses palindromic sequences and leads to the production of sticky and blunt ends.

The DNA sequences used in the construction of recombinant DNA molecules can originate from any species.

For example, plant DNA may be joined to bacterial DNA, or human DNA may be joined with fungal DNA.

In addition, DNA sequences that do not occur anywhere in nature may be created by the chemical synthesis of DNA, and incorporated into recombinant molecules.

Using recombinant DNA technology and synthetic DNA, literally any DNA sequence may be created and introduced into any of a very wide range of living organisms.

Proteins that can result from the expression of recombinant DNA within living cells are termed recombinant proteins.

When recombinant DNA encoding a protein is introduced into a host organism, the recombinant protein is not necessarily produced.

Expression of foreign proteins requires the use of specialized expression vectors and often necessitates significant restructuring by foreign coding sequences]

Recombinant DNA differs from genetic recombination in that the former results from artificial methods in the test tube, while the latter is a normal biological process that results in the remixing of existing DNA sequences in essentially all organisms.

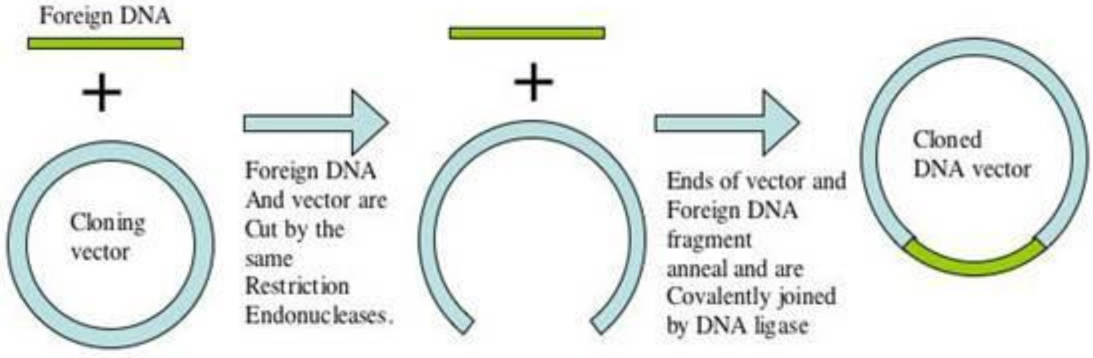

Foreign DNA
+
Cloning vector
Foreign DNA And vector are Cut by the same Restriction Endonucleases.
+
Ends of vector and Foreign DNA fragment anneal and are Covalently joined by DNA ligase
Cloned DNA vector

Fig.12.26 Construction of recombinant DNA, in which a foreign DNA fragment is inserted into a plasmid vector.In this example, the gene indicated by the white color is inactivated upon insertion of the foreign DNA fragment.

PROPERTIES OF ORGANISMS CONTAINING RECOMBINANT DNA

In most cases, organisms containing recombinant DNA have apparently normal phenotypes.

That is, their appearance, behavior and metabolism are usually unchanged, and the only way to demonstrate the presence of recombinant sequences is to examine the DNA itself, typically using a polymerase chain reaction (PCR) test. Significant exceptions exist, and are discussed below.

If the rDNA sequences encode a gene that is expressed, then the presence of RNA and/or protein products of the recombinant gene can be detected, typically using RT-PCR or western hybridization methods.

Gross phenotypic changes are not the norm, unless the recombinant gene has been chosen and modified so as to generate biological activity in the host organism.

Additional phenotypes that are encountered include toxicity to the host organism induced by the recombinant gene product, especially if it is over-expressed or expressed within inappropriate cells or tissues.

In some cases, recombinant DNA can have deleterious effects even if it is not expressed. One mechanism by which this happens is insertional inactivation, in which the rDNA becomes inserted into a host cell's gene.

In some cases, researchers use this phenomenon to "knock out" genes to determine their biological function and importance. Another mechanism by which rDNA insertion into chromosomal DNA can affect gene expression is by inappropriate activation of previously unexpressed host cell genes.

This can happen, for example, when a recombinant DNA fragment containing an active promoter becomes located next to a previously silent host cell gene, or when a host cell gene that functions to restrain gene expression undergoes insertional inactivation by recombinant DNA.

Uses:

Recombinant DNA is widely used in biotechnology, medicine and research. Today, recombinant proteins and other products that result from the use of DNA technology are found in essentially every western pharmacy, doctors or veterinarian's office, medical testing laboratory, and biological research laboratory.

The most common application of recombinant DNA is in basic research, in which the technology is important to most current work in the biological and biomedical sciences.

Recombinant DNA is used to identify, map and sequence genes, and to determine their function. rDNA probes are employed in analyzing gene expression within individual cells, and throughout the tissues of whole organisms.

Recombinant proteins are widely used as reagents in laboratory experiments and to generate antibody probes for examining protein synthesis within cells and organisms.

Many additional practical applications of recombinant DNA are found in industry, food production, human and veterinary medicine, agriculture, and bioengineering. Some specific examples are identified below.

Recombinant chymosin: Found in rennet, chymosin is an enzyme required to manufacture cheese.

Recombinant human insulin: Almost completely replaced insulin obtained from animal sources (e.g. pigs and cattle) for the treatment of insulin-dependent diabetes.

Recombinant human growth hormone (HGH, somatotropin): Administered to patients whose pituitary glands generate insufficient quantities to support normal growth and development.

Recombinant blood clotting factor VIII: A blood-clotting protein that is administered to patients with forms of the bleeding disorder hemophilia, who are unable to produce factor VIII in quantities sufficient to support normal blood coagulation.

Recombinant hepatitis B vaccine: Hepatitis B infection is controlled through the use of a recombinant hepatitis B vaccine, which contains a form of the hepatitis B virus surface antigen that is produced in yeast cells.

Diagnosis of infection with HIV: Each of the three widely used methods for diagnosing HIV infection has been developed using recombinant DNA.

Golden rice: A recombinant variety of rice that has been engineered to express the enzymes responsible for β-carotene biosynthesis.

Herbicide-resistant crops: Commercial varieties of important agricultural crops (including soy, maize/corn, sorghum, canola, alfalfa and cotton) have been developed that incorporate a recombinant gene that results in resistance to the herbicide glyphosate (trade name *Roundup*), and simplifies weed control by glyphosate application.

Insect-resistant crops: *Bacillus thuringeiensis* is a bacterium that naturally produces a protein (Bt toxin) with insecticidal properties. The bacterium has been applied to crops as an insect-control strategy for many years, and this practice has been widely adopted in agriculture and gardening.

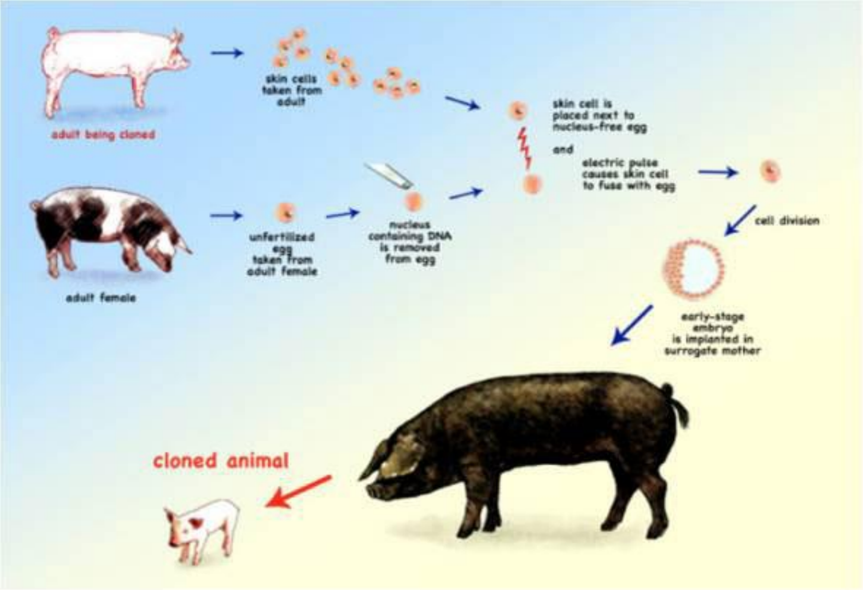

adult being cloned
skin cells taken from adult
skin cell is placed next to nucleus-free egg
and
electric pulse causes skin cell to fuse with egg
adult female
unfertilized egg taken from adult female
nucleus containing DNA is removed from egg
cell division
early-stage embryo is implanted in surrogate mother
cloned animal

Fig.12.26 Diagram of pigs to show how animal cloning is carried out

REFERENCES

- Rosano, Germán L.; Ceccarelli, Eduardo A. (2014-04-17). "Recombinant protein expression in Escherichia coli: advances and challenges". Frontiers in Microbiology.
- Hannig, G.; Makrides, S. (1998). "Strategies for optimizing heterologous protein expression in Escherichia coli". Trends in Biotechnology. 16 (2): 54–60. doi:10.1016/S0167-7799(97)01155-4. PMID 9487731.
- Brown, Terry (2006). Gene Cloning and DNA Analysis: an Introduction. Cambridge, MA: Blackwell Pub. ISBN 1-4051-1121-6.
- Dame RT (May 2005). "The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin". Molecular Microbiolo.
- Biocompare (2016). Microplate readers. Retrieved from January 22, 2015. http://www.biocompare.com/Lab-Equipment/20131-Microplate-Reader-ELISA-Plate-Reader/

- Susan R. Mikkelsen & Eduardo Cortón. Bioanalytical Chemistry, Ch. 13. Centrifugation Methods. John Wiley & Sons, Mar 4, 2004, pp. 247-267.
- Harry W. Lewis & Christopher J. Moody (13 Jun 1989). Experimental Organic Chemistry: Principles and Practice (Illustrated ed.). WileyBlackwell. pp. 159–173. ISBN 978-0-632-02017-1.
- Reich, E.; Schibli A. (2007). High-performance thin-layer chromatography for the analysis of medicinal plants (Illustrated ed.). New York: Thieme. ISBN 3-13-141601-7.

UNIT 13 BIOTECHNOLOGY /BIOTECHNIQUE EXERCISEPART- (II)

Content

Experiment No- 01

OBJECTIVE: Study of prepared slides, models or specimen

INTRODUCTION

Modern biotechnology techniques are based only on the isolation and manipulation of DNA these days. These techniques made it possible to detect diseases and make certain precautionary measures to protect people from these diseases. Some of the techniques are described in this article

DNA Isolation:

DNA isolation techniques totally depend on the experimental organism, but somehow these techniques have following characteristics.

1. A treatment for opening the cells and releasing their DNA
2. Method for inactivating and removing the enzymes which degrade the DNA
3. Method to separate the DNA from proteins and other molecules which contaminate the DNA

4. There are many cheap laboratory exercises for the isolation of DNA which are also very simple. DNA is usually taken from onion, bacteria and other organisms. Heat and detergent is used to isolate the DNA.

These methods break open the cell and inactivate the enzymes which degrade DNA. DNA is separated from other contaminants by using alcohol precipitation. The method of DNA isolation is more reproducible is isolation the DNA from bacteria16.

There are many companies who sell kits to for DNA isolation. DNA is isolated from human body cells to detect genetic diseases in the newborn child, to analyze the forensic evidence and to study genes which are involved in causing cancer.

DNA Transformation

Over the past couple of decades biotechnologists have succeeded in developing DNA transfer techniques.

It is the only the DNA which is required for the transformation. To get the DNA, in most cases tissues are exposed to stresses which enable the plasma membrane to rupture and in other cases cell wall of the plants is used.

Through this process some of the cells will contain the foreign DNA in their genome. This foreign DNA is helpful to make useful products and has been used in pharmaceuticals such as human insulin.

Restriction Enzyme Digestion and Analysis:

Restriction enzymes are actually the proteins. They cut the sequences of bases of DNA at specific places.

Biotechnologists are able to reproduce cut DNA molecules into small well defined fragments through these enzymes.

In biotechnology, restriction enzymes are used to make and analyze new DNA molecules. Agarose gel electrophoresis is the technique through which biotechnologists are able to make new DNA molecules. Agarose gel makes new DNA molecules through following methods
1) Chains of sugars make agarose.
2) DNA is moved in the agarose gel through electric current and this current cuts them into smaller pieces according to their size.
3) Special kind of dye makes possible the identification of small DNA molecules in the gel.

Polymerase chain reaction:

Polymerase chain reaction or PCR is a common technique in biotechnology. This technique is used for producing multiple copies of short DNA molecules. PCR runs through following process.

1) DNA molecule that will act as a template in the reaction
2) Short and single stranded DNA molecules which are called as primers. Primers are the starting points for making new DNA molecule.
3)An enzyme is needed which makes DNA
4)Nucleotides which are the building blocks of DNA
5) Thermal cycler, a machine, which subjects the reaction mixture to varying temperatures repeatedly.

DNA SEQUENCING:

DNA sequencing is the most important thing in the human genome project. Many biotechnology techniques are dependent on this method.

To determine the exact order of bases in the DNA molecule is this technique's purpose. Nature of protein encoded in the DNA molecule can be determined through this technique.

Microarrays: -

It is a very useful technique of biotechnology. In this technique, expression of many genes can be analyzed simultaneously.

Microarrays can be made on a glass slide or a chip and they contain hundreds and thousands of parts of genes in a small section of the slide.

Expressed genes can be isolated from the cells and then are detected. This detection shows that which DNAs bind to which genes and are taken from which cells.

8.1. STUDY OF PREPARED SLIDES, MODELS OR SPECIMEN

Biotechnology is not mere facts, terms, and concepts. In the present era of biological research these two applied subjects are making significant contributions towards scientific knowledge and solving human problems.

In order to be an accomplished biotechnologist a thorough knowledge of Microbiology is essential. Microorganisms were probably the first living forms to appear on the earth.

Microorganisms are omnipresent in biosphere. Most microorganisms are free living and perform useful activities. Biosphere in general comprises of two categories: animate or living and inanimate or non-living beings.

Biology is the science that deals with the living organisms. The branch of biology that deals with microscopic forms of living organisms (microbes) is termed as the "microbiology".

The emergence and development of microbiology has been highly erratic in the beginning. Credit for introduction to microbial world goes to Antony van Leeuwenhoek and to its as of microbial activities to Louis Pasteur and Robert Koch.

Microbiology considers the microscopic forms of life as to their occurrence in nature.

Their reproduction and physiology, their harmful and beneficial relationship with other living things and their significance in science and industry

The methods used to study bacteria and other l1licrobes include direct microscopic examination, cultivation, biochemical tests, animal inoculation, serological reactivity and recent molecular biological techniques.

Currently because of their high versatility and rapid turnover microorganisms are used extensively) in genetic engineering and biotechnological processes and new processes are being developed to produce variety of enzymes of industrial importance, vaccines, insecticides, pharmaceuticals and other biological products of interest to mankind, animals and plants.

STUDY OF PREPARED SLIDES:

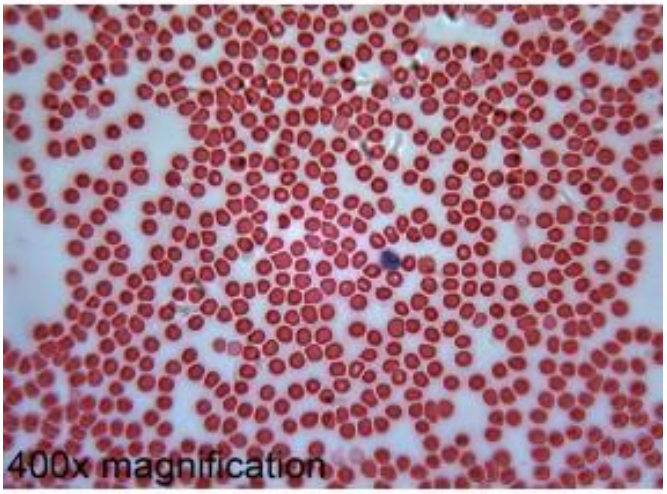


Fig 13.1 Smear of human blood showing both erythrocytes (red blood cells) and different types of leukocytes (white blood cells). Erythrocytes do not have a cell nucleus. Stained with Wright's blood stain

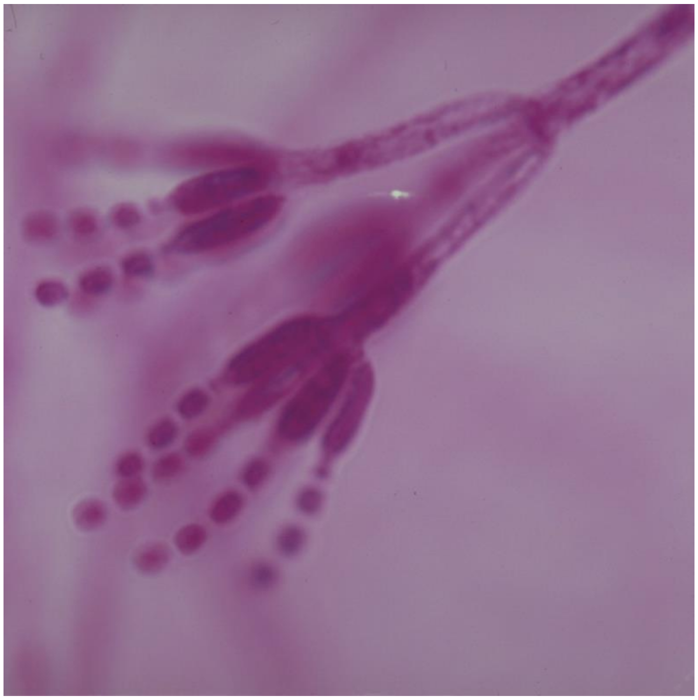
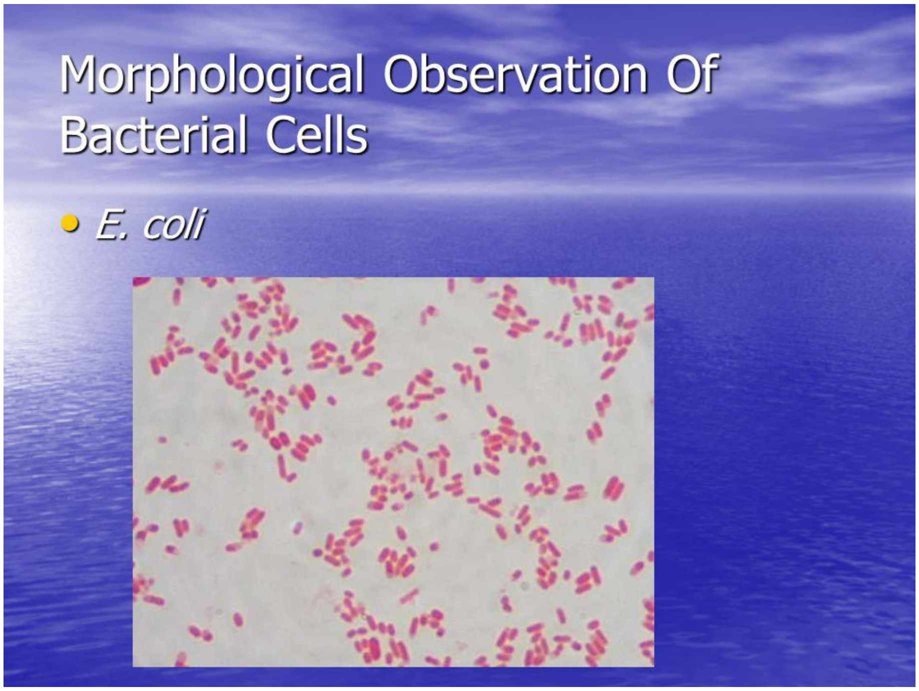


Fig.13.2 Shows mycelium and conidiophores with conidia



Morphological Observation Of Bacterial Cells

- *E. coli*

Fig.13.3 Microscopic view of Escherichia coli

EXPERIMENT NO-02

OBJECTIVE: Study of *Escherichia coli*

INTRODUCTION

Escherichia coli, a normal inhabitant of the human intestinal tract, is the most thoroughly studied of all organisms.

Studies of the mechanisms of genetic exchange and the biology of plasmids and bacteriophages of E. coli have been crucial in understanding many aspects of DNA replication and the expression of genetic material.

These studies have led to the ability to insert DNA from unrelated organisms into E. coli plasmids and bacteriophages, and to have that DNA replicated by the bacteria, with the genetic information it contains expressed by the bacteria.

It is thus possible for bacteria to become living factories for scarce biological products such as human insulin, interferon, and growth hormone. This process is called genetic engineering.

ROLE IN BIOTECHNOLOGY

Because of its long history of laboratory culture and ease of manipulation, *E. coli* plays an important role in modern biological engineering and industrial microbiology.

The work of Stanley Norman Cohen and Herbert Boyer in *E. coli*, using plasmids and restriction enzymes to create recombinant DNA, became a foundation of biotechnology.

E. coli is a very versatile host for the production of heterologous proteins, and various protein expression systems have been developed which allow the production of recombinant proteins in *E. coli*.

Researchers can introduce genes into the microbes using plasmids which permit high level expression of protein, and such protein may be mass-produced in industrial fermentation processes.

One of the first useful applications of recombinant DNA technology was the manipulation of *E. coli* to produce human insulin.

Many proteins previously thought difficult or impossible to be expressed in *E. coli* in folded form have been successfully expressed in *E. coli*.

For example, proteins with multiple disulphide bonds may be produced in the periplasmic space or in the cytoplasm of mutants rendered sufficiently oxidizing to allow disulphide-bonds to form, while proteins requiring post-translational modification such as glycosylation for stability or function have been expressed using the N-linked glycosylation system of *Campylobacter jejuni* engineered into *E. coli*.

Modified *E. coli* cells have been used in vaccine development, bioremediation, production of biofuels, lighting, and production of immobilised enzymes.

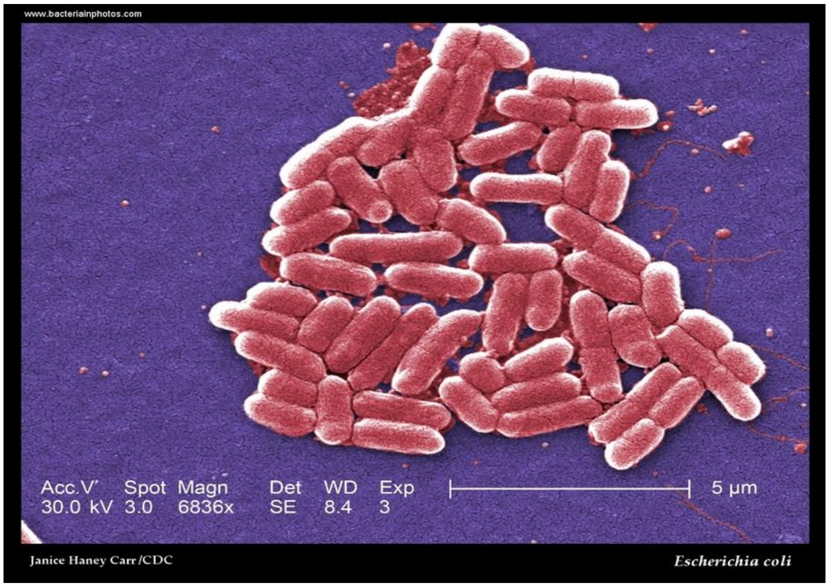


Fig.13.4Escherichia coli (strain O157:H7)

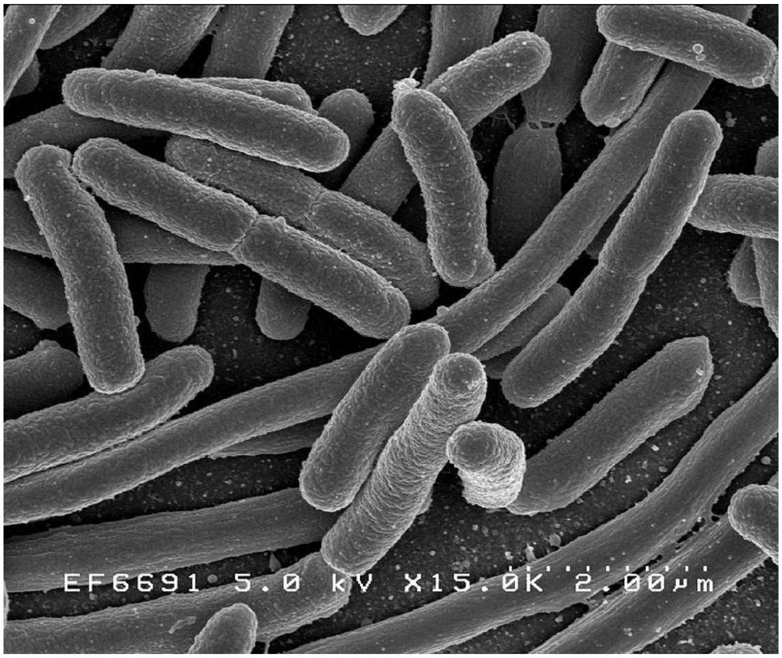


Fig.13..4 E. coli Bacteria

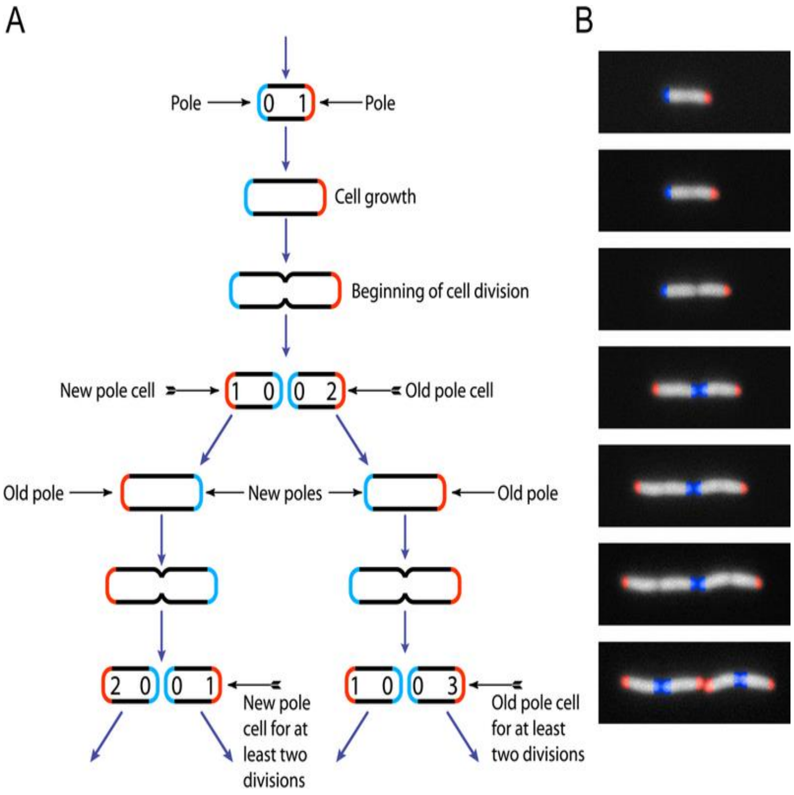

A
B
Pole
Pole
0 1
Cell growth
Beginning of cell division
New pole cell
1 0
0 2
Old pole cell
Old pole
New poles
Old pole
2 0
0 1
New pole cell for at least two divisions
1 0
0 3
Old pole cell for at least two divisions

Fig.13.4 Successive Binary fission in E. coli

CULTURE GROWTH

Optimum growth of *E. coli* occurs at 37 °C (98.6 °F), but some laboratory strains can multiply at temperatures up to 49 °C (120 °F).

E. coli grows in a variety of defined laboratory media, such as lysogeny broth, or any medium that contains glucose, ammonium phosphate, monobasic, sodium chloride, magnesium sulfate, potassium phosphate, dibasic, and water.

Growth can be driven by aerobic or anaerobic respiration, using a large variety of redox pairs, including the oxidation of pyruvic acid, formic acid, hydrogen, and amino acids, and the reduction of substrates such as oxygen, nitrate, fumarate, dimethyl sulfoxide, and trimethylamine N-oxide. *E. coli* is classified as a facultative anaerobe.

It uses oxygen when it is present and available. It can, however, continue to grow in the absence of oxygen using fermentation or anaerobic respiration. The ability to continue growing in the absence of oxygen is an advantage to bacteria because their survival is increased in environments where water predominates.

Exercise 1: Determination of size of bacteria

Objective: To determine the size of bacteria.

The size of microscopic objects including bacteria is expressed in microns (I 0-6m) or nanometers (I 0-9m). Such measurements are done with the ocular micrometer and a stage micrometer (for calibrating ocular micrometer).

Ocular micrometer is placed in the ocular region of the eyepiece. The ruled divisions superimposing specific distance on stage micrometer are counted.

By determining the number of divisions of the ocular micrometer that superimpose a known distance marked on the stage micrometer, one is able to calculate precisely the distance of each division on ocular micrometer.

After calibration, the ocular micrometer can be used for determining the size of various microscopic objects. The size of bacteria is normally determined in viable stained state (intra-vital staining).

Requirements

a. Bacterial culture: *E.coli*

b. Ocular micrometer and stage micrometer

c. Intra-vital stain (crystal violet 1: 120000).

Procedure

1. Remove the ocular lens and insert the ocular micrometer in ocular tube and replace the ocular lens and mount the eyepiece into the optical tube.

2. Mount the stage micrometer on the microscope stage.

3. Center the scale of the stage micrometer (with low power objective in position) while observing through eyepiece.

4. Bring oil immersion objective into position for observation.

5. Rotate the ocular micrometer containing eye piece so that the lines on it superimpose upon the stage micrometer divisions. Now make the lines of two micrometers coincide at one end.

6. Count the number of ocular micrometer divisions coinciding with stage micrometer exactly. Each stage micrometer corresponds to 10 microns.

7. Replace the stage micrometer with bacterial smears and count the number of divisions in the ocular scale that cover the bacterium.

8. Focus under oil immersion objective and record the observations for calculating the size of bacteria.
9. Measure the size of 5-6 different cells to find the average size of bacterial cell.

EXPERIMENT NO-03

OBJECTIVE: Study of Bacteriophage

INTRODUCTION

A **bacteriophage** is a virus that infects and replicates within a bacterium. Bacteriophages are composed of proteins that encapsulate a DNA or RNA genome, and may have relatively simple or elaborate structures.

Their genomes may encode as few as four genes, and as many as hundreds of genes. Phages replicate within the bacterium following the injection of their genome into its cytoplasm.

Bacteriophages are among the most common and diverse entities in the biosphere.

Phages are widely distributed in locations populated by bacterial hosts, such as soil or the intestines of animals.

One of the densest natural sources for phages and other viruses is sea water, where up to 9×10^8 virions per milliliter have been found in microbial mats at the surface, and up to 70% of marine bacteria may be infected by phages.

They have been used for over 90 years as an alternative to antibiotics in the former Soviet Union and Central Europe, as well as in France.

They are seen as a possible therapy against multi-drug-resistant strains of many bacteria Nevertheless, phages of Inoviridae have been shown to complicate biofilms involved in pneumonia and cystic fibrosis, shelter the bacteria from drugs meant to eradicate disease and promote persistent infection.

HISTORY

In 1896, Ernest Hanbury Hankin reported that something in the waters of the Ganges and Yamuna rivers in India had marked antibacterial action against cholera and could pass through a very fine porcelain filter. In 1915, British bacteriologist Frederick Twort, superintendent of the Brown Institution of London, discovered a small agent that infected and killed bacteria. He believed the agent must be one of the following:

1. A stage in the life cycle of the bacteria;
2. An enzyme produced by the bacteria themselves; or
3. A virus that grew on and destroyed the bacteria.

Replication

Bacteriophages may have a lytic cycle or a lysogenic cycle, and a few viruses are capable of carrying out both. With *lytic phages* such as the T4 phage, bacterial cells are broken open (lysed) and destroyed after immediate replication of the virion.

As soon as the cell is destroyed, the phage progeny can find new hosts to infect. Lytic phages are more suitable for phage therapy.

Some lytic phages undergo a phenomenon known as lysis inhibition, where completed phage progeny will not immediately lyse out of the cell if extracellular phage concentrations are high.

This mechanism is not identical to that of temperate phage going dormant and is usually temporary.

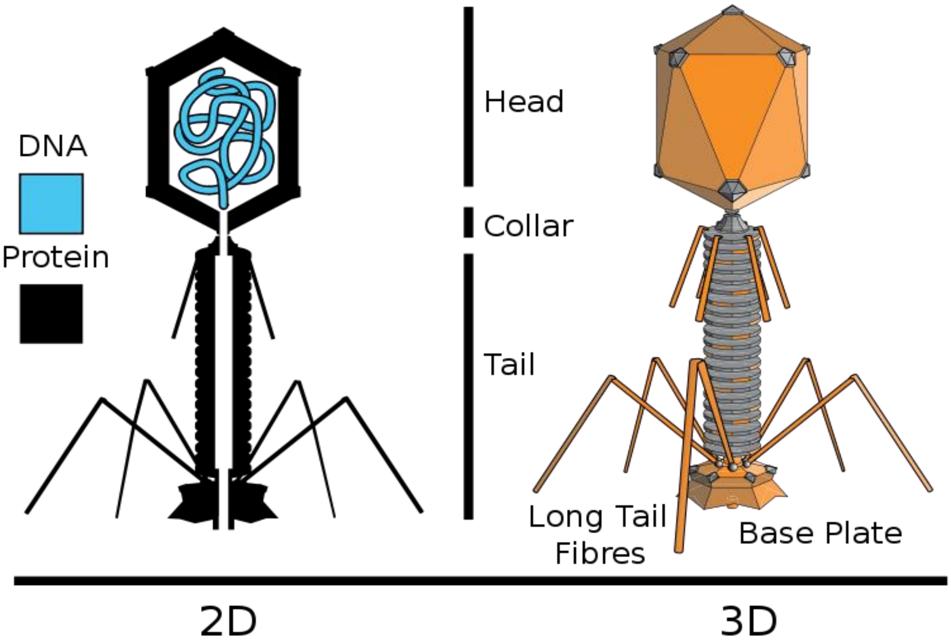


Fig.13.4 Typical tailed bacteriophage structure

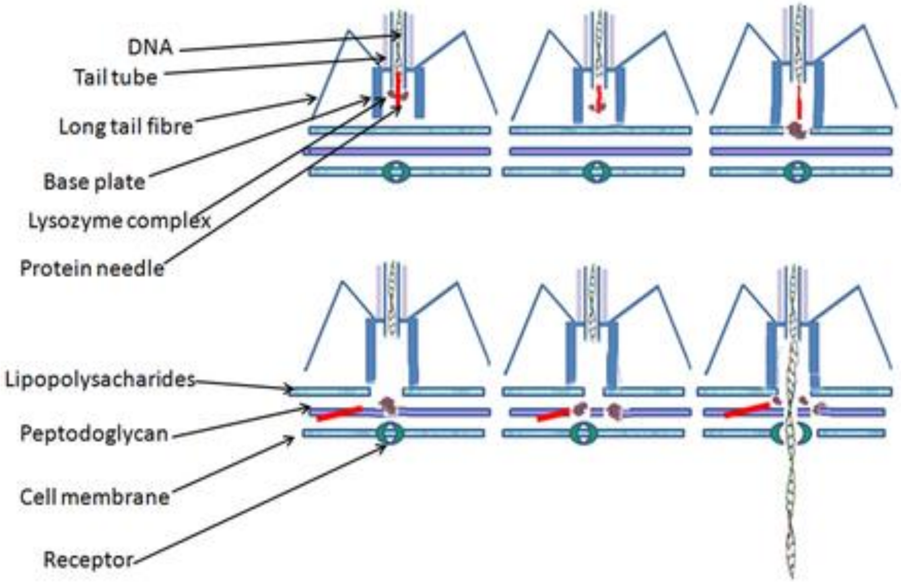


Fig.13.5 Diagram of the DNA injection process

EXPERIMENT NO-05

OBJECTIVE: Study of Plasmid

INTRODUCTION

A **plasmid** is a small DNA molecule within a cell that is physically separated from a chromosomal DNA and can replicate independently.

They are most commonly found in bacteria as small circular, double-stranded DNA molecules; however, plasmids are sometimes present in archaea and eukaryotic organisms.

In nature, plasmids often carry genes that may benefit the survival of the organism, for example antibiotic resistance.

While the chromosomes are big and contain all the essential genetic information for living under normal conditions, plasmids usually are very small and contain only additional genes that may be useful to the organism under certain situations or particular conditions.

Artificial plasmids are widely used as vectors in molecular cloning, serving to drive the replication of recombinant DNA sequences within host organisms.

Plasmids are considered *replicons*, a unit of DNA capable of replicating autonomously within a suitable host. However, plasmids, like viruses, are not generally classified as life.

Plasmids can be transmitted from one bacterium to another (even of another species) via three main mechanisms: transformation, transduction, and conjugation.

This host-to-host transfer of genetic material is called horizontal gene transfer, and plasmids can be considered part of the mobilome.

Unlike viruses (which encase their genetic material in a protective protein coat called a capsid), plasmids are "naked" DNA and do not encode genes necessary to encase the genetic material for transfer to a new host.

However, some classes of plasmids encode the conjugative "sex" pilus necessary for their own transfer.

The size of the plasmid varies from 1 to over 200 kbp, and the number of identical plasmids in a single cell can range anywhere from one to thousands under some circumstances.

The relationship between microbes and plasmid DNA is neither parasitic nor mutualistic, because each implies the presence of an independent species living in a detrimental or commensal state with the host organism. Rather, plasmids provide a mechanism for horizontal gene transfer within a population of microbes and typically provide a selective advantage under a given environmental state.

Plasmids may carry genes that provide resistance to naturally occurring antibiotics in a competitive environmental niche, or the proteins produced may act as toxins under similar circumstances, or allow the organism to utilize particular organic compounds that would be advantageous when nutrients are scarce.

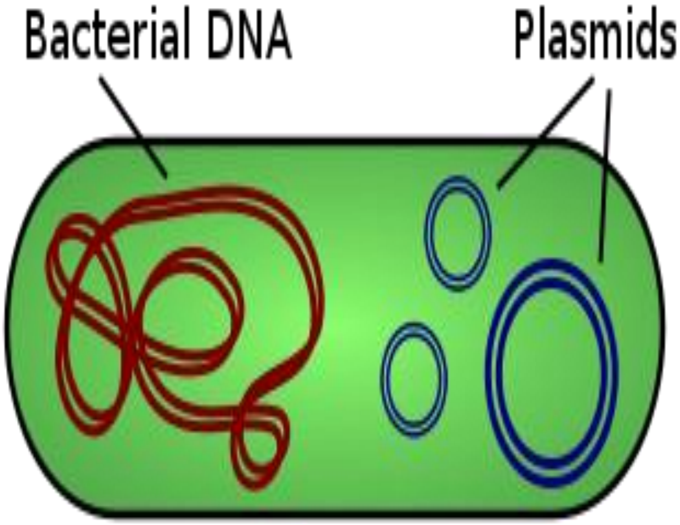


Fig.13.6 Illustration of a bacterium showing chromosomal DNA and plasmids

PROPERTIES AND CHARACTERISTICS

There are two types of plasmid integration into a host bacterium: Non-integrating plasmids replicate as with the top instance, whereas episomes, the lower example, can integrate into the host chromosome.

In order for plasmids to replicate independently within a cell, they must possess a stretch of DNA that can act as an origin of replication.

The self-replicating unit, in this case the plasmid, is called a replicon. A typical bacterial replicon may consist of a number of elements, such as the gene for plasmid-specific replication initiation protein (Rep), repeating units called iterons, DnaA boxes, and an adjacent AT-rich region.

Smaller plasmids make use of the host replicative enzymes to make copies of themselves, while larger plasmids may carry genes specific for the replication of those plasmids.

A few types of plasmids can also insert into the host chromosome, and these integrative plasmids are sometimes referred to as episomes in prokaryotes.

Plasmids almost always carry at least one gene. Many of the genes carried by a plasmid are beneficial for the host cells, for example: enabling the host cell to survive in an environment that would otherwise be lethal or restrictive for growth.

Some of these genes encode traits for antibiotic resistance or resistance to heavy metal, while others may produce virulence factors that enable a bacterium to colonize a host and overcome its defences, or have specific metabolic functions that allow the bacterium to utilize a particular nutrient, including the ability to degrade recalcitrant or toxic organic compounds. Plasmids can also provide bacteria with the ability to fix nitrogen.

Some plasmids, however, have no observable effect on the phenotype of the host cell or its benefit to the host cells cannot be determined, and these plasmids are called cryptic plasmids.

Naturally occurring plasmids vary greatly in their physical properties. Their size can range from very small mini-plasmids of less than a 1 kilobase pairs (Kbp), to very large megaplasmids of several megabase pairs (Mbp).

At the upper end, little can differentiate between a megaplasmid and a minichromosome. Plasmids are generally circular, however examples of linear plasmids are also known. These linear plasmids require specialized mechanisms to replicate their ends.

Plasmids may be present in an individual cell in varying number, ranging from one to several hundreds. The normal number of copies of plasmid that may be found in a single cell is called the copy number, and is determined by how the replication initiation is regulated and the size of the molecule. Larger plasmids tend to have lower copy numbers.

Low-copy-number plasmids that exist only as one or a few copies in each bacterium are, upon cell division, in danger of being lost in one of the segregating bacteria.

Such single-copy plasmids have systems that attempt to actively distribute a copy to both daughter cells. These systems, which include the parABS system and parMRC system, are often referred to as the partition system or partition function of a plasmid.

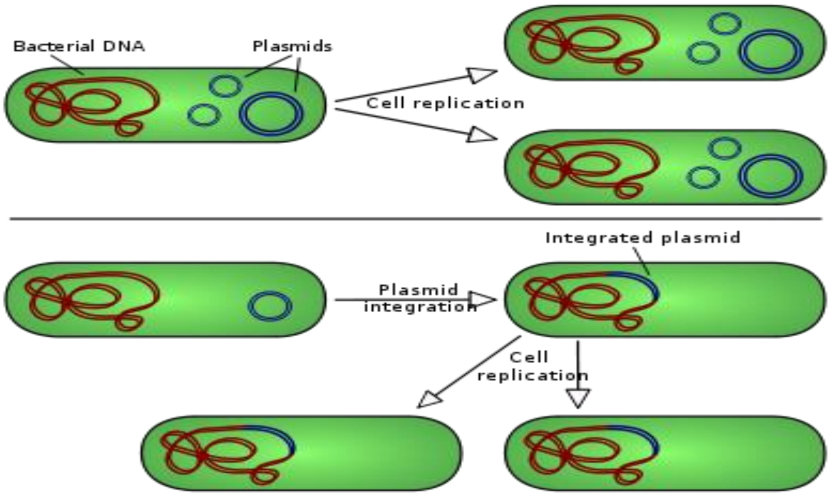


Fig.13.6 there are two types of plasmid integration into a host bacterium: Non-integrating plasmids replicate as with the top instance, whereas episomes, the lower example, can integrate into the host chromosome

EXPERIMENT NO- 06

***OBJECTIVE*:** To identify the Southern Blotting Principle, Procedure and Application

A **Southern blot** is a method used in molecular biology for detection of a specific DNA sequence in DNA samples. Southern blotting combines transfer of electrophoresis-separated DNA fragments to a filter membrane and subsequent fragment detection by probe hybridization.

The method is named after its inventor, the British biologist Edwin Southern.Other blotting methods (i.e., western blot, northern blot, eastern blot, southwestern blot) that employ similar principles, but using RNA or protein, have later been named in reference to Edwin Southern's name.

As the label is eponymous, Southern is capitalised, as is conventional for proper nouns. The names for other blotting methods may follow this convention, by analogy.

METHOD

1. Restriction endonucleases are used to cut high-molecular-weight DNA strands into smaller fragments.
2. The DNA fragments are then electrophoreses on an agarose gel to separate them by size.
3. If some of the DNA fragments are larger than 15 kb, then prior to blotting, the gel may be treated with an acid, such as dilute HCl. This depurinates the DNA fragments, breaking the DNA into smaller pieces, thereby allowing more efficient transfer from the gel to membrane.
4. If alkaline transfer methods are used, the DNA gel is placed into an alkaline solution (typically containing sodium hydroxide) to denature the double-stranded DNA.
5. The denaturation in an alkaline environment may improve binding of the negatively charged thymine residues of DNA to a positively charged amino groups of membrane, separating it into single DNA strands for later hybridization to the probe (see below), and destroys any residual RNA that may still be present in the DNA.
6. The choice of alkaline over neutral transfer methods, however, is often empirical and may result in equivalent results.]
7. A sheet of nitrocellulose (or, alternatively, nylon) membrane is placed on top of (or below, depending on the direction of the transfer) the gel.
8. Pressure is applied evenly to the gel (either using suction, or by placing a stack of paper towels and a weight on top of the membrane and gel), to ensure good and even contact between gel and membrane.
9. If transferring by suction, 20X SSC buffer is used to ensure a seal and prevent drying of the gel. Buffer transfer by capillary action from a region of high water potential to a region of low water potential (usually filter paper and paper tissues) is then used to move the DNA from the gel onto the membrane; ion exchange interactions bind the DNA to the membrane due to the negative charge of the DNA and positive charge of the membrane.
10. The membrane is then baked in a vacuum or regular oven at 80 °C for 2 hours (standard conditions; nitrocellulose or nylon membrane) or exposed to ultraviolet radiation (nylon membrane) to permanently attach the transferred DNA to the membrane.
11. The membrane is then exposed to a hybridization probe—a single DNA fragment with a specific sequence whose presence in the target DNA is to be determined. The probe DNA

is labelled so that it can be detected, usually by incorporating radioactivity or tagging the molecule with a fluorescent or chromogenic dye.

12. In some cases, the hybridization probe may be made from RNA, rather than DNA. To ensure the specificity of the binding of the probe to the sample DNA, most common hybridization methods use salmon or herring sperm DNA for blocking of the membrane surface and target DNA, deionized formamide, and detergents such as SDS to reduce non-specific binding of the probe.
13. After hybridization, excess probe is washed from the membrane (typically using SSC buffer), and the pattern of hybridization is visualized on X-ray film by autoradiography in the case of a radioactive or fluorescent probe, or by development of colour on the membrane if a chromogenic detection method is used.

RESULT

Hybridization of the probe to a specific DNA fragment on the filter membrane indicates that this fragment contains DNA sequence that is complementary to the probe.

The transfer step of the DNA from the electrophoresis gel to a membrane permits easy binding of the labeled hybridization probe to the size-fractionated DNA.

It also allows for the fixation of the target-probe hybrids, required for analysis by autoradiography or other detection methods. Southern blots performed with restriction enzyme-digested genomic DNA may be used to determine the number of sequences (e.g., gene copies) in a genome.

A probe that hybridizes only to a single DNA segment that has not been cut by the restriction enzyme will produce a single band on a Southern blot, whereas multiple bands will likely be observed when the probe hybridizes to several highly similar sequences (e.g., those that may be the result of sequence duplication).

Modification of the hybridization conditions (for example, increasing the hybridization temperature or decreasing salt concentration) may be used to increase specificity and decrease hybridization of the probe to sequences that are less than 100% similar.

APPLICATIONS

Southern transfer may be used for homology-based cloning on the basis of amino acid sequence of the protein product of the target gene.

Oligonucleotides are designed that are similar to the target sequence. The oligonucleotides are chemically synthesised, radiolabeled, and used to screen a DNA library, or other collections of cloned DNA fragments.

Sequences that hybridize with the hybridization probe are further analysed, for example, to obtain the full length sequence of the targeted gene.

Southern blotting can also be used to identify methylated sites in particular genes. Particularly useful are the restriction nucleases *MspI* and *HpaII*, both of which recognize and cleave within the same sequence.

However, *HpaII* requires that a C within that site be methylated, whereas *MspI* cleaves only DNA unmethylated at that site. Therefore, any methylated sites within a sequence analyzed with a particular probe will be cleaved by the former, but not the latter, enzyme.

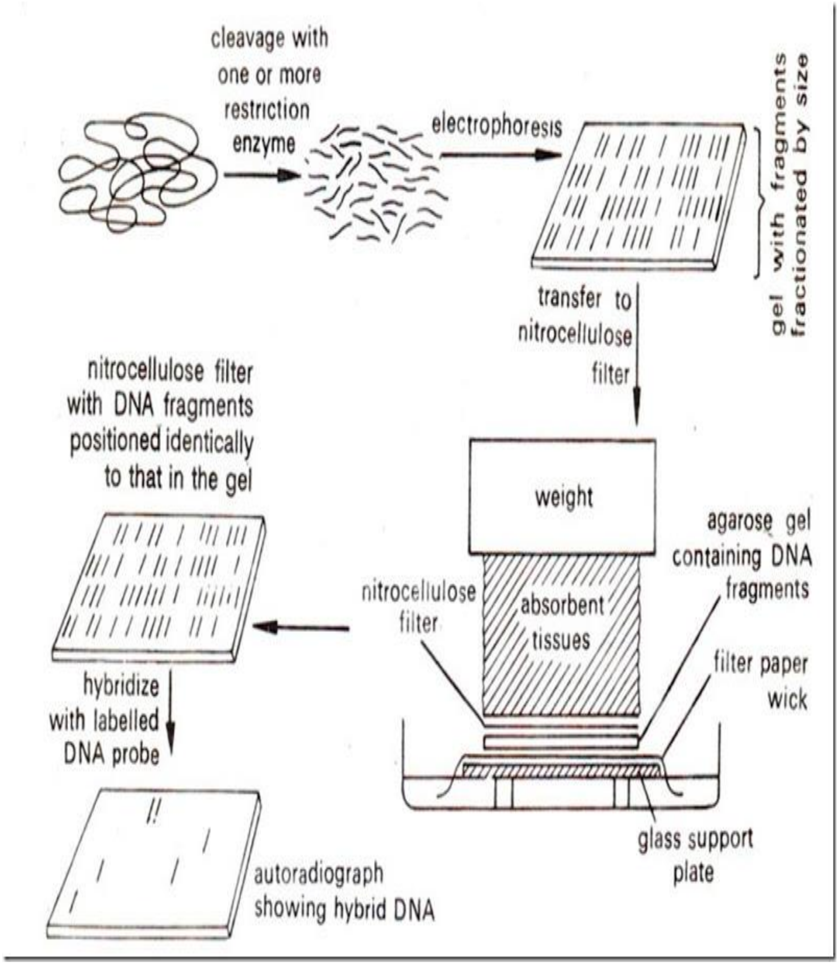


Fig.13.7 southern blotting process

EXPERIMENT NO- 07

OBJECTIVE: Study of DNA Isolation/ DNA Replication

INTRODUCTION

DNA isolation is a process of purification of DNA from sample using a combination of physical and chemical methods. The first isolation of DNA was done in 1869 by Friedrich Miescher. Currently it is a routine procedure in molecular biology or forensic analyses.

BASIC PROCEDURE

There are three basic and two optional steps in a DNA extraction:

- Cells which are to be studied need to be collected.
- Breaking the cell membranes open to expose the DNA along with the cytoplasm within (cell lysis).
    - Lipids from the cell membrane and the nucleus are broken down with detergents and surfactants.
    - Breaking proteins by adding a protease (optional).
    - Breaking RNA by adding an RNase (optional).
- The solution is treated with concentrated salt solution to make debris such as broken proteins, lipids and RNA to clump together.
- Centrifugation of the solution, which separates the clumped cellular debris from the DNA.
- DNA purification from detergents, proteins, salts and reagents used during cell lysis step. The most commonly used procedures are:

- Ethanol precipitation usually by ice-cold ethanol or isopropanol. Since DNA is insoluble in these alcohols, it will aggregate together, giving a *pellet* upon centrifugation. Precipitation of DNA is improved by increasing of ionic strength, usually by adding sodium acetate.
- Phenol–chloroform extraction in which phenol denatures proteins in the sample. After centrifugation of the sample, denaturated proteins stay in the organic phase while aqueous phase containing nucleic acid is mixed with the chloroform that removes phenol residues from solution.
- Minicolumn purification that relies on the fact that the nucleic acids may bind (adsorption) to the solid phase (silica or other) depending on the pH and the salt concentration of the buffer.

Cellular and histone proteins bound to the DNA can be removed either by adding a protease or by having precipitated the proteins with sodium or ammonium acetate, or extracted them with a phenol-chloroform mixture prior to the DNA-precipitation.

After isolation, the DNA is dissolved in slightly alkaline buffer, usually in the TE buffer, or in ultra-pure water.

Special types

Specific techniques must be chosen for isolation of DNA from some samples. Typical samples with complicated DNA isolation are:

- archaeological samples containing partially degraded DNA, see ancient DNA
- samples containing inhibitors of subsequent analysis procedures, most notably inhibitors of PCR, such as humic acid from soil, indigo and other fabric dyes or haemoglobin in blood
- samples from microorganisms with thick cellular wall, for example yeast

Extrachromosomal DNA is generally easy to isolate, especially plasmids may be easily isolated by cell lysis followed by precipitation of proteins, which traps chromosomal DNA in insoluble fraction and after centrifugation, plasmid DNA can be purified from soluble fraction.

A Hirt DNA Extraction is an isolation of all extrachromosomal DNA in a mammalian cell. The Hirt extraction process gets rid of the high molecular weight nuclear DNA, leaving only low molecular weight mitochondrial DNA and any viral episomes present in the cell.

Detecting DNA

A diphenylamine (DPA) indicator will confirm the presence of DNA. This procedure involves chemical hydrolysis of DNA: when heated (e.g. ≥95 °C) in acid, the reaction requires a deoxyribose sugar and therefore is specific for DNA.

Under these conditions, the 2-deoxyribose is converted to w-hydroxylevulinyl aldehyde, which reacts with the compound, diphenylamine, to produce a blue-colored compound. DNA concentration can be determined measuring the intensity of absorbance of the solution at the 600 nm with a spectrophotometer and comparing to a standard curve of known DNA concentrations.

Measuring the intensity of absorbance of the DNA solution at wavelengths 260 nm and 280 nm is used as a measure of DNA purity. DNA absorbs UV light at 260 and 280 nanometres, and aromatic proteins absorb UV light at 280 nm; a pure sample of DNA has a ratio of 1.8 at 260/280 and is relatively free from protein contamination. A DNA preparation that is contaminated with protein will have a 260/280 ratio lower than 1.8.

DNA can be quantified by cutting the DNA with a restriction enzyme, running it on an agarose gel, staining with ethidium bromide or a different stain and comparing the intensity of the DNA with a DNA marker of known concentration.

Using the Southern blot technique, this quantified DNA can be isolated and examined further using PCR and RFLP analysis. These procedures allow differentiation of the repeated sequences within the genome. It is these techniques which forensic scientists use for comparison, identification, and analysis.

DNA REPLICATION

The double helix is un'zipped' and unwound, then each separated strand (turquoise) acts as a template for replicating a new partner strand (green).

Nucleotides (bases) are matched to synthesize the new partner strands into two new double helices.

In molecular biology, **DNA replication** is the biological process of producing two identical replicas of DNA from one original DNA molecule.

This process occurs in all living organisms and is the basis for biological inheritance. DNA is made up of a double helix of two complementary strands. During replication, these strands are separated.

Each strand of the original DNA molecule then serves as a template for the production of its counterpart, a process referred to as semiconservative replication.

Cellular proofreading and error-checking mechanisms ensure near perfect fidelity for DNA replication.

In a cell, DNA replication begins at specific locations, or origins of replication, in the genome. Unwinding of DNA at the origin and synthesis of new strands results in replication forks growing bi-directionally from the origin.

A number of proteins are associated with the replication fork to help in the initiation and continuation of DNA synthesis.

Most prominently, DNA polymerase synthesizes the new strands by adding nucleotides that complement each (template) strand. DNA replication occurs during the S-stage of interphase.

DNA replication can also be performed *in vitro* (artificially, outside a cell).

DNA polymerases isolated from cells and artificial DNA primers can be used to initiate DNA synthesis at known sequences in a template DNA molecule.

The polymerase chain reaction (PCR), a common laboratory technique, cyclically applies such artificial synthesis to amplify a specific target DNA fragment from a pool of DNA.

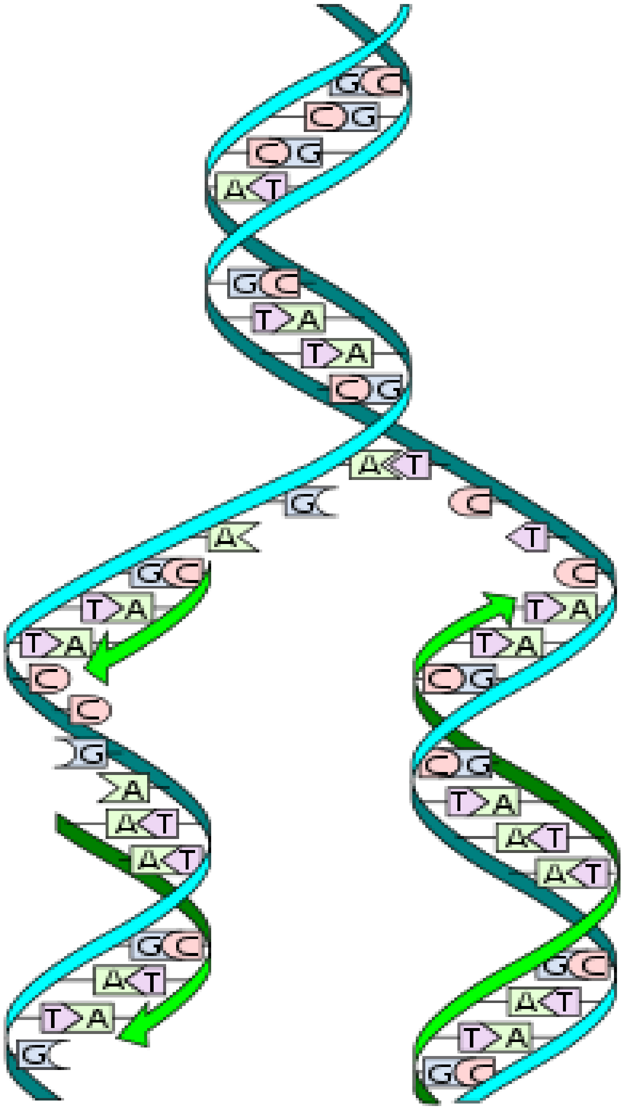


Fig.13.8 DNA replication: The double helix is un'zipped' and unwound, then each separated strand (turquoise) acts as a template for replicating a new partner strand (green). Nucleotides (bases) are matched to synthesize the new partner strands into two new double helices.

DNA structures

DNA usually exists as a double-stranded structure, with both strands coiled together to form the characteristic double-helix. Each single strand of DNA is a chain of four types of nucleotides.

Nucleotides in DNA contain a deoxyribose sugar, a phosphate, and a nucleobase. The four types of nucleotide correspond to the four nucleobases adenine, cytosine, guanine, and thymine, commonly abbreviated as A,C, G and T. Adenine and guanine are purine bases, while cytosine and thymine are pyrimidines.

These nucleotides form phosphodiester bonds, creating the phosphate-deoxyribose backbone of the DNA double helix with the nuclei bases pointing inward (i.e., toward the opposing strand).

Nucleotides (bases) are matched between strands through hydrogen bonds to form base pairs. Adenine pairs with thymine (two hydrogen bonds), and guanine pairs with cytosine (stronger: three hydrogen bonds).

DNA strands have a directionality, and the different ends of a single strand are called the "3' (three-prime) end" and the "5' (five-prime) end". By convention, if the base sequence of a single strand of DNA is given, the left end of the sequence is the 5' end, while the right end of the sequence is the 3' end.

The strands of the double helix are anti-parallel with one being 5' to 3', and the opposite strand 3' to 5'. These terms refer to the carbon atom in deoxyribose to which the next phosphate in the chain attaches.

Directionality has consequences in DNA synthesis, because DNA polymerase can synthesize DNA in only one direction by adding nucleotides to the 3' end of a DNA strand.

The pairing of complementary bases in DNA (through hydrogen bonding) means that the information contained within each strand is redundant. Phosphodiester (intra-strand) bonds are stronger than hydrogen (inter-strand) bonds.

This allows the strands to be separated from one another. The nucleotides on a single strand can therefore be used to reconstruct nucleotides on a newly synthesized partner strand.

REPLICATION PROCESS

DNA replication, like all biological polymerization processes, proceeds in three enzymatically catalyzed and coordinated steps: initiation, elongation and termination.

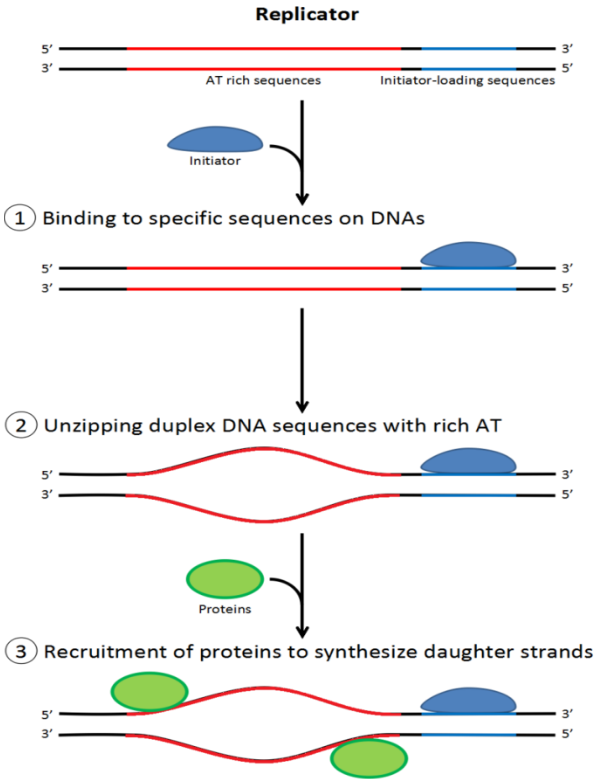

Replicator
5'
3'
3'
5'
AT rich sequences
Initiator-loading sequences
Initiator
1 Binding to specific sequences on DNAs
2 Unzipping duplex DNA sequences with rich AT
Proteins
3 Recruitment of proteins to synthesize daughter strands

Fig.13.6Role of initiators for initiation of DNA replication

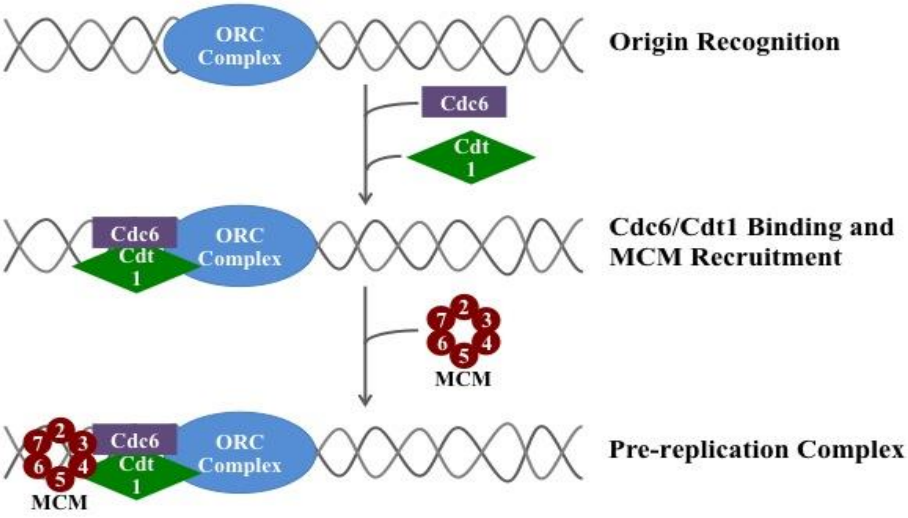


Fig.13.8 Formation of pre-replication complex

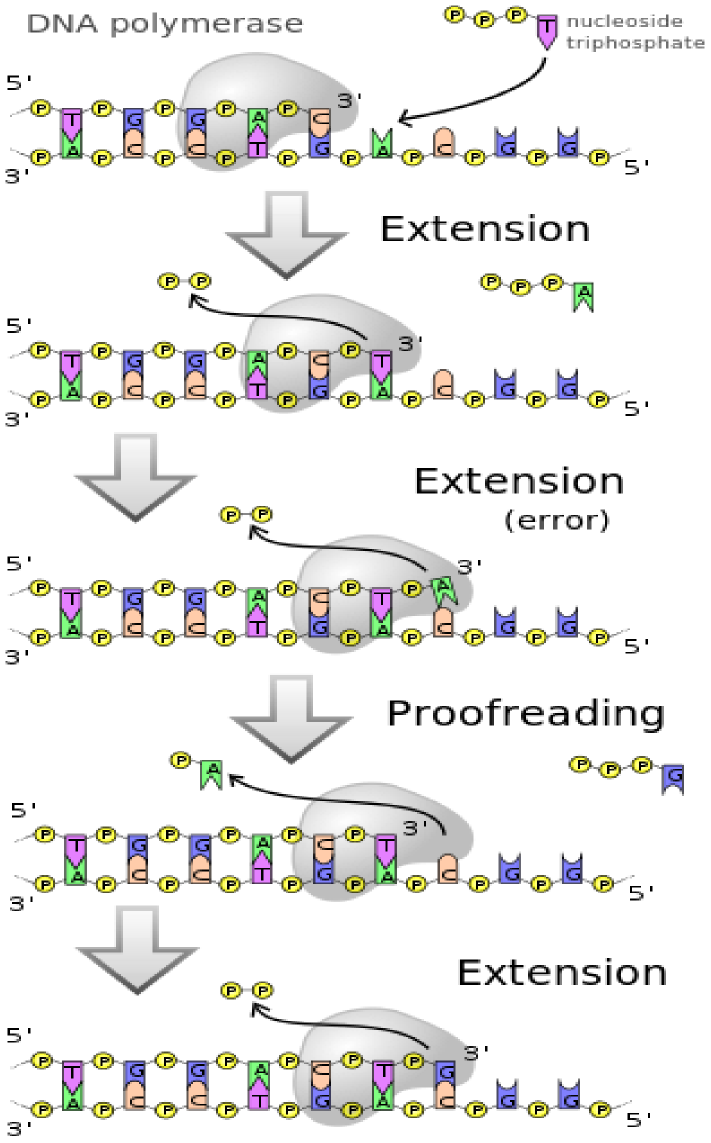


Fig.13.9 DNA polymerases add nucleotides to the 3' end of a strand of DNA. If a mismatch is accidentally incorporated, the polymerase is inhibited from further extension. Proofreading removes the mismatched nucleotide and extension continues

REFERENCES

1. Advanced Light Microscopy vol. 1 Principles and Basic Properties by Maksymilian Pluta, Elsevier (1988)
2. Mc Grath S and van Sinderen D (editors). (2007). Bacteriophage: Genetics and Molecular Biology (1st ed.). Caister Academic Press. ISBN 978-1-904455-14-1. [1].
3. Keen, Eric C.; Bliskovsky, Valery V.; Malagon, Francisco; Baker, James D.; Prince, Jeffrey S.; Klaus, James S.; Adhya, Sankar L.; Groisman, Eduardo A. (2017). "Novel "Superspreader" Bacteriophages Promote Horizontal Gene Transfer by Transformation".
4. Woese CR; Kandler O; Wheelis ML (1990). "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya". Proceedings of the National Academy of Sciences of the United States of America. **87** (12): 4576–9. Bibcode:1990PNAS...87.4576W. doi:10.1073/pnas.87.12.4576. PMC 54159. PMID 2112744.
5. Petitjean, C.; Deschamps, P.; López-García, P. & Moreira, D. (2014). "Rooting the Domain archaea by phylogenomic analysis supports the foundation of the new kingdom proteoarchaeota". Genome Biol. Evol. **7** (1): 191–204. doi:10.1093/gbe/evu274. PMC 4316627. PMID 25527841.
6. Dahm, R (January 2008). "Discovering DNA: Friedrich Miescher and the early years of nucleic acid research.". Human Genetics. **122** (6): 565–81. doi:10.1007/s00439-007-0433-0. PMID 17901982
7. E-Coli virus in France was linked to the one in Germany, but tests are ongoing.
8. "coli". Oxford English Dictionary (3rd ed.). Oxford University Press. September 2005.

9. Singleton P (1999). Bacteria in Biology, Biotechnology and Medicine (5th ed.). Wiley. pp. 444–454. ISBN 0-471-98880-4.

10. "Escherichia coli". CDC National Center for Emerging and Zoonotic Infectious Diseases. Retrieved 2012-10-02.

11. Mc Grath S and van Sinderen D (editors). (2007). Bacteriophage: Genetics and Molecular Biology (1st ed.). Caister Academic Press. ISBN 978-1-904455-14-1. [1].

UNIT 14: BIOSTATICS EXERCISE

INTRODUCTION

Biostatistics is the application of statistics to a wide range of topics in biology. The science of biostatistics encompasses the design of biological experiments, especially in medicine, pharmacy, agriculture and fishery; the collection, summarization, and analysis of data from those experiments; and the interpretation of, and inference from, the results.

A major branch of this is medical biostatistics, which is exclusively concerned with medicine and health.

BACKGROUND

In the early 1900s, after the rediscovery of Gregor Mendel's Mendelian inheritance work, the gaps in understanding between genetics and evolutionary Darwinism led to vigorous debate among biometricians, such as Walter Weldon and Karl Pearson, and Mendelians, such as Charles Davenport, William Bateson and Wilhelm Johannsen. By the 1930s, statisticians and models built on statistical reasoning had helped to resolve these differences and to produce the neo-Darwinian modern evolutionary synthesis.

The leading figures in the establishment of population genetics and this synthesis all relied on statistics and developed its use in biology.

These individuals and the work of other biostatisticians, mathematical biologists, and statistically inclined geneticists helped bring together evolutionary biology and genetics into a consistent, coherent whole that could begin to be quantitatively modeled.

In parallel to this overall development, the pioneering work of D'Arcy Thompson in On Growth and Form also helped to add quantitative discipline to biological study.

Despite the fundamental importance and frequent necessity of statistical reasoning, there may nonetheless have been a tendency among biologists to distrust or deprecate results which are not qualitatively apparent.

One anecdote describes Thomas Hunt Morgan banning the Friden calculator from his department at Caltech, saying "Well, I am like a guy who is prospecting for gold along the banks of the Sacramento River in 1849. With a little intelligence, I can reach down and pick up big nuggets of gold. And as long as I can do that, I'm not going to let any people in my department waste scarce resources in placer mining.

APPLICATIONS OF BIOSTATICS

- Public health, including epidemiology, health services research, nutrition, environmental health and healthcare policy & management.
- Design and analysis of clinical trials in medicine
- Assessment of severity state of a patient with prognosis of outcome of a disease.
- Population genetics and statistical genetics in order to link variation in genotype with a variation in phenotype.
- This has been used in agriculture to improve crops and farm animals (animal breeding).
- In biomedical research, this work can assist in finding candidates for gene alleles that can cause or influence predisposition to disease in human genetics
- Analysis of genomics data, for example from microarray or proteomics experiments.
- Often concerning diseases or disease stages.
- Ecology, ecological forecasting
- Biological sequence analysis
- Systems biology for gene network inference or pathways analysis.

14.1 CALCULATION OF MEAN

“The mean is the average of the numbers. It is easy to calculate: add up all the numbers, then divide by how many numbers there are. In other words it is the sum divided by the count.”

In mathematics, the "mean" is a kind of average found by dividing the sum of a set of numbers by the count of numbers in the set. While it isn't the only kind of average, the mean is the one most people think of when speaking about an average.

You can use means for all kinds of useful purposes in your daily life, from calculating the time it takes you to get home from work, to working out how much money you spend in an average week.

WHAT IS AN ARITHMETIC MEAN?

There are six kindergarten classrooms in a small school district in Florida. The class sizes of each of these kindergartens are 26, 20, 25, 18, 20 and 23.

A researcher writing a report about schools in her town wants to come up with a figure to describe the typical kindergarten class size in this town. She asks a friend for help and her friend suggests that she calculate the average of these class sizes.

To do this, the researcher finds out that she needs to add the kindergarten class sizes together and then divide this sum by six, which is the total number of schools in the district.

Adding the six kindergarten class sizes together gives the researcher a total of 132. If she then divides 132 by six, she gets 22. Therefore, the average kindergarten class size in this school district is 22.

$$Average = \frac{26+20+25+18+20+23}{6} = \frac{132}{6} = 22$$

This average is also known as the arithmetic mean of a set of values.

The Arithmetic Mean Formula

The **arithmetic mean** of a set of values is the ratio of their sum to the total number of values in the set. Thus, if there are a total of *n* numbers in a data set whose values are given by a group of *x*-values, then the arithmetic mean of these values, represented by 'm', can be found using this formula:

$$m = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

In our kindergarten class size example, *n* is 6, or the number of kindergarten classrooms, while the *x*-values are given by the class sizes in each of the kindergartens within the school district. If you recall adding the total number of students in the six classrooms gave us 132. We can plug these values into our formula, dividing 132 by six, and find once again that the average class size is 22. Let's take a look at a couple more examples of how to calculate the arithmetic mean of a group of values.

Examples

A pediatrician has four 9-year-old patients who are boys. Their heights in inches are 54, 57, 53 and 52. She finds out that according to national statistics, the average height of a nine year old boy is 55 inches, or 4 feet and 7 inches. What is the mean or average height of these four boys?

Remember when finding the average, or arithmetic mean, we add together the given values in a data set, then divide by the number of given values. In this example there four different heights given for four different boys. Since we are considering *n* = 4 boys, we add the four heights together and divide the result by 4:

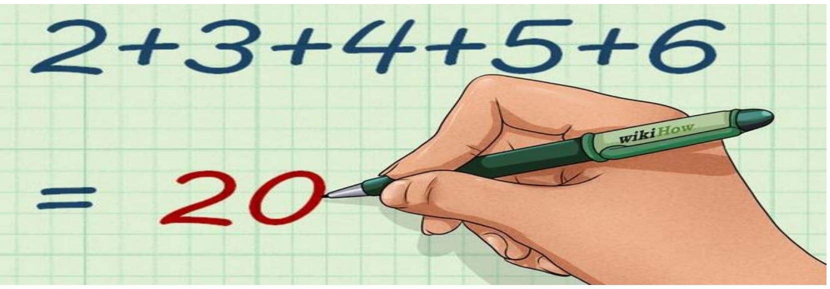
$$Average = \frac{54 + 57 + 53 + 52}{4} = \frac{216}{4} = 54 \text{ inches}$$

As we can see, the average height of these four boys is 54 inches, which is less than the national statistical average, but only by an inch.

Determine the set of values you want to average.

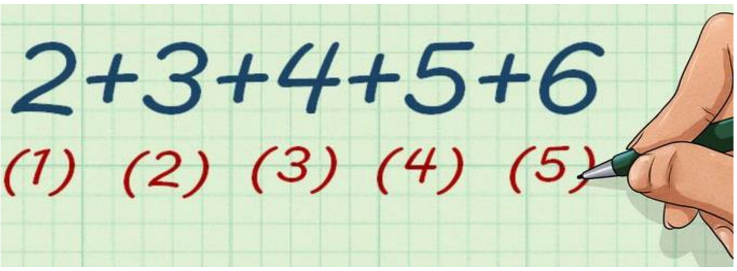
These numbers can be big or small, and there can be as many of them as you want. Just make sure you are using real numbers and not variables.

- Example: 2,3,4,5,6.



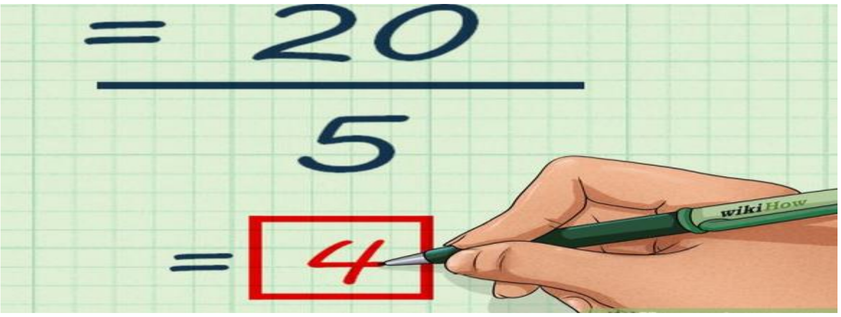
Add your values together to find the sum. You can use a calculator or a spreadsheet, or do it by hand if the set is simple enough.

- Example: 2+3+4+5+6=20.



Count the quantity of values in your group. If you have values that repeat in your set, each one still counts in determining your total.

- Example: 2,3,4,5, and 6 make for a total of five values.



Divide the sum of the set by the count of values. The result is the mean, or average, of your set. This means that if each number in your set was the mean, they would add up to the same total.

- Example: 20 ÷ 5 = 4

Therefore 4 is the mean of the numbers.

1.2: Calculation of Median

The **median** is the value separating the higher half of a data sample, a population, or a probability distribution, from the lower half.

In simple terms, it may be thought of as the "middle" value of a data set. For example, in the data set {1, 3, 3, 6, 7, 8, 9}, the median is 6, the fourth number in the sample.

The median is a commonly used measure of the properties of a data set in statistics and probability theory.

The basic advantage of the median in describing data compared to the mean (often simply described as the "average") is that it is not skewed so much by extremely large or small values, and so it may give a better idea of a 'typical' value.

For example, in understanding statistics like household income or assets which vary greatly, a mean may be skewed by a small number of extremely high or low values. Median income, for example, may be a better way to suggest what a 'typical' income is.

Because of this, the median is of central importance in robust statistics, as it is the most resistant statistic, having a breakdown point of 50%: so long as no more than half the data are contaminated, the median will not give an arbitrarily large or small result.

Basic procedure

The median of a finite list of numbers can be found by arranging all the numbers from smallest to greatest.

If there is an odd number of numbers, the middle one is picked. For example, consider the set of numbers:

1, 3, 3, 6, 7, 8, 9

This set contains seven numbers. The median is the fourth of them, which is 6.

If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of the two middle values. For example, in the data set:

1, 2, 3, 4, 5, 6, 8, 9

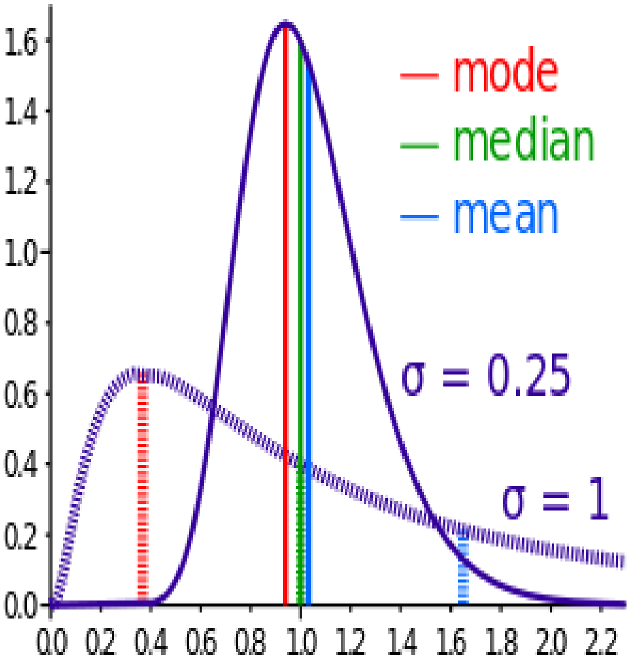
The median is the mean of the middle two numbers: this is (4 + 5) ÷ 2, which is 4.5. (In more technical terms, this interprets the median as the fully trimmed mid-range.)

The formula used to find the middle number of a data set of *n* numbers is (n + 1) ÷ 2. This either gives the middle number (for an odd number of values) or the halfway point between the two middle values. For example, with 14 values, the formula will give 7.5, and the median will be taken by averaging the seventh and eighth values.

You will also be able to find the median using the Stem-and-Leaf Plot.

There is no widely accepted standard notation for the median, but some authors represent the median of a variable *x* either as $\tilde{x}$ or as $\mu_{1/2}$ sometimes also *M*. In any of these cases, the use of these or other symbols for the median needs to be explicitly defined when they are introduced.

DISCUSSION



Comparison of mean, median and mode of two log-normal distributions with different skewnes Calculation of medians is a popular technique in summary statistics and summarizing statistical data, since it is simple to understand and easy to calculate, while also giving a measure that is more robust in the presence of outlier values than is the mean.

The widely cited empirical relationship between the relative locations of the mean and the median for skewed distributions is, however, not generally true. There are, however, various relationships for the *absolute* difference between them; see below.

The median does not identify a specific value within the data set, since more than one value can be at the median level and with an even number of observations (as shown above) no value need be exactly at the value of the median.

Nonetheless, the value of the median is uniquely determined with the usual definition. A related concept, in which the outcome is forced to correspond to a member of the sample, is the medoid.

In a population, at most half have values strictly less than the median and at most half have values strictly greater than it. If each group contains less than half the population, then some of the population is exactly equal to the median.

For example, if $a < b < c$, then the median of the list $\{a, b, c\}$ is b, and, if $a < b < c < d$, then the median of the list $\{a, b, c, d\}$ is the mean of b and c; i.e., it is $(b + c)/2$. Indeed, as it is based on the middle data in a group, it is not necessary to even know the value of extreme results in order to calculate a median. For example, in a psychology test investigating the time needed to solve a problem, if a small number of people failed to solve the problem at all in the given time a median can still be calculated.

The median can be used as a measure of location when a distribution is skewed, when end-values are not known, or when one requires reduced importance to be attached to outliers, e.g., because they may be measurement errors.

A median is only defined on ordered one-dimensional data, and is independent of any distance metric. A geometric median, on the other hand, is defined in any number of dimensions.

The median is one of a number of ways of summarizing the typical values associated with members of a statistical population; thus, it is a possible location parameter. The median is the 2nd quartile, 5^{th} decile, and 50th percentile. Since the median is the same as the *second quartile*, its calculation is illustrated in the article on quartiles.

A median can be worked out for ranked but not numerical classes (e.g. working out a median grade when students are graded from A to F), although the result might be halfway between grades if there are an even number of cases.

When the median is used as a location parameter in descriptive statistics, there are several choices for a measure of variability: the range, the interquartile range, the mean absolute deviation, and the median absolute deviation.

For practical purposes, different measures of location and dispersion are often compared on the basis of how well the corresponding population values can be estimated from a sample of data.

The median, estimated using the sample median, has good properties in this regard. While it is not usually optimal if a given population distribution is assumed, its properties are always reasonably good.

For example, a comparison of the efficiency of candidate estimators shows that the sample mean is more statistically efficient than the sample median when data are uncontaminated by data from heavy-tailed distributions or from mixtures of distributions, but less efficient otherwise, and that the efficiency of the sample median is higher than that for a wide range of distributions.

More specifically, the median has a 64% efficiency compared to the minimum-variance mean (for large normal samples), which is to say the variance of the median will be ~50% greater than the variance of the mea

Easy explanation of the sample median

In individual series (if number of observation is very low) first one must arrange all the observations in order. Then count(n) is the total number of observation in given data.

If n is odd then Median (M) = value of ((n + 1)/2)th item term.

If n is even then Median (M) = value of [(n/2)th item term + (n/2 + 1)th item term]/2

For an odd number of values:

As an example, we will calculate the sample median for the following set of observations: 1, 5, 2, 8, 7.

Start by sorting the values: 1, 2, 5, 7, 8.

In this case, the median is 5 since it is the middle observation in the ordered list.

The median is the ((n + 1)/2)th item, where n is the number of values. For example, for the list {1, 2, 5, 7, 8}, we have n = 5, so the median is the ((5 + 1)/2)th item.

median = (6/2)th item

median = 3rd item

median = 5

For an even number of values:

As an example, we will calculate the sample median for the following set of observations: 1, 6, 2, 8, 7, 2.

Start by sorting the values: 1, 2, 2, 6, 7, 8.

In this case, the arithmetic mean of the two middlemost terms is (2 + 6)/2 = 4. Therefore, the median is 4 since it is the arithmetic mean of the middle observations in the ordered list.

We also use this formula MEDIAN = {(n + 1)/2}th item . n = number of values

As above example 1, 2, 2, 6, 7, 8 n = 6 Median = {(6 + 1)/2}th item = 3.5th item. In this case, the median is average of the 3rd number and the next one (the fourth number). The median is (2 + 6)/2 which is 4.

CALCULATING THE MEDIAN

Information identified as archived is provided for reference, research or recordkeeping purposes. It is not subject to the Government of Canada Web Standards and has not been altered or updated since it was archived.

If observations of a variable are ordered by value, the median value corresponds to the middle observation in that ordered list. The median value corresponds to a cumulative percentage of 50% (i.e., 50% of the values are below the median and 50% of the values are above the median). The position of the median is

$\{(n + 1) \div 2\}^{th}$ value, where **n** is the number of values in a set of data.

In order to calculate the median, the data must first be ranked (sorted in ascending order). The median is the number in the middle.

Median = the middle value of a set of ordered data.

The median is usually calculated for numeric variables, but may also be calculated for categorical variables that are sequenced, such as the categories in a satisfaction survey: excellent, good, satisfactory and poor. These qualitative categories can be ranked in order, and thus, are considered ordinal.

RAW DATA

In raw data, the median is the point at which exactly half of the data are above and half below. These halves meet at the median position.

If the number of observations is odd, the median fits perfectly and the depth of the median position will be a whole number. If the number of observations is even, the depth of the median position will include a decimal. You need to find the midpoint between the numbers on either side of the median position.

Example 1 – Raw data (discrete variables)

Imagine that a top running athlete in a typical 200-metre training session runs in the following times:

26.1, 25.6, 25.7, 25.2 et 25.0 seconds.

How would you calculate his median time?

First, the values are put in ascending order: 25.0, 25.2, 25.6, 25.7, 26.1. Then, using the following formula, figure out which value is the middle value. Remember that n represents the number of values in the data set.

Median = $\{(n + 1) \div 2\}^{th}$ value

= $(5 + 1) \div 2$

= 3

The third value in the data set will be the median. Since 25.6 is the third value, 25.6 seconds would be the median time.

= 25.6 secondes

Example 2 – Raw data (discrete variables)

Now, if the runner sprints the sixth 200-metre race in 24.7 seconds, what is the median value now?

Again, you first put the data in ascending order: 24.7, 25.0, 25.2, 25.6, 25.7, 26.1. Then, you use the same formula to calculate the median time.

Median = $\{(n + 1) \div 2\}^{th}$ value

= (6 + 1) ÷ 2

= 7 ÷ 2

= 3,5

Since there is an even number of observations in this data set, there is no longer a distinct middle value. The median is the 3.5th value in the data set meaning that it lies between the third and fourth values. Thus, the median is calculated by averaging the two middle values of 25.2 and 25.6. Use the formula below to get the average value.

Average=(value below median + value above median) ÷ 2

= (third value + fourth value) ÷ 2

= (25.2 + 25.6) ÷ 2

= 50.8 ÷ 2

= 25.4

The value 25.4 falls directly between the third and fourth values in this data set, so 25.4 seconds

Would be the median time.

UNGROUPED FREQUENCY DISTRIBUTION

In order to find the median using cumulative frequencies (or the number of observations that lie above or below a particular value in a data set), you must calculate the first value with a cumulative frequency greater than or equal to the median.

If the median's value is exactly 0.5 more than the cumulative frequency of the previous value, then the median is the midpoint between the two values.

Example 3 – Ungrouped frequency table (discrete variables)

Imagine that your school baseball team scores the following number of home runs in 10 games:

4, 5, 8, 5, 7, 8, 9, 8, 8, 7

If you were to place the total home runs in a frequency table, what would the median be?

First, put the scores in ascending order:

4, 5, 5, 7, 7, 8, 8, 8, 8, 9

Then, make a table with two columns. Label the first column "Number of home runs" and then list the possible number of home runs the team could get.

You can start from 0 and list up until the number 10, but since the team never scored less than 4 home runs, you may wish to start listing at the number 4.

Label the second column "Frequency." In this column, record the numbers of times 4 home runs were scored, 5 home runs were scored and so on.

In this case, there was only one time that 4 home runs were scored, but two times that 5 home runs were scored.

If you add all of the numbers in the Frequency column, they should equal 10 (for the 10 games played).

Table 1. Number of home runs in 10 baseball games

Number of home runs (x)	Frequency (f)
4	1
5	2
6	0
7	2
8	4
9	1

To find the median, again use the same formula:

Median = {(n + 1) ÷ 2}th value

= (10 + 1) ÷ 2

= 11 ÷ 2

= 5.5

= the median is the 5.5th value in the data set

To get the median, add up the numbers in the Frequency column until you get to 5 (and since the total number of games is 10, the remaining numbers in that column should also equal 5). You will reach 5 after adding all of the frequencies up to and including those for the 7 home runs. The next set of five will begin with the frequencies for 8 home runs. The median (the 5.5th value) lies between the fifth value and the sixth value. Thus, the median lies between 7 home runs and 8 home runs.

If you calculate the average of these values (using the same formula used in Example 2), the result is 7.5.

Average = (middle value before + middle value after) ÷ 2

= (fifth value + sixth value) ÷ 2

= (7 + 8) ÷ 2

= 15 ÷ 2

= 7.5

Technically, the median should be a possible variable. In the above example, the variables are discrete and always whole numbers. Therefore, 7.5 is not a possible variable—no one can hit 7 and a half home runs. Thus, this number only makes sense statistically. Some mathematicians may argue that 8 is a more appropriate median.

GROUPED FREQUENCY DISTRIBUTION

Sometimes it does not make sense to list each individual variable when a frequency distribution table would be long and cumbersome to work with.

In order to simplify this, divide the range of data into intervals and then list the intervals in a frequency distribution table, including a column for the cumulative percentage. (For more information, refer to the Cumulative frequency section.)

The calculation to find the median is a little longer because the data have been grouped into intervals and, therefore, all of the original information has been lost.

Some textbooks simply take the midpoint of the interval as the median. However, that method is an over-simplification of the true value. Use the following calculations to find the median for a grouped frequency distribution.

1. Figure out which interval contains the median by using the **(n + 1) ÷ 2** formula. Take whatever value the calculation gives you and then add up the numbers in the frequency column until you come to that value (just like Example 3). For example, if your median is the 13.5th value, add up the frequencies until you come to the 13th and 14th values. Whichever interval contains these values is called the median group.
2. Find the cumulative percentage of the interval preceding the median group. Label this value **A**.
3. Using this cumulative percentage, calculate how many numbers are needed in order to add up to 50% of the total cumulative percentage. This value will be labeled **B**. Use the following formula to calculate **B**:
   B = 50 - A

4. Figure out the range (how many numbers the interval covers). Call this value **C**. Then, find the percentage for the median interval. Call this value **D**.
5. Calculate how many data values you have to count in the median group to get 50% of the total data set by using the following formula. Call this value **E**.

E = (B ÷ D) x C

6. Find out what the median value is by adding the value for E to the lower value of the median interval:

   Median = lower value + E

   Since **E = (B ÷ D) x C**, this formula can also be written as:

   Median = lower value + (B ÷ D) x C

If the cumulative frequency for an interval is exactly 50%, then the median value would be the endpoint of this interval.

Let's make this clear with an example!

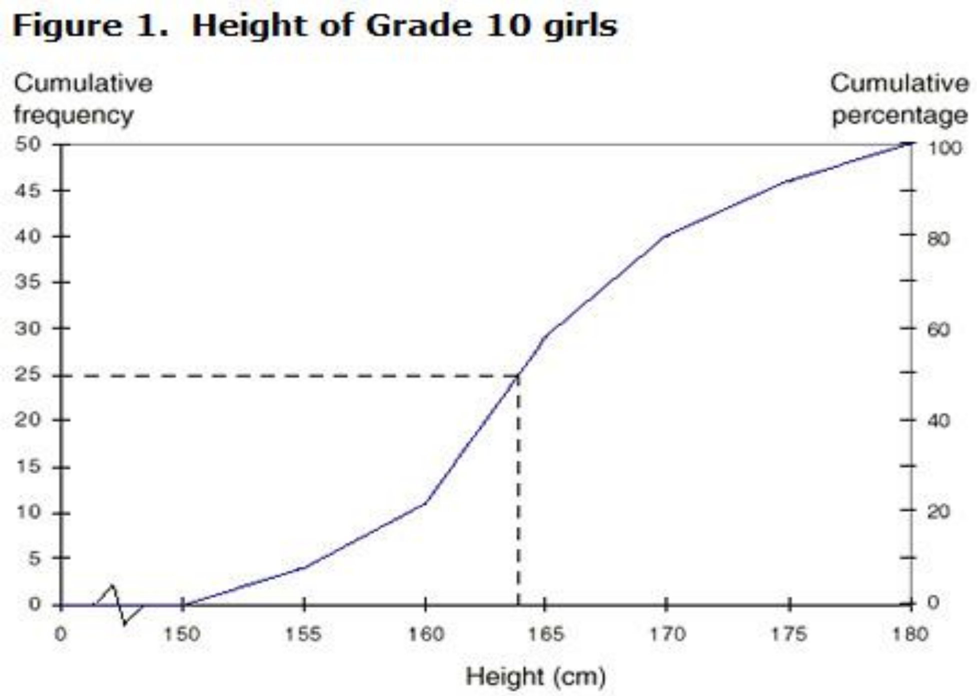
Example 4 – Grouped variables - frequency distribution (continuous or discrete)

Using the same information from Example 4 in the Mean section, imagine that you surveyed 50 Grade 10 girls to find out how tall each one is in centimetres. After gathering all of your data, you created a frequency distribution table that looked like this:

Table 2. Height of Grade 10 girls

Height (cm)	**Frequency (f)**	**Endpoint (x)**	**Cumulative frequency**	**Percentage**	**Cumulative percentage**
150 to < 155	4	155	4	8	8
155 to < 160	7	160	11	14	22
160 to < 165	18	165	29	36	58
165 to < 170	11	170	40	22	80
170 to < 175	6	175	46	12	92
175 to < 180	4	180	50	8	100

Using the grouped data, you created a cumulative frequency graph to accompany your table. The endpoints of the height intervals, the numbers for cumulative frequency and the numbers for cumulative percentage have been plotted on the graph.



By just looking at the graph, you can try to find the median value. The median is the point where the x-axis (Height) intersects with the midpoint (25) of the y-axis (Cumulative frequency). You will see that the median value is approximately 164 cm. Using mathematical calculations; you can find out that the value is actually 163.9 cm. Here's how:

1. According to the information provided in Table 2:

   Median = {(n + 1) ÷ 2}th value
2. = (50 + 1) ÷ 2
3. = 51 ÷ 2
4. = 25.5
5. By adding up the frequencies, we find that the median (25.5) lies in the median group of the 160 to < 165 cm interval.
6. The cumulative percentage of the preceding interval (**A**) is 22.

7. The percentage needed in order to get 50% of the total cumulative percentage (**B**) is 28.

   B = 50 – A
8. = 50 – 22
9. = 28
10. The range of the median interval (**C**) is 5 and the percentage for the median interval (**D**) is 36.
11. The number of values to count down within the interval in order to get to 50% of the total data set is 3.9.
12. **E = (B ÷ D) x C**
13. = (28 ÷ 36) x 5
14. **= 3.9**
15. Since the lower value of the median interval is 160, when you add the value of **E** to that you get a median of 163.9 cm.
16. **Median = lower value of median interval + (B ÷ D) x C**
17. = 160 + (28 ÷ 36) x 5
18. = 160 + 3.9
19. **= 163.9 cm**

COMPARING THE MEAN AND MEDIAN

It is possible for the mean and median of a distribution to have the same value. This is always the case if distribution is symmetrical as in a normal distribution. If the distribution is roughly symmetrical, then the two values will be close together.

In the example of the heights of the 50 Grade 10 girls, the mean (164.5 cm) is very close to the value of the median (163.5 cm). This is because the distribution is roughly symmetrical.

However, one number can alter the mean without affecting the median.

Example 6 – Comparing the mean and median

Consider the following sets of data that represent the number of points scored by 3 players in 11 lacrosse games.

Eileen: 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3
Mean = 22 ÷ 11 = **2**
Median = **2**

Jeremy: 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 4
Mean = 23 ÷ 11 = **2.1**
Median = **2**

Randy: 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 14
Mean = 33 ÷ 11 = **3**
Median = **2**

The three sets of data above are identical except for the last observation values (3, 4 and 14).

14.3: MODE

The **mode** is the value that appears most often in a set of data. The mode of a discrete probability distribution is the value *x* at which its probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled.

The mode of a continuous probability distribution is the value *x* at which its probability density function has its maximum value, so the mode is at the peak.

Like the statistical mean and median, the mode is a way of expressing, in a (usually) single number, important information about a random variable or a population.

The numerical value of the mode is the same as that of the mean and median in a normal distribution and it may be very different in highly skewed distributions.

The mode is not necessarily unique to a given distribution, since the probability mass function or probability density function may take the same maximum value at several points x_1, x_2, etc.

The most extreme case occurs in uniform distributions, where all values occur equally frequently.

When a probability density function has multiple local maxima it is common to refer to all of the local maxima as modes of the distribution. Such a continuous distribution is called multimodal.

In symmetric unimodal distributions, such as the normal distribution, the mean (if defined), median and mode all coincide. For samples, if it is known that they are drawn from a symmetric distribution, the sample mean can be used as an estimate of the population mode.

MODE OF A SAMPLE

The mode of a sample is the element that occurs most often in the collection. For example, the mode of the sample [1, 3, 6, 6, 6, 6, 7, 7, 12, 12, 17] is 6.

Given the list of data [1, 1, 2, 4, 4] the mode is not unique - the dataset may be said to be bimodal, while a set with more than two modes may be described as multimodal.

For a sample from a continuous distribution, such as [0.935..., 1.211..., 2.430..., 3.668..., 3.874...], the concept is unusable in its raw form, since no two values will be exactly the same, so each value will occur precisely once. In order to estimate the mode, the usual practice is to discretize the data by assigning frequency values to intervals of equal distance, as for making a histogram, effectively replacing the values by the midpoints of the intervals they are assigned to. The mode is then the value where the histogram reaches its peak. For small or middle-sized samples the outcome of this procedure is sensitive to the choice of interval width if chosen too narrow or too wide; typically one should have a sizable fraction of the data concentrated in a relatively small number of intervals (5 to 10), while the fraction of the data falling outside these intervals is also sizable. An alternate approach is kernel density estimation, which essentially blurs point samples to produce a continuous estimate of the probability density function which can provide an estimate of the mode.

Comparison of mean, median and mode

Comparison of common averages of values { 1, 2, 2, 3, 4, 7, 9 }

Type	Description	Example	Result
Arithmetic mean	Sum of values of a data set divided by number of values:	(1+2+2+3+4+7+9) / 7	**4**
Median	Middle value separating the greater and lesser halves of a data set	1, 2, 2, **3**, 4, 7, 9	**3**
Mode	Most frequent value in a data set	1, **2**, **2**, 3, 4, 7, 9	**2**

USE

Unlike mean and median, the concept of mode also makes sense for "nominal data" (i.e., not consisting of numerical values in the case of mean, or even of ordered values in the case of median).

For example, taking a sample of Korean family names, one might find that "Kim" occurs more often than any other name. Then "Kim" would be the mode of the sample.

In any voting system where a plurality determines victory, a single modal value determines the victor, while a multi-modal outcome would require some tie-breaking procedure to take place.

Unlike median, the concept of mode makes sense for any random variable assuming values from a vector space, including the real numbers (a one-dimensional vector space) and the integers (which can be considered embedded in the reals).

For example, a distribution of points in the plane will typically have a mean and a mode, but the concept of median does not apply.

The median makes sense when there is a linear order on the possible values. Generalizations of the concept of median to higher-dimensional spaces are the geometric median and the centerpoint.

PROPERTIES

Assuming definedness, and for simplicity uniqueness, the following are some of the most interesting properties.

- All three measures have the following property: If the random variable (or each value from the sample) is subjected to the linear or affine transformation which replaces *X* by *aX+b*, so are the mean, median and mode.
- Except for extremely small samples, the mode is insensitive to "outliers" (such as occasional, rare, false experimental readings). The median is also very robust in the presence of outliers, while the mean is rather sensitive.
- In continuous unimodal distributions the median lies, as a rule of thumb, between the mean and the mode, about one third of the way going from mean to mode. In a formula, median ≈ (2 × mean + mode)/3. This rule, due to Karl Pearson, often applies to slightly non-symmetric distributions that resemble a normal distribution, but it is not always true and in general the three statistics can appear in any order.

- For unimodal distributions, the mode is within standard deviations of the mean, and the root mean square deviation about the mode is between the standard deviation and twice the standard deviation.

CALCULATING THE MODE

Information identified as archived is provided for reference, research or recordkeeping purposes. It is not subject to the Government of Canada Web Standards and has not been altered or updated since it was archived.

In a set of data, the mode is the most frequently observed data value. There may be no mode if no value appears more than any other.

There may also be two modes (bimodal), three modes (trimodal), or four or more modes (multimodal).

In the case of grouped frequency distributions, the modal class is the class with the largest frequency.

Mode = *the most frequently observed data value*

As a set of data can have more than one mode, the mode does not necessarily indicate the centre of a data set.

The mode will be close to the mean and median if the data have a normal or near-normal distribution.

In fact, if the distribution is symmetrical and unimodal, then the mean, the median and the mode may have the same value.

CATEGORICAL OR DISCRETE VARIABLES

For categorical or discrete variables, the mode is simply the most observed value. To work out the mode, observations do not have to be placed in order, although for ease of calculation it is advisable to do so.

Example 1 – Categorical or discrete variables

During a hockey tournament, Anne scored 7, 5, 0, 7, 8, 5, 5, 4, 1, and 5 points in 10 games. The mode of her data set is 5 because this value occurred the most often (four times). This can be interpreted to mean that if one game were selected at random, a good guess would be that Anne would score 5 points.

Example 2 – Categorical or discrete variables

During Marco's 12-game basketball season, he scored 14, 14, 15, 16, 14, 16, 16, 18, 14, 16, 16 and 14 points. This data set is bimodal; there are two modes, 14 and 16, because both of them occur the most often (five times).

Example 3 – Categorical or discrete variables

The following data set represents the number of touchdowns scored by Jerome in his high-school football season:

0, 0, 1, 0, 0, 2, 3, 1, 0, 1, 2, 3, 1, 0

First, put the data set in order:

0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 2, 2, 3, 3

Find and compare the mean, median and mode.

Mode = most frequently observed data value = 0

The mode is 0 because this value occurs most often. If one game were selected at random, the mode tells us that a good guess would be that Jerome would not score a touchdown.

Mean =

Σ

x ÷ n

= 14 ÷ 14

= **1**

However, on average (mean), Jerome will score one touchdown per game even though the mode indicates he did not score a touchdown in a lot of games. In this case, the mode does not provide a useful measure of the player's performance.

Median = (n + 1) ÷ 2th value

= (14 + 1) ÷ 2

= 15 ÷ 2

= 7.5

Average = (value below median + value above median) ÷ 2
= (seventh value + eighth value) ÷ 2
= (1 + 1) ÷ 2
= **1**

Because the number of values in the data set is even, the median does not fit perfectly in the centre of the data set. Instead, the median had to be found using the above equations. According to the results, the median states that Jerome will score one touchdown per game.

Grouped variables (continuous or discrete)

When continuous or discrete variables are grouped in tables, the mode is defined as the class interval where most observations lie. This is called the modal-class interval.

In the example of the height of 50 Grade 10 girls, the *modal-class interval* would be 160 –< 165 cm, as this interval has the most observations in it.

The mode is rarely used as a measure of central tendency for numeric variables. However, for categorical variables, the mode is more useful because the mean and median do not make sense.

Next, you could determine the midrange of the modal class. The *midrange* is simply the midpoint between the highest and lowest values in a class. The mode is not used very often in conjunction with the midrange because it gives only a very poor estimate of the average.

The mode can be used with categorical data, but the mean and median cannot. The mode may or may not exist, and there may be more than one value for the mode.

14.4: STANDARD DEVIATION

In statistics, the standard deviation (SD, also represented by the Greek letter sigma σ or the Latin letters) is a measure that is used to quantify the amount of variation or dispersion of a set of data values.

A low standard deviation indicates that the data points tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

The standard deviation of a random variable, statistical population, data set, or probability distribution is the square root of its variance.

It is algebraically simpler, though in practice less robust, than the average absolute deviation. A useful property of the standard deviation is that, unlike the variance, it is expressed in the same units as the data.

There are also other measures of deviation from the norm, including mean absolute deviation, which provide different mathematical properties from standard deviation.

In addition to expressing the variability of a population, the standard deviation is commonly used to measure confidence in statistical conclusions. For example, the margin of error in polling data is determined by calculating the expected standard deviation in the results if the same poll were to be conducted multiple times.

This derivation of a standard deviation is often called the "standard error" of the estimate or "standard error of the mean" when referring to a mean.

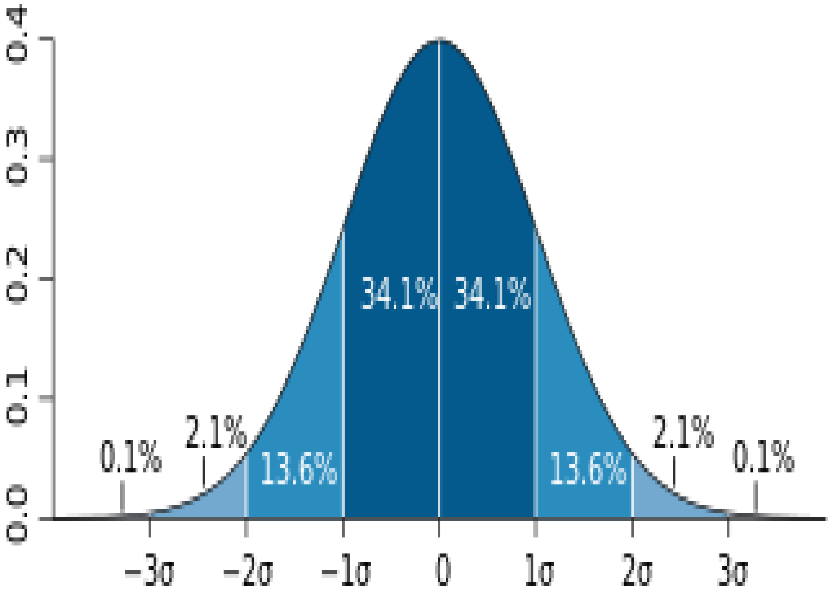
It is computed as the standard deviation of all the means that would be computed from that population if an infinite number of samples were drawn and a mean for each sample were computed.

It is very important to note that the standard deviation of a population and the standard error of a statistic derived from that population (such as the mean) are quite different but related (related by the inverse of the square root of the number of observations).

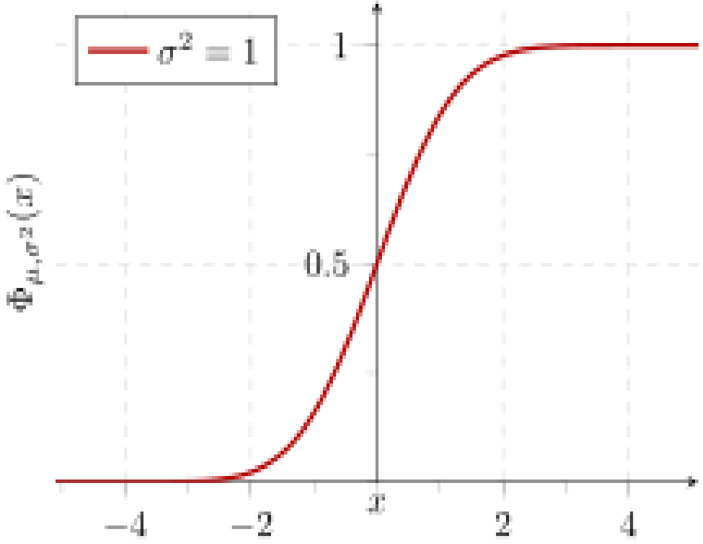
The reported margin of error of a poll is computed from the standard error of the mean (or alternatively from the product of the standard deviation of the population and the inverse of the square root of the sample size, which is the same thing) and is typically about twice the standard deviation—the half-width of a 95 percent confidence interval.

In science, researchers commonly report the standard deviation of experimental data, and only effects that fall much farther than two standard deviations away from what would have been expected are considered statistically significant—normal random error or variation in the measurements is in this way distinguished from likely genuine effects or associations. The standard deviation is also important in finance, where the standard deviation on the rate of return on an investment is a measure of the volatility of the investment.

When only a sample of data from a population is available, the term standard deviation of the sample or sample standard deviation can refer to either the above-mentioned quantity as applied to those data or to a modified quantity that is an unbiased estimate of the population standard deviation (the standard deviation of the entire population).



A plot of normal distribution (or bell-shaped curve) where each band has a width of 1 standard deviation – See also: 68–95–99.7 rule



Cumulative probability of a normal distribution with expected value 0 and standard deviation 1

DEFINITION OF POPULATION VALUES

Let X be a random variable with mean value μ:

Here the operator E denotes the average or expected value of X. Then the **standard deviation** of X is the quantity (derived using the properties of expected value).

In other words, the standard deviation σ (sigma) is the square root of the variance of X; i.e., it is the square root of the average value of $(X - \mu)^2$.

The standard deviation of a (univariate) probability distribution is the same as that of a random variable having that distribution. Not all random variables have a standard deviation, since these expected values need not exist.

For example, the standard deviation of a random variable that follows a Cauchy distribution is undefined because its expected value μ is undefined

DISCRETE RANDOM VARIABLE

In the case where X takes random values from a finite data set x_1, x_2, ..., x_N, with each value having the same probability, the standard deviation is or, using summation notation,If, instead of having equal probabilities, the values have different probabilities, let x_1 have probability p_1, x_2 have probability p_2, ..., x_N have probability p_N.

CONTINUOUS RANDOM VARIABLE

The standard deviation of a continuous real-valued random variable X with probability density function $p(x)$ isand where the integrals are definite integrals taken for x ranging over the set of possible values of the random variable X. In the case of a parametric family of distributions, the standard deviation can be expressed in terms of the parameters. For example, in the case of the log-normal distribution with parameters μ and σ^2, the standard deviation is $[(\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)]^{1/2}$.

ESTIMATION

One can find the standard deviation of an entire population in cases (such as standardized testing) where every member of a population is sampled.

In cases where that cannot be done, the standard deviation σ is estimated by examining a random sample taken from the population and computing a statistic of the sample, which is used as an estimate of the population standard deviation.

Such a statistic is called an estimator, and the estimator (or the value of the estimator, namely the estimate) is called a **sample standard deviation,** and is denoted by s (possibly with modifiers).

However, unlike in the case of estimating the population means, for which the sample means, is a simple estimator with many desirable properties (unbiased, efficient, maximum likelihood), there is no single estimator for the standard deviation with all these properties, and unbiased estimation of standard deviation is a very technically involved problem.

Most often, the standard deviation is estimated using the *corrected sample standard deviation* (using $N - 1$), defined below, and this is often referred to as the "sample standard deviation", without qualifiers.

However, other estimators are better in other respects: the uncorrected estimator (using N) yields lower mean squared error, while using $N - 1.5$ (for the normal distribution) almost completely eliminates bias.

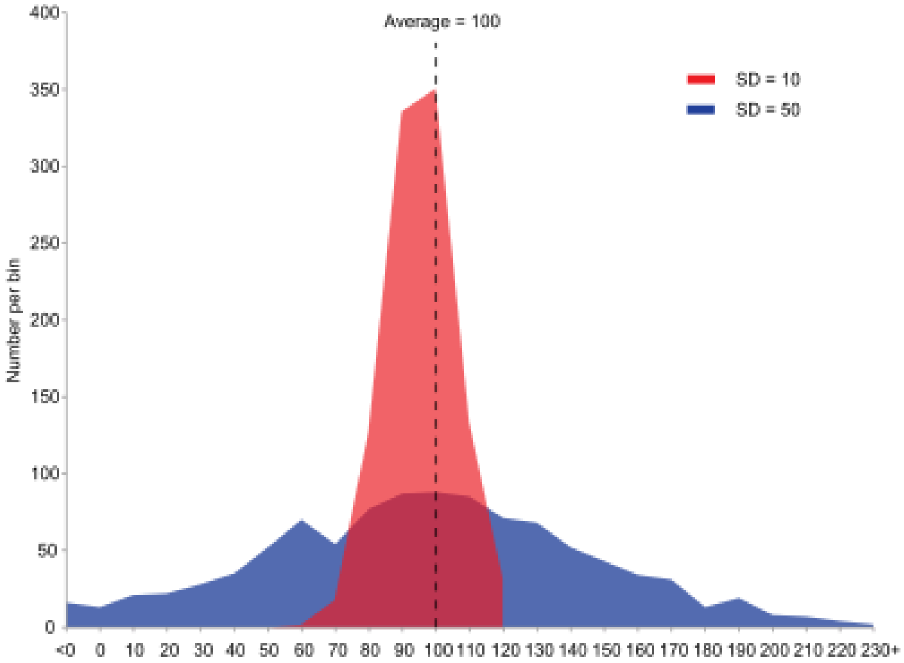
UNCORRECTED SAMPLE STANDARD DEVIATION

Firstly, the formula for the *population* standard deviation (of a finite population) can be applied to the sample, using the size of the sample as the size of the population (though the actual population size from which the sample is drawn may be much larger).

This estimator, denoted by sN, is known as the uncorrected sample standard deviation, or sometimes the standard deviation of the sample (considered as the entire population), and is

defined as follows.s N = 1 N ∑ i = 1 N (x i − x ¯) 2 , where { x 1 , x 2 , … , x N } are the observed values of the sample items and x ¯ is the mean value of these observations, while the denominator N stands for the size of the sample: this is the square root of the sample variance, which is the average of the squared deviations about the sample mean.

INTERPRETATION AND APPLICATION



Example of samples from two populations with the same mean but different standard deviations. Red population has mean 100 and SD 10; blue population has mean 100 and SD 50.

A large standard deviation indicates that the data points can spread far from the mean and a small standard deviation indicates that they are clustered closely around the mean.

For example, each of the three populations {0, 0, 14, 14}, {0, 6, 8, 14} and {6, 6, 8, 8} has a mean of 7.

Their standard deviations are 7, 5, and 1, respectively. The third population has a much smaller standard deviation than the other two because its values are all close to 7. It will have the same units as the data points themselves.

If, for instance, the data set {0, 6, 8, 14} represents the ages of a population of four siblings in years, the standard deviation is 5 years.

As another example, the population {1000, 1006, 1008, and 1014} may represent the distances traveled by four athletes, measured in meters. It has a mean of 1007 meters, and a standard deviation of 5 meters.

Standard deviation may serve as a measure of uncertainty. In physical science, for example, the reported standard deviation of a group of repeated measurements gives the precision of those measurements.

When deciding whether measurements agree with a theoretical prediction, the standard deviation of those measurements is of crucial importance: if the mean of the measurements is too far away from the prediction (with the distance measured in standard deviations), then the theory being tested probably needs to be revised.

This makes sense since they fall outside the range of values that could reasonably be expected to occur, if the prediction were correct and the standard deviation appropriately quantified.

While the standard deviation does measure how far typical values tend to be from the mean, other measures are available.

An example is the mean absolute deviation, which might be considered a more direct measure of average distance, compared to the root mean square distance inherent in the standard deviation.

14.5: STANDARD ERROR

The **standard error of the mean** (**SEM**) can be seen to depict the relationship between the dispersion of individual observations around the population mean (the standard deviation), and the dispersion of sample means around the population mean (the standard error). Different samples drawn from that same population would in general have different values of the sample mean, so there is a distribution of sampled means (with its own mean and variance).

The relationship with the standard deviation is defined such that, for a given sample size, the standard error equals the standard deviation divided by the square root of the sample size.

As the sample size increases, the dispersion of the sample means clusters more closely around the population mean and the standard error decreases.

In regression analysis, the term "standard error" is also used in the phrase standard error of the regression to mean the ordinary least squares estimate of the standard deviation of the underlying errors.

The term may also be used to refer to an estimate of that standard deviation, derived from a particular sample used to compute the estimate.

STANDARD ERROR OF THE MEAN

The **standard error of the mean** (SEM) is the standard deviation of the sample-mean's estimate of a population mean. (It can also be viewed as the standard deviation of the error in the sample mean with respect to the true mean, since the sample mean is an unbiased estimator.)

SEM is usually estimated by the sample estimate of the population standard deviation (sample standard deviation) divided by the square root of the sample size (assuming statistical independence of the values in the sample): where

s is the sample standard deviation (i.e., the sample-based estimate of the standard deviation of the population), and n is the size (number of observations) of the sample.

This estimate may be compared with the formula for the true standard deviation of the sample mean: SD x ‾ = σ n where σ is the standard deviation of the population.

This formula may be derived from what we know about the variance of a sum of independent random variables.

If X 1 , X 2 , … , X n are n independent observations from a population that has a mean μ

and standard deviation σ, then the variance of the total T = (X 1 + X 2 + ⋯ + X n) is n σ 2 .

The variance of T / n must be 1 n 2 n σ 2 = σ 2 n.

- And the standard deviation of must be .
- Of course, is the sample mean .

Note: the standard error and the standard deviation of small samples tend to systematically underestimate the population standard error and deviations: the standard error of the mean is a biased estimator of the population standard error. With n = 2 the underestimate is about 25%, but for n = 6 the underestimate is only 5%. Gurland and Tripathi (1971) provide a correction and equation for this effect. Sokal and Rohlf (1981)[give an equation of the correction factor for small samples of *n* < 20.

A practical result: Decreasing the uncertainty in a mean value estimate by a factor of two requires acquiring four times as many observations in the sample. Or decreasing standard error by a factor of ten requires a hundred times as many observations.

Student approximation when σ value is unknown

In many practical applications, the true value of σ is unknown. As a result, we need to use a distribution that takes into account that spread of possible σ's.

When the true underlying distribution is known to be Gaussian, although with unknown σ, then the resulting estimated distribution follows the Student t-distribution.

The standard error is the standard deviation of the Student t-distribution. T-distributions are slightly different from Gaussian, and vary depending on the size of the sample.

To estimate the standard error of a student t-distribution it is sufficient to use the sample standard deviation "s" instead of σ, and we could use this value to calculate confidence intervals.

Note: The Student's probability distribution is approximated well by the Gaussian distribution when the sample size is over 100. For such samples one can use the latter distribution, which is much simpler.

ASSUMPTIONS AND USAGE

If its sampling distribution is normally distributed, the sample mean, its standard error, and the quantiles of the normal distribution can be used to calculate confidence intervals for the mean. The following expressions can be used to calculate the upper and lower 95% confidence limits, where x ¯ is equal to the sample mean, S E is equal to the standard error for the sample mean, and 1.96 is the 0.975 quantile of the normal distribution: Upper 95% limit = x ¯ + (SE × 1.96) , and Lower 95% limit.

In particular, the standard error of a sample statistic (such as sample mean) is the estimated standard deviation of the error in the process by which it was generated. In other words, it is the standard deviation of the sampling distribution of the sample statistic. The notation for standard error can be any one of SE, SEM (for standard error of *measurement* or *mean*), or S_E.

Standard errors provide simple measures of uncertainty in a value and are often used because:

- If the standard error of several individual quantities is known then the standard error of some function of the quantities can be easily calculated in many cases;
- Where the probability distribution of the value is known, it can be used to calculate a good approximation to an exact confidence interval; and
- Where the probability distribution is unknown, relationships like Chebyshev's or the Vysochanskiï–Petunin inequality can be used to calculate a conservative confidence interval
- As the sample size tends to infinity the central limit theorem guarantees that the sampling distribution of the mean is asymptotically normal.

STANDARD ERROR OF MEAN VERSUS STANDARD DEVIATION

In scientific and technical literature, experimental data are often summarized either using the mean and standard deviation or the mean with the standard error. This often leads to confusion about their interchangeability.

However, the mean and standard deviation are descriptive statistics, whereas the standard error of the mean describes bounds on a random sampling process.

Despite the small difference in equations for the standard deviation and the standard error, this small difference changes the meaning of what is being reported from a description of the variation in measurements to a probabilistic statement about how the number of samples will provide a better bound on estimates of the population mean, in light of the central limit theorem.

Put simply, the **standard error** of the sample mean is an estimate of how far the sample mean is likely to be from the population mean, whereas the **standard deviation** of the sample is the degree to which individuals within the sample differ from the sample mean.

If the population standard deviation is finite, the standard error of the mean of the sample will tend to zero with increasing sample size, because the estimate of the population mean will improve, while the standard deviation of the sample will tend to approximate the population standard deviation as the sample size increases.

CORRECTION FOR FINITE POPULATION

The formula given above for the standard error assumes that the sample size is much smaller than the population size, so that the population can be considered to be effectively infinite in size.

This is usually the case even with finite populations, because most of the time, people are primarily interested in managing the processes that created the existing finite population; this is called an analytic study, following W. Edwards Deming. If people are interested in managing an existing finite population that will not change over time, then it is necessary to adjust for the population size; this is called an enumerative study.

When the sampling fraction is large (approximately at 5% or more) in an enumerative study, the estimate of the standard error must be corrected by multiplying by a "finite population correction" to account for the added precision gained by sampling close to a larger percentage of the population. The effect of the FPC is that the error becomes zero when the sample size n is equal to the population size N.

RELATIVE STANDARD ERROR

The **relative standard error** of a sample mean is the standard error divided by the mean and expressed as a percentage. It can only be calculated if the mean is a non-zero value.

As an example of the use of the relative standard error, consider two surveys of household income that both results in a sample mean of $50,000.

If one survey has a standard error of $10,000 and the other has a standard error of $5,000, then the relative standard errors are 20% and 10% respectively.

The survey with the lower relative standard error can be said to have a more precise measurement, since it has proportionately less sampling variation around the mean. In fact, data organizations often set reliability standards that their data must reach before publication.

For example, the U.S. National Center for Health Statistics typically does not report an estimated mean if its relative standard error exceeds 30%. (NCHS also typically requires at least 30 observations – if not more – for an estimate to be reported.)

INTRODUCTION TO THE STANDARD ERROR FOR NOVICES

The standard error is a quantitative measure of uncertainty. Consider the following scenarios. Scenario 1.

For an upcoming national election, 2000 voters are chosen at random and asked if they will vote for candidate A or candidate B. Of the 2000 voters, 1040 (52%) state that they will vote for candidate A.

The researchers report that candidate A is expected to receive 52% of the final vote, with a margin of error of 2%.

In this scenario, the 2000 voters are a sample from all the actual voters. The sample proportion of 52% is an estimate of the true proportion who will vote for candidate A in the actual election.

The margin of error of 2% is a quantitative measure of the uncertainty – the possible difference between the true proportion who will vote for candidate A and the estimate of 52%.

Scenario 2. A medical research team tests a new drug to lower cholesterol. They report that, in a sample of 400 patients, the new drug lowers cholesterol by an average of 20 units (mg/dL).

The 95% confidence interval for the average effect of the drug is that it lowers cholesterol by 18 to 22 units. In this scenario, the 400 patients are a sample of all patients who may be treated with the drug. The confidence interval of 18 to 22 is a quantitative measure of the uncertainty – the possible difference between the true average effect of the drug and the estimate of 20 mg/dL.

In each of these scenarios, a sample of observations is drawn from a large population. The proportion or the mean is calculated using the sample.

Because of random variation in sampling, the proportion or mean calculated using the sample will usually differ from the true proportion or mean in the entire population.

A quantitative measure of uncertainty is reported: a margin of error of 2%, or a confidence interval of 18 to 22. The margin of error and the confidence interval are based on a quantitative measure of uncertainty: the standard error.

The standard error of a proportion and the standard error of the mean describe the possible variability of the estimated value based on the sample around the true proportion or true mean

Standard errors are used in many hypothesis tests, such as t-tests. They may be used to calculate confidence intervals.

STANDARD ERROR OF THE MEAN (SEM)

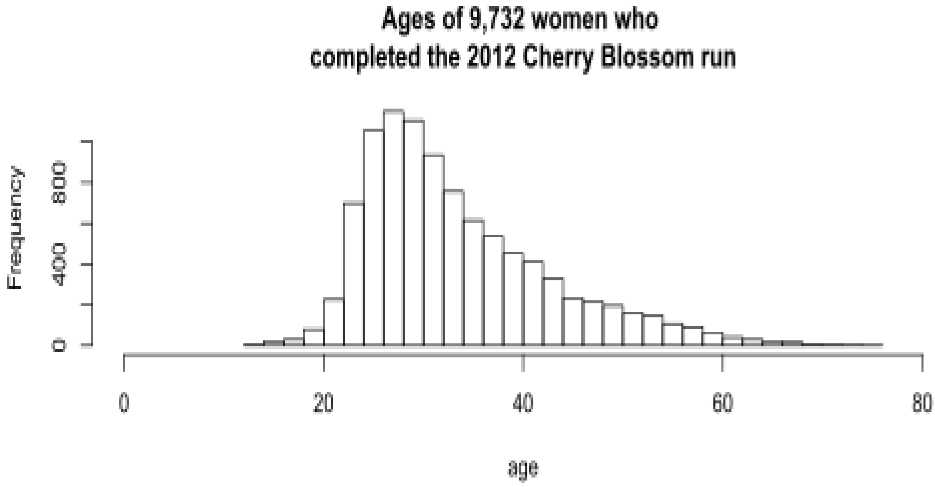
This section will focus on the standard error of the mean. Later sections will present the standard error of other statistics, such as the standard error of a proportion, the standard error of the difference of two means, the standard error of the difference of two proportions and so on.

The concept of a sampling distribution is key to understanding the standard error. Two data sets will be helpful to illustrate the concept of a sampling distribution and its use to calculate the standard error. As will be shown, the standard error is the standard deviation of the sampling distribution.

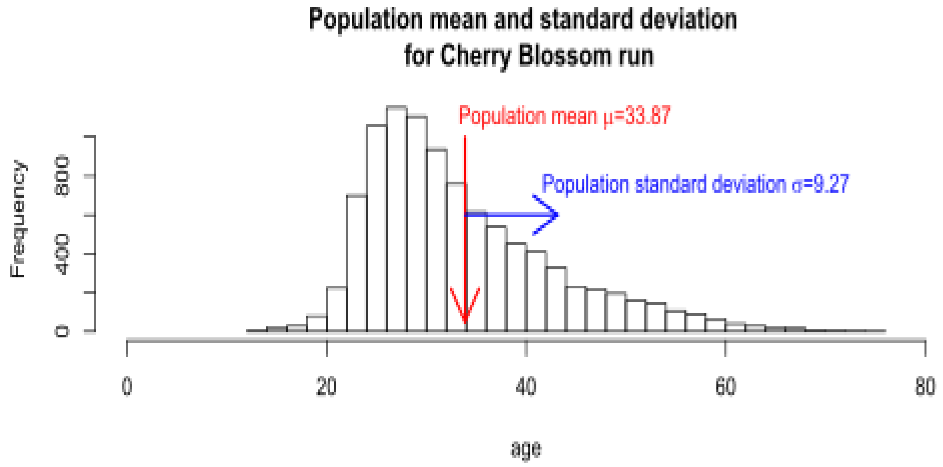
Sampling from a distribution with a large standard deviation

The first data set consists of the ages of 9,732 women who completed the 2012 Cherry Blossom run, a 10-mile race held in Washington each spring.

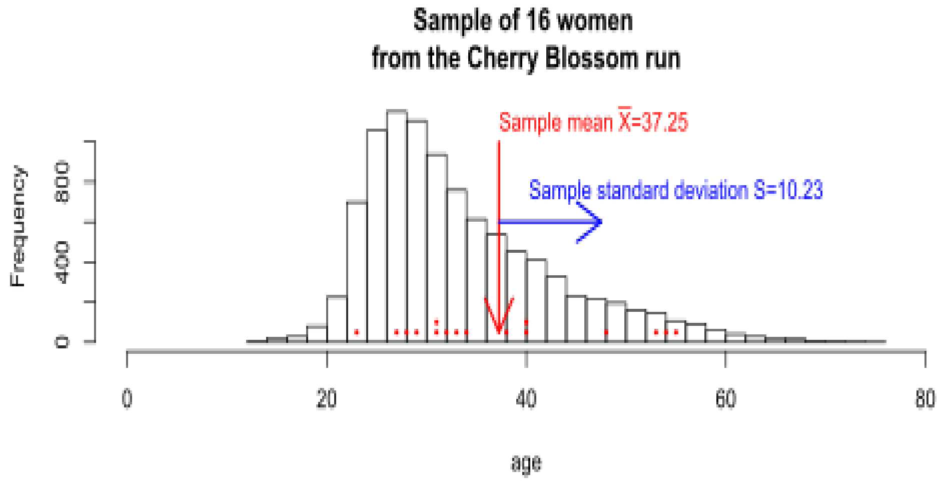
The age data are in the data set run10 from the R package openintro that accompanies the textbook by Dietz .The graph shows the distribution of ages for the runners.


Ages of 9,732 women who completed the 2012 Cherry Blossom run

For the purpose of this example, the 9,732 runners who completed the 2012 run are the entire population of interest. The mean age was 33.88 years. The standard deviation of the age was 9.27 years. Because the 9,732 runners are the entire population, 33.88 years is the population mean, , and 9.27 years is the population standard deviation, σ. Greek letters indicate that these are population values.



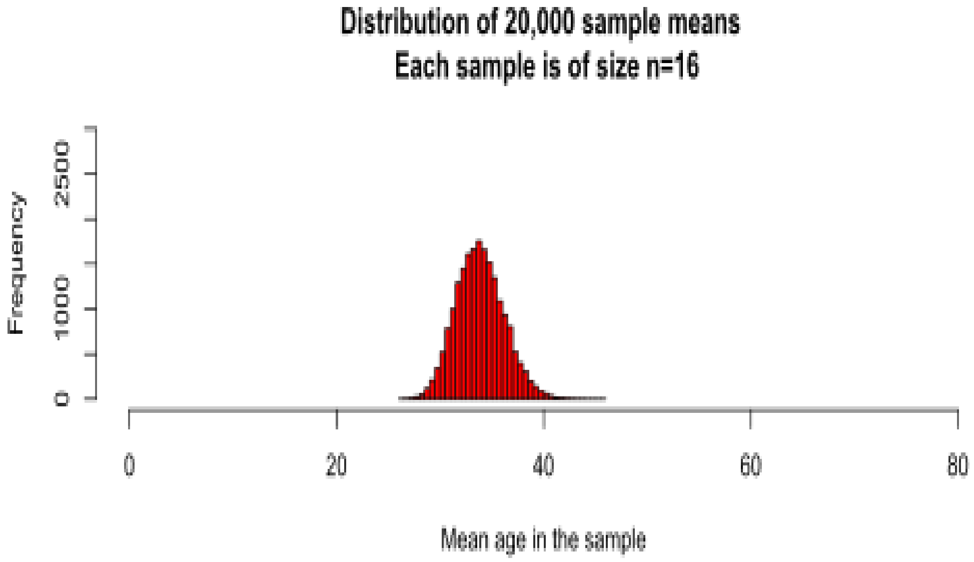
Consider a sample of n=16 runners selected at random from the 9,732. The ages in one such sample are 23, 27, 28, 29, 31, 31, 32, 33, 34, 38, 40, 40, 48, 53, 54, and 55. The graph shows the ages for the 16 runners in the sample, plotted on the distribution of ages for all 9,732 runners.



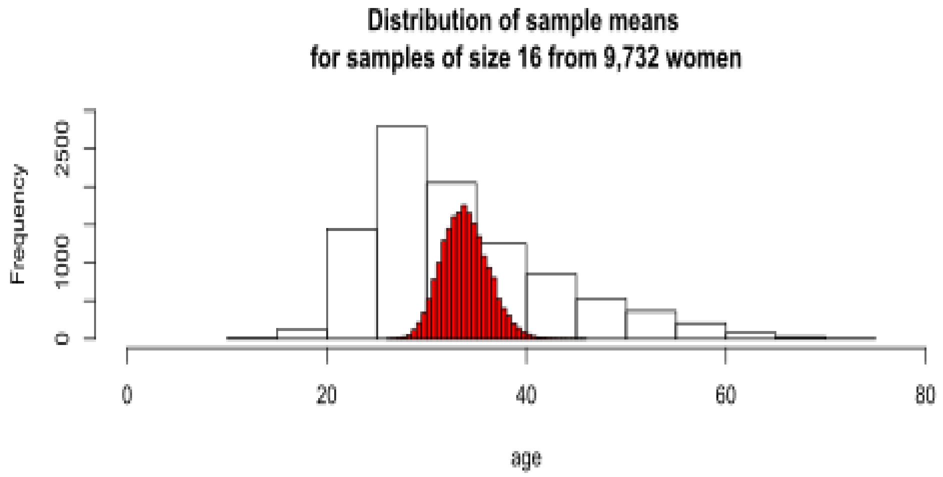
The mean age for the 16 runners in this particular sample is 37.25. The standard deviation of the age for the 16 runners is 10.23. Because these 16 runners are a sample from the population of 9,732 runners, 37.25 is the sample mean, and 10.23 is the sample standard deviation, s. Roman letters indicate that these are sample values.

The sample mean = 37.25 is greater than the true population mean = 33.88 years. The sample standard deviation s = 10.23 is greater than the true population standard deviation σ = 9.27 years. For any random sample from a population, the sample mean will very rarely be equal to the population mean. Similarly, the sample standard deviation will very rarely be equal to the population standard deviation.

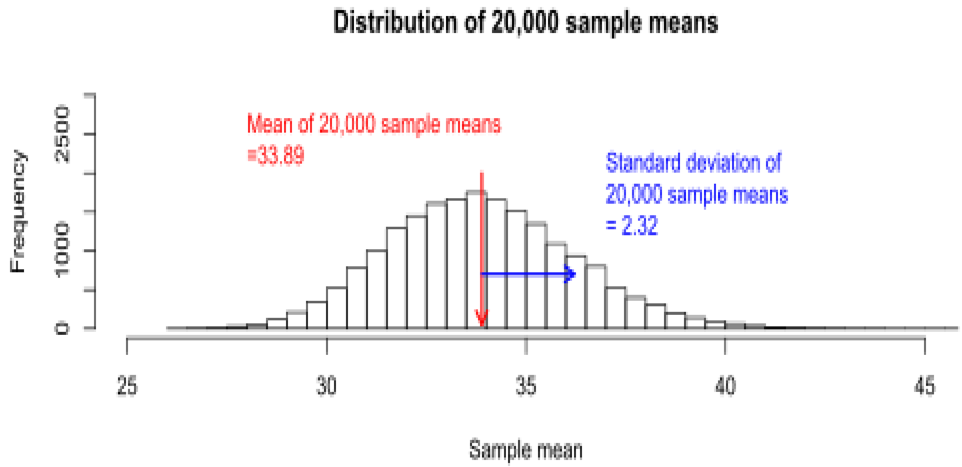
Next, consider all possible samples of 16 runners from the population of 9,732 runners. For each sample, the mean age of the 16 runners in the sample can be calculated. The distribution of the mean age in all possible samples is called the sampling distribution of the mean. For illustration, the graph below shows the distribution of the sample means for 20,000 samples, where each sample is of size n=16.



The next graph shows the sampling distribution of the mean (the distribution of the 20,000 sample means) superimposed on the distribution of ages for the 9,732 women.

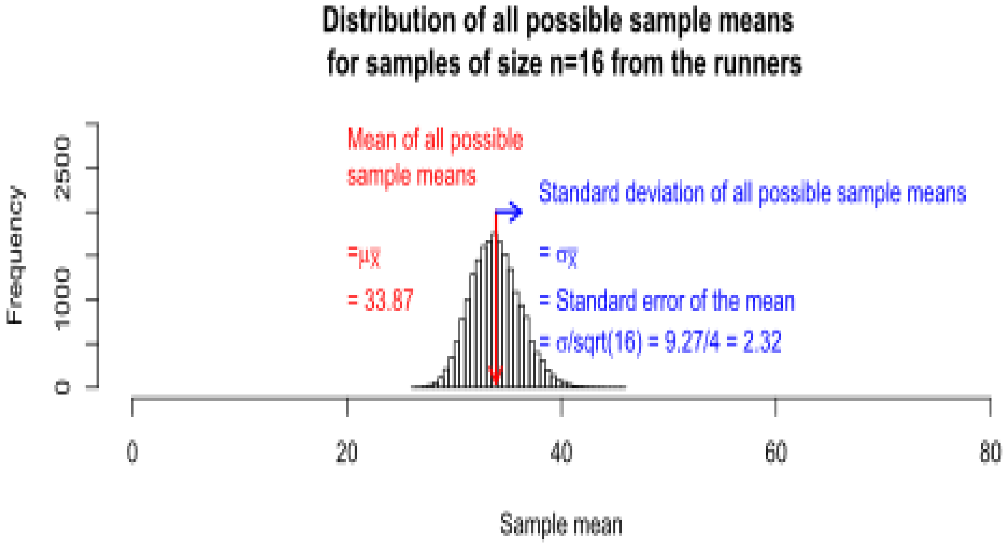


The distribution of these 20,000 sample means indicate how far the mean of a sample may be from the true population mean. A natural way to describe the variation of these sample means around the true population mean is the standard deviation of the distribution of the sample means.



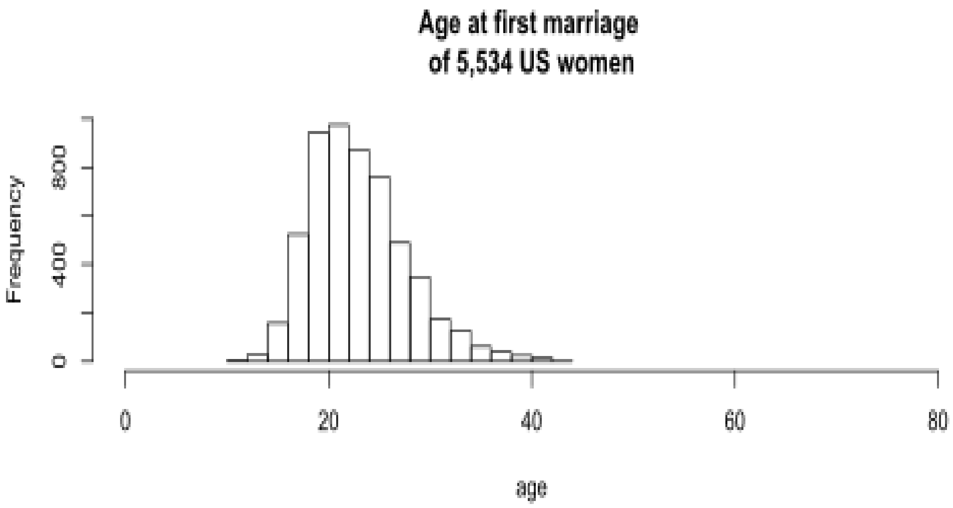
The mean of all possible sample means is equal to the population mean. For the runners, the population means age is 33.87, and the population standard deviation is 9.27. It will be shown that the standard deviation of all possible sample means of size n=16 is equal to the population standard deviation, σ, divided by the square root of the sample size, 16. This gives

9.27/sqrt(16) = 2.32. The standard deviation of all possible sample means is the standard error, and is represented by the symbol.

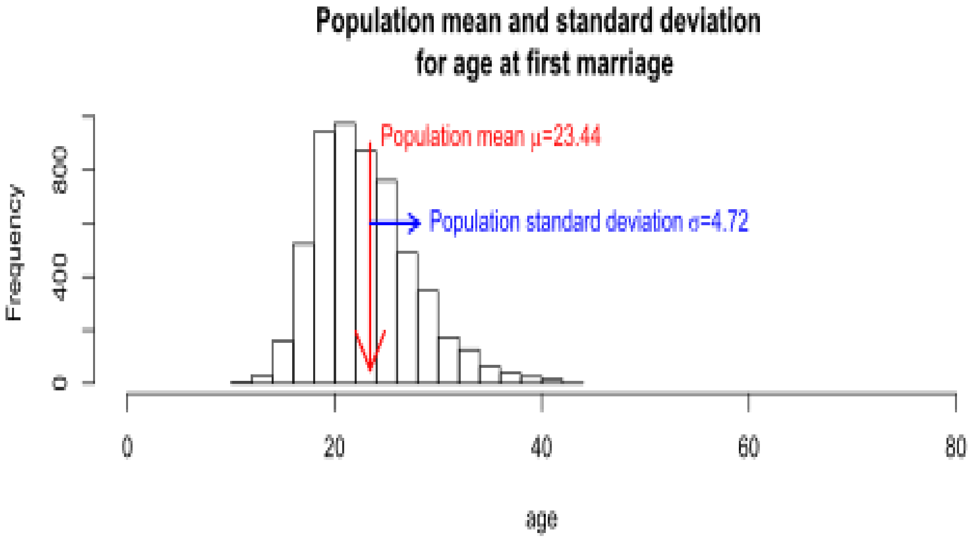


SAMPLING FROM A DISTRIBUTION WITH A SMALL STANDARD DEVIATION

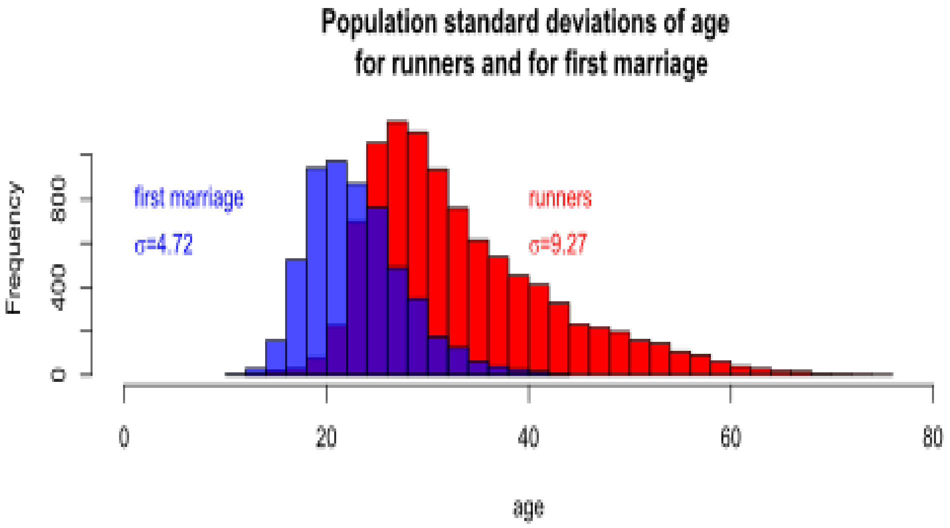
The second data set consists of the age at first marriage of 5,534 US women who responded to the National Survey of Family Growth (NSFG) conducted by the CDC in the 2006 and 2010 cycle. The data set is ageAtMar, also from the R package openintro from the textbook by Dietz et al.



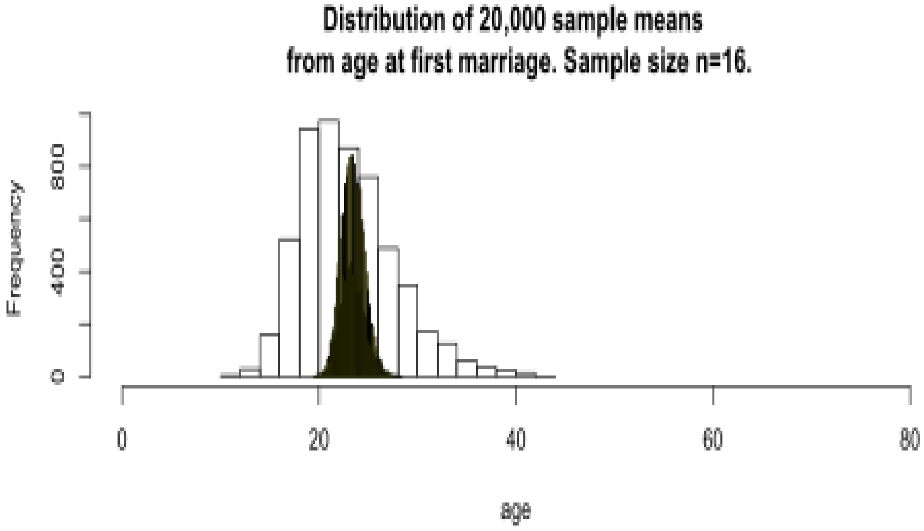
For the purpose of this example, the 5,534 women are the entire population of interest. The mean age was 23.44 years. The standard deviation of the age was 4.72 years. Because the 5,534 women are the entire population, 23.44 years is the population mean, and 4.72 years is the population standard deviation, .



Notice that the population standard deviation of 4.72 years for age at first marriage is about half the standard deviation of 9.27 years for the runners. The smaller standard deviation for age at first marriage will result in a smaller standard error of the mean.

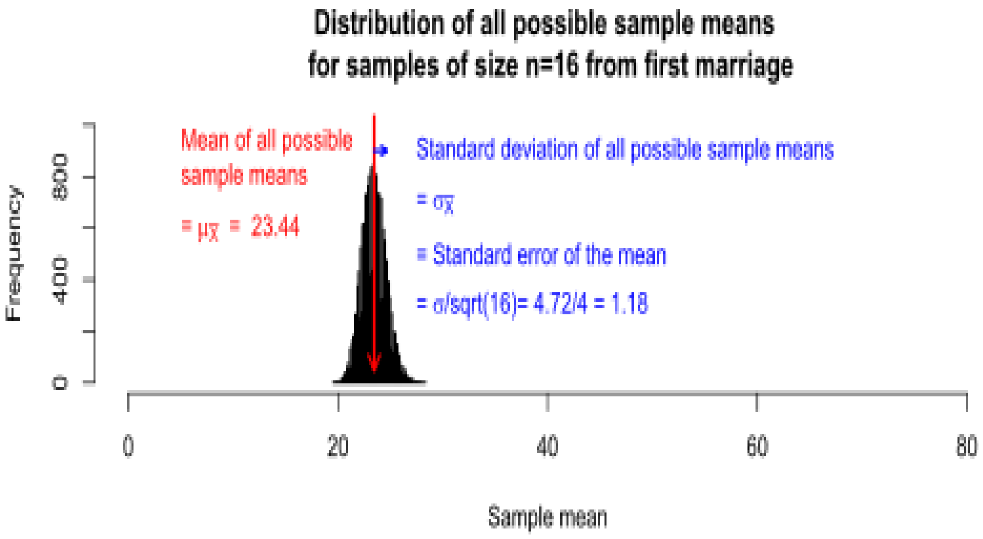


Repeating the sampling procedure as for the Cherry Blossom runners, take 20,000 samples of size n=16 from the age at first marriage population. The graph below shows the distribution of the sample means for 20,000 samples, where each sample is of size n=16.

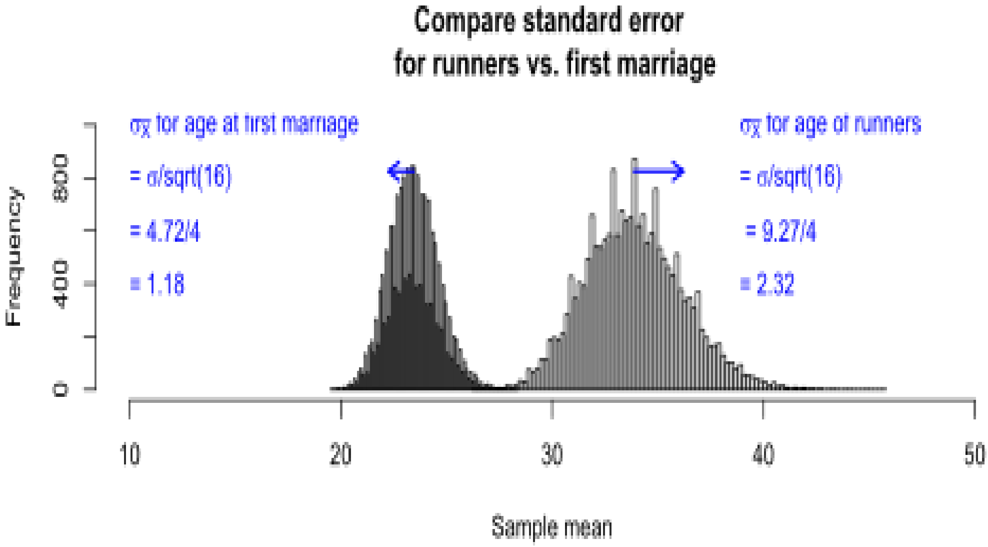


The mean of these 20,000 samples from the age at first marriage population is 23.44, and the standard deviation of the 20,000 sample means is 1.18.

As will be shown, the mean of all possible sample means is equal to the population mean. For the age at first marriage, the population mean age is 23.44, and the population standard deviation is 4.72. It will be shown that the standard deviation of all possible sample means of size n=16 is equal to the population standard deviation, σ, divided by the square root of the sample size, 16, which is 4.72/sqrt(16) = 1.18. The standard deviation of all possible sample means of size 16 is the standard error.



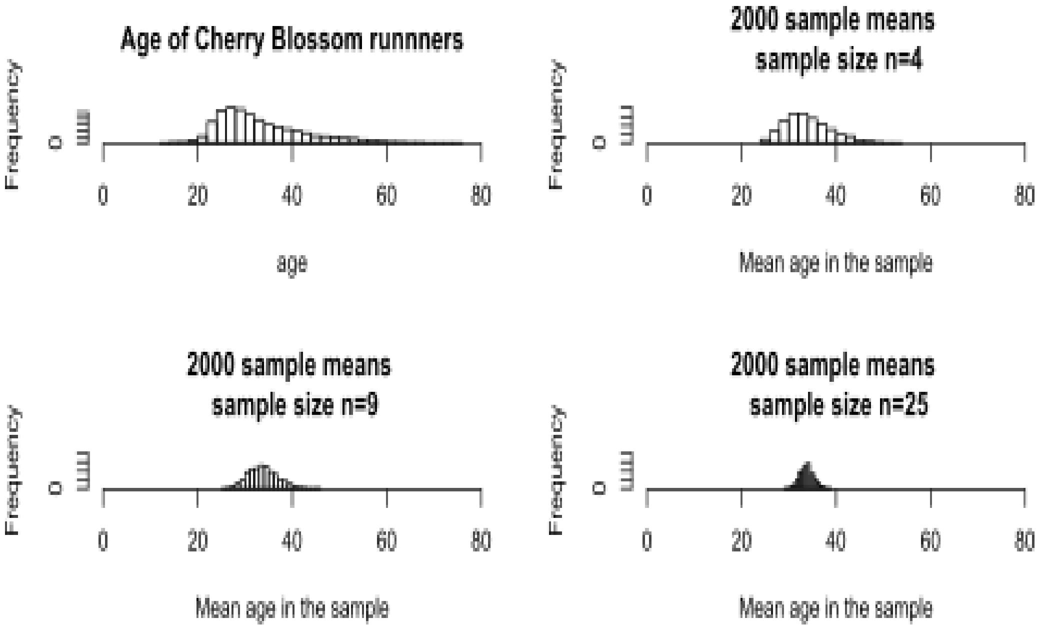
It is useful to compare the standard error of the mean for the age of the runners versus the age at first marriage, as in the graph.



Because the age of the runners has a larger standard deviation (9.27 years) than does the age at first marriage (4.72 years), the standard error of the mean is larger for the runners than for first marriage.

LARGER SAMPLE SIZES GIVE SMALLER STANDARD ERRORS

As would be expected, larger sample sizes give smaller standard errors. The graphs below show the sampling distribution of the mean for samples of size 4, 9, and 25. As the sample size increases, the sampling distribution becomes narrower, and the standard error decreases.



USING A SAMPLE TO ESTIMATE THE STANDARD ERROR

In the examples so far, the population standard deviation σ was assumed to be known. If σ is known, the standard error is calculated using the formula

Where

σ is the standard deviation of the population.

n is the size (number of observations) of the sample.

It is rare that the true population standard deviation is known. However, the sample standard deviation, s, is an estimate of σ. If σ is not known, the standard error is estimated using the formula:

Where *s* is the sample standard deviation

n is the size (number of observations) of the sample.

Notice that is only an estimate of the true standard error, .

In an example above, n=16 runners were selected at random from the 9,732 runners. The ages in that sample were 23, 27, 28, 29, 31, 31, 32, 33, 34, 38, 40, 40, 48, 53, 54, and 55.

The standard deviation of the age for the 16 runners is 10.23, which is somewhat greater than the true population standard deviation σ = 9.27 years. Compare the true standard error of the mean to the standard error estimated using this sample.

The true standard error of the mean, using σ = 9.27, is

The standard error of the mean estimated by using the sample standard deviation, s = 10.23, is

The true standard error of the mean is 2.32. The standard error estimated using the sample standard deviation is 2.56. For the purpose of hypothesis testing or estimating confidence intervals, the standard error is primarily of use when the sampling distribution is normally distributed, or approximately normally distributed.

These assumptions may be approximately met when the population from which samples are taken is normally distributed, or when the sample size is sufficiently large to rely on the Central Limit Theorem.

14.6 SUMMARY

Circumstances generally dictate which measure of central tendency—mean, median or mode—is the most appropriate.

If you are interested in a total, the mean tends to be the most meaningful measure of central tendency because it is the total divided by the number of data.

For example, the mean income of the individuals in a family tells you how much each family member can spend on life's necessities.

The median measure is good for finding the central value and the mode is used to describe the most typical case.

14.7: REFERENCES

- Hayden, Erika Check (8 February 2012). "Biostatistics: Revealing analysis". Nature. 482 (7384): 263–265. Doi: 10.1038/nj7384-263a.
- Efron, Bradley (February 2008). "Microarrays, Empirical Bayes and the Two-Group Model". Statistical Science. **23** (1): 1–22. Doi: 10.1214/07-STS236.
- Helen Causton; John Quackenbush; Alvis Brazma (2003). Statistical Analysis of Gene Expression Microarray Data. Wiley-Blackwell.
- Terry Speed (2003). Microarray Gene Expression Data Analysis: A Beginner's Guide. Chapman & Hall/CRC.
- Frank Emmert-Streib; Matthias Dehmer (2010). Medical Biostatistics for Complex Diseases. Wiley-Blackwell. ISBN 3-527-32585-9.
- Warren J. Ewens; Gregory R. Grant (2004). Statistical Methods in Bioinformatics: An Introduction. Springer.

- Matthias Dehmer; Frank Emmert-Streib; Armin Graber; Armindo Salvador (2011). Applied Statistics for Network Biology: Methods in Systems Biology. Wiley-Blackwell. ISBN 3-527-32750-9.
- Hippel, Paul T. von (2005). "Mean, Median, and Skew: Correcting a Textbook Rule". J. of Statistics Education. 13 (2).
- Bottomley, H. (2004). "Maximum distance between the mode and the mean of a unimodal distribution" (PDF). Unpublished preprint.
- Simon, Laura J.; "Descriptive statistics", Statistical Education Resource Kit, Pennsylvania State Department of Statistics.
- David J. Sheskin (27 August 2003). Handbook of Parametric and Nonparametric Statistical Procedures: Third Edition. CRC Press. pp. 7–. ISBN 978-1-4200-3626-8. Retrieved 25 February 2013.
- Everitt, B.S. (2003) The Cambridge Dictionary of Statistics, CUP. ISBN 0-521-81099-X.
- Kenney, J. and Keeping, E.S. (1963) Mathematics of Statistics, van Nostrand, p. 187.
- Zwillinger D. (1995), Standard Mathematical Tables and Formulae, Chapman&Hall/CRC. ISBN 0-8493-2479-3 p. 626.
- T.P. Hutchinson, Essentials of statistical methods in 41 pages.
- Gurland, J; Tripathi RC (1971). "A simple approximation for unbiased estimation of the standard deviation". American Statistician. American Statistical Association. 25 (4): 30–32. doi:10.2307/2682923. JSTOR 2682923.
- Sokal and Rohlf (1981) Biometry: Principles and Practice of Statistics in Biological Research , 2nd ed. ISBN 0-7167-1254-7 , p 53.
- Barde, M. (2012). "What to use to express the variability of data: Standard deviation or standard error of mean?". Perspect Clin Res. 3 (3): 113–116. doi:10.4103/2229-3485.100662.
- Isserlis, L. (1918). "On the value of a mean as calculated from a sample". Journal of the Royal Statistical Society. Blackwell Publishing. 81 (1): 75–81. Doi: 10.2307/2340569. JSTOR 2340569. (Equation 1).
- Derek Bissell (1994). Statistical Methods for Spc and Tqm. CRC Press. pp. 26–. ISBN 978-0-412-39440-9. Retrieved 25 February 2013. "Journal of Statistics Education, v13n2: Paul T. von Hippel". amstat.org.

- Robson, Colin (1994). Experiment, Design and Statistics in Psychology. Penguin. pp. 42–45. ISBN 0-14-017648-9.
- "AP Statistics Review - Density Curves and the Normal Distributions". Retrieved 16 March 2015.
- Newman, Mark EJ. "Power laws, Pareto distributions and Zipf's law." Contemporary physics 46.5 (2005): 323–351.
- Stroock, Daniel (2011). Probability Theory. Cambridge University Press. p. 43. ISBN 978-0-521-13250-3."An Error Occurred Setting Your User Cookie". siam.org.
- Mallows, Colin (August 1991). "Another comment on O'Cinneide". The American Statistician. 45 (3): 257. doi:10.1080/00031305.1991.10475815.
- Apache Lucene. [(20 October 2009, date last accessed)]. http://lucene.apache.org/java/docs/index.html.
- Cochrane G, Akhtar R, Bonfield J, et al. Petabyte-scale innovations at the European Nucleotide Archive. Nucleic Acids Res. 2009; 37:D19–25. [PMC free article]
- The Universal Protein Resource (UniProt) 2009. The UniProt Consortium. Nucleic Aci.