SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOTECHNOLOGY

UNIT – I -STATISTICAL METHODS– SBT1002

I. Introduction

1.1 Research methodology-basic concepts-types

1.1.1. Definitions and Meanings of Research

- Research in common parlance refers to a search for knowledge.
- Research is actually a voyage of discovery.
- Redman defined research as a "systematized effort to gain new knowledge"
- Research is the pursuit of truth with the help of study, observation , comparison and experiment.
- Research is thus, an original contribution to the existing stock of knowledge making for its advancement.

1.1.2. Objectives of Research

- The research objective is to gain familiarity with a phenomenon or to achieve new insight into it.
- The main aim of the research is to find out the truth which is hidden and which has not been discovered yet.
- To test a hypothesis of a causal relationship between variables (hypothesis testing)

1.1.3. Motivation in Research

- Directives of the Government and Organizations
- Curiosity about new things
- Desire to understand causal relationship
- Social thinking, awakening
- Employment and economic conditions
- Desire to get respect in the society
- Desire to get intellectual joy of creativity

1.2. Types of Research

- Descriptive Research
- Analytical Research
- Applied Research
- Fundamental or Basic Research
- Quantitative Research

- Qualitative Research
- Empirical Research
- Conceptual Research

1.2.1. Descriptive research:-it includes surveys and fact-finding enquiries of different kinds. The major purpose of descriptive research is description of the state of affairs as it exists at present. In social science and business research we quite often use the term EX-Post-Facto research for descriptive research studies.

The main characteristic of this method is that the researcher has no control over the variables, he can only report what has happened or what is happening. Most Ex-Post-Facto research projects are used for descriptive studies in which the researcher seeks to measure such items as for e.g:- frequency of shopping, preferences of people, or similar data.

1.2.2. **Analytical research**:- on the other hand, the researcher has to use facts or information already available, and analyze these to make a critical evaluation of the subject or material

1.2.3. **Applied Research**:- it aims at finding a solution for an immediate problem facing a society or an industry

1.2.4. **Fundamental or Basic research**:- it is mainly concerned with generalizations and with the formulation of a theory. "Gathering knowledge for knowledge's sake is termed 'pure' or 'basic' research.

1.2.5. **Quantitative research**:-is based on the measurement of quantity or amount. It is applicable to phenomena that can be expressed in terms of quantity.

1.2.6. **Quantitative research**:-is based on the measurement of quantity or amount. It is applicable to phenomena that can be expressed in terms of quantity.

1.2.7. **Conceptual research** is that related to some abstract idea or theory. It is generally used by philosophers and thinkers to develop new concepts or to reinterpret existing ones.

1.2.8. **Empirical Research**:- relies on experience or observation alone, often without due regard for system and theory. It is data based research, coming up with conclusions which are capable of being verified by observation or experiment.

1.2.9. **Some Other Types of Research**:

One-time research or longitudinal research: In the former case the research is confined to a single time-period, whereas in the latter case the research is carried on over several time-periods. Research can be field-setting research or laboratory research or simulation research, depending upon the environment in which it is to be carried out.

Clinical or diagnostic research. Such research follow case-study methods or indepth approaches to reach the basic causal relations. Such studies usually go deep into the causes of things or events that interest us, using very small samples and very deep probing data gathering devices.

Exploratory or it may be formalized. The objective of exploratory research is the development of hypotheses rather than their testing, whereas formalized research studies are those with substantial structure and with specific hypotheses to be tested.

Historical research is that which utilizes historical sources like documents, remains, etc. to study events or ideas of the past, including the philosophy of persons and groups at any remote point of time.

Conclusion-oriented and decision-oriented. While doing conclusionoriented research, a researcher is free to pick up a problem, redesign the enquiry as he proceeds and is prepared to conceptualize as he wishes. Decision-oriented research is always for the need of a decision maker and the researcher in this case is not free to embark upon research according to his own inclination. Operations research is an example of decision oriented research since it is a scientific method of providing executive departments with a quantitative basis for decisions regarding operations under their control.

1.3. Significance of Research

- All progress is born of inquiry.
- Doubt is better than over-confidence, for it leads to inquiry, and inquiry leads to invention. (famous Hudson Maxim).
- Research provides the basis for all Govt policies in our economic system.
- Research is “For Prognosis” (for the prediction of future.)
- To understand the problems and plights of farmers, labors, small business people, consumers, students, political issues, defense forces, Space, Health, Social Reforms

1.4. Approaches to Research

- There are two types of approaches to research, i.e. quantitative approach and qualitative approach.
- The Quantitative approach be further classified into inferential, experimental, simulation approach to research.
- Qualitative approach to research is concerned with subjective assessment of attitudes, opinions and behavior.

Thus, research is the fountain of knowledge for the sake of knowledge and an important source of providing guidelines for solving different business, governmental and social problems. It is a sort of formal training which enables one to understand the new development in one’s field in a better way

2. Research Methods versus Methodology

2.1. **Research methods** may be understood as all those methods/techniques that are used

for conduction of research. Research methods or techniques*, thus, refer to the methods the researchers use in performing research operations.

In other words, all those methods which are used by the researcher during the course of studying his research problem are termed as research methods.

Since the object of research, particularly the applied research, it to arrive at a solution for a given problem, the available data and the unknown aspects of the problem have to be related to each other to make a solution possible.

Keeping this in view, research methods can be put into the following three groups:

1. In the first group we include those methods which are concerned with the collection of data. These methods will be used where the data already available are not sufficient to arrive at the required solution;

2. The second group consists of those statistical techniques which are used for establishing relationships between the data and the unknowns;

3. The third group consists of those methods which are used to evaluate the accuracy of the results obtained.

2.2. Research methodology is a way to systematically solve the research problem. It may be understood as a science of studying how research is done scientifically.

In it we study the various steps that are generally adopted by a researcher in studying his research problem along with the logic behind them.

It is necessary for the researcher to know not only the research methods/techniques but also the methodology.

Researchers not only need to know how to develop certain indices or tests, how to calculate the mean, the mode, the median or the standard deviation or chi-square, how to apply particular research techniques, but they also need to know which of these methods or techniques, are relevant and which are not, and what would they mean and indicate and why.

Researchers also need to understand the assumptions underlying various techniques and they need to know the criteria by which they can decide that certain techniques and procedures will be applicable to certain problems and others will not. All this means that it is necessary for the researcher to design his methodology for his problem as the same may differ from problem to problem.

For example, an architect, who designs a building, has to consciously evaluate the basis of his decisions, i.e., he has to evaluate why and on what basis he selects particular size, number and location of doors, windows and ventilators, uses particular materials and not others and the like.

Similarly, in research the scientist has to expose the research decisions to evaluation before they are implemented. He has to specify very clearly and precisely what decisions he selects and why he selects them so that they can be evaluated by others also.

From what has been stated above, we can say that research methodology has many dimensions and research methods do constitute a part of the research methodology. The scope of research methodology is wider than that of research methods.

Thus, when we talk of research methodology we not only talk of the research methods but also consider the logic behind the methods we use in the context of our research study and explain why we are using a particular method or technique and why we are not using others so that research results are capable of being evaluated either by the researcher himself or by others. Why a research study has been undertaken, how the research problem has been defined, in what way and why the hypothesis has been formulated, what data have been collected and what particular method has been adopted, why particular technique of analysing data has been used and a host of similar other questions are usually answered when we talk of research methodology concerning a research problem or study.

3. Research Process

Research is a search for knowledge. It helps in taking appropriate decisions. Research involves asking a question and then trying to find an answer to it. Research is essentially a systematic, scientific and structured inquiry seeking facts through objective methods. Therefore a research must have a clearly defined step-by-step process. A knowledge of the research process is essential both for those who conduct the research and for those who wish to be benefited by the conclusions drawn from the research. A meaningful knowledge should have a definite purpose and direction. In developing a research process, one would like to list the sequence of step-bystep activities. In a research process these steps are inter- dependent and may overlap each other. They may not follow a strict sequence and the researcher has to be vigilant of their order continuously through out the research process. However, one can broadly enlist the main steps involved in a research process as a procedural guideline to the researcher. These steps are:

- Problem formulation
- Literature survey
- Development of hypothesis
- Research design
- Choice of sample design
- Data collection
- Analysis and interpretation of data
- Hypothesis testing
- Interpretation of results
- Report writing

A brief description of the above steps is given below.

3.1. Problem formulation

Formulation of a problem is the first and foremost step in a research process. It is not always easy to identify and define a problem in an ever-changing business environment. A researcher not only discovers and defines a problem area but also a specific problem within that area concerning his interest in business. The problem should be clearly and precisely stated. The statement of the problem must be complete.

The problems in a business may sometimes be obvious and one can pinpoint them. Many a problems is not so apparent and needs explorations. Thus, first of all one has to identify a problem specifically and thoroughly, and then it has to be expressed in scientific terms so that statistical analysis can be performed on that problem.

3.2. Literature survey

After the formulation and identification of a problem, the next important step is the review of literature survey. An exhaustive and critical review of professional literature familiarizes the researcher with the current state of knowledge. It helps in understanding of the problems and hypothesis that others have studied. It clarifies the concepts, theories, major variables involved, operational definitions and research methods used in the past. This contributes to the cumulative nature of scientific knowledge.

Every year thousands of articles, books and monographs are published in any field of study. Therefore, it is important to sort out the relevant literature connected with the field of one's interest. It is best to begin any search for literature with one of the guides to published literature. These guides are increasingly computerized and include bibliographies, indexes and abstracts.

With the advent of Internet the modern life has changed drastically. One can find an ocean of information within the four walls of one's study room through Internet. Some of the popular search engines like google, yahoo and rediff are becoming more popular in searching for literature on any topic.

3.3. Development of hypothesis

Once a problem is defined and a review of literature is made, the next step is to define clearly the hypothesis in a research problem. A hypothesis is a tentative assumption in a research problem, which has to be tested empirically with the help of

observed data. When formulating a hypothesis, a researcher does not know whether it will be rejected or accepted. A hypothesis is constructed and tested; if it is rejected, another one is formulated; if it is accepted, it is incorporated in the scientific body of knowledge.

One should arrive at a clear and specific working hypothesis for which research methods already exist. A good hypothesis states a research problem in concise and precise terms so that the researcher is focused on the problem at hand.

3.4. Research design

A research design is a well-defined plan of action. It is a planned sequence of the entire research process. It is a blue print of research activity. In a big business scenario designing a research study is very complex. Therefore, a research design may change during the operation of a project. A good research design must use minimum of resources like time, money and manpower. A research design must be able to translate the general scientific model into a practical research operation.

A scientifically developed research design possess the characteristics like (i) objectivity (ii) Reliability (iii) Validity (iv) Generalization .5. Choice of sample design

In any investigation the group of all items, objects or individuals under study is called ‘population’ or ‘universe’. For all purpose of determining population characteristics, instead of enumerating entire population, some items of the population, called a sample, are observed. Then the sample characteristics are utilized to approximately determine or estimate the population. For example, on examining the sample of particular product, we arrive at a decision of purchasing or rejecting that product. There will be surely some error in such an approximation and is inherent and unavoidable in any and every sampling scheme. But samples results in considerable gains in terms of time, money, accuracy and efforts.

Drawing a sample of some predetermined size from an entire population is not a child’s play. These have to be a systematic plan to choose the sample items. This plan or a technique of drawing a sample is known as sample design or sample plan or sampling technique. Researchers have suggested various sample designs. One research situation may be different from another, therefore, simple random sample, though most

popular, may not be suitable in each case. Depending on the requirement of a situation one can choose one of the following sample designs:

1. Purposive or Judgment Sampling
2. Simple Random Sampling
3. Stratified Random Sampling
4. Systematic Sampling
5. Cluster Sampling
6. Area Sampling
7. Multi-stage Sampling
8. Multi-phase Sampling

6. Data collection:

Once a sample design is formulated, the next step in a research process is the collection of relevant data.

There can be two sources of data (i) Internal data, that refers to the happenings and functions of a business organization. For example, the salepurchases details of a company. (ii) External data, which is related to outside sources and external agencies.

There are two types of data (a) Primary data (b) Secondary data.
Primary data are those observations, which are collected by an investigator for the first time, Secondary data are already available in the records and have been collected by some other researcher for the purpose of studying a similar problem.

Methods of collecting primary data are:

(i) by direct observation and experimentations
(ii) by direct personal interview
(iii) by direct interviews through phone, sms and email.
(iv) by indirect personal interview
(v) by mailed / emailed Questionnaire

(vi) by schedules through enumeratorsMethods of collecting secondary data:

(i) International organizations like WHO, UNO etc.

(ii) Government publications like economic survey, CSO, NSSO.

(iii) Journal and Newspapers

(iv) Research articles

(v) Reports of business organization and financial institutions.

A method and source of data collection is chosen by an investigator taking into account the objectives and requirements of the inquiry. The adopted method should incur minimum cost and time should have a reasonable level of accuracy and unbiasedness.

3.7. Analysis and interpretation of data

After the collection of data, what we have is a huge chunk of observations and numerical values. The data at the beginning are in raw form. For the purpose of applying further statistical techniques, one has to put the raw data in a useful form by classification, tabulation and categorization of data. If one has to feed the data in a computer, the data should bear the same form as required by the software used. This kind of processing of data involves one or more of the following activities:

(i) coding (ii) labeling (iii)editing (iv) tabulation (v) classification.

3.8. Hypothesis testing

After analyzing and processing of data, it is time now to test the hypothesis that were formed in step 3 of the research process. A hypothesis is skeptically formulated regarding the relationship between phenomena and variables involved in a study. Then by empirical investigation the hypothesis is tested for possible acceptance or rejection. In other words, the researcher decides on the basis of the observed facts that he has collected, whether or not an assumption is valid. A hypothesis is tested by making use of a predefined decision rules established in statistical methods. Some of the popular statistical tests are, Z- test, Chi- square test, t- test and F- test.

In a situation where no hypothesis is formulated in a study, the observations are made on the data directly and conclusions are drawn to formulate new generalizations and assumptions for future purposes.

3.9. Interpretations of results

After the data collection and testing of hypothesis one has to reach to the conclusions of the research study. These conclusions are the most vital outcomes of the study and have to be dealt with very carefully. On the basis of findings of the research work done we draw inferences about the phenomenon under study. This is a useful activity as without any outcome a research study is fruitless. The results obtained from the analysis of data are to be interpreted skillfully. A wrong interpretation my lead to wrong decisions. Interpretation may also lead to generalizations of the phenomena understudy. It may also help in developing new theories and can suggest new research problems to be explored in future.

3.10. Report writing

Last but not the least is the step of reporting the facts and findings of the research study. A report is a summary of the whole research process. The layout of a report must be attractive. The words used in the text must be easily comprehensive to a reader. Even a non- technical person understands a good report. In the beginning of the report one should give the title, time period of work, acknowledgement and preface. In the main text an introduction to the problem, summary of findings, results and inferences, and then the recommendations of the researcher are given. The report should conclude with appendices, bibliography and a subject or / and author indexes.

4. Research Problem

4.1. Meaning of Research Problem

Research is a scientific, systematic and purposeful search, for new knowledge or for re-interpretation of existing knowledge. It is a journey, which starts with a problem and ends with a solution. Identifying a research problem is the first and foremost step in a research process. The statement of research problem is the axis around which the whole research revolves, because it explains in brief the aims and objectives of the research.

A research problem is a specific statement in the general area of investigation. It is a precise identification of a problem situation in a certain context involving what, why, who, where and when of the problem area.

Who– means the person or business organization that is facing a problem.

Why– means that there is a purpose, goal aim or objective to solve this problem.

How– means the options of actions one can take to solve the problem.

When–means the time frame in which the problem is to be solved.

Where–means the environment in which the problem exists.

What– means the optimum action that is to be taken in solving the problem to attain the best results.

4.2. Selecting a Research Problem

The old saying goes, "Necessity is the mother of invention". A research study is another form of invention. Thus, if there is some necessity or a difficulty a business organization is facing, it forms a research problem that is to be investigated in order to fulfill that necessity or remove that difficulty. However, simple it may look, but selecting a research problem is a big problem in itself. Specially, when there is no apparent problem in an organization or when a students wants to select a research problem for the purpose of a dissertation or thesis for a degree or a diploma it becomes utmost important that the research problem must be very carefully chosen.

Some of the guidelines that researcher must follow in selecting a research problem are listed below.

4.1. Sources of problems: First of all one should look at the sources from which one can select a research problem. Those may be readily available problem that has been identified by a person or an organization. If that is not the case, one can make use of the experience of experts of that field. The survey of related literature may also help in selection of a research problem.

4.2. Potential to be a research problem: One must ensure that the problem one has undertaken has potential to be called as a research problem. One should avoid trivial or meaningless problem. A research problem must provide solution to an existing problem or contribute to the body of knowledge.

4.3. Select non-controversial issues: Unless the problem is specifically related to issues such as religion, dogmas, beliefs, sexual preferences etc; one should try to avoid taking up controversial subjects.

4.4. Researcher's interest and competency: In selection of research problem a researcher should choose a topic of his own field of study in which he has independent mastery in both the subject and method. The problem should sustain his interest, stimulate his imagination and should be within the range of his competencies.

4.5. Resources available: To conduct a research study various resources are needed. A researcher must ensure the following points about the inputs of the problem:

(a) Time: The study must be completed in the allotted time frame.

(b) Funds: The amount of funds available from the sponsoring agencies must be known in advance and the study must be completed within given budget.

(c) Size of research: The size of the investigation must be manageable and should not be too large to handle. It should also not be too small to appear as a trivial problem.

(d) Co-operation of others: A researcher must ensure the necessary cooperation of colleagues and operational help of administrative authorities is available to the problem he is selecting.

(e) Literature and Material: Related literature is readily available to compare and support the research findings.

(f) Obtainable data: The selection of a problem should be such that the information or data needed for it is either readily available or obtainable. In short a researcher must select a research problem of his area of interest and should ensure the availability of all the comments needed to conduct the study. The consultation and guidance of experts, who have experience in that area, is a must. A researcher must seek help of such supervisors and guides in selecting a research problem

5. statistical methods

Statistics is defined as collection , compilation ,analysis and interpretation of numerical data .

5.1. Why Statistics?

• To develop an appreciation for variability and how it effect product ,process and system.

• It is estimating the present ; predicting the future

• Study methods that can be used to solve problems ,build knowledge.

• Statistics make data into information

• Develop an understanding of some basic ideas of statistical reliability , stochastic process (probability concepts) .

• Statistics is very important in every aspects of society (Govt., People or Business)

5.2. Basic terms

(i) **Measurement** :assignment of numbers to something

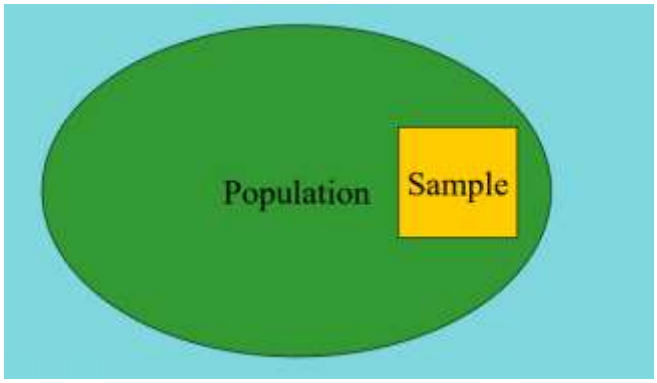
(ii) **Data**: collection of measurements

(iii) **Population** : The set of data (numerical or otherwise) corresponding to the entire collection of units about which information Is sought

(iv) **Sample**: A Subset of a population.
Representative Sample

– Has the characteristics of the population

Census - A Sample that Contains all Items in the Population



(v) Why Sampling?

In most studies, it is difficult to obtain information from the entire population because of various reasons. We rely on samples to make estimates or inferences related to the population.

(vi) Statistical Inference

Drawing Conclusions (Inferences) about a Population Based on an examination of a Sample taken from the population

(vii) Variables :

property with respect to which data from a sample differ in some measurable way

eg: length of a fish is variable, weight, blood glucose level, number of seeds in fruit, oxygen content in water

Types of variables: There are 2 types

a) **Qualitative variable**: it is one that generally cannot be expressed in numbers, it is an attribute and is descriptive in nature

Eg: sex, skin colour and behaviour of animals, colour and smell of flowers

b) **Quantitative variable**: it is one whose differing status can be expressed in numbers

Eg: length of fish, weight of frog, number of children in a family

A quantitative variable can be discrete variable or continuous variable

Discrete variables: these variables can assume certain fixed values without no intermediate values possible

Eg: number of children in a family is counted as 0.1.2.3…so on

But cannot be expressed as 1.4 or 5.6

Continuous variables: The variables can assume at least theoretically, infinite values between any fixed points. These variables are obtained by measurements using some scale such as weighing balance or sophisticated instruments

Eg: length of fish can be measured as 1.5, 1.6. 1.7 cm …soon

c) Ranked variables:

These variables cannot be measured but can be ordered or ranked by their magnitude

Eg: Activity of pancreas gland as assessed in microscopic observations

Ranking of students in class

d) Derived variables

Those which are calculated based on two or more independently measured variables

They show relationship between variables

Eg: rates, ratios, percentages, indices etc

e) Variate

It is single observation of a variable

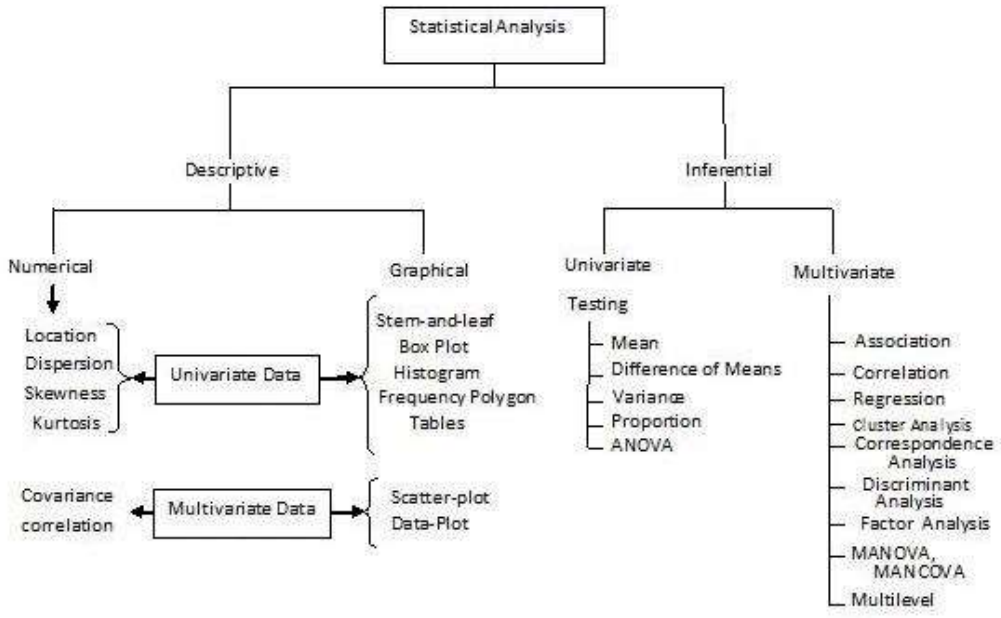
Eg: if the weight of 6 catla are measured in grams as 23, 24, 23, 34, 33 and 32 each value is variate. 23g is a variate of Variable weight.

5.3. Types of statistics

• Descriptive statistics : use to organize and describe a sample /population

• Inferential statistics : use to extrapolate (estimate) from a sample to larger

Population



5.3.1. DESCRIPTIVE STATISTICS

There are several measures to describe a data

(i) Measure of central tendency: When given a set of raw data one of the most useful ways of summarising that data is to find an average of that set of data. An average is a measure of the centre of the data set.

There are three common ways of describing the centre of a set of numbers. They are the mean, the median and the mode and are calculated as follows.

The mean − add up all the numbers and divide by how many numbers there are.

For n values in a set of data namely as x_1, x_2, x_3, ... x_n, the mean of data is given as:

$$\bar{x} = \frac{x_1 + x_2 + x_3 \dots \dots \dots . . x_n}{n}$$

It can also be denoted as:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The median − is the middle number. It is found by putting the numbers in order and taking the actual middle number if there is one,
or the average of the two middle numbers if not.

The mode − is the most commonly occurring number.

Problem 1: Let's illustrate these by calculating the mean, median and mode for the following data.
Weight of luggage presented by airline passengers at the check-in (measured to the nearest kg).

18 23 20 21 24 23 20 20 15 19 24

Mean = 18 + 23 + 20 + 21 + 24 + 23 + 20 + 20 + 15 + 19 + 24/ 11 = 20.64.

Median = 20.

15 18 19 20 20 20 21 23 23 24 24

↑ middle value

Mode = 20.

The number 20 occurs here 3 times.

Here the mean, median and mode are all appropriate measures of central tendency

Frequency distribution

For calculating the mean when the frequency of the observations is given, such that x1, x2, x3,… xn is the recorded observations, and f1, f2, f3 … fn is the respective frequencies of the observations then;

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 \dots \dots \dots f_nx_n}{f_1 + f_2 + f_3 \dots \dots \dots f_n}$$

This can be expressed briefly as:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

The above method of calculating the arithmetic mean is used when the data is ungrouped in nature. For calculating the mean of grouped data, we calculate the class mark. For this, the midpoints of the class intervals are calculated as:

$$Class\ Mark = \frac{Upper\ Limit + Lower\ Limit}{2}$$

After calculating the class mark, the mean is calculated as discussed earlier. This method of calculating the mean is known as the direct method.

Problem 2: In a class of 30 students, marks obtained by students in mathematics out of 50 is tabulated below. Calculate the mean of the data.

Marks Obtained	Number of student	Class Mark	$f_i x_i$
10-20	5	15	75
20-30	5	25	125
30-40	8	35	280
40-50	12	45	540
Total	$\Sigma f_i = 30$		$\sum f_i x_i = 1020$

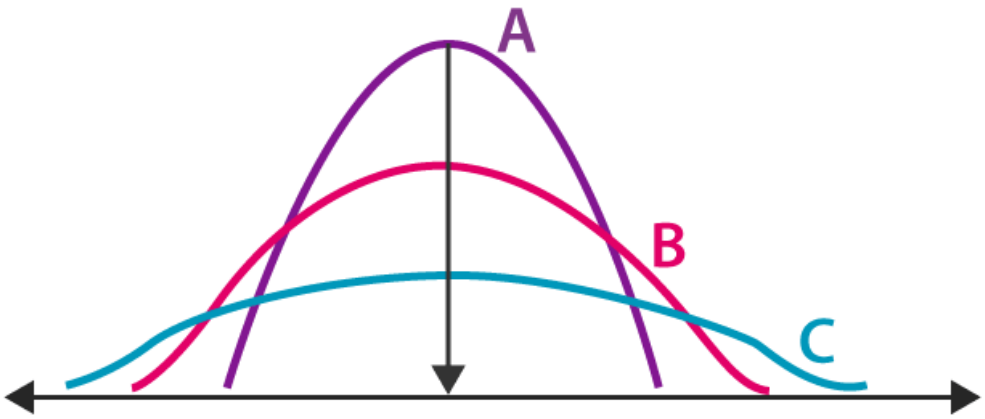
Solution:

The mean of the data given above is,

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{1020}{30} = 34$$

(ii) Measure of dispersion

Dispersion is the state of getting dispersed or spread. Statistical dispersion means the extent to which a numerical data is likely to vary about an average value. In other words, dispersion helps to understand the distribution of the data.



Absolute Measure of Dispersion

An absolute measure of dispersion contains the same unit as the original data set. Absolute dispersion method expresses the variations in terms of the average of deviations of observations like standard or means deviations. It includes range, standard deviation, quartile deviation, etc.

The types of absolute measures of dispersion are:

Range: It is simply the difference between the maximum value and the minimum value given in a data set. Example: 1, 3,5, 6, 7 => Range = 7 -1= 6

Standard Deviation: it is defined as the square root of the arithmetic mean of the squared deviations of the various items from arithmetic mean. In short it is called root mean square deviation. The mean of squared deviation is called variance.

Variance: Deduct the mean from each data in the set then squaring each of them and adding each square and finally dividing them by the total no of values in the data set is the variance.

Variance $(\sigma 2)=\sum(X-\mu)2/N$

Therefore, The square root of the variance is known as the standard deviation i.e. S.D. = $\sqrt{\sigma}$.

Or

Standard Deviation Formula	
Population	**Sample**
$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X-\bar{x})^2}{n-1}}$
X – The Value in the data distribution μ – The population Mean N – Total Number of Observations	X – The Value in the data distribution $\bar{x}$ – The Sample Mean n - Total Number of Observations

Problem 3: Compute the standard deviation of the exam scores of a national standardised test to see if many people scored close to the mean or not. Use the following dataset.

Test taker	Score
1	20
2	40
3	60
4	60
5	75
6	80
7	70
8	65
9	70
10	90

Solution

Test taker	Score (X)	$(X-\bar{X})$	$(X-\bar{X})^2$
1	20	-43	1849
2	40	-23	529
3	60	-3	9
4	60	-3	9
5	75	12	144
6	80	17	289
7	70	7	49
8	65	2	4
9	70	7	49
10	90	27	729
	$\sum X=630$ $\bar{X}=63$		$\sum(X-\bar{X})^2$ =3660

Next, we simply put it into the **formula.**

$$sd = \sqrt{\frac{3660}{10-1}} = 20.2$$

Interpretation of standard deviation

Problem 4: Suppose you are going through a research paper and you score across a statement that the mean IQ of children is 100 with a SD of 13 (100±13). How do you interpret these values?

About 68.26% of the children have an IQ of 100 ± 1(13) i.e., mean± 1SD.

100-13=87

100+13=113

The IQ of 68.26% children lies between 87-110

About 95.45% of the children have an IQ of 100 $\pm$ 2(13) i.e., mean$\pm$2.

100-26=74

100+26=126

The IQ of 95.45% children lies between 74-126

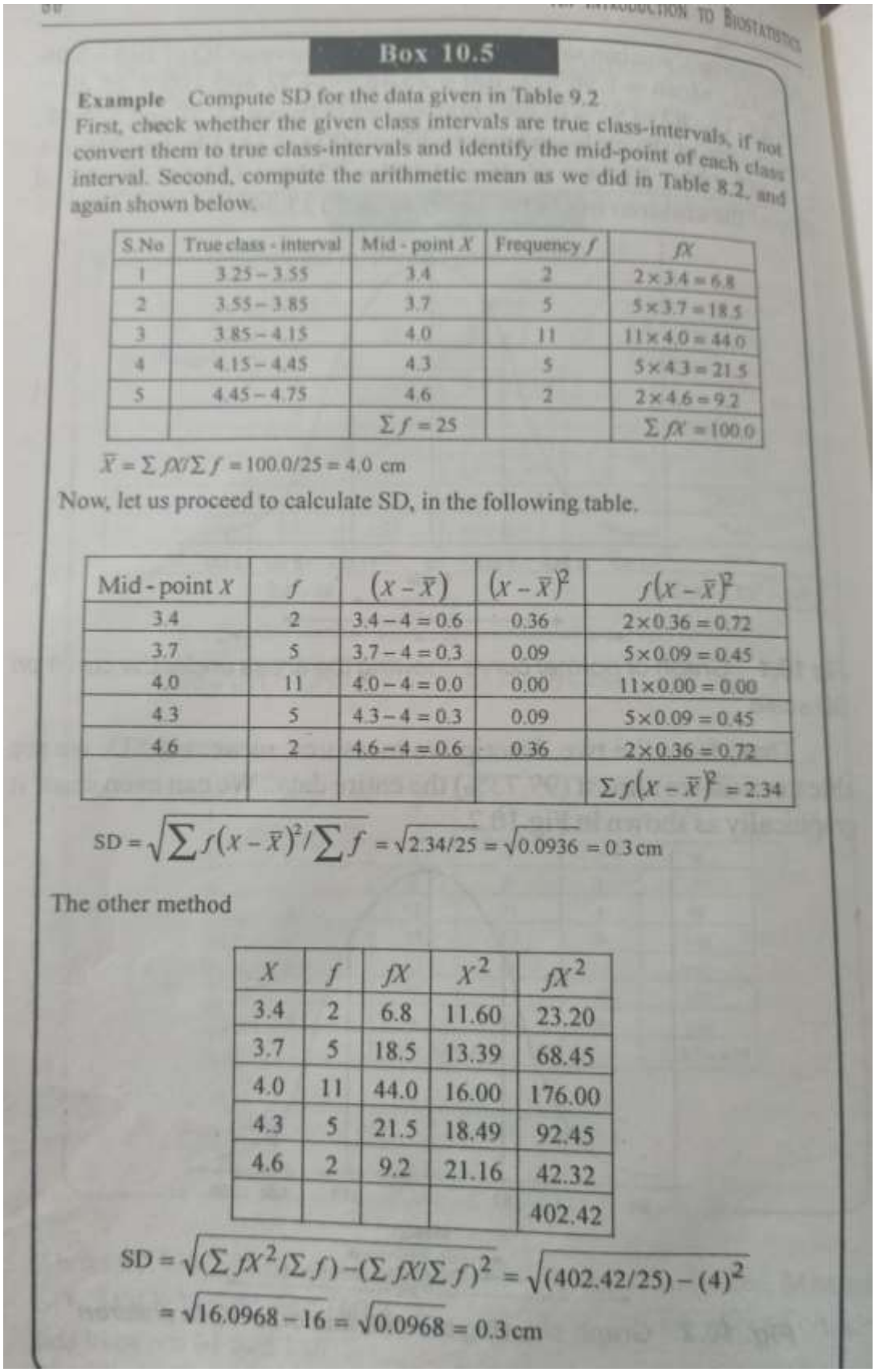
About 99.73% of the children have an IQ of 100 $\pm$ 3(13) i.e., mean$\pm$3SD.

100-39=61

100+39=139

The IQ of 99.73% children lies between 61-139

Problem 4:



Box 10.5

Example Compute SD for the data given in Table 9.2

First, check whether the given class intervals are true class-intervals, if not convert them to true class-intervals and identify the mid-point of each class interval. Second, compute the arithmetic mean as we did in Table 8.2, and again shown below.

S.No	True class - interval	Mid - point X	Frequency f	fX
1	3.25 – 3.55	3.4	2	2×3.4 = 6.8
2	3.55 – 3.85	3.7	5	5×3.7 = 18.5
3	3.85 – 4.15	4.0	11	11×4.0 = 44.0
4	4.15 – 4.45	4.3	5	5×4.3 = 21.5
5	4.45 – 4.75	4.6	2	2×4.6 = 9.2
		$\Sigma f = 25$		$\Sigma fX = 100.0$

$\bar{X} = \Sigma fX/\Sigma f = 100.0/25 = 4.0$ cm

Now, let us proceed to calculate SD, in the following table.

Mid - point X	f	$(X - \bar{X})$	$(X - \bar{X})^2$	$f(X - \bar{X})^2$
3.4	2	3.4 – 4 = 0.6	0.36	2×0.36 = 0.72
3.7	5	3.7 – 4 = 0.3	0.09	5×0.09 = 0.45
4.0	11	4.0 – 4 = 0.0	0.00	11×0.00 = 0.00
4.3	5	4.3 – 4 = 0.3	0.09	5×0.09 = 0.45
4.6	2	4.6 – 4 = 0.6	0.36	2×0.36 = 0.72
				$\Sigma f(X - \bar{X})^2 = 2.34$

$$SD = \sqrt{\sum f(X - \bar{X})^2 / \sum f} = \sqrt{2.34/25} = \sqrt{0.0936} = 0.3 \text{ cm}$$

The other method

X	f	fX	X^2	fX^2
3.4	2	6.8	11.60	23.20
3.7	5	18.5	13.39	68.45
4.0	11	44.0	16.00	176.00
4.3	5	21.5	18.49	92.45
4.6	2	9.2	21.16	42.32
				402.42

$$SD = \sqrt{(\Sigma fX^2/\Sigma f) - (\Sigma fX/\Sigma f)^2} = \sqrt{(402.42/25) - (4)^2}$$
$$= \sqrt{16.0968 - 16} = \sqrt{0.0968} = 0.3 \text{ cm}$$

Quartiles and Quartile Deviation: The quartiles are values that divide a list of numbers into quarters. The quartile deviation is half of the distance between the third and the first quartile.

Problem 5:

We can find the interquartile range or IQR in four simple steps:

1. Order the data from least to greatest
2. Find the median
3. Calculate the median of both the lower and upper half of the data
4. The IQR is the difference between the upper and lower medians

Step 1: Order the data

In order to calculate the IQR, we need to begin by ordering the values of the data set from the least to the greatest. Likewise, in order to calculate the median, we need to arrange the numbers in ascending order (i.e. from the least to the greatest).

Let's sort an example data set with an odd number of values into ascending order.

$$\text{Odd data set} : 9, 3, 2, 5, 6, 11, 4, 3, 2$$

$$\text{Odd data set (ascending)} : 2, 2, 3, 3, 4, 5, 6, 9, 11$$

Now, let's perform this task with another example data set that is comprised of an even number of values.

$$\text{Even data set} : 11, 2, 4, 3, 8, 1, 2, 7, 4, 9$$

Rearrange into ascending order.

$$\text{Even data set (ascending)} : 1, 2, 2, 3, 4, 4, 7, 8, 9, 11$$

Step 2: Calculate the median

Next, we need to calculate the median. The median is the "center" of the data. If the data set has an odd number of data points, then the mean is the centermost number. On the other hand, if the data set has an even number of values, then we will need to take the arithmetic average of the two centermost values. We will calculate this average by adding the two numbers together and then dividing that number by two.

First, we will find the median of a set with an odd number of values. Cross out values until you find the centermost point

$$\text{Odd data set}: \cancel{2}, \cancel{2}, \cancel{3}, \cancel{3}, 4, \cancel{5}, \cancel{6}, \cancel{9}, \cancel{11}$$

The median of the odd valued data set is four.

Now, let's find the mean of the data set with an even number of values. Cross out values until you find the two centermost points and then calculate the average the two values.

$$\text{Even data set}: \cancel{1}, \cancel{2}, \cancel{2}, \cancel{3}, 4, 4, \cancel{7}, \cancel{8}, \cancel{9}, \cancel{11}$$

Find the average of the two centermost values.

$$\text{Average} = \frac{4+4}{2}$$

$$\text{Average} = \frac{8}{2}$$

$$\text{Average} = 4$$

The median of the even valued set is four.

Step 3: Upper and lower medians

Once we have found the median of the entire set, we can find the medians of the upper and lower portions of the data. If the data set has an odd number of values, we will omit the median or centermost value of the set. Afterwards, we will find the individual medians for the upper and lower portions of the data.

$$\text{Odd data set}: 2, 2, 3, 3, 4, 5, 6, 9, 11$$

Omit the centermost value.

$$\text{Odd data set}: 2, 2, 3, 3, \mid 5, 6, 9, 11$$

Find the median of the lower portion.

$$\text{Odd data set}: \cancel{2}, 2, 3, \cancel{3}, \mid 5, 6, 9, 11$$

Calculate the average of the two values.

$$\text{Average} = \frac{2+3}{2}$$

$$\text{Average} = \frac{5}{2}$$

$$\text{Average} = 2.5$$

The median of the lower portion is 2.5

Find the median of the upper portion.

$$\text{Odd data set} : 2, 2, 3, 3, \not{5}, 6, 9, \not{11}$$

Calculate the average of the two values.

$$\text{Average} = \frac{6+9}{2}$$

$$\text{Average} = \frac{15}{2}$$

$$\text{Average} = 7.5$$

The median of the upper potion is 7.5

If the data set has an even number of values, we will use the two values used to calculate the original median to divide the data set. These values are not omitted and become the largest value of the lower data set and the lowest values of the upper data set, respectively. Afterwards, we will calculate the medians of both the upper and lower portions.

$$\text{Even data set} : 1, 2, 2, 3, 4 \mid 4, 7, 8, 9, 11$$

Find the median of the lower portion.

$$\text{Even data set} : \not{1}, \not{2}, 2, \not{3}, \not{4} \mid 4, 7, 8, 9, 11$$

The median of the lower portion is two.

Find the median of the upper portion.

$$\text{Even data set} : 1, 2, 2, 3, 4 \mid \not{4}, \not{7}, 8, \not{9}, \not{11}$$

The median of the upper portion is eight.

Step 4: Calculate the difference

Last, we need to calculate the difference of the upper and lower medians by subtracting the lower median from the upper median. This value equals the IQR.

Let's find the IQR of the odd data set.

$$\text{IQR of the odd data set} = 7.5 - 2.5$$

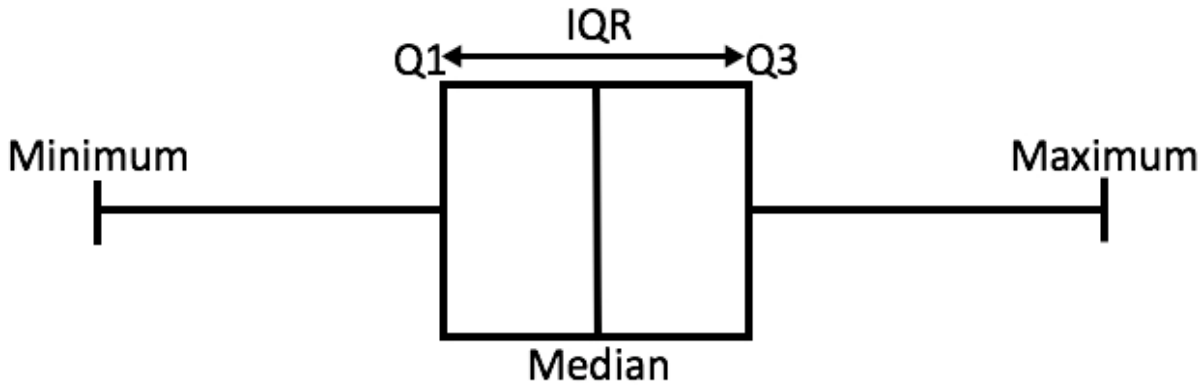
$$\text{IQR} = 5$$

Finally, we will find the IQR of the even data set.

$$\text{IQR of the even data set} = 8 - 2$$

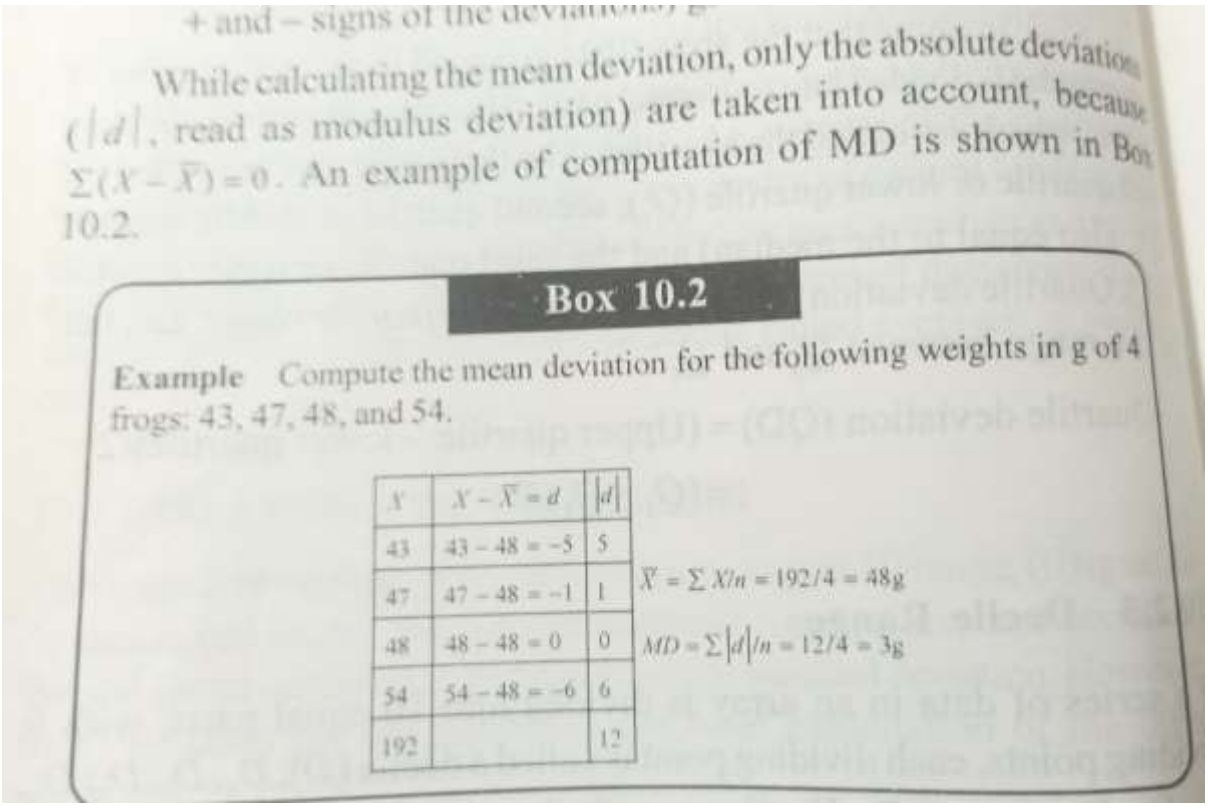
$$\text{IQR} = 6$$

In order to better illustrate these values, their positions in a box plot have been labeled in the provided image.



Mean and Mean Deviation: The average of numbers is known as the mean and the arithmetic mean of the absolute deviations of the observations from a measure of central tendency is known as the mean deviation (also called mean absolute deviation).

Problem 6:



+ and – signs of the deviations) are

While calculating the mean deviation, only the absolute deviations ($|d|$, read as modulus deviation) are taken into account, because $\Sigma(X - \bar{X}) = 0$. An example of computation of MD is shown in Box 10.2.

Box 10.2

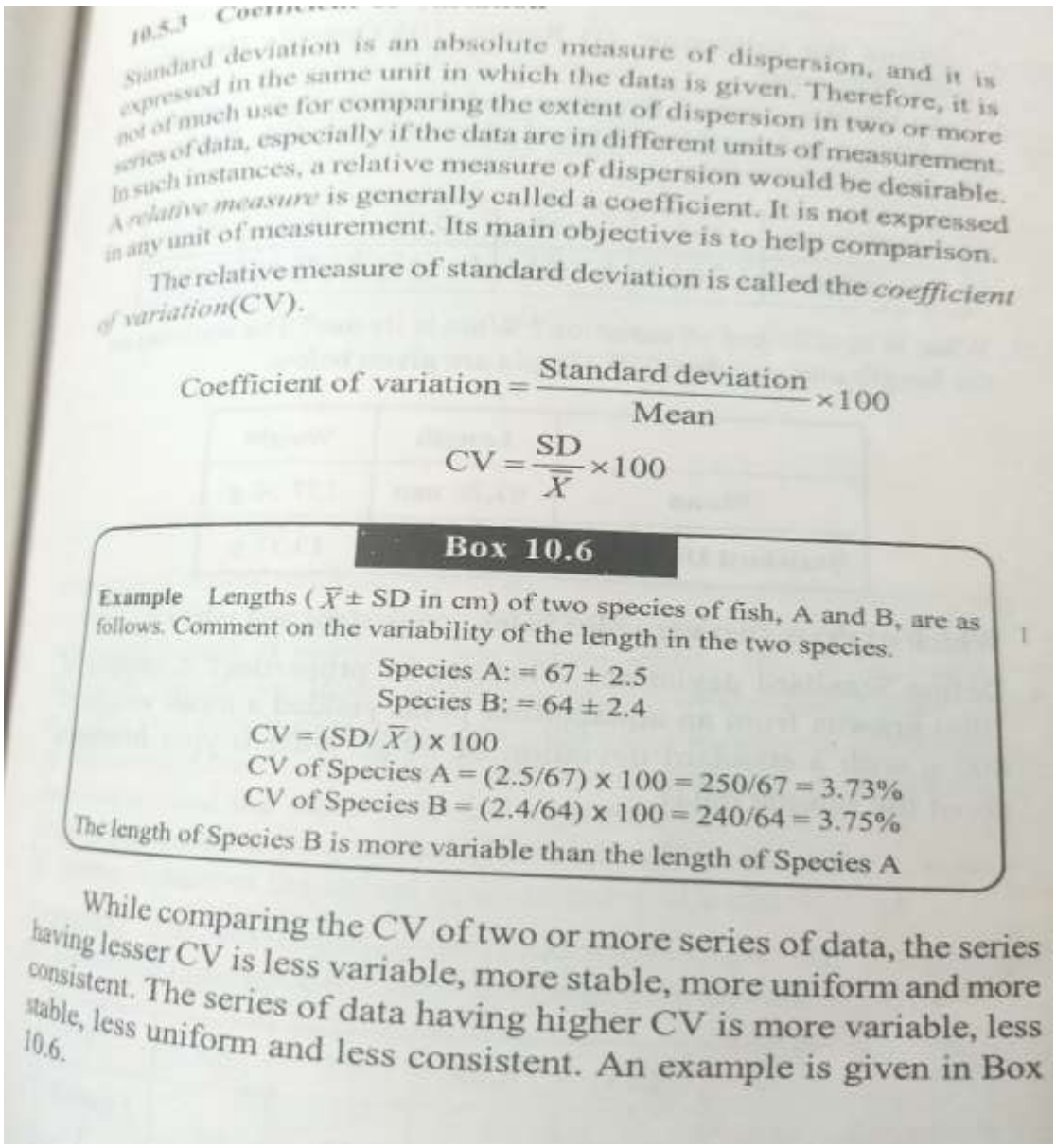
Example Compute the mean deviation for the following weights in g of 4 frogs: 43, 47, 48, and 54.

X	$X - \bar{X} = d$	$\lvert d \rvert$
43	43 – 48 = –5	5
47	47 – 48 = –1	1
48	48 – 48 = 0	0
54	54 – 48 = –6	6
192		12

$\bar{X} = \Sigma X/n = 192/4 = 48$g

$MD = \Sigma|d|/n = 12/4 = 3$g

Coefficient of variation



10.5.3

Standard deviation is an absolute measure of dispersion, and it is expressed in the same unit in which the data is given. Therefore, it is not of much use for comparing the extent of dispersion in two or more series of data, especially if the data are in different units of measurement. In such instances, a relative measure of dispersion would be desirable. A *relative measure* is generally called a coefficient. It is not expressed in any unit of measurement. Its main objective is to help comparison.

The relative measure of standard deviation is called the *coefficient of variation*(CV).

$$\text{Coefficient of variation} = \frac{\text{Standard deviation}}{\text{Mean}} \times 100$$

$$CV = \frac{SD}{\bar{X}} \times 100$$

Box 10.6

Example Lengths ($\bar{X} \pm$ SD in cm) of two species of fish, A and B, are as follows. Comment on the variability of the length in the two species.

Species A: = 67 ± 2.5
Species B: = 64 ± 2.4

$CV = (SD/\bar{X}) \times 100$
CV of Species A = (2.5/67) x 100 = 250/67 = 3.73%
CV of Species B = (2.4/64) x 100 = 240/64 = 3.75%
The length of Species B is more variable than the length of Species A

While comparing the CV of two or more series of data, the series having lesser CV is less variable, more stable, more uniform and more consistent. The series of data having higher CV is more variable, less stable, less uniform and less consistent. An example is given in Box 10.6.

5.3.2. Correlation and Regression

Correlation means association - more precisely it is a measure of the extent to which two variables are related. There are three possible results of a correlational study: a positive correlation, a negative correlation, and no correlation.

A positive correlation is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases, or one

variable decreases while the other decreases. An example of positive correlation would be height and weight. Taller people tend to be heavier.

A negative correlation is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. An example of negative correlation would be height above sea level and temperature. As you climb the mountain (increase in height) it gets colder (decrease in temperature).

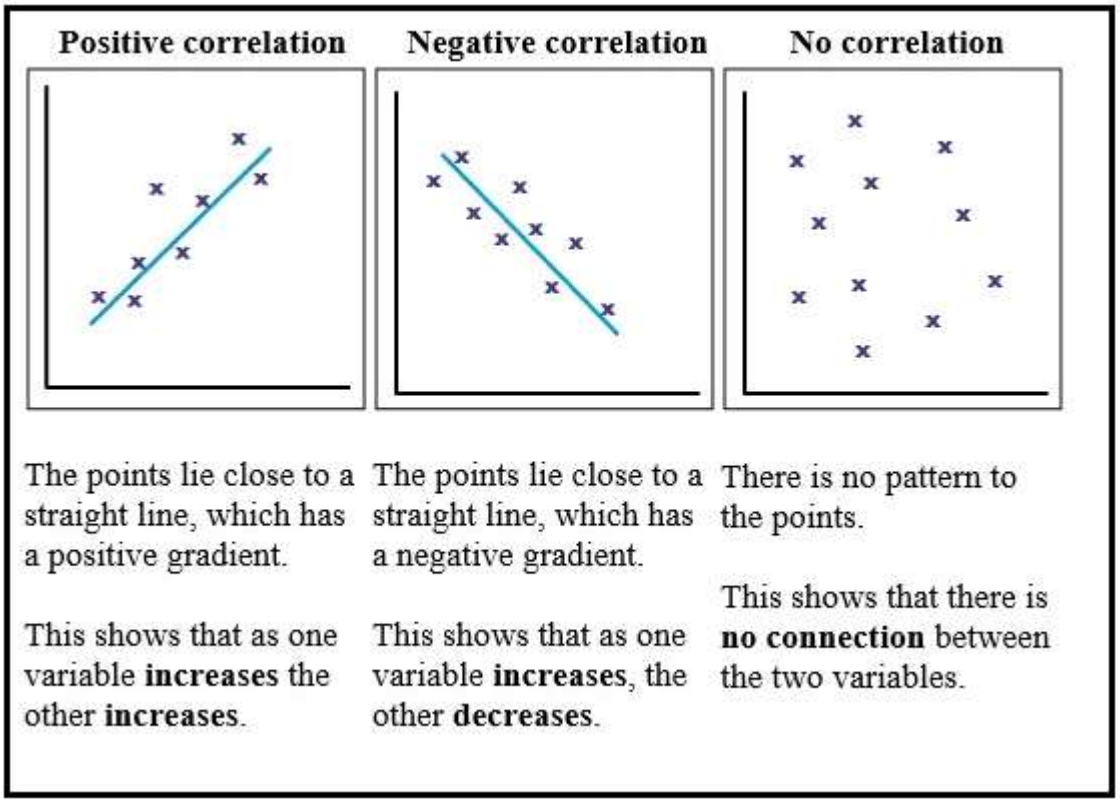
A zero correlation exists when there is no relationship between two variables. For example there is no relationship between the amount of tea drunk and level of intelligence.

Scattergrams

A correlation can be expressed visually. This is done by drawing a scattergram (also known as a scatterplot, scatter graph, scatter chart, or scatter diagram).

A scattergram is a graphical display that shows the relationships or associations between two numerical variables (or co-variables), which are represented as points (or dots) for each pair of score.

A scattergraph indicates the strength and direction of the correlation between the co-variables.



When you draw a scattergram it doesn't matter which variable goes on the x-axis and which goes on the y-axis.

Remember, in correlations we are always dealing with paired scores, so the values of the 2 variables taken together will be used to make the diagram.

Decide which variable goes on each axis and then simply put a cross at the point where the 2 values coincide.

Some uses of Correlations

- **Prediction**

If there is a relationship between two variables, we can make predictions about one from another.

- **Validity**

Concurrent validity (correlation between a new measure and an established measure).

- **Reliability**

Test-retest reliability (are measures consistent).

Inter-rater reliability (are observers consistent).

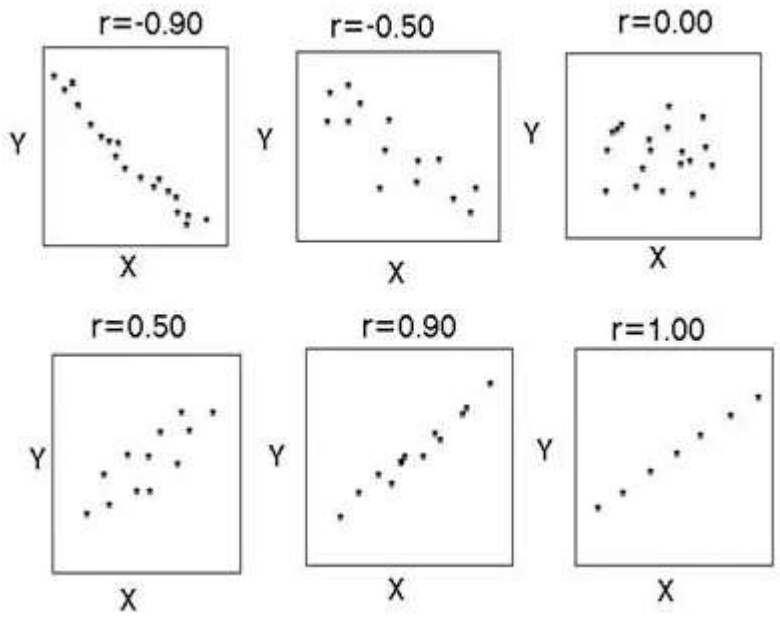
- **Theory verification**

Predictive validity.

Correlation Coefficients:

Determining Correlation Strength

Instead of drawing a scattergram a correlation can be expressed numerically as a coefficient, ranging from -1 to +1. When working with continuous variables, the correlation coefficient to use is Pearson's r.



The correlation coefficient (r) indicates the extent to which the pairs of numbers for these two variables lie on a straight line. Values over zero indicate a positive correlation, while values under zero indicate a negative correlation.

A correlation of –1 indicates a perfect negative correlation, meaning that as one variable goes up, the other goes down. A correlation of +1 indicates a perfect positive correlation, meaning that as one variable goes up, the other goes up.

There is no rule for determining what size of correlation is considered strong, moderate or weak. The interpretation of the coefficient depends on the topic of study.

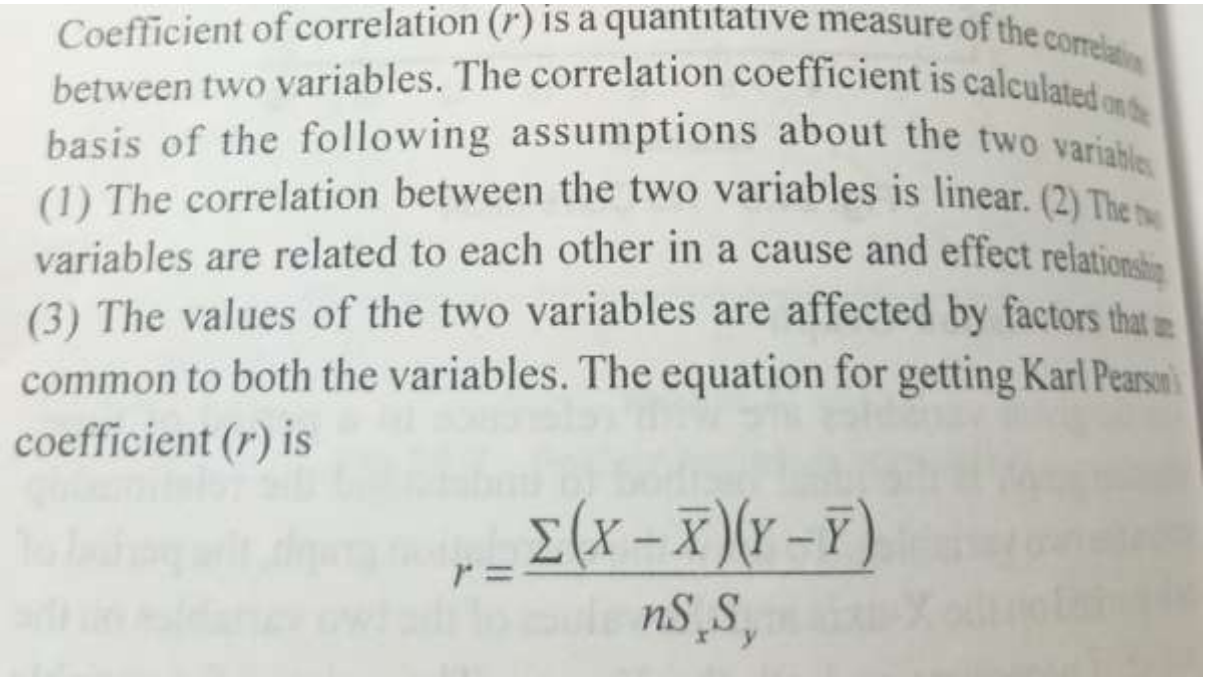
When studying things that are difficult to measure, we should expect the correlation coefficients to be lower (e.g. above 0.4 to be relatively strong). When we are studying things that are more easier to measure, such as socioeconomic status, we expect higher correlations (e.g. above 0.75 to be relatively strong).)

In these kinds of studies, we rarely see correlations above 0.6. For this kind of data, we generally consider correlations above 0.4 to be relatively strong; correlations between 0.2 and 0.4 are moderate, and those below 0.2 are considered weak.

When we are studying things that are more easily countable, we expect higher correlations. For example, with demographic data, we we generally consider correlations above 0.75 to be relatively

strong; correlations between 0.45 and 0.75 are moderate, and those below 0.45 are considered weak.

Karl Pearson's coefficient of correlation



Coefficient of correlation (r) is a quantitative measure of the correlation between two variables. The correlation coefficient is calculated on the basis of the following assumptions about the two variables: (1) The correlation between the two variables is linear. (2) The two variables are related to each other in a cause and effect relationship. (3) The values of the two variables are affected by factors that are common to both the variables. The equation for getting Karl Pearson's coefficient (r) is

$$r = \frac{\Sigma\left(X - \bar{X}\right)\left(Y - \bar{Y}\right)}{nS_x S_y}$$

where, r = Coefficient of correlation,

X = variable X,

$\bar{X}$ = Mean of variable X,

Y = variable Y,

$\bar{Y}$ = Mean of variable Y,

n = number of pairs of variables,

S_x = SD of variable X, and

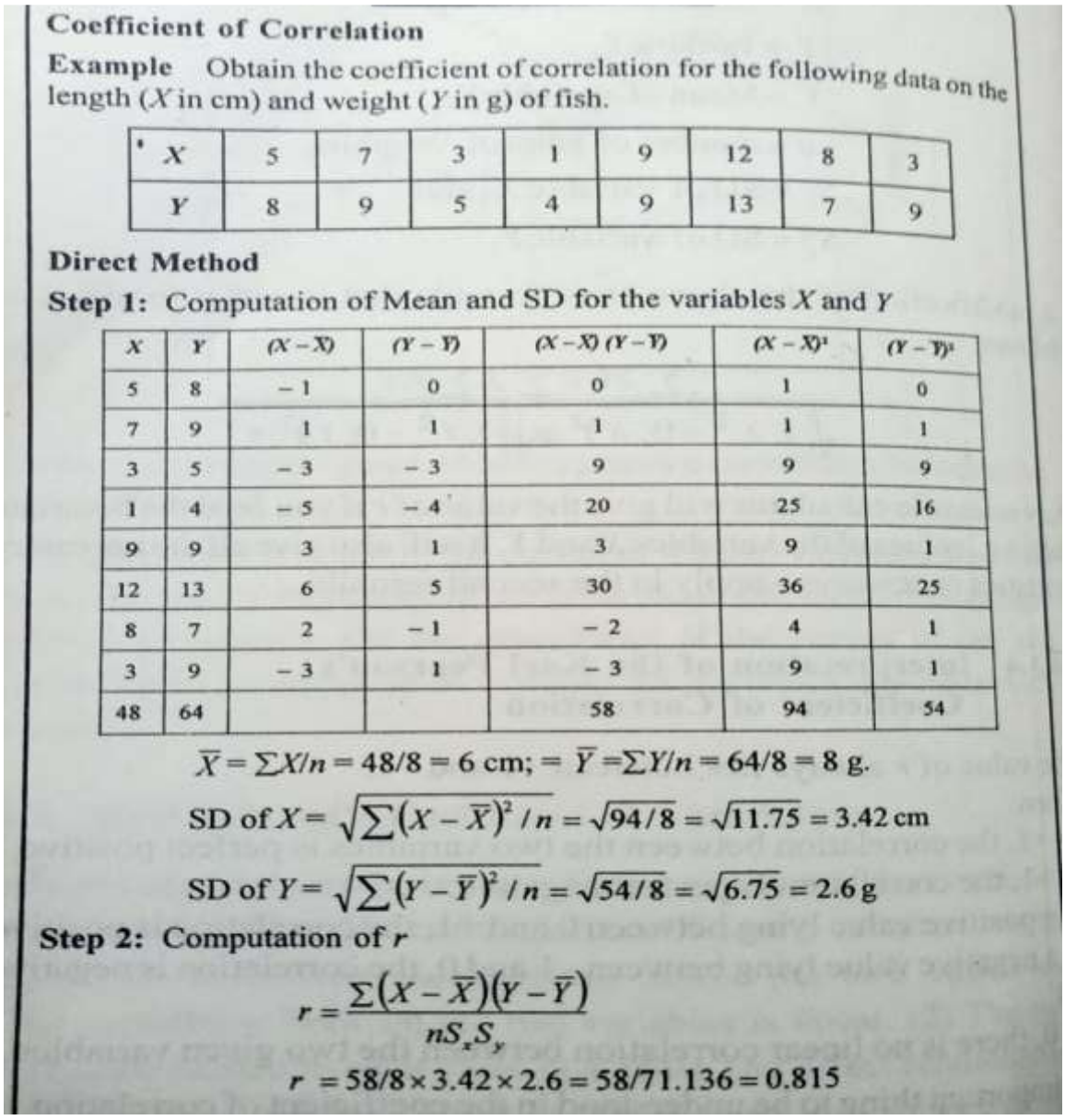
S_y = SD of variable Y.

A modification of the above basic formula that is easier to use is as follows:

$$r = \frac{\sum XY - \sum X \sum Y/n}{\sqrt{\left[\sum X^2 - (\sum X)^2/n\right]\left[\sum Y^2 - (\sum Y)^2/n\right]}}$$

Any scientific calculator will give the value of r if you feed the bivariate data, i.e., values of the variables X and Y. It will also give all the necessary statistics necessary to apply in the second formula.

Problem 8



Coefficient of Correlation

Example Obtain the coefficient of correlation for the following data on the length (X in cm) and weight (Y in g) of fish.

X	5	7	3	1	9	12	8	3
Y	8	9	5	4	9	13	7	9

Direct Method

Step 1: Computation of Mean and SD for the variables X and Y

X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$
5	8	− 1	0	0	1	0
7	9	1	1	1	1	1
3	5	− 3	− 3	9	9	9
1	4	− 5	− 4	20	25	16
9	9	3	1	3	9	1
12	13	6	5	30	36	25
8	7	2	− 1	− 2	4	1
3	9	− 3	1	− 3	9	1
48	64			58	94	54

$\bar{X} = \Sigma X/n = 48/8 = 6$ cm; $= \bar{Y} = \Sigma Y/n = 64/8 = 8$ g.

SD of $X = \sqrt{\sum (X-\bar{X})^2 / n} = \sqrt{94/8} = \sqrt{11.75} = 3.42$ cm

SD of $Y = \sqrt{\sum (Y-\bar{Y})^2 / n} = \sqrt{54/8} = \sqrt{6.75} = 2.6$ g

Step 2: Computation of r

$$r = \frac{\Sigma (X-\bar{X})(Y-\bar{Y})}{nS_x S_y}$$

$$r = 58/8 \times 3.42 \times 2.6 = 58/71.136 = 0.815$$

Regression analysis

1. Predicts the value of one variable, given the value of another variable

When those variables are related

2. Mathematical measure of average relationship between 2 or more variables
3. Estimates of values of dependent variable from the values of independent variable
4. Measure of error using regression line

5. Degree of association or correlation that exists between 2 variables

Regression line

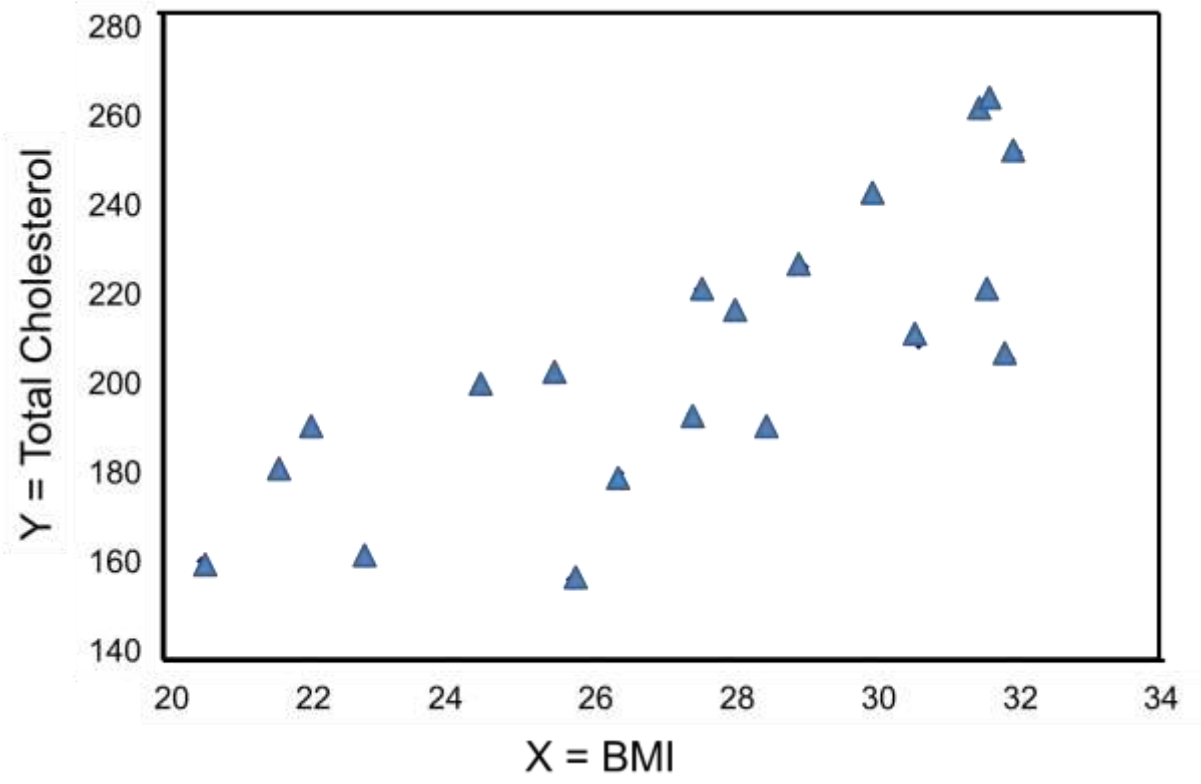
- Best estimate of one variable from the value of any other given variable
- Average relation between two variables in mathematical form

Simple Linear Regression

Simple linear regression is a technique that is appropriate to understand the association between one independent (or predictor) variable and one continuous dependent (or outcome) variable. For example, suppose we want to assess the association between total cholesterol (in milligrams per deciliter, mg/dL) and body mass index (BMI, measured as the ratio of weight in kilograms to height in meters2) where total cholesterol is the dependent variable, and BMI is the independent variable. In regression analysis, the dependent variable is denoted Y and the independent variable is denoted X. So, in this case, Y=total cholesterol and X=BMI.

When there is a single continuous dependent variable and a single independent variable, the analysis is called a simple linear regression analysis . This analysis assumes that there is a linear association between the two variables. (If a different relationship is hypothesized, such as a curvilinear or exponential relationship, alternative regression analyses are performed.)

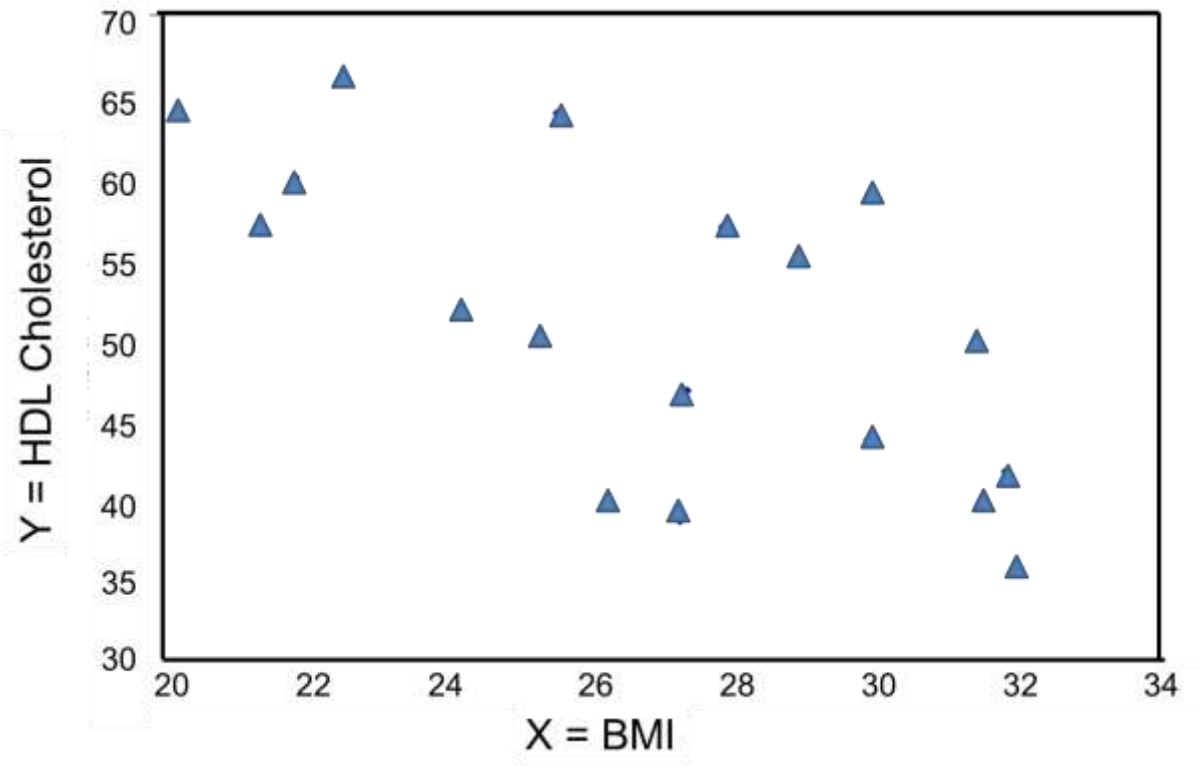
The figure below is a scatter diagram illustrating the relationship between BMI and total cholesterol. Each point represents the observed (x, y) pair, in this case, BMI and the corresponding total cholesterol measured in each participant. Note that the independent variable (BMI) is on the horizontal axis and the dependent variable (Total Serum Cholesterol) on the vertical axis



The graph shows that there is a positive or direct association between BMI and total cholesterol; participants with lower BMI are more likely to have lower total cholesterol levels and participants with higher BMI are more likely to have higher total cholesterol levels. In contrast, suppose we examine the association between BMI and HDL cholesterol.

In contrast, the graph below depicts the relationship between BMI and **HDL cholesterol** in the same sample of n=20 participants.

BMI and HDL Cholesterol



This graph shows a negative or inverse association between BMI and HDL cholesterol, i.e., those with lower BMI are more likely to have higher HDL cholesterol levels and those with higher BMI are more likely to have lower HDL cholesterol levels.

For either of these relationships we could use simple linear regression analysis to estimate the equation of the line that best describes the association between the independent variable and the dependent variable. The simple linear regression equation is as follows:

$$\hat{Y} = b_0 + b_1 X$$

where **Y** is the predicted or expected value of the outcome, **X** is the predictor, $\mathbf{b_0}$ is the estimated Y-intercept, and $\mathbf{b_1}$ is the estimated slope. The Y-intercept and slope are estimated from the sample data, and they are the values that minimize the sum of the squared differences between the observed and the predicted values of the outcome, i.e., the estimates minimize:

$$\Sigma\left(Y - \hat{Y}\right)^2$$

These differences between observed and predicted values of the outcome are called **residuals**. The estimates of the Y-intercept and slope minimize the sum of the squared residuals, and are called the **least squares estimates**.[1]

Residuals Conceptually, if the values of X provided a perfect prediction of Y then the sum of the squared differences between observed and predicted values of Y would be 0.

That would mean that variability in Y could be completely explained by differences in X. However, if the differences between observed and predicted values are not 0, then we are unable to entirely account for differences in Y based on X, then there are residual errors in the prediction. The residual error could result from inaccurate measurements of X or Y, or there could be other variables besides X that affect the value of Y.

Based on the observed data, the best estimate of a linear relationship will be obtained from an equation for the line that minimizes the differences between observed and predicted values of the outcome. The **Y-intercept** of this line is the value of the dependent variable (Y) when the independent variable (X) is zero. The **slope** of the line is the change in the dependent variable (Y) relative to a one unit change in the independent variable (X). The least squares estimates of the y-intercept and slope are computed as follows:

$$b_1 = r\frac{s_y}{s_x}$$

and

$$b_0 = \overline{Y} - b_1\overline{X}$$

where

- r is the sample correlation coefficient,
- the sample means are $\overline{X}$ and $\overline{Y}$
- and S_x and S_y are the standard deviations of the independent variable x and the dependent variable y, respectively.

BMI and Total Cholesterol

The least squares estimates of the regression coefficients, b_0 and b_1, describing the relationship between BMI and total cholesterol are $b_0 = 28.07$ and $b_1=6.49$. These are computed as follows:

$$b_1 = r\frac{s_y}{s_x} = 0.78\,\frac{30.8}{3.7} = 6.49$$

and

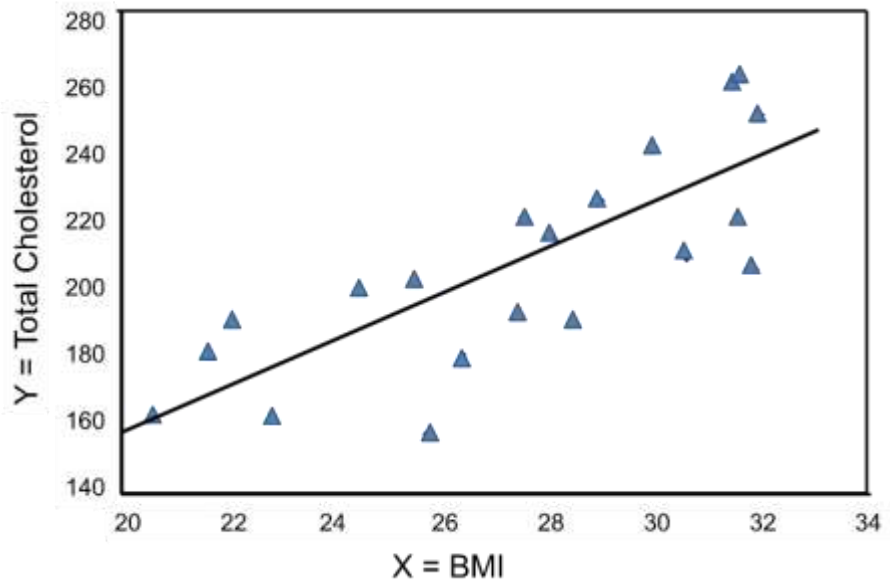
$$b_0 = \overline{Y} - b_1\overline{X} = 205.9 - 6.49\left(27.4\right) = 28.07$$

The estimate of the Y-intercept ($b_0 = 28.07$) represents the estimated total cholesterol level when BMI is zero. Because a BMI of zero is meaningless, the Y-intercept is not informative. The estimate of the slope ($b_1 = 6.49$) represents the change in total cholesterol relative to a one unit change in BMI. For example, if we compare two participants whose BMIs differ by 1 unit, we would expect their total cholesterols to differ by approximately 6.49 units (with the person with the higher BMI having the higher total cholesterol).

The equation of the regression line is as follows:

$$\hat{Y} = 28.07 + 6.49\ \text{BMI}$$

The graph below shows the estimated regression line superimposed on the scatter diagram.



The regression equation can be used to estimate a participant's total cholesterol as a function of his/her BMI. For example, suppose a participant has a BMI of 25. We would estimate their total cholesterol to be 28.07 + 6.49(25) = 190.32. The equation can also be used to estimate total cholesterol for other values of BMI. However, the equation should only be used to estimate cholesterol levels for persons whose BMIs are in the range of the data used to generate the regression equation. In our sample, BMI ranges from 20 to 32, thus the equation should only be used to generate estimates of total cholesterol for persons with BMI in that range.

There are statistical tests that can be performed to assess whether the estimated regression coefficients (b_0 and b_1) are statistically significantly different from zero. The test of most interest is usually H_0: $b_1=0$ versus H_1: $b_1 \neq 0$, where b_1 is the population slope. If the population slope is significantly different from zero, we conclude that there is a statistically significant association between the independent and dependent variables.

BMI and HDL Cholesterol

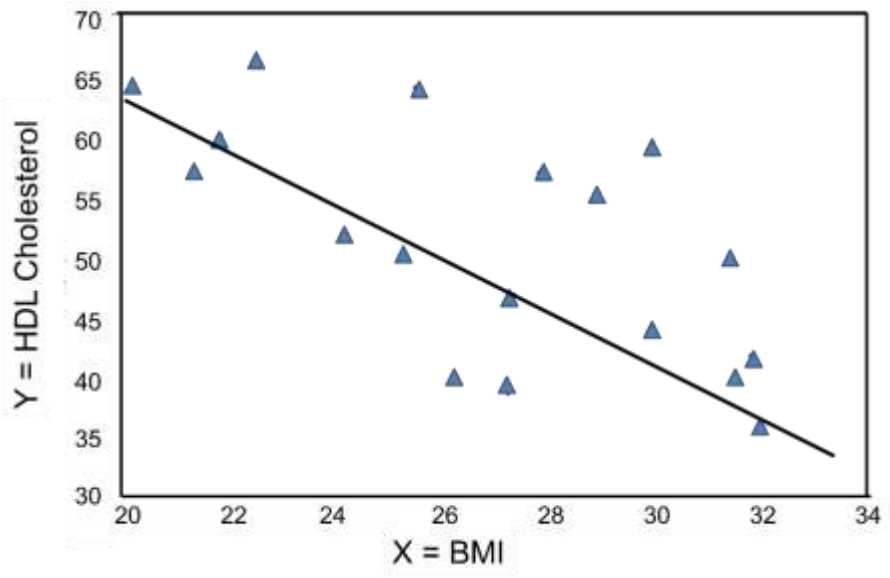
The least squares estimates of the regression coefficients, b_0 and b_1, describing the relationship between BMI and HDL cholesterol are as follows: $b_0 = 111.77$ and $b_1 = -2.35$. These are computed as follows:

$$b_1 = r\frac{s_y}{s_x} = -0.72\frac{12.1}{3.7} = -2.35$$

and

$$b_0 = \bar{Y} - b_1\bar{X} = 47.4 - (-2.35)(27.4) = 111.79$$

Again, the Y-intercept in uninformative because a BMI of zero is meaningless. The estimate of the slope (b_1 = -2.35) represents the change in HDL cholesterol relative to a one unit change in BMI. If we compare two participants whose BMIs differ by 1 unit, we would expect their HDL cholesterols to differ by approximately 2.35 units (with the person with the higher BMI having the lower HDL cholesterol. The figure below shows the regression line superimposed on the scatter diagram for BMI and HDL cholesterol.



Linear regression analysis rests on the assumption that the dependent variable is continuous and that the distribution of the dependent variable (Y) at each value of the independent variable (X) is approximately normally distributed. Note, however, that the independent variable can be continuous (e.g., BMI) or can be dichotomous (see below).

6. Standard Statistical Distributions (e.g. Normal, Poisson, Binomial) and their uses

Statistics: Distributions

Summary

Normal distribution describes continuous data which have a symmetric distribution, with a characteristic 'bell' shape.

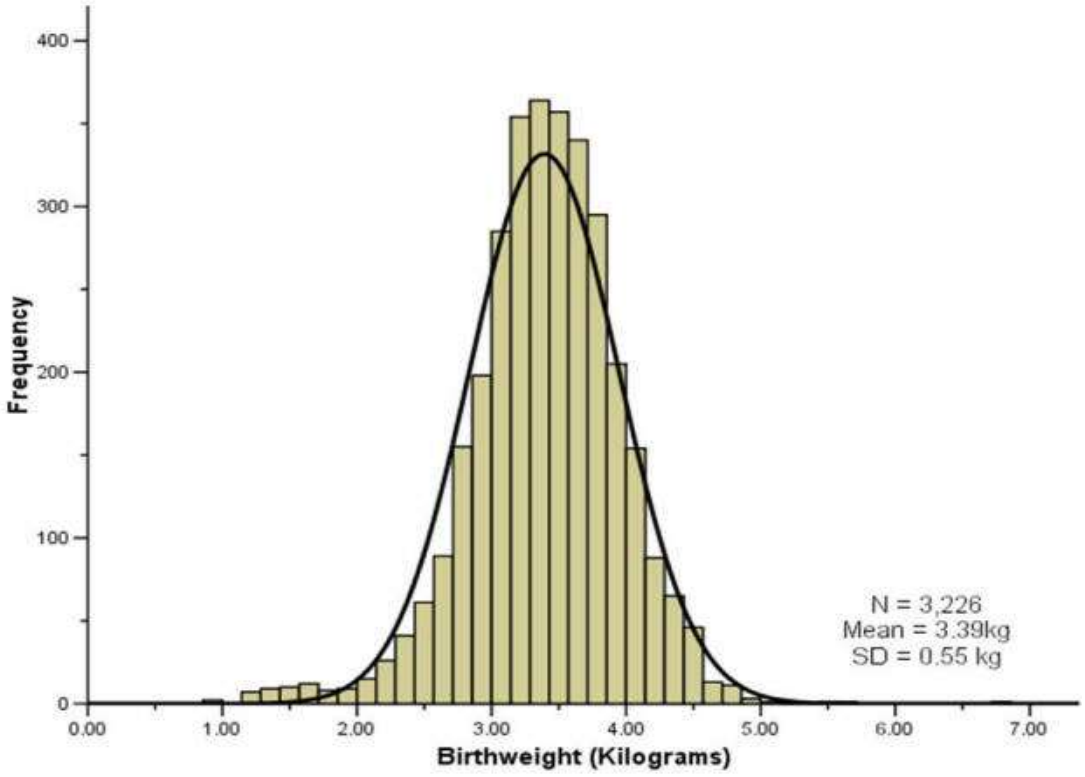
Binomial distribution describes the distribution of binary data from a finite sample. Thus it gives the probability of getting r events out of n trials.

Poisson distribution describes the distribution of binary data from an infinite sample. Thus it gives the probability of getting r events in a population.

6.1. The Normal Distribution

It is often the case with medical data that the histogram of a continuous variable obtained from a single measurement on different subjects will have a characteristic `bell-shaped' distribution known as a Normal distribution. One such example is the histogram of the birth weight (in kilograms) of the 3,226 new born babies shown in Figure 1.

Figure 1 Distribution of birth weight in 3,226 newborn babies (data from O' Cathain et al 2002)



To distinguish the use of the same word in normal range and Normal distribution we have used a lower and upper case convention throughout.

The histogram of the sample data is an estimate of the population distribution of birth weights in new born babies. This population distribution can be estimated by the superimposed smooth `bell-shaped' curve or `Normal' distribution shown. We presume that if we were able to look at the entire population of new born babies then the distribution of birth weight would have exactly the Normal shape. We often infer, from a sample whose histogram has the approximate Normal shape, that the population will have exactly, or as near as makes no practical difference, that Normal shape.

The Normal distribution is completely described by two parameters μ and σ, where μ represents the population mean, or centre of the distribution, and σ the population standard deviation. It is symmetrically distributed around the mean. Populations with small values of the standard deviation σ have a distribution concentrated close to the centre μ; those with large standard deviation have a distribution widely spread along the measurement axis. One mathematical property of the Normal distribution is that exactly 95% of the distribution lies between

μ−(1.96xσ)andμ+(1.96xσ)μ−(1.96xσ)andμ+(1.96xσ)

Changing the multiplier 1.96 to 2.58, exactly 99% of the Normal distribution lies in the corresponding interval.

In practice the two parameters of the Normal distribution, μ and σ, must be estimated from the sample data. For this purpose a random sample from the population is first taken. The sample mean $\bar{x}$ and the sample standard deviation, SD($\bar{x}$)=SSD($\bar{x}$)=S , are then calculated. If a sample is taken from such a Normal distribution, and provided the sample is not too small, then approximately 95% of the sample lie within the interval:

$\bar{x}$−[1.96×SD($\bar{x}$)]$\bar{x}$−[1.96×SD($\bar{x}$)] to $\bar{x}$+[1.96×SD($\bar{x}$)]$\bar{x}$+[1.96×SD($\bar{x}$)]

This is calculated by merely replacing the population parameters μ and σ by the sample estimates $\bar{x}$ and *s* in the previous expression.

In appropriate circumstances this interval may estimate the reference interval for a particular laboratory test which is then used for diagnostic purposes.

We can use the fact that our sample birth weight data appear Normally distributed to calculate a reference range. We have already mentioned that about 95% of the observations (from a Normal distribution) lie within ±1.96 SDs of the mean. So a reference range for our sample of babies, using the values given in the histogram above, is:

3.39 - [1.96 x 0.55] to 3.39 + [1.96 x 0.55]

2.31kg to 4.47kg

A baby's weight at birth is strongly associated with mortality risk during the first year and, to a lesser degree, with developmental problems in childhood and the risk of various diseases in adulthood. If the data are not Normally distributed then we can base the normal reference range on the observed percentiles of the sample, i.e. 95% of the observed data lie between the 2.5 and 97.5 percentiles. In this example, the percentile-based reference range for our sample was calculated as 2.19kg to 4.43kg.

Most reference ranges are based on samples larger than 3500 people. Over many years, and millions of births, the WHO has come up with a normal birth weight range for new born babies. These ranges represent results than are acceptable in newborn babies and actually cover the middle 80% of the population distribution, i.e. the 10th to 90th centiles. Low birth weight babies are usually defined (by the WHO) as weighing less than 2500g (the 10th centile) regardless of gestational age, and large birth weight babies are defined as weighing above 4000kg (the 90th centile). Hence the normal birth weight range is around 2.5kg to 4kg. For our sample data, the 10th to 90th centile range was similar, 2.75 to 4.03kg.

6.2. The Binomial Distribution

If a group of patients is given a new drug for the relief of a particular condition, then the proportion *p* being successively treated can be regarded as estimating the population treatment success rate π.

The sample proportion *p* is analogous to the sample mean $\bar{x}$, in that if we score zero for those *s* patients who fail on treatment, and 1 for those *r* who succeed, then $p=r/n$, where $n=r+s$ is the total number of patients treated. Thus *p* also represents a mean.

Data which can take only a binary (0 or 1) response, such as treatment failure or treatment success, follow the binomial distribution provided the underlying population response rate does not change. The binomial probabilities are calculated from:

$$P\,(r \text{ responses out of } n) = \frac{n!}{r!(n-r)!}\,\pi^r(1-\pi)^{n-r}$$

…for successive values of R from 0 through to n. In the above, *n!* is read as "n factorial" and *r!* as "r factorial". For *r*=4, *r!*=4×3×2×1=24. Both 0! and 1! are taken as equal to 1. The shaded area marked in Figure 2 (below) corresponds to the above expression for the binomial distribution calculated for each of *r*=8,9,...,20 and then added. This area totals 0.1018. So the probability of eight or more responses out of 20 is 0.1018.

For a fixed sample size *n* the shape of the binomial distribution depends only on π. Suppose $n = 20$ patients are to be treated, and it is known that on average a quarter, or $\pi = 0.25$, will respond to this particular treatment. The number of responses actually observed can only take integer values between 0 (no responses) and 20 (all respond). The binomial distribution for this case is illustrated in Figure 2.

The distribution is not symmetric, it has a maximum at five responses and the height of the blocks corresponds to the probability of obtaining the particular number of responses from the 20 patients yet to be treated. It should be noted that the expected value for *r*, the number of successes yet to be observed if we treated *n* patients, is ($n\pi$). The potential variation about this expectation is expressed by the corresponding standard deviation:

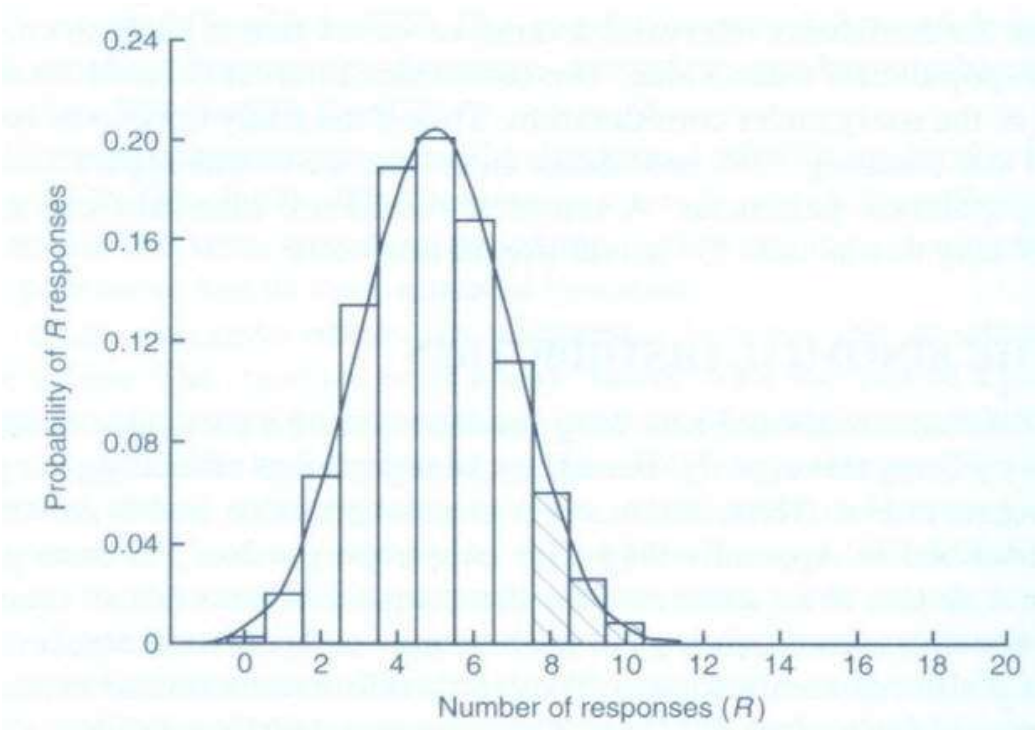
$$\mathrm{SD}\,(r) = \sqrt{n\pi\,(1-\pi)}$$

Figure 2 also shows the Normal distribution arranged to have $\mu = n\pi = 5$ and $\sigma = \sqrt{[n\pi(1 - \pi)]} = 1.94$, superimposed on to a binomial distribution with $\pi = 0.25$ and $n = 20$. The Normal distribution describes fairly precisely the binomial distribution in this case. If n is small, however, or π close to 0 or 1, the disparity between the Normal and binomial distributions with the same mean and standard deviation increases and the Normal distribution can no longer be used to approximate the binomial distribution. In such cases the probabilities generated by the binomial distribution itself must be used.

It is also only in situations in which reasonable agreement exists between the distributions that we would use the confidence interval expression given previously. For technical reasons, the expression given for a confidence interval for a proportion is an approximation. The approximation will usually be quite good provided *p* is not too close to 0 or 1, situations in which either almost

none or nearly all of the patients respond to treatment. The approximation improves with increasing sample size *n*.

Figure 2: Binomial distribution for n=20 with π=0.25 and the Normal approximation



6.3. Poisson Distribution

The Poisson distribution is used to describe discrete quantitative data such as counts in which the population size *n* is large, the probability of an individual event π is small, but the expected number of events, nπ, is moderate (say five or more). Typical examples are the number of deaths in a town from a particular disease per day, or the number of admissions to a particular hospital.

Example

Wight et al (2004) looked at the variation in cadaveric heart beating organ donor rates in the UK. They found that there were 1330 organ donors, aged 15-69, across the UK for the two years 1999 and 2000 combined. Heart-beating donors are patients who are seriously ill in an intensive care unit (ICU) and are placed on a ventilator.

Now it is clear that the distribution of the number of donors takes integer values only, thus the distribution is similar in this respect to the binomial. However, there is no theoretical limit to the number of organ donors that could happen on a particular day. Here the population is the UK population aged 15-69, over two years, which is over 82 million person years, so in this case each member can be thought to have a very small probability of actually suffering an event, in this case being admitted to a hospital ICU and placed on a ventilator with a life threatening condition.

The mean number of organ donors per day over the two year period is calculated as:

$$r = \frac{1330}{(365+365)} = \frac{1330}{730} = 1.82 \text{ organ donations per day}$$

It should be noted that the expression for the mean is similar to that for π, except here multiple data values are common; and so instead of writing each as a distinct figure in the numerator they are first grouped and counted. For data arising from a Poisson distribution the standard error, that is the standard deviation of r, is estimated by SE(r) = $\sqrt{(r/n)}$, where n is the total number of days (or an alternative time unit). Provided the organ donation rate is not too low, a 95% confidence interval for the underlying (true) organ donation rate λ can be calculated in the usual way:

$$r - [1.96 \times \text{SE}(r)] \text{ to } r + [1.96 \times \text{SE}(r)]$$

In the above example r=1.82, SE(r)=$\sqrt{(1.82/730)}$=0.05, and therefore the 95% confidence interval for λ is 1.72 to 1.92 organ donations per day. Exact confidence intervals can be calculated as described by Altman et al. (2000).

The Poisson probabilities are calculated from:

$$P(r \text{ responses}) = \frac{\lambda^r}{r!} e^{-\lambda}$$

…for successive values of r from 0 to infinity. Here e is the exponential constant 2.7182…, and λ is the population rate which is estimated by r in the example above.

Example

Suppose that before the study of Wight et al. (2004) was conducted it was expected that the number of organ donations per day was approximately two. Then assuming $\lambda = 2$, we would anticipate the probability of 0 organ donations in a given day to be $(2^0/0!)e^{-2} = e^{-2} = 0.135$. (Remember that 2^0 and 0! are both equal to 1.) The probability of one organ donation would be $(2^1/1!)e^{-2} = 2(e^{-2}) = 0.271$. Similarly the probability of two organ donations per day is $(2^2/2!)e^{-2} = 2(e^{-2}) = 0.271$; and so on to give for three donations 0.180, four donations 0.090, five donations 0.036, six donations 0.012, etc. If the study is then to be conducted over 2 years (730 days), each of these probabilities is multiplied by 730 to give the expected number of days during which 0, 1, 2, 3, etc. donations will occur. These expectations are 98.8, 197.6, 197.6, 131.7, 26.3, 8.8 days. A comparison can then be made between what is expected and what is actually observed.

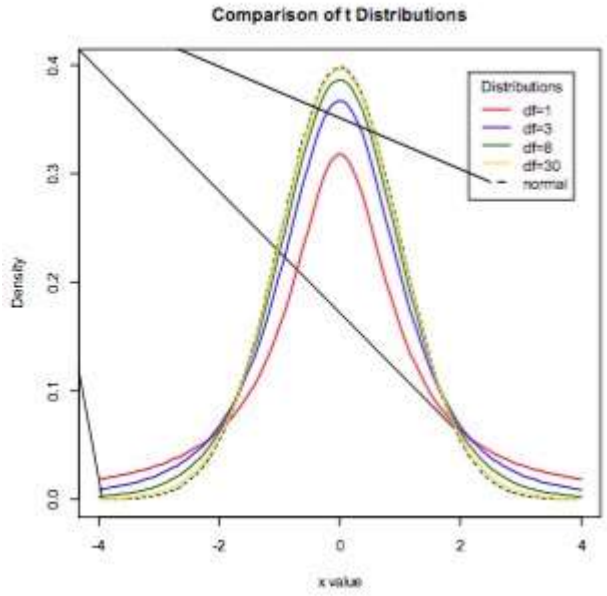
Other Distributions

A brief description of some other distributions are given for completeness.

6.4. t-distribution

Student's *t*-distribution is a continuous probability distribution with a similar shape to the Normal distribution but with wider tails. *t*-distributions are used to describe samples which have been drawn from a population, and the exact shape of the distribution varies with the sample size. The smaller the sample size, the more spread out the tails, and the larger the sample size, the closer the *t*-distribution is to the Normal distribution (Figure 3). Whilst in general the Normal distribution is used as an approximation when estimating means of samples from a Normally-distribution population, when the same size is small (say n<30), the *t*-distribution should be used in preference.

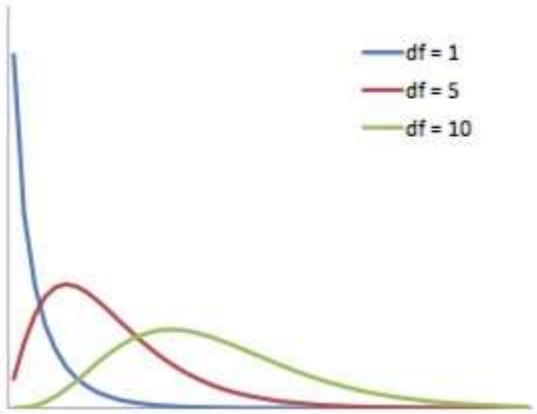
Figure 3. The t-distribution for various sample sizes. As the sample size increases, the t-distribution more closely approximates the Normal.



6.5. Chi-squared distribution

The chi-squared distribution is continuous probability distribution whose shape is defined by the number of degrees of freedom. It is a right-skew distribution, but as the number of degrees of freedom increases it approximates the Normal distribution (Figure 4). The chi-squared distribution is important for its use in chi-squared tests. These are often used to test deviations between observed and expected frequencies, or to determine the independence between categorical variables. When conducting a chi-squared test, the probability values derived from chi-squared distributions can be looked up in a statistical table.

Figure 4. The chi-squared distribution for various degrees of freedom. The distribution becomes less right-skew as the number of degrees of freedom increases.



6. What is hypothesis?

A hypotheses is any assumption/presupposition that the researcher makes about the probable direction of the results that might be obtained on the completion of the research process

- Descriptive hypotheses: This is simply a statement about the magnitude, trend, or behaviour of a population under study.
- Relational hypotheses: These are the typical kind of hypotheses which state the expected relationship between two variables.

A new medicine you think might work.

A way of teaching you think might be better.

A possible location of new species.

A fairer way to administer standardized tests.

It can really be anything at all as long as you can put it to the test.

6.1. What is a Hypothesis Statement?

If you are going to propose a hypothesis, it's customary to write a statement. Your statement will look like this:
"If I…(do this to an independent variable)….then (this will happen to the dependent variable)."
For example:

- If I (decrease the amount of water given to herbs) then (the herbs will increase in size).
- If I (give patients counseling in addition to medication) then (their overall depression scale will decrease).
- If I (give exams at noon instead of 7) then (student test scores will improve).
- If I (look in this certain location) then (I am more likely to find new species).

A good hypothesis statement should:

- Include an "if" and "then" statement (according to the University of California).

- Include both the independent and dependent variables.
- Be testable by experiment, survey or other scientifically sound technique.
- Be based on information in prior research (either yours or someone else's).
- Have design criteria (for engineering or programming projects).

6.2. What is hypothesis testing?

Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results. You're basically testing whether your results are valid by figuring out the odds that your results have happened by chance. If your results may have happened by chance, the experiment won't be repeatable and so has little use.

Hypothesis testing can be done by:

- Figure out your null hypothesis,
- State your null hypothesis,
- Choose what kind of test you need to perform,
- Either support or reject the null hypothesis.

What is the Null Hypothesis?

Simple examples of null hypotheses that are generally accepted as being true are:

- DNA is shaped like a double helix.
- There are 8 planets in the solar system (excluding Pluto).
- Taking Vioxx can increase your risk of heart problems (a drug now taken off the market).

How do I State the Null Hypothesis?

A researcher thinks that if knee surgery patients go to physical therapy twice a week (instead of 3 times), their recovery period will be longer. Average recovery times for knee surgery patients is 8.2 weeks.

The hypothesis statement in this question is that the researcher believes the average recovery time is more than 8.2 weeks. It can be written in mathematical terms as:

$H1: \mu > 8.2$

Next, you'll need to state the null hypothesis . That's what will happen if the researcher is wrong. In the above example, if the researcher is wrong then the recovery time is less than or equal to 8.2 weeks. In math, that's:

H0 $\mu \leq 8.2$

Rejecting the null hypothesis

Ten or so years ago, we believed that there were 9 planets in the solar system. Pluto was demoted as a planet in 2006. The null hypothesis of "Pluto is a planet" was replaced by "Pluto is not a planet." Of course, rejecting the null hypothesis isn't always that easy—the hard part is usually figuring out what your null hypothesis is in the first place.

7. Inferential Statistics

The arrangement of statistical tests which analysts use to make inference from the data given- inferential statistics

These tests make decisions on the basis of observed pattern from data

There is wide range of statistical tests

The choice of which statistical test to utilise relies upon the structure of data, the distribution of data and variable type

There are many different types of tests like t-test, Z-test, chi-square test, ANOVA test, binomial test, one sample median test etc.

7.1 Choosing a Statistical test-

Parametric tests are used if the data is normally distributed. A parametric statistical tests makes an assumption about the population parameters and the distributions that the data came from. These types of test includes t-tests, z-tests and anova tests, which assume data is from normal distribution

7.1.1. Z-test-A z-test is a statistical test used to determine whether two population means are different when variances are known and the sample size is large. In z-test mean of the population is compared. The parameters used are population mean and population standard deviation. Z-test is used to validate a hypothesis tha the sample drawn belongs to the same population

Ho: sample mean is same as the population mean (Null hypothesis)

Ha: sample mean is not same as the population mean (alternate hypothesis)

$$Z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

$\overline{X}$ = sample mean

μo= population mean

$\sigma/\sqrt{n}$ =population standard deviatiom

If the z value is less that critical value accept null hypothesis else reject null hypothesis

Problem 1: A principal at a certain school claims that the students in his school are above average intelligence. A random sample of thirty students IQ scores have a mean score of 112.5. Is there sufficient evidence to support the principal's claim? The mean population IQ is 100 with a standard deviation of 15.

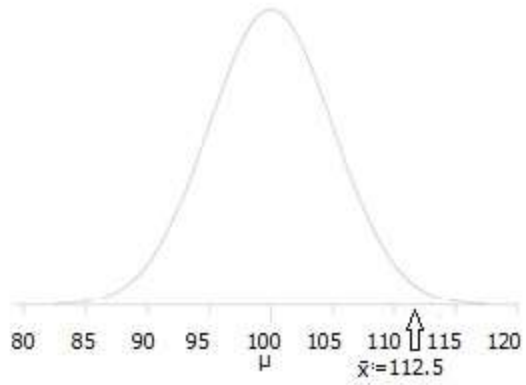
Step 1: State the Null hypothesis. The accepted fact is that the population mean is 100, so: H0: $\mu=100$.

Step 2: State the Alternate Hypothesis. The claim is that the students have above average IQ scores, so:

H1: $\mu > 100$.

The fact that we are looking for scores "greater than" a certain point means that this is a one-tailed test.

Step 3: Draw a picture to help you visualize the problem.



Step 4: State the alpha level. If you aren't given an alpha level, use 5% (0.05).

Step 5: Find the rejection region area (given by your alpha level above) from the z-table. An area of .05 is equal to a z-score of 1.645.

Step 6: Find the test statistic using this formula:z score formula

For this set of data: $z= (112.5 – 100) / (15/\sqrt{30}) = 4.56$.

Step 6: If Step 6 is greater than Step 5, reject the null hypothesis. If it's less than Step 5, you cannot reject the null hypothesis. In this case, it is greater (4.56 > 1.645), so you can reject the null.

Problem 2: Blood glucose levels for obese patients have a mean of 100 with a standard deviation of 15. A researcher thinks that a diet high in raw cornstarch will have a positive or negative effect on blood glucose levels. A sample of 30 patients who have tried the raw cornstarch diet have a mean glucose level of 140. Test the hypothesis that the raw cornstarch had an effect.

Step 1: State the null hypothesis: $H0:\mu=100$

Step 2: State the alternate hypothesis: $H1:\neq100$

Step 3: State your alpha level. We'll use 0.05 for this example. As this is a two-tailed test, split the alpha into two.

0.05/2=0.025

Step 4: Find the z-score associated with your alpha level. You're looking for the area in one tail only. A z-score for 0.75(1-0.025=0.975) is 1.96. As this is a two-tailed test, you would also be considering the left tail (z = 1.96)

Step 5: Find the test statistic using this formula:z score formula

$z = (140 – 100) / (15/\sqrt{30}) = 14.60.$

Step 6: If Step 5 is less than -1.96 or greater than 1.96 (Step 3), reject the null hypothesis. In this case, it is greater, so you can reject the null

7.1.2. T-test : In t-test the mean of the two given samples are compared. A test is used when the population parameters (mean and standard deviation) are not known

Paired t-test: tests for the difference between two variables from the same population (pre- and post test score). For example before and after completion of a program

Independent T-test: The independent t-test which is also called the two sample t-test or student's t-test, is a statistical difference between the means in two unrelated groups. For example comapring boys and girls in a population

One sample t-test: The mean of a single group is compared with a given mean. For example to check the increase and decrease in sales if the average sales is given

Steps in checking hypothesis

1. Check assumptions and write hypotheses

Data must be quantitative. In order to use the t distribution to approximate the sampling distribution either the sample size must be large (≥30) or the population must be known to be normally distributed. The possible combinations of null and alternative hypotheses are:

Research Question	Is the mean different from μ_0?	Is the mean greater than μ_0?	Is the mean less than μ_0?
Null Hypothesis, H_0	$\mu = \mu_0$	$\mu = \mu_0$	$\mu = \mu_0$
Alternative Hypothesis, H_a	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$
Type of Hypothesis Test	Two-tailed, non-directional	Right-tailed, directional	Left-tailed, directional

where μ_0 is the hypothesized population mean.

2. Calculate the test statistic

For the test of one group mean we will be using a test statistic:

Test Statistic: One Group Mean

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

$\overline{x}$ = sample mean
μ_0 = hypothesized population mean
s = sample standard deviation
n = sample size

3. Determine the p-value

When testing hypotheses about a mean or mean difference, a t- distribution is used to find the -value. These distributions are indexed by a quantity called degrees of freedom df=n-1, calculated as for the situation involving a test of one mean or test of mean difference.

4. Make a decision

If the calculated t value is less or equal to table t value -accept null hypothesis

If the calculated t value is greater that table t value-reject null hypothesis

Applications of t-distribution

Estimation of population mean

Problem 1: A sample of 10 adult female rana hexadactyla yielded a mean blood sugar level of 45mg% with a standard deviation of 8mg%. Estimate the population mean of the blood-sugar with 95% confidence

Step1: n=10; $\overline{X}$=45mg%, SD=8mg%

Step 2: SE of mean=SD$\sqrt{n}$ =8/$\sqrt{10-1}$ =2.26

Step 3: table value of t at 0.05 LS with 10-1=9DF-2.26

Step 4: $\mu = \frac{\bar{X} \pm t_{\alpha f} S}{\sqrt{n}}$

$= 45 \pm 2.26(2.67) =$

$45 \pm 6.03 = 38.97$ to 51.03mg%

Matched paired analysis

Problem 2:

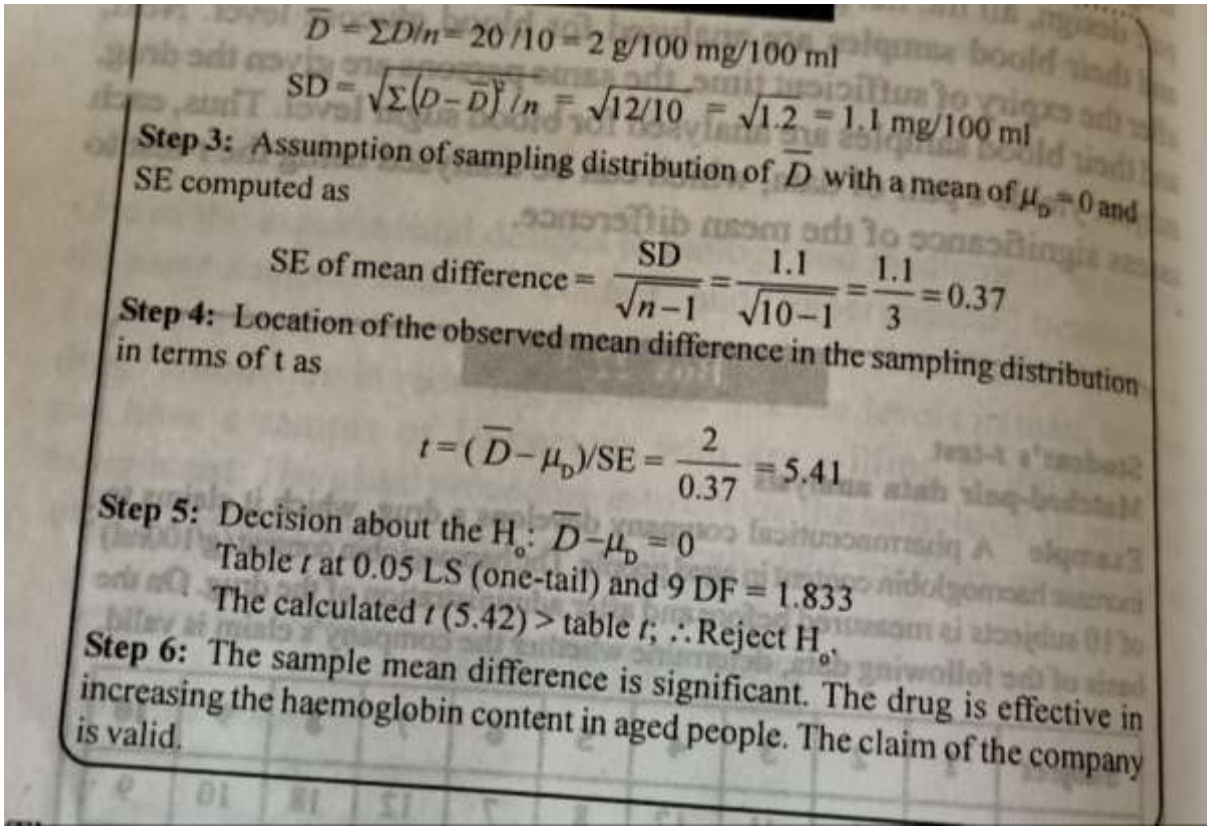
Example A pharmaceutical company develops a drug, which it claims to increase haemoglobin content in aged people. The haemoglobin content (g/100ml) of 10 subjects is measured before and after administration of the drug. On the basis of the following data, determine whether the company's claim is valid.

Subject	1	2	3	4	5	6	7	8	9	10
Before	10	9	11	12	8	7	12	18	10	9
After	12	11	13	14	9	10	12	14	11	12

Step 1: $H_o: \overline{D} - \mu_D = 0$; 0.05 LS one-tailed test with $10 - 1 = 9$ DF.

Step 2: Computation of $\overline{D} = \Sigma D/n$ and SD $= \sqrt{\Sigma(D - \overline{D})^2/n}$

Before	After	D	$D - \overline{D}$	$(D - \overline{D})^2$
10	12	2	0	0
9	11	2	0	0
11	13	2	0	0
12	14	2	0	0
8	9	1	−1	1
7	10	3	1	1
12	12	0	−2	4
10	14	4	2	4
10	11	1	−1	1
9	12	3	1	1
		20		12



$\bar{D} = \Sigma D/n = 20/10 = 2$ g/100 mg/100 ml

$SD = \sqrt{\Sigma(D-\bar{D})^2/n} = \sqrt{12/10} = \sqrt{1.2} = 1.1$ mg/100 ml

Step 3: Assumption of sampling distribution of $\bar{D}$ with a mean of $\mu_D = 0$ and SE computed as

$$\text{SE of mean difference} = \frac{SD}{\sqrt{n-1}} = \frac{1.1}{\sqrt{10-1}} = \frac{1.1}{3} = 0.37$$

Step 4: Location of the observed mean difference in the sampling distribution in terms of t as

$$t = (\bar{D} - \mu_D)/SE = \frac{2}{0.37} = 5.41$$

Step 5: Decision about the H_o: $\bar{D} - \mu_D = 0$

Table t at 0.05 LS (one-tail) and 9 DF = 1.833

The calculated t (5.42) > table t; $\therefore$ Reject H_o.

Step 6: The sample mean difference is significant. The drug is effective in increasing the haemoglobin content in aged people. The claim of the company is valid.

7.1.3. F test

In biology we come across situations where in we have to compare the parameters of more than two populations. Use of student's t-test is rather difficult in such situations for the obvious reason of the increasing number of tests we will have to perform to compare each pair of means in all possible combinations. Further, the probability of error would become enormous, proportional to number of tests

Eg: Suppose 3 samples for eg: comparing 3 drugs or 2 drugs with the control or 1 drug with control placebo . So more than 2 sets of samples we cannot so t-test.

In such cases F test is performed. This statistical method compares variances to test the hypothesis

Hypotheses and alternatives

1. $H_0: \sigma_1^2 = \sigma_2^2$
   $H_1: \sigma_1^2 \neq \sigma_2^2$
2. $H_0: \sigma_1^2 = \sigma_2^2$
   $H_1: \sigma_1^2 > \sigma_2^2$

Test statistics

$$F = \frac{s_1^2}{s_2^2}, \quad (s_1^2 > s_2^2)$$

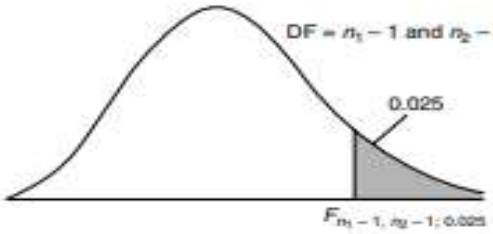
where s_1^2 and s_2^2 are sample variances.
(If, in 2, $s_1^2 < s_2^2$, do not reject H_0.)

When used

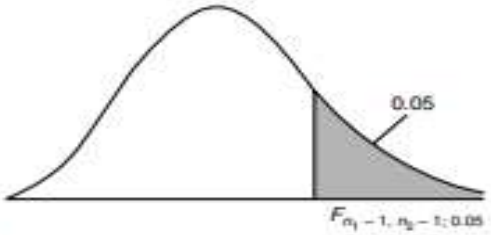
Given two sample with unknown variances σ_1^2 and σ_2^2 and normal populations.

Critical region and degrees of freedom

1.



2.



Data

$H_0: \sigma_1^2 = \sigma_2^2$
$n_1 = 11, n_2 = 16$
$s_1^2 = 6.0, s_2^2 = 3.0$
$\therefore F = 2.0$

Conclusion

1. $F_{10,15;0.025} = 3.06$.
   Do not reject H_0.
2. $F_{10,15;0.05} = 2.54$. [see Table 3].
   Do not reject H_0.

Problem1:

7 women from a population of women and 12 men from a population of men were tested for blood glucose levels. Table shows the standard deviation of blood glucose in each sample

Population	Population std deviation σ	Sample standard deviation s
Women	30	35
men	50	45

$F=(s_1^2/ \sigma_1^2)/ (s_2^2/ \sigma_2^2)$

Women/men= $[(35)^2-(30)^2]/ [(45)^2-(50)^2] = 1.68$ men/women= $[(45)^2-(50)^2]/ [(35)^2-(30)^2] = 4.03$

Numerator df=7-1=6

Denominator df=12-1=11

F_{table}= 3.09 p=0.05

Cannot reject null hypothesis

Numerator df=12-1=11

Denominator df=7-1=6

F_{table}= 4.03 p=0.05

Cannot reject null hypothesis

Problem2:

During clinical trails 9 blood samples were tested in different labs located in different parts of the world. The standard deviation of results in lab 1 and lab 2 were 9.34 and 6.59(mg/ml) resp. Ascertain by F test whether there is a statistical difference in results?

Ho=$\sigma_1^2 = \sigma_2^2$

Ha=$\sigma_1^2 \neq \sigma_2^2$

F= $(9.34)^2/(6.59)^2$= 2.01

F table= 8/8 = 3.44 p=0.05

Since table vale is greater than calculated F value null hypothesis cannot be rejected

ANOVA ("Analysis of Variance"):

F- test is also used in the context of analysis of variance (ANOVA) for judging the significance of more than two sample means at one and the same time. It is also used for judging the significance of multiple correlation coefficients.

Assumptions:

1. The groups are independent simple random samples

2. The populations are normally distributed

3. The population variances are equal $\sigma1 = \sigma2 = \sigma3 = \cdots = \sigma k$

We must estimate two types of variances:

• Variance between the groups

• Variance within the group

$$F = \frac{\text{Estimate of population variance based on between samples variance}}{\text{Estimate of population variance based on within samples variance}}$$

This value of F is to be compared to the F-table for given degrees of freedom. If the F value we work out is equal or exceeds* the F-table value we may say that there are significant differences between the sample means

The variance within samples is large the F value will be small and we cannot reject null hypothesis.

Problem 1: Set up an analysis of variance table for the following per acre production data for three varieties of wheat, each grown on 4 plots and state if the variety differences are significant.

Plot of land	*Per acre production data*		
	Variety of wheat		
	A	*B*	*C*
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

Solution through direct method: First we calculate the mean of each of these samples:

$$\bar{X}_1 = \frac{6+7+3+8}{4} = 6$$

$$\bar{X}_2 = \frac{5+5+3+7}{4} = 5$$

$$\bar{X}_3 = \frac{5+4+3+4}{4} = 4$$

Mean of the sample means or $\quad \bar{\bar{X}} = \dfrac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{k}$

$$= \frac{6+5+4}{3} = 5$$

Now we work out *SS* between and *SS* within samples:

$$SS \text{ between} = n_1\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 + n_2\left(\bar{X}_2 - \bar{\bar{X}}\right)^2 + n_3\left(\bar{X}_3 - \bar{\bar{X}}\right)^2$$

$$= 4(6-5)^2 + 4(5-5)^2 + 4(4-5)^2$$
$$= 4 + 0 + 4$$
$$= 8$$

$$SS \text{ within} = \Sigma\left(X_{1i} - \bar{X}_1\right)^2 + \Sigma\left(X_{2i} - \bar{X}_2\right)^2 + \Sigma\left(X_{3i} - \bar{X}_3\right)^2, \qquad i = 1, 2, 3, 4$$

$$= \{(6-6)^2 + (7-6)^2 + (3-6)^2 + (8-6)^2\}$$
$$+ \{(5-5)^2 + (5-5)^2 + (3-5)^2 + (7-5)^2\}$$
$$+ \{(5-4)^2 + (4-4)^2 + (3-4)^2 + (4-4)^2\}$$
$$= \{0+1+9+4\} + \{0+0+4+4\} + \{1+0+1+0\}$$
$$= 14 + 8 + 2$$
$$= 24$$

$$SS \text{ for total variance} = \Sigma\left(X_{ij} - \bar{\bar{X}}\right)^2 \qquad i = 1, 2, 3\ldots$$

$$j = 1, 2, 3\ldots$$

$$= (6-5)^2 + (7-5)^2 + (3-5)^2 + (8-5)^2$$
$$+ (5-5)^2 + (5-5)^2 + (3-5)^2$$
$$+ (7-5)^2 + (5-5)^2 + (4-5)^2$$
$$+ (3-5)^2 + (4-5)^2$$
$$= 1 + 4 + 4 + 9 + 0 + 0 + 4 + 4 + 0 + 1 + 4 + 1$$
$$= 32$$

Alternatively, it (*SS* for total variance) can also be worked out thus:

SS for total = *SS* between + *SS* within

$$= 8 + 24$$
$$= 32$$

We can now set up the ANOVA table for this problem:

Table 11.2

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit (from the F-table)*
Between sample	8	$(3-1)=2$	$8/2=4.00$	$4.00/2.67=1.5$	$F(2,9)=4.26$
Within sample	24	$(12-3)=9$	$24/9=2.67$		
Total	32	$(12-1)=11$			

The above table shows that the calculated value of F is 1.5 which is less than the table value of 4.26 at 5% level with d.f. being v1 = 2 and v2 = 9 and hence could have arisen due to chance. This analysis supports the null-hypothesis of no difference is sample means. We may, therefore, conclude

that the difference in wheat output due to varieties is insignificant and is just a matter of chance

Non parametric statistical test: Non parameteric tests are used when data is normally distributed. Non parametric tests include chi-square test

Chi-square test (χ2-test)

Chi square test is used to compare categorical variables. Calculating the chi-square static value and comparing it against a critical value from the chi-square distribution allows to assess whether the observed frequency are significantly different from the expected frequency

The hypothesis being tested for chi-square is

Ho= variable x and variable y are independent

Ha= variable x and variable y are not independent

Chi Square formula is

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

O = the frequencies observed

E = the frequencies expected

$\sum$ ***= the 'sum of'***

E is computed by

$$E = \frac{\text{row total} \times \text{column total}}{\text{sample size}}$$

We will compare the calculated value to the table value of with degree of freedom = (row total - 1) (column - 1), and reject the null hypothesis if calculated value is greater than table value

The chi square value from the table is calculated by using p=0.05 and degrees of freedom

Problem: We are investigating the seed shape of pea plant. We think round seed is caused by dominat allele (R) and wrinkled seed is caused by recessive allele (r). we crossed parents who are heterozygous for seed shape which produced 5474 round seeds and 1850 wrinkled seeds in the

offspring. Are these observed values similar to expected values for a cross between heterozygous individuals?

Phenotype	**Observed (O)**	**Expected (E)**	**(O-E)**	**$(O-E)^2$**	**$(O-E)^2/E$**
Round seeds	5474				
Wrinkled seeds	1850				
Sum the last values to calculate **$\chi 2$**					**$\chi 2$=**

To determine the expected values perform punnet square which allows to determine the probability of certain phenotype will occur in the offspring.

Let us do punnet square the heterozygous parents Rr x Rr

	R	**r**
R	RR	Rr
r	Rr	rr

Possible offsprings in the F1 generation

From this ¾ or 75% offsprings are expected to be round and ¼ or 25% are expected to be wrinkled

Total number of offsprings observed= 5474+1850=7324

Expected number of round off springs = 7324X0.75 (i.e., 75%) = 5493

Expected number of wrinkled off springs= 7324x0.25 (i.e., 25%) = 1831

Phenotype	**Observed (O)**	**Expected (E)**	**(O-E)**	**$(O-E)^2$**	**$(O-E)^2/E$**
Round seeds	5474	5493	-19	361	361/5493=0.066
Wrinkled seeds	1850	1831	19	361	361/1831=0.197
Sum the last values to calculate **$\chi 2$**					**$\chi 2$=0.263**

The calculate **$\chi 2$** value is **0.263**

If the p value is small $p<0.05$ statistically significant between observed and expected values

To determine the table $\chi 2$ value the degree of freedom is determined by (column-1)(rows-1)= 1

At p value of 0.05 and df=1 the table $\chi 2$ value is 3.84

Since the table value is greater that calculated value the null hypothesis is accepted

UNIT – II -DESIGN OF EXPERIMENT, SAMPLING AND DATA COLLECTION– SBT1002

2. RESEARCH DESIGN

2.1. Meaning of Research Design

A research design is a controlling plan for a research study in which the methods and procedures for collecting and analyzing the information to be collected is specified. It is a framework or plan for study that guides the collection and analysis of data.

The word 'design' means to work out the structure of form', as by making a sketch or plan. Thus, 'Research Design' is planning a strategy or drawing a blue print of conducting research. It is a guideline for collecting and utilizing data so that desired information can be obtained with sufficient precision and hypothesis can be tested properly. A research is designed for the purpose of producing results that may be applied to real world situations. It not only enables a researcher to anticipate potential problems that can occur during the actual operation of the research, but also to limit boundaries of research study.

2.2. Definitions of Research Design

Some of the popular definitions of research design are:

1. "Research design is the planned sequence of the entire process involved in conducting a research study." By Miller.

2. "Research design is a catalogue of the various phases and facts relating to the formulation of a research effort. It is an arrangement of the essential conditions for collection and analysis of data in a form that aims to combine relevance to research purpose with economy in the procedure". By Selltiz and others.

3. "A research designates the logical manner in which individuals or other units are compared and analyzed, it is the basis of making interpretations from the data". By Anonymous.

4. "Also known as a market research briefing, this is a basic plan which guides the data collection and analysis phased of the research project. It acts a frame work which details the type of information to be collected, the data sources and the data collection procedure". By Market Intelligence Group, India

In short, research design is a plan of what data to gather, from whom, how and when to collect the data, and how to analyze the data obtained.

2.3 Need for Research Design

Research is a scientific investigation of a problem for which we need a systematic planning of research. For a successful research we need a research design because it includes (i) the formulation of a strategy to resolve a particular question (ii) the collection and recording of information and evidence (iii) the processing and analysis of these data and their interpretation and (iv) the publication of results.

A research design states structure and process of conducting a research process. Thus, it shows a path to researcher without which he may be lost or confused as to what next step he has to take. More so, it also takes care of budget and time frame of the research study. All this planning can only make a research study a success story

2.4 Features of a Good Research Design

It is a challenge to translate general scientific model into a practical research operation. Therefore, designing a research study is not a simple task. There is nothing like completely correct design or completely incorrect design. A design may work very well for one research problem and may not work at all for the other. There are some features, however, a good research design should posses.

They are:

1. Freedom from bias: A good research design should ensure that the method of data collection and analysis would not cause the data to vary in a systematic way. That is to say that the data should be free from systematic errors.

2. Freedom from confounding: In a good research design the variables involved in the study are separated from each other so that they do not influence each other.

3. Control of extraneous variables: In a well-designed research study the variables that are not under scrutiny do not influence the experimental variables in a systematic way for example things like temperature, time of day etc.

4. Statistical precision for testing hypothesis: A research design should ensure that the data are recorded at a level of precision that will yield statistically meaningful results.

5. With in resources: A design should draw limits of a research study so that it could be completed within available resources like time, money and staff.

6. Optimality: The best research design is one, which yields maximum precision in terms of bias and variance using minimum resources in terms sample size, time and money.

7. Objectivity: If operated by more than one researcher a good research design obtains same results. Thus, a good research design should be free from the subjectivity of its performer.

8. Flexibility: It is often observed that one has to deviate from the basic research design during the operation of the research study due to real world problems. A good research design is one, which not only has the potential to predict such practical problems, but also is flexible enough to incorporate changes in it whenever needed.

2.5. Different Research Designs

After the formulation and definition of research problem, the next step is to choose an appropriate research design. Every research study is unique in itself, but there are certain things common in these studies. On the basis of these commonalities one can categorize the research studies by research methods and procedures used to collect and analyze data. Accordingly a research design is chosen. There are three basic types of research designs:

1. Exploratory

2. Descriptive or diagnostic research design

3. Causal or Hypothesis testing or Experimental research design

1. Exploratory

Exploratory research is defined as collecting information in an unstructured and informal way. For example, a restaurant owner may regularly visit other competing restaurants in order to gather information about menu selection, prices and service quality.

In exploratory type of research, the investigation may be conducted because a problem has not been clearly defined. It helps in determining the best research design, data collection method and selection of subjects. Usually exploratory research is qualitative in nature. Some times exploratory research may even conclude that a perceived problem does not actually exist.

Generally an exploratory research design helps in finding out the feasibility of the research problem, getting familiar with various components of the study, generating new ideas, and formulating the hypothesis. Exploratory research designs are of different forms depending on the nature and objectives of the study. The following three forms are most popular:

(a) Literature Survey: In any research, review of literature is an essential part. The literature survey is carried out at a preliminary stage of the research. Through the review, one understands the work that has already been done and what more can be explored in one's chosen field. The theories and techniques used in the existing literature can be used in the present analysis or these can be modified to give better results.

The literature are the documentary sources of information which are contained in the published and unpublished documents, reports, statistics, manuscripts, letters, diaries, and so on. It is important for scientific workers to scrutinize these sources very closely. Since not all documents can be consulted, it is best to start a selective process early. The theory and techniques of the literature must serve useful purpose in the present study. And, their meaning should not have altered with changing circumstances with the passage of time.

(b) Expert Survey: Expert Survey or experience survey means consulting the experienced researchers who are experts in the field of study. One should not be shy in taking advice and guidance of such people. They should be given sometime with the problem before asking them questions about the study, so that they can give their opinion after a good thought on the problem.

(c) Example Survey: In case of a new type of studies sometimes neither much literature nor expert advice is available. In such situations it is advisable to go through some case studies performed in the past. This refers to 'insight stimulating examples'. Single cases or a group of cases, as may be relevant to the research study are selected and studied in order to collect data for main study.

2. Descriptive or diagnostic research design

Descriptive research refers to a set of methods and procedures that describe the study variables. Descriptive studies portray these variables by answering who, what, why and how questions. These types of research studies may describe such things as consumer's attitudes, intentions, behaviors or the number of competitors and their strategies.

Descriptive research is also known as statistical research or diagnostic research. It describes data and characteristics about the population or phenomenon being studied. The description is used for frequencies, averages and other statistical calculations.

The process of conducting descriptive research can be linked to that of passing an idea through an hourglass. The research starts with a consideration of the larger issues of interest, and these are then narrowed into a specific questions (hypothesis) that can only be evaluated with some degree of

control. The components of the hypothesis are operationalized into observable units and behaviors to ensure that the independent and dependent variables can be observed and measured. Research is then conducted to observe the relationships of interest, in the context of the specified research environment. Observations are made, and data are collected to reflect behaviors, changes and other indicators of interest.

The data are filtered and analyzed in order to generate conclusions that may support or refute the hypothesis, and then everything is considered in the context of the bigger picture, which usually includes reference and association to the board issues that started the process.

Although data description is factual, accurate and systematic, the research cannot describe what caused a situation. Thus, descriptive research cannot be used to create causal relationship where one variable affects another.

3. Causal or Hypothesis testing or Experimental research design

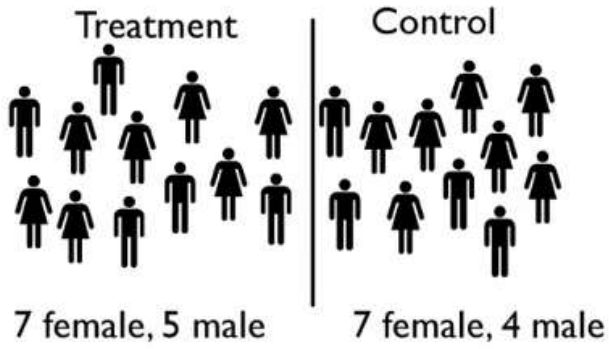
Causal research designs are used in hypothesis testing research or experimental research studies. This type of research design is conducted by controlling various factors to determine which factors are causing the problem. It isolates causes and effects. By changing one factor, say price you can monitor its effects on a key consequence such as sales. Although experimental research can give a high level of understanding of the variables under study, the designs often require experiments that are complex and expensive. Some of the popular experimental designs are:

(a) Completely randomized design: A completely randomized design means that treatments will be randomly assigned to individual participants in an experiment. However, a disadvantage of this design is that treatment and control groups could have disproportionate representations of the population.

Eg: Let’s say you developed a new drug to combat the symptoms of acid reflux. You want to see if it’s more effective than what is currently available. So you get 500 volunteers and write “1” on 250 slips of papers and “2” on the other 250 slips of paper. You put all 500 sheets of papers into a hat, mix them up, and the volunteers retrieve one slip of paper each.

Those who selected “1” will receive the new drug and those who selected “2” receive the drug that's currently available. This is the simplest way to assign subjects to treatments. However, it's not necessarily ideal for every scenario.

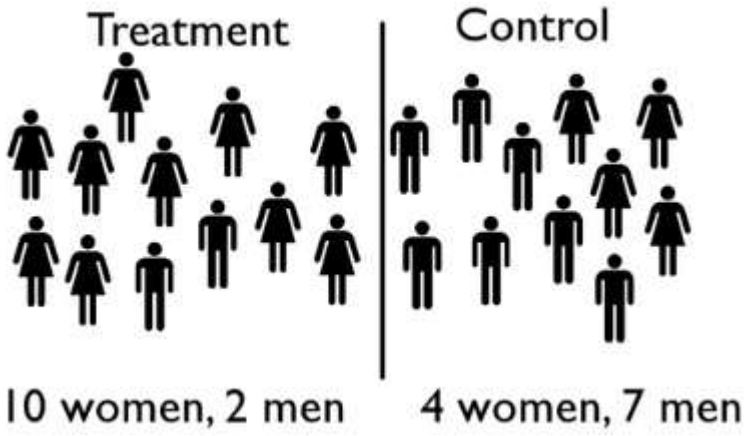
Let’s say that the acid reflux drug is more effective for men than it is for women. It’s not really a problem if you divide the treatment control groups like this:



Roughly the same amount of females and males in the treatment group and the control group.

In this particular case, you can see there is roughly the same amount of females and males in the treatment group and the control group. Since there is a relative equal assignment on each side, it will be easy to see if the new drug is more effective for males than for females. Problems occur when the random assignment doesn't match the proportions of the population equally.

Consider for a moment if this happened:



Disproportionate amount of women in the first group and a disproportionate amount of men in the second group.

Both groups are roughly the same size. Will you be able to determine if the treatment is more effective for men? Why not?

If the drug were more effective for men and than women, you actually wouldn't notice because there aren't that many men in the treatment group. The proportions are way out of whack. This sometimes happens with random assignment.

(b) Randomized block design:

With a randomized block design, the experimenter divides subjects into subgroups called blocks, such that the variability within blocks is less than the variability between blocks. Then, subjects within each block are randomly assigned to treatment conditions.

Steps: In randomized block design, first participants are divided into homogeneous groups-- groups that are the same across some variable of interest. Some examples are age, race, income, location, job, or gender.

Once the participants are in their similar groups, they are randomly assigned to treatment and control within that group.

An advantage is that it controls for variables that would otherwise be confounding. If we think that job has an effect, this way we can make sure that a proportionate number of people who have the same job are assigned to treatment and control groups.

A disadvantage is that it reduces the sample size for each group.

Example: A researcher wants to study the effects of the two main types of laser eye surgery on participants. He thinks that age is a significant factor. So first the researcher would divide the participants into groups based on age.

These are some groups he could use. He could use under 30, 30 to 40, 40 to 50, 50 to 60, and 60 plus. Now, once people are sorted on age, you then randomly assign people within the group to one of the two treatments, so that that way there are a group of people under 30 who receive the first treatment, and who receive the second treatment. There are a group of people between 30 and 40 who receive each of the two types of treatments. It helps the researcher to examine what the effect of age is, and what kind of role that plays in how well the treatments work out for the people.

Example 2: A farmer wants to test out fertilizer for a company. The company has three different types of fertilizer, and wants to know what's best. The farmer knows his land really well. He knows that it's not all the same.

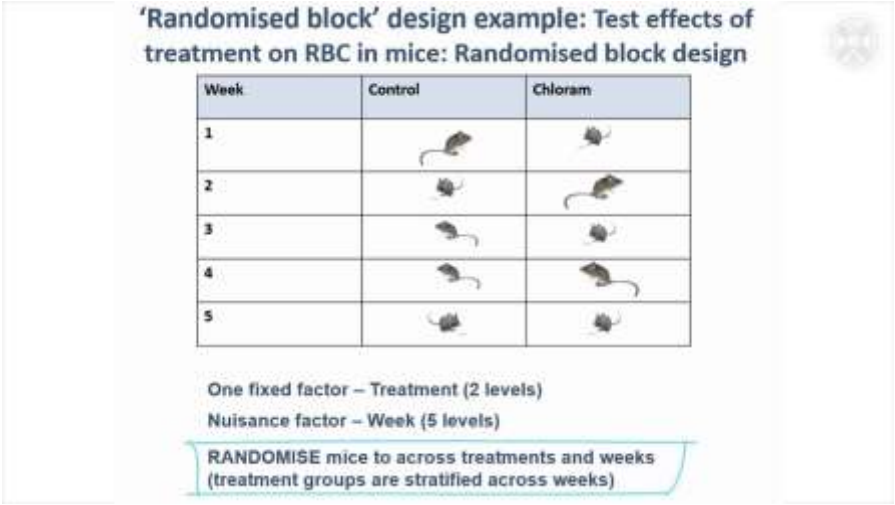
He has a chunk by his house that gets a lot more sunlight, a chunk up by a river that gets a lot more water, and a chunk in the backfield that gets a lot more dry. If the farmer just applied the treatments randomly and had no concern for these differences of his land, the differences of the land would affect the results of the study.

If, for example, Treatment A got mostly in the water area, and Treatment A for some reason was really ineffective in a watery area, but overall was best everywhere else in his field, he wouldn't be able to determine that from the study. So instead, the farmer makes blocks out of his land.

He creates blocks based on what he knows to be similar. So he has a block down here for the really sunlit land, the block up here for the really watered land, and a block over here for the very dry land. Once he set up the blocks, he assigns the treatments randomly within each block.

Now, this is a very small-scale example. Obviously a company wouldn't just use one farmer. Most likely the farmer wouldn't assign just one set of block treatments. But it can give you an idea of what would happen.

In order to minimize the effects of the differences between the groups, we first make blocks, and then randomly assign the treatments within the block.



(c) Latin square design

What is a Latin Square?

A Latin square is an ancient puzzle where you try to figure out how many ways Latin letters can be arranged in a set number of rows and columns (a matrix); each symbol appears only once in each row and column. It's called a Latin square because it was developed based on Leonard Euler's works, which used Latin symbols. However, any letters can be used.

A	B	C
C	A	B
B	C	A

A 3×3 square.

What is a Balanced Latin Square Design?

While a Latin square ensures that each letter appears an equal number of times, it doesn't protect against order effects. For example, in the above 3 x 3 example square, treatment B follows A three times (in the rows). It follows C zero times. A balanced Latin square evens out the what-follows-what scenario, protecting against order effects.

If you have an even number of experimental conditions, constructing the square is very easy.

Making the Square

The following algorithm can be used to make a balanced square for an even number of testing conditions:

1	2	n	3	n-1	4	...
2	3	1	4	n	5	...
3	4	2	5	1	6	...
⋮	⋮	⋮	...	⋮	⋮	⋮
n-1	n	n-2	1	n-3	2	...
n	1	n-1	2	n-2	3	...

Balanced square algorithm (UF).

Example

Make a balanced square design for six participants A B C D E F with six testing conditions.

Step 1: Make the first row using the formula: **row_1 = 1,2,n,3,n-1,n-2**.

			Order of Testing Conditions			
Participant	1st	2nd	3rd	4th	5th	6th
A	1	2	6	3	5	4
B						
C						
D						
E						
F						

Step 2: Fill in the first column sequentially.

			Order of Testing Conditions			
Participant	1st	2nd	3rd	4th	5th	6th
A	1	2	6	3	5	4
B	2					
C	3					
D	4					
E	5					
F	6					

Step 2: Continue filling in the columns sequentially until the square is completed.

			Order of Testing Conditions			
Participant	1st	2nd	3rd	4th	5th	6th
A	1	2	6	3	5	4
B	2	3	1	4	6	5
C	3	4	2	5	1	6
D	4	5	3	6	2	1
E	5	6	4	1	3	2
F	6	1	5	2	4	3

A completed balanced square design with an even number of conditions.

For an **odd number of conditions**, making the balanced square becomes a lot more complicated (the experiment will also likely become more time consuming and expensive). Each participant must be tested twice for each condition, doubling the number of treatments given. For example, if you have 5 treatments, participant A might be given {1,2,5,3,4,3,2,1,5,4}.

(d) Factorial design.

Factorial design is a type of research methodology that allows for the investigation of the main and interaction effects between two or more independent variables and on one or more outcome variable(s).

Such **designs** are classified by the number of levels of each factor and the number of factors. So a 2x2 **factorial** will have two levels or two factors and a 2x3 **factorial** will have three factors each at two levels.

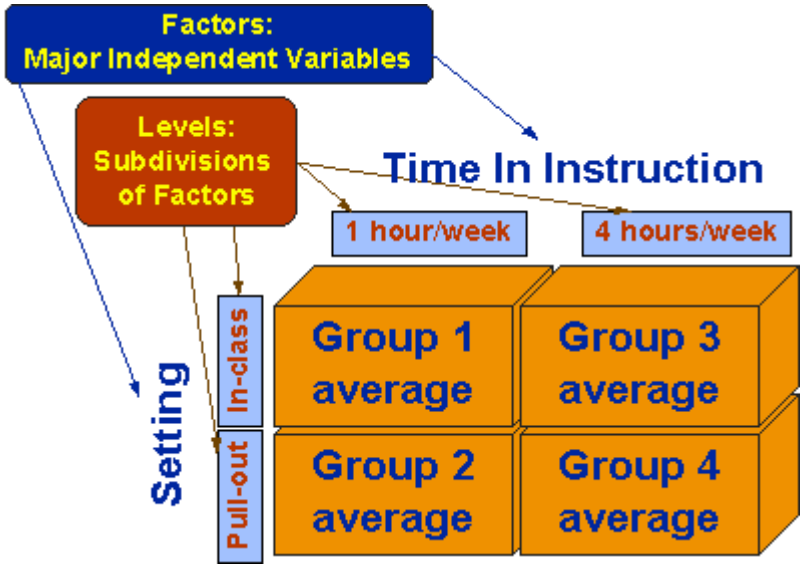
Uses of a factorial design:

Factorial designs can test limits; to test whether an independent variable effects different kinds of people, or people in different situations, the same way.

Example, we would like to vary the amount of time the children receive instruction with one group getting 1 hour of instruction per week and another getting 4 hours per week. And, we'd like to vary the setting with one group getting the instruction in-class (probably pulled off into a corner of the classroom) and the other group being pulled-out of the classroom for instruction in another room. We could think about having four separate groups to do this, but when we are varying the amount of time in instruction, what setting would we use: in-class or pull-out? And, when we were studying setting, what amount of instruction time would we use: 1 hour, 4 hours, or something else?

With factorial designs, we don't have to compromise when answering these questions. We can have it both ways if we cross each of our two time in instruction conditions with each of our two settings. Let's begin by doing some defining of terms. In factorial designs, a **factor** is a major independent variable. In this example we have two factors: time in instruction and setting. A **level** is a subdivision of a factor. In this example, time in instruction has two levels and setting has two levels. Sometimes we depict a factorial design with a numbering notation. In this example, we can say that we have a 2 x 2 (spoken "two-by-two) factorial design. In this notation, the *number of numbers* tells you how many factors there are and the *number values* tell you how many levels. If I said I had a 3 x 4 factorial design, you would know that I had 2 factors and that one factor had 3 levels while the other had 4. Order of the numbers makes no difference and we could just as easily

term this a 4 x 3 factorial design. The number of different treatment groups that we have in any factorial design can easily be determined by multiplying through the number notation. For instance, in our example we have 2 x 2 = 4 groups. In our notational example, we would need 3 x 4 = 12 groups.



3. **Data Collection: Primary and Secondary**

3.1. Introduction

It is a term used to describe a processing of preparing and collecting data

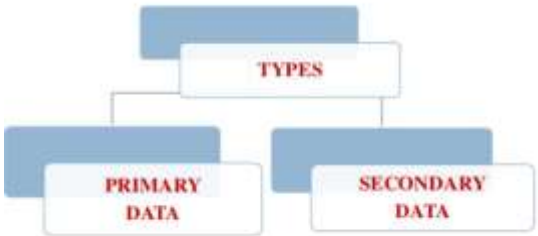
Systemic gathering of data for a particular purpose from various sources, that has been systemically observed, recorded and organised

Data are the basic inputs to any decision-making process in research

3.2. Purpose of data collection

- To obtain information
- To keep on record
- To make decisions about important issues
- To pass on information to others

3.3. Classification of data



3.3.1 Primary data:

It is a term for data collected at source. This type of information is obtained directly from first hand sources by means of surveys, observations and experimentation and not subjected to any processing or manipulation and also called primary data.

Primary data means original data that has been collected specially for the purpose in mind. It means someone collected the data from the original source first hand. Primary data has not been published yet and is more reliable, authentic and objective.

For example population census conducted by the government of India after every 10 years.

3.3.1.1. COLLECTION OF PRIMARY DATA

Primary data is collected in the course of doing experimental or descriptive research by doing experiments, performing surveys or by observation or direct communication with respondents. Several methods for collecting primary data are given below-

I. Observation Method It is commonly used in studies relating to behavioural science. Under this method observation becomes scientific tool and the method of data collection for the researcher, when it serves a formulated research purpose and is systematically planned and subjected to checks and controls.

(a) Structured (descriptive) and unstructured (exploratory) observation- When a observation is characterized by careful definition of units to be observed, style of observer, conditions of or observation and selection of pertinent data of observation it is a structured observation. When there characteristics are not thought of in advance or not present. it is a unstructured observation.

(b) Participant, Non-participant and disguised observation- When the observer observes by making himself more or less, the member of the group he is observing, it is participant observation but when the observer observes by detaching himself from the group under observation it is non participant observation. If the observer observes in such manner that his presence is unknown to the people he is observing it is disguised observation.

(c) Controlled (laboratory) and uncontrolled(exploratory) observation- If the observation takes place in the natural setting it is a uncontrolled observation but when observation takes place according to some pre-arranged plans ,involving experimental procedure it is a controlled observation.

Advantages-

- Subjective bias is eliminated.
- Data is not affected by past behaviour or future intentions.
- Natural behaviour of the group can be recorded.

Limitations-

- Expensive methodology.
- Information provided is limited.
- Unforeseen factors may interfere with the observational task

II. INTERVIEW METHOD

This method of collecting data involves presentation of oral verbal stimuli and deeply in terms of oral- verbal responses. It can be achieved by two ways:-

(A)Personal interview- It requires a person known as interviewer to ask questions generally in a face to face contact to the other person.

It can be –

Direct personal investigation- The interviewer has to collect the information personally from the services concerned.

Indirect oral examination- The interviewer has to cross examine other persons who are suppose to have a knowledge about the problem.

Structured interviews- Interviews involving the use of pre-determined questions and of highly standard techniques of recording

Unstructured interviews- It does not follow a system of pre-determined questions and is characteirzsed by flexibility of approach to questioning.

Focussed interview- It is meant to focus attention on the given experience of the respondent and its effect. The interviewer may ask questions in any manner or sequence with the aim to explore reasons and motives of the respondent.

Clinical interviews- It is concerned with broad underlying feeling and motives or individuals life experience which are used as method to collect information under this method at the interviewer direction.

Non directive interview- The interviewer`s function is to encourage the respendent to talk about the given topic with a bare minimum of direct questioning.

Advantages-

- More information and in depth can be obtained.
- Samples can be controlled.
- There is greater flexibility under this method
- Personal information can as well be obtained.
- Mis-interpretation can be avoided by unstructured interview.

Limitations

- It is an expensive method.
- More time consuming.
- Possibility of imaginary info and less frank responses.
- High skilled interviewer is required

(B) Telephonic interviews- It requires the interviewer to collect information by contacting respondents on telephone and asking questions or opinions orally.

III. QUESTIONNAIRE

In this method a questionnaire is sent (mailed) to the concerned respondents who are expected to read, understand and reply on their own and return the questionnaire. It consists of a number of questions printed or typed in a definite order on a form or set of forms. It is advisable to conduct a 'pilot study' which is the rehearsal of the main survey by experts for testing the questionnaire for weaknesses of the questions and techniques used. Essential of a good questionnaire-

- It should be short and simple.
- Questions should processed in a logical sequence.
- Technical terms and vauge expressions must be avoided.
- Control questions to check the reliability of the respondent must be present.
- Adequate space for answers must be provided.
- Brief directions with regard to filling up of questionnaire must be provided.
- The physical appearances-quality of paper, colour etc must be good to attract the attention of the respondent

Advantages

- Free from bias of interviewer.
- Respondents have adequate time to give answers
- Respondents are easily and conveniently approachable
- Large samples can be used to be more reliable.

LIMITATIONS

- Low rate of return of duly filled questionnaire.
- Control over questions is lost once it is sent.
- It is inflexible once it is sent.
- Possiblitty of ambiguous omission of replies.
- Time taking and slow process.

IV. SCHEDULES

This method of data collection is similar to questionnaire method with difference that schedule are being filled by the enumerations specially appointed for the purpose. Enumerations explain the aims and objects of the investigation and may remove any misunderstanding and help the respondents to record answer. Enumerations should be well trained to perform their job,he/she should be honest hardworking and patient. This type of data is helpful in extensive enquiries however it is very expensive

3.2. Secondary data:

It refers to the data collected by someone other than the user i.e. the data is already available and analysed by someone else. Common sources of secondary data include various published or unpublished data, books, magazines, newspaper, trade journals etc.

Collection of secondary data

A researcher can obtain secondary data from various sources.Secondary data may either

be published data or unpublished data. Published data are available in:

a. Publications of government. b. Technical and trade journals. c. Reports of various businesses, banks etc. d. Public records. e. Stastistical or historical documents. Unpublished data may be found in letters, diaries, unpublished biographies or work.

Before using secondary data it must be checked for the following characteristics- 1. Reliability of data- Who collected the data? From what source? Which method? Time?

Possibility of bias? Accuracy?

2. Suitability of data- The object scope and nature of the original enquiry must be studies and then carefully scrutinize the data for suitability.

3. Adequency- The data is considered inadequate if the level of accuracy achieved in data is found inadequate or if they are related to an area which may be either narrower or wider than the area of the present enquiry

Advantages of using Secondary data

- The data is already there- no hassles of data collection
- It is less expensive
- The investigator is not personally responsible for the quality of data

Disadvantages of using Secondary data

- The investigator cannot decide what is collected (if specific data about something is required, for instance).
- One can only hope that the data is of good quality
- Obtaining additional data (or even clarification) about something is not possible (most often)

Primary Data vs Secondary Data

Character	Primary Data	Secondary Data
Definition	Primary data refers to the first-hand data gathered by the researcher himself.	Secondary data means data collected by someone else earlier.
Data	Real time data	Past Data
Process	Very Involved	Quick and easy
Source	Surveys, observations, experiments, questionnaire, personal interview, etc.	Government publications, websites, books, journal articles, internal records etc.
Cost-effectiveness	Expensive	Economical
Collection time	Long	Short
Specificity	Always specific to the researcher's needs.	May or may not be specific to the researcher's need.
Form	Available in the crude form	Available in the refined form
Accuracy and Reliability	More	Less

4. Introduction to sampling methods

When you conduct research about a group of people, it's rarely possible to collect data from every person in that group. Instead, you select a sample. The sample is the group of individuals who will actually participate in the research.

To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. There are two types of sampling methods:

- Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.
- Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

You should clearly explain how you selected your sample in the methodology section of your paper or thesis.

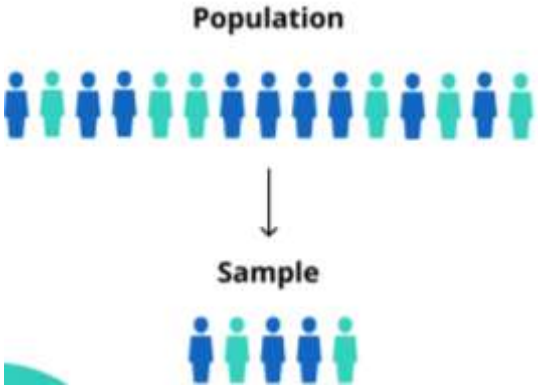
Population vs sample

First, you need to understand the difference between a population and a sample, and identify the target population of your research.

- The population is the entire group that you want to draw conclusions about.

- The sample is the specific group of individuals that you will collect data from.

The population can be defined in terms of geographical location, age, income, and many other characteristics.



It can be very broad or quite narrow: maybe you want to make inferences about the whole adult population of your country; maybe your research focuses on customers of a certain company, patients with a specific health condition, or students in a single school.

It is important to carefully define your target population according to the purpose and practicalities of your project.

If the population is very large, demographically mixed, and geographically dispersed, it might be difficult to gain access to a representative sample.

Sampling frame

The sampling frame is the actual list of individuals that the sample will be drawn from. Ideally, it should include the entire target population (and nobody who is not part of that population).

Example

You are doing research on working conditions at Company X. Your population is all 1000 employees of the company. Your sampling frame is the company's HR database which lists the names and contact details of every employee.

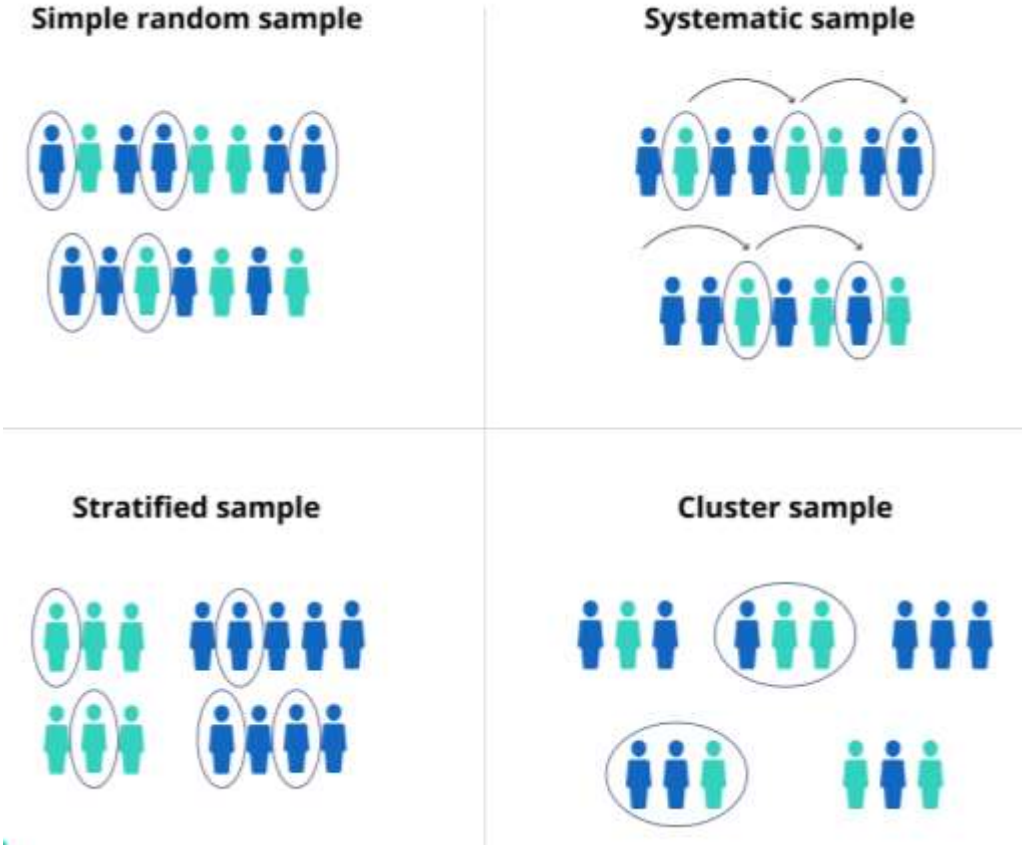
Sample size

The number of individuals you should include in your sample depends on various factors, including the size and variability of the population and your research design. There are different sample size calculators and formulas depending on what you want to achieve with statistical analysis.

Probability sampling methods

Probability sampling means that every member of the population has a chance of being selected. It is mainly used in quantitative research. If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice.

There are four main types of probability sample.



1. Simple random sampling

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

Example

You want to select a simple random sample of 100 employees of Company X. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

2. Systematic sampling

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Example

All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team

members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

3. Stratified sampling

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g. gender, age range, income bracket, job role).

Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

Example

The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

4. Cluster sampling

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above.

This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It's difficult to guarantee that the sampled clusters are really representative of the whole population.

Example

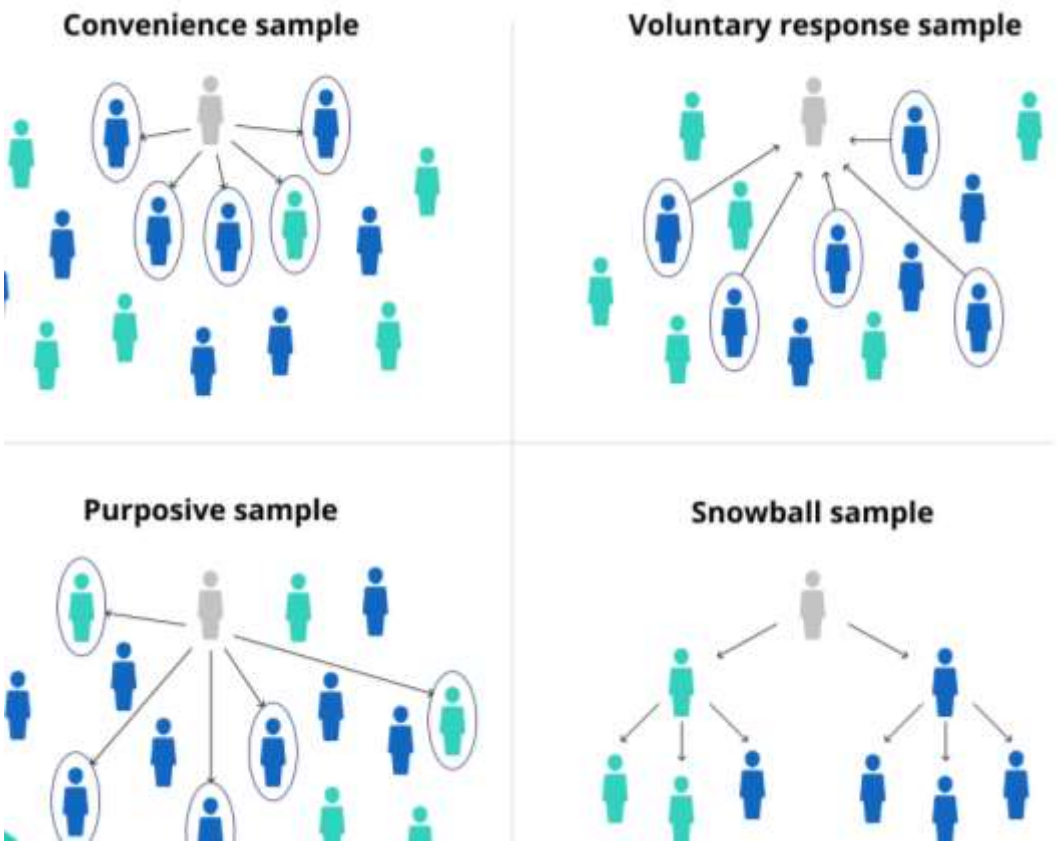
The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don't have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.

Non-probability sampling methods

In a non-probability sample, individuals are selected based on non-random criteria, and not every individual has a chance of being included.

This type of sample is easier and cheaper to access, but it has a higher risk of sampling bias. That means the inferences you can make about the population are weaker than with probability samples, and your conclusions may be more limited. If you use a non-probability sample, you should still aim to make it as representative of the population as possible.

Non-probability sampling techniques are often used in exploratory and qualitative research. In these types of research, the aim is not to test a hypothesis about a broad population, but to develop an initial understanding of a small or under-researched population.



1. Convenience sampling

A convenience sample simply includes the individuals who happen to be most accessible to the researcher.

This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can't produce generalizable results.

Example

You are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a survey on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.

2. Voluntary response sampling

Similar to a convenience sample, a voluntary response sample is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g. by responding to a public online survey).

Voluntary response samples are always at least somewhat biased, as some people will inherently be more likely to volunteer than others.

Example

You send out the survey to all students at your university and a lot of students decide to complete it. This can certainly give you some insight into the topic, but the people who responded are more likely to be those who have strong opinions about the student support services, so you can't be sure that their opinions are representative of all students.

3. Purposive sampling

This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.

It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion.

Example

You want to know more about the opinions and experiences of disabled students at your university, so you purposefully select a number of students with different support needs in order to gather a varied range of data on their experiences with student services.

4. Snowball sampling

If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to "snowballs" as you get in contact with more people.

Example

You are researching experiences of homelessness in your city. Since there is no list of all homeless people in the city, probability sampling isn't possible. You meet one person who agrees to participate in the research, and she puts you in contact with other homeless people that she knows in the area.

UNIT – III -ANALYTICAL METHODS AND SPECTROSCOPY– SBT1002

3.BIOPHYSICAL TECHNIQUES

The characterization of molecular structure, the measurement of molecular properties, and the observation of molecular behavior presents an enormous challenge for biological scientists. A wide range of biophysical techniques have been developed to study molecules in crystals, in solution, in cells, and in organisms. These biophysical techniques provide information about the electronic structure, size, shape, dynamics, polarity, and modes of interaction of biological molecules. Some of the most exciting techniques provide images of cells, subcellular structures, and even individual molecules. It is now possible, for example, to directly observe the biological behavior and physical properties of single protein or DNA molecules within a living cell and determine how the behavior of the single molecule influences the biological function of the organism.

Much biophysical research involves either the development of novel techniques to investigate the structure, properties, and biological functions of biomolecules or the application of these techniques to monitor how the structure and dynamics of biomolecules enables specific biological functions. Information about specific biophysical techniques is provided here

3.1. X-ray diffraction

X-ray diffraction analysis (XRD) is a technique used in materials science to determine the crystallographic structure of a material. XRD works by irradiating a material with incident X-rays and then measuring the intensities and scattering angles of the X-rays that leave the material.A primary use of XRD analysis is the identification of materials based on their diffraction pattern. As well as phase identification, XRD also yields information on how the actual structure deviates from the ideal one, owing to internal stresses and defects.

Principle

Max von Laue, in 1912, discovered that crystalline substances act as three-dimensional diffraction gratings for X-ray wavelengths similar to the spacing of planes in a crystal lattice. X-ray diffraction is now a common technique for the study of crystal structures and atomic spacing.

X-ray diffraction is based on constructive interference of monochromatic X-rays and a crystalline sample. These X-rays are generated by a cathode ray tube, filtered to produce monochromatic radiation, collimated to concentrate, and directed toward the sample. The interaction of the incident rays with the sample produces constructive interference (and a diffracted ray) when conditions satisfy Bragg's Law ($n\lambda=2d \sin \theta$). This law relates the wavelength of electromagnetic radiation to the diffraction angle and the lattice spacing in a crystalline sample. These diffracted X-rays are then detected, processed and counted. By scanning the sample through a range of 2θangles, all possible diffraction directions of the lattice should be attained due to the random orientation of the powdered material. Conversion of the diffraction peaks to d-spacings allows identification of the mineral because each mineral has a set of unique d-spacings. Typically, this is achieved by comparison of d-spacings with standard reference patterns. All diffraction methods are based on the generation of X-rays in an X-ray tube. These X-rays are directed at the sample, and the diffracted rays are collected.

A key component of all diffraction is the angle between the incident and diffracted rays. Powder and single crystal diffraction vary in instrumentation beyond this.

X-ray diffraction

X-rays for chemical analysis are commonly obtained by rotating anode generators (in-house) or synchrotron facilities (Fig. 2). In rotating anode generators, a rotating metal target is bombarded with high-energy (10–100 keV) electrons that knock out core electrons. An electron in an outer shell fills the hole in the inner shell and emits the energy difference between the two states as an X-ray

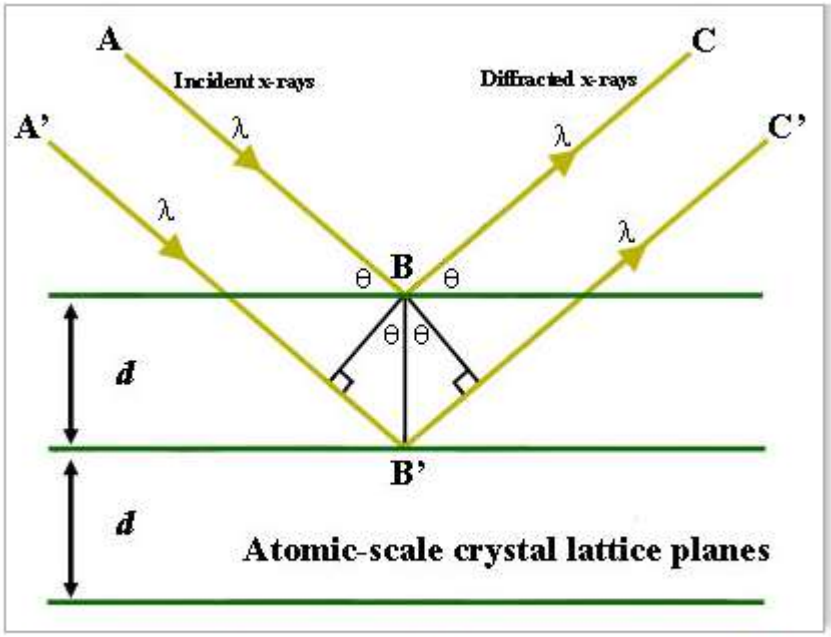


Figure 1. Bragg's Law reflection. The diffracted X-rays exhibit constructive interference when the distance between paths ABC and A'B'C' differs by an integer number of wavelengths (λ).

photon. Common targets are copper, molybdenum and chromium, which have strong distinct X-ray emission at 1.54 A˚ , 0.71 A˚ and 2.29 A˚ , respectively, that is superimposed on a continuous spectrum known as Bremsstrahlung. In synchrotrons, electrons are accelerated in a ring, thus producing a continuous spectrum of X-rays. Monochromators are required to select a single wavelength. As X-rays are diffracted by electrons, the analysis of X-ray diffraction data sets produces an electron density map of the crystal. Since hydrogen atoms have very little electron density, they are not usually determined experimentally by this technique. Unfortunately, the detection of light beams is restricted to recording the intensity of the beam only. Other properties, such as polarization, can only be determined with rather complex measurements. The phase of the light waves is even systematically lost in the measurement. This phenomenon has thus been termed the phase problem owing to the essential information contained in the phase in diffraction and microscopy experiments. The X-ray diffraction data can be used to calculate the amplitudes of the three-dimensional Fourier transform of the electron density. Only together with the phases can the electron density be calculated, in a process called Fourier synthesis. Different methods to overcome the phase problem in X-ray crystallography have been developed, including:

- Molecular replacement, where phases from a structurally similar molecule are used;
- Experimental methods that require incorporation of heavy element salts (multiple isomorphous replacement);
- Experimental methods where methionine has been replaced by seleno-methionine in proteins (multi-wavelength anomalous diffraction);
- Experimental methods using the anomalous diffraction of the intrinsic sulphur in proteins (single wavelength anomalous diffraction);
- Ddirect methods, where a statistical approach is used to determine phases. This approach is limited to very high resolution data sets and is the main method for small molecule crystals as these provide high-quality diffraction with relatively few numbers of reflections.

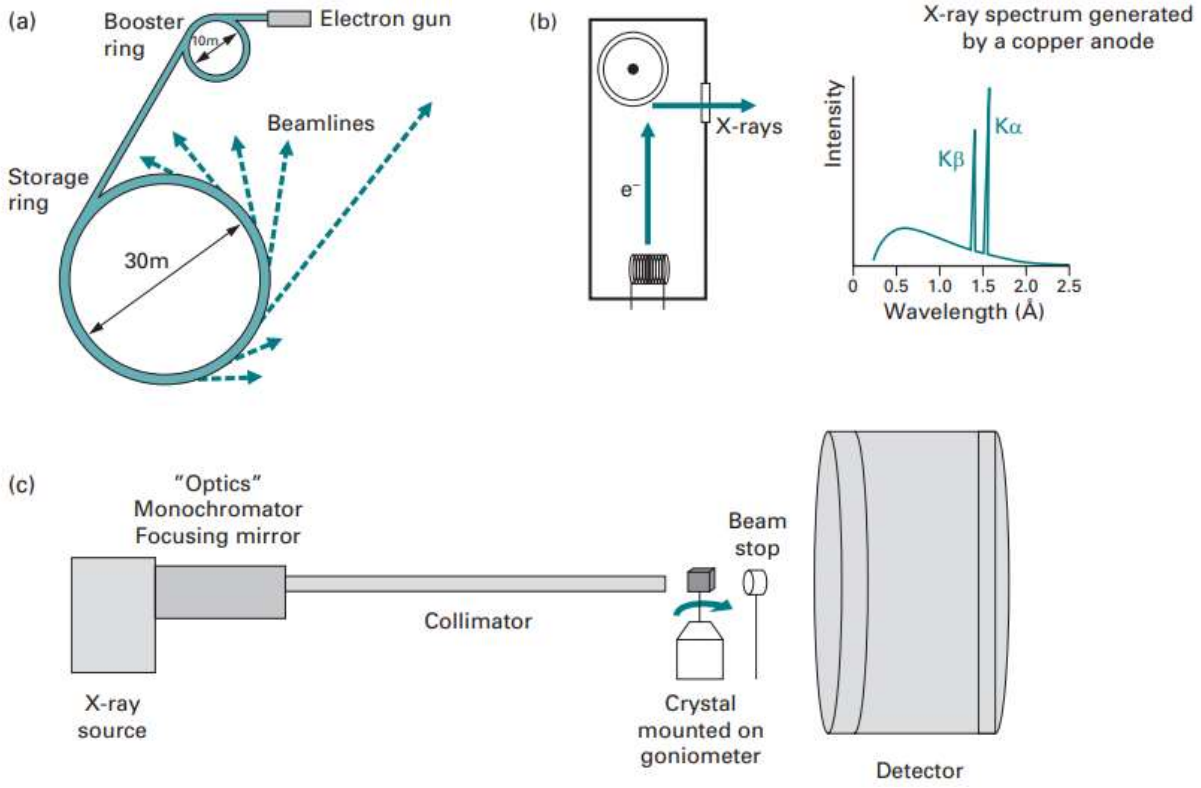


Fig.2 Instrumentation for X-ray diffraction. The most common X-ray sources are (a) particle storage rings which produce synchrotron radiation, and (b) rotating anode tubes. The schematics of an X-ray diffractometer are shown in (c).

Applications

Single-crystal diffraction A crystal is a solid in which atoms or molecules are packed in a particular arrangement within the unit cell which is repeated indefinitely along three principal directions in space. Crystals can be formed by a wide variety of materials, such as salts, metals, minerals and semiconductors, as well as various inorganic, organic and biological molecules. A crystal grown in the laboratory is mounted on a goniometer and exposed to X-rays produced by rotating anode generators (in-house) or a synchrotron facility. A diffraction pattern of regularly spaced spots known

as reflections is recorded on a detector, most frequently image plates or CCD cameras for proteins, and moveable proportional counters for small molecules. An incident X-ray beam is diffracted by a crystal such that beams at specific angles are produced, depending on the X-ray wavelength, the crystal orientation and the structure of the crystal (i.e. unit cell). To record a data set, the crystal is gradually rotated and a diffraction pattern is acquired for each distinct orientation. These two-dimensional images are then analysed by identifying the appropriate reflection for each lattice plane and measuring its intensity, measuring the cell parameters of the unit cell and determining the appropriate space group. If information about the phases is available, this data can then be used to calculate a three-dimensional model of the electron density within the unit cell using the mathematical method of Fourier synthesis. The positions of the atomic nuclei are then deduced from the electron density by computational refinement and manual intervention using molecular graphics.

Fibre diffraction: Certain biological macromolecules, such as DNA and cytoskeletal components, cannot be crystallised, but form fibres. In fibres, the axes of the long polymeric structures are parallel to each other. While this can be an intrinsic property, for example in muscle fibres, in some cases the parallel alignment needs to be induced. As fibres show helical symmetry, by analysing the diffraction from oriented fibres one can deduce the helical symmetry of the molecule, and in favourable cases the molecular structure. Generally, a model of the fibre is constructed and the expected diffraction pattern is compared with the observed diffraction. Historically, fibre diffraction was of central significance in enabling the determination of the three-dimensional structure of DNA by Crick, Franklin, Watson and Wilkins. Two classes of fibre diffraction patterns can be distinguished. In crystalline fibres (e.g. A form of DNA), the long fibrous molecules pack to form thin micro-crystals randomly arranged around a shared common axis. The resulting diffraction pattern is equivalent to taking a long crystal and spinning it about its axis during the X-ray exposure. All Bragg reflections are recorded at once. In non-crystalline fibres (e.g. B form of DNA), the molecules are arranged parallel to each other but in a random orientation around the common axis. The reflections in the diffraction pattern are now a result of the periodic repeat of the fibrous molecule. The diffraction intensity can be calculated via Fourier–Bessel transformation replacing the Fourier transformation used in single-crystal diffraction.

Powder diffraction: Powder diffraction is a rapid method to analyse multicomponent mixtures without the need for extensive sample preparation. Instead of using single crystals, the solid material is analysed in the form of a powder where, ideally, all possible crystalline orientations are equally represented. From powder diffraction patterns, the interplanar spacings d of the lattice planes are determined and then compared to a known standard or to a database (Powder Diffraction File by the International Centre for Diffraction Data or the Cambridge Structural Database) for identification of the individual components.

X-ray powder diffraction is most widely used for the identification of unknown crystalline materials (e.g. minerals, inorganic compounds). Determination of unknown solids is critical to studies in geology, environmental science, material science, engineering and biology.

Other applications include:

- characterization of crystalline materials
- identification of fine-grained minerals such as clays and mixed layer clays that are difficult to determine optically
- determination of unit cell dimensions
- measurement of sample purity

With specialized techniques, XRD can be used to:

- determine crystal structures using Rietveld refinement
- determine of modal amounts of minerals (quantitative analysis)
- characterize thin films samples by:
    - determining lattice mismatch between film and substrate and to inferring stress and strain
    - determining dislocation density and quality of the film by rocking curve measurements
    - measuring superlattices in multilayered epitaxial structures
    - determining the thickness, roughness and density of the film using glancing incidence X-ray reflectivity measurements
- make textural measurements, such as the orientation of grains, in a polycrystalline sample.

3.2. FLUORESCENCE SPECTROSCOPY

Fluorescence is an emission phenomenon where an energy transition from a higher to a lower state is accompanied by radiation. Only molecules in their excited forms are able to emit fluorescence; thus, they have to be brought into a state of higher energy prior to the emission phenomenon.

The molecules possess discrete states of energy. Potential energy levels of molecules have been depicted by different Lennard–Jones potential curves with overlaid vibrational (and

rotational) states. Such diagrams can be abstracted further to yield Jablonski diagrams (Figure 3).

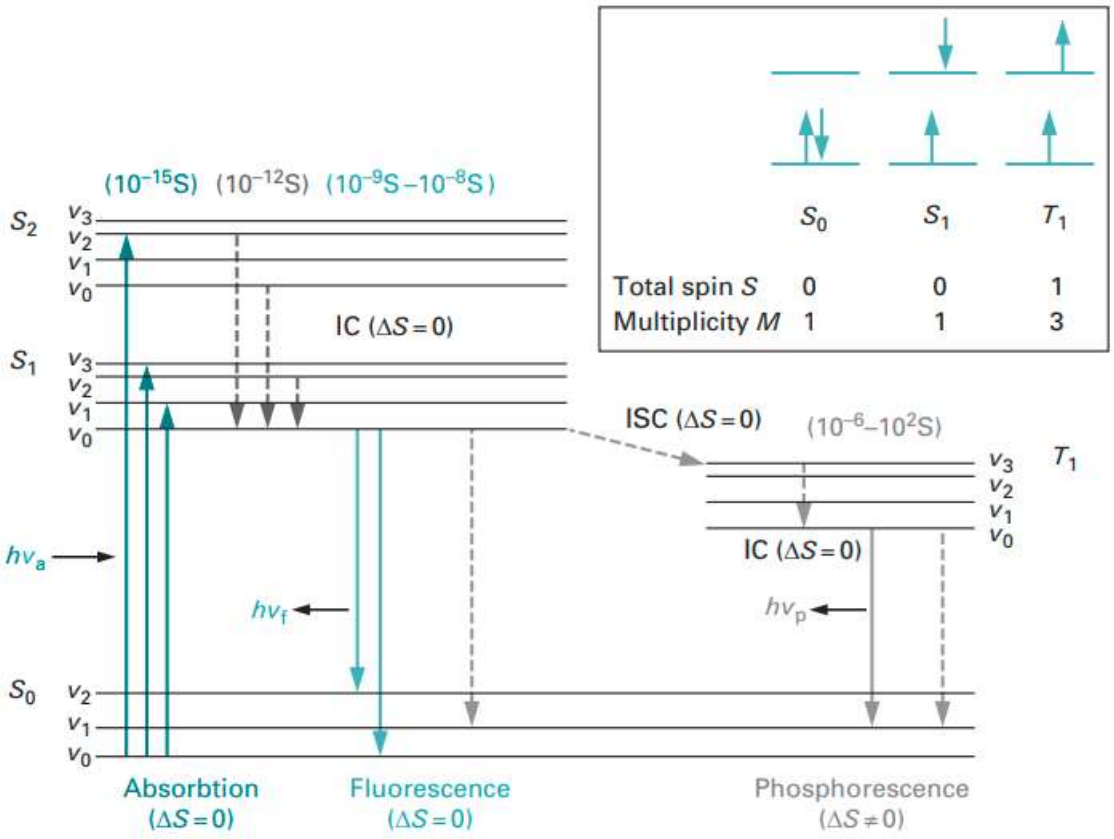


Fig. 3 Jablonski diagram. Shown are the electronic ground state (S0), two excited singlet states (S1, S2) and a triplet state (T1). Vibrational levels (v) are only illustrated exemplarily. Solid vertical lines indicate radiative transitions, dotted lines show non-radiative transitions. The inset shows the relationship between electron configurations, total spin number S and multiplicity M

In these diagrams, energy transitions are indicated by vertical lines. Not all transitions are possible; allowed transitions are defined by the selection rules of quantum mechanics. A molecule in its electronic and vibrational ground state (S0v0) can absorb photons matching the energy difference of its various discrete states. The required photon energy has to be higher than that required to reach the vibrational ground state of the first electronic excited state (S1v0). The excess energy is absorbed as vibrational energy (v > 0), and quickly dissipated as heat by collision with solvent molecules. The molecule thus returns to the vibrational ground state (S1v0). These relaxation processes are non-radiating transitions from one energetic state to another with lower energy, and are called internal conversion (IC). From the lowest level of the first electronic excited state, the molecule returns to the ground state (S0) either by emitting light (fluorescence) or by a non-radiative transition. Upon radiative

transition, the molecule can end up in any of the vibrational states of the electronic ground state (as per quantum mechanical rules).

If the vibrational levels of the ground state overlap with those of the electronic excited state, the molecule will not emit fluorescence, but rather revert to the ground state by non-radiative internal conversion. This is the most common way for excitation energy to be dissipated and is why fluorescent molecules are rather rare. Most molecules are flexible and thus have very high vibrational levels in the ground state. Indeed, most fluorescent molecules possess fairly rigid aromatic rings or ring systems. The fluorescent group in a molecule is called a fluorophore.

Since radiative energy is lost in fluorescence as compared to the absorption, the fluorescent light is always at a longer wavelength than the exciting light (Stokes shift). The emitted radiation appears as band spectrum, because there are many closely related wavelength values dependent on the vibrational and rotational energy levels attained. The fluorescence spectrum of a molecule is independent of the wavelength of the exciting radiation and has a mirror image relationship with the absorption spectrum. The probability of the transition from the electronic excited to the ground state is proportional to the intensity of the emitted light.

An associated phenomenon in this context is phosphorescence which arises from a transition from a triplet state (T1) to the electronic (singlet) ground state (S0). The molecule gets into the triplet state from an electronic excited singlet state by a process called intersystem crossing (ISC). The transition from singlet to triplet is quantum-mechanically not allowed and thus only happens with low probability in certain molecules where the electronic structure is favourable. Such molecules usually contain heavy atoms. The rate constants for phosphorescence are much longer and phosphorescence thus happens with a long delay and persists even when the exciting energy is no longer applied.

The fluorescence properties of a molecule are determined by properties of the molecule itself (internal factors), as well as the environment of the protein (external factors). The fluorescence intensity emitted by a molecule is dependent on the lifetime of the excited state. The transition from the excited to the ground state can be treated like a decay process of first order, i.e. the number of molecules in the excited state decreases exponentially with time. In analogy to kinetics, the exponential coefficient kr is called rate constant and is the reciprocal of the lifetime: $Tr = kr^{-1}$. The lifetime is the time it takes to reduce the number of fluorescence emitting molecules to N0/e, and is proportional to $\boldsymbol{\lambda}^3$.

The effective lifetime of excited molecules, however, differs from the fluorescence lifetime Tr since other, non-radiative processes also affect the number of molecules in the excited state T. is dependent on all processes that cause relaxation: fluorescence emission, internal conversion, quenching, fluorescence resonance energy transfer, reactions of the excited state and intersystem crossing. The ratio of photons emitted and photons absorbed by a fluorophore is called quantum yield Φ . It equals the ratio of the rate constant for fluorescence emission kr and the sum of the rate constants for all six processes mentioned above

$$\Phi = \frac{N(\mathrm{em})}{N(\mathrm{abs})} = \frac{k_\mathrm{r}}{k} = \frac{k_\mathrm{r}}{k_\mathrm{r} + k_\mathrm{IC} + k_\mathrm{ISC} + k_\mathrm{reaction} + k_Q\, c(Q) + k_\mathrm{FRET}} = \frac{\tau}{\tau_\mathrm{r}}$$

The quantum yield is a dimensionless quantity, and, most importantly, the only absolute measure of fluorescence of a molecule. Measuring the quantum yield is a difficult process and requires comparison with a fluorophore of known quantum yield. In biochemical applications, this measurement is rarely done. Most commonly, the fluorescence emissions of two or more related samples are compared and their relative differences analysed.

Instrumentation

Fluorescence spectroscopy works most accurately at very low concentrations of emitting fluorophores. UV/Vis spectroscopy, in contrast, is least accurate at such low concentrations. One major factor adding to the high sensitivity of fluorescence applications is the spectral selectivity. Due to the Stokes shift, the wavelength of the emitted light is different from that of the exciting light. Another feature makes use of the fact that fluorescence is emitted in all directions. By placing the detector perpendicular to the excitation pathway, the background of the incident beam is reduced.

The schematics of a typical spectrofluorimeter are shown in Fig. 12.9. Two monochromators are used, one for tuning the wavelength of the exciting beam and a second one for analysis of the fluorescence emission. Due to the emitted light always having a lower energy than the exciting light, the wavelength of the excitation monochromator is set at a lower wavelength than the emission monochromator. The better fluorescence spectrometers in laboratories have a photon-counting detector yielding very high sensitivity. Temperature control is required for accurate work as the emission intensity of a fluorophore is dependent on the temperature of the solution. Two geometries are possible for the measurement, with the 90^0 arrangement most commonly used. Pre- and post-filter effects can arise owing to absorption of light prior to reaching the fluorophore and the reduction of emitted radiation. These phenomena are also called inner filter effects and are more evident in solutions with high concentrations. As a rough guide, the absorption of a solution to be used for fluorescence experiments should be less than 0.05. The use of microcuvettes containing less material can also be useful. Alternatively, the front-face illumination geometry (Fig. 4) can be used which obviates the inner filter effect. Also, while the 90^0 geometry requires cuvettes with two neighbouring faces being clear (usually, fluorescence cuvettes have four clear faces), the front-face illumination technique requires only one clear face, as excitation and emission occur at the same face. However, front-face illumination is less sensitive than the 90 illumination.

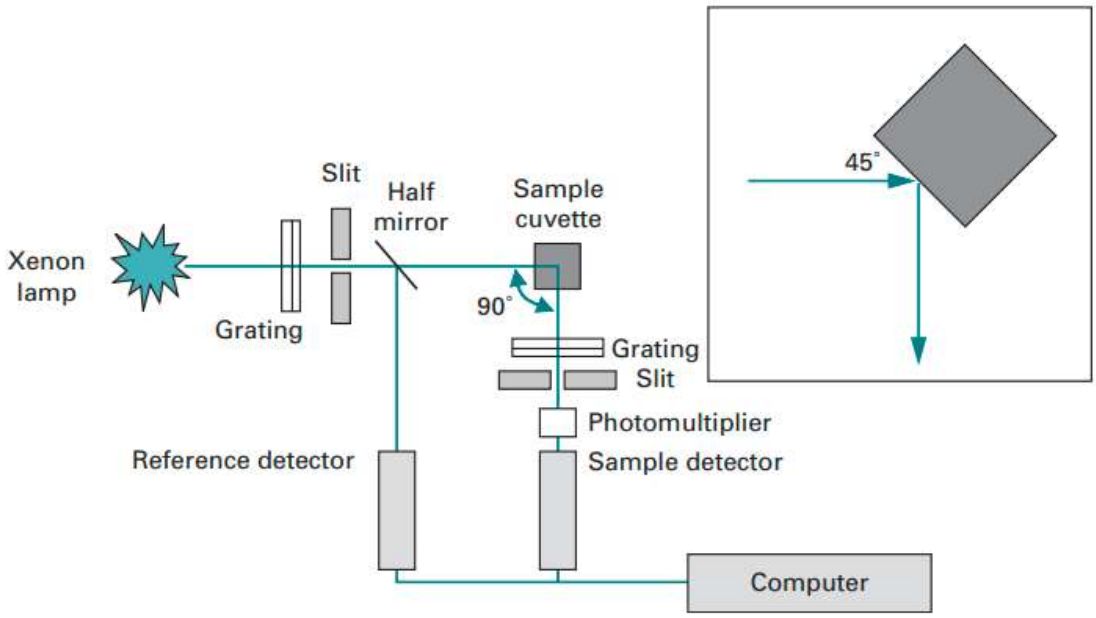


Fig. 4 Schematics of a spectrofluorimeter with 'T' geometry (90). Optical paths are shown as green lines. Inset: Geometry of front-face illumination.

Applications

There are many and highly varied applications for fluorescence despite the fact that relatively few compounds exhibit the phenomenon. The effects of pH, solvent composition and the polarisation of fluorescence may all contribute to structural elucidation. Measurement of fluorescence lifetimes can be used to assess rotation correlation coefficients and thus particle sizes. Non-fluorescent compounds are often labelled with fluorescent probes to enable monitoring of molecular events. This is termed extrinsic fluorescence as distinct from intrinsic fluorescence where the native compound exhibits the property. Some fluorescent dyes are sensitive to the presence of metal ions and can thus be used to track changes of these ions in in vitro samples, as well as whole cells. Since fluorescence spectrometers have two monochromators, one for tuning the excitation wavelength and one for analysing the emission wavelength of the fluorophore, one can measure two types of spectra: excitation and emission spectra. For fluorescence excitation spectrum measurement, one sets the emission monochromator at a fixed wavelength (lem) and scans a range of excitation wavelengths which are then recorded as ordinate (x-coordinate) of the excitation spectrum; the fluorescence emission at lem is plotted as abscissa. Measurement of emission spectra is achieved by setting a fixed excitation wavelength (lexc) and scanning a wavelength range with the emission monochromator. To yield a spectrum, the emission wavelength lem is recorded as ordinate and the emission intensity at lem is plotted as abscissa.

Intrinsic protein fluorescence

Proteins possess three intrinsic fluorophores: tryptophan, tyrosine and phenylalanine, although the latter has a very low quantum yield and its contribution to protein fluorescence emission

is thus negligible. Of the remaining two residues, tyrosine has the lower quantum yield and its fluorescence emission is almost entirely quenched when it becomes ionised, or is located near an amino or carboxyl group, or a tryptophan residue. Intrinsic protein fluorescence is thus usually determined by tryptophan fluorescence which can be selectively excited at 295–305 nm. Excitation at 280 nm excites tyrosine and tryptophan fluorescence and the resulting spectra might therefore contain contributions from both types of residues. The main application for intrinsic protein fluorescence aims at conformational monitoring. We have already mentioned that the fluorescence properties of a fluorophore depend significantly on environmental factors, including solvent, pH, possible quenchers, neighbouring groups, etc. A number of empirical rules can be applied to interpret protein fluorescence spectra:

- As a fluorophore moves into an environment with less polarity, its emission spectrum exhibits a hypsochromic shift (lmax moves to shorter wavelengths) and the intensity at lmax increases.
- Fluorophores in a polar environment show a decrease in quantum yield with increasing temperature. In a non-polar environment, there is little change.
- Tryptophan fluorescence is quenched by neighbouring protonated acidic groups.

When interpreting effects observed in fluorescence experiments, one has to consider carefully all possible molecular events. For example, a compound added to a protein solution can cause quenching of tryptophan fluorescence. This could come about by binding of the compound at a site close to the tryptophan (i.e. the residue is surfaceexposed to a certain degree), or due to a conformational change induced by the compound. The comparison of protein fluorescence excitation and emission spectra can yield insights into the location of fluorophores. The close spatial arrangement of fluorophores within a protein can lead to quenching of fluorescence emission; this might be seen by the lower intensity of the emission spectrum when compared to the excitation spectrum (Fig. 5).

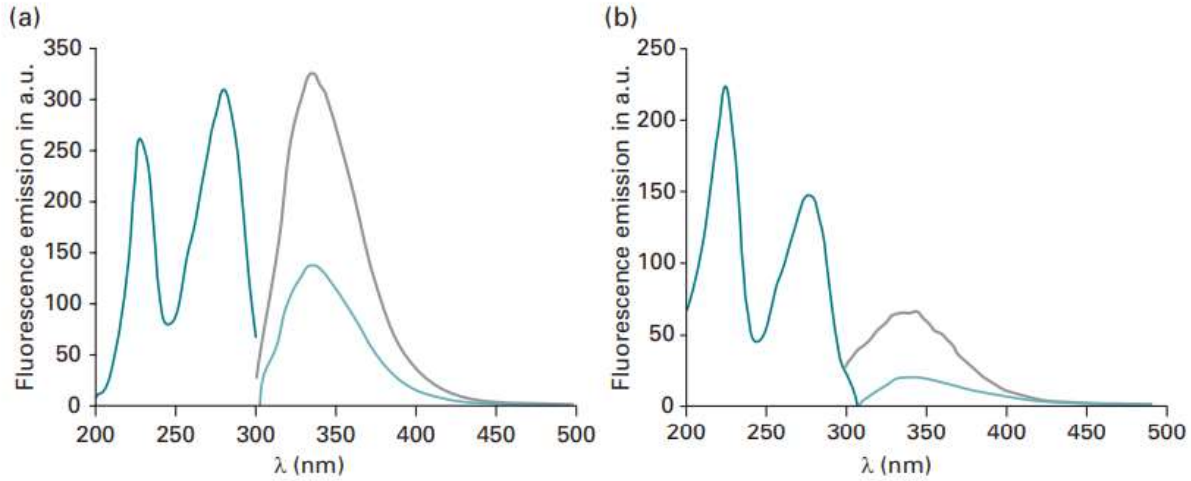


Fig. 5 Comparison of fluorescence excitation and emission spectra can yield insights into internal quenching. Excitation spectra with emission wavelength 340 nm are shown in dark green. Emission spectra with excitation wavelength 295 nm are shown in light green; emission spectra with excitation wavelength 280 nm are grey. (a) PDase homologue (Escherichia coli). (b) CPDase (Arabidopsis thaliana); in this protein, the fluorophores are located in close proximity to each other which leads to the effect of intrinsic quenching, as

obvious from the lower intensity of the emission spectrum as compared to the excitation spectrum.

Extrinsic fluorescence

Frequently, molecules of interest for biochemical studies are non-fluorescent. In many of these cases, an external fluorophore can be introduced into the system by chemical coupling or non-covalent binding. Some examples of commonly used external fluorophores are shown in Fig. 6. Three criteria must be met by fluorophores in this context. Firstly, it must not affect the mechanistic properties of the system under investigation. Secondly, its fluorescence emission needs to be sensitive to environmental conditions in order to enable monitoring of the molecular events. And lastly, the fluorophore must be tightly bound at a unique location. A common non-conjugating extrinsic chromophore for proteins is 1-anilino-8- naphthalene sulphonate (ANS) which emits only weak fluorescence in polar environment, i.e. in aqueous solution. However, in non-polar environment, e.g. when bound to hydrophobic patches on proteins, its fluorescence emission is significantly increased and the spectrum shows a hypsochromic shift; lmax shifts from 475 nm to 450 nm. ANSis thus a valuable tool for assessment of the degree of non-polarity. It can also be used in competition assays to monitor binding of ligands and prosthetic groups.

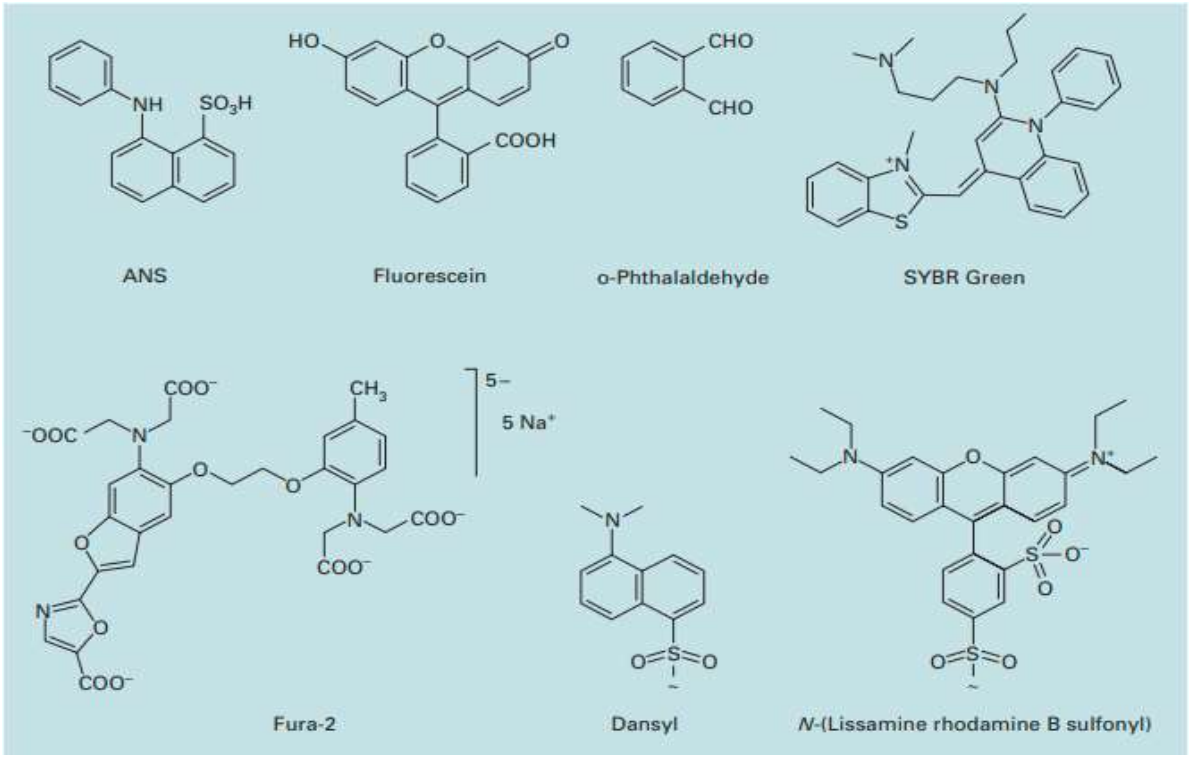


Fig. 6: Structures of some extrinsic fluorophores. Fura-2 is a fluorescent chelator for divalent and higher valent metal ions (Ca2þ, Ba2þ, Sr2þ, Pb2þ, La3þ, Mn2þ, Ni2þ, Cd2þ).

Reagents such as fluorescamine, o-phthalaldehyde or 6-aminoquinolyl-N-hydroxysuccinimidyl carbamate have been very popular conjugating agents used to derivatise amino acids for analysis (see Section 8.4.2). o-Phthalaldehyde, for example, is a non-fluorescent compound that reacts with primary amines and b-mercaptoethanol to yield a highly sensitive fluorophore. Metal-chelating compounds with fluorescent properties are useful tools for a variety of assays, including monitoring of metal homeostasis in cells. Widely used probes for calcium are the chelators Fura-2, Indo-1 and Quin-1. Since the chemistry of such compounds is based on metal chelation, cross-reactivity of the probes with other metal ions is possible. The intrinsic fluorescence of nucleic acids is very weak and the required excitation wavelengths are too far in the UV region to be useful for practical applications. Numerous extrinsic fluorescent probes spontaneously bind to DNA and display enhanced emission. While in earlier days ethidium bromide was one of the most widely used dyes for this application, it has nowadays been between the base pairs of double-helical DNA. Their fluorescence emission in water is very weak and increases about 30-fold upon binding to DNA.replaced by SYBR Green, as the latter probe poses fewer hazards for health and environment and has no teratogenic properties like ethidium bromide. These probes bind DNA by intercalation of the planar aromatic ring systems.

Quenching

The quantum yield of a fluorophore is dependent on several internal and external factors. One of the external factors with practical implications is the presence of a quencher. A quencher molecule decreases the quantum yield of a fluorophore by non-radiating processes. The absorption (excitation) process of the fluorophore is not altered by the presence of a quencher. However, the energy of the excited state is transferred onto the quenching molecules. Two kinds of quenching processes can be distinguished:

- Dynamic quenching which occurs by collision between the fluorophore in its excited state and the quencher; and
- Static quenching whereby the quencher forms a complex with the fluorophore. The complex has a different electronic structure compared to the fluorophore alone and returns from the excited state to the ground state by non-radiating processes.

Fluorescence microscopy, high-throughput assays

Fluorescence emission as a means of monitoring is a valuable tool for many biological and biochemical applications. We have already seen the usage of fluorescence monitoring in DNA sequencing; the technique is inseparably tied in with the success of projects such as genome deciphering. Fluorescence techniques are also indispensable methods for cell biological applications with fluorescence microscopy (see Sections 4.6 and 17.3.2). Proteins (or biological macromolecules) of interest can be tagged with a fluorescent label such as e.g. the green fluorescent protein (GFP) from the jelly fish Aequorea victoria or the red fluorescent protein from Discosoma striata, if spatial and temporal tracking of the tagged protein is desired. Alternatively, the use of GFP spectral variants such as cyan fluorescent protein (CFP) as a fluorescence donor and yellow fluorescent protein (YFP) as an acceptor allows investigation of mechanistic questions by using the FRET phenomenon. Specimens with cells expressing the labelled proteins are illuminated with light of the excitation

wavelength, and then observed through a filter that excludes the exciting light and only transmits the fluorescence emission. The recorded fluorescence emission can be overlaid with a visual image computationally, and the composite image then allows for localisation of the labelled species. If different fluorescence labels with distinct emission wavelengths are used simultaneously, even co-localisation studies can be performed.

Time-resolved fluorescence spectroscopy

The emission of a single photon from a fluorophore follows a probability distribution. With time-correlated single photon counting, the number of emitted photons can be recorded in a time-dependent manner following a pulsed excitation of the sample. By sampling the photon emission for a large number of excitations, the probability distribution can be constructed. The time-dependent decay of an individual fluorophore species follows an exponential distribution, and the time constant is thus termed the lifetime of this fluorophore. Curve fitting of fluorescence decays enables the identification of the number of species of fluorophores (within certain limits), and the calculation of the lifetimes for these species. In this context, different species can be different fluorophores or distinct conformations of the same fluorophore.

3.3. UV-Vis Spectroscopy

Spectroscopy

Spectroscopy involves investigating the interaction of electromagnetic field with matter. Historically, it originated as the study of wavelength based dispersion of visible light by a prism. The idea was expanded to include all the interactions involving radiative energy variation with wavelength or frequency. Spectroscopic data is represented as an emission spectrum of wavelength or frequency dependent response. Spectroscopy can be broadly classified into two categories – (a) techniques based on energy transfer between photon and sample, and (b) reflections, refraction, diffraction, dispersion, or scattering from the sample altering the amplitude, phase angle, polarization, or direction of propagation of the electromagnetic radiation.

Over the years, spectroscopy has evolved as a potential tool for experiments and analyses conducted in research laboratories and industries. This technique is essentially considered by analysts as an apparent solution. The objective should also be to use these spectroscopic techniques in control and industrial laboratories and to develop fully recognised spectroscopic techniques. This chapter reviews the interactions of ultraviolet, visible, and infrared radiations with matter. Irrespective of differences in the instrumentation, all spectroscopic techniques have many common attributes. The similarities as well as differences between various spectroscopic techniques have been outlined.

1. Principles and instrumentation for UV-Vis-IR

Ultraviolet (UV) spectroscopy is an important physical tool which exploits light in ultraviolet, visible, and near infrared range of electromagnetic spectrum. Beer-Lambert law establishes a linear relationship between absorbance, concentration of absorbers (or absorbing species) in the solution and the path length. Therefore, UV-Vis spectroscopy can be employed for determining the concentration of the absorbing species, for a fixed path length [1]. This is a very simple, versatile, fast, accurate and costeffective technique. Instrument employed for ultraviolet−visible (or UV-Vis) spectroscopy is called UV−Vis−NIR Spectrophotometer. This can be used to analyze liquids, gases and solids by using radiative energy corresponding to far and near ultraviolet (UV), visible (Vis) and near infrared (NIR) regions of electromagnetic spectrum. Consequently, predetermined wavelengths in these regions have been defined as: UV: 300 - 400 nm; Vis: 400 - 765 nm; and NIR: 765 - 3200 nm.

Principle: A light beam is passed through an object and wavelength of the light reaching the detector is measured. The measured wavelength provides important information about chemical structure and number of molecules (present in intensity of the measured signal). Thus, both quantitative and qualitative information can be gathered. Information may be obtained as transmittance, absorbance or reflectance of radiation in 160 to 3500 nm wavelength range [2,3]. The absorption of incident energy promotes electrons to excited states or the anti-bonding orbitals. For this transfer to occur, photon energy must match the energy needed by electron to be promoted to next higher energy state. This process forms the basic operating principle of absorption spectroscopy. Potentially, there may be three types of ground state orbitals involved:

1. σ (bonding) molecular orbital
2. π (bonding) molecular orbital
3. n (non-bonding) atomic orbital Besides, the anti-bonding orbitals are:
    i. σ* (sigma star) orbital
    ii. π* (pi star) orbital

A transition involving excitation of an electron from s bonding orbital to σ anti-bonding orbital is called σ to σ* transition. Likewise, π to π* represents the excitation of an electron of a lone pair (non-bonding electron pair) to an antibonding π orbital. Electronic transitions occurring due to absorption of UV and visible light are:

σ to σ*;

n to σ*;

n to π*;

π to π*.

The transitions s to σ* and n to σ* involve higher energies and thus usually occur in far UV region or weakly in 180 to 240nm region. Thus, saturated groups do not show strong absorption in UV

region. Molecules with unsaturated centres undergo n to π* and π to π* transitions; these transitions involve lesser energies and thus occur at longer wavelengths than transitions to σ* anti-bonding orbitals.

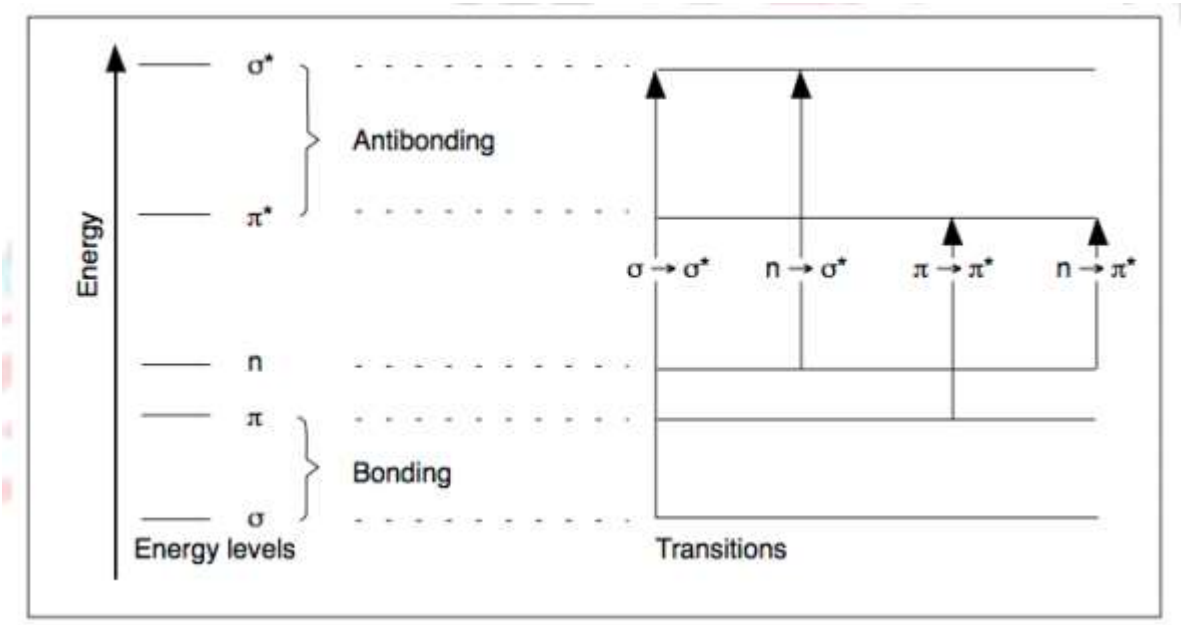


Figure 7: Electron transitions in UV-Vis spectroscopy.

When light having specific wavelength and energy is focused onto the sample, it absorbs some energy of the incident wave. A photodetector measures energy of transmitted light from sample, and registers absorbance of the sample. The absorption or transmission spectrum of the light absorbed or transmitted by the sample against the wavelength is formed. Bouguer−Beer law or the Lambert−Beer rule (Figure 8) is basic principle of quantitative analysis, and it establishes that absorbance of a solution scales directly with analyte concentration. For a given wavelength, the absorbance (unit less) A is described as the molar absorptivity of the absorbing species(M^{-1} cm^{-1}), b is path length ofsample holder (normally 1 cm), and c is the concentration of the solution (M).

$$A = a \cdot b \cdot c$$

UV−Vis−NIR spectrometer can monitor absorbance or transmittance in UV – visible wavelength range. The relation between incident light of intensity 'Io' and transmitted light of intensity 'I' is described as follows.

Transmittance (T) is given by I/I_o and $(I/I_o)*100$ gives transmission rate (T%). Absorbance (abs) is the inverse of transmittance and given by $\log(1/T) = log(I_o/I)$.

$$T = I/I_o = 10^{-kcl}$$

$$abs = \log (1/T) = \log(I_o/I) = - kcl$$

here, k represents constant of proportionality. Transmittance does not depend on sample concentration, whereas absorbance shows proportionality with concentration of sample (Beer's law) and optical path (Bouguer's law). Additionally, when optical path is 1 cm and concentration of targeted material is 1mol/l, k is described as molar absorption coefficient and denoted as 'ε'. Under specific conditions, molar absorption coefficient is typical of the material. Bouguer–Beer rule assumes the absence of any stray, generated, scattered, or reflected light.

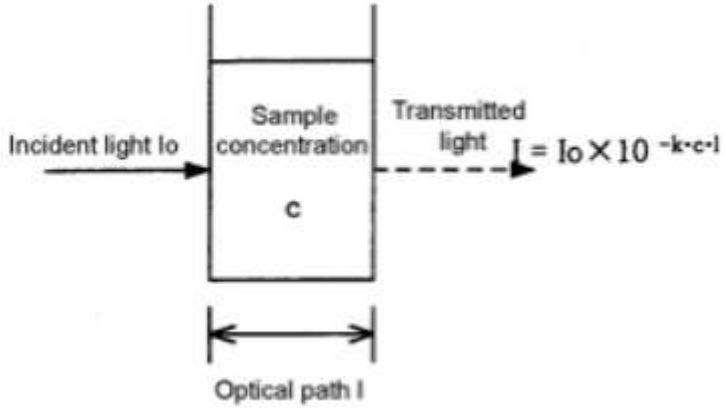


Figure 8 Bouguer– Beer Rule.

The UV-Vis spectrum can by recorded via the following types of absorbance instruments:

a. Single bram spectrometer

b. Double beam spectrometer

c. Simultaneous spectrometer

Light source (mostly tungsten lamps), small holder and detector are common to all the three type of spectrometers. However, a filter may be used, in addition, to select one wavelength at a time. This filter is often termed as the monochromator. Single beam spectrometer (shown in Figure 9) includes a monochromator between the light source and specimen. The specimen is analysed individually for all wavelengths. Double beam spectrometer (Figure 10) uses a single light source,

monochromator, a splitter and a series of mirrors, to direct the beam towards the reference and the sample under investigation. Whereas, a simultaneous spectrometer (Figure 11) uses an array of diodes for simultaneous detection of absorbance at all wavelengths. This is the fastest and most efficient of the three.

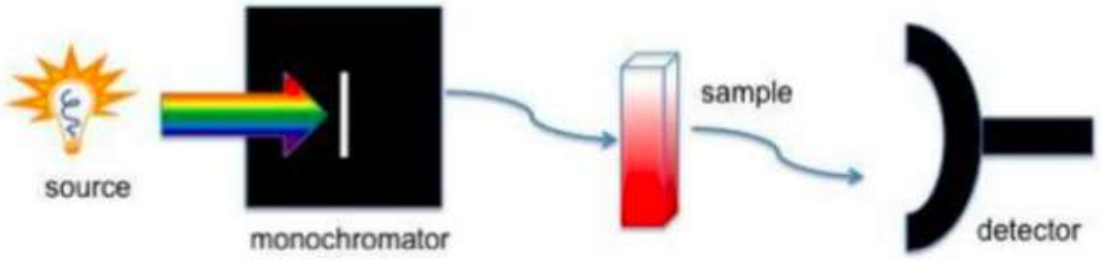


Figure 9: Schematics of a single beam UV-Visible spectrometer.

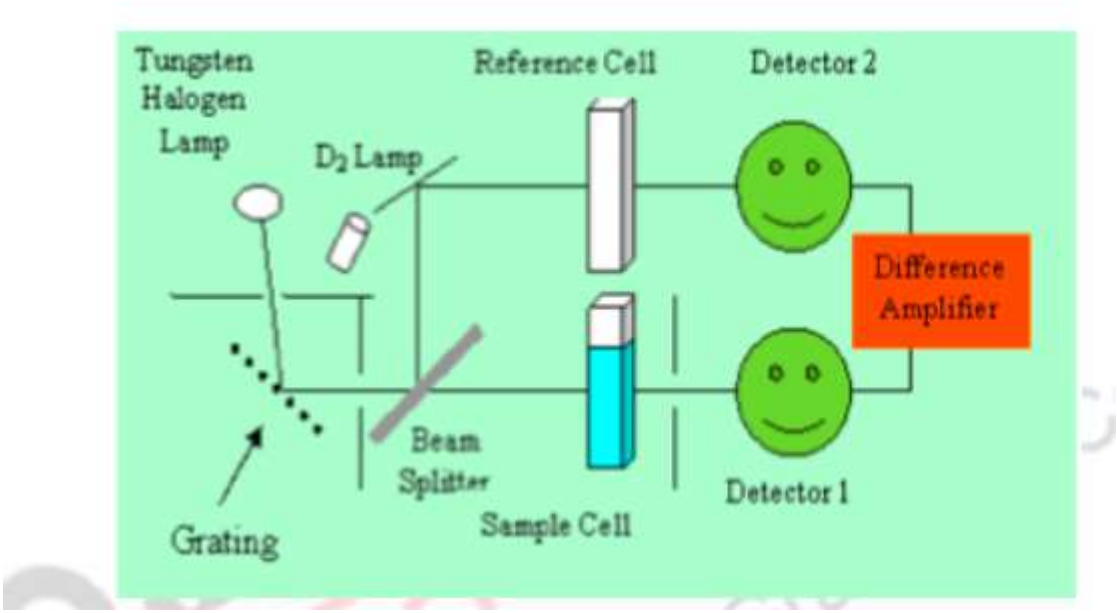


Figure 10: Double beam UV-Visible spectrometer.

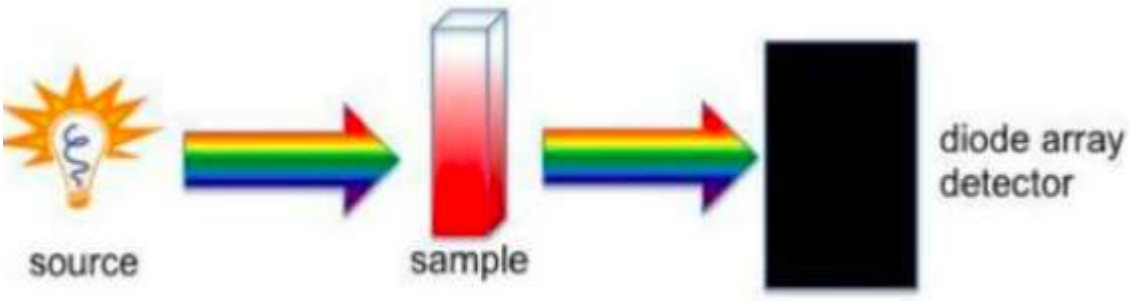


Figure 11: Simultaneous UV-Visible spectrometer.

1.2 Instrumentation:

The basic components of a spectrometer include: light source (UV and visible), monochromator (wavelength selector),sample stage, and detector. A tungsten filament, continuous over UV region is generally used as light source. Detector is usually a photodiode or CCD. Photodiodes go with monochromators to filter light of a particular wavelength, to be fed to the detector. While monitoring the absorbance in UV spectrum, the visible lamp must be turned off, and vice-versa. Figure 12 includes schematic UV−Vis−NIR Spectrometer.

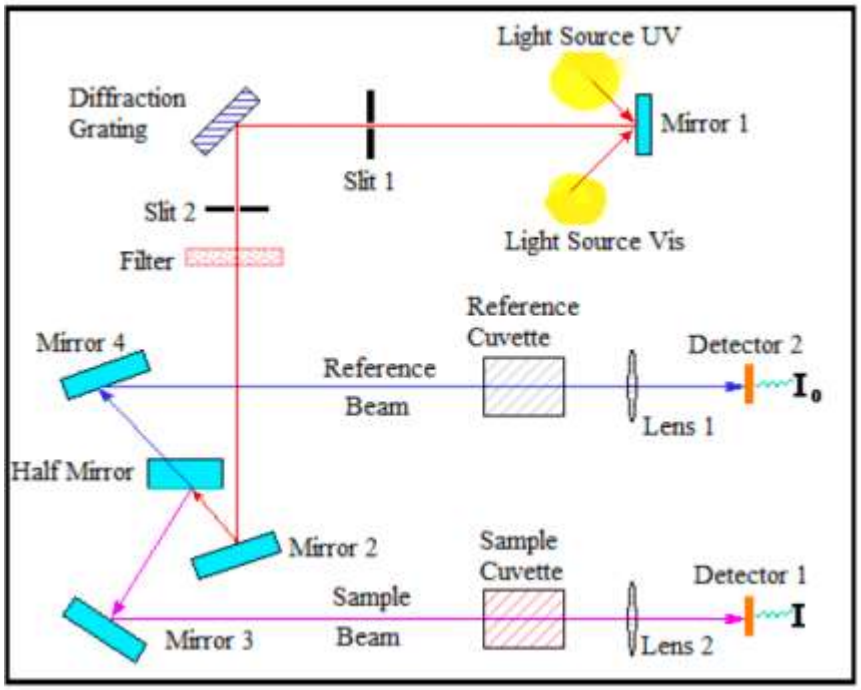


Figure 12: UV– Vis– NIR Spectrometer.

Instrumental components

1. UV Source

The power of radiating source should not vary in its operating wavelength range. Continuous UV spectrum is produced by electrically exciting deuterium or hydrogen at low pressures. The mechanism for generation of UV light includes creating an excited molecular species, that breaks into two atomic species and a UV photon. The emission wavelengths of both deuterium and hydrogen lamps are in 160 to 375 nm range. The material of the cuvettes needs to selected such that it does not absorb the light incident, because this will result in errors in obtained absorption spectrum. Thus, quartz is usually used.

2. Visible Light Source

Tungsten filament lamp is used as visible light source. This lamp can produce light in 350 to 2500 nm wavelength range. In a tungsten filament lamp, energy emitted is proportional to the fourth power of the operating voltage. Thus, in order to get stable emission, a highly stable voltage must be applied to the lamp. The stability of voltage is ensured by using electronic voltage regulators or constant-voltage transformers. Tungsten/halogen lamps include small quantities of iodine embedded within a quartz 'envelope', which also contains the tungsten filament. The iodine reacts with gaseous tungsten, formed by sublimation, and produces a volatile compound WI2. As WI2 molecules hit the filament, they decompose, and redeposit tungsten back on the filament. The tungsten/halogen lamps usually have lifetime twice to the conventional tungsten filament lamp. Tungsten/halogen lamps are used in modern spectrophotometers owing to their high efficiency, and their output extends to UV region as well.

3. Cuvettes

Monochromator source is used; before reaching sample, light is divided in two parts of similar intensity with a half-mirror splitter. One part (or sample beam), travels via the cuvette having the solution of material to be examined in transparent solvent. Second beam, or reference beam, travels via similar cuvette having only solvent. Reference and sample solution containers have to be transparent towards passing beam.

4. Detectors

Detector detects intensity of light transmitted by cuvettes and sends this data to a meter to record and display the values. Electronic detectors calculate and compare the intensities of light beams. Several UV−Vis spectrophotometers have two detectors – a phototube and a photomultiplier tube, and reference and sample beams are monitored simultaneously. The photomultiplier tube is the extensively used detector in UV-Vis instruments. It includes a photoemissive cathode (electrons are emitted from the cathode when photons strike it), several dynodes (a dynode emits multiple electrons when one electron strikes it) and an anode. The incident photon, after entering the tube, strikes the cathode. The cathode then emits multiple electrons, which are then accelerated towards the first dynode (whose potential is 90V more positive than cathode). The electrons strike the first dynode, leading to the emission of several electrons for each incident electron. These electrons are then accelerated towards the second dynode, to produce more electrons which are accelerated towards dynode three and so on. All the electrons are eventually collected at the anode. By this time, each original photon has produced 106 – 107 electrons. The resulting current is amplified and measured. Photomultipliers are highly sensitive towards UV and visible radiations and have fast response times. However, photomultipliers are used only at low power radiation as high power light may damage them.The linear photodiode array is an example of a multichannel photon detector. These detectors can simultaneously measure all elements of a beam of dispersed radiation. A linear photodiode array consists of several small silicon photodiodes created on a single silicon chip. The number of photodiodes can vary between 64 to 4096 sensor elements on a chip, however, 1024 photodiodes is most common. For each diode, there is also a storage capacitor and a switch. The individual diodecapacitor circuits can be sequentially scanned.

Charge-Coupled Devices (CCDs) are like diode array detectors, but instead of diodes, they consist of an array of photocapacitors [4]. Reference beam intensity should suffer little or no absorption, and termed I0 whereas that of sample beam is called I. The spectrophotometer automatically examines all wavelength components in a short time. This technique is good to evaluate the concentration as well as molecular structure or structural changes. It may also be used to examine

the vibrational and conformational energy levels alterations before and after an interaction with a substrate, or a molecule.

Applications

The usual procedure for (colorimetric) assays is to prepare a set of standards and produce a plot of concentration versus absorbance called calibration curve. This should be linear as long as the Beer–Lambert law applies. Absorbances of unknowns are then measured and their concentration interpolated from the linear region of the plot. It is important that one never extrapolates beyond the region for which an instrument has been calibrated as this potentially introduces enormous errors.

Qualitative and quantitative analysis Qualitative analysis may be performed in the UV/Vis regions to identify certain classes of compounds both in the pure state and in biological mixtures (e.g. protein-bound). The application of UV/Vis spectroscopy to further analytical purposes is rather limited, but possible for systems where appropriate features and parameters are known. Most commonly, this type of spectroscopy is used for quantification of biological samples either directly or via colorimetric assays. In many cases, proteins can be quantified directly using their intrinsic chromophores, tyrosine and tryptophan. Protein spectra are acquired by scanning from 500 to 210 nm. The characteristic features in a protein spectrum are a band at 278/280 nm and another at 190 nm . The region from 500 to 300 nm provides valuable information about the presence of any prosthetic groups or coenzymes. Protein quantification by single wavelength measurements at 280 and 260 nm only should be avoided, as the presence of larger aggregates (contaminations or protein aggregates) gives rise to considerable Rayleigh scatter that needs to be corrected.

Solvent perturbation

As we have mentioned above, aromatic amino acids are the main chromophores of proteins in the UV region of the electromagnetic spectrum. Furthermore, the UV absorption of chromophores depends largely on the polarity in its immediate environment. A change in the polarity of the solvent changes the UV spectrum of a protein by bathochromic or hypsochromic effects without changing its conformation. This phenomenon is called solvent perturbation and can be used to probe the surface of a protein molecule. In order to be accessible to the solvent, the chromophore has to be accessible on the protein surface. Practically, solvents like dimethyl-sulfoxide, dioxane, glycerol, mannitol, sucrose and polyethylene glycol are used for solvent perturbation experiments, because they are miscible with water. The method of solvent perturbation is most commonly used for determination of the number of aromatic residues that are exposed to solvent.

Spectrophotometric and colorimetric assays

For biochemical assays testing for time- or concentration-dependent responses of systems, an appropriate read-out is required that is coupled to the progress of the reaction (reaction coordinate). Therefore, the biophysical parameter being monitored (read-out) needs to be coupled to the biochemical parameter under investigation. Frequently, the monitored parameter is the absorbance of a system at a given wavelength which is monitored throughout the course of the experiment. Preferably, one should try to monitor the changing species directly (e.g. protein absorption, starting product or generated product of a reaction), but in many cases this is not possible and a secondary reaction has to be used to generate an appropriate signal for monitoring. A common application of the latter approach is the determination of protein concentration by Lowry or Bradford assays, where a secondary reaction is used to colour the protein. The more intense the colour, the more protein is

present. These assays are called colorimetric assays and a number of commonly used ones are listed in Table

Use of UV-Visible-

-Structure determinaon

– Peaks give informaon on molecular orbitals.

- Idencaon

– Molecular 7ngerprint.

- Solvent informaon

– Solvents aHects spectrum.

- Toxicology

– DiHerenate between very similar compounds.

- Detecon of impuries.

- Quantave and qualitave analysis.

- As detector in HPLC.

- As a method of measuring rates of reaction

3.4. Optical Rotatory Dispersion and Circular Dichroism

Circular Dichroism, an absorption spectroscopy, uses circularly polarized light to investigate structural aspects of optically active chiral media. It is mostly used to study biological molecules, their structure, and interactions with metals and other molecules.

introduction

Circular Dichroism (CD) is an absorption spectroscopy method based on the differential absorption of left and right circularly polarized light. Optically active chiral molecules will preferentially absorb one direction of the circularly polarized light. The difference in absorption of the left and right circularly polarized light can be measured and quantified. UV CD is used to determine aspects of protein secondary structure. Vibrational CD, IR CD, is used to study the structure of small organic molecules, proteins and DNA. UV/Vis CD investigates charge transfer transitions in metal-protein complexes.

Circular Polarization of Light

Electromagnetic radiation consists of oscillating electric and magnetic fields perpendicular to each other and the direction of propagation. Most light sources emit waves where these fields oscillate in all directions perpendicular to the propagation vector. Linear polarized light occurs when the electric field vector oscillates in only one plane. In circularly polarized light, the electric field vector rotates around the propagation axis maintaining a constant magnitude. When looked at down the axis of propagation the vector appears to trace a circle over the period of one wave frequency (one full

rotation occurs in the distance equal to the wavelength). In linear polarized light the direction of the vector stays constant and the magnitude oscillates. In circularly polarized light the magnitude stays constant while the direction oscillates.

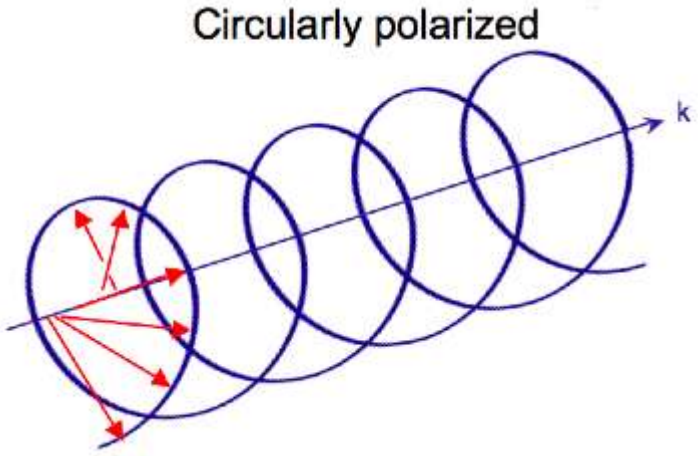


Figure 13: Diagram of linearly polarized and circularly polarized light

As the radiation propagates the electric field vector traces out a helix. The magnetic field vector is out of phase with the electric field vector by a quarter turn. When traced together the vectors form a double helix.

Light can be circularly polarized in two directions: left and right. If the vector rotates counterclockwise when the observer looks down the axis of propagation, the light is left circularly polarized (LCP). If it rotates clockwise, it is right circularly polarized (RCP). If LCP and RCP of the same amplitude, they are superimposed on one another and the resulting wave will be linearly polarized.

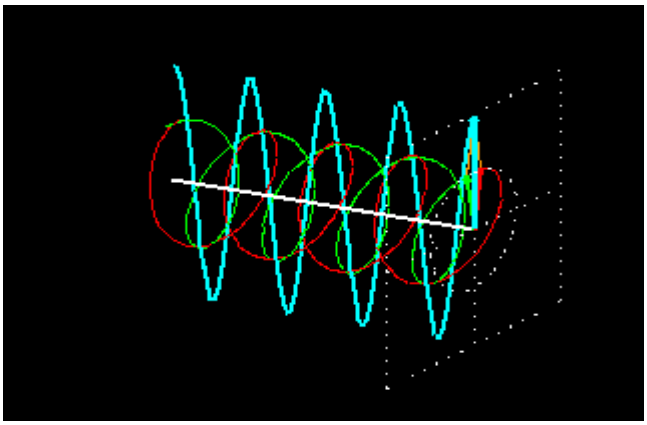


Figure 14: The superposition of LCP and RCP light of the same amplitude produces linearly polarized light

Instrumentation

The basic layout of a CD spectrometer follows that of a single-beam UV absorption spectrometer. Owing to the nature of the measured effects, an electro-optic modulator, as well as a more sophisticated detector are needed, though. Mean residue ellipticity (1° cm2 dmol–1) 190 –60 x 103 200 210 220 230 240 250 –40 x 103 –20 x 103 20 x 103 0 40 x 103 60 x 103 80 x 103 100 x 103 (nm) Fig. 12.18 Circular dichroism spectra for three standard secondary structures according to Fasman. An a-helical peptide is shown in dark green, a peptide adopting b-strand structure in grey, and a random coil peptide in light green. 512 Spectroscopic techniques: I Photometric techniques Generally, left and right circularly polarised light passes through the sample in an alternating fashion. This is achieved by an electro-optic modulator which is a crystal that transmits either the left- or right-handed polarised component of linearly polarised light, depending on the polarity of the electric field that is applied by alternating currents. The photomultiplier detector produces a voltage proportional to the ellipticity of the resultant beam emerging from the sample. The light source of the spectrometer is continuously flushed with nitrogen to avoid the formation of ozone and help to maintain the lamp. CD spectrometry involves measuring a very small difference between two absorption values which are large signals. The technique is thus very susceptible to noise and measurements must be carried out carefully. Some practical considerations involve having a clean quartz cuvette, and using buffers with low concentrations of additives. While this is sometimes tricky with protein samples, reducing the salt concentrations to values as low as 5 mM helps to obtain good spectra. Also, filtered solutions should be used to avoid any turbidity of the sample that could produce scatter. Saturation of the detector must be avoided, this becoming more critical with lower wavelengths. Therefore, good spectra are obtained in a certain range of protein concentrations only where enough sample is present to produce a good signal and does not saturate the detector. Typical

protein concentrations are 0.03–0.3 mg cm^{-3} . In order to calculate specific ellipticities (mean residue ellipticities) and be able to compare the CD spectra of different samples with each other, the concentration of the sample must be known. Provided the protein possesses sufficient amounts of UV/Vis-absorbing chromophores, it is thus advisable to subject the CD sample to a protein concentration determination by UV/Vis Section.

Applications

The main application for protein CD spectroscopy is the verification of the adopted secondary structure. The application of CD to determine the tertiary structure is limited, owing to the inadequate theoretical understanding of the effects of different parts of the molecules at this level of structure.

Rather than analysing the secondary structure of a 'static sample', different conditions can be tested. For instance, some peptides adopt different secondary structures when in solution or membrane-bound. The comparison of CD spectra of such peptides in the absence and presence of small unilamellar phospholipid vesicles shows a clear difference in the type of secondary structure. Measurements with lipid vesicles are tricky, because due to their physical extensions they give rise to scatter. Other options in this context include CD experiments at lipid monolayers which can be realised at synchrotron beam lines, or by usage of optically clear vesicles (reverse micelles).

CD spectroscopy can also be used to monitor changes of secondary structure within a sample over time. Frequently, CD instruments are equipped with temperature control units and the sample can be heated in a controlled fashion. As the protein undergoes its transition from the folded to the unfolded state, the CD at a certain wavelength (usually 222 nm) is monitored and plotted against the temperature, thus yielding a thermal denaturation curve which can be used for stability analysis.

3.5. ELECTRON PARAMAGNETIC RESONANCE

Prior to any detailed discussion of electron paramagnetic resonance (EPR) and nuclear magnetic resonance (NMR) methods, it is worthwhile considering the more general phenomena applicable to both.

Magnetic phenomena

Magnetism arises from the motion of charged particles. This motion is controlled by internal forces in a system. For the purpose of this discussion, the major contribution to magnetism in molecules is due to the spin of the charged particle. In chemical bonds of a molecule, the negatively charged electrons have a spin controlled by strict quantum rules. A bond is constituted by two electrons with opposite spins occupying the appropriate molecular orbital. According to the Pauli principle, the two electrons must have opposite spins, leading to the term paired electrons. Each of the spinning electronic charges generates a magnetic effect, but in electron pairs the effect is almost self-cancelling. In atoms, a value for magnetic susceptibility may be calculated and is of the order of 10^{-6} g^{-1} . This diamagnetism is a property of all substances, because they all contain the minuscule magnets, i.e. electrons. Diamagnetism is temperature independent.

If an electron is unpaired, there is no counterbalancing opposing spin and the magnetic susceptibility is of the order of $+10^{-3}$ to $+10^{-4}$ g^{-1} . The effect of an unpaired electron exceeds the 'background' diamagnetism, and gives rise to paramagnetism. Free electrons can arise in numerous cases. The most notable example is certainly the paramagnetism of metals such as iron, cobalt and nickel, which are the materials that permanent magnets are made of. The paramagnetism of these metals is called ferromagnetism. In biochemical investigations, systems with free electrons (radicals) are frequently used as probes.

Similar arguments can be made regarding atomic nuclei. The nucleus of an atom is constituted by protons and neutrons, and has a net charge that is normally compensated by the extra-nuclear electrons. The number of all nucleons (Z) is the sum of the number of protons (P) and the number of neutrons (N). P and Z determine whether a nucleus will exhibit paramagnetism. Carbon-12 (^{12}C), for example, consists of six protons (P ¼ 6) and six neutrons (N ¼ 6) and thus has Z ¼ 12. P and Z are even, and therefore the ^{12}C nucleus possesses no nuclear magnetism. Another example of a nucleus with no residual magnetism is oxygen-16 (^{16}O). All other nuclei with P and Z being uneven possess residual nuclear magnetism.

The way in which a substance behaves in an externally applied magnetic field allows us to distinguish between dia- and paramagnetism. A paramagnetic material is attracted by an external magnetic field, while a diamagnetic substance is rejected. This principle is employed by the Guoy balance, which allows quantification of magnetic effects. A balance pan is suspended between the poles of a suitable electromagnet supplying the external field. The substance under test is weighed in air with the current switched off. The same sample is then weighed again with the current (i.e. external magnetic field) on. A paramagnetic substance appears to weigh more, and a diamagnetic substance appears to weigh less.

The resonance condition

In both EPR and NMR techniques, two possible energy states exist for either electronic or nuclear magnetism in the presence of an external magnetic field. In the low-energy state, the field generated by the spinning charged particle is parallel to the external field. Conversely, in the high-energy state, the field generated by the spinning charged particle is antiparallel to the external field. When enough energy is input into the system to cause a transition from the low- to the high-energy state, the condition of resonance is satisfied. Energy must be absorbed as a discrete dose (quantum) h, where h is the Planck constant and is the frequency (see equation 12.1). The quantum energy required to fulfil the resonance condition and thu

$$hv=g\beta B$$

where g is a constant called spectroscopic splitting factor, β is the magnetic moment of the electron (termed the Bohr magneton), and B is the strength of the applied external magnetic field. The frequency v of the absorbed radiation is a function of the paramagnetic species β and the applied magnetic field B. Thus, either or B may be varied to the same effect. With appropriate external magnetic fields, the frequency of applied radiation for EPR is in the microwave region, and for NMR in the region of radio frequencies. In both techniques, two possibilities exist for determining the absorption of electromagnetic energy (i.e. enabling the resonance phenomenon):

- constant frequency v is applied and the external magnetic field B is swept; or
- constant external magnetic field B is applied and the appropriate frequency is selected by sweeping through the spectrum.

For technical reasons, the more commonly used option is a sweep of the external magnetic field.

Principles

The absorption of energy is recorded in the EPR spectrum as a function of the magnetic induction measured in Tesla (T) which is proportional to the magnetic field strength applied. The area under the absorption peak is proportional to the number of unpaired electron spins. Most commonly, the first derivative of the absorption peak is the signal that is actually recorded. For a delocalised electron, as observed e.g. in free radicals, the g value is 2.0023; but for localised electrons such as in transition metal atoms, g varies, and its precise value contains information about the nature of bonding in the environment of the unpaired electron within the molecule. When resonance occurs, the absorption peak is broadened owing to interactions of the unpaired electron with the rest of the molecule (spin–lattice interactions). This allows further conclusions as to the molecular structure.

High-resolution EPR may be performed by examining the hyperfine splitting of the absorption peak which is caused by interaction of the unpaired electron with adjacent nuclei, thus yielding information about the spatial location of atoms in the molecule. The proton hyperfine splitting for free radicals occurs in the range of $0–3*10^{-3}$ T, and yields data analogous to those obtained in high-resolution NMR.

The effective resolution of an EPR spectrum can be considerably improved by combining the method with NMR, a technique called electron nuclear double resonance (ENDOR). Here, the sample is irradiated simultaneously with microwaves for EPR and radio frequencies (RF) for NMR. The RF signal is swept for fixed points in the EPR spectrum, yielding the EPR signal height versus nuclear RF. This approach is particularly useful when there are a large number of nuclear levels that broaden the normal electron resonance lines.

The technique of electron double resonance (ELDOR) finds an application in the separation of overlapping multiradical spectra and the study of relaxation phenomena, for example chemical spin exchange. In ELDOR, the sample is irradiated with two microwave frequencies simultaneously. One is used for observation of the EPR signal at a fixed point in the spectrum, the other is used to sweep other parts of the spectrum. The recorded spectrum is plotted as a function of the EPR signal as a function of the difference of the two microwave frequencies.

Instrumentation

Figure 13.5 shows the main components of an EPR instrument. The magnetic fields generated by the electromagnets are of the order of 50 to 500 mT, and variations of less than 10^{-6} are required for highest accuracy. The monochromatic microwave radiation is produced in a klystron oscillator with wavelengths around 3 cm (9 GHz). The samples are required to be in the solid state; hence biological samples are usually frozen in liquid nitrogen. The technique is also ideal for investigation of membranes and membrane proteins. Instead of plotting the absorption A versus B, it is the first-order differential (dA/dB) that is usually plotted against B (Fig. 13.6). Such a shape is called a ‘line’ in EPR spectroscopy. Generally, there are relatively few unpaired electrons in a molecule, resulting in fewer than 10 lines, which are not closely spaced.

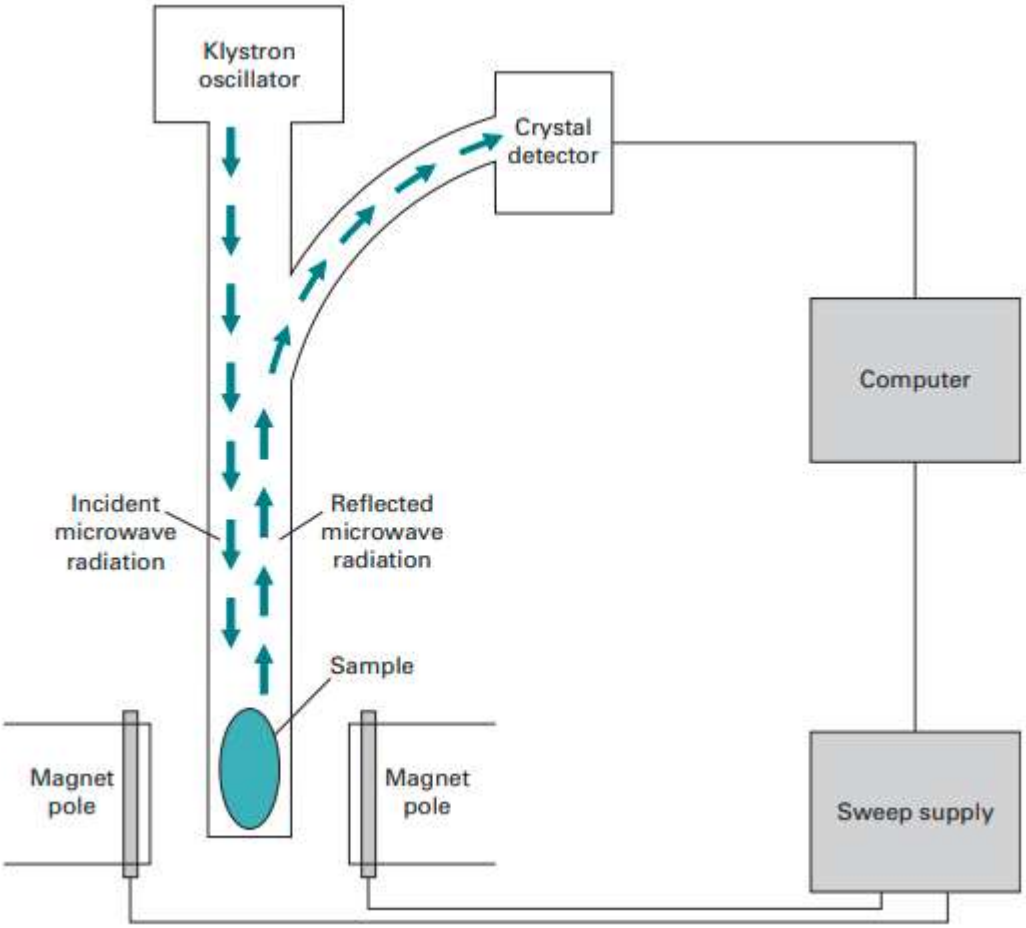


Figure. 15 Diagram of an EPR spectrometer.

Applications

Metalloproteins

EPR spectroscopy is one of the main methods to study metalloproteins, particularly those containing molybdenum (xanthine oxidase), copper (cytochrome oxidase, copper blue enzymes) and iron (cytochrome, ferredoxin). Both copper and non-haem iron, which do not absorb in the UV/Vis region, possess EPR absorption peaks in one of their oxidised states. The appearance and disappearance of their EPR signals are used to monitor the activity of these proteins in the multi-enzyme systems of intact mitochondria and chloroplasts, as well as in isolated enzymes. In many metalloproteins, the ligands coordinating the metal ion are the amino acid residues of the protein. Coordination chemistry requires a specific stereochemical structure of the ligands, and EPR studies show that the geometry is frequently distorted in proteins when compared to model systems. Such distortions may be related to biological function.

Spin labels Spin labels are stable and non-reactive unpaired electrons used as reporter groups or probes for EPR. The procedure of spin labelling is the attachment of these probes to biological molecules that lack unpaired electrons. The label can be attached to either a substrate or a ligand.

Often, a spin label contains the nitric oxide moiety. These labels enable the study of events that occur with a frequency of 10^7 to 10^{11} s^{-1} . If the motion is restricted in some directions, only anisotropic motion (movement in one particular direction) may be studied, for example in membrane-rigid spin labels in bilayers. Here, the label is attached so that the NO group lies parallel to the long axis of the lipid. Intramolecular motions and lateral diffusion of lipid through the membrane, as well as the effect of proteins and other factors on these parameters may be observed. Quantification of effects often involves calculation of the order parameter Z. Spinlabelled lipids are either concentrated into one region of the bilayer or randomly incorporated into model membranes. The diffusion of spin labels allows them to come into contact with each other, which causes line-broadening in the spectrum. Labelling of phospholipids with 2,2,6,6-tetramethylpiperidine-1-oxyl (TEMPOL) is used for measurement of the flip rate of phospholipids between inner and outer surfaces as well as lateral diffusion.

Free radicals

Molecules in their triplet states (Fig. 12.8) have unpaired electrons and thus are amenable to EPR spectroscopy. Such molecules possess the property of phosphorescence and EPR may deliver data complementary to the UV/Vis region of the spectrum. For instance, free radicals due to the triplet state of tryptophan have been observed in cataractuous lenses.

Spin trapping is a process whereby an unstable free radical is being stabilised by reaction with a compound such as 5,5-dimethylpyrroline-1-oxide (DMPO). Hyperfine splittings (Fig. 13.6) are observed that depend upon the nature of the radical.

Carcinogenesis is an area where free radicals have been implicated. While free radicals promote the generation of tumours through damage due to their high reactivity, there is, in general, a lower concentration of radicals in tumours than in normal tissue. Also, a gradient has been observed with higher concentrations of radicals in the peripheral non-necrotic surface layers than in the inner regions of the tumour. EPR has been used to study implanted tumours in mice, but also in evaluation of potential chemical carcinogens. Polycyclic hydrocarbons, such as naphthalene, anthracene and phenantrene, consist of multiple aromatic ring systems. These extended aromatic systems allow for single free electrons to be accommodated and thus yield long-lived free radicals, extending the periods of time in which damage can be done. Many of the precursors of these radicals exist in natural sources such as coal tar, tobacco smoke and other products of combustion, hence the environmental risk. Another source of free radicals is irradiation with UV light or gamm-rays. Ozone is an oxygen radical that is present as a protective shield around the Earth, filtering the dangers of cosmic UV irradiation by complex radical chemistry. The pollution of the Earth’s atmosphere with radical-forming chemicals has destroyed large parts of the ozone layer, increasing the risk of skin cancer from sun exposure. EPR can be used to study biological materials, including bone or teeth, and detect radicals formed due to exposure to high energy radiation. Another major application for EPR is the examination of irradiated foodstuffs for residual free radicals, and it is mostly used to establish whether packed food has been irradiated.

3.6. NUCLEAR MAGNETIC RESONANCE

The essential background theory of the phenomena that allow NMR to occur have been introduced in Section. However, the miniature magnets involved here are not electrons, but the nuclei. The specific principles, instrumentation and applications are discussed below.

Principles

Most studies in organic chemistry involve the use of 1 H, but NMR spectroscopy with ^{13}C, ^{15}N and ^{31}P isotopes is frequently used in biochemical studies. The resonance condition in NMR is satisfied in an external magnetic field of several hundred mT, with absorptions occurring in the region of radio waves (frequency 40 MHz) for resonance of the 1 H nucleus. The actual field scanned is small compared with the field strengths applied, and the radio frequencies absorbed are specifically stated on such spectra.

Similar to other spectroscopic techniques discussed earlier, the energy input in the form of electromagnetic radiation promotes the transition of 'entities' from lower to higher energy states (Fig. 16). In case of NMR, these entities are the nuclear magnetic spins which populate energy levels according to quantum chemical rules. After a certain time-span, the spins will return from the higher to the lower energy level, a process that is known as relaxation.

The energy released during the transition of a nuclear spin from the higher to the lower energy state can be emitted as heat into the environment and is called spin– lattice relaxation. This process happens with a rate of T_1^{-1} , and T_1 is termed the longitudinal relaxation time, because of the change in magnetisation of the nuclei parallel to the field. The transverse magnetisation of the nuclei is also subject to change over time, due to interactions between different nuclei. The latter process is thus called spin–spin relaxation and is characterised by a transverse relaxation time T_2

The molecular environment of a proton governs the value of the applied external field at which the nucleus resonates. This is recorded as the chemical shift (delta) and is measured relative to an internal standard, which in most cases is tetramethylsilane (TMS; $(H_3C)_4Si$) because it contains 12 identical protons. The chemical shift arises from the applied field inducing secondary fields of about 0.15–0.2 mT at the proton by interacting with the adjacent bonding electrons.

- If the induced field opposes the applied field, the latter will have to be at a slightly higher value for resonance to occur. The nucleus is said to be shielded, the magnitude of the shielding being proportional to the electron-withdrawing power of proximal substituents.
- Alternatively, if the induced and applied fields are aligned, the latter is required to be at a lower value for resonance. The nucleus is then said to be deshielded.

Usually, deuterated solvents such as CDCl3 are used for sample preparation of organic compounds. For peptides and proteins D2O is the solvent of choice. Because the stability of the magnetic field is critical for NMR spectroscopy, the magnetic flux needs to be tuned, e.g. by locking with deuterium resonance frequencies. The use of deuterated solvents thus eliminates the need for further experiments. The chemical shift is plotted along the x-axis, and measured in p.p.m. instead of the actual magnetic field strengths. This conversion makes the recorded spectrum independent of the magnetic field used. The signal of the internal standard TMS appears at δ= 0 p.p.m. The type of proton giving rise to a particular band may thus be identified

by the resonance peak position, i.e. its chemical shift, and the area under each peak is proportional to the number of protons of that particular type. Figure 16 shows an H NMR spectrum of ethyl alcohol, in which there are three methyl, two methylene and one alcohol group protons. The peak areas are integrated, and show the proportions 3 : 2 : 1. Owing to the interaction of bonding electrons with like or different spins, a phenomenon called spin-spin coupling (also termed scalar or J-coupling) arises that can extend to nuclei four or five bonds apart. This results in the splitting of the three bands in Fig. 13.8 into several finer bands (hyperfine splitting). The hyperfine splitting yields valuable information about the near-neighbour environment of a nucleus. NMR spectra are of great value in elucidating chemical structures. Both qualitative and quantitative information may be obtained. The advances in computing power have made possible many more advanced NMR techniques. Weak signals can be enhanced by running many scans and accumulating the data. Baseline noise, which is random, tends to cancel out whereas the signal increases. This approach is known as computer averaging of transients or CAT scanning, and significantly improves the signal-to-noise ratio. Despite the value and continued use of such 'conventional' ^{1}H NMR, much more structural information can be obtained by resorting to pulsed input of radio frequency energy, and subjecting the output to Fourier transform. This approach has given rise to a wide variety of procedures using multidimensional spectra, ^{13}C and other odd-isotope NMR spectra and the determination of multiplicities and scan images.

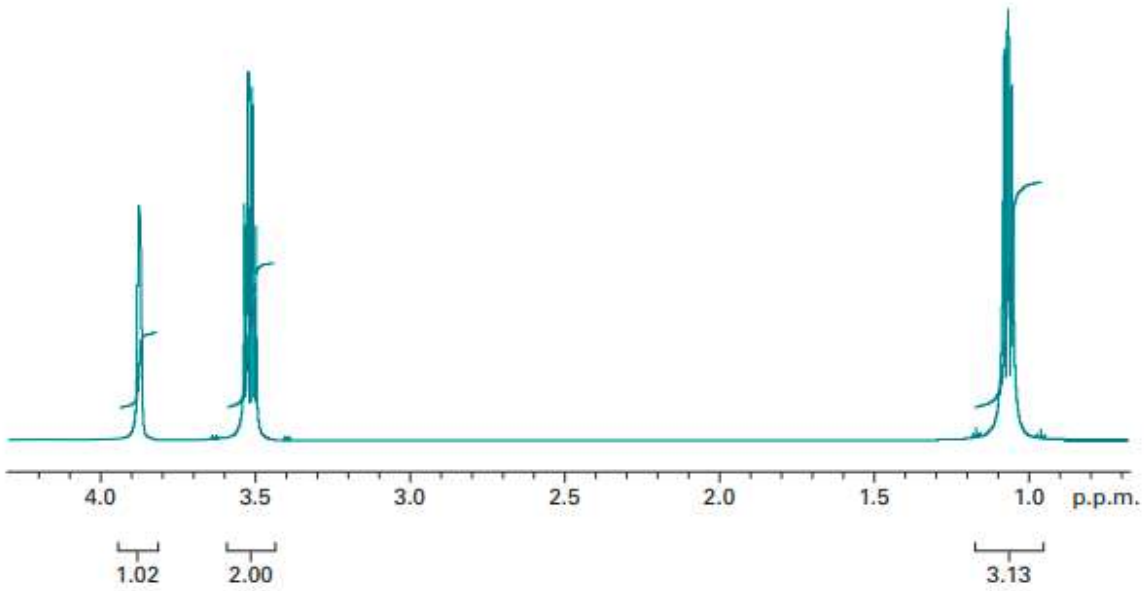


Figure. 16 ^{1}H NMR spectrum of ethyl alcohol (H3C–CH2–OH) with integrated peaks.

Pulse-acquire and Fourier transform methods

In 'conventional' NMR spectroscopy, the electromagnetic radiation (energy) is supplied from the source as a continuously changing frequency over a preselected spectral range (continuous wave method). The change is smooth and regular between fixed limits. Figure 17a illustrates this approach. During the scan, radiation of certain energy in the form of a sine wave is recorded. By using the mathematical procedure of Fourier transform, the 'time domain' can be resolved into a 'frequency domain'. For a single-frequency sine wave, this procedure yields a single peak of fixed amplitude. However, because the measured signal in NMR is the re-emission of energy as the nuclei return from their high-energy into their low-energy states, the recorded radiation will decay with time, as fewer

and fewer nuclei will return to the ground state. The signal measured is thus called the free induction decay (FID). Figure 17b shows the effect of the FID on the corresponding Fourier transform. The frequency band broadens, but the peak position and the amplitude remain the same. The resolved frequency peak represents the chemical shift of a nucleus resonating at this energy.

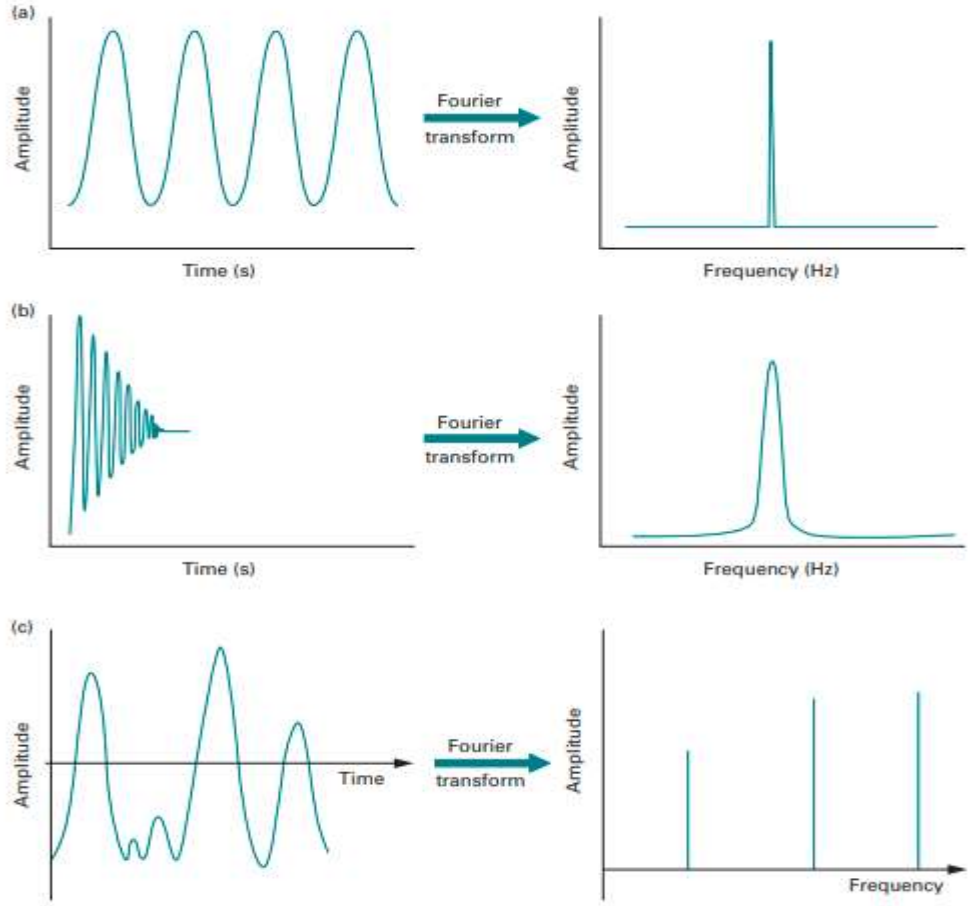


Figure. 17 Diagrammatic representation of the Fourier transformation of (a) a single frequency sine wave, (b) a single frequency FID, and (c) a three-sine-wave combination.

Alternatively, the total energy comprising all frequencies between the fixed limits can be put in all at the same time. This is achieved by irradiating the sample with a broadband pulse of all frequencies at one go. The output will measure all resonance energies simultaneously and will result in a very complicated interference pattern. However, Fourier transform is able to resolve this pattern into the constituting frequencies (Fig. 17c).

In the presence of an external magnetic field, nuclear spins precess around the axis of that field with the so-called Larmor frequency. The vector sum of all nuclear magnetic moments yields a magnetisation parallel to the external field, i.e. a longitudinal magnetisation. When a high-frequency pulse is applied, the overall magnetisation is forced further off the precession by a pulse angle. This introduces a new vector component to the overall magnetisation which is perpendicular to the external field; this component is called transverse magnetisation. The FID measured in pulseacquired spectra is, in fact, the decay of that transverse magnetisation component.

Nuclear Overhauser effect

It has already been mentioned above that nuclear spins generate magnetic fields which can exert effects through space, for example as observed in spin–spin coupling. This coupling is mediated through chemical bonds connecting the two coupling spins. However, magnetic nuclear spins can also exert effects in their proximal neighbourhood via dipolar interactions. The effects encountered in the dipolar interaction are transmitted through space over a limited distance on the order of 0.5 nm or less. These interactions can lead to the nuclear Overhauser effects (NOEs), as observed in a changing signal intensity of a resonance when the state of a near neighbour is perturbed from the equilibrium. Because of the spatial constraint, this information enables conclusions to be drawn about the three-dimensional geometry of the molecule being examined.

^{13}C NMR

Due to the low abundance of the ^{13}C isotope, the chance of finding two such species next to each other in a molecule is very small (see Chapter 9). As a consequence, ^{13}C–^{13}C couplings (homonuclear couplings) do not arise. While ^{1}H–^{13}C interactions (heteronuclear coupling) are possible, one usually records decoupled 13C spectra where all bands represent carbon only. 13C spectra are thus much simpler and cleaner when compared to 1 H spectra. The main disadvantage though is the fact that multiplicities in these spectra cannot be observed, i.e. it cannot be decided whether a particular ^{13}C is associated with a methyl (H_3C), a methylene (H_2C) or a methyne (HC) group. Some of this information can be regained by irradiating with an off-resonance frequency during a decoupling experiment. Another routinely used method is called distortionless enhancement by polarisation transfer (DEPT), where sequences of multiple pulses are used to excite nuclear spins at different angles, usually 45^0, 90^0 or 135^0. Although interactions have been decoupled, in this situation the resonances exhibit positive or negative signal intensities dependent on the number of protons bonded to the carbon. In DEPT-135, for example, a methylene group yields a negative intensity, while methyl and methyne groups yield positive signals.

Multidimensional NMR

As we learned above, the observable in pulse-acquired Fourier transform NMR is the decay of the transverse magnetisation, called free induction decay (FID). The detected signal thus is a function of the detection time t_2. Within the pulse sequence, the time t_1 (evolution time) describes the time between the first pulse and signal detection. If t_1 is systematically varied, the detected signal becomes a function of both t_1 and t_2, and its Fourier transform comprises two frequency components. The two components form the basis of a two-dimensional spectrum.

Correlated 2D-NMR spectra show chemical shifts on both axes. Utilising different pulse sequences leads to different methods, such as correlated spectroscopy (COSY), nuclear overhauser effect spectroscopy (NOESY), etc. Such methods yield the homonuclear 1 H couplings. The 1D-NMR spectrum now appears along the diagonal and long-range couplings between particular nuclei appear as off-diagonal signals (Figure 18).

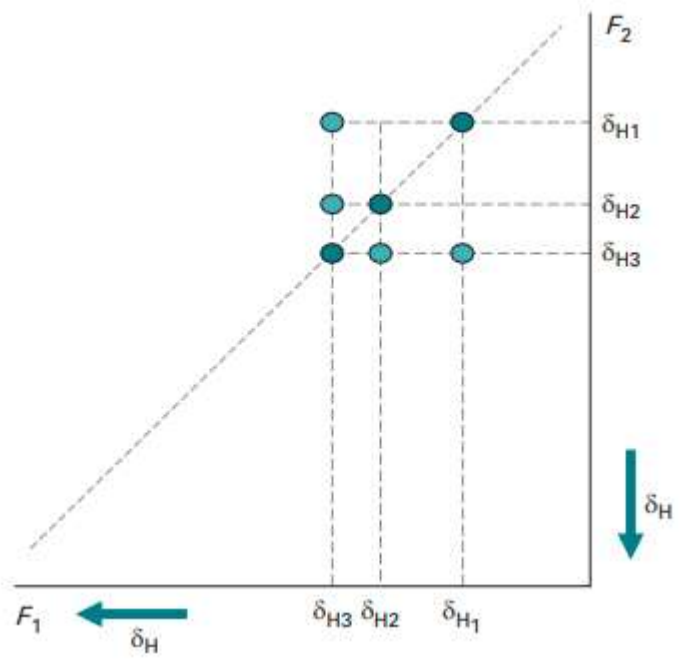


Figure. 18: Schematics of a correlated 2D-1 H NMR spectrum. H3 couples with H2 and H1. H1 and H2 show no coupling.

Instrumentation

Schematically, an analytical NMR instrument is very similar to an EPR instrument, except that instead of a klystron generating microwaves two sets of coils are used to generate and detect radio frequencies (Fig. 19). Samples in solution are contained in sealed tubes which are rotated rapidly in the cavity to eliminate irregularities and imperfections in sample distribution. In this way, an average and uniform signal is reflected to the receiver to be processed and recorded. In solid samples, the number of spin–spin interactions is greatly enhanced due to intermolecular interactions that are absent in dissolved samples due to translation and rotation movements. As a result, the resonance signals broaden significantly. However, high-resolution spectra can be obtained by spinning the solid sample at an angle of 54.7^0 (magic angle spinning). The sophisticated pulse sequences necessary for multidimensional NMR require a certain geometric layout of the radio frequency coils and sophisticated electronics. Advanced computer facilities are needed for operation of NMR instruments, as well as analysis of the acquired spectra.

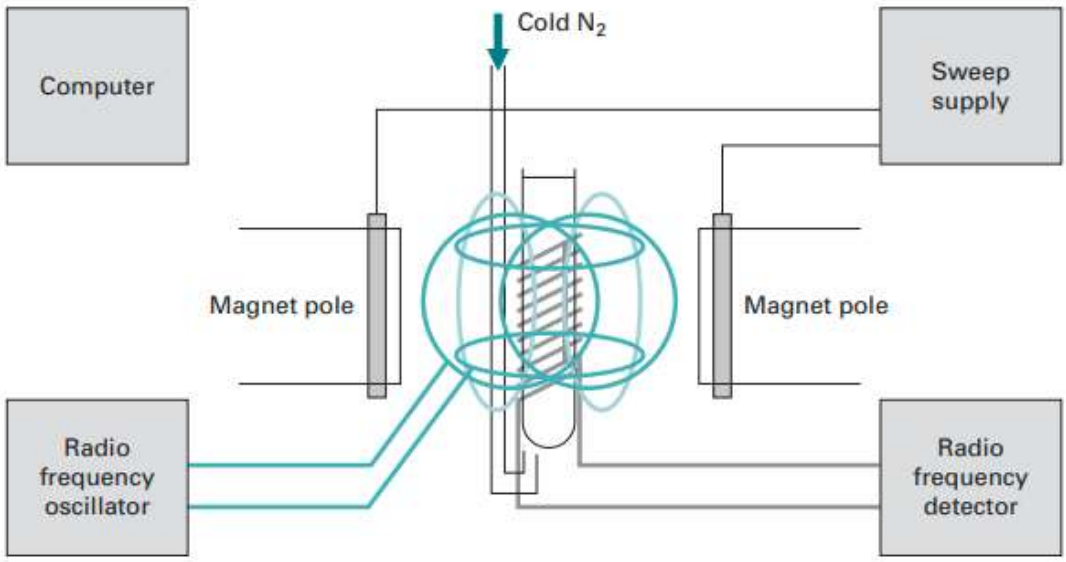


Figure. 19 Schematic diagram of an NMR spectrometer with cryoprobe.

Applications

Molecular structure determination

Traditionally, NMR spectroscopy is the main method of structure determination for organic compounds. The chemical shift provides a clue about the environment of a particular proton or carbon, and thus allows conclusions as to the nature of functional groups. Spin–spin interactions allow conclusions as to how protons are linked with the carbon skeleton. For structure determination, the fine structure usually is the most useful information because it provides a unique criterion while chemical shifts of some groups can vary over an extended range. Additionally, the signal intensity provides information as to how many protons contribute to a particular signal.

Solution structure of proteins and peptides

The structures of proteins up to a mass of about 50 kDa can be determined with biomolecular NMR spectroscopy. The development of magnets with very high field strengths (currently 900 MHz) continues to push the size limit. The preparation of proteins or selected domains for NMR requires recombinant expression and isotopic labelling to enrich the samples with ^{13}C and ^{15}N; 2 H labelling might be required as well. Sample amounts in the order of 10 mg used to be required for NMR experiments; however, the introduction of cryoprobe technology has reduced the sample amount significantly. Heteronuclear multidimensional NMR spectra need to be recorded for the assignment of all chemical shifts (^{1}H, ^{13}C, ^{15}N). For interproton NOEs, ^{13}C- and ^{15}N-edited 3D NOESY spectra are required. The data acquisition can take several weeks, after which spectra are processed (Fourier transformation) and improved with respect to digital resolution and signal-to-noise. Assignment of chemical shifts and interatomic distances is carried out with the help of software programs. All experimentally derived parameters are then used as restraints in a molecular dynamics or simulated annealing structure calculation. The result of a protein NMR structure is an ensemble of structures, all of which are consistent with the experimentally determined restraints, but converge to the same fold.

Magnetic resonance imaging

The basic principles of NMR can be applied to imaging of live samples. Because the proton is one of the more sensitive nuclides and is present in all biological systems abundantly, ^{1}H resonance is used almost exclusively in the clinical environment. The most important compound in biological samples in this context is water. It is distributed differently in different tissues, but constitutes about 55% of body mass in the average human. In soft tissues, the water distribution varies between 60% and 90%. In NMR, the resonance frequency of a particular nuclide is proportional to the strength of the applied external magnetic field. If an external magnetic field gradient is applied then a range of resonant frequencies are observed, reflecting the spatial distribution of the spinning nuclei. Magnetic resonance imaging (MRI) can be applied to large volumes in whole living organisms and has a central role in routine clinical imaging of large-volume soft tissues.

The number of spins in a particular defined spatial region gives rise to the spin density as an observable parameter. This measure can be combined with analysis of the principal relaxation times (T1 and T2). The imaging of flux, as either bulk flow or localised diffusion, adds considerably to the options available. In terms of whole-body scanners, the entire picture is reconstructed from images generated in contiguous slices. MRI can be applied to the whole body or specific organ investigations on head, thorax, abdomen, liver, pancreas, kidney and musculoskeletal regions. The use of contrast agents with paramagnetic properties has enabled investigation of organ function, as well as blood flow, tissue perfusion, transport across the blood– brain barrier and vascular anatomy. Resolution and image contrast are major considerations for the technique and subject to continuing development. The resolution depends on the strength of the magnetic field and the availability of labels that yield high signal strengths. MRI instruments used for clinical imaging typically operate with field strengths of up to 3 T, but experimental instruments can operate at more than 20 T, allowing the imaging of whole live organisms with almost enough spatial and temporal resolution to follow regenerative processes continuously at the singlecell level. Equipment cost and data acquisition time remain important issues. On the other hand, according to current knowledge, MRI has no adverse effects on human health, and thus provides a valuable diagnostic tool, especially due to the absence of the hazards of ionising radiation.

3.7. ATOMIC SPECTROSCOPY

So far, all methods have dealt with probing molecular properties. In Section we discussed the general theory of electronic transitions and said that molecules give rise to band spectra, but atoms yield clearly defined line spectra. In atomic emission spectroscopy (AES), these lines can be observed as light of a particular wavelength (colour). Conversely, black lines can be observed against a bright background in atomic absorption spectroscopy (AAS). The wavelengths emitted from excited atoms may be identified using a spectroscope with the human eye as the ‘detector’ or a spectrophotometer.

Principle

In a spectrum of an element, the absorption or emission wavelengths are associated with transitions that require a minimum of energy change. In order for energy changes to be minimal, transitions tend to occur between orbitals close together in energy terms. For example, excitation of a sodium atom and its subsequent relaxation gives rise to emission of orange light (‘D-line’) due to the transition of an electron from the 3s to the 3p orbital and return (Fig. 20). Electron transitions in

an atom are limited by the availability of empty orbitals. Filling orbitals with electrons is subject to two major rules:

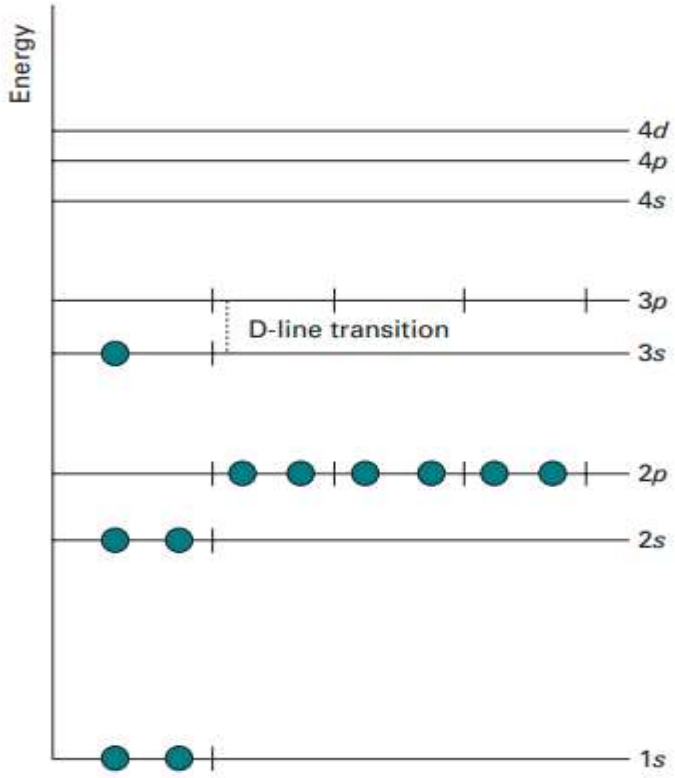


Figure20: Energy levels of atomic orbitals in the sodium atom. Each atomic orbital can be occupied by electrons following the rules of quantum chemistry until the total number of electrons for that element is reached (in case of sodium: 11 electrons). The energy gap between the 3s and the 3p orbitals in the sodium atom is such that it can be overcome by absorption of orange light. samples and standards, polyethylene bottles are used, since glass can absorb and release metal ions, and thus impact the accuracy of this sensitive technique. Cyclic analysis may be performed that involves the estimation of each interfering substance in a mixture. Subsequently, the standards for each component in the mixture are doped with each interfering substance. This process is repeated two or three times with refined estimates of interfering substance, until self-consistent values are obtained for each component. Flame instability requires experimental protocols where determination of an unknown sample is bracketed by measurements of the appropriate standard, in order to achieve the highest possible accuracy. Biological samples are usually converted to ash prior to determination of metals. Wet ashing in solution is often used, employing an oxidative digestion similar to the Kjeldahl method.

Applications

Atomic emission and atomic absorption spectrophotometry

Sodium and potassium are assayed at concentrations of a few p.p.m. using simple filter photometers. The modern emission spectrophotometers allow determination of about 20 elements in biological samples, the most common being calcium, magnesium and manganese. Absorption spectrophotometers are usually more sensitive than emission instruments and can detect less than 1 p.p.m. of each of the common elements with the exception of alkali metals. The relative precision is about 1% in a working range of 20–200 times the detection limit of an element. AES and AAS have

been widely used in analytical chemistry, such as environmental and clinical laboratories. Nowadays, the technique has been superseded largely by the use of ion-selective electrodes.

Atomic fluorescence spectrophotometry

Despite being limited to only a few metals, the main importance of atomic fluorescence spectrophotometry (AFS) lies in the extreme sensitivity. For example, zinc and cadmium can be detected at levels as low as 1–2 parts per 10^{10}. AFS uses the same basic setup as AES and AAS. The atoms are required to be vaporised by one of three methods (flame, electric, ICP). The atoms are excited using electromagnetic radiation by directing a light beam into the vaporised sample. This beam must be intense, but not spectrally pure, since only the resonant wavelengths will be absorbed, leading to fluorescence.

3.8. Plasma Emission Spectroscopy

The plasma is used as an atomisation source for emission spectroscopic method is known as plasma emission spectroscopy. Plasma is defined as a cloud of highly ionised gas which is composed of electrons. In plasma emission spectroscopy, the gas usually used is argon which is ionised by the influence of a strong electrical field either by a direct current or by radio frequency.Based on the current applied, plasma emission spectroscopy is classified as the following:

- Direct current plasma emission spectroscopy.
- Inductively coupled plasma emission spectroscopy.

Plasma was first identified by Sir William Crookes in 1879 in a Crookes tube.

PRINCIPLE

The main principle involved here is atomisation process. It is carried out by the following steps:

1. An aqueous solution in a flame gets *desolvated* and converted into a solid form.

$$M(H_2O)_n^+X^- \xrightarrow{\text{desolvation}} (MX)_n\,(\text{solid})$$

2. Then it is *vaporised*.

$$(MX)_n \xrightarrow{\text{vaporisation}} MX\,(\text{gas})$$

3. This obtained gas is subsequently heated for *atomisation*.

$$MX \xrightarrow{\text{atomisation}} M\,(\text{elemental})$$

4. And this element undergoes *ionisation*.

$$M \xrightarrow{\text{ionisation}} M^{+} \text{ (ion formation)}$$

5. On excitation, an electron moves from ground state to excited state which is having higher energy.

$$_{1}e^{0} + M^{+} \xrightarrow{\text{excitation}} M^{+} \text{(additional electron)}$$

6. This excited state is not a stable one so the electron returns to the ground state by emitting the energy.

$$M^+ \xrightarrow{\text{decay}} {}_{-1}e^0 + M^+ \xrightarrow{h\nu} \text{(atomic emission)}$$

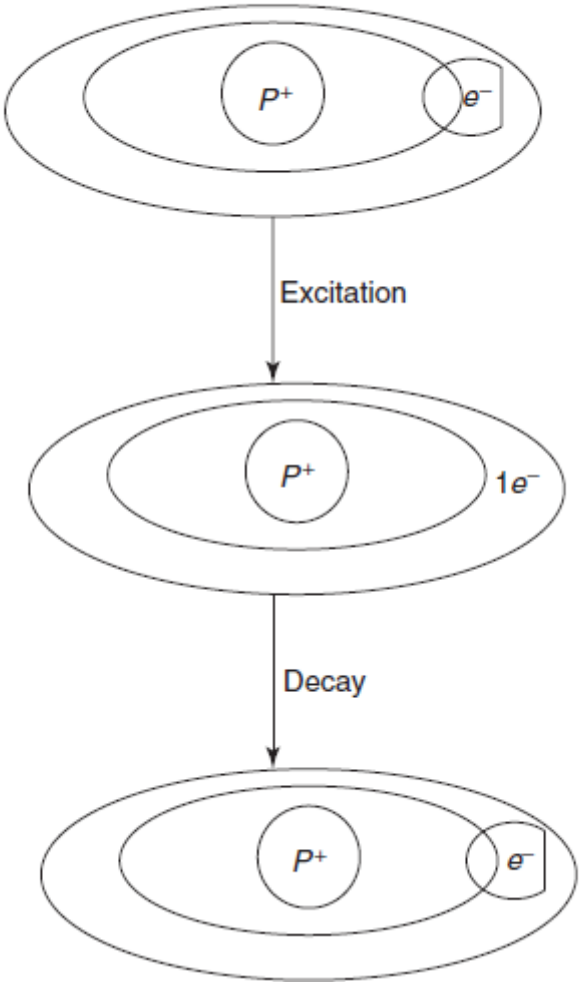


Diagram for the plasma formation

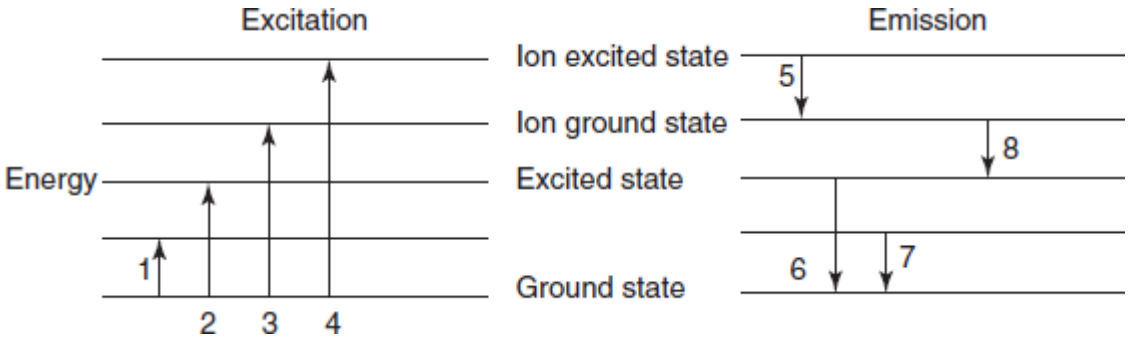
THEORY

A transition is given by the following equation:

$$E = hn$$

where E is the energy; h is Plank's constant; v is the frequency.

In plasma emission spectroscopy, there are no of transitions which can be depicted in the energy level diagrams.



Energy transitions in the ESR

Direct current plasma: The direct plasma source consists of high voltage discharge between the two electrodes. The recent developments are observed by incorporating the third electrode which is in inverted Y shape. The incorporation of the third electrode increases the stability of the discharge. The sample is carried the argon carrier gas. The argon is ionised before mixing with the sample solution. The argon gas is ionised by the high voltage discharge.

Advantages of Direct Current Plasma

- Less expensive
- Simple to handle

Disadvantages of Direct Current Plasma

- Low detection limits.
- Replacement of the electrodes is necessary after every use.

Inductively coupled plasma: This consists of the three concentric silica quartz tubes with opening at the top. The sample is mixed with the argon gas and is allowed for the formation of aerosol. This aerosol is again mixed with the argon gas and passed into the tubes from the bottom. The gas is excited by the radio frequency power. The plasma gas flows as the helical pattern which provides the stability of the plasma. Then the plasma is initiated by the spark from a tesla probe.

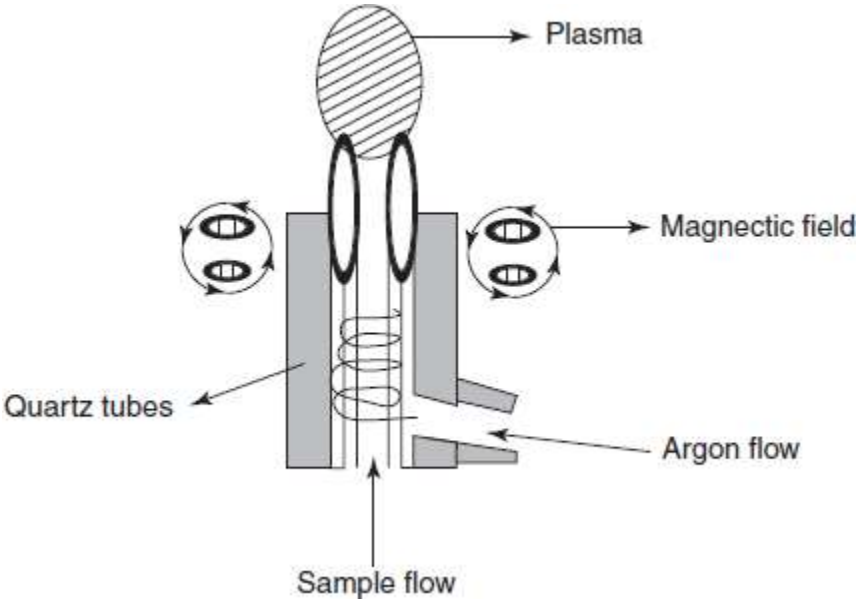
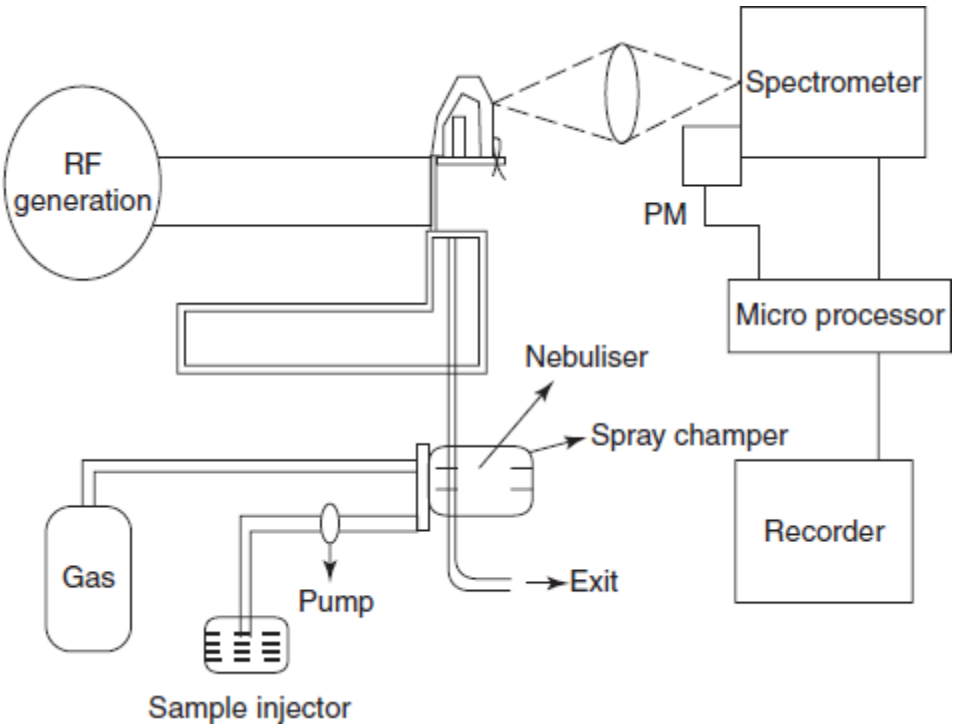


Diagram for the ICP source

INSTRUMENTATION

The instrument of the plasma emission spectroscopy contains the following components:

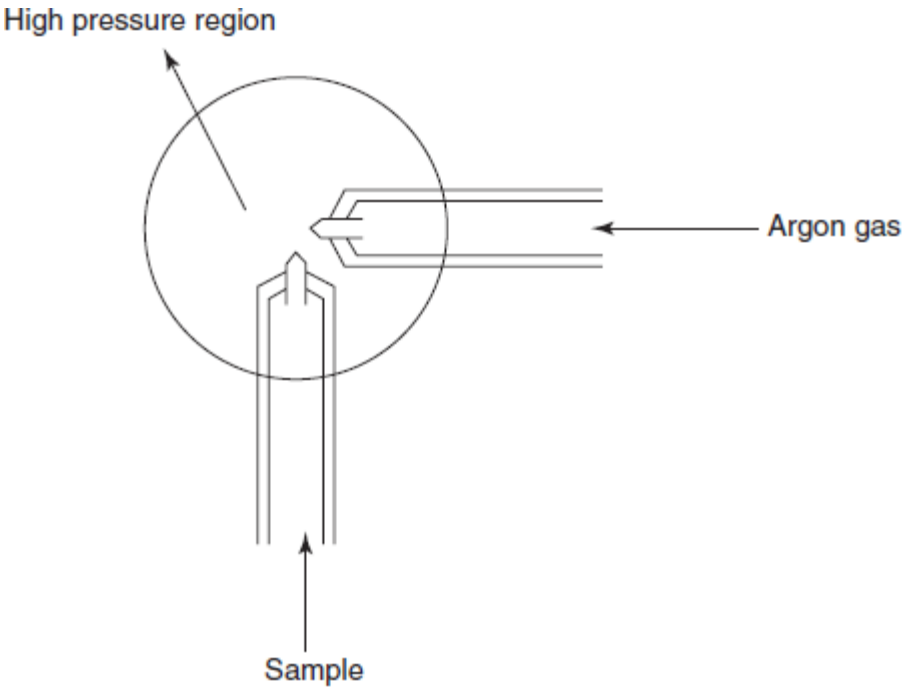
- Nebulisers
- Pumps
- Spray chamber
- Sample injector
- Torch
- RF generator
- Detector



Instrumentation of ESR

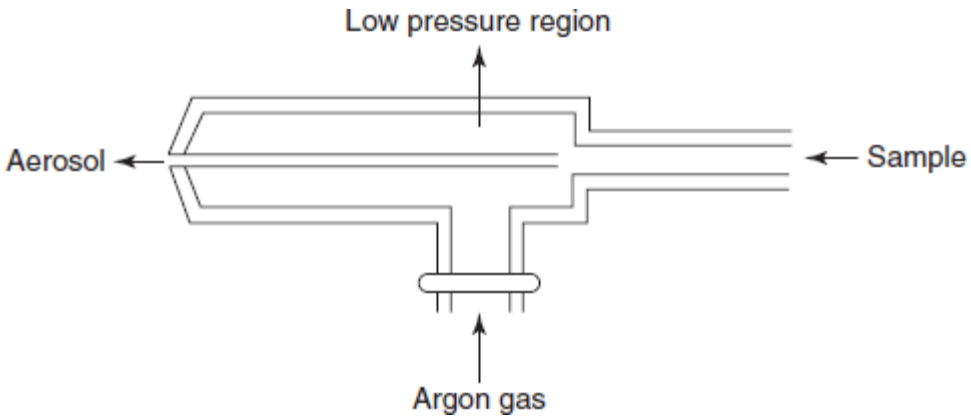
Nebuliser: These are used for the conversion of the liquid into aerosol. This aerosol is carried out into the plasma which is an excitation source. The commonly employed nebulisers are the following:

- **Pneumatic nebulisers**: The principle involved in this is the sample is subjected to the high-speed gas flow to form an aerosol.



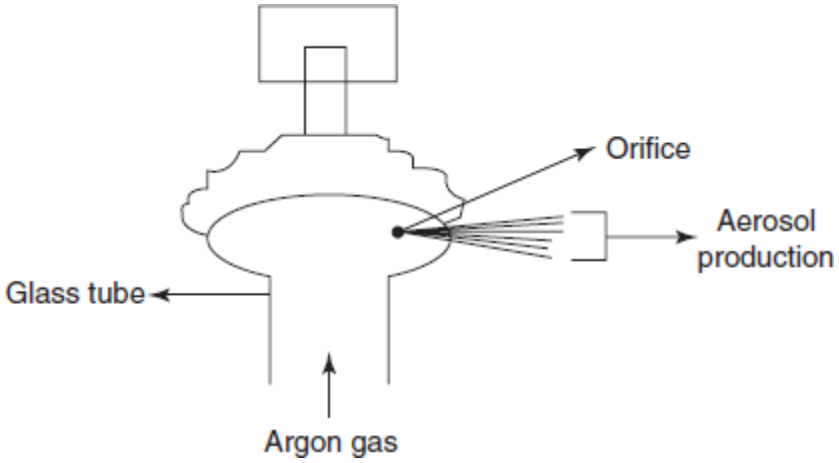
Pneumatic nebuliser

- **Concentric nebulisers**: The main principle involved in this is the sample is introduced through the capillary tube to a low-pressure region which produces the aerosol.



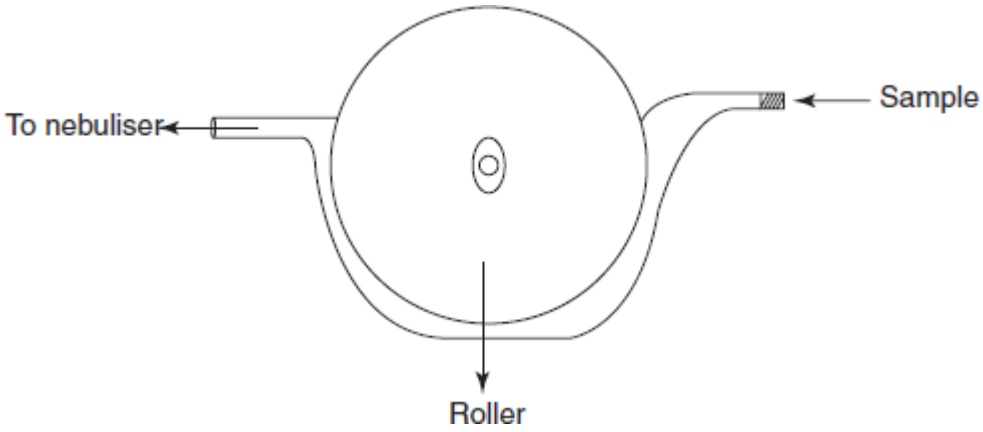
Concentric nebuliser

- **Babington nebuliser**: The principle involved in the liquid to flow over a smooth surface with a small hole in it. High-speed argon gas from the hole shears the sheet of liquid into small drops. This is very sensitive nebuliser.



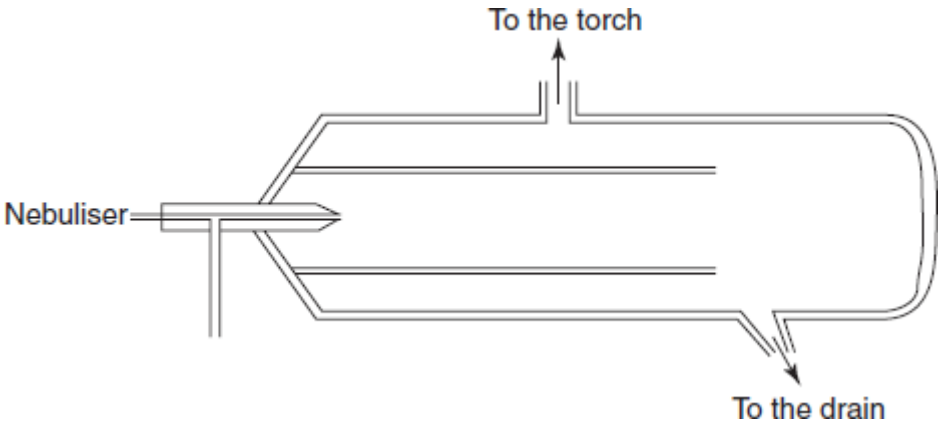
Babington nebuliser

- **Pumps**: These are mainly used to introduce the samples into the nebuliser. Commonly the peristaltic pumps are employed. These are composed of the rollers which pushes the sample solution through the tube. This process is known as the peristalsis.



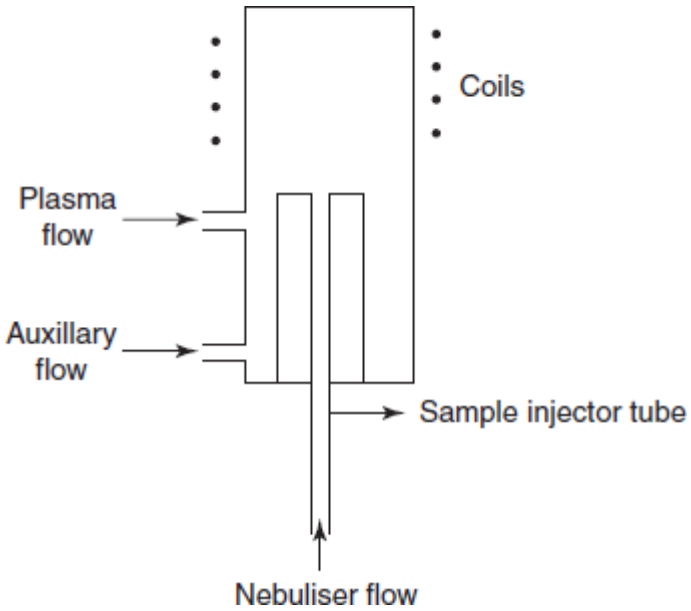
Peristaltic pump diagram

- **Spray chambers**: The spray chamber is used to transport the aerosol into plasma which is placed between the nebuliser and torch. This also helps in the removal of the large droplets. These must be made up of corrosion-resistant material.



Spray chamber diagram

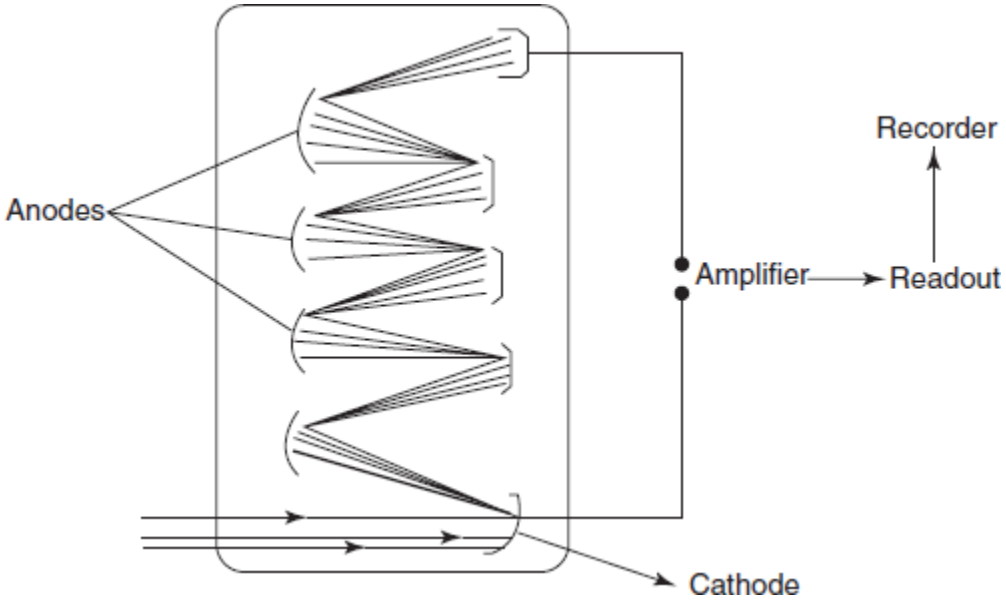
- **Drains**: This is mainly used to drain the sample from the spray chamber to waste container.
- **Torch**: This is mainly used to produce the plasma. This contains the three concentric tubes for argon gas flow and the sample aerosol injection. The gas flow carries the sample aerosol injected into the plasma through the sample injector tube.



Torch or plasma source diagram

- **RF generator**: This is mainly used for the generation and sustainment of the plasma discharge. It ranges from the 700 to 1500 watts. The commonly used RF generators are the following:
    1. **Crystal controlled generators**: This consists of the piezoelectric quartz crystal to produce an RF oscillating signal.
    2. **Free running generators**: This operates at an oscillation frequency.

- **Detectors**: The photomultiplier tubes are generally employed as the detectors. The principle involved in the detector is the secondary emission of the electrons when the light falls on the cathode and the replication of the anodes. The secondary emission of the electrons produces the current signals.



Photomultiplier tube detector diagram

ADVANTAGES

- High resolution
- Low stray light
- Wide dynamic range
- High accuracy
- High precision
- Highly reproducible

LIMITATIONS

- Time consuming
- Temperature dependent
- Less stability

APPLICATIONS

- Used in the trace metal analysis.
  Example: Copper, iron, manganese, magnesium and calcium
- Used in the estimation of the aluminium in blood.
  Example: Estimation of traces of the aluminium in blood
- Used in the estimation of the Cu in brain tissue.
  Example: Used in the scanning of the brain tissue incase of brain tumor
- Used in the estimation of Na in the breast milk.
  Example: Salts estimation in the breast milk
- Used in the analysis of agricultural products.
  Example: Pesticidal residues analysis
- Used in the study of the earth sciences.
- Used in the steel analysis.
  Example: Hardness of the steel
- Used in the alloys analysis.

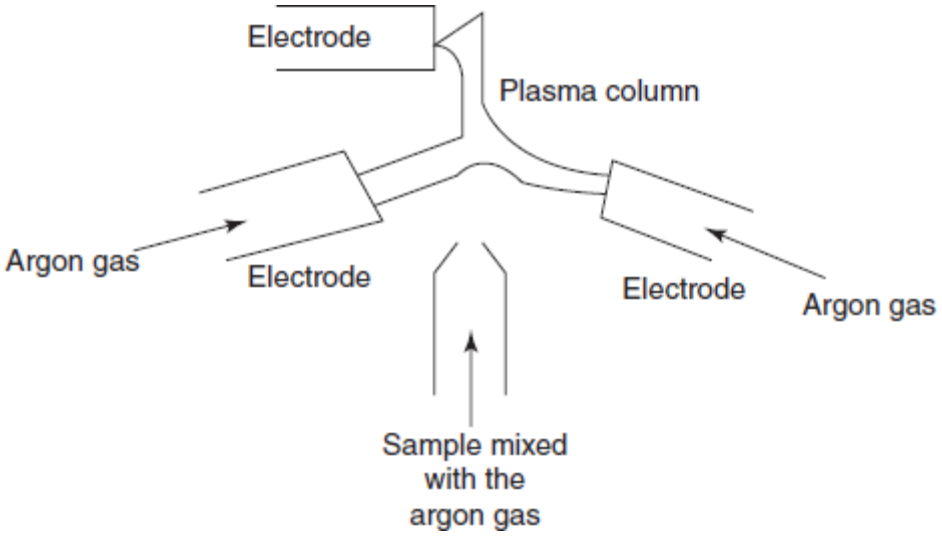


Diagram of the direct current plasma source

UNIT – IV -MOLECULAR TECHNIQUES FOR BIOTECHNOLOGY– SBT1002

4.1. NUCLEIC ACID HYBRIDIZATION

4.1.1. Basic Considerations of Nucleic Acid Hybridization:

Nucleic acid hybridization is a basic tool in molecular genetics which exploits the ability of single stranded nucleic acid molecules to hybridize with complementary sequences to form double stranded molecules. The principle of hybridization in general is the addition of a probe to a complex mixture of target DNA. The mixture is incubated under conditions that promote the formation of hydrogen bonds between complementary strands.

The factors that affect hybridization characteristics are:

i. Strand Length:

The longer the probe the more stable the duplex.

ii. Base Composition:

The % G:C base pairs are more stable than A:T.

iii. Chemical Environment:

The concentration of Na+ ions stabilizes.

4.1.2. Chemical denaturants (formamide or urea) destabilize hydrogen bonds.

Therefore, in nucleic acid hybridization, single-stranded nucleic acids (DNA or RNA) are allowed to interact so that complexes (hybrids) are formed by molecules with complementary sequences. Through nucleic acid hybridization, the degree of sequence identity between nucleic acids can be determined and specific sequences can be detected. The most important factor in nucleic acid hybridization is the high degree of base complementarity.

In nucleic acid hybridization assay systems, a labeled nucleic acid molecule (probe) is used to identify a complementary nucleic acid (DNA or RNA) molecule amongst a mixture of unlabelled nucleic acid molecules. The labeled nucleic acid molecule is known as probe and the nucleic acid sequence identified by labeled nucleic acid molecule (because of complementarity) is known as target nucleic acid.

Nucleic acid hybridizations can be done in all combinations-DNA-DNA (DNA can be rendered single-stranded by denaturation), DNA-RNA or RNA-RNA. In situ hybridization involves hybridizing a labeled nucleic acid to prepared cells or histological sections. This is used particularly to look for specific transcription/localization of genes to specific chromosomes via fluorescent in situ hybridization (FISH) analysis. Nucleic acid hybridization is used in cloning and PCR also.

4.1.3. Nucleic Acid Hybridization Method:

This method was developed in 1975 by Grunstein and Hogness. It is used to determine the colony which contains the sequence from hundreds of clones. In this method, bacterial colonies are transferred from the agar plate onto the nitrocellulose mem-brane and lysed by alkaline solution.

After the bacterial cell wall lysis, the proteins and DNA adhere to the nitrocellulose membrane because of the membrane's negative charge. Then the mem-brane is placed in proteinase K solution to remove the protein bound to the membrane and DNA. Then DNA is exposed to UV rays so that the DNA gets fixed onto the nitrocellulose membrane. After baking, the membrane is exposed to labeled RNA for hybridization. Then RNA-DNA hybrids are monitored by autoradiography. A colony which gives a positive result can then be picked from the master plate.

In general, nucleic acid hybridization method involves the following steps:

1. A nucleic acid probe is needed which will anneal to the target nucleic acid,

2. The target is attached to a solid matrix e.g., membrane,

3. Both the probe and target are denatured,

4. Denatured probe is added to the target in a solution,

5. If there is sequence homology between the target and the probe, the probe will hybridize/anneal to the target, and

6. The hybridized probe is the detected by autoradiography, chemiluminescence or colorimetric.

A variety of techniques utilize hybridization of DNA or RNA probes**Such as:**

i. ASO,

ii. Southern Blot,

iii. RFLP,

iv. VNTRs,

v. Northern Blot,

vi. Tissue specific expression, and

vii. In situ hybridization.

a. Chromosome location and integrity.

b. Tissue specific expression.

4.1.4. Hybridization Probes:

A hybridization probe is a nucleic-acid (DNA or RNA) fragment that is complementary to another nucleic-acid sequence and thus, when labeled (with radioisotope, fluorescent dye, etc.) can be used to identify complementary segments present in the nucleic-acid sequences of various

micro-organisms. The probe actually hybridizes to single-stranded nucleic acid (DNA or RNA) molecules because of complementarity between the probe and target.

The labeled probe is first denatured into single DNA strands and then hybridized to the target DNA (Southern blotting) or RNA (northern blotting) immobilized on a membrane or in situ. To detect probe hybridization to its target sequence, the probe is labeled with a molecular marker (radioactive -32P, a radioactive isotope of phosphorus incorporated into the phosphodiester bond in the probe DNA; or fluorescent molecules – Digoxigenin).

Nucleic acid probes are synthesized in the laboratory. Nucleic acid probes can be synthesized as single stranded or double stranded probes, but a working nucleic acid probe should be single stranded only (so as to bind with complementary target sequence).

Nucleic acid hybridization probes are of three types:

1. DNA Probe:

A single-stranded DNA molecule is used in laboratory experiments to detect the presence of a complementary sequence among a mixture of other singled-stranded nucleic acid molecules. Therefore, DNA probe is a short sequence of DNA labeled isotopically or chemically that is used for the detection of a complementary nucleotide sequence.

2. RNA Probe [Riboprobes]:

A single-stranded RNA molecule is used in laboratory experiments to detect the presence of a complementary sequence among a mixture of other singled-stranded nucleic acid molecules. Therefore, RNA probe is a short sequence of RNA labeled isotopically or chemically that is used for the detection of a complementary nucleotide sequence.

Single-stranded RNA probes also called complementary RNA (cRNA) or riboprobes and are often used for in situ hybridization because of high sensitivity; RNA-RNA hybrids are more stable than DNA-RNA hybrids; and non-specific tissue signals can be removed after hybridization with RNase A since RNA duplexes are resistant to degradation by RNase A.

3. Oligonucleotide Probe:

Oligonucleotide probe is a short sequence of nucleotides synthesized to match a region where a mutation is known to occur and then used as a molecular probe to detect the mutation.

Oligonucleotide probes are synthesized chemically in the laboratory. Initially a mononucleotide is bound to a solid support and then more mononucleotides are added one by one to the 3'end of the membrane bound mononucleotide. Oligonucleotides are generally labeled by adding a marker molecule (radioactive or fluorescent dye) at the 5'end.

Labeling of Hybridization Probes:

Hybridization probes can be labeled by two methods:

1. In Vivo Labeling:By supplying labeled nucleotides to the cultured cells.

2. In Vitro Labeling: An enzyme is used to incorporate a labeled nucleotide in the probe.

(a) Strand Synthesis Labeling for DNA Probes:

(i) Labeling by nick translation.

(ii) Random primed labeling.

(iii) PCR mediated labeling.

(b) Strand Synthesis Labeling for RNA Probes:

(i) In vitro transcription from cloned DNA inserts.

(ii) 3′-end labeling RNA.

(c) End labeling for oligonucleotide probes:

(i) Kinase end labeling.

(ii) Fill-in end labeling.

(iii) Primer mediated 5'end labeling.

(a) Strand Synthesis Labeling for DNA Probes:

Strand synthesis labeling is the most common labeling method for DNA. In this method DNA polymerase enzyme is used to incorporate labeled nucleotides in the DNA copies of the starting DNA. Any one of the nucleotides is labeled and is then incorporated in the DNA synthesis reaction mixture.

Labeling of DNA is generally done by any one of the following methods:

(i) Labeling by Nick Translation:

Nick translation is a procedure for making a DNA probe in which a DNA fragment is treated with DNase to produce single-stranded nicks, followed by incorporation of labeled nucleotides from the nicked sites by DNA polymerase I. This process is called nick translation because the DNA to be processed is treated with DNase to produce single-stranded "nicks."

This is followed by replacement in nicked sites by DNA polymerase I, which elongates the 3′ hydroxyl terminus, removing nucleotides by 5′-3′ exonuclease activity, replacing them with labeled dNTPs.

This nick is then sealed by DNA ligase. If the reaction is carried at 15°C (low temperature), then it will renew only one copy of the original nucleotide strand. The advantage of this method is that it allows the control of probe length which is important in in-situ hybridization applications.

(ii) Random Primed Labeling:

The method of "random primed" DNA labeling was introduced by Feinberg and Vogelstein in 1983. This method is based on the hybridization of oligonucleotides of all possible sequences to the denatured template DNA to be labeled. The complementary DNA strand is synthesized by a "Klenow" fragment of DNA polymerase I, using the random oligonucleotides as primers.

A labeled nucleotide is added in the reaction mixture. Therefore, the newly synthesized complementary DNA will be labeled. The reaction mixture will contain random oligonucleotides, a Klenow fragment of DNA polymerase I, dATP, dGTP, dTTP. This rapid labeling is accomplished with the use of the Klenow fragment, which lacks 5-3′ exonuclease activity, and by the use of hexamer/nonamer primers giving more efficient priming from the template at 37°C.

Random primed labeling method produces labeled DNA of high specific activity.

The advantages of this method are:

1. It produces very sensitive probes.
2. It can be scaled up tenfold.
3. It can label templates of almost any length.

(iii) **PCR Mediated Labeling:**

Amplification of DNA using the polymerase chain reaction (PCR) provides the opportunity to label the resulting product with either modified nucleotides or oligonucleotide primers. Therefore, we can incorporate either the labeled nucleotides in the reaction mixture or the labeled primers in the PCR reaction mixture which will result in the production of labeled PCR product throughout its length. PCR labeling produces a very high yield of labeled probe from very little template.

The advantages of this method are:

i. It requires only a small amount of template,

ii. Impure templates can be used,

iii. It requires less optimization than other methods,

iv. Has high yield of labeled probe,

v. It is recommended for very short probes (< 100 bp), and

vi. It produces very sensitive probes.

(b) **Strand Synthesis Labeling for RIMA Probes:**

Strand synthesis labeling is a common labeling method for RNA. In this method RNA polymerase enzyme is used to incorporate labeled nucleotides in the RNA copies of the starting RNA. Anyone of the nucleotides is labeled and is then incorporated in the RNA synthesis reaction mixture.

Labeling of RNA is done in dark by:

(i) In Vitro Transcription from Cloned DNA Inserts:

RNA probes are prepared by in vitro transcription. The RNA probe is transcribed from a linear DNA template using bacteriophage DNA-dependent RNA polymerases from the Salmonella bacteriophage SP6, and the E. coli., bacteriophages T3 and T7 (RNA polymerase T7, T3 or SP6).

We should have sufficient quantities of a plasmid carrying the gene sequence of interest that can be used as the template for RNA probe synthesis. Before a riboprobe can be transcribed, the correct RNA polymerase promoter sequences must be available in the plasmid in the correct orientation with respect to the template sequence. The transcription vectors pGEM (SP6 and T7 promoters) and pBluescript (T3 and T7 promoters) are commonly used. For in vitro transcription reactions the plasmid must also be in a linear form.

We use restriction enzymes to linearize the plasmid. For example, the plasmid vector (pSP64) contains SP6 promoter sequence adjacent to multiple cloning sites (MCS). Then transcription is initiated from a specific start point in the promoter sequence transcribing the DNA sequence that has been inserted in the multiple cloning sites (MCS). Labeled sense and anti-sense riboprobes can be generated and are widely used in tissue in situ hybridization.

The advantages of this method are:

i. It generates large amounts of probe,

ii. Labeled probe is completely free of vector sequences,

iii. RNA probes are single-stranded and cannot re-anneal to an opposite strand,

iv. RNA probes are more sensitive than DNA probes for analyzing northern blots, and

v. DIG-labeled RNA probes can easily be fragmented for in situ hybridization.

(ii) 3′-End Labeling RNA:

A short DNA template is designed to anneal to the 3′-end of the RNA, with a two nucleotide 5′ overhang of 3′-TA-5′, 3′-TG-5′ or 3′-TC-5′. The Klenow fragment of DNA polymerase I can then cleanly and efficiently extend the 3′-end of the RNA by the incorporation of a single a-32P-labeled dATP residue. This method can be used to label one RNA in a mixture of RNAs, or to label 5′-blocked RNAs such as mRNA.

(c) End Labeling for Oligonucleotide Probes:

Oligonucleotide probes are sensitive enough to detect single copy gene sequences in complex genomes if sufficient target DNA (e.g., 10 µg human genomic DNA) is present on the blot. For some applications, such as in situ hybridization, a labeled synthetic oligonucleotide is the best hybridization probe. In addition to in situ hybridizations, labeled oligonucleotides may be used as hybridization probes.

i. Dot/slot blots,

ii. Library screening,

iii. Detection of repeated gene sequences on Southern blots, and

iv. Detection of abundant mRNAs on Northern blots.

Advantages of oligonucleotide probes are:

i. Oligonucleotides are small and diffuse readily into cells, making them ideal for in situ hybridization applications.

ii. Small probe hybridizes to target faster, so hybridization times are reduced.

iii. Oligonucleotide probes are single-stranded, so they cannot self-hybridize (re-nature) during the hybridization reaction.

iv. Oligonucleotide sequences can be custom synthesized, so target recognition is optimized.

Methods used for labeling of oligonucleotides are:

(i) Kinase end labeling.

(ii) Fill-in end labeling.

(iii) Primer mediated 5'end labeling.

(i) Kinase End Labeling:

This is the most common method for oligonucleotide labeling. Polynucleotide kinase (therefore named kinase end labeling) is used to add 32P at the gamma-phosphate position of ATP and the polynucleotide catalyses an exchange reaction with the 5′-terminal phosphate.

(ii) Fill-in End Labeling:

Initially, the selected DNA is digested with such restriction endonucleases that after digestion 5′-overhangs are produced (e.g., EcoRI). These 5′-overhangs will act as primer and polymerase activity of Klenow fragment (of E.coli DNA polymerase I) is used to fill in the recessed ends.

Any one of the nucleotides provided in the reaction mixture is labeled. Hence, we get labeled oligonucleotide at 3′-end. Another specific restriction endonuclease can be used to cleave internal site so as to get fragments of different sizes which can be size fractionated.

The advantages of this method are:

i. Requires only a small amount of template,

ii. Labeled probes can be used without purification, and

iii. Reaction can be scaled up indefinitely (if incubation time is increased to 1 hour).

(iii) Primer Mediated 5′-End Labeling:

In primer mediated 5′-end labeling, a primer attached to a labeled group at its 5′-end is used in the normal PCR reaction. As PCR amplification proceeds, the 5′-end labeled primer is incorporated in the amplified PCR product.

The advantages of this method are:

i. Requires only a small amount of template.

ii. Produces more sensitive probes than end labeling.

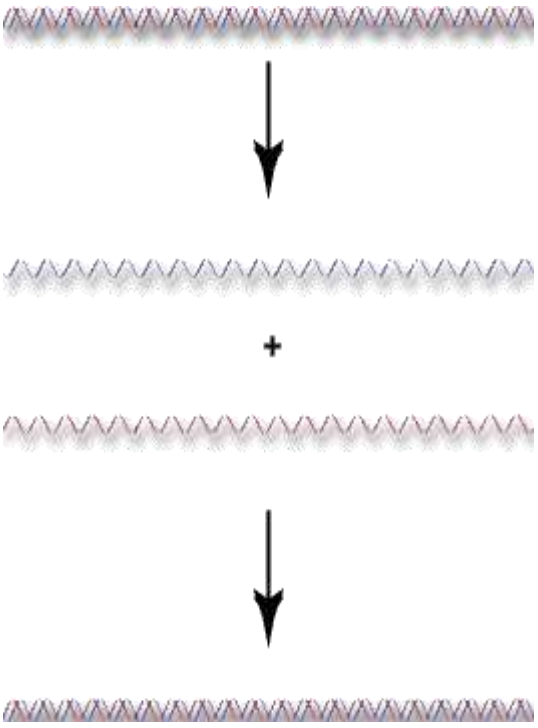
iii. Labeled probes can be used without purification.

iv. Reaction can be scaled up indefinitely.

4.1.5. Cot Curve

Cot analysis was first developed and utilized in the mid 1960s by Roy Britten, EricDavidson, and associates. It is based upon the principles of DNA renaturation kinetics.

What happens if we cool our denatured DNA sample and allow the two complementary strands to reanneal, renature, hybridize…



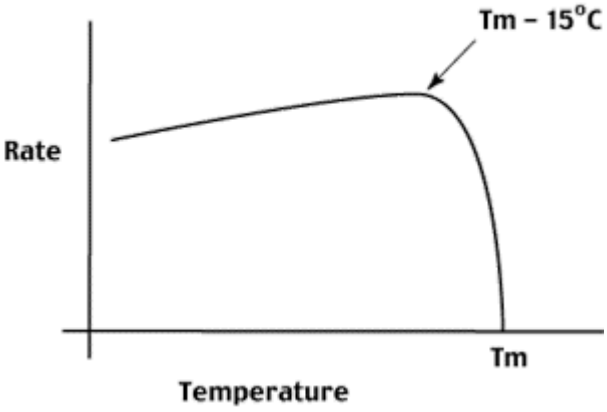
The **melting** of DNA into two single complementary strands is a **mono-molecular** reaction. As a result, the **rate** at which double-stranded DNA molecules denature is independant of DNA concentration.

In contrast, the reannealing of two single strands is a bi-molecular reaction.Two complementary single strands must meet one another and then complementary base pairs form between the two strands.

Because this is a bi-molecular reaction, the rate of the reaction depends on the concentration of the reactants - the two complementary strands.

Because reannealing is a bimolecular reaction, we are primarily interested in the rate at which the reaction occurs.

Reannealing reaction conditions are therefore chosen to MAXIMIZE the rate at which hybrids form.



The curve to the right shows how the rate of reannealing depends on temperature.To maximize the rate of reannealing, we typically hybridize 15 ^{o}C below the Tm of the hybrid. Hybridization is also done at high salt to minimize the repulsion of the sugar-phosphate backbones.

Hybridization conditions are always set to maximize the rate of hybrid formation. The rate at which hybridization occurs is therefore a direct measurement of the concentration of the complementary strands in solution.

In order to visualize this concept, consider the reannealing of a variety of DNA samples.

To standardize conditions,

Each DNA sample is sheared to random fragments 500 bp long.

Each DNA sample is at the same DNA (50 ug/ml) and NaCl concentration(1M).

Each sample is heated to boiling to denature it and then held at (Tm-15 oC) while the amount of DNA remaining single stranded is monitored.

The % single stranded is graphed as a function of

Cot - the initial DNA concentration (Co) times time (t)

Since the reaction follows bimolecular kinetics, we can obtain a value

$Cot_{1/2}$ = Cot at which 1/2 of the DNA has reannealed

$Cot_{1/2}$ measures the rate of reannealing which is inversely proportional to the concentration of complementary sequences being examined.

Small $Cot_{1/2}$ values indicate that the complementary sequences are at high concentration.

(they reanneal very quickly, t is small)

Large $Cot_{1/2}$ values indicate that the complementary sequences are at low concentration.

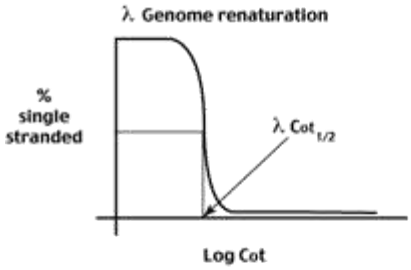
(they reanneal very slowly, t is large)

The following examples are ment to illustrate "Cot analysis

The lambda genome consists of 50,000 bp of unique sequence.

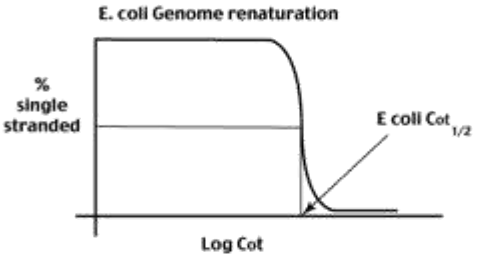
The DNA reanneals as a single kinetic component (a single sigmoidal curve)

The midpoint of the curve provides the value for $Cot_{1/2}$ - the time required for half of the single strands to be in double stranded form at a defined initial DNA concentration.



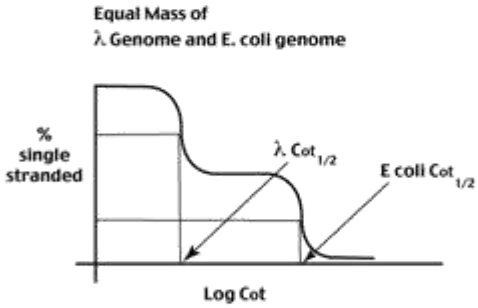
In contrast to the above example, the E. coli genome consists of 5,000,000 bp of unique sequence.

Again, plotting the % single stranded as a function of Cot, we observe a single sigmoidal curve confirming that the E coli genome reanneals as a single kinetic class.



Notice that the Cot curve has shifted to the right relative to the lambda Cot curve. The increase in $Cot_{1/2}$ reflects the number of copies of the E coli genome in solution as compared to the lambda genome. Since the E coli genome is 2 orders of magnitude larger than the lambda genome, any given 500 bp fragment will be present at 100 x lower concentration in the E coli genome sample.

This can be illustrated by considering what happens when the DNA sample consists of equal mass amounts of the lambda and E coli genomes (each at 25 ug/ml)

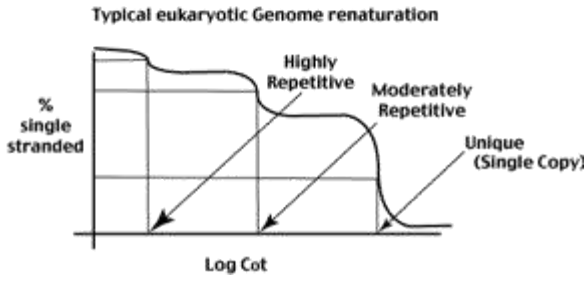


The Cot curve now has two kinetic components,

each comprising about 1/2 of the total DNA in the sample,

one with a $Cot_{1/2}$ of the lambda genome,

one with a $Cot_{1/2}$ of the E coli genome.



What about eukaryotes? What does the reannealing of eukaryotic genomes look like?

Unlike most prokaryotes, the Cot curves of eukaryotic genomes are complex curves. Modeling based on bimolecular kinetics allows us to 'fit' the data to define three kinetic classes that differ in their repetition frequency in the genome.

The first class represents a small portion of the genome (typically 10% or so) but is very highly repeated (10,000s of copies). These are short highly clustered repeated sequences found at eukaryotic telomeres and centromeres. Dispersed copies of these simple sequence repeats are also common.

The second class is called moderately repetitive. This class contains longer sequences which are repeated 100s to 1000s of times in the genome. Ribosomal RNA genes, histone genes and a few others fit into this class. Another major component are the transposable elements - a class of sequence which has the capacity to replicate and move to new positions in the genome. The portion of the genome in this portion of the Cot curve varies widely.

The third kinetic class is the unique sequence of the genome. In this class we find most the the protein coding sequences of the genome.

4.2. Protein Sequencing Methods

Protein sequencing denotes the process of finding the amino acid sequence, or primary structure of a protein. Sequencing plays a very vital role in Proteomics as the information obtained can be used to deduce function, structure, and location which in turn aids in identifying new or novel proteins as well as understanding of cellular processes. Better understanding of these processes allows for creation of drugs that target specific metabolic pathways among other things. Though several methods exist to sequence proteins the two dominant methods are Mass Spectrometry and Edman Degradation. Other methods that are not as frequently used still can serve very specific roles, such as overcoming inadequacies or acting as a preliminary, that compliment the two predominant methods.

History The advent of protein sequencing can be traced to two almost parallel discoveries by Frederick Sanger and Pehr Edman. Pehr Edman began his work in the Northrop-Kunitz laboratory at the Princeton branch of the Rockefeller Institute of Medical Research in 1947 where he attempted to find a method to decode the amino acid sequence of a protein using chemicals; specifically he had early success with fluorodinitrobenzene (FDNB) and phenylisothiocyanate (PITC). Throughout his

year at Princeton, Edman was able to conduct enough experiments to understand that it was feasible to use reagents like FDNB and PITC to determine amino acid sequence. Edman returned to Sweden in 1947 and after two more years of work he was able to publish his paper that would describe the first successful method to sequence proteins. This groundbreaking paper described a method to determine the amino acid sequence of a protein and would come to be known as the Edman Degradation. Five years earlier, Frederick Sanger had demonstrated a method to determine the amino acid residue located on the N-terminal end of a polypeptide chain by using the reagent fluorodinitrobenzene. While it was thought, that at most, this method could only provide the sequences found on the N-terminal, Sanger was able to take the method one step further. By using several proteolytic enzymes, partial hydrolysis and early version of chromatography, Sanger was able to cleave the protein into fragments and piece together the residues like a jigsaw puzzle. It wasn't until 1955 that Sanger was able to present the complete sequence of insulin which led to him being awarded a Nobel Prize in Chemistry in 1958. Mass Spectrometry, as a tool for the analysis of individual molecules, had been available many years before either Sanger or Edman began their work on protein sequencing. From its humble beginnings in the late 1800's it has undergone many changes in its hardware and its software and has proven to be so critically important to the field of sequencing that several more Nobel Prizes were awarded to those that were able to improve upon this technology. Despite its importance today, it wasn't until 1966 when K. Biemann, C. Cone, B.R. Webster, and G.P. Arsenault sequenced several oligopeptides containing glycine, alanine, serine, proline, and several other amino acids that the importance of mass spectrometry was fully realized. Further developments were to come in the late 80's as the Mass Spectrometer became a more robust piece of instrumentation in the laboratory. 1989 saw the first demonstration of fast atom bombardment ionization with Tandem Mass Spectroscopy as applied towards the identification of protein sequences. The work here was the early foundation for the Protein Mass Fingerprinting procedure that would come to use in the early 90's. With the arrival of MALDI and electrospray ionization as two new ionization methods the dynamic range of mass spectroscopy was greatly improved and paved the way for the mass spectrometer to be a dominant tool in the use of protein sequencing. The advent of Proteomics in 1996 saw many rapid developments such as increasing computational power, the growth of the world wide web and protein databases, and the advances in mass spectroscopy multi-quadrupole systems that enabled MALDI-MS/MS and other tandem mass spectrometer methods.

4.2.1. Sequencing Methods

N-terminal Residue Identification

N-terminal residue identification encompasses a technique which chemically determines which amino acid forms the N-terminus of a peptide chain. This information can be used to aid in ordering of individual peptide sequences that were generated using other sequencing techniques that fragment the peptide chain. Frequently, the first round of Edman Degradation will also contain impurities that may make identification of the N-terminus residue difficult. The general process of N-terminal residue identification is described below

1. The free unprotonated α-amino groups are labeled using a reagent that will selectively label the terminal amino acid. Reagents that can accomplish this include 2,4-dinitrofluorobenzene (DFNB - Sanger's reagent), dansyl chloride, and phenylisothiocyanate (Edman's reagent).
2. The labeled peptide is hydrolyzed with acid which yields the N-terminal residue and other free amino acids.
3. Each of these derivative N-terminal residues can be separated and identified using chromatography.

These methods can be used to identify the N-terminal residue of the peptide. This is time consuming process which has decreased in usefulness now that more efficient sequencing techniques are now available. Further complicating the issue: certain reagents used in the process can also degrade amino acid residues to the point where they are unrecognizable. Of the reagents available to label the N-terminal residue dansyl chloride is about 100 times more sensitive than FDNB due to its highly fluorescent nature which makes it easily detectable in minute amounts. The use of Edman's reagent is also advantageous as it leaves the remaining residues in the peptide chain untouched as described in the next section.

Sequencing steps

1. Splitting polypeptide chain

Protein moleculars should be separated and purified. Several polypeptides are combined together by non-covalent bond, which is known as oligomeric protein. For example. 8 mol/L urea or 6mol / L guanidine hydrochloride can be used to deal with tetramer—Hb and dimer—Enolase.

2. Detecting the number of polypeptide in protein moleculars

The number of polypeptides can be determined by detecting the relationship between the number of moles of amino acid residues and protein molecular weight.

3. Breaking disulfide bonds

Several polypeptides chains are linked by disulfide bonds. Disulfide bonds will be reduced to thiol with excessive & [beta]- mercaptoethanol under the condition of 8mol / L urea or 6mol / L guanidine hydrochloride. And then it should be protected by alkyl reagents from re-oxidation.

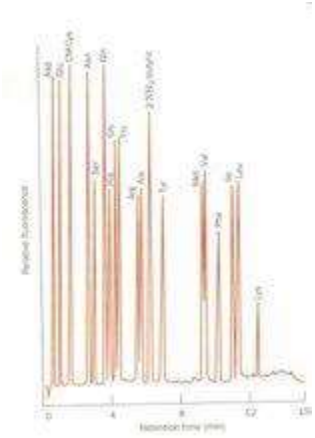
Cleaving and protecting disulfide bonds

A.performic acid: -CH2SO3H

B.Reduction + oxidation: Mercaptoethanol, DTT +iodoacetic acid, -S-CH2-COOH

C.Sulfurous acid decomposition: -R1-S-S-R2 + HSO3-, R1-S- + R2-S-SOH3

4.Detecting the amino acid composition of polypeptide chains and calculating the molecular ratio of amino acid composition.



5.Sequencing N-terminal and C-terminal of polypeptide chains

Amino acid of polypeptides is divided into two categories: amino-terminal and carboxyl-terminal. The N-terminal is much more important in the analysis of amino acid sequence of peptide chains than C-terminal.

6.Polypeptide can be cleaved into several small peptides. More than two methods can be used to break peptide samples into two or more sets of peptides or peptide fragments and then separate them.

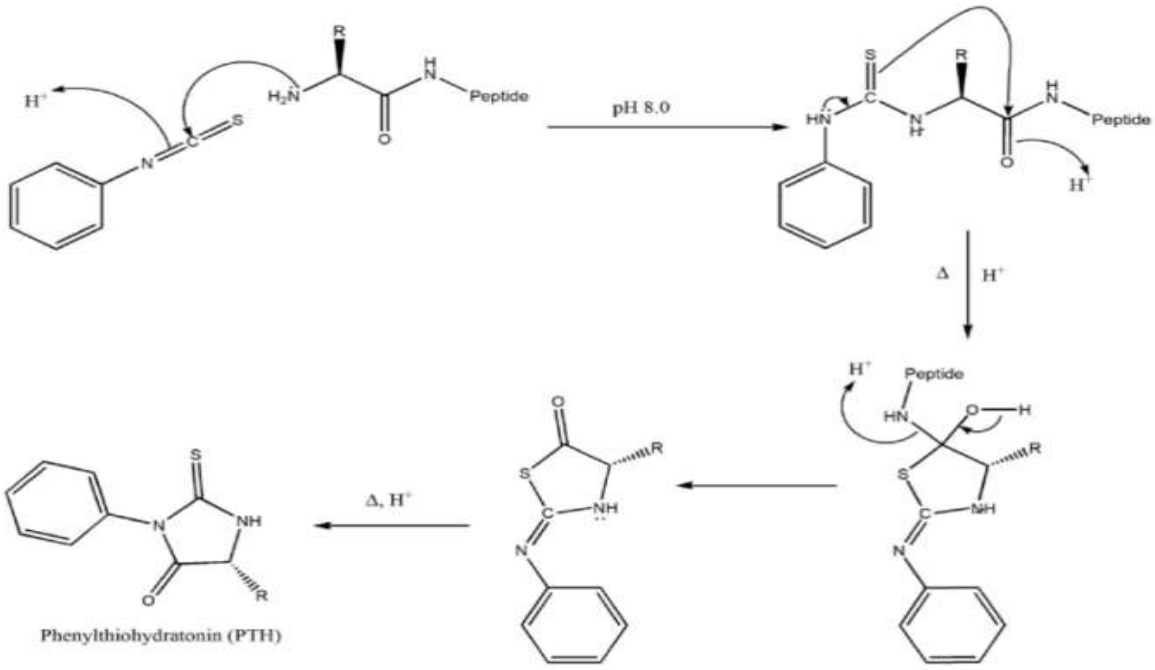
7.Determining the amino acid sequencing of each peptide

8.Determining the sequence of peptide fragments in polypeptide chains

9.Determining the position of disulfide bonds in the original polypeptide chain

Generally, pepsin will be used to deal with those polypeptide chains with disulfide bonds. And then 2D-electrophoresis technology will be used to separate each peptide fragment. Analyzing the composition and sequence of peptide fragments, which may contain disulfide bonds. And then comparing it with other peptide fragments, which are analyzed by other methods, to determine the position of disulfide bonds.

4.2.2. Edman Degradation



Edman Degradation with generic amino acid peptide chain.

The Edman Degradation method is based on the principal that single amino acid residues can be modified chemically such that they can be cleaved from the chain without disrupting the bonds between any other residues.The procedure can be achieved with very minute amounts of peptide, usually amounts on the order of 10-100 picomoles will allow for successful completion. Samples must contain only one protein component and should be free of any reagents that interfere with the degradation process such as glycine, glycerol, sucrose, guanidine, ethanolamine, ammonium sulfate, and ammonium salts. The general method is described below

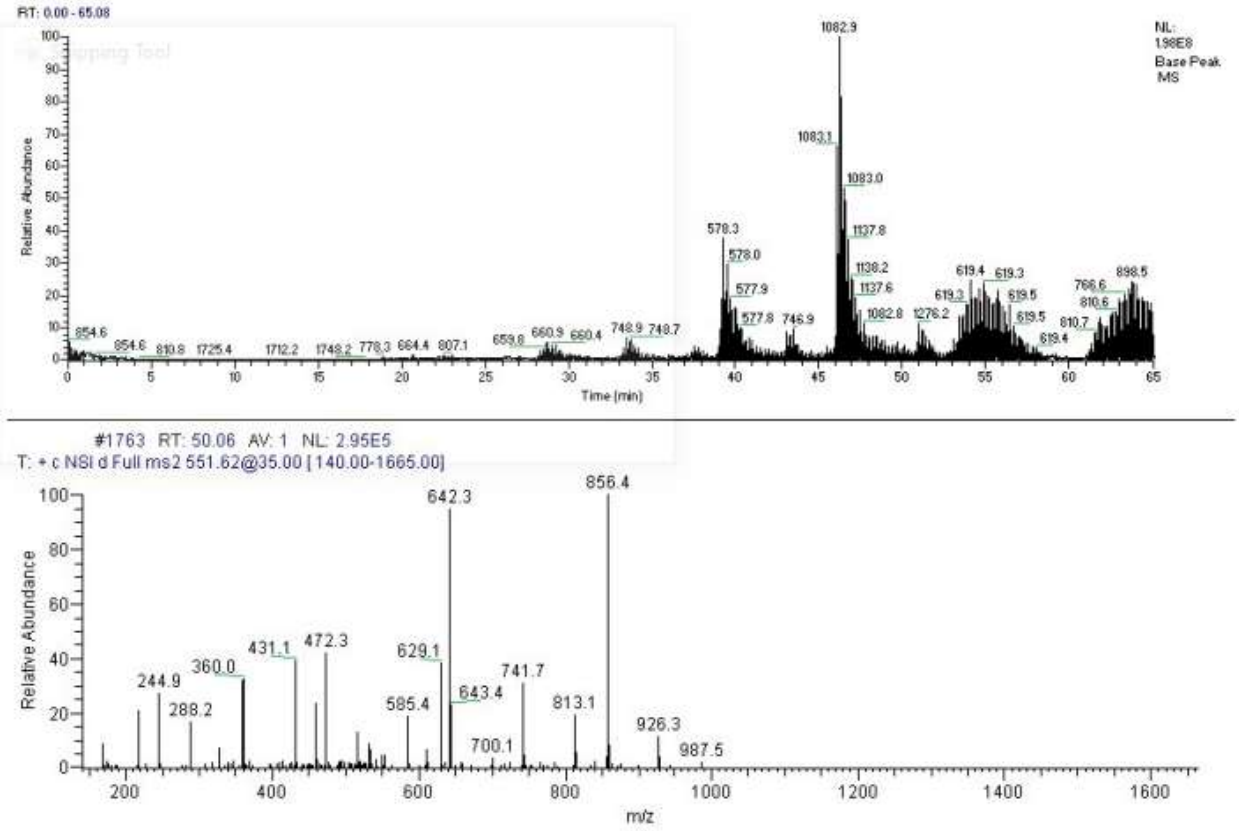
1. The peptides to be sequenced must first be immobilized by being absorbed onto a chemically modified glass or by electroblotting onto a porous polyvinylidene fluoride (PVDF) membrane.

2. Under mildly alkaline conditions phenylisothiocyanate (PITC) is reacted with an uncharged terminal group on the amino acid chain to form a phenylthiocarbamoyl derivative.
3. This phenylthiocarbamoyl derivative is then cleaved using Trifluoroacetic acid producing its anilinothiazolinone derivative (ATZ-amino acid). The next terminal amino acid is now exposed and ready for the same reactions to occur.
4. A wash is performed to remove excess buffers and reagents and the ATZ amino acid is selectively extracted with ethyl acetate and converted to a more stable phenylthiohydantoin (PTH)- amino acid derivative.
5. Identification of the PTH amino acid derivative is accomplished using chromatography or electrophoresis.
6. The process can now be repeated for the remaining residues of the chain.

Automation of the Edman Degradation procedure was initiated in 1967 and continues to be a favorable sequencing method due to its sensitivity and rapid completion. Sequencers that can automate the Edman Degradation procedure include many models of the Applied Biosystems Procise or Protein Sequencer families. The major drawback of the procedure remains the length of the peptide chain. If the chain exceeds a length of 50-60 residues (30 residues in practice) the procedure tends to fail due to the incompletion of the cyclical derivitization. This can be solved by taking the larger peptide chain and cleaving it into smaller fragments using cyanogen bromide, trypsin, chemotrypsin or any enzyme/chemical which can break peptide chains.

4.2.3. Mass Spectrometry

Mass spectrometry is quickly becoming the gold standard by which to identify protein sequences due to its ease of automation and extreme accuracy. The use of mass spectroscopy now dominates the process of sequencing proteins because prior problems of delivery were solved by John B. Fenn and Koichi Tanaka with their Nobel Prize winning electrospray ionization procedure. The two most popular methods to identify protein sequences using Mass Spectrometry are Peptide Mass Fingerprinting and Tandem Mass Spectrometry



Full MS and MS^2 spectra of a peptide.

Peptide Mass Fingerprinting

This method, also known as Protein fingerprinting, was developed in 1993 by several groups and functions by cleaving an unknown protein into smaller fragments so that these smaller fragments can then be accurately measured with a mass spectrometer. A generalized procedure is shown below:

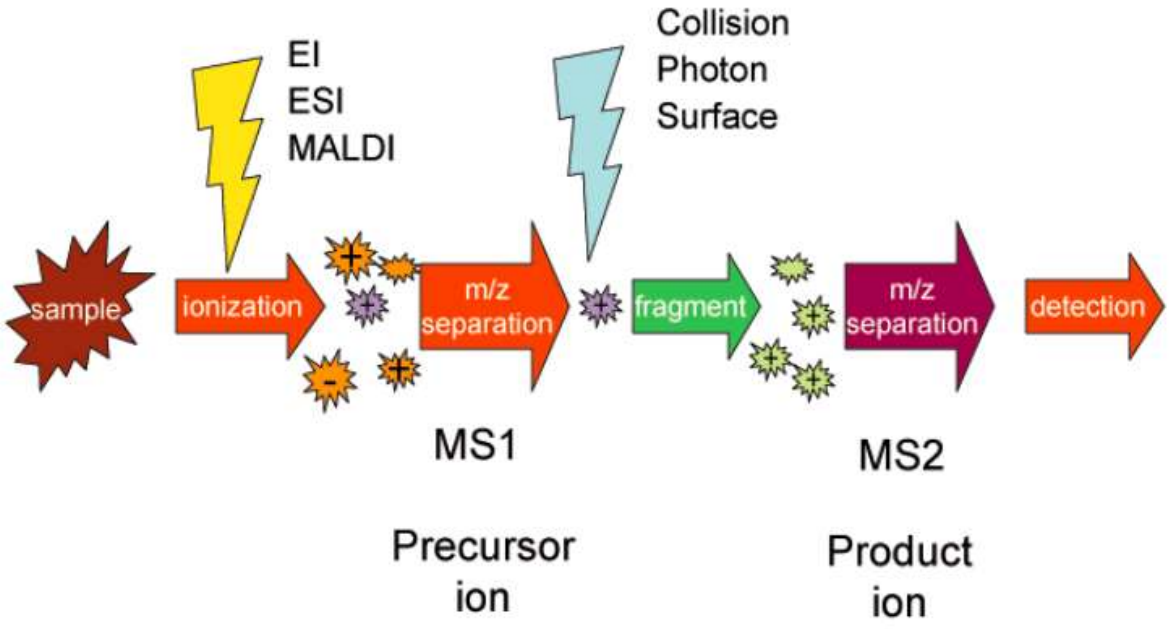
1. Protein samples are broken up into several smaller peptide fragments by proteolytic enzymes.
2. The resulting fragments are extracted using acetonitrile and dried by vacuum. The peptides are then dissolved in distilled water and ready for analysis.
3. The peptides are then inserted into the vacuum chamber of a mass spectrometer such as ESI-TOF or MALDI-TOF.

The mass spectrometer produces a peak list (i.e. a list of molecular weights), which is then compared against databases such as SwissProt or GeneBank to find close matches. Software is used to translate the retrieved genomic data into proteins which then undergo simulated cleavage by the same enzyme used to cleave the unknown protein. The mass is calculated of these fragments and then compared to that of the unknown protein.

Tandem Mass Spectrometry

Tandem Mass Spectrometry describes the partitioning of mass spectroscopy into separate steps where fragmentation occurs in between these steps. These separations can occur either

physically in space, by separate chambers called quadrupoles, using either multiple mass spectrometers or in a single mass spectrometer by time. The generalized procedure is described below



Schematic of tandem mass spectrometry

1. Enzymatic or chemical degradation of target protein to produce peptides.
2. Fractionation of peptides by high-performance liquid chromatography.
3. Resulting fragments fed into mass spectrometer for analysis.

Analysis of fragments by Tandem Mass Spectrometry occurs in two or more quadrupole systems with the first quadrupole filtering select ions that will undergo further analysis. These filtered ions are transferred to the second quadrupole which acts as a collision center to induce further fragmentation at amide linkages. A third quadrupole is then used to separate these fragments by mass. Tandem Mass Spectrometry mainly generates peptides of the N- and C- terminal types, which are represented in 2D via mass/charge vs. intensity graphs. The spectra produced by a mass spectrometer containing all the molecular weights of the fragments is called a peptide map and can serve as a means of identifying proteins analyzed by a mass spectrometer. Other approaches to identifying protein sequences include Protein Sequence tags and de novo methods. Peptide sequence tags, proposed by Matthias Wilm and Matthias Mann at the EMBL, function by sampling of masses at random points during a Tandem Mass Spectrometry experiment. These handfuls of masses are then used as unique tags to identify specific peptides following further fragmentation by the mass spectrometer. This aids not only in identification but also during the process of attempting to stitch back together the peptides into a full sequence. De novo approaches to protein sequencing identification are also employed along side similarity searches. These de novo methods do not take into account any prior knowledge of the amino acid sequence being analyzed and approach the identification of peptide sequences in novel ways. Examples include Hidden Markov Models, which takes a statistical approach to protein sequence identification, and graph searches of the problem space that helps minimize the search space as to speed up the time needed to identify sequences via a database.

4.3. DNA sequencing

DNA sequencing is the process of determining the sequence of nucleotides within a DNA molecule. Every organism's DNA consists of a unique sequence of nucleotides. Determining the sequence can help scientists compare DNA between organisms, which can help show how the organisms are related.

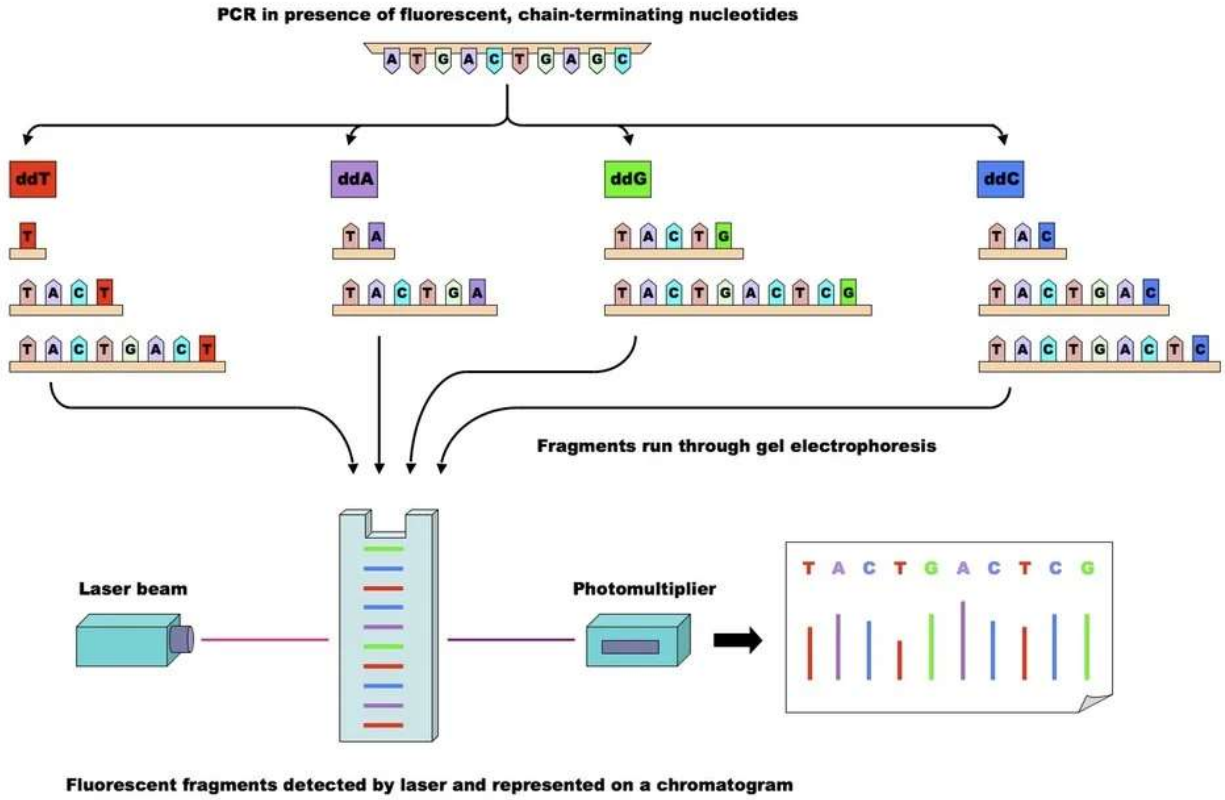
DNA sequencing is the determination of the precise sequence of nucleotides in a sample of DNA. Before the development of direct DNA sequencing methods, DNA sequencing was difficult and indirect. The DNA had to be converted to RNA, and limited RNA sequencing could be done by the existing cumbersome methods. Thus, only shorter DNA sequences could be determined by this method. Using this method, Walter Gilbert and Alan Maxam at Havard University determined that the Lac operator is a 27 bp long sequence.The development of direct DNA sequencing techniques changed the scope of biological research. The evolution of DNA sequencing technology from plus-minus sequencing to pyro-sequencing within about 20 years parallels the progress in biology from molecular biology to genomics.

The development of DNA sequencing techniques with enhanced speed, sensitivity and throughput are of utmost importance for the study of biological systems. Sequence determination is most commonly performed using di-deoxy chain termination technology. Pyro-sequencing, a non-electrophoretic real- time bio-luminometric method for DNA sequencing has emerged as a state of the art sequencing technology.

This technology has the advantage of accuracy, ease of use, and high flexibility for different applications. Pyro-sequencing allows the analysis of genetic variations including SNPs, insertion/deletions and short repeats, as well as assessing RNA allelic imbalance, DNA methylation status and gene copy number.

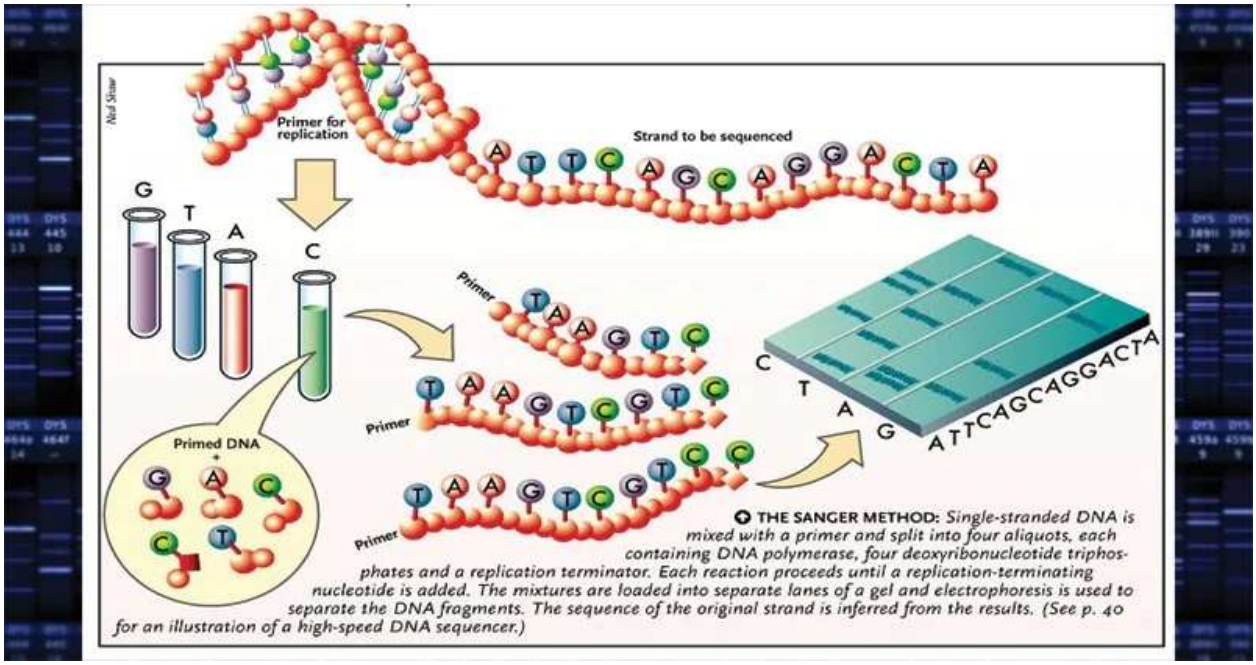
1. **Chain Termination Method (Sanger Dideoxy Method)**

- The chain terminator method is more efficient and uses fewer toxic chemicals and lower amount of radioactivity than the method of Maxam and Gilbert.
- The key principle of the Sanger method was the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators.
- The chain termination method requires a single-stranded DNA template, a DNA primer,a DNA polymerase, radioactively or fluorescently labelled nucleotides,and modified nucleotides that terminate DNA strand elongation.



Fluorescent fragments detected by laser and represented on a chromatogram

- The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP, dTTP) and the DNA polymerase.
- To each reaction is added only one of the four dideoxynucleotide (ddATP, ddGTP, ddCTP, ddTTP) which are the chain terminating nucleotides, lacking a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, thus terminating DNA strand extension and resulting in DNA fragments of varying length.
- The newly synthesized and labelled DNA fragments are heat denatured, and separated by size by gel electrophoresis on a denaturing polyacrylamide-urea gel with each of the four reactions run in one of the four individual lanes (lanes A, T, G, C).
- The DNA bands are then visualized by autoradiography or UV light, and the DNA sequence can be directly read off the X-ray film or gel image.
- A dark band in a lane indicates a DNA fragment that is result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP).
- The relative position of the different bands among the four lanes are then used to read (from bottom to top) the DNA sequence.
- The technical variations of chain termination sequencing include tagging with nucleotides containing radioactive phosphorus for labelling, or using a primer labelled at the 5' end with a fluorescent dye.
- Dye- primer sequencing facilitates reading in an optical system for faster and more economical analysis and automation.



Advantage

- Chain termination methods have greatly simplified DNA sequencing.

Limitations

- Non-specific binding of the primer to the DNA, affecting accurate read-out of the DNA sequence.
- DNA secondary structures affecting the fidelity of the sequence.

2. **Chemical Cleavage Method (Maxam–Gilbert Method)**
    - In 1976-1977, Allan Maxam and Walter Gilbert developed a DNA sequencing method based on chemical modification of DNA and subsequent cleavage at specific bases.
    - The method requires radioactive labelling at one end and purification of the DNA fragment to be sequenced.
    - Chemical treatment generates breaks at a small proportions of one or two of the four nucleotide based in each of four reactions (G,A+G, C, C+T).
    - Thus a series of labelled fragments is generated, from the radiolabelled end to the first ‘cut’ site in each molecule.
    - The fragments in the four reactions are arranged side by side in gel electrophoresis for size separation.
    - To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each corresponding to a radiolabelled DNA fragment, from which the sequence may be inferred.

Advantages

- Purified DNA can be read directly
- Homopolymeric DNA runs are sequenced as efficiently as heterogeneous DNA sequences
- Can be used to analyze DNA protein interactions (i.e. footprinting)
- Can be used to analyze nucleic acid structure and epigenetic modifications to DNA

Disadvantages

- It requires extensive use of hazardous chemicals.
- It has a relatively complex set up / technical complexity.
- It is difficult to "scale up" and cannot be used to analyze more than 500 base pairs.
- The read length decreases from incomplete cleavage reactions.
- It is difficult to make Maxam-Gilbert sequencing based DNA kits.

3. **Hybridization Method:**

Ed Southern's (1990) sequencing by hybridization technique relies on detection of specific DNA sequences using hybridization of complementary probes. It utilizes a large number of short nested oligonucleotides immobilized on a solid support to which the labeled sequencing template is hybridized. The target sequence is deduced by computer analysis of the hybridization pattern of the sample DNA.

DNA sequence can also be analyzed by sequencing by synthesis. Sequencing by hybridization makes use of a universal DNA microarray, which harbors all nucleotides of length k (called "k-words", or simply words when k is clear). These oligonucleotides are hybridize to an unknown DNA fragment, whose sequence one would like to determine.

Under ideal conditions, this target molecule will hybridize to all words whose Watson-Crick complements occur somewhere along its sequence. Thus, in principle, one would determine in a single microarray reaction the set of all k-long substrings of the target and try to infer the sequence from those data.

The average length of a uniquely resconstructible sequence using an 8-mer array is <200 bases, far below a single read length on commercial gel-lane machine. The main weakness of sequencing by hybridization is ambiguous solutions-when several sequences have the same spectrum; there is no way to determine the true sequence.

4. **Pal Nyren's Method:**

In 1996, Pal Nyren's group reported that natural nucleotide can be used to obtain efficient incorporation during a sequencing-by-synthesis protocol. The detection was based on the pyrophosphate (inorganic biphosphate) released during the DNA polymerase reaction, the quantitative conversion of pyrophosphate to ATP by sulfurylase and the subsequent production of visible light by firefly luciferase.

The first major improvement was inclusion of dATPaS in place of dATP in the polymerization reaction, which enabled the pyrosequencing reaction to be performed in homogeneous phase in real time.

The non-specific signals were attributed to the fact that dATP is a substrate for luciferase. Conversely, dATPaS was found to be inert for luciferase, yet could be incorporated efficiently by all DNA polymerases tested. The second improvement was the introduction apyrase to the reaction to make a four-enzyme system. Apyrase allows nucleotides to be added sequentially without any intermediate washing step.

Pyrosequencing nonelectrophoretic real-time DNA sequencing method is based on sequencing by synthesis based on the pyrophosphate (inorganic biphosphate) released during the DNA polymerase reaction. In a cascade of enzymatic reaction, visible light is generated that is proportional to the number of incorporated nucleotides. The cascade starts with a nucleic acid polymerization reaction in which inorganic bip-hosphate (PPi) is released as a result of nucleotide incorporation by polymerase.

The released PPi is subsequently converted to ATP by ATP sulfurylase, which provides the energy to luciferase to oxidize luciferin and generate light. The light so generated is captured by a CCD camera and recorded in the form of peaks known as pyrogram (compared with electropherograms in Sanger's method). Because the added nucleotide is known the sequence of template can be determined.

Standard pyrosequencing uses the Klenow fragment of E. coli DNA pol I, which is relatively slow polymerase. The ATP sulfurylase used in pyrosequencing is a recombinant version from the yeast and the luciferase is from the American firefly. The overall reaction from polymerization to light detection takes place within three to four seconds at real time.

One pmol of DNA in a pyrosequencing reaction yields 6 x 1011 ATP molecules which in turn, generate more than 6 x 109 photons at a wavelength of 560 nm. This amount of light is easily detected by a photodiode, photomultiplier tube or a CCD camera. Pyrosequencing technology has been further improved into array-based massively parallel microfluidic sequencing platform.

5. Automatic DNA Sequencer:

A variant of the above dideoxy-method was developed, which allowed the production of automatic sequencers. In this new approach, different fluorescent dyes are tagged either to the oligonucleotide primer (dye primers) in each of the four reaction tubes (blue for A, red for C, etc), or to each of the four ddNTPs (dye terminators) used in a single reaction tube: when four tubes are used, they are pooled.

After the PCR reaction is over, the reaction mixture is subjected to separation of synthesized fragments through electrophoresis (Fig. 23.9). Depending upon the electrophoretic system used, whether slab gel electrophoresis or capillary electrophoresis, following two types of automatic sequencing systems have been designed.

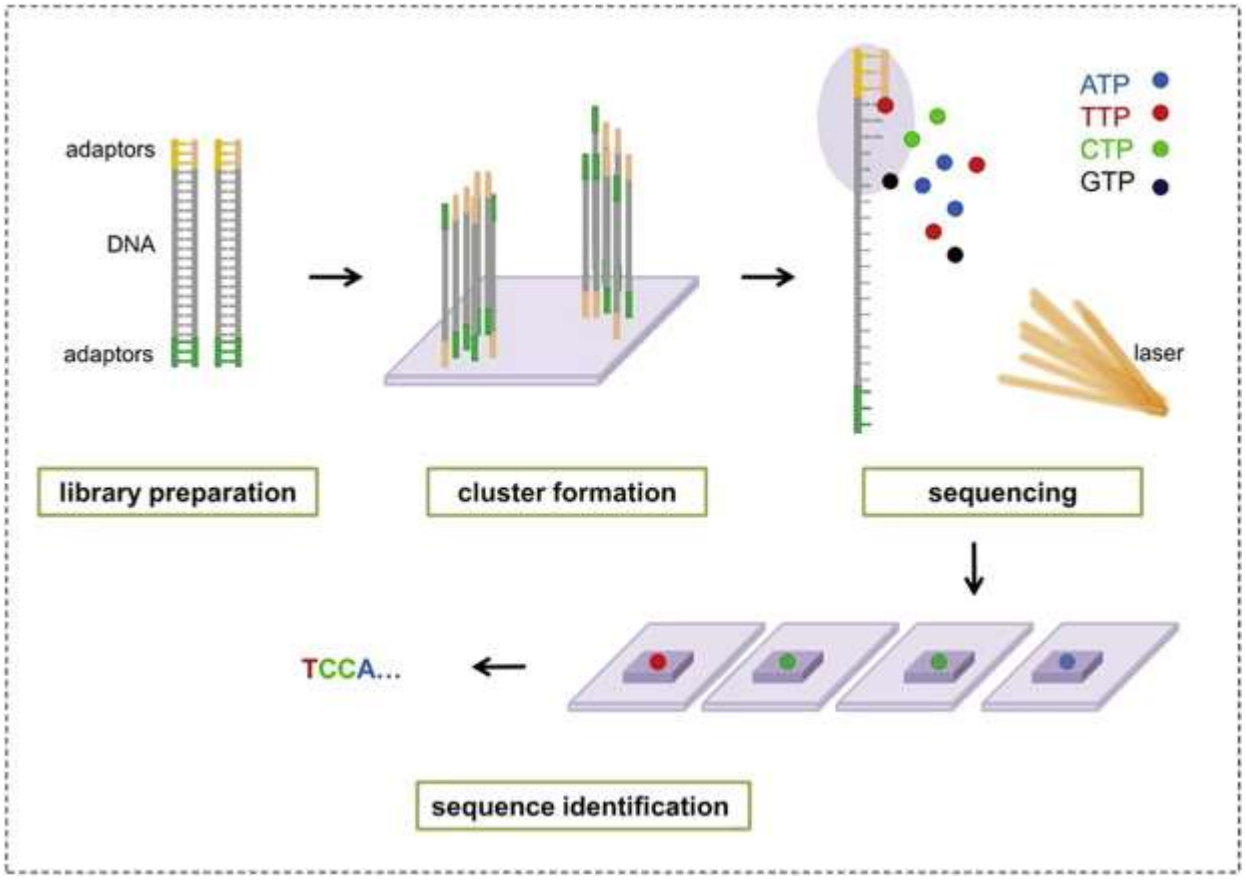
6. Capillary Gel Electrophoresis:

In these systems, slab gel electrophoresis is replaced by capillary gel electrophoresis to analyse DNA samples. In these systems, instead of scanning DNA as it migrates through 96 lanes each in a series of 96 capillary tubes, DNA fragments pass are scanned.

In the original models of the above old slab gel machines, gels must be poured and reagents frequently reloaded, interrupting the sequencing. In capillary gel sequencing systems, on the other hand, the robot moves the DNA samples and reagents through the tubes continuously, requiring attention only once a day. The system produces a steady flow of data, each signal representing one of the four DNA bases (adenine, cytosine, guanine and thymine).

7. Illumnia Clustering

Illumina short read technology is currently the most popular sequencing technology. Reads are sequenced in clusters, which are generated using a process called bridge amplification. First, DNA is sheared and tagged with adapters. Then, with a series of cycles, the DNA fragment is complemented to produce an additional strand. The DNA is then folded over to another receptor, and the receptors on opposite strands are detached, to form two strands.



This process is then repeated to form large clusters of DNA strands simultaneously. Finally, in order to read the bases, a process called sequencing by synthesis is used. Through a variety of cycles, the complement base pair is added to the strands of DNA, and the signal from that complement base pair is recorded. Then, the process is repeated to sequence the next base pair, and is repeated until the entire strand has been sequenced. Overall, the read error is less than one percent with this technology. While this approach is the standard due to its affordable

price point, the technology has a variety of disadvantages. First, this technology is limited to reads of roughly 150 bp. Additionally, this technology has a large sequence bias, so regardless of the coverage, some regions of the genome will always be poorly sequenced.

8. PacBio SMRT Technology

While Illumina reads are rather short, PacBio SMRT technology is used to generate reads with a mean size of 10 - 15 kilobases, with a maximum size of 30 kilobases. This is accomplished by reading a single DNA molecule at a time, but at the cost of having a very high error rate. These long reads, along with a high error rate, make processing and assembling the reads a different, much more difficult, computational task. However, long reads provide a variety of advantages compared to short reads. Structural variants, such as an insertion, or a deletion, are difficult to detect if they are longer than the size of the read. Therefore, longer reads allow us to detect many more structural variants. Additionally, longer reads are better for haplotype phasing, which we will discuss in a future lecture. Finally, the errors associated with this long read technology are more stochastic, and less bias, than the popular Illumina technology. Therefore, long reads enable us to sequence certain areas of the genome with much higher accuracy. While PacBio sequencing is not very popular in human sequencing, it is popular in agriculture, where the high quality of long reads is more suitable for the field's specific needs.

9. Oxford Nanopore

Similar to PacBio, Oxford Nanopore technology is able to read a single base pair at a time using its proprietary nanopore technology. While the error rate is currently even higher than PacBio, this technology is interesting because it is a portable device that attaches to a computer via USB. Additionally, the device will be free after agreeing to purchase $75,000 worth of reactants.

10. Moleculo

Moleculo is another long read technology that was developed here at Stanford by Steven Quake's group, and then purchased by Illumina in 2013. First, DNA is sheared to roughly lengths of roughly 10kb. Then, the fragments are diluted, and distributed into 384 wells. The fragments in each well are then amplified with PCR, cut into short fragments, and tagged with the individual well's unique barcode. Finally, these short fragments are pooled together, and sequenced. This is advantageous because the barcodes allow for greater accuracy during assembly, as well as haplotype phasing.

4.4. Northern Blotting

The trend set by Southern blotting (in 1975) to detect specific DNA brought new ideas in the field of modern molecular biology. The method got modified in 1977, to develop something very similar to the southern blot when James Alwin, David Kemp and George Stark at Stanford University repeated the design of the southern blot. The major difference was the use of RNA sample to detect a specific RNA molecule within that sample. This was performed to transfercellular RNA to chemically activated cellulose paper and given the name northern blot for detection of RNA. However, the technique still made use of a radio-labelled DNA probe.

Definition

The technique that is used in molecular biology research to study gene expression by detection of RNA or isolated mRNA in a sample is called northern blotting (RNA blotting). It is a classical method for analysis of the size and steady state level of a specific RNA in a complex sample.

It is relatively simple to perform, inexpensive and not obstructed by artefacts. However, there are many technical difficulties that may be encountered in northern blotting because of the process of gel fractionation of the RNA and probe preparation to variation in the quality of the blotting membrane utilized.

Principle

The fundamental principle of northern blotting is to separate RNA based on their size using gel electrophoresis and identified on a cellular membrane by means of hybridization probe with a base sequence corresponding to all or a part of the chain of the target RNA. Initially, we extract RNA from a tissue through chaotropic agents such as guanidinium isothiocynate to cause disrupting in cells and denaturing the proteins as well as dissolving the RNA. Some cases require the isolation of mRNA from total RNA using a poly-A+ selection procedure.The derived RNA then undergoes agarose-gel electrophoresis where it gets separated which is then proceeded by blotting onto a nylon membrane. The RNA on the membrane must be immobilized after blotting through baking or exposure under UV light avoiding nucleic acid to wash away later. Finally, hybridization probe is made ready and the probe is hybridized with the membrane. A post hybridization wash is required so that the probe is bounded to target mRNA or not is ensured. The signals are then detected and visualized using x-films and other methods.

Procedure

1. **RNA Isolation**

- Multiple ways of isolating RNA but all have some common attributes such as cellular lysis and membrane disruption, inhibiting of ribonuclease activity, deprotenization and recovery of intact RNA.
- Some of the common methods of RNA extraction are:
  RNA isolation by acid guanidinium thiocyanate phenol chloroform extraction.

  RNA isolation by column based technology.
- Assessment of quantity and quality of RNA through spectrophotometry.

2. Separation of RNA using gel Electrophoresis

- RNA have secondary structure formed by intramolecular base pairing that prevents RNA from separation according to their size.
- Denaturing agent (formaldehyde or glyoxal/DMSO) disrupts the secondary structure.
- Separation of RNA is better in glyoxal/DMSO system as sharper band for specific RNA is detected by hybridization.
- Size of RNA fragments are compared by migration distance with those of molecular weight markers.
- Gels are stained with ethidium bromide after electrophoresis to detect molecular markers and rRNAs.

3. Transfer of RNA to a membrane

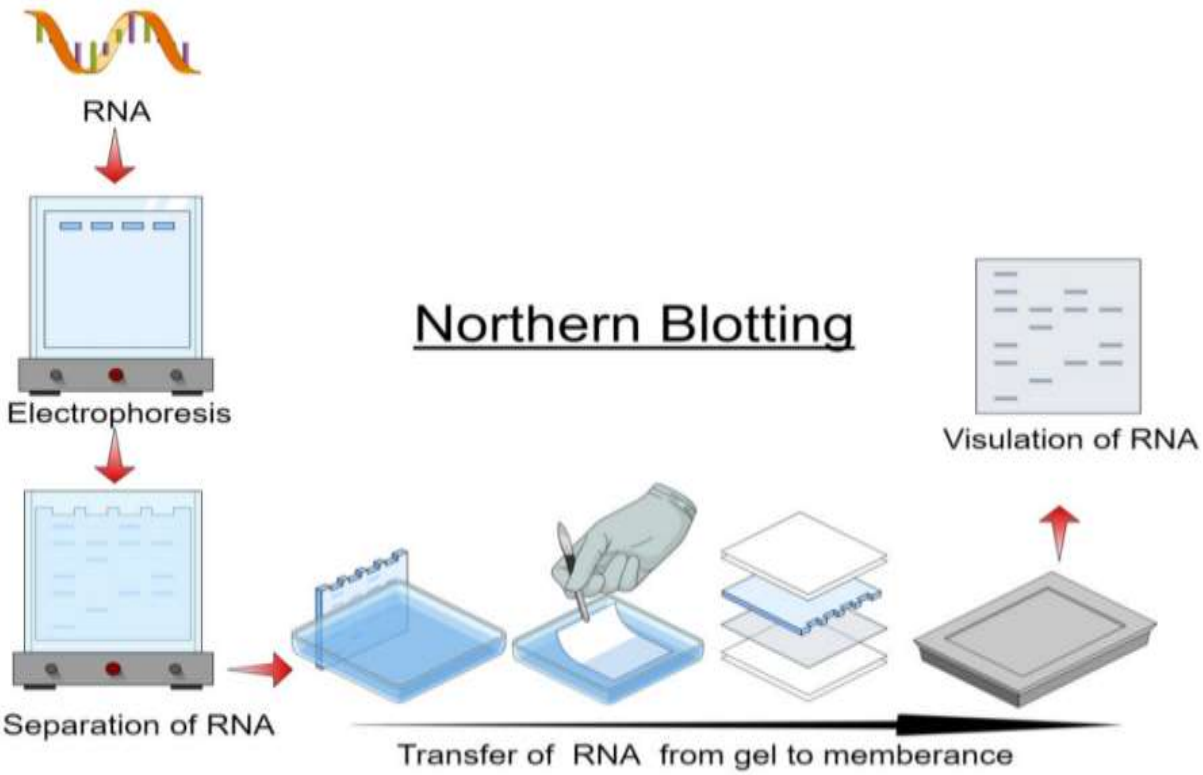
- Nitrocellulose and nylons are commonly used membranes. (Nylon membrane are more durable)
- RNA are transferred via capillary or vacuum transfer. Vacuum transfer are more efficient but require special transfer apparatus.
- Electrophoretic transfer method is available only for nylon membrane due to low concentration of salts required to bind the RNA.
- Transferred RNA are then immobilized to membrane by baking at 80 °C or through UV crosslinking in case of nylon membrane.

4. Hybridization and Washing

- Hybridization is performed using radio or fluorescently labelled probe to identify specific RNA immobilization.
- Pre-hybridization blocks single stranded probe from binding on non-specific sites on the membrane.
- Hybridization solution should contain 50% formamide to ensure hybridization at lower temperature and minimize RNA degradation.
- The membrane is washed in the buffer containing lower concentration of salt to remove excess probes.

5. Visualization

- Detection of specific transcript through autoradiography.
- Membranes are place over X-ray film.
- The X-ray film darkens where fragments are corresponding to the radioactive probes.



Northern blotting Figure adapted from: Detection of cytosolic tRNA in mammal by Northern blot analysis

Applications

- Standard for studying and inspecting gene manifestationdesign between tissues, organs, developmental stages, pathogen infection, and over the course of treatment.
- Diagnosis of several diseases (Crohn's diseases) including viral infection.
- Exhibit overexpression of oncogenes and down regulation of tumor suppressor genes in cancerous cells.
- Study of RNA degradation and splicing.
- Detect specific mRNA molecular weights and contents in a sample.
- Identify mRNA produced by the transgene to protect recombinants

Advantage and Disadvantages

Advantages

- Relatively simple, cost effective, reduced artifacts.
- Spotted membranes can be stripped of the probes and are reused for hybridization.
- Detect slightest of gene expression changes due to its sensitivity.
- Due to dilution and housekeeping genes control, the results are highly reliable.

Disadvantages

- Time consuming (Only one gene can be analyzed at a time).
- RNA degradation risks because of RNases contamination in the work environment.
- Relatively expensive for large scale analysis as huge amounts of RNA and reagents are required.

4.5. Southern Blotting

Edwin Southern, the inventor of Southern blotting started a trend to his invention after him. It was introduced as a technique to detect a particular sequence of DNA in DNA samples. He first published the article in 1975. Southern integrated three innovations to create the Southern blot – restriction endonucleases, gel electrophoresis and blotting through methods.DNA fragments were differentiated using electrophoresis based on size, then transferred to a membrane and hybridized with a radio labeled DNA probe.

Definition

An example of RFLP(restriction fragment length polymorphism), southern blotting can be defined as an analytical technique for identifying specific sequences of DNA by separating fragments on a gel and transferring them to a second medium (carrier membrane) on which hybridization testing may be carried out. During southern blotting, the DNA fragments are immobilized as a result, the membrane carries a semi-permanent reproduction of the banding pattern of the gel. The DNA are then exposed to hybridization analysis allowing bands with sequence resemblance to a labeled probe to be identified.

There are different types of membrane, transfer buffer and transfer methods to set up a southern blot. The most common and popular membranes are made of nitrocellulose, uncharged nylon positively charged nylon but they are interchangeable depending on the applications.

Principle

It is based on the principle of transfer of separated DNA fragments to a carrier membrane (usually nitrocellulose) using gel electrophoresis and subsequent identification of specific DNA fragmentsby labelled probe hybridization. Hybridization is a technique in which a double stranded DNA molecule is formed in between a single stranded DNA probe and a target single stranded DNA. The probes are labeled with a marker and complementary to the target DNA as a result we can detect one molecule of target in a mixture of millions after hybridization as the reactions are specific.

Procedure

1. Extract and purify DNA from cells
    - We separate the DNA to be tested from the rest of the cellular material in the nucleus.
    - We then incubate the specimen with detergent to promote cell lysis (frees cellular proteins and DNA).
    - Proteins are removed through organic and non-organic extraction.
    - We then use alcohol precipitation to purify the DNA from the solution.
    - Visible DNA fibers are removed and suspended in a buffer.
2. **DNA fragmentation**
    - We use restriction endonuclease enzyme to break long nucleotide sequence into smaller fragments for purification or identification process.

- Each restriction enzymes are validated with universal buffers (L, M, H, K, or T (+BSA)) and supplied with recommended buffer.
- Before appropriate DNA concentration and establishing a restriction digestion with preferable enzymes, we keep reagent necessary for digestion process on ice.
- The components are then added into a PCR tube and mixedby absorbing the contents with the help of pipette slowly avoiding formation of any bubbles.
- PCR amplifies the number of fragments of DNA obtained from the restriction digest which are easily separated using gel electrophoresis.

3**. Gel electrophoresis**

- Sorts the complex mixture of DNA fragments according to size.
- The percentage and size of the gel to be used must be determined.
- Gels consist of microscopic pores and are solid (usually agarose or polyacrimide). Generally, 0.7 – 2% gel is considered to be adequate for most of the applications.
- Nucleic acids have negative charge and move from left to right. The large molecules are held up while smaller ones move faster causing separation by size.
- Gels are stained with ethidium bromide to permit photography under UV light.
- A single band form is given to intact high quality DNA (small amount of degradation is tolerable).

4**. Denature DNA**

- DNA obtained are double stranded in nature.
- Alkalis are used to denature the restriction fragments in the gel that makes double stranded DNA to become single stranded.
- To avoid re-hybridization, we use NaCl so that DNA is neutralized.

5. **Blotting**

- Transfer DNA from the gel to solid support (carrier membrane).
- We dry the blot (around 80°C) or use UV radiation to make it permanent.

6. **Hybridization**

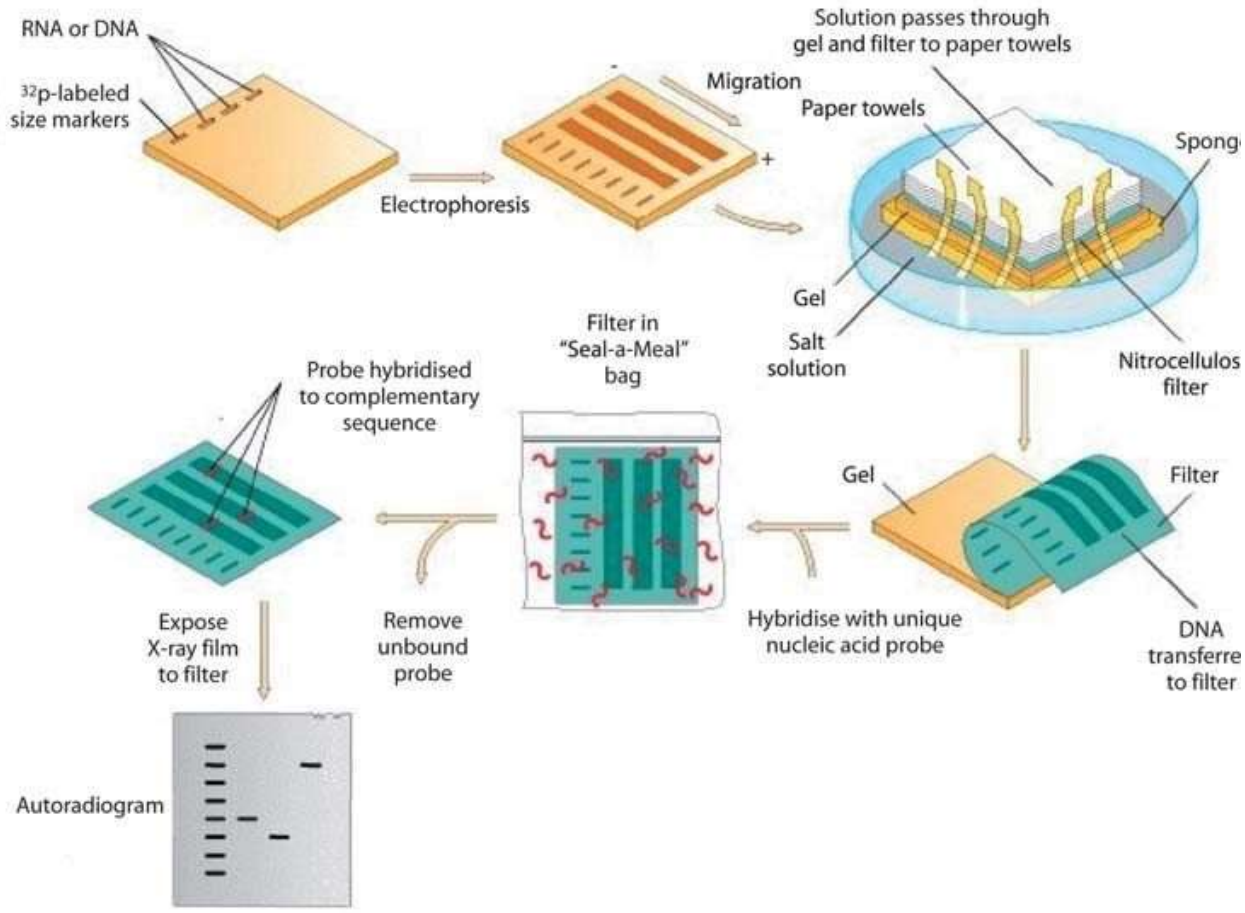
- The membrane bounded with DNA are incubated after adding the labelled probe.
- Usually requires 1-16 hours depending on the complexity of the probe and concentration.
- The probe then binds with complementary DNA on the membrane with the help of BSA or casein (blocks all other non-specific binding).

7. **Washing**

- Despite using blocking agents, some excess probe binds to the membrane.
- Wash buffers containing NaCl and detergent washes away the excess probe.

8. **Autoradiography**

- The particles are exposed to X-ray film when we use radioactive probe or fluorescent probe.
- If a chromogenic detection method is used, we can see development of color on the membrane.



Southern Blotting Figure adapted from: The mycology of the Basidiomycetes

Applications

- Identify specific DNA from a DNA sample.
- Identification of viral and bacterial infections.
- Important in the study of gene mutation, deletion and rearrangements.
- DNA fingerprinting (maternity and paternity analysis, forensic studies and personal identification).
- Diagnosis of neonatal and genetic diseases including cancer.
- Discovery of RFLP (restriction fragment length polymorphism) to map crucial genomes.

Advantage and Disadvantages

Advantages

- Less degraded compared with protein and mRNA as DNA are very stable.
- Effective way to detect specific DNA sequence from large complex samples.
- Increased sensitivity of fragments detection because of probe label used for amplifying signals.
- Only way to diagnose FSHD (Facioscapulohumerals Muscular Dystrophy).

Disadvantages

- Not applicable in routine diagnostic setting.

- Time consuming and requires large amount of DNA.
- Does not allow morphologic preservation of tissue so historic evaluation features are not available.

4.6. Western Blotting

Western blotting is a widely used technique in molecular biology and immuno-genetics for the detection and analyses of proteins.This method is also called “immune blotting” because of its nature to use an antibody for specifically identifying its antigen and also protein blotting.A qualitative and semi quantitative data can be produced using western blotting for the desired protein of interest.

Principle

- Western blotting technique principle relies on the specificity of binding between a molecule of interest and a probe to allow detection of the molecule of interest in the mixture of many other similar molecules.
- Here, molecule of interest is protein whereas the probe is typically an antibody raised against that particular protein.
- SDS PAGE is a prerequisite for western blotting .
- Proteins get spilt up by their size by a process called SDS-polyacrylamide gel electrophoresis.
- A primary antibody (an antibody precisely for the target protein) is then used to probe and wash the membrane with transferred protein.
- This primary antibody treated membranes are then reacted with a secondary antibody, usually an antibody enzyme conjugate (e.g. horseradish peroxidase).
- The target protein is visualized as band on blotting paper, X-ray film or imaging system.

Procedure

1. **Sample Preparation**

- Use different samples to extract protein(e.g. tissues or cells)
- Use homogenizer or sonication to breakdown samples.
- Prevent sample digestion at cold temperature through protease and phosphates.
- Finally, observe concentration of proteins and use spectrophotometer for protein concentration.

2. Gel Electrophoresis

- Proteins are separated on the basis of their size, shape and charge.
- In SDS gel electrophoresis, protein samples are separated according to their molecular weights.
- Load protein samples to polyacrylamide gel (*Higher percentage of gels are used for low molecular weight problems and vice versa*).

3. Protein Transfer

- A solid support membrane is placed where we transfer the separated protein for antibody detection.
- The method is called electro blotting where an electric field is directed perpendicular to the surface of the gel maintain their relative position.
- A transfer sandwich is created made up of: a fiber pad (sponge), filter papers, blotting membrane, gel, filter papers and finally a fiber pad.

4. Protein staining

- The gels must be stained as proteins are not directly visible in the gel.
- Dyes like coomassie blue, silver stain or deep purples are used.
- The gel is imaged with suitable instrument and a permanent record can be made after staining.

5. Blocking

- Prevents non-specific binding of antibodies by blocking unoccupied sites of membrane with inert protein or non-ionic detergent.
- BSA and non-fat dry milk are commonly used typical blockers.
- Blocking agents should possess greater connection towards membrane than the antibodies.

6. Antibody probing

- Incubate the blot with one or more antibodies
- Primary antibodies are specific depending on the antigens to be detected.
- The secondary antibody (monoclonal or polyclonal) is linked to an enzyme that is used to indicate the location of the protein**.**

7. Washing

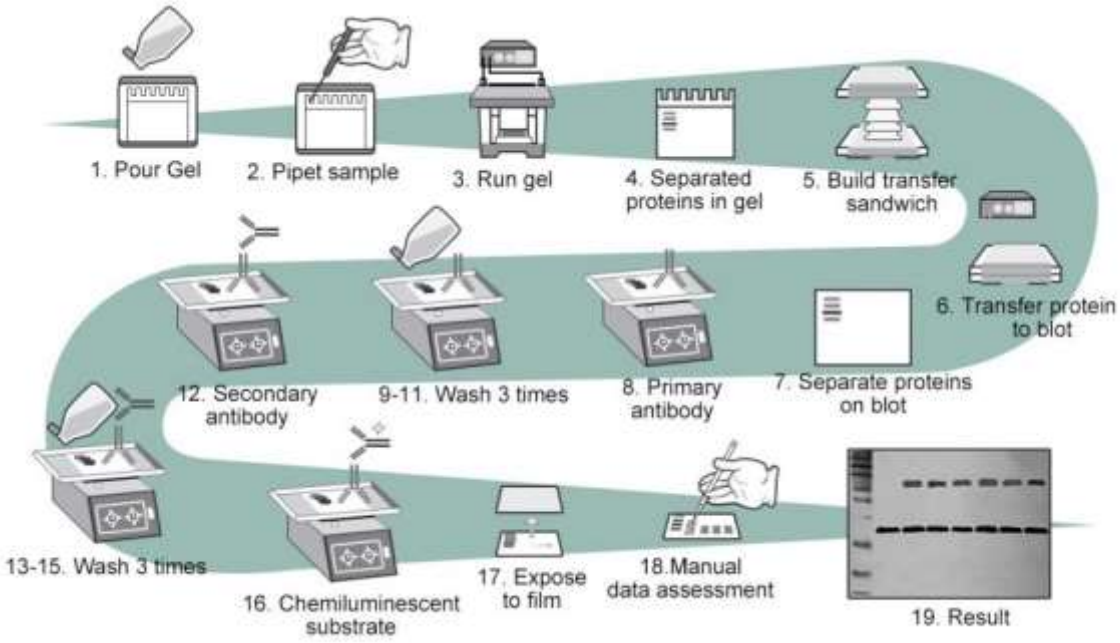
- Removes unbounded antibodies from the membrane.
- Commonly used buffer : a dilute solution of tween-20 in TBS or PBS buffer .

8. Protein Detection

- Alkaline phosphates (AP) and horse radish peroxidase (HRP) are widely used.
- Four types: Chromogenic detection, Chemiluminescence detection, Fluorescent detection and radioactive detection.

9. Analysis and Imaging

- Detection of signals using X-ray film, scanners or CCD.
- Benchmarking with marker protein to estimate the molecular weight of the protein.
- Verification can be done through qualitative and quantitative analysis to show the presence and absence of specific proteins of interests.



Western Blotting: Figure adapted from: https://www.elabscience.com/List-detail-306.html

Applications

- Detection of particular protein from a mixture of proteins.
- Size and amount estimate of proteins in the mixture.
- Verification following a high sensitivity ELISA test for diagnosis of Lyme, HIV infection, BSE, HBV and so on.
- Detect condensed isoforms of proteins as well as tagged proteins.

Advantage and Disadvantages

Advantages

- Effective early diagnostic tool.
- Detect minimal immunogenic response form virus or bacteria.
- Requires fewer antibodies for testing.
- Detect specific protein from a large mixture of different proteins. (Even more than 300,000)

Disadvantages

- Requires specific primary antibodies to perform test on desired protein of interest.
- Challenging and hence requires well trained staffs.
- Poorer results as antibodies may revel off-target bindings.
- Detecting and imaging the results can be expensive as equipment cost is high.

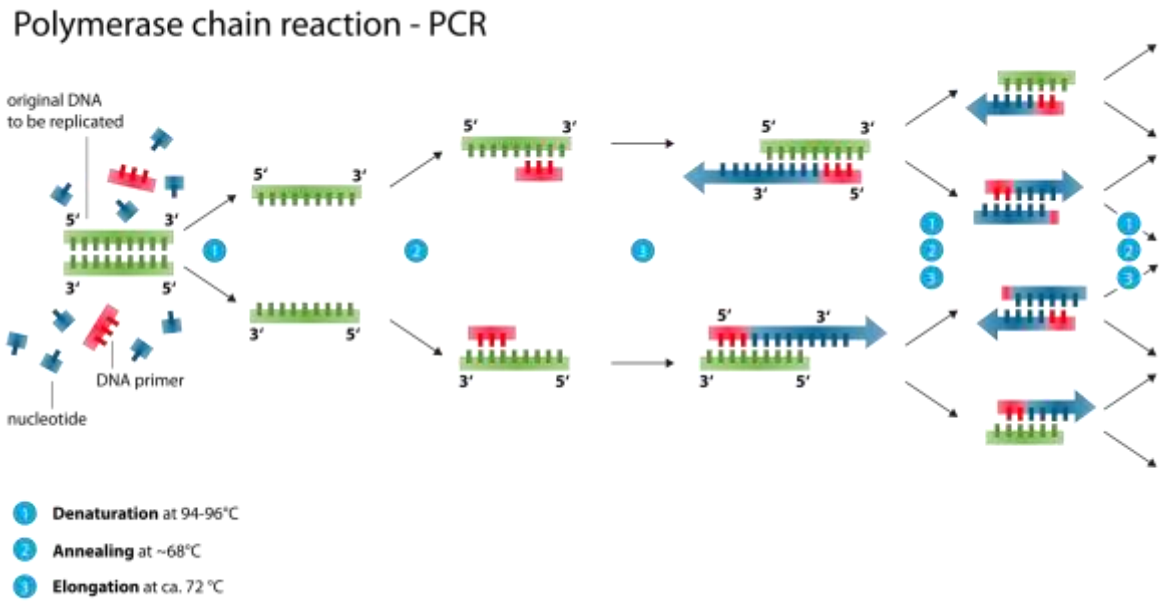
4.7. Polymerase chain reaction (PCR

Principle:

Polymerase chain reaction is method for amplifying particular segments of DNA. It is an enzymatic method and carried out invitro. PCR technique was developed by Kary mullis in 1983. PCR is very simple, inexpensive technique for characterization, analysis and synthesis of specific fragments of DNA or RNA from virtually any living organisms.

Steps or procedures:

PCR consists of three basic steps.



1. **Denaturation:**
    - Two strand of DNA separates (melt down) to form single stranded DNA
    - This step is generaly carried out at 92C-96C for 2 minutes.

2. Annealing:
    - Annealing of primer to each strand is carried out at 45C-55C

3. Extension:
    - DNA polymerase adds dNTPs complementary to templates strands at 3'end of primer.
    - It is carried out at temperature of 72C.

- *These three steps are repeated 20-30 times in an automated thermocycler that can heat and cool the reaction mixture in tube within very short time. This results in exponential accumulation of specific DNA fragments.*
- *The doubling of number of DNA strands corresponding to target sequences can be estimated by amplification number associated with each cycle using the formula.*
- *Amplification=2n, where n=no. of PCR cycle.*
- *PCR can amplify a desired DNA sequences of any origin hundred or millions time in a matter of hour, which is very short in comparison to recombinant DNA technology.*

- *PCR is especially valuable because the reaction is highly specific, easily automated and very sensitive.*
- *It is widely used in the fields like- clinical medicine for medical diagnosis, diagnosis of genetic diseases, forensic science; DNA finger printing, evolutional biology.*

Factors affecting PCR

i. Primer

- PCR reaction needs two primer, a forward and a reverse primer
- Primer are synthesized oligonucleotide usually ranging from 15-30 bases long
- Primers are designed such that at 3'end they donot have more than two bases complementary to each other as this results in PRIMER-DIMER formation.
- The G+C contents is in the range of 40-60%
- The melting temperature (Tm) of both forward and reverse primer is usually the same.
- Low concentration of primer results in poor yield while high concentration may results in non specific amplification. Hence optimal concentration of primer is needed ie 0.1-1µL

ii. Amount of Template DNA

- Optimal amount of template DNA usually in nano gram. Higher concentration inhibit or results in non specific amplification.
- Taq DNA polymerase:
- Taq DNA polymerase is 94 KD thermostable DNA polymerase isolated from Thermus aquaticus.
- Optimal temperature for activity of Taq polymerase is 72° but it can tolerate high temperature and donot affects by denaturating temperature of 94°C.
- Taq DNA polymerase have both 5'-3' polymerase activity and 5'-3' exonuclease activity. But it lacks 3'-5' exonuclease activity (proof reading activity).

Types of PCR

1. Standard PCR:

- Nested PCR
- Random amplified polymorphic DNA
- Long PCR
- Restriction fragment length polymorphism (RFLP)
- Amplified fragment length polymorphism (AFLP)
- Multiplex PCR
- Single cell PCR
- Fast cycling PCR
- In situ PCR
- High fidelity PCR
- Asymmetric PCR
- Repetitive sequence based PCR
- Overlap extension PCR

- Assemble PCR
- Mini primer PCR
- Solid phase PCR
- Touch Down PCR

2. Reverse transcriptase Polymerase chain reaction (RT-PCT): for RNA

- One step RT-PCR
- Two step RT-PCR

3. Real time PCR: for DNA or RNA

- Dye binding to ds DNA
- Fluorescent probes

Application:

1. Forensic science: DNA finger printing, paternity testing and criminal identification
2. Diagnosis: Molecular identification of microorganisms
3. Evolution study: evolutionary biology
4. Fossil study: paleontology
5. Gene cloning and expression
6. Gene sequencing
7. Vaccine production by recombinant DNA technology
8. Drug discovery
9. Mutation study
10. Human genome project

4.8. Methods for Detecting DNA-Protein Interactions

DNA-protein interactions (DPIs) are critical to all living organisms, particularly in the regulation of gene expression, replication, packing, recombination, and repair, as well as RNA transport and translation. Since microscopically observing interactions between proteins and DNA in the late nineteenth century, scientists have been intrigued in the mechanisms by which proteins associate with and control both DNA and RNA. DNA-binding proteins are common and ubiquitous, comprising on average 10% of the proteome/genome of higher plants and animals—about 2000 for the average organism.

Interactions may be specific, that is governed by nucleotide sequences and mediated by hydrogen bonding, ionic interactions, and van der Waal's forces. Control of transcription is a one example. Nucleotide sequence is irrelevant for nonspecific interactions, which occur through attraction between the protein's functional groups and the sugar-phosphate backbone of DNA.

Analytical methods

Most DPIs are only partly understood, but not for want of trying. Their study often begins with identifying and characterizing the protein component. Analytical methods include microscopy and classical biochemical assays like chromatin immunoprecipitation analysis (ChIP), Systematic

Evolution of Ligands by EXponential enrichment (SELEX), Electrophoretic mobility shift assays, DNA footprinting, and protein-binding microarrays.

ChIP causes proteins to bind covalently to their DNA targets, after which they are unlinked and characterized separately. SELEX exposes target proteins to a random library of oligonucleotides. Those genes that bind are separated and amplified by PCR. DNase footprinting locates protein-DNA binding sites followed by limited DNA digestion. Protein-binding DNA microarrays identify gene sequences that associate with labeled target proteins, followed by fluorescence detection.

Microscopic techniques include optical, fluorescence, electron, and atomic force microscopy (AFM), the latter two providing the highest spatial resolution. Where the latter three resolve dynamic interactions, electron microscopy is limited to static observations. AFM is arguably the most versatile microscopic method because It offers sub-nanometer resolution, images samples in liquids, and probes intermolecular forces between single molecules. Sandor Kasas, Ph.D., scientific collaborator at the Swiss Federal Polytechnic University (Lausanne) and an investigator into the microscopy of DPIs, notes that high-speed AFM explores the structures of proteins and associated genes with very high resolution. “With high-speed AFMs, all the dynamics of the interaction can be followed in real time while measuring the interaction forces between the protein and DNA.”This last application involves attaching the protein to the AFM tip and measuring its interaction with free-floating DNA. Depending on what investigators are looking for the protein sample may or may not be in solution, but in any case it must be immobilized onto a substrate to be visible to the AFM. For quick and dirty experiments one can image samples in air. In this case attachment is relatively easy. If one is looking for interaction dynamics, however, imaging must occur in liquids, where attachment is more complicated. “The DNA must be attached to the surface strongly enough to be visible to the AFM, but bound loosely enough to interact with the protein. This compromise is sometimes difficult to achieve,” Kasas adds. A low speed AFM costs about $100,000 to $200,000, which means groups or departments will often pool resources to acquire one. Dedicated AFM facilities are often happy to collaborate with biologists.

High-speed AFMs are another matter

“These are still rare and only few groups possess them, only about five in France and perhaps a dozen in the United States,” Kasas says. Yet for biologists working on the cutting edge of DPIs they are essential, and for their research transformative.

Physical characterization

Biochemical analysis is just one aspect of DPI characterization. Physical methods like X-ray crystallography also provides valuable insights, namely a three-dimensional atomic view of DPIs. The resulting crystal structure may reveal the location of the binding site complex, allowing for identification of essential amino acid residues necessary for complex formation. This can help with downstream pharmaceutical development, in particular, structure-based drug design.

Crystallization of proteins for structural studies has been around for decades. As it is often difficult to predict optimal (and practical) conditions for crystallization, favorable conditions are identified by screening against variations of chemical formulations and concentrations. Yet when conventional techniques are applied to DPIs the resulting crystals often consist of either protein or DNA alone, leading to false positives and wasted time. It appears that the crystallization process itself may cause dissociation of these complexes.

Proteins interact with DNA through electrostatic interactions (salt bridges), dipolar interactions (hydrogen bonding, H-bonds), entropic effects (hydrophobic interactions) and dispersion forces (base stacking). These forces contribute in varying degrees to proteins binding in a sequence-specific or non–sequence-specific manner. Understanding how proteins interact with DNA, determining what proteins are present in these protein–DNA complexes, and identifying the nucleic acid sequence (and possible structure) required to assemble these complexes are vital to understanding the role these complexes play in regulating cellular processes. A number of laboratory techniques have been developed to study the complex interactions of proteins with DNA, each with a unique history, varying utility, and distinct strengths and weaknesses.

Chromatin immunoprecipitation (ChIP) assays

The chromatin immunoprecipitation (ChIP) method can be used to monitor transcriptional regulation through histone modification (epigenetics) or transcription factor–DNA binding interactions. The ChIP assay method allows analysis of DNA–protein interactions in living cells by treating the cells with formaldehyde or other crosslinking reagents in order to stabilize the interactions for downstream purification and detection. Performing ChIP assays requires knowledge of the target protein and DNA sequence that will be analyzed, as researchers must provide an antibody against the protein of interest and PCR primers for the DNA sequence of interest. The antibody is used to selectively precipitate the protein–DNA complex from the other genomic DNA fragments and protein–DNA complexes. The PCR primers allow specific amplification and detection of the target DNA sequence. Quantitative PCR (qPCR) technique allows the amount of target DNA sequence to be quantified. The ChIP assay is amenable to array-based formats (ChIP-on-chip) or direct sequencing of the DNA captured by the immunoprecipitated protein (ChIP-seq).

Strengths	Limitations
• capture a snapshot of specific protein–DNA interactions as they occur in living cells • quantitative when coupled with qPCR analysis • ability to profile a promoter for different proteins	• researcher needs to source ChIP-grade antibodies • requires designing specific primers • difficult to adapt for high-throughput screening

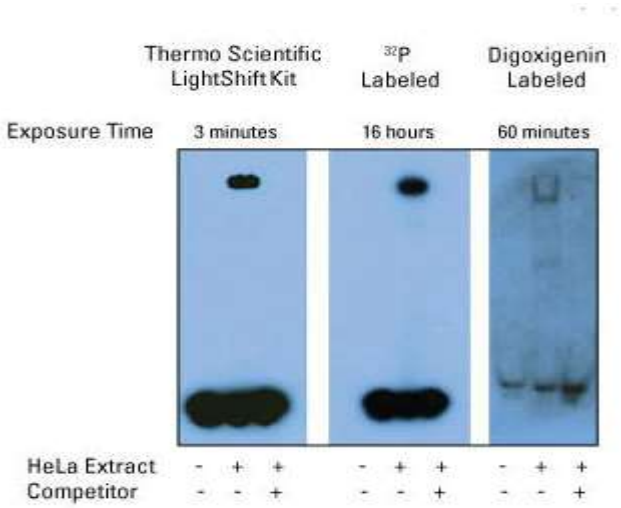
DNA electrophoretic mobility shift assay (EMSA)

The DNA electrophoretic mobility shift assay (EMSA) is used to study proteins binding to known DNA oligonucleotide probes and can be used to assess the degree of affinity or specificity of the interaction. The technique is based on the observation that protein–DNA complexes migrate more slowly than free DNA molecules when subjected to non-denaturing polyacrylamide or agarose gel electrophoresis. Because the rate of DNA migration is shifted or retarded upon protein binding, the assay is also referred to as a gel shift or gel retardation assay. Adding a protein-specific antibody to the binding components creates an even larger complex (antibody–protein–DNA), which migrates even slower during electrophoresis. This is known as a “supershift”, and it can be used to confirm

protein identities. Until conception of the EMSA, protein–DNA interactions were studied primarily by nitrocellulose filter–binding assays using radioactively labeled probes.

Strengths	Limitations
• detect low abundance DNA binding proteins from lysates • test binding site mutations using many probe configurations with the same lysate • test binding affinity through DNA probe mutational analysis • non-radioactive EMSA possible using biotinylated or fluorescently labeled DNA probes	• analyze protein–DNA interactions in vitro • difficult to quantitate • need to perform supershift assay with antibody to be certain of protein identity in a complex

Traditionally, DNA probes have been radiolabeled with ^{32}P by incorporating an [γ-^{32}P]dNTP during a 3' fill-in reaction using Klenow fragment or by 5' end labeling using [γ-^{32}P]ATP and T4 polynucleotide kinase. Following electrophoresis, the gel is exposed to X-ray film to document the results. The Thermo Scientific LightShift Chemiluminescent EMSA Kit is a non-radioactive assay that provides robust and sensitive performance. The kit includes reagents for setting up and customizing DNA-binding reactions, a control set of DNA and protein extract to test the kit system, stabilized streptavidin–HRP conjugate to probe for the biotin-labeled DNA target, and an exceptionally sensitive chemiluminescent substrate module for detection.



Chemiluminescent EMSA of four different DNA–protein complexes. Biotin-labeled target duplexes ranged in size from 21–25 bp. The Oct-1, AP1 and NF-κB transcription factors were derived from HeLa nuclear extract. EBNA-1 extract is provided as a control in the LightShift Chemiluminescent EMSA Kit. Unlabeled specific competitor sequences (where used) were present at a 200-fold molar excess over labeled target. X-ray film exposure times for each system ranged from 2 minutes for EBNA, Oct-1 and AP1, and 5 minutes for NF-κB

DNA pull-down assays

Pull-down assays are used to selectively extract a protein–DNA complex from a sample. Typically, the pull-down assay uses a DNA probe labeled with a high affinity tag, such as biotin, which allows the probe to be recovered or immobilized. A biotinylated DNA probe can be complexed with a protein from a cell lysate in a reaction similar to that used in the EMSA and then used to purify the complex using agarose or magnetic beads. The proteins are then eluted from the DNA and detected by western blot or identified by mass spectrometry. Alternatively, the protein may be labeled with an affinity tag, or the DNA–protein complex may be isolated using an antibody against the protein of interest (similar to a supershift assay). In this case, the unknown DNA sequence bound by the protein is detected by Southern blotting or through PCR analysis.

Strengths	Limitations
• enrichment of low abundant targets • end-labeled DNA can be generated by several methods • isolation of intact complex • compatible with immunoblotting and mass spectrometry analysis	• long DNA probes can show significant nonspecific binding • requires very specific antibodies for native proteins • requires nuclease-free conditions • assay must be performed in vitro

Microplate capture and detection assays

A hybrid of the DNA pull-down assay and enzyme-linked immunosorbent assay (ELISA), microplate capture assays use immobilized DNA probes to capture specific protein–DNA interactions and confirm protein identities and relative amounts with target specific antibodies. Typically, a biotinylated DNA probe is immobilized on the surface of a 96- or 384-well microplate coated with streptavidin. A cellular extract is prepared in binding buffer and added for a sufficient amount of time to allow the putative binding protein to bind to the oligonucleotide. The extract is then removed and each well is washed several times to remove nonspecifically bound proteins. Finally, the protein is detected using a specific antibody labeled for detection. This method can be extremely sensitive when performed with enzyme-labeled antibodies and a chemiluminescent substrate, detecting less than 0.2 pg of the target protein per well. The microplate format is efficient and compatible with high-throughput analysis, allowing statistical mutational and activation assays to be performed. This method may also be utilized for oligonucleotides labeled with other tags, such as primary amines that can be immobilized on microplates coated with an amine-reactive surface chemistry.

Strengths	Limitations
• the use of ELISA-based technology increases speed and throughput • compatible with drug screening • possible to optimize sensitive non-radioactive assays	• requires antibodies with affinity for DNA-bound native proteins (i.e., supershift antibodies) • data only provides relative changes in transcription factor–DNA affinity or abundance • assay kits are available for only a few targets

Reporter assays

Reporter assays provide a real-time in vivo readout of translational activity for a promoter of interest. Reporter genes are fusions of a target promoter DNA sequence and a reporter gene DNA sequence. The promoter DNA sequence is customized by the researcher and the reporter gene DNA sequence codes for a protein with detectable properties such as firefly luciferase, Renilla luciferase or alkaline phosphatase. These genes produce enzymes only when the promoter of interest is activated. The enzyme, in turn, catalyzes a substrate to produce either light, a color change, or other reaction that can be detected by spectroscopic instrumentation. The signal from the reporter gene is used as an indirect determinant for the translation of endogenous proteins driven from the same promoter.

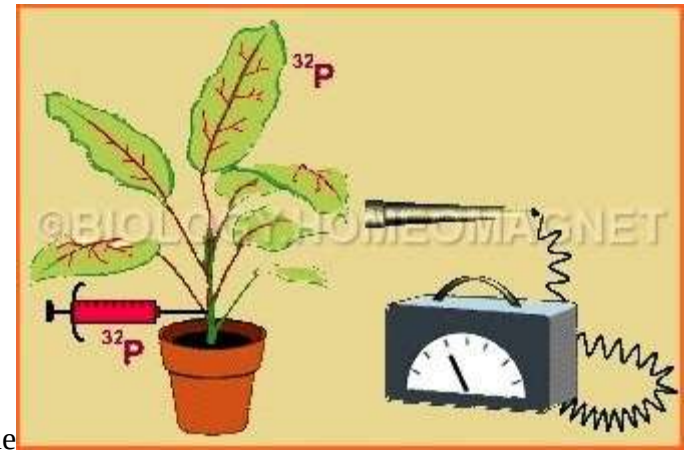
Strengths	Limitations
• *in vivo* monitoring • captures real-time data • powerful tool for mutational analysis of promoters • amenable to high-throughput screening	• uses exogenous DNA • does not address changes due to genomic sequences near the promoter of interest • artifacts due to gene dosage can occur

UNIT – V -RADIO LABELLING AND ANALYSIS– SBT1002

5. TRACER TECHNIQUES IN BIOLOGY

For the study of biological pathway and mechanism Tracer techniques has its importance. Tracer techniques involve use of isotopically labelled molecules and detection of the isotopes for the study.An isotope is a form of an element which has the same atomic number of electrons as the common form of the element but it differs in atomic weight. The difference of atomic weight is due to difference of number of neutrons in its nucleus. An isotope may be stable or radioactive depending on the relative number of protons and neutrons in its nucleus.

Both stable and radioactive isotopes of an element are identical in chemical properties, and thus they undergo all the physical and chemical changes like the ordinary form of the element. Moreover, the can be detected at any time by atomic weight or radioactivity. By Ginger-Muller counter or other sensitive detectors the radioactive isotopes can be detected by their radioactivity even when it present in very small quantity. Detection of stable isotopes can be done by their atomic weight through a mass spectrograph. Stable or radioactive isotopes used for studying the fate of a molecule in physical, chemical or biological processes are called tracer element and the methods for such studies are called tracer techniques . Commonly used radioactive tracers in the study of biology are C14, P12, H3 etc. An important stable isotope used as tracer in biology is O18.



Tracer Technique

At specific positions in their molecules inorganic and organic compounds can be prepared with isotopes. Such compound containing an isotope in its molecule is used as a tracer. In the tracer technique the isotope element in the molecule is said ‘tagged ‘or ‘labelled’. When an isotopically labelled compound is administrated to an animal or a plant or incubated with tissue preparations, it undergoes same fate as the unlabelled form of the compound and the isotopically labelled products can be detected. In this way, this source, metabolic pathway and end products of bio-molecules can be studied with the use of isotopically tagged tracers.

In the process of Tracer techniques tracers are also used for determining the following: –

i) Metabolic turnover of a substance.

ii) Relative proportion of a substance being catabolised through different pathways.

iii) Intestinal absorption of the nutrient.

iv) Volume of body fluids.

v) Blood level of a hormone.

vi) Mechanism and site of action of a hormone.

vii) Cardiac output.

viii) Flow of blood through an organ.

ix) Intracellular distribution, i.e. autoradiography.

The most common radioactive tracer used in metabolic studies is C^{14}. P^{32}, which is a radioactive tracer is mainly used to study the phosphorylation reactions. H^3 or tritium is also a radioactive isotope which is used as a tracer in the form of tritium oxide (THO) for determination of total body fluid volume. O^{18} is a stable isotope and it is used to trace the source of O_2 liberated in photosynthesis. When O^{18} labelled water (H_2O^{18}) is used in photosynthesis, $O^{18}{}_2$ is liberated. It is thus proved that the water is the source of oxygen liberated in photosynthesis.

Radiation dosimeter

A radiation dosimeter is a device, instrument or system that measures or evaluates, either directly or indirectly, the quantities exposure, kerma, absorbed dose or equivalent dose, or their time derivatives (rates), or related quantities of ionizing radiation. A dosimeter along with its reader is referred to as a dosimetry system. Measurement of a dosimetric quantity is the process of finding the value of the quantity experimentally using dosimetry systems. The result of a measurement is the value of a dosimetric quantity expressed as the product of a numerical value and an appropriate unit. To function as a radiation dosimeter, the dosimeter must possess at least one physical property that is a function of the measured dosimetric quantity and that can be used for radiation dosimetry with proper calibration. In order to be useful, radiation dosimeters must exhibit several desirable characteristics. For example, in radiotherapy exact knowledge of both the absorbed dose to water at a specified point and its spatial distribution are of importance, as well as the possibility of deriving the dose to an organ of interest in the patient. In this context, the desirable dosimeter properties will be characterized by accuracy and precision, linearity, dose or dose rate dependence, energy response, directional dependence and spatial resolution. Obviously, not all dosimeters can satisfy all characteristics. The choice of a radiation dosimeter and its reader must therefore be made judiciously, taking into account the requirements of the measurement situation; for example, in radiotherapy ionization chambers are recommended for beam calibrations and other dosimeters, such as those discussed below, are suitable for the evaluation of the dose distribution (relative dosimetry) or dose verification.

Radon Exposure

For many people, one of the largest sources of exposure to radiation is from radon gas (Rn-222). Radon-222 is an α emitter with a half–life of 3.82 days. It is one of the products of the radioactive decay series of U-238 (Chapter 21.3 Radioactive Decay), which is found in trace amounts in soil and rocks. The radon gas that is produced slowly escapes from the ground and gradually seeps into homes and other structures above. Since it is about eight times more dense than air, radon gas accumulates in basements and lower floors, and slowly diffuses throughout buildings (Figure A).

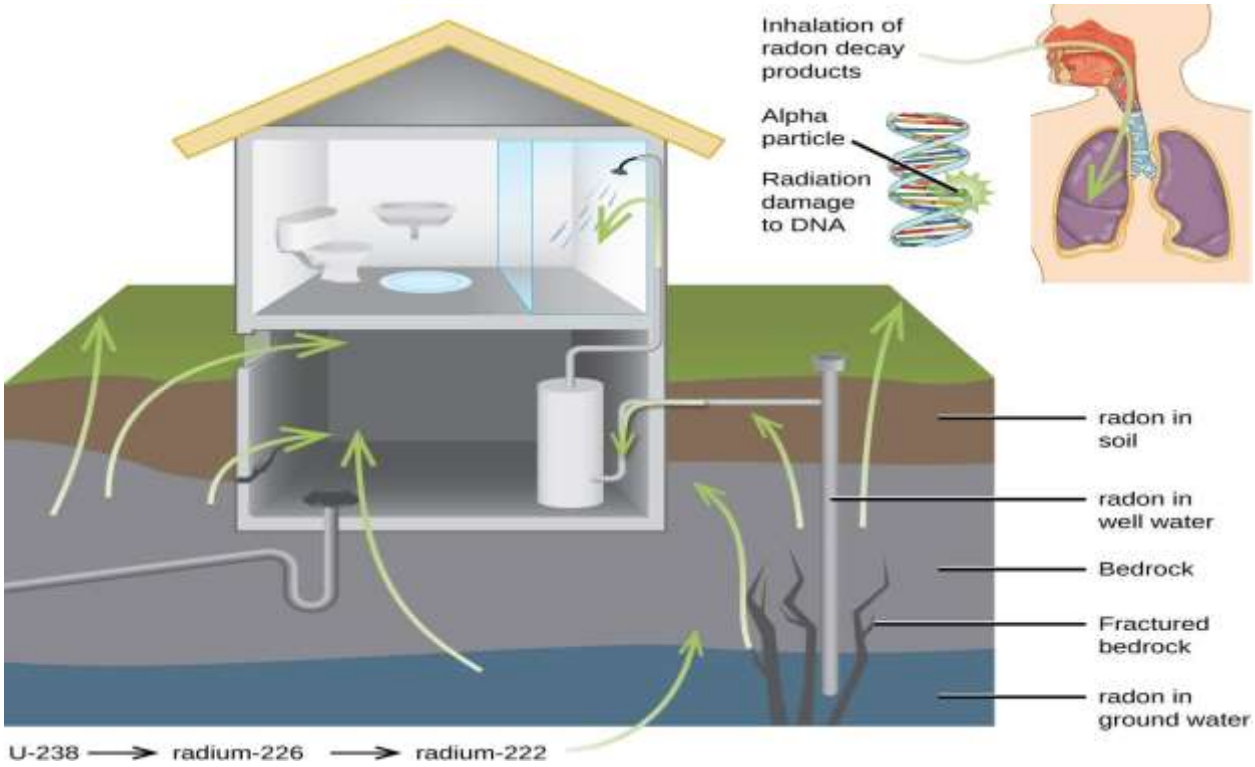


Figure A. Radon-222 seeps into houses and other buildings from rocks that contain uranium-238, a radon emitter. The radon enters through cracks in concrete foundations and basement floors, stone or porous cinderblock foundations, and openings for water and gas pipes.

Radon is found in buildings across the country, with amounts depending on where you live. The average concentration of radon inside houses in the US (1.25 pCi/L) is about three times the levels found in outside air, and about one in six houses have radon levels high enough that remediation efforts to reduce the radon concentration are recommended. Exposure to radon increases one's risk of getting cancer (especially lung cancer), and high radon levels can be as bad for health as smoking a carton of cigarettes a day. Radon is the number one cause of lung cancer in nonsmokers and the second leading cause of lung cancer overall. Radon exposure is believed to cause over 20,000 deaths in the US per year.

5.1.MEASURING RADIATION EXPOSURE

Several different devices are used to detect and measure radiation, including Geiger counters, scintillation counters (scintillators), and radiation dosimeters (Figure B). Probably the best-known radiation instrument, the Geiger counter (also called the Geiger-Müller counter) detects and measures radiation. Radiation causes the ionization of the gas in a Geiger-Müller tube. The rate of ionization is proportional to the amount of radiation. A scintillation counter contains a scintillator—a material that emits light (luminesces) when excited by ionizing radiation—and a sensor that converts the light into an electric signal. Radiation dosimeters also measure ionizing radiation and are often used to determine personal radiation exposure. Commonly used types are electronic, film badge, thermoluminescent, and quartz fiber dosimeters.

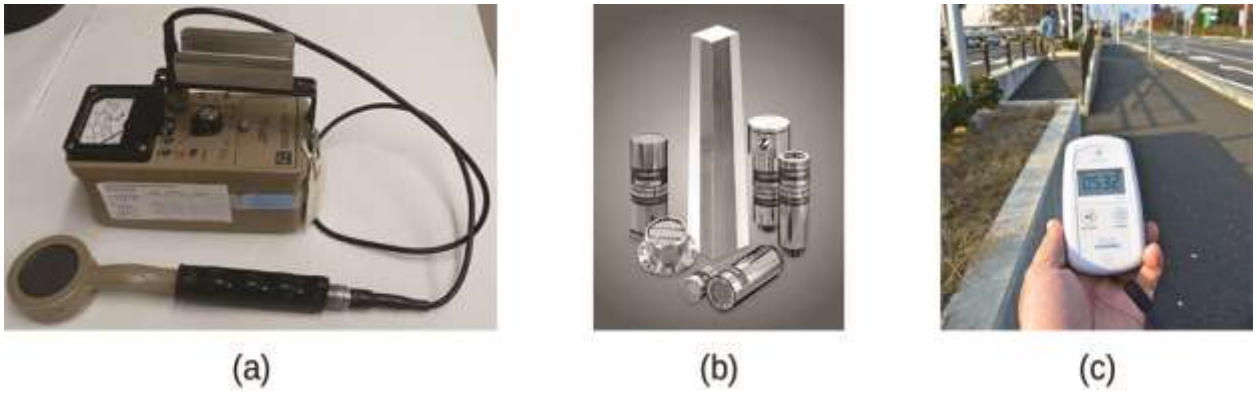


Figure B. Devices such as (a) Geiger counters, (b) scintillators, and (c) dosimeters can be used to measure radiation. (credit c: modification of work by “osaMu”/Wikimedia commons)

A variety of units are used to measure various aspects of radiation (Figure B. The SI unit for rate of radioactive decay is the becquerel (Bq), with 1 Bq = 1 disintegration per second. The curie (Ci) and millicurie (mCi) are much larger units and are frequently used in medicine (1 curie = 1 Ci = 3.7 × 1010 disintegrations per second). The SI unit for measuring radiation dose is the gray (Gy), with 1 Gy = 1 J of energy absorbed per kilogram of tissue. In medical applications, the radiation absorbed dose (rad) is more often used (1 rad = 0.01 Gy; 1 rad results in the absorption of 0.01 J/kg of tissue). The SI unit measuring tissue damage caused by radiation is the sievert (Sv). This takes into account both the energy and the biological effects of the type of radiation involved in the radiation dose. The roentgen equivalent for man (rem) is the unit for radiation damage that is used most frequently in medicine (1 rem = 1 Sv). Note that the tissue damage units (rem or Sv) includes the energy of the radiation dose (rad or Gy) along with a biological factor referred to as the RBE (for relative biological effectiveness) that is an approximate measure of the relative damage done by the radiation. These are related by:

number\;of\;rems=RBE×number\;of\;rads

with RBE approximately 10 for α radiation, 2(+) for protons and neutrons, and 1 for β and γ radiation.

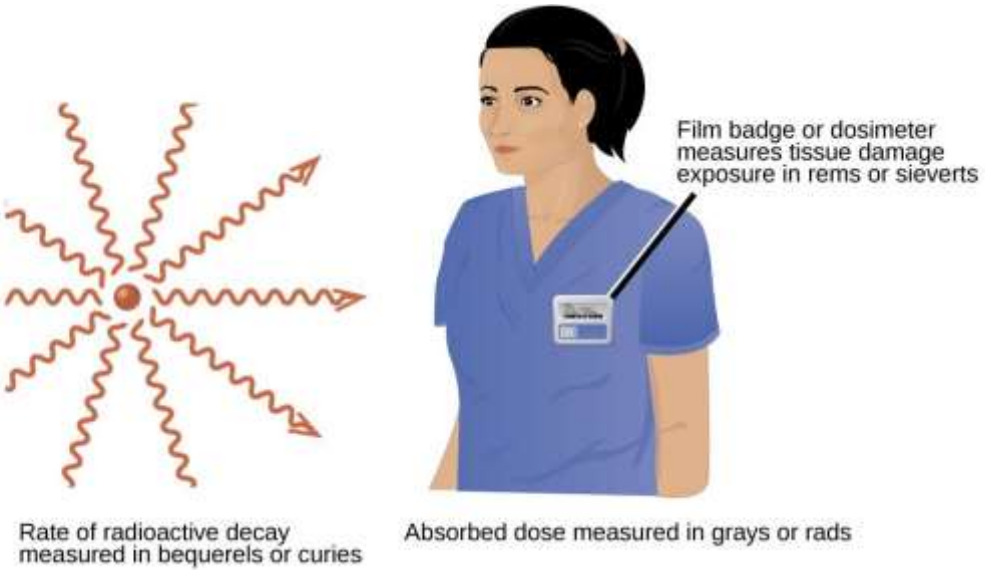


Figure 7. Different units are used to measure the rate of emission from a radioactive source, the energy that is absorbed from the source, and the amount of damage the absorbed radiation does.

5.2. UNITS OF RADIATION MEASUREMENT

Table 1 summarizes the units used for measuring radiation.

Measurement Purpose	Unit	Quantity Measured	Description
activity of source	becquerel (Bq)	radioactive decays or emissions	amount of sample that undergoes 1 decay/second
	curie (Ci)		amount of sample that undergoes 3.7 × 1010 decays/second
absorbed dose	gray (Gy)	energy absorbed per kg of tissue	1 Gy = 1 J/kg tissue
	radiation absorbed dose (rad)		1 rad = 0.01 J/kg tissue
biologically effective dose	sievert (Sv)	tissue damage	Sv = RBE × Gy
	roentgen equivalent for man (rem)		Rem = RBE × rad

5.3. Radioisotopes

An atomic species is defined by two whole numbers: the number of protons in the nucleus (known as Z, or atomic number) and the total number of protons plus neutrons (known as Z, or mass number).

Isotopes are the atoms in an element that have the same atomic number but a different atomic mass; that is, the same number of protons and thus identical chemical properties, but different numbers of neutrons and consequently different physical properties. Isotopes can be stable or unstable or radioisotopes. In the latter, their nuclei have a special property: they emit energy in the form of ionizing radiation while searching for a more stable configuration.

Atomic number (Z) and mass number (A)

The atomic number defines the chemical element that the atom belongs to. Thus, regardless of the number of neutrons they have, all atoms whose nuclei have one proton are hydrogen atoms. All of those with eight protons are oxygen atoms, etcetera.

The mass number is the whole number that is closest to the mass (expressed in atomic mass units) of the atom in question. Thus, all the atoms with A = 2 have a mass of approximately 2 mass units; atoms with A = 235 have a mass of approximately 235 atomic mass units.

Annotation: ${}^{A}Z_{X}$ (where X is the atom's element)

Isotopes

Isotopes (from the Greek *isos* = same and *tópos* = place) are atoms from a same element, whose nuclei have a different number of neutrons and, therefore, differ in mass. That is, they have the same atomic number (Z) but different mass numbers (A).

For instance, carbon is presented in nature as a mix of three isotopes with mass numbers 12, 13 and 14: ${}^{12}C$, ${}^{13}C$ and ${}^{14}C$. The global amounts of carbon in each are respectively 98.99%, 1.11% and traces.

Most chemical elements possess more than one isotope, as is the case of tin, the element with the highest number of stable isotopes. Only 21 elements, like beryllium and sodium, have one single natural isotope.

Types of isotopes:

- *Natural*: Those naturally found in nature. Examples: hydrogen has three natural isotopes (protio, which has no neutrons, deuterium, with one neutron, and tritium, with two). Another element containing very important isotopes is carbon, which includes carbon 12, the referential base of atomic mass in any element; carbon 13, the only carbon with magnetic properties, and radioactive carbon 14, very important since its average life span is 5,730 years and is widely used in archeology to determine the age of organic fossils.
- *Artificial*: These isotopes, manufactured in nuclear laboratories by bombarding of subatomic particles, usually have a short life span, mostly due to their unstable nature and radioactivity. Examples: iridium 192, used to verify that pipe welding is hermetically sealed, especially as regards transport pipes for heavy crude oil and fuels. Some isotopes from uranium are also used for nuclear work such as electric generation.

Isotopes are also subdivided into stable isotopes (there are less than 300) and unstable or radioactive isotopes (there are around 1,200). The concept of stability is not exact, since there are almost stable isotopes. That is, for some time they are unstable and become stable or turn into other stable isotopes.

Radioisotope (also known as radisotope)

These are radioactive isotopes, since they have an unstable atomic nucleus (due to the balance between neutrons and protons) and emit energy and particles when it changes to a more stable form. The energy liberated in the form change can be measured with a Geiger counter or with photographic film. Each radioisotope has a characteristic disintegration or semi-life period. Energy may be liberated mostly in the form of alpha (helium nuclei), beta, (electrons or positrons), or gamma (electromagnetic energy) rays.

Several unstable and artificial radioactive isotopes have medical uses. For instance, a technetium isotope (99mTc) may be used to identify blocked blood vessels. Various natural radioactive isotopes are used to determine chronologies, such as the archeological kind (^{14}C).

Radioactive Half Lives

Each radioactive nuclide has a characteristic, constant **half-life** ($t_{1/2}$), the time required for half of the atoms in a sample to decay. An isotope's half-life allows us to determine how long a sample of a useful isotope will be available, and how long a sample of an undesirable or dangerous isotope must be stored before it decays to a low-enough radiation level that is no longer a problem.

For example, cobalt-60, an isotope that emits gamma rays used to treat cancer, has a half-life of 5.27 years (Figure1). In a given cobalt-60 source, since half of the $^{60}_{27}\mathrm{Co}$ nuclei decay every 5.27 years, both the amount of material and the intensity of the radiation emitted is cut in half every 5.27 years. Note that for a given substance, the intensity of radiation that it produces is directly proportional to the rate of decay of the substance and the amount of the substance. Thus, a cobalt-60 source that is used for cancer treatment must be replaced regularly to continue to be effective.Since every half-life for a radionuclide is the same length of time, we can use the following equation to calculate how much radioactive nuclide is remaining after the passage of any number (n) of half-lives:

A '**half-life**' is defined as the amount of time taken for the number of nuclei present in a sample at a given time to exactly halve. This value does not depend on the moment chosen: the amount of time taken for the nuclei to halve will always be the same. The number of nuclei that have not yet decayed diminishes very rapidly with the number of half-lives that pass. After ten half-lives, for instance, the radioactivity will have gone down by a factor of 1000.

Nuclei which decay easily have shorter half-lives, while those that have more difficulty last for longer.At any given moment, the number of remaining nuclei, the number of decays taking place and the number of various types of radiations emitted are all proportional and decrease at the same rate. This mathematical relationship with time allows for a calculation of the half-life. In fact, while it is almost impossible to count the number of nuclei, the decay process can be followed quite easily by measuring the radiations detected with, say, a Geiger counter .Half-life is a convenient way to assess the rapidity of a decay, but it should not be confused with the **average life span** of a radioactive nucleus. This average life span, in the case of a simple radioactive decay, is 1,443 times the half-life.

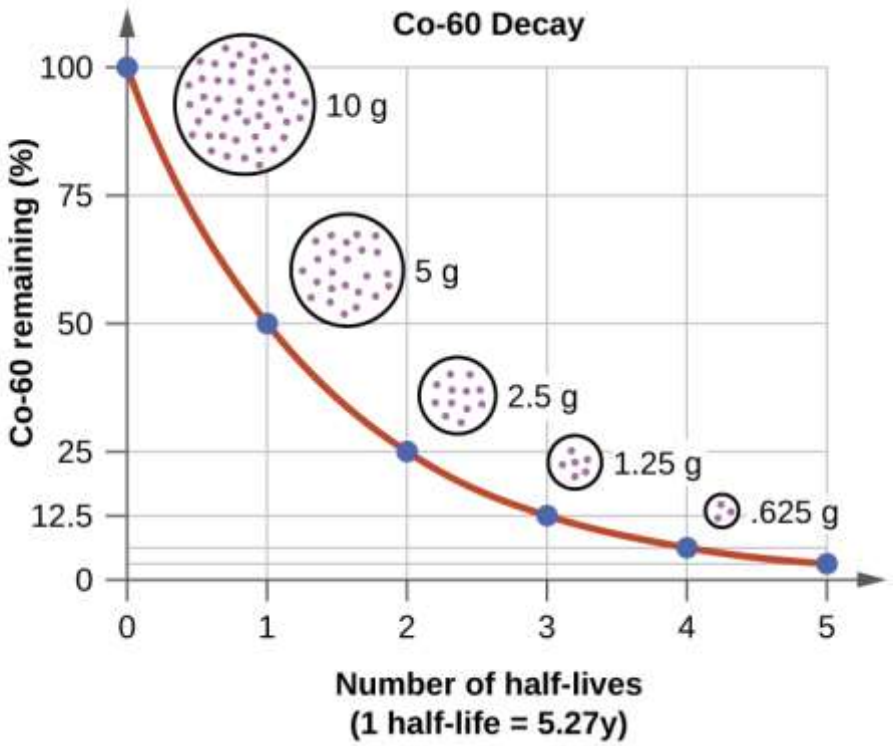


Figure 1. The Decay of Cobalt-60. For cobalt-60, which has a half-life of 5.27 years, 50% remains after 5.27 years (one half-life), 25% remains after 10.54 years (two half-lives), 12.5% remains after 15.81 years (three half-lives), and so on. Note that every half-life is the same length of time.

$$\text{Isotope Remaining} = \left(\frac{1}{2}\right)^{n} \times \text{Starting Material}$$

- *Where n = the number of half-lives determined*

Practice Problem:

Question: The half-life of Zn-71 is 2.4 minutes. If one had 100.0 g at the beginning, how many grams would be left after 7.2 minutes has elapsed?

Solution:

Step 1. Determine the number of half-lives that have passed: number of half-lives = time passed divided by the half-life (Be sure that the time units match!!)

$$\frac{7.2\ \text{min}}{2.4\ \text{min}} = 3\ \text{half-lives}$$

Step 2. Use the Isotope Remaining equation to solve for how much isotope will remain after the number of half-lives determined in step 1 have passed.

$$\text{Isotope Remaining} = \left(\frac{1}{2}\right)^3 \times 100.0\ \text{g}$$
$$= \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) \times 100.0\ \text{g} = 12.5\ \text{g}$$

Biological Effects of Radiation Exposure

There is a large difference in the magnitude of the biological effects of **nonionizing radiation** (for example, light and microwaves) and **ionizing radiation**, emissions energetic enough to knock electrons out of molecules (for example, α and β particles, γ rays, X-rays, and high-energy ultraviolet radiation) (Figure 2).

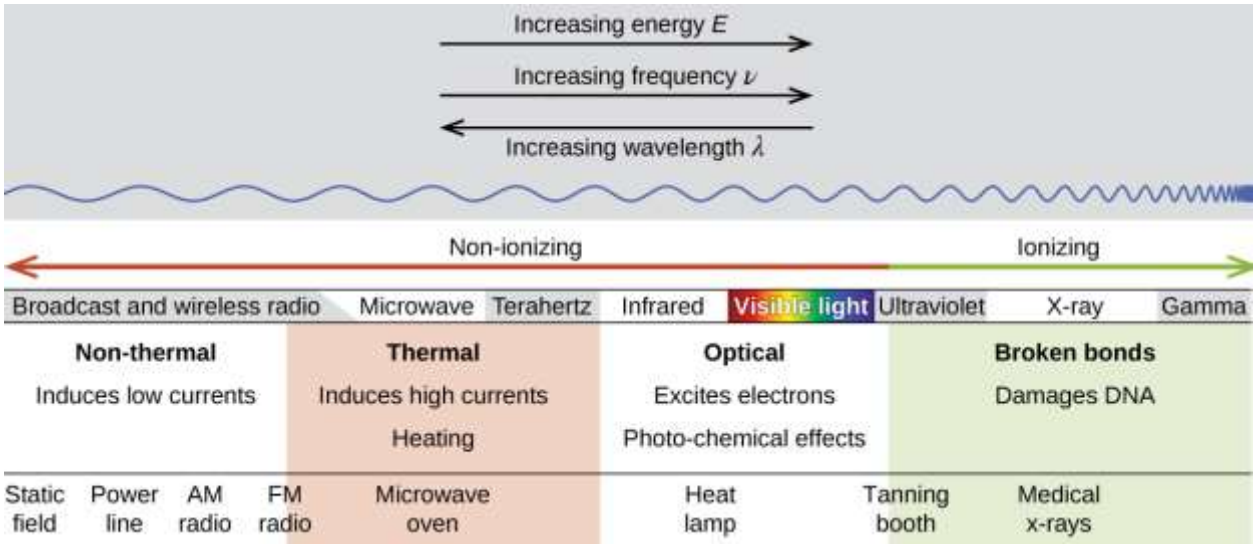


Figure 1. Damaging Effects of Ionizing Radiation. Lower frequency, lower-energy electromagnetic radiation is nonionizing, and higher frequency, higher-energy electromagnetic radiation is ionizing.

Energy absorbed from nonionizing radiation speeds up the movement of atoms and molecules, which is equivalent to heating the sample. Although biological systems are sensitive to heat (as we might know from touching a hot stove or spending a day at the beach in the sun), a large amount of nonionizing radiation is necessary before dangerous levels are reached. Ionizing radiation, however, may cause much more severe damage by breaking bonds or removing electrons in biological molecules, disrupting their structure and function (Figure 3).

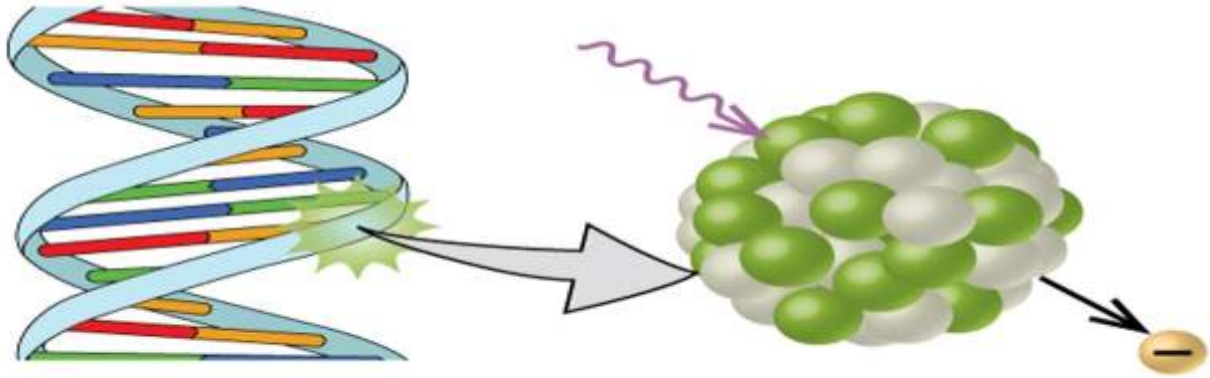


Figure 3. Biological Effects of Ionizing Radiation. Ionizing radiation can directly damage a biomolecule by ionizing it or breaking its bonds

Radiation can harm either the whole body (somatic damage) or eggs and sperm (genetic damage). Its effects are more pronounced in cells that reproduce rapidly, such as the stomach lining, hair follicles, bone marrow, and embryos. This is why patients undergoing radiation therapy often feel nauseous or sick to their stomach, lose hair, have bone aches, and so on, and why particular care must be taken when undergoing radiation therapy during pregnancy.

5.4. BIOLOGICAL EFFECTS OF EXPOSURE TO RADIATION

Radiation can harm either the whole body (somatic damage) or eggs and sperm (genetic damage). Its effects are more pronounced in cells that reproduce rapidly, such as the stomach lining, hair follicles, bone marrow, and embryos. This is why patients undergoing radiation therapy often feel nauseous or sick to their stomach, lose hair, have bone aches, and so on, and why particular care must be taken when undergoing radiation therapy during pregnancy.

Different types of radiation have differing abilities to pass through material (Figure 4). A very thin barrier, such as a sheet or two of paper, or the top layer of skin cells, usually stops alpha particles. Because of this, alpha particle sources are usually not dangerous if outside the body, but are quite hazardous if ingested or inhaled (see the Chemistry in Everyday Life feature on Radon Exposure). Beta particles will pass through a hand, or a thin layer of material like paper or wood, but are stopped by a thin layer of metal. Gamma radiation is very penetrating and can pass through a thick layer of most materials. Some high-energy gamma radiation is able to pass through a few feet of concrete. Certain dense, high atomic number elements (such as lead) can effectively attenuate gamma radiation with thinner material and are used for shielding. The ability of various kinds of emissions to cause ionization varies greatly, and some particles have almost no tendency to produce ionization. Alpha particles have about twice the ionizing power of fast-moving neutrons, about 10 times that of β particles, and about 20 times that of γ rays and X-rays.

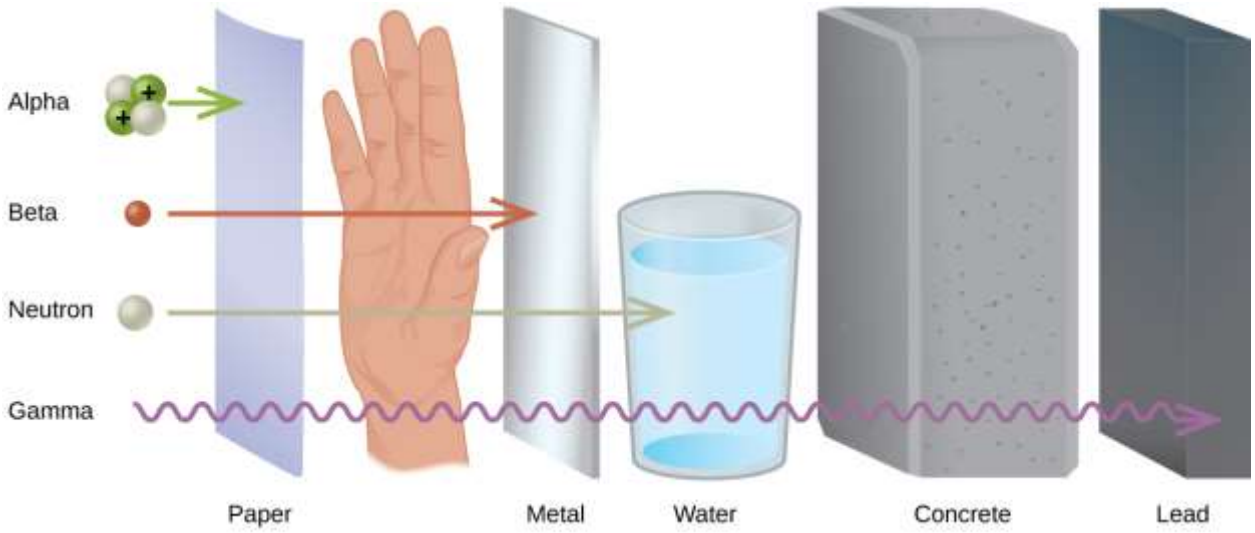


Figure 4. The ability of different types of radiation to pass through material is shown. From least to most penetrating, they are alpha < beta < neutron < gamma.

Radiation is one of the best-investigated hazardous agents. Because of the vast accumulation of quantitative dose-response data, specialists are able to set environmental radiation levels so that applications of nuclear technologies may continue at a level of risk that is much less than with many other technologies.

1. Acute and Delayed Effects

A single accidental exposure to a high dose of radiation during a short period of time is referred to as an acute exposure and may produce biological effects within a short period after exposure. These effects include:

1. Skin damage
2. Nausea and vomiting
3. Malaise and fatigue
4. Increased temperature
5. Blood changes
6. Bone marrow damage
7. Damage to cells lining the small intestine
8. Damage to blood vessels in the brain

The above list is given for information purposes only. The doses that can produce such effects are extremely unlikely even in the event of an accident at the U of T.

The delayed effects of radiation are due to both acute exposure and continuous exposure (chronic exposure). In this case, the negative effects may not be apparent for years. Chronic exposure is likely to be the result of improper or inadequate protective measures.

In the case of inhalation or ingestion of radioactive materials, a single "acute" event may cause a long period of "chronic" internal body exposure due to irradiation of tissue where radioactive material has been fixed.

The most common delayed effects are various forms of cancer (leukaemia, bone cancer, thyroid cancer, lung cancer) and genetic defects (malformations in children born to parents exposed to radiation). In any radiological situation involving the induction of cancer, there is a certain time period between the exposure to radiation and the onset of disease. This is known as the "latency period" and is an interval in which no symptoms of the disease are present. The minimum latency period for leukaemia produced by radiation is 2 years and can be up to 10 years or more for other types of cancer.2

2. **Dose-Effect Relationship**

The connection between the effects of exposure to radiation and dose (i.e., dose-response relationship) is classified into 2 categories, non-stochastic, and stochastic.

The non-stochastic effects, also referred to as deterministic or tissues and organs effects, are specific to each exposed individual. They are characterized by:

1. A certain minimum dose must be exceeded before the particular effect is observed. Because of this minimum dose, the non-stochastic effects are also called Threshold Effects. The threshold may differ from individual to individual
2. The magnitude of the effect increases with the size of the dose received by the individual
3. There is a clear relationship between exposure to radiation and the observed effect on the individual

Stochastic effects are those that occur by chance. They are more difficult to identify since the same type of effects may appear among individuals not working with radioactive materials. The main stochastic effects are cancer and genetic defects. According to current knowledge of molecular biology, a cancer is initiated by damaging chromosomes in a somatic cell. Genetic defects are caused by damage to chromosomes in a germ cell (sperm or ovum). There is no known existing threshold for stochastic effects. One single photon or electron can produce the effect. For these reasons, a stochastic effect is called a Linear or Zero-Threshold Dose-Response Effect.

Stochastic effects can also be caused by many other factors, not only by radiation. Since everybody is exposed to natural radiation, and to other factors, stochastic effects can arise in all of us

regardless of the type of work (working with radiation or not). Whether or not an individual develops the effect is simply a question of chance.

There is a stochastic correlation between the number of cases of cancers developed inside a population and the dose received by the population at relatively large levels of radiation. Attempts have been made to extrapolate the data from these levels of dose to low levels of dose (close to the levels received from background radiation). There is no scientific evidence to prove the results of these attempts.

Since there is no evidence of a lower threshold for the appearance of Stochastic Effects, the prudent course of action is to ensure that all radiation exposures follow a principle known as **ALARA (As Low As Reasonably Achievable)**. We will be referring to the application of this principle at U of T in subsequent modules.

3. Effects of Radiation on Foetus

It is well known that the foetus is more sensitive to the effects of radiation than an adult human. If an irradiation occurs in the first 30 weeks of pregnancy, delayed effects may appear in the child. These include mental and behaviour retardation, with a delay period of approximately 4 years.

Because of these possible effects, dosimetry during pregnancy differs from the usual protocol. Special attention is paid to both external and internal irradiation. A Radiation Safety Officer of the U of T must review procedures for handling radioactive materials when a pregnant worker performs such work.

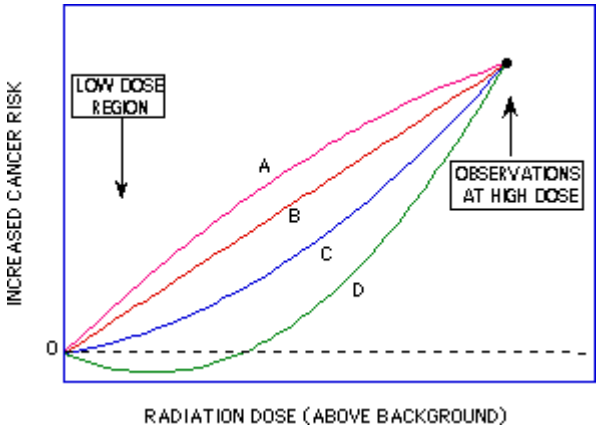
It is not possible to accurately measure the dose to the foetus and so it must be inferred from the exposure to the mother. Radiation protection principles limit exposure to the mother in order to achieve the minimum risk to the foetus.

4. Effects of very Low Radiation Levels

Exposure to very low levels of radiation is a controversial issue, originating many debates throughout the scientific community. What happens at very low levels of radiation exposure (a few percentages of the background), on top of the day to day natural irradiation), it is not known.

As was explained earlier, everybody is exposed to a level of radiation called natural radiation or background radiation. Also, was proved that the background levels vary on earth by a factor greater than 10.

There are not enough data to support extrapolations of the effects of high levels of exposure (like the survivors of Hiroshima and Nagasaki bombing, uranium miners, medical exposures, etc.). However, these extrapolations are made. There are four mathematical models use to describe the effects of the low level of radiation. All are supported (more or less) by controversial epidemiological studies, or by extrapolation conclusions obtained from studies with other mammals to humans.



Model A considers that at a low level there is an increased effect (above proportionality). Model B (also called the LNT model) considers than any increase in dose results in a proportional increase in cancer risk with no threshold. Model C considers that at low doses there is only a slight increase in risk that becomes proportional to dose at higher doses. Model D (also called the Hormesis model), considers that low doses of radiation have a positive effect and decreases the overall risk of cancer. In the absence of clear scientific evidence, the regulators adopted a conservative approach and consider all levels of radiation as being potentially damaging to the human body. Because of this, any procedure that involves radioactive materials must abide by the ALARA principle.

5.5. Autoradiography

Autoradiography is employed for the detection of materials that possess radioactive properties. By using X-ray films, autoradiography determine the relative positions and intensities of radiolabeled bands in a gel or blot. In 1867 the first autoradiography was observed accidently when an emulsion of silver chloride and iodide turns black by uranium salts. With the advent of photographic emulsions and photographic films after World War II, autoradiography was used as a biological technique for the detection of radioactive substances or materials labelled with radioactive isotopes.

Mechanism

Penetration of negatively charged beta particles emitted by radioactive salts through silver halide film emulsion causes activation of silver present in the emulsion. Activated silver crystals are very unstable therefore quickly reduced to black silver particles which is easily detectable. Autoradiography sensitivity is improved by carrying the detection process at 70°C and preflashing the film before use. Preflashing needs only one hit per crystal deposited to increases sensitivity. Autoradiography detection limits vary for different radioisotopes as given in the table below (Table 1) .

Sequential steps of autoradiography

- Brief exposure of living cells to a pulse of specific radioactive material for a variable time.
- Preparation of samples are for microscopy either light or electron.

- Dissection of samples into sections for coverage with thin film of photographic emulsion which are then incubated in the dark for few days for radioactive decay. The exposure time depends on isotope activity, temperature and the background radiation.
- Development of photographic emulsion.
- Toluidine blue is used for counter staining to reveal tissue histology. Instead Osmium or dipping emulsion can be used for pre-staining of the entire tissue before exposure to the photographic emulsion to avoid for individual post- staining each slide.
- Microscopy either light or electron is used to determine the relative position of the silver particles.
- Generation of records in the form of autoradiographs.

Isotope	Count per minute (CPM) for Detection	Energy per Emission (MEV)
^{3}H	$>10^{7}$	0.0055
^{14}C	2000	0.050
^{35}S	1000	0.167
^{32}P	100	0.70
^{125}I	10	Gamma

Table 1: Autoradiography detection limit.

Fluorography

Autoradiography sensitivity is greatly enhanced through fluorography which transforms radioactive emissions into light which efficiently penetrates the film to be readily detected. A number of phosphorus compounds absorb energy from beta particles and re-emit it as light e.g. Autofluor.

Advantages

- Technically easy not much expertise required,
- Highly specific detection tool,
- Unlike tissue bath preparations, pharmacologically characterize and localize receptors in tissues,
- Enables characterization of receptors in different tissues in different animals or brain regions .

Disadvantages

- Lack of assessment criteria to determine whether the binding site really corresponds to an actual receptor,
- Non-physiological significance of high affinity radiolabelled receptor,
- Non-specificity of ligands can easily cause misinterpretation of results.

Autoradiography practical applications

Autoradiography provides qualitative as well as quantitative information regarding a specimen. Some of the following applications of this technique are given below:

- Autoradiography is used to determine receptor distribution and localization while studying neurodegenerative disorders.
- Application of autoradiography in electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets during blotting .
- To study cytogenesis of the forebrain.
- Applications in radiopharmaceutical research.
- Applications in radioimmunoelectroosmophoresis to study viruses.
- In imaging and analyzing rock porosity .
- In matrix-assisted laser desorption/ionization mass spectrometric imaging (MALDI-MSI), and secondary ion mass spectrometric imaging (SIMS-MSI) for pharmaceutical discovery and development.
- In whole body imaging .
- Tool for genetic studies.
- For comparison of complex mixtures of proteins.
- Applications in microbial ecology.
- Determining gross absorption and utilization of foliar applied nutrients etc.

5.6. Cherenkov Radiation

The Phenomenon of Cherenkov Radiation

The Phenomenon of Cherenkov Radiation Bright blue light visible is used in movies from the deep sea to signify something mystic in that area. Have you ever wondered why? It is due to the emission of light called Cherenkov radiation. While filmmakers bright out this strange blue light using visual effects, it is actually due to a pretty interesting phenomenon. The accelerated speed of charged particles, such as electrons, in a dielectric medium, is the cause of this strange radiation. (The high speeds are greater than that of light, that is, more than 299,792,458 m/s.) When a charged particle passes through a dielectric medium, it emits electromagnetic radiation which is termed Cherenkov radiation. This radiation is named after the physicist Pavel Cherenkov.

Properties of Cherenkov Radiation

This radiation has a high frequency and is continuous. Due to its continuity, it does not have any characteristic peaks in its spectrum but is rather constant. Owing to its high frequencies it has short wavelengths and is also very intense. It thus emits blue light that falls in the ultraviolet region of the electromagnetic spectrum. With sucient accelerated charged particles it becomes visible to the naked eye. A common example of Cherenkov radiation is the blue light emitted by underwater nuclear reactors. the Cherenkov effect. Common transparent media where this effect is observed are water and air. The speed of light in water is approximately 200,000 km/sec and in the air, it is about 300,000 km/sec. To travel faster than light in water and exhibit this eect, a charged particle needs energy above 175 keV. Radioactive beta electrons often exhibit this eect while its an impossibility for the heavy and slow alpha particles. In the air, the energy demanded by Cherenkov light from the particles is greater than 21 MeV for a small ash of light. This is a far cry and is never fulfilled by radioactive electrons in the air.

During their journey, the electrons pass through many atoms and molecules that they encounter. The balancing of the medium is done by de-exciting photons. The de-excitation, which leads to the emission of photons, leads to the dissemination of blue light. This emission costs the photons, a mere amount of 2.5 eV energy. Cosmic radiation in the atmosphere exhibits the Cherenkov effect as it possesses electrons, positrons, and the high energy muons that are capable of producing Cherenkov light. The ashes produced from the light is used for the detection of cosmic showers.

Uses of Cherenkov Effect

Many experiments in physics use the Cherenkov effect. These include:

- For the Identification of Nature of Particles in High Energy
- In Astrophysics, While Studying Cosmic Showers

Observations made in astrophysics have shown that using the Cherenkov eect the properties of astronomical objects with highfrequency gamma rays can be determined and cosmic showers in space can be detected.

- Imaging of Radioactive Isotopes in Medicine

Recently, the Cherenkov light has been used to produce images of substances in the body. This attempt was aimed at imaging for diagnostic value demonstration and the radioactive elements used were uorine (13), iodine (131), nitrogen (13), phosphorus (32), and yttrium (90).

- For Detecting Labeled Biomolecules

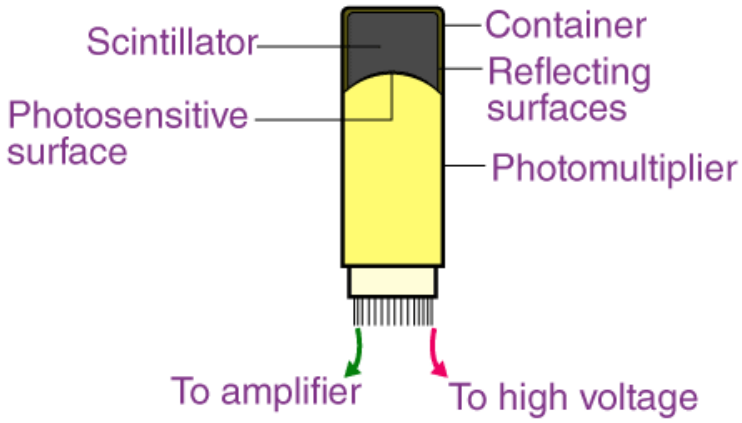
Selective biological molecules of low concentrations can be detected using Cherenkov radiation. On introducing radioactive elements by enzymatic and synthetic procedures, the affinity constants and dissociation rates are determined in the biomolecules.

5.7. Liquid scintillation spectrometry

Scintillation Counter is an instrument that is used for measuring ionizing radiation. “It comprises the scintillator that generates photons in response to incident radiation”, a PMT tube is used to convert an electronics and electric signal to process the signal.

A scintillation counter is used to detect gamma rays and the presence of a particle. It can also measure the radiation in the scintillating medium, the energy loss, or the energy gain. The medium can either be gaseous, liquid, or a solid. The scintillator counter is generally comprised of transparent crystalline material such as glasses, liquids, or plastics. One sector of the scintillators is placed

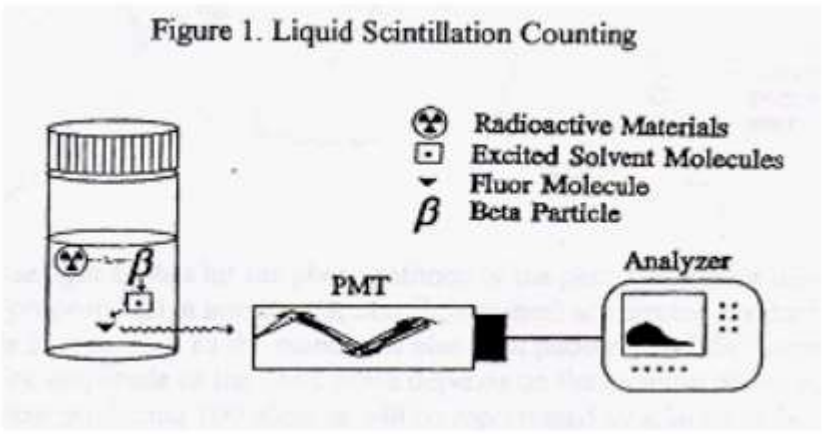
(optical contact) with the pin code.



A charged particle loses energy when passing through the scintillator thus leaving the trail of excited molecules and atoms. A rapid interatomic transfer of electronic excitation energy follows, which leads to the burst of scintillator material luminescence characteristics. The scintillation response, when a particle stops leading to the light output. The energy loss of a particle is measured when a particle passes completely through a scintillator.

LIQUID SCINTILLATION PRINCIPLES Figure I provides a graphic illustration of the way the emitted radiation interacts with the cocktail (a mixture of a solvent and a solute) leading to a count being recorded by the system.

Figure 1. Liquid Scintillation Counting



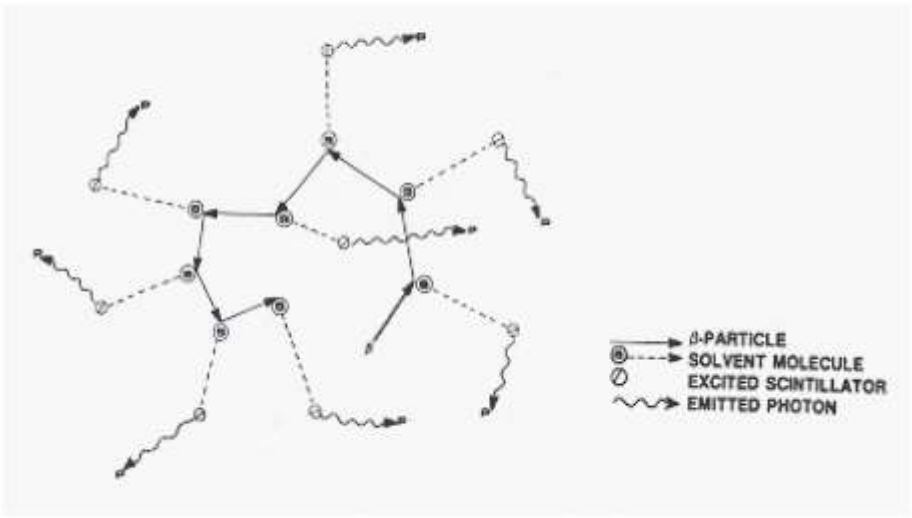
Step 1. Beta particles are emitted in a radioactive decay. To assure efficient transfer of energy between the beta particle and the solution, the solution is a solvent for the sample material.

Step 2. In the relatively dense liquid, the beta particle travels only short distances before all of its kinetic energy is dissipated. Typically a beta particle will take only a few nanoseconds to dissipate all its kinetic energy. The energy is absorbed by the medium in 3 forms: heat, ionization and

excitation. Some of the beta energy is absorbed by solvent molecules making them excited (not ionized).

Step 3. Energy of the excited solvent is emitted as UV light and the solvent molecule returns to ground state. The excited solvent molecules can transfer energy to each other and to the solute (Figure 2). The solute is a fluor. An excited solvent molecule which passes its energy to a solute molecule disturbs the orbital electron cloud of the solute raising it to a state of excitation. As the excited orbital electrons of the solute molecule return to the ground state, a radiation results, in this case a photon of UV light. The UV 2 light is absorbed by fluor molecules which emit blue light flashes upon return to ground state. Nuclear decay events produce approximately 10 photons per keV of energy. The energy is dissipated in a period of time on the order of 5 nanoseconds. The total number of photons from the excited fluor molecules constitutes the scintillation. The intensity of the light is proportional to the beta particle's initial energy.

Figure 2. Illustration of the Collision Process



Step 4. Blue light flashes hit the photo cathode of the photo multiplier tube (PMT). Electrons (proportional in number the blue light pulses) are ejected producing an electrical pulse that is proportional to the number of blue light photons. A LSC normally has two PMT's. The amplitude of the PMT pulse depends on the location of the event within the vial. An event producing 100 photons will be represented by a larger pulse if the event is closer to the PMT than if the event is more remote. The signal from each PMT is fed into a circuit which produces an output only if the 2 signals occur together, that is within the resolving time of the circuit, approximately 20 nanoseconds (coincidence circuit). By summing the amplitude of the pulses from each PMT, an output is obtained which is proportional to the total intensity of the scintillation. This analog pulse rises to its maximum amplitude and falls to zero.

Step 5. The amplitude of the electrical pulse is converted into a digital value and the digital value, which represents the beta particle energy, passes into the analyzer where it is compared to digital values for each of the LSC's channels. Each channel is the address of a memory slot in a multi-

channel analyzer which consists of many storage slots or channels concerting the energy range from 0-2000 keV.

Step 6. The number of pulses in each channel is printed out or displayed on a CRT. In this manner, the sample is analyzed and the spectrum can be plotted to provide information about the energy of the radiation or the amount of radioactive material dissolved in the cocktail.

Applications of Scintillation Counter

- Scintillation Counters are widely used in radioactive contamination, radiation survey meters, radiometric assay, nuclear plant safety, and medical imaging, that are used to measure radiation.
- There are several counters of mounted on helicopters and some pickup trucks for rapid response in case of a security situation due to radioactive waste or dirty bombs.
- Scintillation counters designed for weighbridge applications, freight terminals, scrap metal yards, border security, contamination monitoring of nuclear waste, and ports.
- It is widely used in screening technologies, In vivo and ELISA alternative technologies, cancer research, epigenetics, and Cellular research.
- It also has its applications in Protein interaction and detection, academic research, and Pharmaceutical.
- Liquid Scintillation Counter is a type of scintillation counter that is used for measuring the beta emission from the nuclides.