NETAJI SUBHAS OPEN UNIVERSITY

STUDY MATERIAL

POST GRADUATE
ZOOLOGY

Paper - 6

Group : A(I) & A(II)

Quantitative Biology &
Biotechnology

PREFACE

In the curricular structure introduced by this University for students of Post-Graduate Degree Programme, the opportunity to pursue Post-Graduate course in any subject introduced by this University is equally available to all learners. Instead of being guided by any presumption about ability level, it would perhaps stand to reason if receptivity of a learner is judged in the course of the learning process. That would be entirely in keeping with the objectives of open education which does not believe in artificial differentiation.

Keeping this in view, study materials of the Post-Graduate level in different subjects are being prepared on the basis of a well laid-out syllabus. The course structure combines the best elements in the approved syllabi of Central and State Universities in respective subjects. It has been so designed as to be upgradable with the addition of new information as well as results of fresh thinking and analysis.

The accepted methodology of distance education has been followed in the preparation of these study materials. Co-operation in every form of experienced scholars is indispensable for a work of this kind. We, therefore, owe an enormous debt of gratitude to everyone whose tireless efforts went into the writing, editing and devising of proper lay-out of the materials. Practically speaking, their role amounts to an involvement in 'invisible teaching'. For, whoever makes use of these study materials would virtually derive the benefit of learning under their collective care without each being seen by the other.

The more a learner would seriously pursue these study materials, the easier it will be for him or her to reach out to larger horizons of a subject. Care has also been taken to make the language lucid and presentation attractive so that they may be rated as quality self-learning materials. If anything remains still obscure or difficult to follow, arrangements are there to come to terms with them through the counselling sessions regularly available at the network of study centres set up by the University.

Needless to add, a great deal of these efforts is still experimental–in fact, pioneering in certain areas. Naturally, there is every possibility of some lapse or deficiency here and there. However, these do admit of rectification and further improvement in due course. On the whole, therefore, these study materials are expected to evoke wider appreciation the more they receive serious attention of all concerned.

Professor (Dr.) Subha Sankar Sarkar
Vice-Chancellor

Third Reprint : January, 2020

Printed in accordance with the regulations of the
Distance Education Bureau of the University Grants Commission.

POST GRADUATE ZOOLOGY

[M.Sc]

PAPER : PGZO - 6
GROUP : A (I)

Writer & Editor
Prof. Debajyoti Das

PAPER : PGZO - 6
GROUP : A (II)

	Writer	Editor
Unit 1	Dr. Jana Chakrabarti	Prof. S. K. Dey
Unit 2	Dr. Suman Das	do
Unit 3	Dr. Suman Das	do
Unit 4	Dr. Subir Ch. Dasgupta	do
Unit 5	Dr. Subir Ch. Dasgupta	do
Unit 6	Dr. Jana Chakrabarti	do
Unit 7	Dr. Jana Chakrabarti	do
Unit 8	Dr. Jana Chakrabarti	do
Unit 9	Dr. Suman Das	do
Unit 10	Dr. Subir Ch. Dasgupta	do

Netaji Subhas Open University

PGZO–6
Quantitative Biology & Biotechnology, Immunology & Microbiology

Group A(I)
Quantitative Biology

Group A(II) Biotechnology

Unit 1 ❑ BASIC MATHEMATICS

Structure

1.1 Introduction

In this unit, you will get a glimpse of the basic mathematical principles and techniques which find applications in the exploration, treatment, analysis and interpretation of scientific events and phenomena. During the earlier epochs of development of life sciences, scientific investigations were initially confined largely to qualitative pursuits such as nature study, search for new species, identifications, morphological distinctions, macro and micro anatomical structures, fundamental organ and tissue functions, and systematic classifications of living organisms. But since the early decades of the last century in particular, quantitative approaches to life science problems have been developing by leaps and bounds with commensurate increase in the application of mathematical techniques to the collection, presentation, analysis and interpretation of biological data. Samuel Brody's monumental book, *Bioenergetics and Growth,* eminently deserves mention as a landmark in this field in the early fifties of the twentieth century. Such mathematical applications largely consist of the use of biostatistics. Principles and practice of biostatistics have consequently evolved considerably and received a high impetus from the emergence of successive computer generations. Some basic mathematics, currently in practice in life sciences, will find very brief mention in the present unit.

Objectives

After reading this unit, you should be able to do the following :

- recall some basic mathematical principles and methods which find use in biological sciences,
- define and use common fractions, multiples and exponents,
- recall the types, bases and uses of logarithms,
- understand the properties and uses of arithmetic, geometric and logarithmic series,
- learn the importance and working out of equations of linear relationships,
- know the nature and use of equations and graphs of exponential relations, and
- understand the importance of mathematics in biological work.

1.2 Fractions, multiples and exponents

You may recall here some important points helpful in dealing with common fractions, multiples and powers or exponents.

1.2.1. Common fractions

A common fraction is a ratio of two numbers; it is constituted by dividing a number called the *numerator* with another number called the *denominator*. Its basic properties are as follows.

(i) If the same number (a) is either added to or subtracted from both the numerator (X) and the denominator (Y) of a fraction, the latter changes into a new fraction. Thus,

$$\frac{X}{Y} \neq \frac{X+a}{Y+a}; \quad \frac{X}{Y} \neq \frac{X-a}{Y-a}.$$

(ii) If both the numerator X and the denominator Y are multiplied (or divided) by the same number a, the fraction neither changes in value nor equals its product (or division result) with the constant number a. Thus,

$$\frac{X}{Y} = \frac{Xa}{Ya} \neq a\frac{X}{Y}.$$

(iii) A fraction changes in value if both the numerator X and the denominator Y are either raised to a common power a or reduced to the same root a. Thus,

$$\frac{X}{Y} \neq \frac{X^a}{Y^a}; \quad \frac{X}{Y} \neq \frac{\sqrt[a]{X}}{\sqrt[a]{Y}}, \text{ or } \frac{X}{Y} \neq \frac{X^{\frac{1}{a}}}{Y^{\frac{1}{a}}}.$$

(iv) A low fraction is formed by adding two fractions with the same denominator;

for this, their numerators are added together to form a new numerator while the denominator is left to remain the same as before. Thus,

$$\frac{X}{Y}+\frac{Z}{Y}=\frac{X+Z}{Y}.$$

(v) Similarly, two fractions with a common denominator can be subtracted from one another by the subtraction of the numerator of one from that of the other, while retaining the original denominator. Thus,

$$\frac{X}{Y}-\frac{Z}{Y}=\frac{X-Z}{Y}.$$

(vi) If two common fractions with different denominators (Y and b) are to be added to or subtracted from one another, the lowest common multiple (LCM), say bY, of the two denominators is first computed and then used for working out the resulting sum or difference between the numerators. Thus.

$$\frac{X}{Y}-\frac{a}{b}=\frac{bX-aY}{bY}; \quad \frac{X}{Y}+\frac{a}{b}=\frac{bX+aY}{bY}.$$

(vii) To form a new fraction by multiplying two fractions, the product of the two numerators and that of the two denominators constitute respectively the numerator and the denominator of the new fraction. Thus,

$$\frac{X}{Y}\times\frac{a}{b}=\frac{Xa}{Yb}.$$

(viii) To divide a common fraction by another, the fraction to be divided is multiplied by the reciprocal of the divisor fraction. Thus,

$$\frac{X}{Y}+\frac{a}{b}=\frac{X}{Y}\times\frac{b}{a}=\frac{Xb}{Ya}.$$

1.2.2. Multiples

If a score Y is the multiple of another score X, it means that Y is obtained by multiplying a given value of X by a constant number a : $Y = aX$. Stated otherwise, Y is considered as a *function* of X and varies with the value of the latter. It follows that if there occurs a series of very high or very low Y scores, the scale of the latter may be moderated by downward or upward movement, expressing each Y score of the series as a function of the common denominator X having respectively a very high or a very low value. Such a change of scale of a series of too high or too low numbers makes it convenient to use the series for computation as well as comprehension. In this way, the series of numbers such as 34300, 35300, 36300, ...etc., may be scaled down to the series 3.43×10^4, 3.53×10^4, 3.63×10^4, ...etc. You may recall that similar expressions of physical quantities in multiples of 10 are followed in the scales of standard international (SI) units.

1.2.3. Powers or exponents

An *exponent* or *index* (i) is any power to which a score X may be raised, while the score thus raised to a given power is called the *base* of the latter. For example, in case of 10^4, the base is 10 while the power or exponent is 4. If the raising of a score X (base) to the power a (exponent) gives the product Y, the latter is a *function* of X and varies with the value of the base as also with the value of the exponent; the relation between Y and X may thus be expressed as : $Y = X^a$. Irrespective of the values of X and Y, any Y score may be obtained by raising the base X to a specific power or exponent; the latter is not necessarily an integral number, and may even be a negative, fractional or decimal number. You may, therefore, realise that the Y scores can be expressed as *exponential terms* dependent on the exponent or power of the base X. By varying the exponent of an identical base X, a series of Y scores may be converted to a series of respective exponential terms. If the exponential terms have been derived using the same base, algebraic computations with them become far more convenient than with the Y scores themselves, particularly if the latter scores are too large or too small. Thus, if Y_1, Y_2 and Y_3 scores are changed to exponential terms X^a, X^b and X^c respectively, using the common base X,

$$Y_1Y_2Y_3 = X^aX^bX^c = X^{(a+b+c)};$$

$$Y_1 \div Y_3 = X^a \div X^c = X^{(a-c)}; \quad \frac{Y_1Y_2}{Y_3} = X^{(a+b-c)};$$

$$\frac{Y_1}{\sqrt{Y_3}} = \frac{X^a}{\sqrt{X^c}} = X^{(a-\frac{c}{2})}; \qquad Y_1\sqrt{Y_3} = X^a\sqrt{X^c} = X^{(a+\frac{c}{2})}$$

1.3 Logarithms

Logarithm of any number Y is the exponent, power or index (i) of its exponential term using a base X which is generally chosen to be either 10 in case of *common or Briggsian logarithms* or the base e in case of *natural or Naperian logarithms; e* is an infinitely long number and approximates to 2.7183. The Naperian logarithm ($\log_e$ or ln) of any number X is related to the logarithm (log) of that number to the base 10 in the following way :

$$\log_e X \text{ or } \ln X = 2.303 \log X.$$

If any score Y has an exponential term X^i and if X is made the base of a set of logarithms, then the logarithm of Y to the base X ($\log_X Y$) would be given by the index i of the corresponding exponential term of that Y score. Thus, logarithms to any chosen base — even other than 10 or e — can be worked out and used.

Every logarithmic number is a decimal number with one part called the *characteristic* and another part called the *mantissa,* respectively preceding and following the decimal point. While the mantissa is always positive, the characteristic is either zero or an integral number which may be either positve or negative, the negativity being indicated by a superscript bar above the numerical value of the characteristic.

1.3.1. Using the common logarithm table

The logarithmic table is used to find the log X of a given X score under consideration. Ignoring any decimal in the number X, the first 4 figures from the table, corresponding to the figures of the given number X, are noted as the mantissa, reading the figures of X starting from the extreme left column of the table and then moving to the right along the corresponding row. A decimal point is now placed on the left of the first figure of the noted mantissa.

Next, the number of figures is counted on the left of the decimal point in the original X score. If that figure consists of a single digit, zero is entered to the left of the decimal point before the noted mantissa; but if two or more digits occur on the left of the decimal point in the score X, then the number, less by 1 than that number of digits in X, is entered on the left of the decimal preceding the mantissa; in case only a zero precedes the decimal in the original X score, then the number to be entered before the decimal of the mantissa should exceed by 1 the number of zeroes following the decimal in the X score, and should bear a superscript bar to indicate the negativity of the logarithmic number. Thus,

log 111.4 = 2.0453; log 11.12 = 1.0461; log 2.429 = 0.1750;

log 0.4952 = $\bar{1}$.3127; log 0.0731 = $\bar{2}$.5383.

For any logarithmic number $\log_i X$, the original X score corresponding to the latter is called its *antilogarithm.* To find X from $\log_i X$, the antilogarithm table has to be used : X = antilog ($\log_i X$). For this, the number corresponding to the mantissa of the logarithmic number is noted from that table; then, considering the figure in the characteristic of the logarithmic number and following a procedure reverse to that used in recording the characteristic, a decimal point is inserted in the noted number corresponding to the mantissa. Thus, for the common logarithms to the base 10 ($\log_{10} X$) :

antilog 0.6207 = 4.176; antilog 1.5935 = 39.22;

antilog $\bar{1}$.6554 = 0.4523; antilog $\bar{2}$.5540 = 0.0358.

1.3.2. Logarithms in multiplication, division and power

Logarithmic procedures for these computations are like those using exponents (see Section 1.2.3). Thus, using X and Y secores,

XY = antilog (log X + log Y); $\frac{X}{Y}$ = antilog (log X – log Y);

X^Y = antilog (Y log X); Y^X = antilog (X log Y);

$\sqrt[Y]{X}$ = antilog $\left[\frac{\log X}{Y}\right]$; $\sqrt[X]{Y}$ = antilog $\left[\frac{\log Y}{X}\right]$.

Example 1.3.1.

Solve the following using common logarithms :

$$\frac{7 \cdot 251^3 + 2 \cdot 752}{4 \cdot 351 \times 16 \cdot 35 + 4 \cdot 043^2}.$$

Solution :

log (numerator) = log ($7.251^3 \div 2.752$) = 3 log 7.251 – log 2.752
= 3 × 0.8604 – 0.4396 = 2.1416.

log (denominator) = log ($4.351 \times 16.35 \div 4.043^2$)
= log 4.351 + log 16.35 – 2 log 4.043
= 0.6386 + 1.2135 – 2 × 0.6067 = 0.6387.

$\log \left[\frac{\text{numerator}}{\text{denominator}}\right]$ = log (numerator) – log (denominator)
= 2.1416 – 0.6387 = 1.5029.

antilog 1.5029 = 31.84.

1.4 Arithmetic and geometric progressions

Some important aspects of these two series or progressions are mentioned below.

1.4.1. Arithmetic progressions

Each arithmetic progression (AP) consists of a sequential arrangement of real numbers in a series in which successive numbers *differ* from each other *by an identical number.* Whatever be the position in the series where a pair of such successive numbers is located, the difference between any two successive numbers would be the same and is called the *constant difference* (d). Any term in an AP, except the first one (X_1), is given by the *sum* of the immediately preceding term or number and the constant difference d of the series. Examples of such series are : 4, 9, 14, 19, 24,....; 15, 10, 5, 0, –5, –10,.....; 0.35, 0.76, 1.17, 1.58, 1.99, ...; the constant difference amounts to 5, –5 and 0.41 in them, respectively. The numbers in an AP series would be in an ascending order if d exceeds 0 and in a descending order when d is lower than 0.

It follows that in an AP series, viz., $X_1, X_2, X_3, \ldots\ldots X_n$, where d is the common difference and k is a positive integer, the kth term (X_k) is given by : $X_k = X_1 + (k-1)d$.

If an AP has X_1, X_2 and X_3 as three of its terms, then : $X_2 - X_1 = X_3 - X_2 = d$, or, $2X_2 = X_1 + X_3$, $\therefore X_2 = \frac{X_1 + X_3}{2}$. Thus, X_2 is the *arithmetic mean* (AM) between X_1 and X_3. If X_1 and X_2 are two real numbers, any number of arithmetic means can be inserted between them.

Also, if $X_1, X_2, X_3, \ldots\ldots X_k$ form an AP, a constant number (c) can be added to, subtracted from, multiplied with or used in dividing each term of that series, to form other AP sequences like :

$$\frac{X_1}{c}, \frac{X_2}{c}, \frac{X_3}{c}, \ldots\ldots \frac{X_k}{c}; \quad cX_1, cX_2, cX_3, \ldots\ldots cX_k;$$

$$X_1 + c, X_2 + c, X_3 + c, \ldots\ldots X_k + c; \qquad X_1 - c, X_2 - c, X_3 - c, \ldots\ldots X_k - c.$$

If d is the common difference, and X_1 and X_k are respectively the first and the kth terms of an AP, the sum of the first k number of terms is given by :

$$\Sigma(X_1 + X_2 + \ldots\ldots + X_k) = \frac{k}{2}[2X_1 + (k-1)d] = \frac{k}{2}(X_1 + X_k).$$

For scores of a variable like length and mass, that differ arithmetically from each other, scales are used with divisions arranged in arithmetic progressions.

1.4.2. Geometric progressions

Each geometric progression (GP) consists of a sequential arrangement of non-zero real numbers in a series in which the successive numbers form a *constant ratio*. Except the first term or number of every series, each of the other terms is given by multiplying — not by adding — a constant non-zero real number with the preceding number of the series; this constant non-zero real number, with which to multiply each term to get the next term, is called the *common ratio* (r) of that series and is obtained by dividing the number of any term by that of the preceding term :

$$r = \frac{X_2}{X_1} = \frac{X_3}{X_2} = \ldots\ldots = \frac{X_K}{X_{K-1}}.$$

Moreover, the arithemetic difference between two successive terms varies with the location of that pair in the series. Examples of GP are : 3, 9, 27, 81, ...; and 48, 12, 3, $\frac{3}{4}, \frac{3}{16}, \ldots\ldots$. In the first series, each term is obtained by multiplying the immediately preceding term by the common ratio 3, the ratio of each two successive terms amounts to $\frac{1}{3}$, and differences between the terms of the three successive pairs

amount respectively to 6, 18 and 54; in the second series, each number is obtained by multiplying the immediately earlier one by the common ratio $\frac{1}{4}$, the ratio of successive terms amounts to 4, and differences between the terms of successive pairs amount respectively to –36, –9, $-2\frac{1}{4}$ and $-\frac{9}{16}$.

Where the common ratio (r) is positive and higher than 1, a GP of positive terms has its terms arranged in an ascending order (e.g., the first series quoted above) while a GP of negative terms has its terms in a descending order. On the contrary, a GP of positive terms has its terms decreasing progressively (e.g., the second series quoted above) while that of negative terms has its terms increasing progressively, if the common ratio is positive but lower than 1.

If $X_1, X_2, \ldots\ldots X_k$ form a GP, other GP sequences are formed by treating each of those terms, as follows, with a constant c.

$$\frac{X_1}{c}, \frac{X_2}{c}, \frac{X_3}{c}, \ldots\ldots\ldots \frac{X_k}{c}; \qquad cX_1, cX_2, cX_3, \ldots\ldots\ldots cX_k;$$

$$X_1, \frac{c}{X_2}, \frac{c}{X_3}, \ldots\ldots\ldots \frac{c}{X_k}; \qquad c^{X_1}, c^{X_2}, c^{X_3}, \ldots\ldots\ldots c^{X_k}.$$

If X_1, X_2 and X_3 are in geometric progression as three of its terms, X_2 is the *geometric mean* (GM) between X_1 and X_3.

If r is the common ratio, X_1, X_{k-1} and X_k are respectively the first, $(k-1)$th and kth terms of a GP, and k is a positive integer, then the kth term is given by :

$$X_k = X_1 r^{k-1}.$$

Also, the sum of the first k number of terms of the GP is given by :

$$\Sigma(X_1 + X_2 + \ldots\ldots\ldots + X_k) = \frac{X_1(r^k - 1)}{r - 1}.$$

Differences between an AP and a GP would be clear to you on comparing the statements in Sections 1.4.1 and 1.4.2.

It follows that the numbers forming the successive terms of a geometrical series follow a progressive *exponential change*. For example, the GP of 1, 2, 4, 8, 16, ... may be considered as $2^0, 2^1, 2^2, 2^3, 2^4, \ldots\ldots$; so, using any base for logarithms of the terms of a geometric series, the latter may be changed into a *logarithmic series* in the form of an *arithemetic progression*, such as 0, 1, 2, 3, 4, ... in case of the GP quoted above.

For variables whose values differ exponentially, scales are used with their divisions arranged in the forms of geometric progressions.

Example 1.4..1.

(a) Work out the 6th term and the sum of the first six terms in each of the following series : (i) 4, 9, 14, ...; (ii) 3, 9, 27, ...

(b) Change the following geometric series to a common logarithmic series : 10, 100, 1000, 10000, ...

Solution :

(a) X_6 *and* $\overset{6}{\Sigma}$:

(i) The series 4, 9, 14, ... constitutes an AP.

$d = X_2 - X_1 = 9 - 4 = 5. \qquad k = 6.$

$X_k = X_1 + (k-1)\,d = 4 + (6-1)\,5 = 29.$

$$\Sigma(X_1 + X_2 + \ldots\ldots\ldots + X_k) = \frac{k}{2}[2X_1 + (k-1)d] = \frac{6}{2}[2\times 4 + (6-1)5] = 99;$$

or, $$\Sigma(X_1 + X_2 + \ldots\ldots\ldots + X_k) = \frac{k}{2}(X_1 + X_k) = \frac{6}{2}(4+29) = 99.$$

(ii) The series 3, 9, 27, ... constitutes a GP.

$$r = \frac{X_2}{X_1} = \frac{9}{3} = 3. \qquad k = 6.$$

$$X_k = X_1 r^{k-1} = 3 \times 3^{6-1} = 729.$$

$$\Sigma(X_1 + X_2 + \ldots\ldots\ldots + X_k) = \frac{X_1(r^k - 1)}{r-1} = \frac{3(3^6 - 1)}{3-1} = 1092\ .$$

(b). *Change to log series :*

GP : 10, 100, 1000, 10000, ...

Exponential form : 10^1, 10^2, 10^3, 10^4, ...

Logarithmic series : 1, 2, 3, 4, ...

1.5 Equation for straight line

Where two events (X and Y) bear such an association with one another that equal changes of one are accompanied by uniform changes of the other, the rate of changes of the latter does not vary with the changes of the former, and the plotted points of their scores in a number of cases lie distributed on or close to a straight line. Such an association is called *linear*, and either variable may be considered as a *linear function* of the other. Such linear relations can be graphically represented by plotting straight lines, using the following *linear equation* (Fig. 1.1) : $Y = a + bX$. In this

equation, b is the *slope* of the line; it is the measure of the *average rate of change* of the variable or event Y scaled along the ordinate (y-axis) for unit changes of the event X scaled along the abscissa (x-axis). The higher the value of b, the steeper is

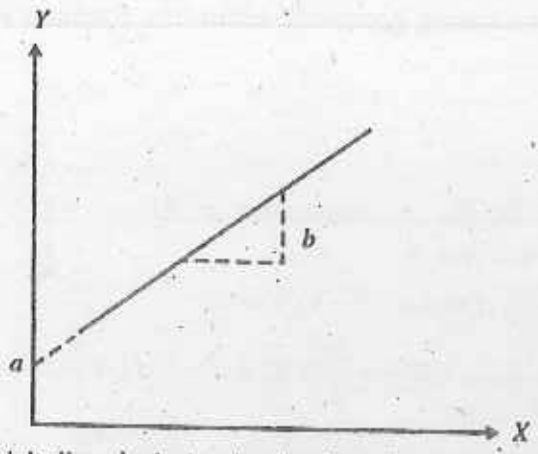


Fig 1.1. A straight line graph showing its slope (b) and its y-intercept (a).

the gradient of the line and the higher is the *magnitude* of the average rate of change of Y, associated with unit changes of X. The algebraic sign of b indicates the *relation* between the directions of changes of the two events; a positive b indicates the simultaneous rises (or falls) in the values of both events with a consequent *ascending gradient* of the line, while a negative b signifies the changes of the two events in opposite directions with a consequent *descending gradient* of the line.

The term a gives the *y-intercept* of the line and indicates its general level. It may be defined as the expected value of Y when X amounts to zero, and depends on the slope b of the line — rise and fall in the value of b are associated respectively with the fall and rise in the y-intercept.

Any point on the line is located by its perpendicular distances, called its *co-ordinates*, from the y and x axes. If the plotted points of Y versus X lie dispersed around the line, the latter is drawn using the *method of least squares* so as to keep the sum of squared vertical distances of the plotted points from the line at a minimum — in other words, the line drawn using the dispersed points would be the *best-fitting straight line* for those points. Examples of three such linear plots are given below.

(i) Rate of diffusion (J) of a solute across a thin permeable membrane is a linear function of the change (Δn) in its concentration across that membrane. Scaling J along the y-axis and Δn along the x-axis, and plotting J against the corresponding Δn values, we get a straight line which has a slope P equalling the permeability coefficient for the solute and its y-intercept at the origin of the y-scale, i.e., at zero values of both J and Δn ($X = 0$, $Y = 0$). This line is given by the following linear equation (Fig.1.2) :

$$J = P\Delta n, \quad \text{or} \quad J = O + P\Delta n,$$

which conforms to the equation $Y = a + bX$, where a or the y-intercept is zero.

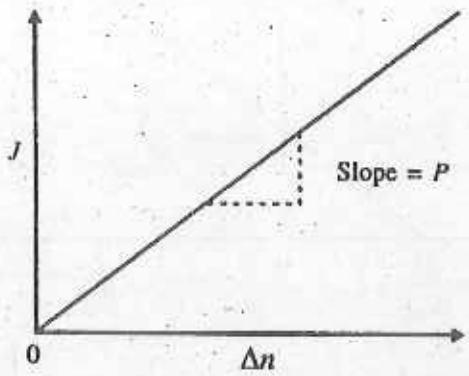


Fig 1.2. Linear plot of diffusion rate (J) against the transmembrane solute concentration gradient (Δn). [From D. Das, *Biophysics and Biophysical Chemistry*, 6th ed., Academic Publishers, 2007.]

(ii) *Lineweaver-Burk double-reciprocal plot* expresses the linear relation between the reciprocals of the initial velocity (V_0) of enzyme action and the substrate concentration [S]. Where V_{max} is the maximum V_0 with saturating [S] concentration, and K_m is the [S] for attaining $\frac{1}{2}V_{max}$, the linear equation of $\frac{1}{V_0}$ versus $\frac{1}{[S]}$ is as follows :

$$\frac{1}{V_0} = \frac{1}{V_{max}} + \frac{K_m}{V_{max}} \times \frac{1}{[S]}.$$

On scaling $1/V_0$ and $1/[S]$ along respectively the y and x axes, and plotting $1/V_0$ against the corresponding $1/[S]$ values, we get a straight line having K_m/V_{max} as its *slope*, $1/V_{max}$ as its *y-intercept* and $-1/K_m$ as its *x-intercept* (Fig. 1.3).

(iii) *Eadie-Hofstee plot* also expresses the linear relation between V_0 and [S] using a different linear equation. (Fig. 1.4).

$$V_0 = V_{max} - \frac{V_0}{[S]} K_m.$$

On scaling V_0 and $V_0/[S]$ along the y-axis and the x-axis respectively, and plotting V_0 against $V_0/[S]$ values, a straight line may be drawn (Fig. 1.4). The *slope* of the straight line, drawn using the linear equation quoted above, is downward and amounts to $-K_m$, the *y-intercept* is V_{max} and the *x-intercept* is V_{max}/K_m.

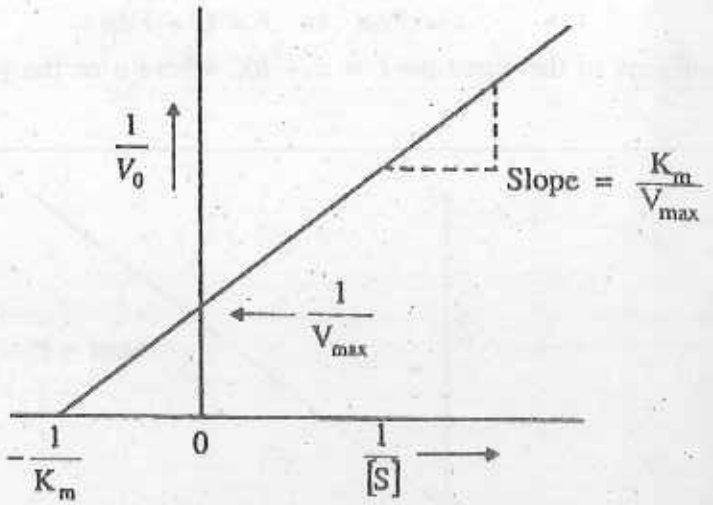


Fig 1.3. Double-reciprocal plot. [From D. Das, *Biochemistry*, 12th ed., Academic Publishers, 2005.]

1.6 Equation for exponential curve

Exponential relation between two variables or events (X and Y) can be expressed by an equation using the base (e) of natural logarithms and two constants (Z and k) :

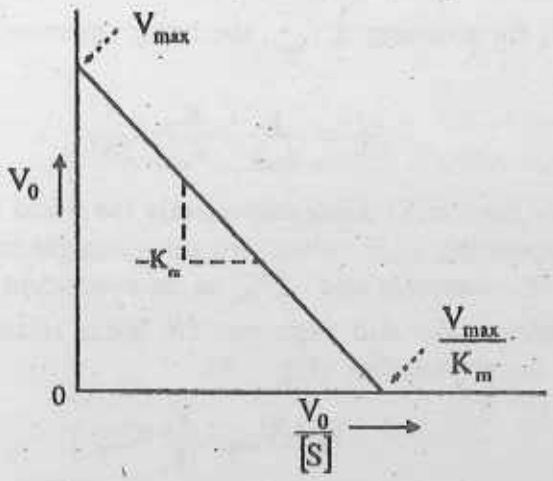


Fig 1.4. Eadie-Hofstee plot. [From D. Das, *Biochemistry*, 12th ed., Academic Publishers, 2005.]

$Y = Ze^{kX}$. This equation is the basis of exponential equations used for graphical representations of exponential distributions. It expresses the scores of an event Y as

the *exponential function* of the scores of another event X. Where the exponent X bears a positive algebraic sign, as in the equation $Y = e^X$, the exponential curve obtained by plotting Y against the corresponding value of X is an ascending one with the gradient of its slope rising progressively with the rise of X, and with its y-intercept at the value 1 of the Y scale on the y-axis. This *ascending exponential curve* thus shows the Y scores increasing exponentially with the rise in values of the positive index X. Such ascending curves result from plotting Y against X where $Y = e^X$, as in the case of systems obeying the *law of continuous growth* linking any event with the positive exponent of the base e of natural logarithm. Examples of such curves are those for rises in bacterial counts in bacterial cultures with time (Fig. 1.5).

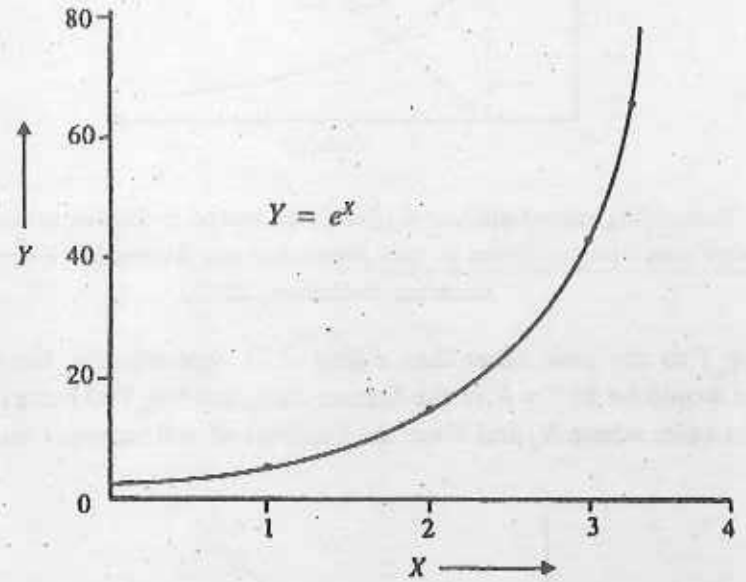


Fig 1.5. Ascending exponential curve ($Y=e^X$) of rise in bacterial count in bacterial culture.

On the contrary, if the power or exponent bears a negative algebraic sign to change the equation into $Y = Ze^{-kX}$, the exponential curve of Y against X is a descending one. Thus, for the equation $Y = e^{-X}$, a *descending exponential curve* results with a downward slope, having a progressively decreasing gradient with the rises in X. This curve depicts an exponential decrease in Y scores with the rise in values of the negative exponent or power. Such descending curves result from plotting Y against X where $Y = e^{-X}$, as in the case of systems obeying the *law of continuous decay* linking any event with the negative power of e. An example is the exponential decline in the relative activity (A/A_0) of a radioisotope with time, A_0 and A being respectively the original activity and the activity left after a time interval (Fig. 1.6).

For the equation $Y = e^X$, log Y to any base like e or 10 may be plotted against the exponent X to transform the exponential curve of Y against X into a straight line.

The latter may be ascending or descending according to the positive or negative algebraic sign of the exponent and has its y-intercept at the origin of the Y scale ($X = 0$, $Y = 0$); the slope of the line equals 1 on using $\log_e Y$, but amounts to 1/constant

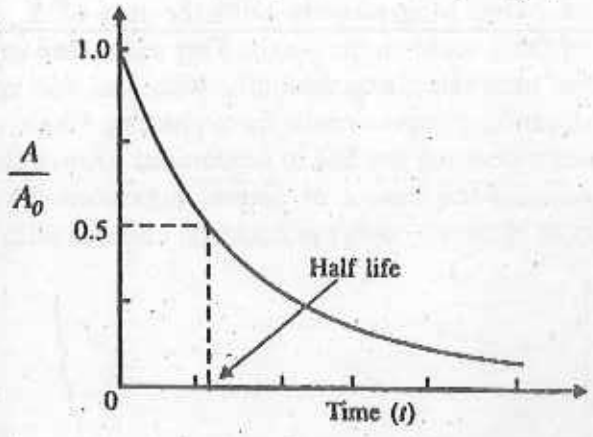


Fig 1.6. Descending exponential curve ($Y=e^{-X}$) of decline in relative activity (A/A_0) of a radioisotope with time (t). [From D. Das, *Biophysics and Biophysical Chemistry*, 6th ed., Academic Publishers, 2007.]

on using $\log_a Y$ to any base other than e (Fig. 1.7). Accordingly, the equation for the straight line would be $\ln Y = X$ in the former case, and $\log_a Y = X/$ constant in the latter case. For example, where N_0 and N are the numbers of radioisotopic nuclei respectively

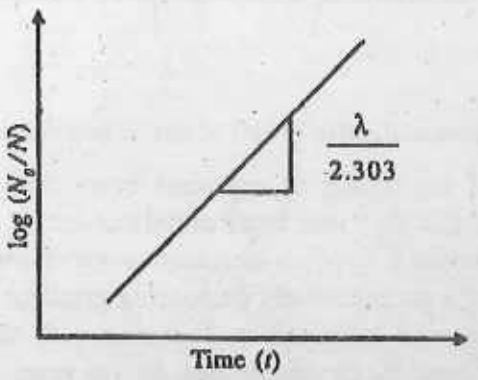


Fig 1.7. Linear relationship between log (N_0/N) and time t. [From D. Das, *Biophysics and Biophysical Chemistry*, 6th ed., Academic Publishers, 2007.]

at the initial stage and after a time interval t (i.e., $t = X$), and λ is their decaying constant (i.e., $\lambda = k$), the exponential expression $Y = e^{\lambda X}$ may be transformed into a linear equation using N_0/N as Y.

$$Y = e^{\lambda x}, \text{ or } \frac{N_0}{N} = e^{\lambda t}, \text{ or } \ln \frac{N_0}{N} = \lambda t,$$

$$\text{or, } 2.303 \log \frac{N_0}{N} = \lambda t, \text{ or } \log \frac{N_0}{N} = \frac{\lambda t}{2 \cdot 303}..$$

On scaling log (N_0/N) and t along the ordinate and abscissa scales respectively, a straight line is obtained with a slope of $\lambda/2.303$ (Fig. 1.7).

1.7 Summary

Common fractions, multiples and exponents have been defined and their properties briefly recalled in this unit. Common and natural logarithms have been defined and basic uses of common logarithm have been described.

Arithmetic, geometric and logarithmic progressions have been explained and their basic properties have been described with examples.

Equations for straight lines and exponential curves have been described briefly. Transformation of an exponential curve into its linear form has been explained using an example.

1.8 Terminal questions

1. (a) Explain what you understand by exponents and multiples.
   (b) Describe the equations and properties of exponential curves.
   (c) Describe with an example how you would transform an exponential curve into a straight line.
2. (a) Give an account of arithmetic progressions and their basic properties.
   (b) Explain what you understand by common and Naperian logarithms. How do you use common logarithms for powers and multiplications.
   (c) Write briefly about ascending and descending exponential curves, citing examples.
3. (a) Define geometric progressions and discuss their basic properties.
   (b) Write briefly about the equation and graphic properties of a straight line.
   (c) Give three examples of linear graphical plots.

1.9 Answers

1. (a) See Sub-sections 1.2.2 and 1.2.3.

(b) See Section 1.6.

(c) See the last two paragraphs of Section 1.6.

2. (a) See Sub-section 1.4.1.

(b) See Section 1.3 and Sub-section 1.3.2.

(c) See the first two paragraphs of Section 1.6.

3. (a) See Sub-Section 1.4.2.

(b) See the first three paragraphs of Section 1.5.

(c) See paragraphs (i), (ii) and (iii) of Section 1.5.

Unit 2 ❑ BIOSTATISTICS : GENERAL PRINCIPLES

Structure

2.1 Introduction

You will learn in this unit what biostatistics is, and how it may be used in the study and investigation of life sciences such as zoology and allied disciplines. You will know about different types of variables such as events, properties, organisms, objects and characters, the variations and interrelations of which are explored and assayed in experiments, using statistics for interpretation of observations. You will get to understand the difference between a population and a sample, and why and how samples are drawn from a population for scientific investigations. You will be told about different types of statistics and how they are used as estimates of population parameters.

Objectives

After studying this unit, you will be able to do the following :

- define biostatistics and summarize its basic applications,
- understand the natures of variables studied in biological experiments,
- define population and explain why samples drawn from the population are used in representing the latter in scientific studies,
- choose and use different methods of drawing representative samples from a population, and
- define parameters of population and understand different types of sample statistics used as estimates of the respective parameters.

2.2 Biostatistics and its aims

You may be aware that *statistics* is the science which applies mathematical principles, models, laws and methods in the scientific designing of experiments and investigations, drawing of samples, conducting surveys and experiments, collecting data from samples, presenting the data in meaningful forms, and interpreting the experimental data to draw inferences from the investigation undertaken.

Biostatistics is that discipline of statistics which applies mathematical laws, principles, models and methods for exploring, studying, describing, analyzing and interpreting events, natures, properties, changes and problems of living organisms.

2.2.1. Aims or basic applications of biostatistics

Biostatistics is used for the following purposes :

(a) planning an experimental design statistically, minimizing experimental errors,

(b) determining the reliability of the proposed experimental test or method for measuring a specific variable consistently,

(c) determining the validity of the proposed experimental test from its ability to measure a specific variable in exclusion of other similar variables,

(d) working out the minimum size of a sample as is necessary in making dependable inferences from the experimental observations,

(e) drawing an unbiased sample, representative of the chosen population, by random sampling methods depending on the laws of probability,

(f) presenting the experimental data in properly arranged, grouped, tabulated, graphically plotted, diagrammatically displayed and easily decipherable forms for universal communication, common perception and subsequent systematic mathematical treatment,

(g) working out mathematically whether the probability is too high or too low for the occurrence of the observed result due merely to random sampling by laws of probability, and inferring therefrom whether or not the result is significant and fit for generalization in the entire population,

(h) finding the probability of error in inferring that the observed result is significant,

(i) working out the degree and direction of association between two or more variables, and

(j) predicting the probable score of a chosen variable in an individual from the observed score(s) of one or more associated variable(s) in the same individual.

2.3 Variables

Anything that varies or undergoes changes is called a *variable*. The latter may be a property, an event, a character, or a phenomenon. A variable may undergo variations either qualitatively or in quantity. It may change from time to time in the same animal, organ, cell, organelle or site; such a change is called a *temporal variation*; for example, the change of body weight of an animal from day to day. A variable may also vary at the same instant from case to case, organ to organ, cell to cell or place to place; such a change is called a *spatial variation;* for exmple, the difference in body weights of two animals at the same moment. It is important to know about different classes of variables because the application of a statistical method or test depends on the specfic class of the variable.

2.3.1. Classification according to nature

Variables may be distinguished into four main classes, viz., nominal, ordinal, continuous measurement and discontinuous measurement variables, according to their distinctive properties or characteristics.

(a) *Nominal or qualitative variables* : Individuals or cases of a population differ qualitatively with respect to such a variable; but their differences in such a variable cannot be measured or expressed quantitatively in amounts, nor can the individuals of a sample or population be graded in ranks in ascending or descending order for the variable. You may, however, identify whether the individuals are identical or different from each other with respect to a nominal variable. Examples include sex, race, skin color, fur color, etc.

(b) *Ordinal or ranked variables* : These variables also cannot be measured quantitatively in any individual; but the individuals can be distinguished as higher or lower than each other with respect to such a variable, and can consequently be ranked in ascending or descending orders. However, by how much two cases differ from each other, cannot be worked out. Examples include ferocity, attentiveness, alertness and docility.

(c) *Continuous measurement variables* : These variables can occur and be measured quantitatively, not only in whole numbers of units, but even in infinitely small fractional units. Thus, there cannot exist any such gap in their scales where no score is located. Moreover, it can be measured quantitatively by how much the score of one case is higher or lower than that of another. Examples include length, mass, volume, temperature and time.

(d) *Discontinuous or discrete measurement variables* : These variables can also be measured quantitatively; by how much one case is higher or lower than another, can also be worked out. But these variables exist and are measurable in only whole

numbers of units; as they cannot occur in any individual in fractional units, there exist real gaps in their scales where no score can occur. Such variables include cell counts, litter sizes, finger digit numbers, respiratory rates, heart rates, etc.

The two afore-mentioned measurement variables have one of two alternative *scales*. (i) *Ratio scales* of many variables such as length, mass and volume possess *true zero points* and a ratio of two scores of such variables can be worked out. (ii) *Interval scales* of some variables such as temperature have no true zero points — these scales start from *arbitrary zero points* and a ratio of two scores of such variables cannot be worked out.

2.3.2. Classification in experiments

An experiment is the investigation into the anticipated changes of a single specific variable in a chosen sample or group of cases, on their exposure to one or more other variables. Variables involved in an experiment may be classified as follows, irrespective of their aforementioned classes.

(a) *Dependent variable* : The single specific variable, whose anticipated changes are studied or measured in an experiment, is the dependent variable in that experiment. So, in an experiment to study changes in oxygen consumption in a sample of insects on exposure to a pesticide, oxygen consumption is the dependent variable.

(b) *Independent variable* : The variable(s), one or more in number, which is/are applied to the cases of a sample to study the anticipated consequent change of the dependent variable, may be called the independent variable(s) for the relevent experiment. In the experiment cited in the preceding paragraph, the levels (viz., doses, amounts, volumes, intensities, magnitudes, etc.) of the pesticide constitute the lone independent variable. Independent variables in experiments may belong to two types. (i) *Fixed experimental treatments* are under the strict control of the investigator, are applied on the cases of the sample in well-planned manner and precise doses or amounts, and do not suffer from random fluctuations or random errors; e.g., doses of applied pesticide, amounts of injected hypoglycemic agent, etc. (ii) *Classification or random variables* are beyond the control of the investigator, may occur naturally in the surroundings, environment, naturally inherited genes, etc., and are free to suffer from random errors; e.g., sex, atmospheric O_2 tension, habitats or environmental temperature, to which the cases of a sample may get naturally exposed.

(c) *Relevant variables* : These are some variables which occur in the experimental system, though not intended by the investigator to be applied on the subjects of the sample. They may occur in the subjects themselves (*subject-relevant variables*), or in the environment and experimental situation (*situational-relevant variables*), or in the sequence of exposure of the subjects to the levels of the independent variable (*sequence-relevant variables*), and may affect the dependent variable to vitiate the

experimental result. Examples include age, sex and body weights of experimental animals (subject-relevant), pH, ionic strength and temperature of the medium or tissue preparation (situational-relevant), and the order of injection of different levels of a hypoglycemic agent (sequence-relevant). Relevant variables should be eliminated or kept constant as far as possible, to exclude their effects on the dependent variable.

2.4 Population and samples

Experiments are done with samples drawn from a specific population.

2.4.1. Population

For any experiment, the population consists of the entire aggregate of all such living organisms, inanimate objects, cases, events or phenomena as possess or exhibit at that time the specific dependent variable for the experiment or investigation. For example, for an experiment on the blood sugar of pancreatectomized rats, the population consists of all existing pancreatectomized rats on the earth; for working on O_2 consumption by a species of dragon flies, all insects of that species living anywhere at that time constitute the population.

Populations belong to two broad classes.

(i) *Infinite populations* are so immensely numerous and so widely dispersed that all the members cannot be reached or counted; e.g., the population of type I diabetic children or that of Jersey breed of cattle.

(ii) *Finite populations* consist of such small limited number of cases located within a given narrow area that all the members of such a population may be reached and counted to get its precise size; e.g., the population of pollutant-affected patients of Bhopal gas disaster, or that of a rare species of salamanders occurring in the waterbodies of a small area of Darjeeling district.

A population, whether finite or infinite, retains its identity with an identical size and unaltered properties, only so long as its members do not undergo any addition, deletion or any other change.

2.4.2. Samples

Because of the vast resources and long durations required for covering the entire population intended to be investigated, and also because of chances of unmitigated errors owing to accidental omission of some of its members from the study, an entire population is seldom subjected to any experiment or investigation. Instead, a small group of a limited number of individuals or cases, called a *sample,* is so chosen *at random* from the population by *laws of probability* as to be representative of that population with respect to the variable under investigation; such a representative

sample is then subjected to the intended investigation. The experimental observations of that investigation are then tested statistically to find their significance. If there is adequate probability of significance of the obtained result, the findings are sought to be generalized in the entire population.

Following criteria have to be gratified in a sample, if it is to be used in an investigation or experiment as the *representative* of the relevant population. (i) Individuals or cases must be included in the sample by being chosen at random from the population depending on laws of probability, so as to ensure the conformity between the population and the sample regarding the proportions of different types of cases. (ii) Variations of scores in the sample should closely conform to the variations of such scores in the population. (iii) Values of any statistic (e.g., mean and variance) of scores of different samples from the population should be so closely distributed that their arithmetic average may be identical with the corresponding population value or parameter (e.g., population mean). (iv) Scores of the variable should be distributed in the sample in conformity with their distribution in the population.

2.5 Sampling

Evidently, for generalizing the findings in a sample for the entire population, the sample should be representative of the latter. You have already learnt about the criteria to be fulfilled by a representative sample. These criteria depend largely on unbiased sampling. Some methods of sampling are briefly described below.

2.5.1. Judgement sampling

In this method, the investigator depends upon his personal judgement in considering some cases with specific properties as representing the population with respect to the intended dependent variable, and chooses arbitrarily some of such cases for inclusion in the sample. Such conscious or unintended subjective preference for some individuals or cases of particular types confines the sampling to only specific types of individuals, excluding other types of them from the chance of getting chosen for the sample. Such *judgement sampling* has a high probability of not drawing a representative sample from the population and is suitable neither for making inferences about the latter, nor for working out sampling errors of statistics computed from sample data.

2.5.2. Probability sampling

In this method, the choice of individuals from the population for inclusion in the sample is left entirely to mathematically devised methods of *random sampling* by *laws of probability*. No scope or role is left for the investigator or any other person to choose any case deliberately or arbitrarily; this minimizes the element of bias in

the sample. Instead, cases of different types have the probabilities of random choice commensurate with their respective frequencies or proportions in the population and independent of the choice of each other. Such probability sampling should yield samples, consisting of different types of cases in such relative proportions as in the population, being fairly representative of the population, and suffering from little or no bias. According to the population used and the intended purposes, probability sampling is designed in different ways.

(a) *Simple random sampling* :

If the sample has to be drawn from a *small, finite and homogeneous population*, not divided into distinct strata or sections, random sampling has to be done from the entire population taken together, choosing at random the requisite number of individuals for the sample successively out of all the individuals of the population. Thus, (i) each individual of the population enjoys an *identical probability of choice* at every step of sampling, (ii) each is chosen at random depending on the *laws of probability*, (iii) each gets chosen totally *independent* of the choice or omission of any other individual, (iv) no individual suffers from any subjective selection or rejection of any other individual, and (v) nor does any choice depend on any other quality or property. These should lead to the conformity between the proportions of different types of cases in the sample and those in the entire population.

In the unsophisticated *card drawal method* of simple random sampling, the sample size to be required for the experiment is first worked out statistically; all individuals of the population are then given successive serial numbers which are entered individually on separate cards; all those cards are mixed up in a container, and the requisite number of cards are next successively picked up blindly from that container. Individuals whose cards are so drawn are included in the sample.

Choices may be made for random sampling in two alternative ways. (i) In the *with-replacement method*, an individual once chosen is again included in the cases still left for subsequent choices and is, therefore, again considered for the subsequent choices. Thus, the probability of choice of each individual remains unaltered from choice to choice. However, it would create difficulty for the practical use of a sample if the same individual gets chosen more than once. (ii) In the *without-replacement method*, an individual once chosen is excluded from subsequent choices so that the probability of getting chosen rises progressively at successive choices; however, this rising difference in the probability of choice may be ignored as too small because of the much larger size of the population than the sample.

In a more scientific *random number method*, after giving identity numbers serially to all members of the population, individuals are chosen in the same order as the successive numbers, arranged at random in any chosen part of a random number table.

Simple random sample would not give a representative sample if a small sample is to be drawn from a stratified population with relatively low proportions of cases in one or more strata having different sizes and characters.

(b) *Stratified random sampling* :

If the population is *large and heterogeneous*, divided into distinct strata differing in properties and sizes, a proportional stratified random sampling is used in drawing a representative sample. This consists of the use of simple random sampling separately for each stratum. First, the required total size of the sample and the proportion of each stratum in the population are worked out. Next, simple random sampling is applied separately on each stratum to draw that number of individuals from it as corresponds to its proportional size in the population. All individuals of a stratum have an identical probability of getting chosen for the sample; but this probability varies from stratum to stratum according to their respective proportional sizes in the population.

For example, to draw a sample of 150 cases from a population having three strata (A, B and C) with the respective proportional sizes of 0.50, 0.40 and 0.10, simple random sampling should be undertaken separately from each stratum to draw respectively 150 × 0.50 or 75, 150 × 0.40 or 60, and 150 × 0.10 or 15 cases from A, B and C to constitute the sample.

(c) *Multistage sampling* :

A *vast population, dispersed over a wide area*, may be sampled by this method. Preferably depending on some pre-existing stages, the vast population is arranged stepwise into a number of levels, leading ultimately to the level of individuals. Simple random sampling is then applied at each of these levels. For example, to draw a sample of *Labeo rohita* fishes from waterbodies of West Bengal, three districts are chosen at random at the first stage out of all the districts; at the second stage, three waterbodies are chosen at random from all the waterbodies of the three chosen districts; finally, at the third stage, forty fishes are sampled at random from each of the three chosen waterbodies to constitute a sample of 120 fishes.

(d) *Fixed interval sampling* :

Sometimes, individuals of a population may arrive, occur or get naturally arranged in a systematic sequence; e.g., netting of successive butterflies from the air by an insect collector, angling of successive fishes in the fishing line of an angler, or arrival of successive patients at the out-patients department of a hospital. Fixed interval sampling consists of simple random sampling of cases depending on such a sequence of their random occurrence, appearance or arrival. To start with, any particular individual or case is chosen at random as the first one from the sequence of cases. Simultaneously, an interval is chosen at random as the gap between subsequent successive cases for the purpose of choices. Each subsequent case is next chosen as

it occurs in the given sequence after the preceding chosen one and separated from the latter by the chosen gap. For example, the fifth fish collected may be chosen as the first member of a sample and thereafter, every seventh fish is chosen maintaining a gap of six fishes between each pair of choices. This is continued until the requisite number of cases have been collected for the sample. However, this type of sampling may fail to yield an unbiased representative sample if the cases have been initially arranged in order of a characteristic related in any way to the variable to be investigated.

(e) *Purposive sampling* :

Random sampling is sometimes *deliberately restricted to a particular section* of the population so long as it is justifiable and logical to assume a truly representative nature of that section for the entire population, and the exclusion of its other sections is not anticipated to affect adversely the generalization of obtained results over the whole of the population.

(f) *Incidental sampling* :

In this method, random sampling is kept confined to a particular section or stratum of the population because of reasons like ready availability, easier manipulation and lower cost, instead of attempting to maintain or improve the representative nature of the sample. Such sampling should not be preferred for any investigation because it would seldom turn out a sample representative of the entire population.

2.6 Parameter and statistic

You may easily realize that as such, the experimental data consisting of one or more sets of numerical values or *scores* can hardly communicate much of precise and meaningful information or contribute much in comparing, analyzing and interpreting the observations. For these, the individual scores have to be presented, on one hand, in classified, tabulated or graphical forms while, on the other hand, some *summary values* like the mean and the standard deviation have to be worked out from those scores for further analysis and interpretation. While the presentation of data will be described in the next unit, you will be introduced in this unit to the basics of such summary values, also known as *numerical indices*, and will also get an initial idea of their roles in biostatistics.

2.6.1. Parameter

Parameters serve as measures of different characteristics of a variable in a population, and consist of numerous *summary values* like the mean and the variance, worked out from the scores of the entire population. Parameters of a population remain unchanged so long as the relevant population exists as such, but may differ from population to population. You are aware, however, that seldom do we work

with an entire population (see Sub-section 2.4.2); our investigations are generally undertaken with samples drawn from the population we want to study. Whenever we work with a sample, the summary values of the scores obtained from the latter are used as estimates of the respective parameters of the corresponding population. Two types of such estimates of parameters may be worked out : (i) a *point estimate* is a single summary value (*statistic*) of the sample, directly accepted as an estimate of the population parameter; e.g., a sample mean is the point estimate of the corresponding parametric or population mean; (ii) an *interval estimate* or *confidence interval* consists of a range of scores around a summary value (statistic) of a sample, within which the population summary value (parameter) has a given probability of occurrence; e.g., a 95% confidence interval has a probability of 0.95 for inclusion of the parametric mean.

2.6.2. Statistic

Statistics (singular : *statistic*) serve as measures of different characteristics of a variable in a sample, and consist of numerous summary values like the mean and the variance, worked out from the scores of that particular sample. As the individual scores vary from time to time in the same sample and also at the same instant from sample to sample drawn from the same population, any statistic varies *temporally* in the same sample and also *spatially* between samples of the same population. Consequently, a particular type of statistics (e.g., the sample mean $\bar{X}$) of different samples differ from the parameter (e.g., the population mean μ) by different amounts called the *sampling errors* (s_e) : $s_e = \bar{X} - \mu$. Because of their different sampling errors, the statistics (e.g., $\bar{X}$) of samples lie dispersed around the parameter (e.g., μ) of that population in the form of a *sampling distribution* with the parameter as its mean; e.g., a sampling distribution of sample means ($\bar{X}$) around the population mean (μ). It also follows that the statistics of different samples from a population serve as *different point estimates* of the same population parameter. Statistics belong to different *classes* according to their purposes.

(a) *Descriptive statistics* :

These statistics of a variable measure and describe three different characteristics of a sample in respect of that variable. (i) *Statistics of location* such as mean, median and mode describe the location of a specific point — particularly a central one — of the distribution of the scores of a variable on the scale of the latter. (ii) *Statistics of dispersion* such as variance and standard deviation are the measures of scatter of the scores of a variable around a central point like the mean of the sample. (iii) *Statistics of correlation* or correlation coefficients measure the degree and direction of the association between two or more variables in the sample. Descriptive statistics belong to a particular sample and do not go beyond the limits of the latter.

(b) *Sampling or inferential statistics* :

These statistics are not confined to the limits of a single sample and participate in comparing two or more samples. The best example of sampling statistics is a standard error of any other statistic like the mean and the correlation coefficient. Sampling statistics are used in measuring *sampling errors* and variabilities of other statistics, in computing *confidence intervals* of parameters, and in testing the *significance* of experimental findings.

(c) *Prediction statistics* :

Such statistics as regression coefficients are used in predicting the likely value of a variable on the basis of the known value(s) of one or more other variables correlated with the predicted one.

2.6.3. Mean, standard deviation and standard error

(a) *Mean and its properties* :

Mean (X) is the arithmetic average of all the scores of a sample and serves as an estimate of the parametric mean (μ) of the corresponding population. For ungrouped data not divided into class intervals, mean is computed by dividing the sum of all scores (ΣX) by the sample size (n) : $X = \Sigma X/n$. For a sample with its scores grouped into class intervals (see Sub-section 3.2.2), mean is worked out using the sample size and the sum of products of observed frequencies (f) and midpoints (X_c) of all the intervals : $X = \Sigma f X_c / n$.

Some *properties* of the mean are as follows : (i) The mean is expressed in the same unit as the scores from which it is computed. (ii) The sum of differences of individual scores from the mean amounts to zero : $\Sigma (X-\bar{X}) = 0$. (iii) Mean, median and mode of a sample are identical if its scores are distributed symmetrically in the two halves of their frequency distribution (Section 3.2). (iv) Presence of more extreme scores in one tail of the distribution than those in the other, extends that tail more than the other and displaces the mean more towards the extended tail, making the mean higher than the median and the mode if the right tail is extended (skewed) but making the mean lower than the median and the mode if the left tail is skewed (Sub-section 3.6.1). (v) If each score of the sample is added, subtracted, multiplied or divided by a constant number, the mean also gets identically treated by that number.

(b) *Variance and standard deviation* :

These are two *statistics of dispersion*. Both are *absolute measures of dispersion*, computed directly from the scores of the variable and expressed in the same units as those scores. Variance, however, is in squared units and worked out as the arithmetic average of the squared differences, i.e., $\Sigma(X-\bar{X})^2$, between the scores and their mean. For ungrouped scores of a large sample, variance (s^2) is computed as follows :

$$s^2 = \frac{\Sigma(X-\bar{X})^2}{n}, \quad \text{or} \quad s^2 = \frac{\Sigma X^2}{n} - \frac{(\Sigma X)^2}{n^2}, \quad \text{or} \quad s^2 = \frac{n\Sigma X^2 - (\Sigma X)^2}{n^2}.$$

where $(X - \bar{X})$ is the *error term* of each score, $\Sigma(X - \bar{X})^2$ is called the *sum of squares*, $(\Sigma X)^2$ is the squared sum of all the scores, and ΣX^2 is the sum of all squared scores.

For ungrouped scores of a small sample ($n < 30$), the *degree of freedom* $(n - 1)$ is used as the denominator instead of n in order to remedy the downward bias in s^2 due to omission of many of the extreme scores in the small sample during sampling. (See Sub-section 2.6.4 for degrees of freedom). For such *unbiased variance* of a small sample,

$$s^2 = \frac{\Sigma(X-\bar{X})^2}{n-1}, \quad \text{or} \quad s^2 = \frac{n\Sigma X^2 - (\Sigma X)^2}{n(n-1)}.$$

The *unbiased variance* of a sample with class intervals is computed using the frequencies (f) and the midpoints (X_c) of the intervals. (See Sub-section 5.2.1.)

$$s^2 = \frac{\Sigma[f(X_c - \bar{X})^2]}{n-1},$$

Standard deviation (*SD* or s) is the positive square root of variance and is more popular for everyday work. For ungrouped data of a large sample ($n \geq 30$),

$$s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n}}, \quad \text{or} \quad s = \sqrt{\frac{\Sigma X^2}{n} - \frac{(\Sigma X)^2}{n^2}}, \text{or} \quad s = \sqrt{\frac{n\Sigma X^2 - (\Sigma X)^2}{n^2}}.$$

Like variance, the *SD* of ungrouped data of a small sample ($n < 30$) is also computed as the *unbiased SD* using the degrees of freedom ($df = n - 1$) to reduce the downward bias as mentioned in case of variance. Unbiased *SD* and unbiased variance may also be computed and used for large samples. Thus, for unbiased *SD* (see Example 2.6.1),

$$s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}, \quad \text{or} \quad s = \sqrt{\frac{n\Sigma X^2 - (\Sigma X)^2}{n(n-1)}}.$$

The unbiased *SD* of a sample with class intervals is worked out using the frequencies (f) and midpoints (X_c) of the intervals. (See Example 3.2.1.)

$$s = \sqrt{\frac{\Sigma[f(X_c - \bar{X})^2]}{n-1}}.$$

The higher the variance and *SD*, the greater is the variability of scores in the sample. The respective parameters are expressed by the symbols σ^2 and σ. Both *SD*

and variance are affected if even a single score is changed in the sample. Multiplication or division of each score of a sample by a constant number results in an identical treatment of the *SD*.

(c) *Sampling errors and standard errors* :

You have already learnt in Sub-section 2.6.2 that due to variations of scores from sample to sample drawn from the same population, any statistic varies from sample to sample and every sample statistic may differ in turn from the population parameter by an amount called the *sampling error* (s_e) of that statistic; thus, for sample means, $s_e = \bar{X} - \mu$. Difference in sampling errors between the statistics of different samples from the same population results in the dispersion of the sample statistics around the corresponding parameter in the form of a *sampling distribution* of such statistics.

Standard error (*SE*) of a statistic measures the dispersion of that type of statistic of different samples around the corresponding population parameter; *SE* thus measures the spread of the sampling distribution of a statistic by estimating the sampling error of the latter. The *SE of means* ($s_{\bar{X}}$) is thus a measure of the average sampling error and the spread of sampling distribution of the sample means around the parametric mean of the corresponding population. Similarly, the *SE of the difference* between sample means ($s_{\bar{X}_1 - \bar{X}_2}$) is a measure of the average sampling error and the spread of the sampling distribution of differences ($\bar{X}_1 - \bar{X}_2$) of sample means about the difference ($\mu_1 - \mu_2$) between the corresponding population means. Standard errors bear the same unit as that of the relevant scores. By estimating the sampling errors of statistics and of their differences, standard errors play important roles in finding (i) the dependability and significance of a computed statistic, (ii) the variability of errors in using the statistics of different samples for estimating the population parameter, (iii) the probability of occurrence of the observed results by mere chances of laws of probability in random sampling, and (iv) the confidence interval for the population parameter (see Sub-section 2.6.1).

The *SE of mean* is worked out as follows for a sample drawn from an *infinite population* by random sampling *with or without replacement*, or from a *finite population with replacement.*

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}, \quad \text{if} \quad s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n-1}}, \text{ and}$$

$$s_{\bar{X}} = \frac{s}{\sqrt{n-1}}, \quad \text{if} \quad s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n}}.$$

But for a sample drawn from a *finite population without replacement* and with N as the population size,

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}, \quad \text{if} \quad s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}, \quad \text{and}$$

$$s_{\bar{X}} = \frac{s}{\sqrt{n-1}}\sqrt{\frac{N-n}{N-1}}, \quad \text{if} \quad s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n}}$$

The *SE of difference* ($s_{\bar{X}_1-\bar{X}_2}$) between two means is worked out from the *SE* values ($s_{\bar{X}_1}$ and $s_{\bar{X}_2}$) of those means.

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{s^2_{\bar{X}_1} + s^2_{\bar{X}_2}}.$$

Example 2.6.1.

Work out the mean, unbiased *SD* and *SE* of the mean for the following trunk length scores (X cm) of a sample of *Tilapia tilapia* fishes : 9.0, 8.4, 6.8, 6.4, 7.8, 8.6, 6.8, 7.8, 8.4, 8.0.

Solution :

Table 2.1. Table for computing mean and SD.

	X	$X - \bar{X}$	$(X - \bar{X})^2$
	9.0	+ 1.2	1.44
	8.4	+ 0.6	0.36
	6.8	− 1.0	1.00
	6.4	− 1.4	1.96
	7.8	0	0
	8.6	+ 0.8	0.64
	6.8	− 1.0	1.00
	7.8	0	0
	8.4	+ 0.6	0.36
	8.0	+ 0.2	0.04
Σ	78.0	—	6.80

$$n = 10. \quad \bar{X} = \frac{\Sigma X}{n} = \frac{78.0}{10} = 7.8 \text{ cm.}$$

$$s^2 = \frac{\Sigma(X-\bar{X})^2}{n} = \frac{6.80}{10-1} = 0.756 \text{ cm}^2.$$

$$s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}} = \sqrt{\frac{6.80}{10-1}} = 0.869 \text{ cm.}$$

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{0.869}{\sqrt{10}} = 0.275 \text{ cm.}$$

2.6.4. Degrees of freedom

During the computation of some statistics such as *SD*, variance, correlation coefficient and Student's *t*, one or more statistics computed earlier need to be used as the estimates of the corresponding parameters, and have, therefore, to be left unchanged like the latter during use in the present computation. For example, in working out the unbiased *SD* from ungrouped data, the sample mean ($\bar{X}$) is used as the estimate of population mean (μ). To keep each such precomputed statistic unchanged during the computation of the next statistic, any one of the sample scores loses its freedom for change; while any change of the remaining scores is mathematically permissible, that one score still left must change only in a fixed amount and a fixed direction to counter the effect of changes of all other scores so as to keep the precomputed statistic constant. The number of remaining scores or cases, still retaining the freedom for change, constitutes the *degrees of freedom* (*df*) for the statistic being computed. The *df* would usually amount to the sample size (n) less the number (m) of precomputed statistics being used as estimates of parameters : $df = n - m$. For example, the unbiased *SD* as computed in Example 2.6.1 has the *df* of $(n - 1)$, one score having lost its freedom for change to keep the $\bar{X}$ constant. On the contrary, when the correlation coefficient (r) is computed between two variables (X and Y) using ungrouped data (Sub-section 6.3.3.), the computed r has the *df* of $(n - 2)$, two cases having lost their freedom for change to keep respectively $\bar{X}$ and $\bar{Y}$ constant.

2.7 Summary

You have learnt in this unit about the basic applications of biostatistics. Variables have been defined and classified according to their natures and their roles in experiments. You have read about populations and have become aware of why samples are used instead of entire populations in scientific investigations. The criteria for a representative sample have also been mentioned.

Methods for sampling have been described. The importance of probability sampling has been emphasized. You have read about the applications of simple

random sampling, stratified random sampling, multistage sampling, fixed interval sampling, purposive sampling and incidental sampling.

Parameters and statistics have been defined. Statistics have been classified into descriptive, sampling and prediction statistics. Descriptive statistics have been further classified into statistics of location, of dispersion and of correlation.

Properties and computations of mean, variance, standard deviation and degrees of freedom have been briefly described. Sampling errors, sampling distributions and standard errors of statistics have been explained.

2.8 Terminal questions

1. (a) Classify variables according to their natures, giving examples and characteristics of each class.
   (b) State briefly the criteria for a representative sample.
   (c) Describe independent variables and their different types in an experiment.
2. (a) Write briefly about parameters and statistics.
   (b) Give a brief classification of statistics.
   (c) Explain what you understand by sampling errors and sampling distributions.
3. (a) Classify variables according to their roles in an experiment.
   (b) Give an account of the method of simple random sampling.
   (c) Define variance and standard deviation, and describe their computations from ungrouped scores of a sample.
4. (a) Explain when and how stratified random and multistage samplings are undertaken.
   (b) Write briefly about the standard error of mean and how it is computed, stating its different formulae.
   (c) Explain briefly what you understand by the sampling distribution of means.
5. (a) Write briefly about populations and parameters.
   (b) Define and classify relevant variables in experiments, with examples.
   (c) Explain the with-replacement and without-replacement methods of simple random sampling.
6. (a) Work out the unbiased standard deviation and the standard error of the mean of the following sample of blood sugar (mg/dl) scores : 170, 165,

123, 147, 110, 88, 97, 132, 143, 155.

(b) Classify sample statistics, giving examples.

(c) Write briefly about fixed interval sampling, confidence interval, and properties of mean.

(d) Explain degrees of freedom with examples.

2.9 Answers

1. (a) See Sub-section 2.3.1.
   (b) See 2nd paragraph of Sub-section 2.4.2.
   (c) See paragraph (b) of Sub-section 2.3.2.
2. (a) See Sub-sections 2.6.1 and 2.6.2.
   (b) See paragraphs (a), (b) and (c) of Sub-section 2.6.2.
   (c) See paragraph (c) of Sub-section 2.6.3.
3. (a) See Sub-section 2.3.2.
   (b) See paragraph (a) of Sub-section 2.5.2.
   (c) See paragraph (b) of Sub-section 2.6.3.
4. (a) See paragraphs (b) and (c) of Sub-section 2.5.2.
   (b) See paragraph (c) of Sub-section 2.6.3.
   (c) See paragraph (c) of Sub-section 2.6.3.
5. (a) See Sub-sections 2.4.1 and 2.6.1.
   (b) See paragraph (c) of Sub-section 2.3.2.
   (c) See paragraph (a) of Sub-section 2.5.2.
6. (a) See Example 2.6.1.
   (b) See Sub-section 2.6.2.
   (c) See paragraph (d) of Sub-section 2.5.2, Sub-section 2.6.1, and paragraph (a) of Sub-section 2.6.3.
   (b) See Sub-section 2.6.4.

Unit 3 ❑ PRESENTATION OF DATA AND PROBABILITY DISTRIBUTIONS

Structure

3.1 Introduction

You will learn in this unit about the presentation of experimental data in tabulated, grouped and graphically displayed forms for universal communication, easy perception, straightway comparison and systematic mathematical treatment. You will get introduced to the idea of probability distributions. You will also learn about the properties and applications of normal, binomial, Poisson and Student's t distributions.

Objectives

After studying this unit, you should be able to do the following :

- understand the use of frequency distributions in presenting the data, and work out appropriate frequency distributions of scores of different types of variables,
- draw graphical displays of frequency distributions in the forms of bar diagrams, histograms and frequency polygons,

- define probability and explain different types of probability distributions,
- define and describe the properties of normal and Student's *t* distributions for probabilities of occurrence of scores of continuous measurement variables,
- understand and describe the unit normal curve,
- describe the properties of binomial probability distributions of dichotomized variables,
- discuss the properties of Poisson probability distributions of events of the rare class of some dichotomous variables, and
- understand and estimate skewness and kurtosis of distributions.

3.2 Frequency distributions

Frequency is the number of occurrences of any individual, animal, event, score or phenomenon among the total number of such cases or in a sample. A *frequency distribution* consists of an arrangement of all the scores or cases of a relatively large sample, tabulated in different respective classes of the relevant variable. Observations of a sample, so distributed in a frequency distribution, constitute the *grouped data.* In contrast, individual scores or cases of a sample, not arranged or classified into a frequency distribution, form *ungrouped data.* Frequency distributions are broadly divided into two types, *qualitative and quantitative frequency distributions,* according to the nature of the variable whose data form the distribution.

3.2.1. Qualitative frequency distribution

Qualitative or nominal variables such as sex, phenotypes, races and blood groups cannot be measured and expressed in the form of scores; nor can the cases of a sample be arranged in ranks according to their higher or lower positions with respect to such variables. Nevertheless, such variables can be divided qualitatively into two or more distinctive classes and the frequencies of cases of a sample in the respective classes can be arranged in a table. Such a tabulated arrangement of frequencies of cases or individuals of a sample in different classes of a nominal variable is called a qualitative frequency distribution (Table 3.1). The cases of each class can be distinguished from those of other classes with regard to the variable, and there is no continuity between the classes. However, the proportion (p) of the sample size in each class can be worked out by dividing the frequency (f) of cases in each class by the sample size (n) : $p = f/n$; the computed p may be called the *relative frequency* of the relevant class.

Table 3.1. A qualitative frequency distribution of phenotypes in a sample of *Drosophila* flies.

Phenotypes	Frequencies (f)	f/n
Grey-body red-eye	108	0.551
Black-body red-eye	40	0.204
Grey-body scarlet-eye	36	0.184
Black-body scarlet-eye	12	0.061
Total	196 (n)	1.000

3.2.2. Quantitative frequency distribution

Both continuous and discontinuous measurement variables can be quantitatively measured and expressed in the form of numerical scores. The scores of such variables in a sample form a *range* which can be divided into groups called *class intervals.* Frequencies of scores belonging to different class intervals are then tabulated in the respective intervals to form a quantitative frequency distribution. The latter not only reveals characteristics of the sample in respect of the variable, but also contributes to the subsequent treatment and interpretation of the data. According to the nature of the variable concerned, quantitative frequency distributions may be either continuous or discontinuous.

1. *Continuous frequency distributions* :

Continuous measurement variables have such distributions. Here, the successive class intervals are continuous with each other, without any gap between the contiguous classes. The steps in forming such a frequency distribution are summarized below.

(a) The total *range* from the lowest to the highest score of the sample is divided into a suitable number of *class intervals* of identical lengths (i) covering 3, 5, 7, 10 or 20 scores, separating the successive intervals by their *score limits*, and entered in column 1 of the frequency table (Table 3.2). The *interval length* (i) may be obtained as the difference between either the lower or the upper limits of the consecutive classes.

(b) To avoid gaps between class intervals, *true limits* or *class boundaries* are worked out, each as the mean of two contiguous score limits of successive intervals, and entered in column 2 of the table. (See Example 3.2.1.)

(c) The *midpoint* (X_c) of each interval is worked out as follows and entered in column 3 of the table. In a frequency distribution with class intervals, all the cases in an interval lose their individual identities and are deemed to possess the score of its midpoint (X_c). For any interval,

X_c = lower score limit + $\frac{1}{2}$[(higher score limit) − (lower score limit)].

(d) Each score of the sample is entered as a tally in column 4 of the table against its class interval.

(e) The total number of tallies of each interval is entered in column 5 as the frequency (f) of that interval.

(f) Frequencies of all the intervals are finally totalled to give the sample size (n) : $\Sigma f = n$.

2. *Discontinuous frequency distributions* :

A discontinuous or discrete measurement variable has such a distribution. Here, the class intervals are separated from each other by gaps because the scores of the variable can be in whole numbers only and not in fractional numbers. A discontinuous frequency distribution is worked out in the same way as a continuous one except that the step (b) for computing the true class limits and the column 2 for entering the latter are omitted to retain the intervening gaps between consecutive class intervals. (See Example 3.2.2.)

Example 3.2.1.

(a) Work out a frequency distribution of the following wing length scores (mm) of a sample of insects : 13, 19, 20, 25, 22, 23, 22, 21, 25, 12, 20, 17, 15, 19, 18, 20, 20, 19, 17, 21.

(b) Compute the mean, *SD* and *SE* of the mean of the data.

Solution :

(a) *Continuous frequency distribution* :

Highest score = 25. Lowest score = 12. $n = 20$. $i = 3$.

Range = (highest score − lowest score) + 1 = (25 − 12) + 1 = 14 scores.

Number of class intervals = $\frac{\text{range}}{i} = \frac{14}{3} = 4.7 \simeq 5$.

True limits are computed as averages of contiguous score limits of successive class intervals and entered in Table 3.2. For example, the true upper limit of the interval 15–17 as well as the true lower limit of the next interval 18–20 is computed as : $\frac{17+18}{2} = 17.5$.

Midpoint (X_c) of each interval is computed using the higher and lower score limits of that interval, and entered in Table 3.2. For example, for the interval 18–20,

$$X_c = 18 + \frac{20 - 18}{2} = 19.$$

Each score of the data is entered as a *tally* against its class interval and the total number of tallies in the latter is entered as its *frequency* (f). The sum of frequencies of all the intervals gives the *sample size* (n).

Table 3.2. Continuous frequency distribution of wing length scores.

Class intervals				
Score limits	True limits	X_c	Tallies	f
12 – 14	11.5 – 14.5	13	||	2
15 – 17	14.5 – 17.5	16	|||	3
18 – 20	17.5 – 20.5	19	𝍸 |||	8
21 – 23	20.5 – 23.5	22	𝍸	5
24 – 26	23.5 – 26.5	25	||	2
Total				20 (n)

(b) *Computation of mean, SD and SE of mean* :

Score limits, midpoints and frequencies of the class intervals are entered in the first three columns of Table 3.3.

Table 3.3. Table for computing mean and *SD*.

Class intervals (score limits)	X_c	f	fX_c	$X_c - \bar{X}$	$(X_c - \bar{X})^2$	$f(X_c - \bar{X})^2$
12 – 14	13	2	26	– 6.3	39.69	79.38
15 – 17	16	3	48	– 3.3	10.89	32.67
18 – 20	19	8	152	– 0.3	0.09	0.72
21 – 23	22	5	110	+ 2.7	7.29	36.45
24 – 26	25	2	50	+ 5.7	32.49	64.98
Σ	–	20 (n)	386	—	—	214.20

$n = 20.$ $\bar{X} = \frac{\Sigma fX_C}{n} = \frac{386}{20} = 19.3$ mm. $s = \sqrt{\frac{\Sigma[f(X_C - \bar{X})^2]}{n-1}} = \sqrt{\frac{214.20}{20-1}} = 3.36$ mm.

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{3.36}{\sqrt{20}} = 0.751 \text{ mm}$$

Example 3.2.2.

Work out a frequency distribution of the following respiratory rates (per minute) of a sample of apes : 17, 9, 21, 14, 13, 18, 22, 16, 13, 17, 16, 15, 17, 15, 17, 19, 16, 10, 9, 11, 18, 20, 16, 16, 15, 18, 15, 12, 14, 15.

Solution :

A *discontinuous frequency distribution* is worked out for respiratory rates, using Table 3.4.

Highest score = 22. Lowest score = 9. $n = 30$. $i = 3$.

Range = (highest score – lowest score) + 1 = (22 – 9) + 1 = 14 scores.

Number of class intervals $= \frac{\text{range}}{i} = \frac{14}{3} = 4.7 = 5.$

Midpoint (X_c) of each class interval is worked out, using the higher and lower score limits of the latter, and entered in Table 3.4. For example, for the interval 15–17,

$$X_c = 15 + \frac{(17-15)}{2} = 16.$$

Each score is entered as a *tally* against its class interval and the total number of tallies of each interval is entered as the *frequency* (f) of that interval in the table. *Sample size* (n) is the sum of frequencies of all the intervals.

Note that true limits are not worked out for the intervals because there should be gaps between the successive class intervals of a discontinuous variable.

Table 3.4. Discontinuous frequency distribution of respiratory rates.

Class intervals (score limits)	X_c	Tallies	f
9 – 11	10	||||	4
12 – 14	13	𝍸	5
15 – 17	16	𝍸 𝍸 ||||	14
18 – 20	19	𝍸	5
21 – 23	22	||	2
Total	—	—	30 (n)

3.3 Graphic presentations

Frequency distributions are frequently presented in graphical forms such as frequency polygons, histograms and bar diagrams for more easily comprehensible visual displays, clearer perception of features of a sample with respect to a particular variable, and comparative study of more than one sample.

3.3.1. Frequency polygon

Frequency distribution of a *continuous measurement variable*, grouped into class intervals, is often presented as a *frequency polygon*. The latter is an *area diagram* — the total area enclosed by the arms of the polygon represents the sample size n. Before drawing the polygon, two additional class intervals — each with zero frequency and the two respectively lower and higher than the original lowest and highest intervals of the distribution — are included with their midpoints in the frequency table (Table

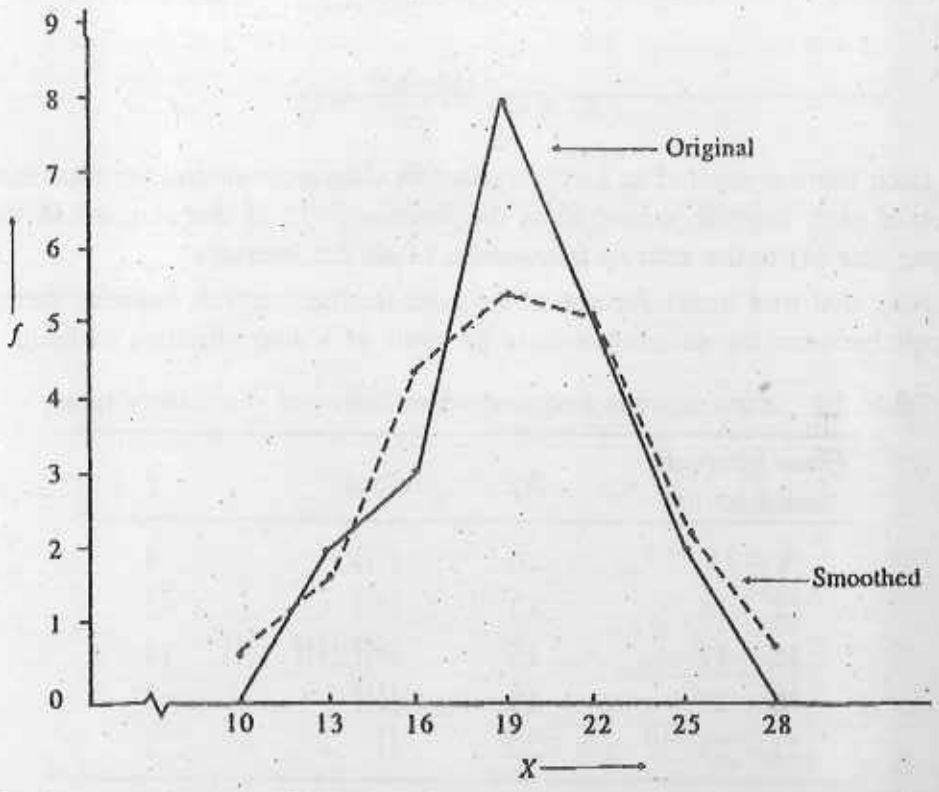


Fig 3.1. Original and smoothed frequency polygons of the data of Table 3.5.

3.5) so as to make the arms of the intended polygon reach the zero baseline of the frequency scale, on the x–axis. The scores of the relevant variable (X) are scaled on a graph paper along the latter axis or abscissa, with the *midpoint* (X_c) of each class interval marked on that axis (Fig. 3.1). Frequencies (f) of scores in class intervals are scaled along the ordinate or y-axis. Because all the cases in a class interval are deemed to possess the score of its midpoint (X_c), the frequency of each interval is plotted against its midpoint. The plotted points are joined by straight lines to draw the polygon. To compare frequency distributions of more than one sample, their respective polygons may be drawn superimposed on each other on the same x-axis.

The outline of a frequency polygon is, however, jagged — the lower the sample size and the longer the class intervals, the higher is the jaggedness of the polygon. This jeopardises the suitability of the latter in visualizing the proportionate score frequencies in different intervals. To decrease such jagged appearance of the polygon, smoothed frequencies (f_s) may be worked out for each interval as the mean of the original recorded frequencies (f) of that interval and of the intervals immediately preceding and following the latter (Table 3.5). These f_s values are than plotted against the respective midpoints (X_c) to get a *smoothed frequency polygon* (Fig. 3.1). For example, for the class interval 18 – 20 of Table 3.5,

$$f_s = \frac{8+3+5}{3} = 5.3.$$

Table 3.5. Continuous distribution of Table 3.2, modified for original and smoothed polygons of Fig. 3.1.

Class intervals		X_c	f	f_s
Score limits	True limits			
9 – 11	8.5 – 11.5	10	0	0.7
12 – 14	11.5 – 14.5	13	2	1.7
15 – 17	14.5 – 17.5	16	3	4.3
18 – 20	17.5 – 20.5	19	8	5.3
21 – 23	20.5 – 23.5	22	5	5.0
24 – 26	23.5 – 26.5	25	2	2.3
27 – 29	26.5 – 29.5	28	0	0.7
Total		—	20 (n)	20.0

3.3.2. Histogram or column diagram

This is an *area diagram* consisting of a number of contiguous rectangular bars or columns for presenting graphically the distribution of frequencies of a *continuous measurement variable* among its class intervals in a sample. To draw the histogram of a distribution with intervals of identical length (i), scores (X) of the variable are

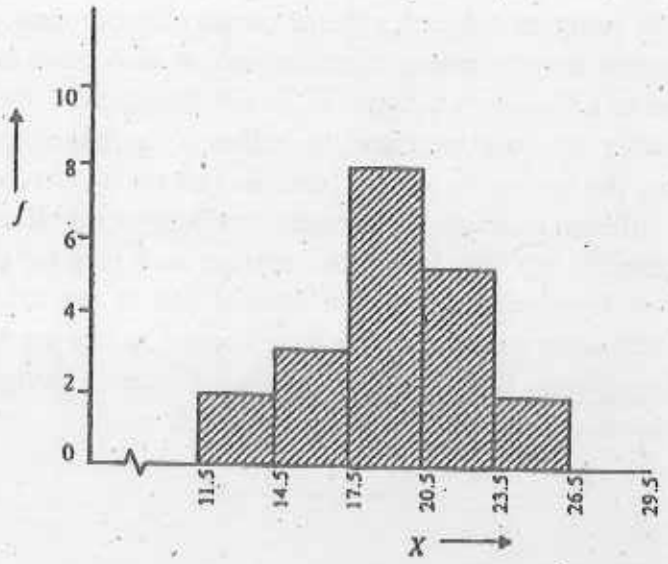


Fig 3.2. Histogram of the frequency distribution of Table 3.2.

scaled along the abscissa or x-axis on a graph paper, marking the *true limits* between successive class intervals while *frequencies* (f) are scaled along the ordinate or y-axis. At each true limit of every class interval, a perpendicular line is raised till it reaches the level of f of that interval; the column thus being formed with its base along the x-axis between the true limits of the interval, is closed at its top by a horizontal line at the level of the f of that interval. In this way, frequencies of all the intervals are depicted as successive columns with no gap between each other and with bases ranging between their respective true limits along the x-axis. As all the intervals are of identical length (i), the bases of all the columns are equal in length; consequently, the areas of the columns are directly proportional to their heights which in turn correspond to the frequencies of the respective intervals (Table 3.2 and Fig. 3.2). The sample size (n) is represented by the total area covered by all the columns. However, unlike frequency polygons, histograms cannot be superimposed on each other and need to be drawn separately side by side to compare the distributions of more than one sample.

3.3.3. Bar diagram

A bar diagram consists of one or more sets of parallel, rectangular columns or bars, used for graphical representation of frequencies of cases in different classes of a *discontinuous variable,* particularly a qualitative one. Each set of bars is meant for

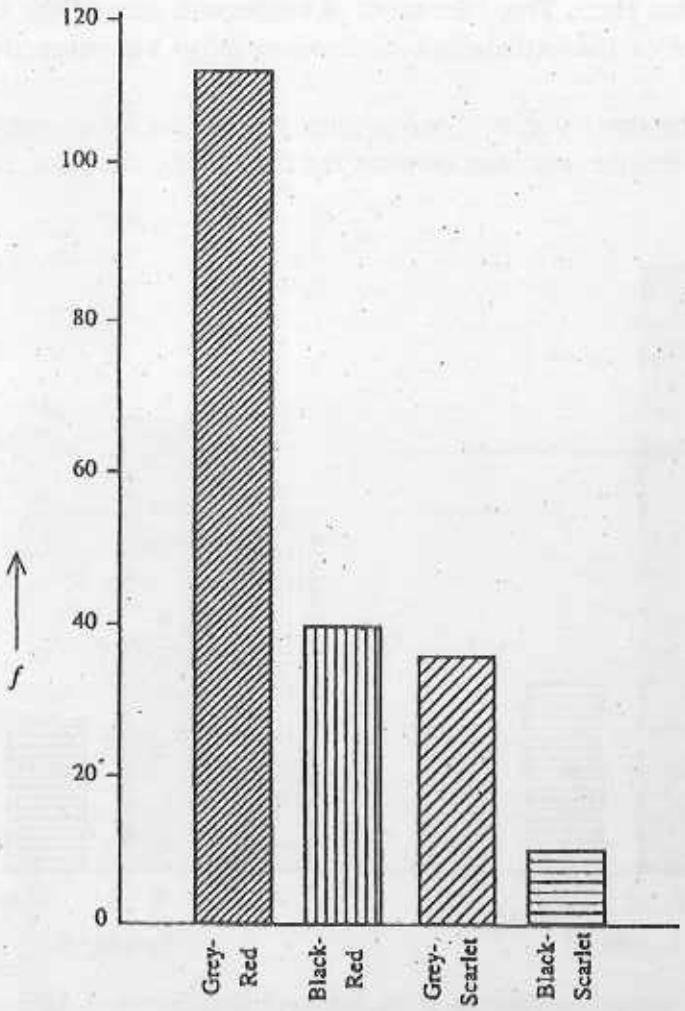


Fig 3.3. A simple bar diagram of the frequency distribution of *Drosophila* phenotypes of Table 3.1.

a particular sample, and the area of every bar or column of the set represents the frequency of cases in a particular class of the relevant variable. A *simple bar diagram* for a single sample consists of one set of bars drawn either vertically or horizontally from respectively a horizontal or vertical baseline, with the bases of the bars being identical in length. The bars are separated by small gaps to indicate that the variable

is not a continuous one, nor do its classes have any continuity with each other. Frequencies of cases are scaled along a line parallel to the bars — each bar extends from the baseline to that length beside the frequency scale which corresponds to the frequency of the corresponding class. Because of an identical width of all the bars, the area of each bar is directly proportional only to its length which in turn corresponds to the frequency of that class. Thus, the areas of successive bars of the set convey a comparative picture of the distribution of frequencies in the respective classes (Fig. 3.3).

A *multiple bar diagram* consists of two or more sets of bars for as many samples, with wider gaps between the sets than between the bars of the same set. Each set of

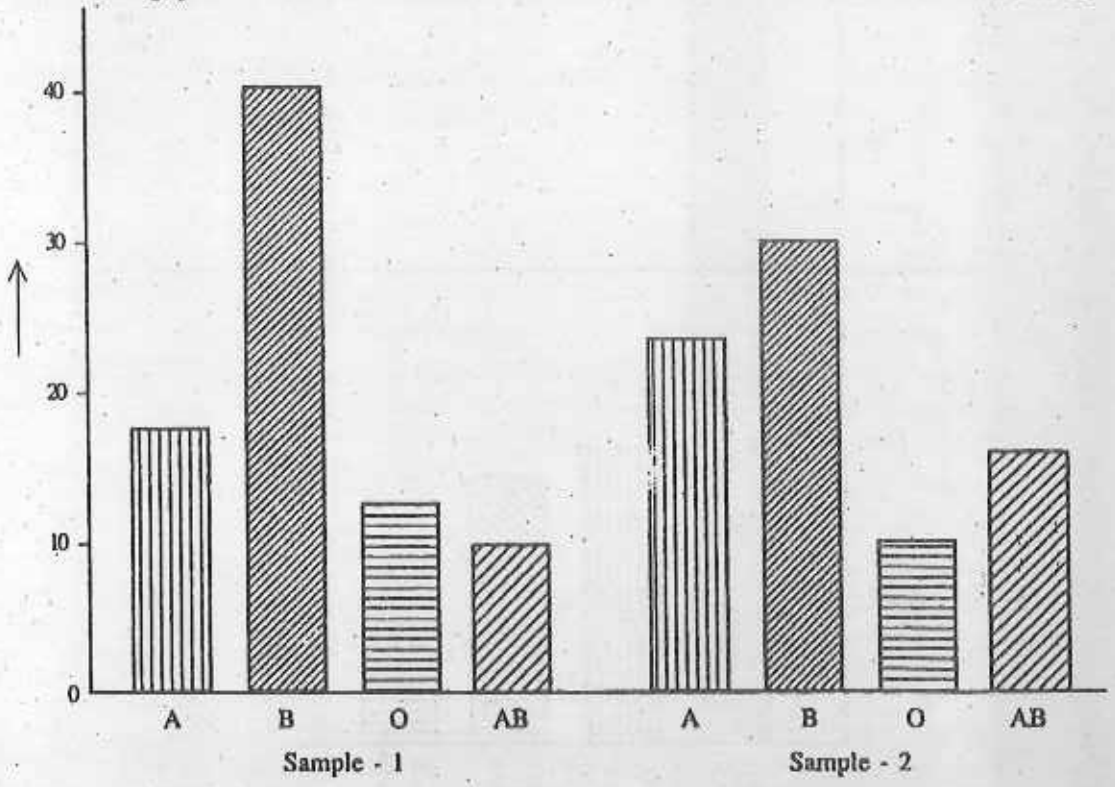


Fig 3.4. A multiple bar diagram of the frequency distributions of ABO blood groups in two samples of humans.

bars presents the frequency distribution of a particular sample. Thus, the multiple bar diagram serves to compare the frequency distributions of a nominal variable in a number of samples (Fig. 3.4).

Like histograms, bar diagrams cannot be superimposed on each other and have to be displayed separately side by side for comparison between samples.

3.4 Probability and probability distributions

In Unit 2, particularly while learning about sampling methods, you have read about probabilities and laws of probability. Probabilities are also involved at numerous stages of scientific investigations such as experimental designs, evaluation of tests, drawing of inferences and statistical predictions of scores of variables. *Probability* may be defined in a simple manner as the *relative frequency* of occurrence of any event, animal, case, score or phenomenon in an almost infinite number of such events, cases or animals. To put simply, the probability (P) of getting one mutant animal in one choice from among the total of 4587 animals of a sample, known to have 150 such mutants, would be $P = f/n = 150/4587 = 0.033$. You should keep in mind that the relative frequencies would differ progressively from the actual probability with the fall in the total number (n) of cases, and would approach the *limiting frequency*, close to the probability, only when n is near infinity. For more details, see Sections 7.5 and 7.6.

3.4.1. Probability distributions

A probability distribution is a distribution of *relative frequencies* (f/n) or *probabilities* of scores or cases in different class intervals of a given variable where the sample size (n) or the total number of events, scores or cases is vast. You have already learnt about the frequency distribution of scores in a sample. If the frequency (f) in each class interval is divided by the total size of a vast sample ($n \simeq \infty$) and that computed relative frequency (f/n) is placed in the corresponding interval, we work out a *probability distribution* of those scores or cases (Table 3.6). To present the latter graphically as a probability distribution curve or a relative frequency polygon, the scores or cases and their relative frequencies are scaled respectively along the abscissa and the ordinate; each relative frequency (f/n) is then plotted against the midpoint (X_c) of the corresponding class interval, and the plotted points are joined by lines to get the curve or polygon.

Table 3.6. Probability distribution of insect wing length scores of Table 3.2.

Class intervals	X_c	f	$P = f/n$
12 – 14	13	2	0.10
15 – 17	16	3	0.15
18 – 20	19	8	0.40
21 – 23	22	5	0.25
24 – 26	25	2	0.10
Σ	—	20 (n)	1.00

Because probabilities may be expressed even in infinitely small fractional units, the ordinate scale for probabilities in a probability distribution is always a continuous one. But the abscissa scale for events, cases or scores maybe either continuous or discontinuous according as the corresponding variable is a continuous measurement variable or a discontinuous one. *Continuous probability distributions* are probability distributions of continuous measurement variables such as blood glucose, wing length, gill weight and oxygen consumption, their x-scale (abscissa) is continuous with no gaps and has scores even in fractional units; e.g., normal probability distributions and Student's t distributions for continuous variables. *Discontinuous probability distributions* are those for discontinuous variables such as heart rates, respiratory rates and litter sizes, so that their scores cannot be in fractional units and the x-scale of their probability distributions have gaps; e.g., binomial distributions for dichotomous variables, and Poisson distributions for cases of the rare class of dichotomous variables.

Probability distribution may again be divided in another way into experimental and theoretical distributions. An *experimental probability distribution* may be worked out using the data actually collected in an experiment or a survey. You may find such a case in Table 3.6; however, you should bear in mind that unlike what is given in that table, a probability distribution should be worked out in real cases with the data of a very large number of cases to minimize errors. On the contrary, a *theoretical probability distribution* is worked out theoretically, following a specific mathematical model, applying the laws of probability and using specifically formulated theoretical equations, but needing no experimental data for its computation. Such theoretical probability distributions are widely used in the assumptions for statistical tests, interpretations of experimental data, drawing of inferences from experimental observations, working out of confidence intervals of parameters, and statistical predictions of scores or events. Such theoretical distributions include binomial distributions based on the binomial equation, Poisson distributions using the equation of S. D. Poisson, normal probability distributions based on the equation of K.F. Gauss, and Student's t distributions on the basis of the equation derived by W. S. Gossett. These four theoretical distributions are described briefly in the following sections.

3.5 Normal distributions

Normal probability distributions are *theoretical and continuous probability distributions* worked out on the basis of the *Gaussian equation* and plotted graphically as *normal curves*. To work out and plot a normal distribution from raw scores (X) of any continuous measurement variable, each X score is first transformed into a

standard score (z score) using the mean (μ or $\bar{X}$) and the standard deviation (σ or s) of those scores.

$$z = \frac{X - \mu}{\sigma}, \quad \text{or} \quad z = \frac{X - \bar{X}}{s}.$$

The z score does not bear the unit of the X score; instead, it is expressed in *SD* units (σ units) and possesses a standard reference value irrespective of mean, *SD* or unit. The probability (Y) of random occurrence of each z score and so, of the corresponding X score, is worked out with the following *Gaussian equation*, using the sample size (n), *SD* (σ), the base (e) of natural logarithms, and the constant ratio (π) between the circumference and diameter of a circle.

$$Y = \frac{n}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

If Y and z are scaled a[illegible]ng respectively an ordinate (y - axis) and an abscissa (x-axis) on a graph paper, and each computed Y is plotted against the corresponding z score, a specific unimodal (single-peaked), bilaterally symmetrical and bell-shaped probability distribution curve results; the latter is called the *normal distribution curve* or normal curve (Fig. 3.5). Variations of n, σ and interval length (i) from sample to

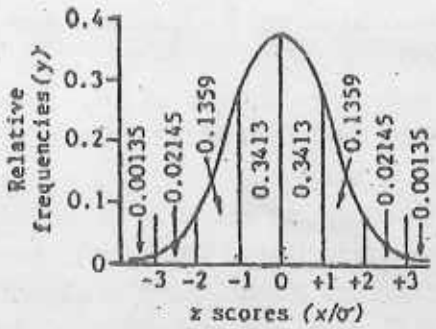


Fig 3.5. A unit normal curve with its fractional areas between the ordinates at several z scores. [From D. Das and A. Das, *Statistics in Biology and Psychology*, 4th ed., Academic Publishers, 2003]

sample yield numerous normal curves. For universal use irrespective of n, σ and i, a standard reference form of the normal curve, called the *unit normal curve*, has been derived using the following modification of the Gaussian equation and taking the values of n, σ and i as 1.00 each.

$$Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

3.5.1. Properties of normal probability distributions

Following properties of normal distribution (unit normal curve) are important for understanding and using the normal curve.

(a) Normal probability distribution is a *continuous probability distribution* with a continuous abscissa scale (x–axis) for z scores without any genuine gap between the latter. So, it can be used in case of continuous measurement variables, but is not applicable to discontinuous variables, ordinal variables and nominal variables.

(b) It is a *theoretical probability distribution* because the probabilities of this distribution can be theoretically computed using the *Gaussian equation.*

(c) According to variations in n, σ and i, there are infinite numbers of normal distributions. So, a *unit normal curve* has been worked out as a standard reference curve using a modification of the Gaussian equation and with n, σ and i amounting to 1.00 each.

(d) It is a single-peaked or *unimodal distribution* with an identical value of zero for its mean, median and mode, because the z score for μ amounts to 0.00 : $z = (\mu - \mu)\sigma = 0.00$. Also note that its mean, median and mode are coincident.

(e) It is *perfectly symmetrical bilaterally* and possesses *no asymmetry or skewness.* Thus, the coefficient of skewness is zero for the unit normal curve.

(f) Because the area enclosed by the unit normal curve represents the sample size n which is taken as 1.00 for that curve, the fractional area in either half of the curve is considered to be 0.5000.

(g) The probability of random occurrence of any z score (as also of the corresponding X score) is given by the height of the ordinate at that z score, as read from the corresponding Y value in the y-axis.

(h) Because of bilateral symmetry, the height of the ordinate at any z score in the right half of the distribution equals that at the same z score in the left half.

(i) The *highest ordinate* of the unit normal curve is located at the zero score of z, and its height (Y), viz., 0.3989, gives the probability of random occurrence of scores identical with the mean, median and mode of the sample.

(j) The fractional area of the unit normal curve between any two z scores in one of its halves is identical with that between the same two z scores in the other half.

(k) Two tails of the unit normal curve are *asymptotic*, i.e., they reach the zero level ($Y = 0.00$) of the y-axis at respectively $-\infty$ and $+\infty$ values of z scores on the x-axis.

(l) Probability (P) of random occurrence of a given z and all other z scores beyond it in *either tail* is known as the *one-tail probability* and is obtained as follows.

$P = 0.5000$ – (area of unit normal curve from its μ to the given z).

(m) Probability (P) of random occurrence of a given z and all other z scores beyond it in *both tails* is known as the *two-tail probability.*

$P = 2$ [0.5000 – (area of unit normal curve from its μ to the given z)].

(n) The normal curve is *mesokurtic*, i.e., has a medium degree of peakedness.

(o) Probability distribution of a continuous measurement variable conforms to the normal distribution, if its scores are determined by the *random effects of many other variables* with no mutual interactions.

(p) Means of samples drawn from a normally distributed population are *distributed normally* around the parametric mean of the latter, forming a sampling distribution of means.

3.6 Skewness and kurtosis

These two properties of a distribution determine its form, shape and many other characteristics.

3.6.1. Skewness

Skewness is a measure of the degree and direction of *bilateral asymmetry* of a distribution. A symmetrical distribution, e.g., normal and t distributions, has no skewness, has its two tails identically extended and equally pointed, and has its mean, median and mode coincide with the centre and peak of the distribution. But a skewed distribution is bilaterally asymmetric with one of its tails more extended

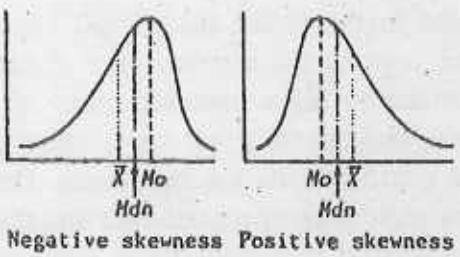


Fig 3.6. Skewed distributions. [From D.Das and A. Das, *Statistics in Biology and Psychology,* 4th ed., Academic Publishers, 2003.]

and pointed than the other tail. This results from the presence of more extreme scores in the extended or skewed tail than in the shorter and blunter tail; the scores are more concentrated in the blunter tail than in the skewed one. The skewness is called *positive* if the right or high-value tail is more drawn out than the left or low-value tail, while *negative* skewness consists of a more drawn-out and sharper left or low-value tail compared to the right tail. Poisson distributions are positively skewed

while binomial distributions are mostly either positively or negatively skewed.

Coefficient of skewness (*Sk*) is a measure of the magnitude and algebraic sign of skewness, indicating thereby the degree and direction of the skewness, respectively.

$$Sk = \frac{\bar{X} - Mo}{s}; \qquad Sk = \frac{3(\bar{X} - Mdn)}{s}.$$

3.6.2. Kurtosis

Kurtosis is a measure of *peakedness* of a ditribution. In assessing kurtosis, the normal distribution is used as the model. Being of a medium degree, its peakedness is known as *mesokurtosis*. Distributions like Student's *t* distributions possess

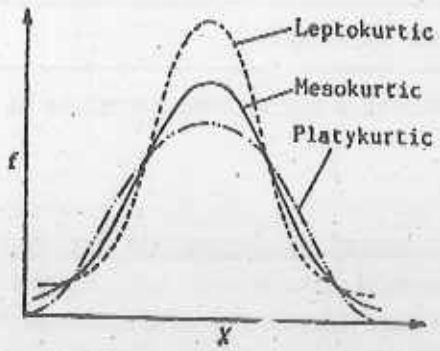


Fig 3.7. Different forms of kurtosis. [From D. Das and A. Das, *Statistics in Biology and Psychology*, 4th ed., Academic Publishers, 2003.]

comparatively sharper and higher peaks, thicker tail ends and thinner intervening regions, and are called *leptokurtic distributions*. Compared to mesokurtosis, leptokurtosis is characterized by higher concentrations of scores in a narrow zone around the peak and at the two tail-ends, but lower score concentrations in the area in between. Poisson and *t* distributions are leptokurtic. Distributions which have a broader and flatter central region, narrower tail-ends and thicker intervening regions, are called *platykurtic*. In contrast to mesokurtosis, platykurtosis is characterized by lower score concentrations in the central region and at the tail-ends, but higher score density in the area in between (Fig. 3.7). Some binomial distributions are platykurtic while some are leptokurtic.

Percentile coefficient of kurtosis (κ) is a measure of kurtosis, worked out using 10th, 25th, 75th and 90th percentiles (P_{10}, P_{25}, P_{75} and P_{90}) which are the scores below which the respective percentages of total scores occur in the sample.

$$\kappa = \frac{P_{75} - P_{25}}{2(P_{90} - P_{10})}.$$

κ amounts to 0.263 in mesokurtosis, is less than that value in leptokurtosis, and exceeds 0.263 in platykurtosis.

3.7 Student's *t* distributions

Scores of a variable are distributed normally in a large ($n \geq 30$) sample, drawn from a population having a normal distribution of scores. But if the sample drawn from such a population is a small one ($n < 30$), the frequency distribution as well as the probability distribution of its scores would take the shape of a unimodal, symmetrical but leptokurtic distribution, conforming to a theoretical probability distribution, originally formulated by W.S. Gossett and known as the *Student's t distribution* because of his pseudonym "Student". To work out the *t* distribution of the scores of a variable in a small sample, each score (X) is first transformed into *t*, basically in the same way as in case of z, and the computed *t* is then used in plotting the *t* distribution.

$$t = \frac{X - \mu}{\sigma}, \quad \text{or} \quad t = \frac{X - \bar{X}}{s}.$$

Similarly, means ($\bar{X}$) of small samples from a normally distributed population may also be transformed into *t*, using their standard errors ($s_{\bar{X}}$), and the computed *t* may then be used in plotting a sampling distribution of means conforming to Student's *t* distribution.

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}.$$

Like the use of z scores in working out a normal distribution, Student's *t* distribution is worked out by computing the probability (Y) of random occurrence of each *t* score in terms of the *Gossett equation*, using the degrees of freedom (df) of the *t* scores.

$$Y = \frac{\frac{[(df-1)/2]!}{[(df-2)/2]!} \times \frac{1}{\sqrt{\pi df}}}{[1 + \frac{t^2}{df}]^{\frac{df+1}{2}}}.$$

Scaling Y and *t* along respectively the ordinate and the abscissa scales on a graph paper, and plotting each computed Y against the corresponding *t*, a Student's *t* distribution curve may be drawn (Fig. 3.8)

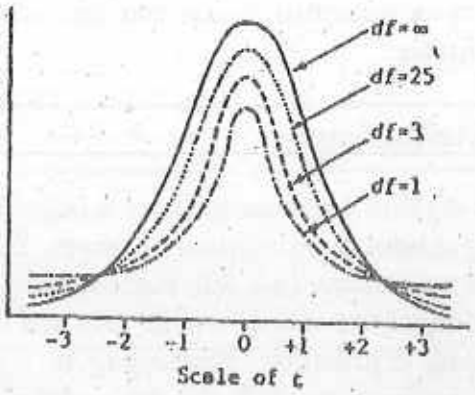


Fig 3.8. Some Student's *t* distributions. [From D. Das and A. Das, *Statistics in Biology and Psychology*, 4th ed., Academic Publishers, 2003.]

3.7.1. Properties of Student's *t* distributions

Following properties of *t* distributions are important for understanding and using them.

(a) Student's *t* distribution is a *continuous probability distribution* with a continuous abscissa scale for *t* scores without any genuine gap between them. So, it can be used in case of continuous measurement variables only, like trunk length and tracheal ventilation volume, and is not applicable to discontinuous measurement variables such as heart rate, respiratory rate, litter size and cell count, ordinal variables like ferocity, and nominal variables like sex.

(b) It is a *theoretical probability distribution* because its probabilities can be theoretically worked out using the *Gossett equation.*

(c) You may have marked in the Gossett equation itself that *t* distributions depend heavily on the *degrees of freedom of t.* So *t* distributions differ from each other according to the *df* of the relevant *t* scores and are consequently very numerous in number.

(d) It is a *unimodal distribution* with its mean, median and mode coincident with each other and amounting to zero, because the *t* score for μ amounts to zero : $t = (\mu - \mu)/\sigma = 0.00$.

(e) It is *perfectly symmetrical bilaterally* and its coefficient of skewness is zero, indicating the absence of skewness.

(f) The *t* distribution has *asymptotic tails,* extending respectively to $-\infty$ and $+\infty$ at the two ends.

(g) The distribution is *leptokurtic,* its percentile coefficient of kurtosis being lower than 0.263; the magnitude of leptokurtosis declines with the rise in *df* so that the *t* distribution with the *df* of ∞ is mesokurtic and identical with normal distribution.

(h) To interpret any computed t score, it should be referred to the specific t distribution with the same df as that of the computed t.

(i) The distribution of scores of a variable in a *small sample*, drawn from a population where such scores are *normally distributed*, would conform to the t distribution.

(j) Means ($\bar{X}$) of a variable in small samples, drawn from a population having a normal distribution of its scores, form a *sampling distribution* conforming to the t distribution around the parametric mean (μ) of that variable in the population.

3.8 Binomial distributions

Some variables are divided into two distinct classes with an intervening gap. They are called *dichotomous variables* and are frequently found among nominal variables such as sex (male-female), HIV positive-negative, pregnant-nonpregnant and Rh plus-minus. With respect to such a dichotomous variable, the individuals or cases of the population belong to either one or the other of the two classes. A random sample from such a population would often include different numbers of cases from both the classes.A *binomial probability distribution* is a distribution of probabilities of random occurrences of different combinations of cases from the two classes of a dichotomous variable, in a sample drawn from such a population. It is used to find the probability of random occurrence of either a given number of cases of one class or a given combination of cases from both classes in a sample, depending on laws of probability.

The binomial distribution is a *theoretical probability distribution* because it can be worked out theoretically using the series of terms of the *binomial equation*. You may find below the series of terms of the binomial expansion. Each of these terms gives the probability of random occurrence of a particular combination of cases from two classes of a dichotomous variable in the sample, where p and q are the proportions of the two classes in the population, the powers of p and q give the numbers of cases of the respective classes in the sample, and n is the sample size.

$$(p+q)^n = p^n + np^{n-1}q + \frac{n(n-1)}{1\times 2}p^{n-2}q^2 + \frac{n(n-1)(n-2)}{1\times 2\times 3}p^{n-3}q^3 + \ldots\ldots$$

$$\ldots\ldots + \frac{n(n-1)(n-2)(n-3)\ldots\ldots\times 2}{1\times 2\times 3\times\ldots\ldots(n-1)}pq^{n-1} + q^n .$$

For a given sample size (n) with X number of cases of the class having the proportion p in the population, that particular term of the expansion should be chosen for computing the probability P of random occurrence of X, which has the power X

of p and the power $(n - X)$ of the proportion q of the other class. P would be obtained on working out that chosen term. Where $n = 10$, $X = 4$, $p = 0.55$ and $q = 0.45$, $n - X = 10 - 4 = 6$.

$$\therefore \quad P = \frac{n(n-1)(n-2)(n-3)(n-4)(n-5)}{1\times2\times3\times4\times5\times6} \cdot p^X q^{n-X}$$

$$= \frac{10(10-1)(10-2)(10-3)(10-4)(10-5)}{1\times2\times3\times4\times5\times6} \times 0.55^4 \times 0.45^6$$

$$= 0.16 .$$

A probability level is next chosen as the *level of significance* (α) to compare with the probability P computed in the above manner. If P exceeds α ($P > \alpha$), the given 4 cases of the p class (along with 6 cases of the q class) are considered to have occurred in the sample due to random sampling; but if it is either equal to or lower than the chosen α ($P \leq \alpha$), the given numbers of cases have not occurred due to random sampling. In the above-mentioned example, if α is chosen to be 0.05, the computed P of 0.16 exceeds the chosen α ; so, the given 4 cases of the p class (along with 6 cases of the q class) have occurred in the sample by mere random sampling ($P > 0.05$).

3.8.1. Properties of binomial distribution

Following properties of binomial distribution are important for its application in biostatistics.

(a) It is a *theoretical probability distribution* because the probabilities of this distribution can be theoretically computed using the *binomial equation.*

(b) It is a *discontinuous probability distribution* because the events or cases scaled in the abscissa occur only in whole numbers (e.g., 1, 2, 3, etc.) with intervening gaps, not in fractional units, and the abscissa or x-scale is consequently discontinuous.

(c) It is a probability distribution of events of *dichotomous variables,* each divided into two classes separated by an intervening gap, *neither of these classes being rare* or having too low a proportion in the population.

(d) Each event or case of any of the two classes occurs in the sample *at random* depending on laws of probability and *independent* of all other events or cases.

(e) The distribution of events of either class, say that with proportion p, has its mean, variance and *SD* depending on the proportions (p and q) of both classes in the population.

$$\mu = np; \quad \sigma^2 = npq; \quad \sigma = \sqrt{npq}.$$

(f) Distribution of events of the class p is *symmetrical* and nonskewed when its

proportion p is 0.50 in the population; but the distribution is *positively skewed* if its proportion is less than 0.50, and has a *negative skewness* if its proportion exceeds 0.50 in the population.

(g) The distribution of either class is *platykurtic* when its proportion in the population falls within the range of 0.21–0.79, but is *leptokurtic* if its proportion is beyond that range on any side.

3.8.2. Assumptions for binomial distribution

The following assumptions should be justifiable if this distribution is to be applied in biostatistics.

(a) The relevant variable is *dichotomous*, i.e., divided into two separate classes.

(b) *Neither of the classes is rare* with too low proportion in the population.

(c) Each event or case of any of the two classes *occurs at random* and *independent* of all other events or cases.

(d) The known proportions of the two classes have remained *unchanged before and during sampling.*

(e) The distribution of events of the p class has its *mean* and *variance* amounting respectively to np and npq.

Example 3.8.1.

Work out the binomial probability of random occurrence of 7 male rats in a sample of 10 rats drawn from a rat population consisting of 48% male and 52% female animals. Interpret your result. ($\alpha = 0.01$.)

Solution :

$$p = \frac{\text{percentage of males}}{100} = \frac{48}{100} = 0.48. \quad q = 1 - p = 1 - 0.48 = 0.52.$$

$n = 10. \quad X = 7 = n - 3. \quad n - X = 10 - 7 = 3.$

$$(p+q)^n = p^n + np^{n-1}q + \frac{n(n-1)}{1\times 2}p^{n-2}q^2 + \frac{n(n-1)(n-2)}{1\times 2\times 3}p^{n-3}q^3 + \ldots\ldots + q^n.$$

$$\therefore \quad P(7) = \frac{n(n-1)(n-2)}{1\times 2\times 3}p^{n-3}q^3 = \frac{10(10-1)(10-2)}{1\times 2\times 3}\times 0.48^7 \times 0.52^3 = 0.10.$$

Alternatively, using the Bernoulli expansion :

$$P(7) = \frac{n!\, p^X q^{n-X}}{X!\,(n-X)!} = \frac{10!\, 0.48^7 \times 0.52^3}{7!\,(10-7)!} = 0.10.$$

$\alpha = 0.01. \quad \therefore \; P > 0.01.$

As P exceeds the chosen α, the given number of 7 males has occurred in the sample due merely to random sampling and has no significance ($P > 0.01$).

3.9 Poisson distributions

Some dichotomous variables are so divided into two classes that one of the latter is a rare one with its proportion very close to zero in the population, while the other class forms a large proportion of the population. *Poisson distribution* is the probability distribution of random occurrences, in the sample, of different numbers of events or cases of the *rare class* of such a dichotomous population. It does not apply to the larger and more frequent class of the variable.

Poisson distribution consists of a series of terms, each for the probability of random occurrence, in a sample, of a specific number of rare events or cases when the proportion (p) of the rare class in the population is known. If X is the number of rare cases in a sample of size n, $P(X)$ is the Poisson probability of random occurrence of X in the sample, and e is the base of natural logarithm,

$$\mu = np\ ; \quad P(X) = \frac{\mu^X}{X!\ e^{\mu}}.$$

The theoretical model worked out by Poisson equation consists of the following series of terms for probabilities of random occurrences of successive numbers of rare events.

No. of rare events (X) :	0	1	2	3	4	5	.....	n
$P(X)$:	$\frac{1}{e^{\mu}}$	$\frac{\mu}{1!e^{\mu}}$	$\frac{\mu^2}{2!e^{\mu}}$	$\frac{\mu^3}{3!e^{\mu}}$	$\frac{\mu^4}{4!e^{\mu}}$	$\frac{\mu^5}{5!e^{\mu}}$	.....	$\frac{\mu^n}{n!e^{\mu}}$.

3.9.1. Properties of Poisson distribution

Following properties should be kept in mind for applying the Poisson distribution to biostatistics.

(a) It is a *theoretical probability distribution* based on the theoretical model of the Poisson equation.

(b) It is a *discontinuous probability distribution* because the rare cases or events scaled in the abscissa or x-axis occur only in whole numbers (e.g., 1, 2, 3, etc.) with intervening gaps, not in fractional units.

(c) It is a probability distribution of random occurrences of events of the *rare class* of a *dichotomous variable,* but not of the other and more frequent class.

(d) Each rare event or case occurs in the sample *at random* obeying laws of prabability, and *independent* of all other rare events.

(e) The rare events may occur either *spatially* at different sites on a particular instant, or *temporally* in the same system at different times. Both types obey the Poisson distribution.

(f) The proportion p of events or cases of the rare class is very low — almost close to zero, while that of the other class is very high and nearly 1.00.

(g) The *mean* (μ) and the *variance* (σ^2) of a Poisson distribution are identical, dependent on the proportion p of rare cases in the population, and less than 5 in value.

$$\mu = \sigma^2 = np < 5.$$

(h) Poisson distributions possess *positive skewness* which declines with the rise in μ.

(i) Poisson distributions possess *leptokurtosis* which also declines with the rise in μ. However, neither the positive skewness nor the leptokurtosis disappears so long as the distribution conforms to the Poisson model.

3.9.2. Assumptions for Poisson distribution

The following assumptions should be justifiable if Poisson distribution has to be applied to the data.

(a) The relevant variable is *dichotomous*, i.e., divided into two classes.

(b) The class, to which Poisson distribution is proposed to be applied, should be a *rare class* with near-zero proportion in the population.

(c) Each event or case of the rare class should occur in the sample *at random* and *independent* of the occurrence or absence of other rare events or cases.

(d) *Mean* and *variance* of the distribution of events of the rare class should be identical, directly proportional to the proportion of such events in the population, and lower than 5 in value.

(e) The known proportions of the two classes in the population should remain *unchanged* before and during sampling.

Example 3.9.1.

Work out the Poisson probability of finding 9 mutant houseflies in a sample of 500 houseflies drawn from a housefly population known to have 80 mutants per 10000. Interpret your result. ($\alpha = 0.05$.)

Solution :

$n = 500.$ $\quad X = 9.$ $\quad p = \dfrac{\text{No. of mutants in given population size}}{\text{Given population size}}$

$$= \frac{80}{10000} = 0.008.$$

$$\mu = np = 500 \times 0.008 = 4.00.$$

$$P(X) = \frac{\mu^X}{X!e^{\mu}}, \quad \text{or} \quad P(9) = \frac{4^9}{9!e^4} = 0.013.$$

The computed Poisson probability is interpreted in the same way as in the case of binomial probability (Section 3.8).

$\alpha = 0.05$. $P = 0.013$. $\therefore P < 0.05$.

Since the computed Poisson probability is lower than the chosen α of 0.05, the given number of 9 mutants has not occurred in the sample due to mere random sampling, and is therefore significant ($P < 0.05$).

3.10 Summary

Scores or cases of a variable in a relatively large sample are often tabulated in a frequency distribution, where they are entered in the respective class intervals into which the entire range of scores or cases has been divided. Qualitative, continuous and discontinuous frequency distributions are worked out for the respective types of variables. Frequency polygons and histograms are drawn for graphic representation and comparison of frequency distributions of continuous measurement variables; simple and multiple bar diagrams are used for discontinuous and qualitative frequency distributions.

Relative frequencies of scores or cases in infinitely large samples may be used in working out their probability distributions. Probability distributions may be continuous or discontinuous according to the continuous or discontinuous natures of the variables. Theoretical probability distributions are worked out theoretically according to specific mathematical models; experimental probability distributions are formed from experimentally obtained data.

Normal probability distributions are theoretical and continuous probability distributions, worked out using Gaussian equation. Continuous measurement variables, whose scores are determined by the random influences of innumerable non-interacting variables, possess normal distributions of their scores in the population. For working out a normal distribution, each score of the relevant variable is transformed into a standard z score expressed in sigma (σ) units. Unit normal curve is a standard reference form of normal distribution, assuming the values of sample size, *SD* and interval length as 1.00 each. The normal curve is a unimodal, bilaterally symmetrical, bell-shaped curve with asymptotic tails, no skewness and medium degree of kurtosis. Skewness gives the degree and direction of bilateral asymmetry of a distribution while kurtosis is the degree of peakedness of the latter.

Student's t distributions are also theoretical and continuous probability distributions which are unimodal, bilaterally symmetrical nonskewed, but leptokurtic.

Distribution of any continuous measurement variable, distributed normally in a population, conforms to the t distribution in small samples from that population. Student's t distributions are worked out theoretically, using Gossett's equation and t scores obtained like z scores by transforming the scores of the relevant variable into standard forms in sigma units. The t distributions are numerous and differ from each other according to the degrees of freedom of the t values used in computing them.

Binomial distributions are theoretical and discontinuous probability distributions. They give the probabilities of random occurrences of the events of either or both classes of a dichotomous variable, provided neither class is a rare one in the population. They are computed theoretically using the binomial equation and the proportions of both the classes in the population. Most binomial distributions are either positively or negatively skewed, and either leptokurtic or platykutic.

Poisson distributions are theoretical and discontinuous probability distributions, giving the probabilities of random occurrences of events of the rare class of a dichotomous variable. They are computed theoretically using Poisson's equation and the low proportion of the rare class in the population. Poisson distributions are positively skewed and leptokurtic.

3.11 Terminal questions

1. (a) What are frequency distributions?
   (b) Describe how you would work out a continuous frequency distribution.
   (c) Write briefly about histograms and their use.
2. (a) What are probability distributions?
   (b) Write briefly about different types of probability distributions with examples.
   (c) Describe the properties of Poisson distributions, quoting the Poisson equation.
3. (a) Write how frequency polygons are drawn, using the original frequencies of scores of a frequency distribution and also their smoothed frequencies.
   (b) Describe the drawing and uses of simple and multiple bar diagrams.
   (c) Write about the skewness of distributions, mentioning the coefficients of skewness.
4. (a) Explain what is meant by a binomial probability distribution.
   (b) Describe the properties of binomial distributions.
   (c) Explain how the binomial probability can be worked out with the binomial expension.

(d) Work out the probability of random occurrence of 12 male pigeons in a sample of 20 pigeons drawn from a pigeon population known to have a male : female ratio of 45 : 55.

5. (a) Compare the assumptions for binomial and Poisson distributions.
   (b) Work out the probability of random ocurrence of 3 thalassemia cases in a sample of 150 humans from a population having 7 thalassemia patients per thousand. Interpret your result. ($\alpha = 0.05$).
   (c) Discuss the properties of Student's t distributions.
6. (a) Explain what you understand by normal distribution and unit normal curve, quoting their equations.
   (b) Discuss the proberties of normal distributions.
7. (a) Work out a frequency distribution of the following body weight scores (kg) of a sample of chimpanzees into five class intervals.
   67, 64, 73, 66, 67, 58, 77, 63, 65, 63, 73, 65, 64, 57, 78, 61, 77, 57, 72, 79, 70, 56, 68, 74, 70, 61, 59, 62, 80, 60.
   (b) Write briefly on skewness and kurtosis, giving examples.
   (c) Discuss the assumptions for binomial distributions.

3.12 Answers

1. (a) See the first paragraph of Section 3.2.
   (b) See Sub-section 3.2.2.
   (c) See Sub-section 3.3.2.
2. (a) See the first pragraph of Sub-section 3.4.1.
   (b) See Sub-section 3.4.1.
   (c) See Section 3.9.
3. (a) See Sub-section 3.3.1.
   (b) See Sub-section 3.3.3.
   (c) See Sub-section 3.6.1.
4. (a) See Section 3.8.
   (b) See Sub-section 3.8.1.
   (c) See Section 3.8.
   (d) See Example 3.8.1.
5. (a) See Sub-section 3.8.2 and 3.9.2.
   (b) See Example 3.9.1.
   (c) See Sub-section 3.7.1.
6. (a) See Section 3.5.
   (b) See Sub-section 3.5.1.
7. (a) See Example 3.2.1.
   (b) See Sub-sections 3.6.1 and 3.6.2.
   (c) See Sub-section 3.8.2.

Unit 4 ❑ STATISTICAL INFERENCE AND HYPOTHESIS TESTING

Structure

4.1 Introduction

In this Unit, you will learn how to apply stastistical principles and methods in carrying out the follow-up of experimental observations to investigate whether and how far they may be meaningful for the population. You will be told about how to explore whether or not the results obtained by working with a sample can be generalized for the entire population. Stepwise use of statistics in the analysis and interpretation of experimental data will be presented and explained to you. You will learn some widely used statistical tests for making inferences using probability distributions such as normal and Student's *t* distributions. You will also know how to work out the probabilities of going wrong in the inferences drawn from experimental results.

Objectives

After studying this unit, you should be able to do the following :

- consider the probability of the observed results arising from sampling errors due to random sampling,
- follow the stepwise use of statistics in interpretation and inference,
- define null hypothesis and alternative hypothesis, and realise their respective roles in conducting an experiment and interpreting the experimental observations,
- understand the importance and the use of levels of significance in finding the probability of the results arising from random sampling,
- distinguish between two types of errors of inference and proceed to limit their probabilities,
- transform the observed scores of large samples to standard z scores and interpret the latter using the unit normal curve,
- transform the observed scores of small samples to Student's t scores and interpret the latter using Student's t distributions with respective degrees of freedom,
- state the probability of error in inferring the observed result as significant,
- use the nonparametric chi square test to find whether or not an observed frequency distribution fits significantly with a model frequency distribution, and
- work out the chi square test of independence to explore whether or not there is any significant association between two given variables.

4.2 Significance tests

Suppose that in exploring the effect of a given independent variable on a specific dependent variable, you have applied two different *levels* (i.e., amplitudes, doses, intensities, concentrations, etc.) of the former on two groups or samples of animals from the same population, and subsequently found a difference $(\bar{X}_1 - \bar{X}_2)$ between the group / sample means of scores of the dependent variable. You have, however, learnt in paragraph (c) of Sub-section 2.6.3 that the means of two groups /samples from the same population, even when not affected at all by any independent variable, may differ from one another owing to their different *sampling errors*. So, it may be that the observed result, viz., the difference $(\bar{X}_1 - \bar{X}_2)$, does not indicate that the independent variable has actually affected and changed the scores of the dependent one; instead,

the observed difference may as well have arisen from the difference between the sampling errors (s_e) of the two means, consequent upon random sampling — there might not have been any difference between the means if the levels of the independent variable would have been applied on the entire population. Such a probability would always persist so long as samples are used instead of the population, whatever precision and caution be used in sampling to make the samples truly representative of the population. You would be right to guess that the result of any such experiment using samples would be open to two alternative inferences. One, the obtained result is not meaningful, i.e., *not significant,* has come from chances associated with random sampling, would not have occurred if the population were used instead of samples, and can thus be explained away by sampling errors; the other, the observed result is meaningful and *significant,* is not the outcome of chances of random sampling, and cannot be explained away by sampling errors. To infer which of these two alternatives may be upheld, a *significance test* has to be undertaken to find statistically whether the probability of the observed result occurring by chance is too high or too low. If this probability is *too high*, it is inferred that the obtained result of the experiment is *not significant* and not fit for generalization in the entire population; on the contrary, if the probability of its chance occurrence is *too low*, the result under consideration is *significant* or meaningful and can be generalized in the population.

For a significance test, in most cases, the experimentally obtained result (e.g., a difference between means, a correlation coefficient, etc.) is first transformed into a standard score (e.g., z, t, χ^2 and F) and the latter is referred to the corresponding probability distribution (e.g., normal, t, χ^2 or F distribution) to find the probability of its chance occurrence. To judge whether that probability is too high or too low, it is compared with a chosen probability level called the *level of significance* (α). You will learn in Sub-sections 4.6.2, 4.7.3, 4.7.4, 5.3.3, 5.5.1, 6.3.4 and 6.4.3 as also in Section 5.6, about significance tests for a number of computed statistics.

4.3 Null hypothesis

Each experiment or investigation is intended, designed and performed to substantiate or prove a proposed conjecture called the experimental hypothesis; the latter is generally known in statistics as the *alternative hypothesis* (H_a) because it is the alternative to and is contested by another hypothesis (H_o) which would be subjected to a significance test. The H_o is called the *null hypothesis* because it contradicts, contests and tries to negate or nullify the assertion of the alternative hypothesis. The *testing of hypothesis* consists basically of the working out of probability of correctness (P) of null hypothesis and finding whether that probability is too high or too low.

In general, the null hypothesis proposes that the experimental result is not significant or meaningful, that it is the outcome of using a sample drawn at random by laws of probability, and that it would not be obtained if the entire population were used instead of the sample. Its elaborate statement, however, varies according to the assertions of diverse alternative hypotheses it contests, For example, where the H_a proposes that there is a significant difference between two means (i.e., $\bar{X}_1 \neq \bar{X}_2$), the H_o contends that there is *no significant difference* between those means (i.e., $\bar{X}_1 = \bar{X}_2$); but if the H_a states that the mean of sample 1 is significantly higher than that of sample 2 (i.e., $\bar{X}_1 > \bar{X}_2$), the H_o proposes that $\bar{X}_1$ is *not significantly higher* than $\bar{X}_2$ (i.e., $\bar{X}_1 \not> \bar{X}_2$). Where the H_a proposes that there is a significant correlation between two given variables, the H_o contends that there is *no significant correlation* between the two. If the H_a states that the frequency distribution of phenotypes observed in the sample does not fit with Mendel's 9 : 3 : 3 : 1 distribution, the H_o proposes that there is a *significant goodness of fit* between the observed distribution and the Mendelian distribution.

In any significance test of the obtained experimental result, the probability (P) of the correctness of H_o is first worked out, and then compared with a chosen probability level called the *level of significance* (α). If P is found to exceed the α, the P is considered *too high;* so, the H_o is then retained, the H_a is rejected and the observed result is *not significant* ($P > \alpha$). But if P is found to be either equal to or lower than the α, the P is considered *too low;* hence, the H_o is then rejected, the H_a is accepted and the observed result is *significant* (either $P = \alpha$ or $P < \alpha$).

The H_o is bound to be tested whenever an experiment is performed with a sample; but the H_o need not be considered or tested when the entire population is subjected to an experiment.

4.4 Levels of significance

A level of significance (α) is that particular level of the probability (P) of correctness of H_o, which is compared with the P worked out in a significance test for considering the rejection or acceptance of the H_o. It is that maximum level of probability, up to which the P worked-out in the significance test is considered too low, and above which the worked-out P is considered too high. In other words, if P exceeds the chosen α, the probability of the H_o being correct is taken as too high so that the H_o cannot be rejected and the observed result is considered *not significant* ($P > \alpha$). But whenever P does not cross the chosen α, i.e., whenever the computed P is either equal to or lower than the α, it is taken to be so low as to warrant the rejection of the H_o ; the observed result is then considered *significant* ($P \leq \alpha$).

In biostatistics, 0.05, 0.02, 0.01 and 0.001 levels of α are usually used for

comparing with the worked-out P. The latter may be compared either with one particular level of α as chosen by the investigator, or successively with different levels of α in a descending order from 0.05. In the second case, the H_o may be rejected and the result considered as significant at or below that lowest level of α which either exceeds or equals the computed P ($P \leq \alpha$). For example; if P is lower than 0.05, it is significant there, but is next compared with the next lower α of 0.02 to find if it is significant even there; this is repeated with successive lower levels of α until that *lowest* α is reached which exceeds or equals P. This process is preferable as the lower the α for significance, the lower is the probability of type I error of inference (Section 4.5). For example, if P is lower than 0.01 ($P < 0.01$), then out of 100 such cases, the result may be wrongly considered significant in less than one of the cases; but if P equals 0.001 ($P = 0.001$), in only one in 1000 such cases the result may be wrongly considered significant. Thus, the lower the α at or below which the result is considered significant, the fewer are the cases wrongly declared significant and consequently the lower is the probability of type I error.

4.5 Errors of inference

Whether or not the experimental result is considered significant, there are probabilities of errors of inference because the inference derived from either the rejection or the acceptance of the H_o depends in both cases on the probabilities, P and α.

Type I error of inference results from the wrong rejection of a correct H_o, thus inferring an experimental result as significant when it is actually not significant. This error arises from the use of the *level of significance* (α) in rejecting the correct H_o and consequently has a probability identical with the α used in considering the computed P as too low ($P \leq \alpha$); so, the probability of type I error has the symbol α identical with that of the level of significance. It follows that the probability of type I error may be lowered by using a lower level of significance in comparing with the probability (P) worked out in the significance test. Thus, if P equals the α of 0.05 ($P = 0.05$), there is a probability of 0.05 for the type I error — out of 100 such cases, such results of any 5 cases would actually be not significant, having resulted from mere random sampling; but if the H_o is rejected because P equals 0.01 ($P = 0.01$), there is a much lower probability of 0.01 of the type I error. (See also the last paragraph of Section 4.4 and Sub-section 4.7.2.)

Type II error of inference is the opposite of the type I error. It is the error resulting from the wrong acceptance of a wrong H_o, thus leading to a wrong inference that the experimental result is not significant when the latter is actually significant. The probability of type II error (β) has a *reverse relation* with that of the type I error

(α) and may amount upto the value of (1 – α). In other words, the type II error (β) is given by that fractional area of H_a distribution which overlaps with the area of H_o distribution beyond the area for α in the latter (Fig. 4.1). Thus, the probability of type

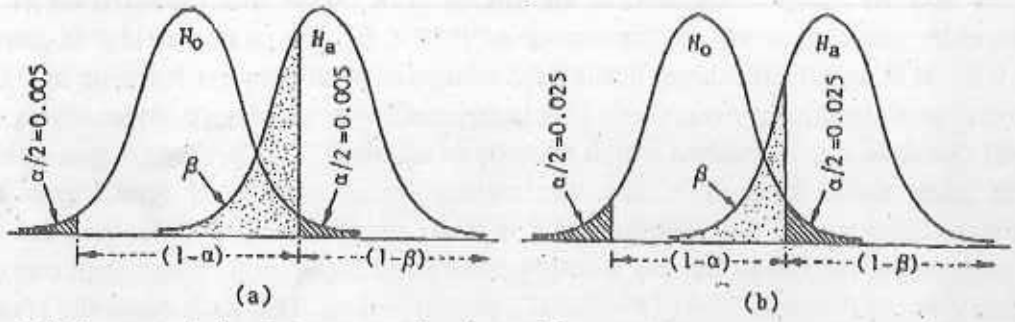


Fig 4.1. Reverse relation between type I and type II errors. [From D. Das and A. Das, *Statistics in Biology and Psychology*, 4th ed., Academic Publishers, 2003.]

II error rises with the fall in the type I error or the lowering of the level of significance. So, if α is chosen to be 0.05, the probability of type II error may be as high as 0.95, but may rise still higher upto 0.99 if α is chosen to be 0.01. β also rises with the increasing proximity and overlap of H_o and H_a distributions (Fig. 4.2)

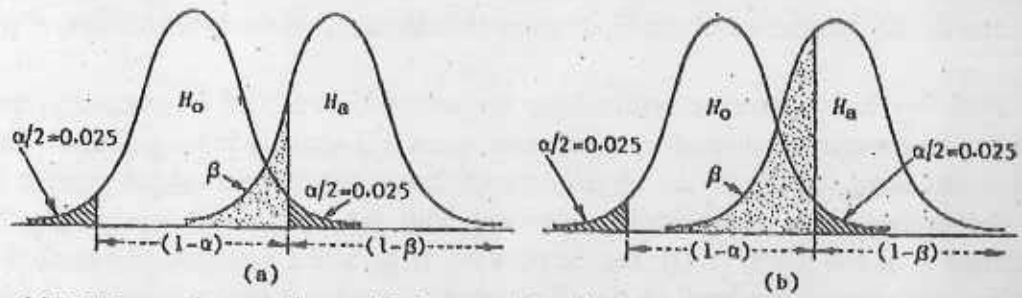


Fig 4.2. Difference in type II error (β) between (a) and (b) because of change in proximity of H_o and H_a distributions. [From D. Das and A. Das, *Statistics in Biology and Psychology*, 4th ed., Academic Publishers, 2003.]

In order to limit the chances of wrong inferences leading to positive consequences, reduction of type I errors by using lower levels of significance is preferred even at the cost of rising chances of type II errors. The latter may be reduced by choosing a relatively more powerful statistical test for interpretation.

The statement of inference of any test should be followed by the mention of either $P < \alpha$, or $P = \alpha$ or $P > \alpha$, as the case may be, and giving the numerical value of the α used, so as to indicate the probability of type I error in the inference.

4.6 Difference between means by z scores

To find the significance of difference between means of a dependent variable in large samples ($n \geq 30$) from a population with normally distributed scores, a significance test is performed by transforming that difference into a standard z score and referring the latter to the unit normal curve for interpretation.

4.6.1. Assumptions for using z scores

Following assumptions should be justifiable if z scores are to be used in finding the significance of difference between two sample means $(\bar{X}_1 - \bar{X}_2)$.

(a) The dependent variable should be a *continuous measurement variable* with its scores occurring even in infinitely small fractional units without any intervening gaps in their scale.

(b) Scores of the dependent variable should be distributed in a *normal distribution* in the population from which the samples have been drawn.

(c) Samples should be *large* in size ($n \geq 30$) so that the distribution is normal and *mesokurtic*, and their means should have a normal sampling distribution.

(d) Each score or case should occur *at random* in the sample, obeying the laws of probability and *independent* of other scores of the variable.

(e) Samples should initially come from the same or similar population(s) so that their statistics such as means and variances are initially *homogeneous*.

4.6.2. Computation and inference

The null hypothesis (H_o) proposes that the observed difference $(\bar{X}_1 - \bar{X}_2)$ between the sample/group means is not significant, has resulted from the choice of the sample by random sampling, and would amount to zero if the experiment were undertaken using the population instead of random samples. To find the probability (P) of the H_o being correct, the difference $(\bar{X}_1 - \bar{X}_2)$ is transformed into the z score. (See Example 4.6.1 also.) Where n_1 and n_2 are the respective sample/group sizes, s_1 and s_2 are the respective *SDs*, $s_{\bar{X}_1}$ and $s_{\bar{X}_2}$ are the SEs of the respective means, and $s_{\bar{X}_1 - \bar{X}_2}$ is the *SE* of the difference,

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1}; \quad \bar{X}_2 = \frac{\Sigma X_2}{n_2}; \quad s_1 = \sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2}{n_1 - 1}}; \quad s_2 = \sqrt{\frac{\Sigma(X_2 - \bar{X}_2)^2}{n_2 - 1}};$$

$$s_{\bar{X}_1} = \frac{s_1}{\sqrt{n_1}}; \quad s_{\bar{X}_2} = \frac{s_2}{\sqrt{n_2}}; \quad s_{\bar{X}_1 - \bar{X}_2} = \sqrt{s^2_{\bar{X}_1} + s^2_{\bar{X}_2}}; \quad z = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}.$$

Next, the fractional area of the unit normal curve from its mean (μ, or z score 0.00) to the computed z score is noted from the *unit normal curve table* and used in working out the probability (P) of the H_0 being correct.

$P = 2$ [0.5000 – (area of unit normal curve from μ to computed z)].

If the computed P is either equal to or lower than the chosen α (0.05 or lower), P is considered too low and the H_0 is rejected — there is thus a *significant difference* between the sample / group means ($P \leq \alpha$). But if P exceeds the α, P is taken to be too high and the H_0 is retained — there is then *no significant difference* between the means ($P > \alpha$).

Example 4.6.1.

Find whether or not there is a significant difference between the mean tracheal ventilations (ml/minute) using the following data of two groups of insects. ($\alpha = 0.05$).

X_1 : 3.0, 3.5, 3.8, 3.7, 3.1, 2.7, 3.0, 3.8, 2.5, 2.9, 3.7, 3.3, 2.6, 3.0, 2.9, 2.4, 2.5, 3.2, 2.6, 3.8.

X_2 : 2.6, 2.7, 2.5, 3.0, 3.3, 2.9, 2.8, 3.5, 3.1, 2.3, 2.5, 3.0, 3.3, 2.5, 2.8, 3.2, 2.6, 3.2, 3.0, 2.2.

z scores :	1.83	1.84	1.85	1.86	1.87	1.88	1.89	1.90
Areas of UNC :	.4664	.4671	.4678	.4686	.4693	.4699	.4706	.4713

Solution :

The data are entered in the first two columns of Table 4.1 for computations.

$$n_1 = 20. \quad n_2 = 20. \quad \bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{62.0}{20} = 3.10 \text{ ml.} \quad \bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{57.0}{20} = 2.85 \text{ ml.}$$

$$s_1 = \sqrt{\frac{\Sigma (X_1 - \bar{X}_1)^2}{n_1 - 1}} = \sqrt{\frac{4.3800}{20-1}} = 0.480 \text{ ml.} \quad s_2 = \sqrt{\frac{\Sigma (X_2 - \bar{X}_2)^2}{n^2 - 1}} = \sqrt{\frac{2.4500}{20-1}} = 0.359 \text{ ml.}$$

$$s_{\bar{X}_1} = \frac{s_1}{\sqrt{n_1}} = \frac{0.480}{\sqrt{20}} = 0.1073 \text{ ml.} \quad s_{\bar{X}_2} = \frac{s_2}{\sqrt{n_2}} = \frac{0.359}{\sqrt{20}} = 0.0803 \text{ ml.}$$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{s^2_{\bar{X}_1} + s^2_{\bar{X}_2}} = \sqrt{0.1073^2 + 0.0803^2} = 0.1340 \text{ ml.}$$

$$z = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{3.10 - 2.85}{0.1340} = 1.87.$$

Table 4.1. Computation of sums of squares.

X_1	X_2	$X_1-\bar{X}_1$	$X_2-\bar{X}_2$	$(X_1-\bar{X}_1)^2$	$(X_2-\bar{X}_2)^2$
3.0	2.6	− 0.10	− 0.25	0.0100	0.0625
3.5	2.7	+ 0.40	− 0.15	0.1600	0.0225
3.8	2.5	+ 0.70	− 0.35	0.4900	0.1225
3.7	3.0	+ 0.60	+ 0.15	0.3600	0.0225
3.1	3.3	0	+ 0.45	0	0.2025
2.7	2.9	− 0.40	+ 0.05	0.1600	0.0025
3.0	2.8	− 0.10	− 0.05	0.0100	0.0025
3.8	3.5	+ 0.70	+ 0.65	0.4900	0.4225
2.5	3.1	− 0.60	+ 0.25	0.3600	0.0625
2.9	2.3	− 0.20	− 0.55	0.0400	0.3025
3.7	2.5	+ 0.60	− 0.35	0.3600	0.1225
3.3	3.0	+ 0.20	+ 0.15	0.0400	0.0225
2.6	3.3	− 0.50	+ 0.45	0.2500	0.2025
3.0	2.5	− 0.10	− 0.35	0.0100	0.1225
2.9	2.8	− 0.20	− 0.05	0.0400	0.0025
2.4	3.2	− 0.70	+ 0.35	0.4900	0.1225
2.5	2.6	− 0.60	− 0.25	0.3600	0.0625
3.2	3.2	+ 0.10	+ 0.35	0.0100	0.1225
2.6	3.0	− 0.50	+ 0.15	0.2500	0.0225
3.8	2.2	+ 0.70	− 0.65	0.4900	0.4225
Σ 62.0	57.0	—	—	4.3800	2.4500

P = 2 [0.5000 − (area of unit normal curve from μ to computed z)]
= 2 [0.5000 − (area of unit normal curve from μ to 1.87)]
= 2 [0.5000 − 0.4693] = 0.06.

$\alpha = 0.05$.

As the computed P exceeds the chosen α of 0.05, P is too high. So, the H_o is retained and there is no significant difference between the group means ($P > 0.05$).

4.7 Student's *t* tests

To find the significance of difference between means of a dependent variable in two small samples ($n < 30$) and even in large samples ($n \geq 30$), drawn from a population with normally distributed scores, Student's t test may be undertaken. For this, the difference between the means ($\bar{X}_1 - \bar{X}_2$), is first transformed into Student's

t by dividing it with the *SE* of the difference ($s_{\bar{x}_1-\bar{x}_2}$). The computed t is then compared with the *critical* t for a chosen level of significance, for finding the probability (P) of the H_o being correct. The method of computing the t score varies according to the nature and size of samples/groups subjected to the t test (see Subsections 4.7.3 to 4.7.5).

4.7.1. Assumptions for t tests

Following assumptions should be justifiable if t tests have to be undertaken.

(a) The dependent variable should be a *continuous measurement variable* with its scores occurring even in infinitely small fractional units without any intervening gaps in their scale.

(b) Scores of the dependent variable should have a *normal distribution* in the population from which the samples or groups have been drawn.

(c) Each score should occur *at random* in the sample obeying the laws of prabability and *independent* of other scores of the variable.

(d) Samples or groups should initially come from the same or similar population(s) so that the variances of their scores are initially homogeneous (*homoscedasticity*).

You should understand from these assumptions that discontinuous measurement variables like heart rates and cell counts, ordinal variables like ferocity and docility, or nominal variables like sex and fur color cannot be subjected to t tests.

4.7.2. Critical scores

After transforming the experimental result, e.g., $\bar{X}_1 - \bar{X}_2$, into the score of a statistic such as t, F or χ^2, the probability (P) of the correctness of H_o has to be worked out by comparing that computed score of the statistic with a critical score of the latter. A *critical score* of such a statistic as t, F or χ^2 for a given level of significance (α) is that score beyond which lies a fractional area equalling the given α in the tail of its probability distribution. Because most statistics such as t, F and χ^2 possess specific degrees of freedom, and because their probability distributions differ according to their *df*, the critical score would differ with the *df*. So, the computed t, F or χ^2 must be compared with its critical score having the same *df*. It may be recalled here that the probability P of random occurrence of any computed score like t or χ^2 is given by the fractional area beyond that score in the tail of its probability distribution; so, if the computed t exceeds or equals the critical t for a given α, the area P beyond the computed t is respectively lower than or equal to the area α beyond the critical t and is consequently considered *too low* ($P \leq \alpha$); but if the computed t is lower than the critical t, the area P beyond the former exceeds the area α beyond the latter and is consequently considered *too high* ($P > \alpha$). The experimental result is thus taken respectively as significant and not significant. The same may be said about F, χ^2, etc., into which the experimental result may be transformed.

Even with the same *df* and for an identical α, the critical score would differ between a two-tail test and a one-tail test. A *one-tail test* investigates whether there is a significant higher-lower concept, such as whether or not $\bar{X}_1$ is lower than $\bar{X}_2$; for such a test, the entire area of α occurs beyond the critical score in a single tail of the probability distribution. But a *two-tail test* investigates simply whether or not the obtained result is significant, with no concern for any higher–lower concept, such as whether or not $(\bar{X}_1 - \bar{X}_2)$ is significant irrespective of its algebraic sign; for such a test, the area of the chosen α is distributed in equal halves beyond respectively the positive and negative values of the critical score in two tails of the probability distribution. Evidently, the critical scores for a given α are different for the two types of tests, that for the one-tail test being lower than that for the two-tail test. Some examples are given in Table 4.2.

Table 4.2. Some critical *t* values.

df	$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.005$	$\alpha = 0.001$
	2-tail	1-tail	2-tail	1-tail	1-tail	2-tail
10	2.228	1.812	3.169	2.764	3.169	4.587
15	2.131	1.753	2.947	2.602	2.947	4.073
20	2.086	1.725	2.845	2.528	2.845	3.850
30	2.042	1.697	2.750	2.457	2.750	3.646
60	2.000	1.671	2.660	2.390	2.660	3.460
∞	1.960	1.645	2.576	2.326	2.576	3.291

4.7.3. *t* test for small independent groups

Student's *t* is worked out by dividing the difference $(\bar{X}_1 - \bar{X}_2)$ between two sample/group means with the *SE* of that difference $(s_{\bar{X}_1 - \bar{X}_2})$. In case each group is smaller than 30 in size and the two groups are independent, consisting of separate sets of cases or animals, the *SD* to be used in computing $s_{\bar{X}_1 - \bar{X}_2}$ is worked out as the *pooled SD* $(\hat{s})$ of both the groups.

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1}; \quad \bar{X}_2 = \frac{\Sigma X_2}{n_2}; \quad \hat{s} = \sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}};$$

$$s_{\bar{X}_1 - \bar{X}_2} = \hat{s}\sqrt{\frac{n_1 + n_2}{n_1 n_2}}; \quad t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}; \quad df = n_1 + n_2 - 2.$$

The computed *t* is next compared with a 2-tail or 1-tail critical $t_{\alpha(df)}$ for a chosen significance level, according as a 2-tail or 1-tail test is being performed. If the

computed t exceeds or equals the critical t_α, P is lower than or equal to that α and the result is considered significant ($P \leq \alpha$); but if the computed t is lower than the critical t_α, P is higher than that α and the result is not significant ($P > \alpha$). Alternatively, the computed t is compared with critical $t_{\alpha(df)}$ values of successive lower levels of significance starting from 0.05 and the lowest level of α is found out, at or below which the computed t exceeds or equals the critical t_α and the result may be considered significant.

Example 4.7.1.

Find if there is a significant difference between the mean body weights (g) of the following two groups of fishes from two different habitats. ($\alpha = 0.05$.)

Group 1 (X_1) : 85, 90, 80, 75, 70, 75, 80, 90, 85, 80.

Group 2 (X_2) : 55, 65, 60, 75, 80, 70, 55, 60, 80, 70.

Critical 2-tail t scores : $t_{0.05(19)} = 2.093$; $t_{0.05(18)} = 2.101$;

$t_{0.05(20)} = 2.086$; $t_{0.05(17)} = 2.110$.

Solution :

Table 4.3. Computation of sums of squares.

X_1	X_2	$X_1 - \bar{X}_1$	$(X_1 - \bar{X}_1)^2$	$X_2 - \bar{X}_2$	$(X_2 - \bar{X}_2)^2$
85	55	+ 4.0	16.00	– 12.0	144.00
90	65	+ 9.0	81.00	–2.0	4.00
80	60	– 1.0	1.00	– 7.0	49.00
75	75	– 6.0	36.00	+ 8.0	64.00
70	80	– 11.0	121.00	+ 13.0	169.00
75	70	– 6.0	36.00	+ 3.0	9.00
80	55	– 1.0	1.00	– 12.0	144.00
90	60	+ 9.0	81.00	– 7.0	49.00
85	80	+ 4.0	16.00	+ 13.0	169.00
80	70	– 1.0	1.00	+ 3.0	9.00
Σ 810	670	—	390.00	—	810.00

$$n_1 = 10. \quad n_2 = 10. \quad \bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{810}{10} = 81.0\text{g}. \quad \bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{670}{10} = 67.0\text{g}.$$

$$\hat{s} = \sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{390.00 + 810.00}{10 + 10 - 2}} = 8.16\text{g}.$$

$$s_{\bar{X}_1 - \bar{X}_2} = \hat{s}\sqrt{\frac{n_1 + n_2}{n_1 n_2}} = 8.16\sqrt{\frac{10+10}{10 \times 10}} = 3.649\text{g}.$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{81.0 - 67.0}{3.649} = 3.837. \quad df = n_1 + n_2 - 2 = 10 + 10 - 2 = 18.$$

Critical $t_{0.05(18)} = 2.101$.

As the computed t is higher than the critical $t_{0.05}$, P is lower than 0.05 and is considered too low. So, the H_0 is rejected and it is inferred that there is a significant difference between the group means. ($P < 0.05$.)

4.7.4. t test for large independent groups

With rise in sample/group size and consequently in the *df*, t distributions come closer to the unit normal curve, and coincide with the latter when *df* amounts to ∞. So, instead of using z scores, t test can also be undertaken to find the significance of difference ($\bar{X}_1 - \bar{X}_2$) between means of two large independent groups ($n_1 \geq 30$, $n_2 \geq 30$) consisting of two separate sets of cases or animals. For such large groups, ($\bar{X}_1 - \bar{X}_2$) is transformed into t by division with $s_{\bar{X}_1 - \bar{X}_2}$ worked out with separate unbiased *SD*s of the individual groups.

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1}; \quad \bar{X}_2 = \frac{\Sigma X_2}{n_2}; \quad s_1 = \sqrt{\frac{\Sigma (X_1 - \bar{X}_1)^2}{n_1 - 1}}; \quad s_2 = \sqrt{\frac{\Sigma (X_2 - \bar{X}_2)^2}{n_2 - 1}};$$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}; \quad t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}; \quad df = n_1 + n_2 - 2.$$

The computed t is compared with critical $t_{\alpha(df)}$. If the computed t exceeds or equals the critical t_α, P is either lower than or equal to that α and is considered too low. So, the H_0 is rejected and the difference between group means is significant ($P \leq \alpha$). But if the computed t falls short of the critical t_α, P exceeds the α ; so, the H_0 is retained and there is no significant difference between the means. ($P > \alpha$).

Example 4.7.2.

Using the data of Example 4.6.1., find if the mean ($\bar{X}_1$) of group 1 is significantly higher than that ($\bar{X}_2$) of group 2 ($\alpha = 0.05$).

(Please note that although this method is to be used for groups equalling or exceeding 30 in size, groups smaller than 30 have been used in this example for the sake of brevity.)

1-tail critical t scores : $t_{0.05(39)} = 1.689$; $t_{0.05(38)} = 1.689$;

$t_{0.05(20)} = 1.725$; $t_{0.05(37)} = 1.689$.

Solution :

Please refer to Table 4.1 for computation of means and sums of squares.

$$n_1 = 20,\ n_2 = 20.\quad \bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{62.0}{20} = 3.10.\quad \bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{57.0}{20} = 2.85.$$

$$s_1 = \sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2}{n_1 - 1}} = \sqrt{\frac{4.3800}{20-1}} = 0.480.\quad s_2 = \sqrt{\frac{\Sigma(X_2 - \bar{X}_2)^2}{n_2 - 1}} = \sqrt{\frac{2.4500}{20-1}} = 0.359$$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{0.480^2}{20} + \frac{0.359^2}{20}} = 0.1340.$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{3.10 - 2.85}{0.1340} = 1.866.\quad df = n_1 + n_2 - 2 = 20 + 20 - 2 = 38.$$

Critical 1-tail $t_{0.05(38)} = 1.689$.

As the computed t exceeds the 1-tail critical $t_{0.05}$, P is lower than 0.05 and considered too low. So, the H_o is rejected and $\bar{X}_1$ is significantly higher than $\bar{X}_2$ ($P < 0.05$).

4.7.5. *t* test for small single-group experiments

If the same set of animals is used once as group 1 and again as group 2, it is a single-group experiment with each individual animal having *a pair of scores correlated with one another.* But if the single group is small in size ($n < 30$), the correlation coefficient (product-moment r) cannot be worked out between the paired scores of the two groups for use in computing t. So, a difference method is used for the t test in such single-group experiments to avoid the necessity of computing the r.

In the *difference method*, the difference D is worked out between the paired scores (X_1 and X_2) of each of the n number of animals in the group, and used in computing the mean difference ($\bar{D}$) and the standard deviation (s_D) of the D scores. $\bar{D}$ is transformed into t using the standard error ($s_{\bar{D}}$) of $\bar{D}$.

$$D = X_1 - X_2;\quad \bar{D} = \frac{\Sigma D}{n};\quad s_D = \sqrt{\frac{\Sigma(D - \bar{D})^2}{n-1}};\quad s_{\bar{D}} = \frac{s_D}{\sqrt{n}};$$

$$t = \frac{\bar{D}}{s_{\bar{D}}};\quad \text{df} = n - 1.$$

The computed t is compared with critical $t_{\alpha(df)}$ values. Only if the computed t

either exceeds or equals the critical t_α, P is respectively lower than or equal to that α and is considered too low — the difference between the means is then considered significant ($P \leq \alpha$).

Example 4.7.3.

Find whether or not there is a significant change in the mean O_2 consumption (ml/100g bodyweight) in the following sample of pigeons after their exposure to antimycin A.

Animal	:	1	2	3	4	5	6	7	8	9	10
Before (X_1)	:	160	176	182	179	184	167	180	158	162	176
After (X_2)	:	130	148	150	145	158	142	143	130	131	147

Critical 2-tail t scores : $t_{0.05(8)}$=2.306; $t_{0.05(9)}$=2.262; $t_{0.01(8)}$=3.355; $t_{0.01(9)}$=3.250; $t_{0.001(8)}$=5.041; $t_{0.001(9)}$=4.781.

Solution :

The data are entered in the first three columns of Table 4.4 and used for working out the mean and the sum of squares of the differences (D).

$$n = 10.\ \bar{D} = \frac{\Sigma D}{n} = \frac{300}{10} = 30.0.\quad s_D = \sqrt{\frac{\Sigma(D-\bar{D})^2}{n-1}} = \sqrt{\frac{120}{10-1}} = 3.65.$$

$$s_{\bar{D}} = \frac{s_D}{\sqrt{n}} = \frac{3.65}{\sqrt{10}} = 1.154.\quad t = \frac{\bar{D}}{s_{\bar{D}}} = \frac{30.0}{1.154} = 25.997.\ df = n - 1 = 10 - 1 = 9.$$

Table 4.4. Computation of sum of squares of differences.

Animal	X_1	X_2	$D = X_1 - X_2$	$D - \bar{D}$	$(D - \bar{D})^2$
1	160	130	30	0	0
2	176	148	28	− 2	4
3	182	150	32	+ 2	4
4	179	145	34	+ 4	16
5	184	158	26	− 4	16
6	167	142	25	− 5	25
7	180	143	37	+ 7	49
8	158	130	28	− 2	4
9	162	131	31	+ 1	1
10	176	147	29	− 1	1
Σ	—	—	300	—	120

Critical 2-tail $t_{0.001(9)}$ = 4.781.

As the computed t exceeds the critical t for 0.001 level of significance, P is lower than 0.001. The P being too low, the H_o is rejected, and it is inferred that the mean O_2 consumption has changed significantly on exposure to antimycin A ($P < 0.001$).

4.8 Chi square tests

Chi square tests are *nonparametric tests*, because neither is any statistic computed earlier used as an estimate of any parameter in working out the chi square, nor is the latter referred to normal or t distributions for interpretation. In contrast to other tests described in this unit, the chi square test is an *analysis of frequencies* and not a test for significance of difference between means. Two principal types of chi square tests consist of *tests for goodness of fit* and *tests for association.*

Chi squares (χ^2) have continuous positively skewed distributions, differing in shape with degrees of freedom — for chi squares with *df* of 1 or 2, the distribition is L-shaped ; but with the rise in *df*, the distribution is unimodal with progressively declining positive skewness. The critical χ^2 is that score, beyond which the fractional area in only the asymptotic right tail of the distribution amounts to the level of significance (α). Chi squares are additive.

4.8.1. Assumptions for chi square tests

For using chi square tests, it should be justifiable to assume that each case occurs in the sample by chance due to *random sampling* depending on *laws of probability,* and *independent* of other cases. Other than this, assumptions for continuous nature of variable, normal distribution in the population, and ability of scores being ranked are not required. So, these tests can be applied on continuous, discontinuous, ordinal or nominal varibles, large or small samples, and normal or non-normal distributions.

4.8.2. Chi square test for goodness of fit

This test explores *by analysis of frequencies* whether or not a distribution of experimentally observed frequencies (f_o) conforms significantly to a distribution of expected frequencies (f_e) based on a proposed theoretical model such as normal, binomial, Poisson, equal-probability and Mendelian distributions.

In this test, for each class interval of the f_o distribution observed in the sample, the corresponding f_e is first computed using the proportional distribution of cases as proposed by a theoretical model. The differences (f_o–f_e) between the observed and the expected frequencies are then used in working out the χ^2 score.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}.$$

The *degrees of freedom* of the computed chi sequare are determined by the difference between the total number (k) of classes of the f_o distribution and the number of classes which have lost their freedom for change to keep the parameters of the proposed distribution and the sample size constant. Thus, the *df* of the computed χ^2 would be (k–1) to keep the sample size constant if the Mendelian phenotype distribution is the proposed one, (k–2) to keep the sample size and the μ constant when either binomial or Poisson distribution is used as the proposed distribution, and (k–3) to keep the sample size, μ and σ constant if the proposed distribution consists of a best-fitting normal distribution.

Yates' correction has to be done to decrease the upward bias in the χ^2 to be computed, if the latter has the *df* of 1 and any one or more classes are lower than 5 in the value of f_e. The correction consists of bringing the (f_o–f_e) difference of each class closer to zero by 0.5, by deducting 0.5 from each positive (f_o–f_e) and adding 0.5 to each negative (f_o–f_e). In effect, for each class,

$$\text{corrected } (f_o - f_e) = |f_o - f_e| - 0.5.$$

In such cases, it is the corrected (f_o–f_e) of each class, which is used in computing the chi square. Thus, with corrected (f_o–f_e) values,

$$\chi^2 = \sum \frac{(|f_o - f_e| - 0.5)^2}{f_e}.$$

The H_o proposes that the computed χ^2 is not significant, having resulted from random sampling only. To find the probability P of correctness of H_o, the computed χ^2 is compared with the critical χ^2 with the computed *df* and for a chosen α. The computed χ^2 is significant only if it either exceeds or equals the critical χ^2 ($P \leq \alpha$), but is not significant if it is lower than the critical χ^2 ($P > \alpha$). Finally, only if the χ^2 is *not significant*, there is a *significant goodness of fit* between the observed and the proposed distributions; a significant χ^2 indicates *no* significant goodness of fit.

Example 4.8.1.

Find whether or not the following observed phenotype distribution in a sample of *Drosophila* has a significant goodness of fit with Mendelian 9 : 3 : 3 : 1 distribution. (α = 0.01.)

Phenotypes :	Grey body-red eye (AB)	Grey body-scarlet eye (Ab)	Black body-red eye (aB)	Black body-scarlet eye (ab)
No. of flies :	104	33	35	20

Critical χ^2 values: $\chi^2_{0.01(4)} = 13.28$; $\chi^2_{0.01(3)} = 11.34$;

$$\chi^2_{0.01(2)} = 9.21;\ \chi^2_{0.01(1)} = 6.64.$$

Solution :

Total of phenotype proportions in Mendelian distribution : 9 + 3 + 3 + 1 = 16.
Sample size (n) = 104 + 33 + 35 + 20 = 192.
Expected proportion (p_e) of each phenotype

$$= \frac{\text{Mendelian proportion of the phenotype}}{\text{Total of all phenotype proportions of Mendel}}.$$

Thus, (i) AB phenotype : $p_e = \frac{9}{16} = 0.5625$;

(ii) Ab phenotype : $p_e = \frac{3}{16} = 0.1875$;

(iii) aB phenotype : $p_e = \frac{3}{16} = 0.1875$;

(iv) ab phenotype : $p_e = \frac{1}{16} = 0.0625$.

Expected frequency (f_e) of each phenotype : $f_e = np_e$.

Thus, (i) AB phenotype : $f_e = np_e = 192 \times 0.5625 = 108$.

(ii) Ab phenotype : $f_e = np_e = 192 \times 0.1875 = 36$.

(iii) aB phenotype : $f_e = np_e = 192 \times 0.1875 = 36$.

(iv) ab phenotype : $f_e = np_e = 192 \times 0.0625 = 12$.

Table 4.5. Computation of chi square for goodness of fit.

Phenotypes	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
Grey-red	104	108	− 4	16	0.1481
Grey-scarlet	33	36	− 3	9	0.2500
Black-red	35	36	− 1	1	0.0278
Black-scarlet	20	12	+ 8	64	5.3333
Σ	192	192	—	—	5.7592

$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 5.76$ (Table 4.5). $df = k - 1 = 4 - 1 = 3$.

Critical $\chi^2_{0.01(3)} = 11.34$.

As the computed χ^2 is lower than the critical χ^2 for the given 0.01 level of significance, $P > 0.01$. So, the H_o is retained and the computed χ^2 is not significant. Hence, there is a significant goodness of fit between the observed phenotype distribution and the Mendelian distribution.

4.8.3. Chi square test of association

Also known as the *chi square test of independence*, this test explores whether or not there is any significant association between two variables.

The data are arranged in this test in a *contingency table* presenting the relation between the two variables. The classes of one variable are arranged along the columns of the table and those of the other along its rows (Table 4.6). Each cell of the table would, therefore, house the observed frequency (f_o) of cases belonging to a specific combination of two particular classes, one for each variable. Where r and c represent the numbers of respectively the rows and columns of the table for the f_o scores, the df of the χ^2 would amount to : $(r - 1)(c - 1)$. The total (f_r) of the f_o scores of each column and the total (f_c) of those of every row are entered in the respective row and column for marginal totals.

On the basis of the H_o of no association between the variables, the f_e scores of that many randomly chosen cells, as given by the df, are calculated using the f_r and the f_c of the respective cells and the sample size (n) : $f_e = f_r f_c / n$. The f_e scores of the remaining cells are worked out by subtracting the already obtained f_e scores from either the f_r or the f_c of the respective cells. The f_e score of each cell is entered in the table against the f_o of that cell. The difference $(f_o - f_e)$ between each f_o and the corresponding f_e is then used in computing the chi square.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}; \qquad df = (r-1)(c-1).$$

Yates' correction is done in the way described in Sub-section 4.8.2 to lower the upward bias of χ^2, if its df amounts to 1 and any f_e is found to be less than 5 in amount. In such a case,

$$\text{corrected } \chi^2 = \sum \frac{(|f_o - f_e| - 0.5)^2}{f_e}.$$

The H_o contends that the computed χ^2 is not significant and has resulted merely from random sampling. To find the probability P of the H_o being correct, the computed χ^2 is compared with the critical $\chi^2_{\alpha(df)}$ for a chosen significance level. Only if the

computed χ^2 either exceeds or equals the critical χ^2, the H_o is rejected and it is inferred that there is a significant association between the variables ($P \leq \alpha$); if the critical χ^2 exceeds the computed one, there is no significant association between the variables and they are significantly independent of one another ($P > \alpha$).

Example 4.8.2.

Out of 40 diabetic monkeys, 20 were hypertensive while out of 60 nondiabetic ones, 15 were hypertensive. Find whether or not there is a significant association between diabetes and hypertension. ($\alpha = 0.02$.)

Critical χ^2 scores: $\chi^2_{0.02(3)} = 9.84$; $\chi^2_{0.02(2)} = 7.82$;

$\chi^2_{0.02(1)} = 5.41$; $\chi^2_{0.02(4)} = 11.67$.

Solution :

Table 4.6. Contingency table for test of association.

	Nonhypertensive		Hypertensive		
	f_o	f_e	f_o	f_e	f_r
Diabetic	20	26	20	14	40
Nondiabetic	45	39	15	21	60
f_c	65	65	35	35	100 (n)

$r = 2$. $c = 2$. $df = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$.

Because the df amounts to 1, only one cell — that for diabetic-hypertensive — is chosen for direct computation of f_e. (See Table 4.6.)

$$f_e = \frac{f_r f_c}{n} = \frac{40 \times 35}{100} = 14.$$

The f_e scores of other cells are worked out by subtracting the already obtained f_e score(s) from either the f_r or the f_c of the respective cells. For example, for the cell for diabetic-nonhypertensive class,

f_e = (f_r of that cell) − (f_e of diabetic-hypertensive cell) = 40 − 14 = 26.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(20-26)^2}{26} + \frac{(20-14)^2}{14} + \frac{(45-39)^2}{39} + \frac{(15-21)^2}{21} = 6.59.$$

Critical $\chi^2_{0.02(1)} = 5.41$

Because the computed χ^2 exceeds the critical $\chi^2_{0.02}$, P is considered too low. So, the H_o is rejected. It is, therefore, inferred that there is a significant association between diabetes and hypertension ($P < 0.02$).

4.9 Summary

Significance tests are undertaken to find whether the experimental result is genuine and can be generalized in the population, or whether the result can be explained by sampling errors. For any investigation using a sample, the null hypothesis (H_o) contests the alternative hypothesis (H_a) being explored, and proposes that the observed result is not significant and comes from the use of a random sample. Only if a significance test indicates that the probability (P) of correctness of H_o does not surpass a chosen probability level called the level of significance (α), the obtained result is considered significant and accepted for the entire population. Any such inference is, however, subject to a type I error of inference for the wrong rejection of a correct H_o. The type I error can be reduced by using a lower significance level (α) in interpreting the result.

A difference between the means of two samples, exposed to two different levels of an independent variable, may be transformed into the standard z score which is referred to the unit normal curve to find the P of correctness of H_o for drawing an inference about the significance of the difference. A difference between two means may also be transformed into Student's t which is compared with critical t scores for different significance levels (α) to draw the inference regarding the significance of the difference. Chi square test is a nonparametric analysis of frequencies often used in finding whether or not an observed frequency distribution has a significant goodness of fit with an expected frequency distribution worked out on a proposed theoretical model. Chi square test may also be used for exploring whether or not two variables have a significant association.

4.10 Terminal questions

1.(a) Write the assumptions for t tests.

(b) Describe mentioning mathematical formulae, how you would work out the t test for a small single-group experiment.

(c) Work out Student's t test to find whether or not the exposure to an arsenic compound has produced significant changes in the body weights (g) of the following sample of crabs. ($\alpha = 0.05$.)

Animal :	1	2	3	4	5	6	7	8	9	10
Before (X_1) :	5.0	4.5	5.5	4.8	6.0	5.8	5.6	5.2	4.7	4.7
After (X_2) :	4.0	3.7	4.5	3.7	4.8	4.3	4.1	3.9	3.8	3.6

Critical t scores : $t_{0.05(19)} = 2.093$; $t_{0.05(9)} = 2.262$;

$t_{0.05(18)} = 2.101$; $t_{0.05(8)} = 2.306$.

2.(a) Describe the assumptions for using z scores.

(b) Discuss how the z score is worked out and interpreted in finding a significance

of difference between two sample means, mentioning all computational formulae.

(c) Work out the z score using the following body weight data of two samples of fishes to find whether or not the sample means differ significantly. ($\alpha = 0.05$.)

Sample 1 : $\bar{X}_1 = 78.0$ g ; $s_1 = 8.50$ g ; $n_1 = 36$.

Sample 2 : $\bar{X}_2 = 74.4$ g ; $s_2 = 6.75$ g ; $n_2 = 49$.

z scores	2.08	2.09	2.10	2.11	2.12
Areas of normal curve	0.4812	0.4817	0.4821	0.4826	0.4830

3.(a) Write briefly about the following : (a) Null hypothesis, (b) Levels of significance. (c) Errors of inference. (d) Yates' correction. (e) Critical scores.

4.(a) Describe, mentioning all computational formulae, how to work out Student's *t* test for small independent groups.

(b) Discuss where *t* test can be used and where not, describing the assumptions for *t* tests.

(c) Work out the *t* test to find whether or not the mean gill weights (mg) differ in the following two small samples of crabs from two habitats. ($\alpha = 0.05$.)

Group 1 (X_1) : 100, 80, 75, 65, 75, 110, 95, 78, 92, 80.

Group 2 (X_2) : 77, 63, 70, 68, 62, 55, 75, 70, 60, 60.

Critical *t* scores : $t_{0.05(19)} = 2.093$; $t_{0.05(18)} = 2.101$;
$t_{0.05(9)} = 2.262$; $t_{0.05(8)} = 2.306$.

5.(a) Discuss the assumptions for chi square tests. Mention important properties of chi squares.

(b) Describe, mentioning the relevant formulae, how to compute chi square for goodness of fit and interpret it.

(c) Work out chi square test to find whether or not the following observed frequency (f_o) distribution of serum iron concentrations (µg/dl) in a given sample of chimpanzees has a significant goodness of fit with the expected frequency (f_e) distribution computed from the normal probability distribution. ($\alpha = 0.05$.)

Class intervals	f_o	f_e
101-110	5	4
111-120	10	10
121-130	12	15
131-140	18	20
141-150	15	16
151-160	14	10
161-170	6	5

Critical χ^2 scores : $\chi^2_{0.05(1)} = 3.84$; $\chi^2_{0.05(2)} = 5.99$;

$\chi^2_{0.05(3)} = 7.82$; $\chi^2_{0.05(4)} = 9.49$; $\chi^2_{0.05(5)} = 11.07$.

6. (a) Describe how you would work out the chi square test of association and interpret the computed chi square to draw an inference.

   (b) Discuss the difference between the computations of degrees of freedom for chi square test of association and different chi square tests for goodness of fit.

   (c) Out of 55 hypercholesterolemic cases, 25 suffer from hypertension while out of 45 cases with normal serum cholesterol, 15 are hypertensive. Use chi square test to find whether or not there is a significant association between hypercholesterolemia and hypertension. ($\alpha = 0.05$.)

   Critical χ^2 scores : $\chi^2_{0.05(1)} = 3.84$; $\chi^2_{0.05(2)} = 5.99$;

   $\chi^2_{0.05(3)} = 7.82$; $\chi^2_{0.05(4)} = 9.49$; $\chi^2_{0.05(5)} = 11.07$.

4.11 Answers

1. (a) See Sub-section 4.7.1.
   (b) See Sub-section 4.7.5.
   (c) See Example 4.7.3.
2. (a) See Sub-section 4.6.1.
   (b) See Sub-section 4.6.2.
   (c) See Example 4.6.1.
3. (a) See Section 4.3.
   (b) See Section 4.4.
   (c) See Section 4.5.
   (d) See Sub-sections 4.8.2 and 4.8.3.
   (e) See Sub-section 4.7.2.
4. (a) See Sub-section 4.7.3.
   (b) See Sub-section 4.7.1.
   (c) See Example 4.7.1.
5. (a) See Sub-section 4.8.1 and second paragraph of Section 4.8.
   (b) See Sub-section 4.8.2.
   (c) See Example 4.8.1.
6. (a) See Sub-section 4.8.3.
   (b) See second paragraph of Sub-section 4.8.3 and third paragraph of Sub-section 4.8.2.
   (c) See Example 4.8.2.

Unit 5 ❑ ANALYSIS OF VARIANCE

Structure

5.1 Introduction

For any experiment done with a sample to substantiate a specific idea called the alternative hypothesis (H_a), the null hypothesis (H_o) contends that the experimental results have been obtained because of chances associated with random sampling, would not occur on using the population instead of a sample, and are consequently not significant. You have learnt earlier that Student's t tests may be performed to find whether or not there are significant differences between sample means. In the present unit, you will learn about analysis of variance (anova), Kruskal-Wallis test and multiple comparison tests, which are more powerful alternatives to the t tests.

Objectives

After studying this unit, you should be able to do the following :

- understand the importance of variances in biostatistics,
- know why the anova is preferable to t test,
- distinguish between different ways and models of anova,
- learn about the assumptions for anova and judge where to apply the latter,
- work out one-way anova by computing the F ratio and finding its significance,
- follow up a significant F ratio by working out omega square in model I anova and added variance component in model II anova,
- follow up a significant F ratio in anova for more than two groups, with Scheffe's F test for multiple comparison between chosen group means,

- work out Kruskal-Wallis nonparametric anova where assumptions for parametric anova are not justifiable, and
- follow up a significant Kruskal-Wallis test with more than two groups, by Mann-Whitney multiple comparison U test between chosen group means.

5.2 Analysis of variance

You are aware that statistics like means and variances of the scores of a dependent variable differ from group to group from the same population, owing to their varying *sampling errors*, even when not affected by any independent variable. In addition, if different groups are exposed to different *levels* (magnitudes, intensities, concentrations, etc.) of an independent variable affecting the dependent one, statistics like the variance of the latter would differ further *between the groups* than what may result from their sampling errors. You have learnt earlier that Student's t test for differences between group means may indicate whether or not such differences can be explained away by the sampling errors of the means, and may thus help in inferring whether or not the group means differ significantly. In the present unit, a more powerful and widely applicable test, viz., the analysis of variance (anova), will be offered for testing simultaneously the differences in variances of the dependent variable scores between two or more groups, exposed to different levels of independent variables, for inferring if the dependent variable has changed significantly on exposure to the latter.

5.2.1. Variances in anova

You may recall that variance or mean square (s^2 or MS) is the squared standard deviation of scores of a sample from its mean, and serves as a better *absolute measure of dispersion* of the scores around the sample mean (see paragraph (b) of Sub-section 2.6.3). It may be defined as the mean of squared deviations of scores from the sample mean, i.e., as the mean of the *sum of squares* (SS), although it is now usually worked out using the df instead of the sample size to minimize its downward bias, particularly in small samples : $s^2 = SS/df = \sum (X-\bar{X})^2/(n-1)$.

Variances of different groups/samples, drawn initially from the same population, possess homogeneity (*homoscedasticity*); they differ only to limited extents because of random variations of scores in different groups and owing to their differing *sampling errors* only, and function as estimates of the same population variance (σ^2). But if the scores of different groups are affected by their subsequent exposure to different levels of an independent variable, their variances become heterogeneous (*heteroscedastic*), differing from each other by far more than their differing sampling errors and behaving now like estimates of parametric variances of different populations.

In the experimental data from two or more groups, the *total variance* (s_t^2) is the

variance of dependent variable scores of all the groups from their grand mean ($\bar{X}$). The total variance may be computed from the *total sum of squares* (SS_t) which is the sum of the squared deviations, viz., $\Sigma (X_i - \bar{X})^2$, of the scores (X_i) of all groups from the grand mean ($\bar{X}$). Where N is the total number of scores or cases of all the groups,

$$SS_t = \Sigma(X_i - \bar{X})^2; \quad s_t^2 = \frac{SS_t}{N-1}; \quad df_t = N-1.$$

In an experiment, all the cases of a group are exposed to the same specific level of independent variable while different groups are exposed to different specific levels of the latter. Scores of each group differ from the mean ($\bar{X}_i$) of that group by their respective *error terms*, viz., $(X_i - \bar{X}_i)$, because of random variation of each score from the group mean, but not due to the effect of independent variable as all the cases in a group are exposed to the same level of the latter. The *within-groups variance* (s_w^2) is the variance of dependent variable scores of all groups from their respective group means and is worked out from the *within-groups sum of squares* (SS_w), viz., $\Sigma (X_i - \bar{X}_i)^2$. Where k is the number of groups,

$$SS_w = \Sigma (X_i - \bar{X}_i)^2 ; \quad s_w^2 = \frac{SS_\omega}{N-k} ; \quad df_w = N - k.$$

The *between-groups variance* (s_b^2) is the variance of scores belonging to different groups and is computed from the *between-groups sum of squares* (SS_b). The latter is worked out as the sum of the products of the respective group sizes (n_i) with the squared differences between the respective group means ($\bar{X}_i$) and the grand mean ($\bar{X}$). The between-groups variance may result from two factors, viz., *random variations* of scores from their respective group means — given also by the within-groups variance — and an *added variance* between the groups due to the exposure of different groups to different levels of independent variable. The added variance factor would be absent from the between-groups variance if the independent variable has not produced any change in the dependent variable scores.

$$SS_b = \Sigma n_i(\bar{X}_i - \bar{X})^2; \quad s_b^2 = \frac{SS_b}{k-1}; \quad df_b = k-1.$$

Anova resolves the total variance (s_t^2) into the between-groups variance (s_b^2) and the within-groups variance (s_w^2) by partitioning the total sum of squares (SS_t) into the between-groups sum of squares (SS_b) and the within-groups sum of squares (SS_w). The *variance ratio* (F ratio) is then computed using s_b^2 and s_w^2 : . The computed F is then compared with the critical F value (F_α) for a chosen significance level (α),

for finding whether or not there is any significant difference between the groups (see Sub-section 5.3.3).

$$F = \frac{s_b^2}{s_w^2};\ df \text{ of } F \text{ ratio} : df_b, df_w = k-1, N-k.$$

Inference is made as follows :

(i) if computed $F \geq F_\alpha$, $P \leq \alpha$ and H_o is rejected;

(ii) if computed $F < F_\alpha$, $P > \alpha$ and H_o is retained.

5.2.2. Reasons for preferring anova

So long as the assumptions for anova are justifiable, the latter should be preferred to Student's *t* test and its alternatives for finding the significance of difference between group means, because of the following reasons. (i) Anova can be applied *at a time to any number of groups,* two or more, to search simultaneously for any significant difference between any or all pairs of groups. (ii) A significant *F* ratio in anova may be followed up by working out either the *strength of association* between dependent and independent variables in model I anova, or the *added variance component* in the between-groups variance in model II anova. (iii) Because of its strong assumptions, anova is much more *powerful* than *t* test and its other alternatives. (iv) Its strong assumptions require meticulous pre-planning and scientific designing of the experiment, which eliminate many experimental errors.

5.2.3. Classes of anova

According to the number of independent variables in the experiment, different *classes or ways of anova* are to be used with respective experimental designs, statistical treatments and interpretations. (i) A *one-way anova* is applied if the effect of a single independent variable is being investigated. For example, a one-way anova would be worked out to explore the significance of difference in the tracheal ventilation of two groups of grasshoppers treated with two respective levels of a pesticide, the latter being the *only independent variable* in this experiment. (ii) A *two-way anova* has to be used if the groups are exposed to combinations of two different independent variables. For example, a two-way anova is undertaken for the difference in serum Ca^{2+} levels of three groups of hermit crabs administered three respective levels of combinations of *two independent variables,* viz., doses of parathormone and growth hormone. (iii) A *three-way of anova* is used where *three independent variables* have been applied on the groups under study.

5.2.4. Models of anova

Three alternative models of anova are chosen from, according to the natures of the independent variables. (i) *Model I* or *fixed model anova* is applied to explore the

significance of change in a dependent variable, when exposed to the chosen levels of one or more *fixed experimental treatments.* For example, a *one-way model I anova* is undertaken for the change in tracheal ventilation of locusts on administration of three chosen levels of an insecticide whose application is strictly under control of the investigator. In this model, a significant *F* ratio indicates the existence of a *cause-effect relationship* between the dependent variable and the fixed experimental treatment, and enables the working out of the strength of association between the two variables. (ii) *Model II* or *random model anova* is used when the dependent variable is deemed to be affected by one or more *random variables* beyond the control of the investigator. For example, a *one-way model II anova* has to be used to explore the difference in blood hemoglobin concentration between the two sexes, because sex is a variable beyond the control of the investigator in this experiment. Cause-effect relation and strength of association cannot be explored here between the two types of variables, because the independent one suffers from random errors beyond the control of the investigator. Instead, an *added variance component* may be worked out as a follow-up of a significant *F* ratio in a model II anova. Similarly, a *two-way model II anova* is used for changes in blood thyroxine level on exposure of the groups to changes of both atmospheric temperature and humidity. (iii) *Model III* or *mixed model anova* is applied when exploring the change of a dependent variable in the groups exposed to chosen levels of a fixed experimental treatment and different levels of a random variable at the same time. An example is the *two-way model III anova* for change in arterial O_2 tension on exposure to the fixed chosen levels of O_2 tension in inhaled gas mixture and the prevailing blood hemoglobin concentrations of the subjects — the Po_2 of inhaled gas mixture and the blood hemoglobin are respectively fixed treatment and random variables.

5.2.5. Assumptions for anova

Numerous rigorous assumptions for anova make it more powerful as a test and also serve to decrease experimental errors by requiring a well-planned experimental design.

(a) Each score of the dependent variable — stated otherwise, each *error term* of the latter — should *occur at random* in any group obeying the laws of probability, to ensure that the groups are representative of the population.

(b) Each error term should occur in the group, *independent* of the occurrence of any other error term in it.

(c) Error terms of the dependent variable should have a *normal distribution* in the population.

(d) Initially, all the groups to be used in the experiment should have homogeneous

variances or *homoscedasticity* to ensure that the group variances are different estimates of the same population variance.

(e) To avoid the *order effect* of application of the independent variable, different levels of the latter should be applied to different cases or individuals of the sample/ group in *randomly varying orders* instead of an identical sequence.

(f) The total variation of any score of the dependent variable should arise from the *additivity* of its numerous variations owing to various factors including random relevant variables and the independent variable. This justifies the partitioning of the total sum of squares into its components like between-groups and within-groups sums of squares during anova.

5.3 One-way anova

One-way anova is worked out for testing the significance of change in a dependent variable on exposure of the groups of cases to different respective levels of a *single independent variable*. Computation and interpretation of a one-way anova depend on the *number of groups*, i.e., on the number of levels of the independent variable used in the experiment, and also on the *model* of the anova, i.e., on the nature of the independent variable. You should recall that the number of groups used in an experiment corresponds to the *number of levels* of the independent variable, because each specific level of the latter would be applied to all the cases of a particular group only. Moreover, a one-way anova may belong to either model I if the independent variable is a fixed experimental treatment under the investigator's control, or model II if the independent variable is beyond the control of the investigator and suffers from random changes.

A one-way anova is worked out in the following steps.

(a) First, *sums of squares* are partitioned and used in working out the respective variances and their ratio (*F ratio*); the latter is compared with a critical F value for significance.

(b) If the computed F turns out to be significant, the next step is to work out either the *strength of association* in case of model I anova, or the *added variance component* in case of model II anova.

(c) If the number of groups exceeds two in the relevant experiment, the next step is a *multiple comparison test* to find whether or not there is any significant difference between the means of every chosen pair of groups.

5.3.1. Partitioning of sums of squares

To partition the total sum of squares (SS_t) into between-groups and within-

groups sums of squares (SS_b and SS_w) in a one-way anova with k number of groups, having $n_1, n_2, n_3, \ldots\ldots n_k$ as the respective group sizes, $X_1, X_2, \ldots\ldots X_k$ as the respective group scores, $\bar{X}_1, \bar{X}_2, \ldots\ldots \bar{X}_k$, as the respective group means, and $\bar{X}$ as the grand mean,

$$N = n_1 + n_2 + \ldots\ldots + n_k. \quad \bar{X} = \frac{\Sigma X_1 + \Sigma X_2 + \ldots + \Sigma X_k}{N}.$$

$$SS_t = \Sigma(X_1 - \bar{X})^2 + \Sigma(X_2 - \bar{X})^2 + \ldots\ldots + \Sigma(X_k - \bar{X})^2;$$

$$\text{or, } SS_t = \Sigma X_1^2 + \Sigma X_2^2 + \ldots\ldots + \Sigma X_k^2 - \frac{(\Sigma X_1 + \Sigma X_2 + \ldots\ldots + \Sigma X_k)^2}{N}.$$

$$df_t = N - 1.$$

$$SS_b = n_1(\bar{X}_1 - \bar{X})^2 + n_2(\bar{X}_2 - \bar{X})^2 + \ldots\ldots + n_k(\bar{X}_k - \bar{X})^2,$$

$$\text{or, } SS_b = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \ldots\ldots + \frac{(\Sigma X_k)^2}{n_k} - \frac{(\Sigma X_1 + \Sigma X_2 + \ldots + \Sigma X_k)^2}{N}.$$

$$df_b = k - 1.$$

$$SS_w = \Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2 + \ldots\ldots + \Sigma(X_k - \bar{X}_k)^2;$$

$$\text{or, } SS_w = SS_t - SS_b. \quad df_w = N - k.$$

5.3.2. Computation of variances and *F* ratio

Between-groups and within-groups variances (s_b^2 and s_w^2) are next computed by dividing the respective sums of squares by their degrees of freedom. The within-groups variance (s_w^2) is used as the *error variance* in one-way anova to work out the F ratio.

$$s_b^2 = \frac{SS_b}{df_b} = \frac{SS_b}{k-1}; \quad s_w^2 = \frac{SS_w}{df_w} = \frac{SS_w}{N-k}; \quad F = \frac{s_b^2}{s_w^2};$$

df of F : $df_b, df_w = k - 1, N - k$.

5.3.3. Significance test of computed *F*

The H_o proposes that the computed F is not significant, has resulted from the use of a sample drawn by random sampling, and would not differ significantly from 1.00 if the entire population is used for the experiment.

To test the significance of the computed F, the latter is compared with the critical F_α value ($df : k - 1, N - k$) for a chosen level of significance. If the computed F either exceeds or equals the critical F_α, the probability (P) of the H_o being correct is considered too low ($P \leq \alpha$), the H_o is rejected and there is some significant difference between the group means; in other words, there is a significant *added*

variance between the groups, which is not present within the groups. But if the critical F_α exceeds the computed F value, P exceeds α and there is no significant difference between the group means. (See Example 5.3.1.)

Example 5.3.1.

Work out a one-way anova to find whether or not there is any significant difference between the mean tracheal ventilation scores (ml/minute) of the following two groups of locusts from two different habitats. ($\alpha = 0.05$.)

Group 1 (X_1) : 75, 78, 80, 75, 81, 85, 82, 78, 80, 86.

Group 2 (X_2) : 71, 72, 70, 70, 73, 74, 73, 65, 70, 82.

Critical F scores : $F_{0.05(2,19)} = 3.52$; $F_{0.05(1,18)} = 4.41$;

$$F_{0.05(1,19)} = 4.38; \quad F_{0.05(2,18)} = 3.55.$$

Solution :

$N = n_1 + n_2 = 10 + 10 = 20$.

$$SS_t = \Sigma X_1^2 + \Sigma X_2^2 - \frac{(\Sigma X_1 + \Sigma X_2)^2}{N} = 64124 + 52008 - \frac{(800+720)^2}{20} = 612.$$

$$SS_b = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} - \frac{(\Sigma X_1 + \Sigma X_2)^2}{N} = \frac{800^2}{10} + \frac{720^2}{10} - \frac{(800+720)^2}{20} = 320.$$

$df_b = k - 1 = 2 - 1 = 1$.

$SS_w = SS_t - SS_b = 612 - 320 = 292$. $\quad df_w = N - k = 20 - 2 = 18$.

Table 5.1. One-way anova between two groups.

	X_1	X_2	X_1^2	X_2^2
	75	71	5625	5041
	78	72	6084	5184
	80	70	6400	4900
	75	70	5625	4900
	81	73	6561	5329
	85	74	7225	5476
	82	73	6724	5329
	78	65	6084	4225
	80	70	6400	4900
	86	82	7396	6724
Σ_i	800	720	64124	52008

$$s_b^2 = \frac{SS_b}{df_b} = \frac{320}{1} = 320. \quad s_w^2 = \frac{SS_w}{df_w} = \frac{292}{18} = 16.22. \quad F = \frac{s_b^2}{s_w^2} = \frac{320}{16.22} = 19.73.$$

df of F : $k - 1$, $N - k = 1, 18$. Critical $F_{0.05(1,18)} = 4.41$.

As the computed F is higher than critical $F_{0.05}$, P is too low. So, H_o is rejected. Hence, there is a significant difference between the group means ($P < 0.05$).

As the independent variable, viz., habitat, is a random variable beyond the control of the investigator, the *added variance component* (s_a^2) is computed. (See Sub-section 5.3.5.) Using the identical size (n) of each group,

$$n = 10; \quad s_a^2 = \frac{s_b^2 - s_w^2}{n} = \frac{320 - 16.22}{10} = 30.38.$$

Therefore, an added variance component of 30.38, absent within the groups, occurs between the groups.

5.3.4. Strength of association

If the preliminary F test yields a significant F ratio and the anova undertaken belongs to model I using a fixed experimental treatment as the independent variable, then the strength of association has to be worked out between dependent and independent variables as the *omega square* (ω^2), irrespective of the anova being worked out for two or more groups. Using the number (k) of groups in the experiment, the total (N) of all the group sizes and the computed F value,

$$\omega^2 = \frac{(k-1)(F-1)}{N + (k-1)(F-1)}.$$

Omega square is not to be computed if the F ratio turns out to be not significant, or if a model II anova is being done. (See Example 5.4.1.)

5.3.5. Added variance component

In case a model II anova yields a significant F ratio, the *added variance component* (s_a^2) has to be worked out, irrespective of the anova being worked out for two or more groups. It serves as an estimate of such variance due to random variable(s) as is a part of the between-groups variance (s_b^2), but not a part of the within-groups variance (s_w^2). When all the groups have an identical size (n),

$$s_a^2 = \frac{s_b^2 - s_w^2}{n}.$$

If the F ratio is not significant, or if a model I anova has been undertaken, s_a^2 is not to be computed. (See Example 5.3.1.)

5.4 Multiple comparison Scheffe's F test

In case of a one-way anova with more than two groups ($k > 2$), a significant F ratio has to be followed up by a multiple comparison test to find whether or not the means of two groups of each chosen pair differ significantly. The multiple comparison test has to be worked out separately for each chosen pair of groups and has, therefore, to be repeated as many times as the number of chosen pairs. Multiple comparison tests need not be done if the F ratio has turned out to be not significant, or if the one-way anova has been undertaken between only two groups. Frequently, either the t test or the F test of Scheffe is carried out for multiple comparison. Multiple comparison *Scheffe's F test*, more powerful and preferable of the two, is briefly described below.

For Scheffe's F test between the means ($\bar{X}_2$ and $\bar{X}_3$) of groups 2 and 3, for example, the standard error of difference ($s_{\bar{X}_2-\bar{X}_3}$) between those means is first worked out using the within-groups variance (s_w^2) computed in the earlier preliminary F test. The squared SE of difference is then used in transforming the squared difference between the means into the Scheffe's F score.

$$s_{\bar{X}_2-\bar{X}_3}=\sqrt{\frac{s_w^2}{n_2}+\frac{s_w^2}{n_3}};\quad F=\frac{(\bar{X}_2-\bar{X}_3)^2}{s^2_{\bar{X}_2-\bar{X}_3}},\quad \text{or}\quad F=\frac{(\bar{X}_2-\bar{X}_3)^2}{\frac{s_w^2}{n_2}+\frac{s_w^2}{n_3}}.$$

The computed Scheffe's F is then compared with the critical F_α' value for the chosen level of significance; the critical F_α' has to be worked out in turn from the critical F_α value as follows. Where k is the number of groups,

$$F_\alpha' = (k-1)F_{\alpha(k-1,N-k)}\,.$$

Only if the computed Scheffe's F exceeds or equals the critical F_α', the probability P of the H_o being correct is considered too low ($P \le \alpha$), the H_o is rejected, and the two means being tested are considered to differ significantly. (See Example 5.4.1.)

Example 5.4.1.

(a) Work out a one-way anova to find whether or not there is a significant difference between the mean tracheal ventilations (ml/minute) of the following group of cockroaches, treated with three different levels of a pesticide. ($\alpha = 0.01$.)

(b) If the preliminary F test shows significant differences, find whether or not the means of the first and second group differ significantly. ($\alpha = 0.01$.)

Animal :	1	2	3	4	5	6	7	8	9	10
X_1 ml :	90	120	90	120	80	97	100	97	130	86
X_2 ml :	55	70	54	68	35	60	60	58	80	40
X_3 ml :	28	30	25	35	15	30	38	28	50	21

Critical F values : $F_{0.01(2,26)} = 5.53$; $F_{0.01(2,27)} = 5.49$;

$F_{0.01(2,28)} = 5.45$; $F_{0.01(1,29)} = 4.18$; $F_{0.01(1,26)} = 4.22$.

Solution :

Table 5.2. One-way anova between three groups.

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2
90	55	28	8100	3025	784
120	70	30	14400	4900	900
90	54	25	8100	2916	625
120	68	35	14400	4624	1225
80	35	15	6400	1225	225
97	60	30	9409	3600	900
100	60	38	10000	3600	1444
97	58	28	9409	3364	784
130	80	50	16900	6400	2500
86	40	21	7396	1600	441
Σ 1010	580	300	104514	35254	9828

$$N = n_1 + n_2 + n_3 = 10 + 10 + 10 = 30. \quad \bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{1010}{10} = 101.0.$$

$$\bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{580}{10} = 58.0. \quad \bar{X}_3 = \frac{\Sigma X_3}{n_3} = \frac{300}{10} = 30.0.$$

$$SS_t = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 - \frac{(\Sigma X_1 + \Sigma X_2 + \Sigma X_3)^2}{N},$$

or, $SS_t = 104514 + 35254 + 9828 - \frac{(1010+580+300)^2}{30} = 30526.$

$$SS_b = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} - \frac{(\Sigma X_1 + \Sigma X_2 + \Sigma X_3)^2}{N}$$

$$= \frac{1010^2}{10} + \frac{580^2}{10} + \frac{300^2}{10} - \frac{(1010+580+300)^2}{30} = 25580.$$

$df_b = k - 1 = 3 - 1 = 2.$

$SS_w = SS_t - SS_b = 30526 - 25580 = 4946. \quad df_w = N - k = 30 - 3 = 27.$

$s_b^2 = \frac{SS_b}{df_b} = \frac{25580}{2} = 12790. \quad s_w^2 = \frac{SS_w}{df_w} = \frac{4946}{27} = 183.19.$

$F = \frac{s_b^2}{s_w^2} = \frac{12790}{183.19} = 69.82. \quad df \text{ for } F : df_b, df_w = 2, 27.$

Critical $F_{0.01(2,27)}$=5.49.

As the computed F exceeds the critical $F_{0.01}$, P is too low ($P < 0.01$). So H_o is rejected and there are significant differences between the group means.

Because the independent variable, viz., chosen levels of the pesticide, is a fixed treatment variable, *omega square* is computed for the strength of association between the dependent and independent variables.

$$\omega^2 = \frac{(k-1)(F-1)}{N + (k-1)(F-1)} = \frac{(3-1)(69.82-1)}{30 + (3-1)(69.82-1)} = 0.82.$$

Thus, 0.82 proportion of the total variance of the dependent variable is associated with the independent variable.

As the preliminary F test yielded a significant F score, the difference $(\bar{X}_1 - \bar{X}_2)$ is subjected to the multiple comparison *Scheffe's F test* to find its significance.

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_w^2}{n_1} + \frac{s_w^2}{n_2}} = \sqrt{\frac{183.19}{10} + \frac{183.19}{10}} = 6.053.$$

$$F = \frac{(\bar{X}_1 - \bar{X}_2)^2}{s^2_{\bar{X}_1 - \bar{X}_2}} = \frac{(101.0 - 58.0)^2}{6.053^2} = 50.47.$$

$$F'_{0.01} = (k-1)\, F_{0.01(df_b, df_w)} = (3-1) F_{0.01(2,\,27)}$$

$$= (3\text{–}1) \times 5.49 = 10.98.$$

As the computed F exceeds the critical $F'_{0.01}$, P is too low. So, the H_o is rejected and there is a significant difference between $\bar{X}_1$ and $\bar{X}_2$ ($P < 0.01$).

Example 5.4.2.

Work out a one-way anova to find whether or not there is a significant difference between the blood sugar scores (mg/dl) of a group of alloxan-diabetic monkeys treated with two different levels of a hypoglycemic agent. ($\alpha = 0.01$.)

Animal :	1	2	3	4	5	6	7	8	9	10
X_1 mg :	160	180	200	170	197	220	162	180	201	190
X_2 mg :	110	140	160	125	162	168	120	135	185	165

Critical F values : $F_{0.01(2,\,19)} = 5.93$; $F_{0.01(1,\,19)} = 8.18$.

$F_{0.01(1,18)} = 8.28$; $F_{0.01(2,18)} = 6.01$.

Solution :

Table 5.3. One-way anova between two groups.

	X_1	X_2	X_1^2	X_2^2
	160	110	25600	12100
	180	140	32400	19600
	200	160	40000	25600
	170	125	28900	15625
	197	162	38809	26244
	220	168	48400	28224
	162	120	26244	14400
	180	135	32400	18225
	201	185	40401	34225
	190	165	36100	27225
Σ	1860	1470	349254	221468

$N = n_1 + n_2 = 10 + 10 = 20.$

$$SS_t = \Sigma X_1^2 + \Sigma X_2^2 - \frac{(\Sigma X_1 + \Sigma X_2)^2}{N} = 349254 + 221468 - \frac{(1860+1470)^2}{20} = 16277.$$

$$SS_b = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} - \frac{(\Sigma X_1 + \Sigma X_2)^2}{N} = \frac{1860^2}{10} + \frac{1470^2}{10} - \frac{(1860+1470)^2}{20} = 7605$$

$df_b = k - 1 = 2 - 1 = 1.$

$SS_w = SS_t - SS_b = 16277 - 7605 = 8672.$ $df_w = N - k = 20 - 2 = 18.$

$$s_b^2 = \frac{SS_b}{df_b} = \frac{7605}{1} = 7605. \quad s_w^2 = \frac{SS_w}{df_w} = \frac{8672}{18} = 481.78$$

$$F = \frac{s_b^2}{s_w^2} = \frac{7605}{481.78} = 15.79. \quad df \text{ of } F : df_b, df_w = 1,18$$

Critical $F_{0.01(1,18)} = 8.28.$

As the computed F is higher than critical $F_{0.01}$, $P < 0.01$. So, H_o is rejected and the group means differ significantly.

Because the independent variable consists of chosen levels of a hypoglycemic agent, it is a fixed treatment variable. So, *omega square* is computed as the measure of the strength of association between the dependent and independent variables.

$$\omega^2 = \frac{(k-1)(F-1)}{N + (k-1)(F-1)} = \frac{(2-1)(15.79-1)}{20 + (2-1)(15.79-1)} = 0.43.$$

Thus, 0.43 proportion of the total variance of the dependent variable is associated with the independent variable.

5.5 Kruskal-Wallis nonparametric anova

Kruskal-Wallis nonparametric anova is a *rank-dependent one-way anova* for two or more groups. It is a useful alternative to the parametric method of one-way anova, with few assumptions, wide applications and easy computation. It can be used when the assumptions for the parametric anova are not justifiable. But it is less powerful than the parametric anova.

5.5.1. Assumptions for Kruskal-Wallis anova

Being a rank-dependent method, Kruskal-Wallis anova may be worked out for ordinal variables, and also for continuous or discontinuous variables, after ranking the cases of the groups in a composite manner (see Sub-section 5.5.2). Following assumptions should be justifiable for computing the Kruskal-Wallis H. (i) The dependent variable should be either an ordinal variable with the cases of the groups already expressed in ranks, or a continuous or discontinuous measurement variable whose scores can be converted into ranks. (ii) Each score/rank or its error term

should occur at random depending on laws of probability and independent of all other scores/ranks or error terms. (iii) No assumption is needed for the continuous nature of the variable. (iv) No assumption is necessary for the normality of distribution of its scores or error terms. Kruskal-Wallis test connot be applied to nominal variables as ranks cannot be given to them. Moreover, this test is less powerful than the parametric anova and should, therefore, be used only if the assumptions for the latter are seen not to be justifiable.

5.5.2. Computation of Kruskal-Wallis *H*

Ranks are first given in an ascending order and a *composite manner* to the scores or ranks of all the k number of groups taken together. Even if the variable is an ordinal one with the cases already ranked separately in each group, fresh composite ranking must be done again taking all the groups together. Identical scores, whether occurring in the same group or in separate groups, constitute a *tied set*; each member of a tied set is given an *average rank* identical with the arithmetic mean of the ranks the tied cases would have got if they were separate consecutive scores. Moreover, the case coming next to a tied set is given that rank which it would have got if it followed a separate untied case. You may realise that Kruskal-Wallis anova would be less powerful and more prone to error because of (i) the *average rank* given to the cases of each tied set, and (ii) the *unequal differences* between successive scores ranked consecutively.

After the composite ranking of the cases of all the k number of groups taken together, the ranks given to each group are totalled separately to give the *rank sum* of the corresponding group (e.g., $R_1, R_2, \ldots R_k$). Each rank sum is divided by the size ($n_1, n_2, \ldots n_k$) of that group to give the *mean rank* ($\bar{R}_1, \bar{R}_2, \ldots \bar{R}_k$) of the latter; the *grand mean* ($\bar{R}$) of all the groups is worked out dividing the total of all the rank sums by the total size (N) of all the groups. These values are then used in working out the statistic H.

$$N = n_1 + n_2 + \ldots\ldots\ldots + n_k. \quad \bar{R}_1 = \frac{R_1}{n_1}; \quad \bar{R}_2 = \frac{R_2}{n_2}, \ldots\ldots \bar{R}_k = \frac{R_k}{n_k}.$$

$$\bar{R} = \frac{R_1 + R_2 + \ldots\ldots + R_k}{N}.$$

$$H = \frac{12}{N(N+1)}[n_1(\bar{R}_1 - \bar{R})^2 + n_2(\bar{R}_2 - \bar{R})^2 + \ldots\ldots\ldots\ldots + n_k(\bar{R}_k - \bar{R})^2].$$

$$df = k - 1.$$

The H_0 proposes that the computed H is not significant and has resulted from chances associated with random sampling using laws of probability. To find the

probability P of this H_0 being correct, the computed H is compared with the critical χ^2_α with the computed *df* and for the chosen α. The computed H is considered significant, only if it either exceeds or equals the critical χ^2_α ($P \leq \alpha$). A significant H indicates the existence of significant difference between the group means. The Kruskal-Wallis anova for only two groups ends here. But if the anova is worked out with more than two groups, a significant H has to be followed up by the *multiple comparison Mann-Whitney U test* to explore the significance of difference between the groups of each chosen pair. (See Examples 5.5.1 and 5.6.1.)

Example 5.5.1.

Apply Kruskal-Wallis anova to find whether or not there is a significant difference between mean corpuscular hemoglobin scores (picograms) of a group of microcytic anemia patients (Group 1) and a group of macrocytic anemia patients (Group 2). ($\alpha = 0.05$.)

Group 1 (X_1) : 23 18 19 17 25 20 32 22 33 16.

Group 2 (X_2) : 32 35 25 28 34 32 30 31 40 42.

Critical chi square scores : $\chi^2_{0.05(19)} = 30.14$; $\chi^2_{0.05(18)} = 28.87$;

$$\chi^2_{0.05(1)} = 3.84; \quad \chi^2_{0.05(2)} = 5.99 .$$

Solution :

Table 5.4. Composite ranking for Kruskal-Wallis anova between two groups.

X_1	Ranks	X_2	Ranks
23	7	32	14
18	3	35	18
19	4	25	8.5
17	2	28	10
25	8.5	34	17
20	5	32	14
32	14	30	11
22	6	31	12
33	16	40	19
16	1	42	20
Σ —	66.5 (R_1)	—	143.5 (R_2)

$n_1 = 10;\ \ n_2 = 10;\ \ N = n_1 + n_2 = 10 + 10 = 20.$

$R_1 = 66.5;\ \ R_2 = 143.5.\ \ \bar{R}_1 = \frac{R_1}{n_1} = \frac{66.5}{10} = 6.65;\ \ \bar{R}_2 = \frac{R_2}{n_2} = \frac{143.5}{10} = 14.35.$

$$\bar{R} = \frac{R_1 + R_2}{N} = \frac{66.5 + 143.5}{20} = 10.50.$$

$$H = \frac{12}{N(N+1)}[n_1(\bar{R}_1 - \bar{R})^2 + n_2(\bar{R}_2 - \bar{R})^2]$$

$$= \frac{12}{20(20+1)}[10(6.65 - 10.50)^2 + 10(14.35 - 10.50)^2] = 8.47.$$

$df = k - 1 \quad = 2 - 1 \ = 1.$

$\chi^2_{0.05(1)} = 3.84.$

As the computed H exceeds the critical $\chi^2_{0.05}$, P is too low ($P < 0.05$). So, H_o is rejected and there is a significant difference in MCH between microcytic and macrocytic patients.

5.6 Multiple comparison Mann-Whitney *U* test

This is a *rank-dependent test* used as a powerful alternative to Student's *t* test between two groups, and also as a *multiple comparison test* between the groups of each chosen pair after a significant H test with more than two groups ($k > 2$).

Assumptions, applications and inaccuracies of this test are similar to those of the Kruskal-Wallis H test.

For a multiple comparison test between two chosen group means (say, $\bar{X}_1$ and $\bar{X}_2$), ranks are given in an ascending order and a *composite manner* to the scores of both the groups taken together, and *rank sums* (R_1 and R_2) of the respective groups are worked out separately. The rank sum and the group size (n_1 or n_2) of any of the groups are used in computing the corresponding statistic U. Thus,

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1,$$

or, $$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2.$$

The U_e, expected in terms of the H_o, and the SE (s_U) of U are computed using the group sizes. These are in turn used in transforming either U_1 or U_2 into the z score.

$$U_e = \frac{n_1 n_2}{2}; \quad s_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}};$$

$$z_1 = \frac{U_1 - U_e}{s_U}, \quad \text{or} \quad z_2 = \frac{U_2 - U_e}{s_U}.$$

Either of the two z scores, identical in magnitude but opposite in algebraic sign, may be used in finding the probability P of the H_o being correct.

P = 2 [0.5000 – (area of unit normal curve from μ to the computed z)].

Only if the P thus worked out does not exceed the chosen level of significance (α), P is considered too low ($P \leq \alpha$) and there is a significant difference between the relevant group means. (See Example 5.6.1.)

Example 5.6.1.

(a) Apply Kruskal-Wallis anova to find whether or not there is any significant difference between the mean knee-jerk scores (degrees of arc) of the following three groups of rhesus monkeys from three different habitats. (α = 0.001.)

Group 1 (X_1) : 33, 37, 31, 34, 33, 29, 30, 35, 32, 28, 27, 36.

Group 2 (X_2) : 27, 22, 18, 28, 23, 25, 28, 21, 20, 19, 24, 19.

Group 3 (X_3) : 18, 16, 14, 10, 11, 18, 12, 15, 13, 17, 14, 9.

Critical chi square values : $\chi^2_{0.001(3)} = 16.27$; $\chi^2_{0.001(2)} = 13.82$;

$$\chi^2_{0.001(30)} = 59.70; \quad \chi^2_{0.001(29)} = 58.30.$$

(b) In case the Kruskal-Wallis H is significant, work out Mann-Whitney multiple comparison U test to find if there is any significant difference between the mean knee-jerks of groups 2 and 3. (α = 0.001).

z scores	:	3.60	3.70	3.80	3.90	4.00
Areas of unit normal curve	:	0.4998	0.4999	0.49993	0.49995	0.49997

Solution :

n_1 = 12; n_2 = 12 ; n_3 = 12 ; $N = n_1 + n_2 + n_3 = 12 + 12 + 12 = 36$.

Table 5.5. Composite ranking for Kruskal-Wallis anova between three groups.

X_1	Ranks	X_2	Ranks	X_3	Ranks
33	31.5	27	22.5	18	12
37	36	22	18	16	9
31	29	18	12	14	6.5
34	33	28	25	10	2
33	31.5	23	19	11	3
29	27	25	21	18	12
30	28	28	25	12	4
35	34	21	17	15	8
32	30	20	16	13	5
28	25	19	14.5	17	10
27	22.5	24	20	14	6.5
36	35	19	14.5	9	1
Σ	—	362.5 (R_1)	—	224.5 (R_2)	—	79.0 (R_3)

$R_1 = 362.5;\ R_2 = 224.5;\ R_3 = 79.0.$

$$\bar{R}_1 = \frac{R_1}{n_1} = \frac{362.5}{12} = 30.21;\ \bar{R}_2 = \frac{R_2}{n_2} = \frac{224.5}{12} = 18.71;\ \bar{R}_3 = \frac{R_3}{n_3} = \frac{79.0}{12} = 6.58;$$

$$\bar{R} = \frac{(R_1 + R_2 + R_3)}{N} = \frac{(362.5 + 224.5 + 79.0)}{36} = 18.50.$$

$$H = \frac{12}{N(N+1)}[n_1(\bar{R}_1 - \bar{R})^2 + n_2(\bar{R}_2 - \bar{R})^2 + n_3(\bar{R}_3 - \bar{R})^2]$$

$$= \frac{12}{36(36+1)}[12(30.21-18.50)^2 + 12(18.71-18.50)^2 + 12(6.58-18.50)^2]$$

$=30.19.$

$df = k - 1 \quad = 3 - 1 = 2.$

Critical $\chi^2_{0.001(2)} = 13.82.$

Because the computed H is higher than the critical $\chi^2_{0.001}$, P is too low ($P < 0.001$). So, the H_o is rejected and there are significant differences between the means of the three groups.

As there are more than two groups, *Mann-Whitmey U test* is undertaken to test the significance between $\bar{X}_2$ and $\bar{X}_3$ (Table 5.6).

Table 5.6. Composite ranking for Mann-Whitney test.

X_2	Ranks	X_3	Ranks
27	22	18	12
22	18	16	9
18	12	14	6.5
28	23.5	10	2
23	19	11	3
25	21	18	12
28	23.5	12	4
21	17	15	8
20	16	13	5
19	14.5	17	10
24	20	14	6.5
19	14.5	9	1
Σ	—	221.0 (R_2)	—	79.0 (R_3)

After the composite ranking of all X_2 and X_3 scores of the two relevant groups in Table 5.6, the rank sums (R_2 and R_3) of the respective groups are worked out and used along with the group sizes (n_2 and n_3) for computing either of two respective U scores.

$$U_2 = n_2 n_3 + \frac{n_2(n_2+1)}{2} - R_2 = 12\times12 + \frac{12(12+1)}{2} - 221.0 = 1.0, \quad \text{or}$$

$$U_3 = n_2 n_3 + \frac{n_3(n_3+1)}{2} - R_3 = 12\times12 + \frac{12(12+1)}{2} - 79.0 = 143.0.$$

U_e and s_U are then computed and used along with either U_2 or U_3 to work out the z score.

$$U_e = \frac{n_2 n_3}{2} = \frac{12\times12}{2} = 72.0; \quad s_U = \sqrt{\frac{n_2 n_3 (n_2+n_3+1)}{12}} = \sqrt{\frac{12\times12(12+12+1)}{12}} = 17.32.$$

$$z_2 = \frac{U_2 - U_e}{s_U} = \frac{1.0-72.0}{17.32} = -4.10, \quad \text{or}$$

$$z_3 = \frac{U_3 - U_e}{s_U} = \frac{143.0-72.0}{17.32} = 4.10.$$

Any of the two computed z scores is used for finding the significance, ignoring any negative sign of z. Because the computed z is higher than even 4.00, the probability P of the correctness of H_o is worked out as follows.

If the computed z were 4.00, P would have been lower than 0.0001 :

$P = 2$ [0.5000 – (area of unit normal curve from μ to the z score of 4.00)]

$= 2[0.5000 - 0.49997] = 0.00006.$

Because either computed z amounts to 4.10, the corresponding P would be lower than 0.00006 and so, much lower than the chosen α of 0.001. So, the H_o is rejected. Hence, there is a significant difference between $\bar{X}_2$ and $\bar{X}_3$. ($P << 0.001$).

5.7 Summary

Anova is used to test the significance of difference between the group means of a dependent variable, on exposure of two or more groups of cases to the levels of one or more independent variables. It has more rigorous assumptions and higher power than Student's t test and can be applied at a time to two or any higher number of groups. Anova may be one-way, two-way or of a still higher order according to the number of independent variables, and model I, model II or model III according to the nature(s) of the latter.

In a one-way anova, the total variance of scores of the dependent variable is resolved into between-groups and within-groups variances, and the F ratio of those two variances is compared with critical F values for different levels of significance to find whether or not the group means differ significantly. A significant F ratio in a one-way anova for more than two groups is followed up by multiple comparison Scheffe's F test to find whether or not the means of groups of any chosen pair differ significantly. A significant F ratio is also followed by the working out of either the omega square in a model I anova, or the added variance component for a model II anova.

Where the assumptions for one-way parametric anova are not justifiable, Kruskal-Wallis nonparametric rank-dependent anova may be undertaken. If this test yields a significant statistic H, and if it involves more than two groups, Mann-Whitney multiple comparison U test is used for testing the significance of difference between the means of any two chosen groups.

5.8 Terminal questions

1. (a) State the assumptions for Kruskal-Wallis one-way anova.

(b) Describe how you would rank the scores of a measurement variable and work out Kruskal-Wallis anova for two groups.

(c) Work out Kruskal-Wallis anova to find whether or not there is a significant difference between the tracheal ventilation scores (ml/min) of the following groups of locusts. ($\alpha = 0.05$.)

Group 1 : 73, 76, 86, 76, 83, 78, 73, 79, 78.

Group 2 : 70, 69, 73, 65, 63, 69, 76, 60, 62, 66, 67.

Critical chi square scores : $\chi^2_{0.05(2)} = 5.99$; $\chi^2_{0.05(1)} = 3.84$;

$\chi^2_{0.05(19)} = 30.14$; $\chi^2_{0.05(18)} = 28.87$.

2. (a) Explain what you understand by one-way anova and model I anova.

(b) Discuss the assumptions for anova.

(c) Work out a one-way anova to find if there is a significant difference between the mean wing length scores (mm) of the following groups of houseflies from two different habitats. ($\alpha = 0.01$)

Group 1 : 4.8, 4.6, 5.4, 4.6, 5.3, 4.7, 4.8, 5.0, 5.2, 5.0.

Group 2 : 3.8, 3.0, 3.3, 4.1, 3.7, 4.0, 3.9, 3.5, 4.3, 3.8.

Critical F values : $F_{0.01(2,19)} = 5.93$; $F_{0.01(1,18)} = 8.28$;

$F_{0.01(2,18)} = 6.01$; $F_{0.01(1,19)} = 8.18$.

3. (a) Write briefly about the uses and computations of omega square and added variance component.

(b) Describe briefly the variances involved in a one-way anova and their partitioning and uses, mentioning the relevant formulae.

(c) Using the data of Question 1 (c), work out a one-way anova to find whether or not there is a significant difference between the given tracheal ventilation scores of the two groups of locusts, exposed to two different fixed levels of amobarbital. ($\alpha = 0.05$)

Critical F values : $F_{0.05(2,19)} = 3.52$; $F_{0.05(1,18)} = 4.41$;

$F_{0.05(1,19)} = 4.38$.

4. (a) State the reasons for preferring anova to t test.

(b) Describe different models of anova with examples.

(c) Using a one-way anova, find whether or not there is any significant

difference between the following blood sugar scores (mg per 100 ml) of a group of diabetic orangutans treated with three different levels of a postulated antidiabetic agent. ($\alpha = 0.01$).

Animal :	1	2	3	4	5	6	7	8	9	10
X_1 mg :	165	192	185	170	200	197	220	180	210	201
X_2 mg :	120	165	140	125	160	162	178	150	163	157
X_3 mg :	85	120	95	87	123	120	135	115	130	120

Critical F values : $F_{0.01(2,27)} = 5.49$; $F_{0.01(2,28)} = 5.45$;

$F_{0.01(1,27)} = 7.68$.

(d) If the computed F is found significant, work out Scheffe's F test to find whether or not there is a significant difference between the mean blood sugars of groups 1 and 3 in the preceding experiment. ($\alpha = 0.01$).

5. (a) Describe briefly Scheffe's multiple comparison F test, quoting its computational formulae.

(b) Give an account of how you would work out the Mann-Whitney U test and find the significance of the computed U, stating the relevant formulae.

(c) Using the data of Question 2(c), work out the Kruskal-Wallis anova to find if there is any significant difference between the given wing length scores of the two groups. ($\alpha = 0.01$)

Critical chi square values : $\chi^2_{0.01(2)} = 9.21$; $\chi^2_{0.01(1)} = 6.64$;

$\chi^2_{0.01(19)} = 36.19$; $\chi^2_{0.01(18)} = 34.80$.

5.9 Answers

1. (a) See Sub-section 5.5.1.
   (b) See Sub-section 5.5.2.
   (c) See Example 5.5.1.
2. (a) See Sub-sections 5.2.3 and 5.2.4.
   (b) See Sub-section 5.2.5.
   (c) See Example 5.3.1.
3. (a) See Sub-sections 5.3.4 and 5.3.5.
   (b) See Sub-sections 5.2.1, 5.3.1 and 5.3.2.

(c) See Example 5.4.2.

4. (a) See Sub-section 5.2.2.
   (b) See Sub-section 5.2.4.
   (c) See Example 5.4.1.
   (d) See Example 5.4.1.
5. (a) See Section 5.4.
   (b) See Section 5.6.
   (c) See Example 5.5.1.

Unit 6 ❑ CORRELATION AND REGRESSION

Structure

6.1 Introduction

You are aware that any system— be it an organism, a cell, an organelle, a tissue, a body fluid, a body cavity, a space, a waterbody, a farmland or a forest— contains or consists of innumerable variables. Between two or more variables in a system, there may exist different types of relations or associations. In this unit, you will learn to explore and assess such associations between variables in any system. You will also come to know here how the likely value or score of one such variable may be mathematically predicted by measuring the score(s) of one or more other variables associated with the former one.

Objectives

On studying this unit, you should be able to do the following :

- understand what is meant by correlation and know its different types,
- describe the general properties of correlation,
- understand the assumptions for product-moment correlation and its applications,
- describe the properties of product-moment r,
- work out product-moment r between two given variables and find its significance,
- understand where Kendall's rank correlation should be used instead of product-

moment r,

- compute Kendall's *tau* between two variables and find its significance,
- know when and how partial and multiple linear correaltions are worked out and their significances are found out,
- understand what is meant by regression and know its different types and models,
- describe the assumptions and properties of simple linear regression,
- work out simple linear regression for predicting the score of one variable on the measured score of another, and
- know when and how to work out multiple linear regression of one variable on the combination of observed scores of two or more other variables.

6.2 Correlation

Correlation is the quantitative estimation and numerical expression of the magnitude or strength as well as the algebraic sign or direction of the association between two or more variables in a system. The correlation coefficient serves basically as a measure of the intensity or degree of association between the variances (Subsections 2.6.3 and 5.2.1) of two or more variables in the cases of the sample, while its algebraic sign is the indicator of whether those variables vary in the same direction or in opposite directions. The correlation coefficient is the sample statistic for correlation and ranges from -1.00 to $+1.00$ in value.

6.2.1. Types of correlation

Correlation may be *simple* or *multiple*, according as it is computed between two variables or more than two variables. For example, there may be a *simple correlation* between trunk length and wing length in a sample of cockroaches ; on the contrary, there may be a *multiple correlation* between oxygen consumption and the combination of atmospheric oxygen tension and tracheal ventilation volume in a sample of locusts.

Correlation may again be *linear* and *nonlinear*, according as the relation between the variables conforms to a straight line equation and a linear graph, or follows the equation of a curved line and a nonlinear graph. For example, there may be a *simple linear correlation* between body weight and gill weight in a sample of fishes ; but the initial velocity of an enzyme action and the corresponding substrate concentration may have a *simple nonlinear correlation*, conforming to a rectangular hyperbola.

Correlation may also be either *positive* or *negative*. If high scores of a variable are mostly accompanied by high scores of another variable while low scroes of one are usually associated with low scores of the other, the two variables are varying in the same direction and are said to bear a *positive correlation* with each other ; an example is the positive correlation between body height and body weight in many

samples. But if high scores of one variable are mostly accompanied by low scores of another variable while high scores of the latter are usually associated with low scores of the former, the scores of the variables are usually varying in opposite directions and the variables bear a *negative correlation* with each other ; for example, a negative correlation exists between blood sugar level and blood insulin level in animal samples.

While the magnitude of correlation is expressed numerically, ranging from −1.00 to +1.00, the algebraic +/− sign preceding the numerical value indicates whether the correlation is positive or negative.

6.2.2. Properties of correlations

General properties of correlations are summarized below. (i) A correlation coefficient worked out with a sample drawn from a population would hold good only within the limits of the particular stratum or class of the population from which the sample has been drawn, and would also be confined within other conditions and situations prevailing during the work. Thus, a correlation coefficient worked out with a sample of adults may not hold good for children of the same population, or that worked out with a sample of females may not apply to males. (ii) A correlation coefficient between two variables does not necessarily indicate that variations of one of them may be either the cause or the effect of variations of the other ; their correlation may very well have arisen from the association of some other variable in common with both of them. (iii) A correlation coefficient *cannot directly predict* the score of one of the variables from that of the other in the same individual. (iv) The correlation coefficient between two variables varies from sample to sample even when they have been drawn from the same population ; so, the sample correlation coefficients (r) lie dispersed around the population correlation coefficient (ρ) to form a *sampling distribution* of r values, owing to their respective *sampling errors.*

6.3 Product-moment correlation

Karl Pearson's *product-moment correlation coefficient* or Pearson's r is a *simple linear correlation coefficient,* used in correlating two variables which have a linear association with each other.

6.3.1. Assumptions for product-moment r

Product-moment r can be applied for correlating two variables, only if it can be *logically assumed* that the following conditions or criteria are fulfilled in the case under investigation. (i) Both the variables being correlated should be *continuous measurement variables,* with their scores quantitatively measurable and occurring even in infinitely small fractional units, with no gaps in the respective scales of

scores. (ii) Scores of each variable should be distributed in the population in a unimodal, bilaterally symmetric or almost symmetric, *normal or nearly normal distribution* with not much skewness of its tails. (iii) There should exist a *linear association* between the variations of the two variables. (iv) The pair of scores of the two variables for each individual or case should have occurred in the sample *at random*, obeying the laws of probability and *independent* of all other similar pairs of scores ; this last assumption ensures that the sample may be a representative of the population, enabling the inference made from the sample to be generalized for the corresponding population.

It follows from these assumptions that the product moment r cannot be used in correlating such variables as are not associated linearly, or are discontinuous in nature (e.g., heart rate, cell count and litter size) or are ordinal variables (e.g., ferocity) or nonmeasurable qualitative variables (e.g., sex and race), or have prominently skewed or non-normal distributions in the population.

6.3.2. Properties of product-moment r

(a) The magnitude of the computed r is a measure of the *strength of association* between the variables while its algebraic sign indicates whether the variables vary in the same direction *(positive)* or in opposite directions *(negative)*. Thus, + 0.80 indicates a high positive correlation, – 0.72 shows a high negative correlation, +0.14 indicates a low positive correlation, while 0.00 means the absence of any linear correlation.

(b) If every score of any or each variable is multiplied, divided, added or subtracted by a constant number, it does not result in any change in the r value between the two variables.

(c) Correlation depends on that proportion of total variance of each variable which is associated with the variance of the other. This makes the value of r directly proportional to the *covariance* of the two variables. Where X and Y are the scores, $\bar{X}$ and $\bar{Y}$ are the means, s_x and s_y are the unbiased standard deviations, and $Cov(X,Y)$ is the covariance of two variables, and n is the sample size,

$$s_x = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}};\ s_y = \sqrt{\frac{\Sigma(Y-\bar{Y})^2}{n-1}};\ Cov(X,Y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{n-1};$$

$$r = \frac{Cov(X,Y)}{s_x s_y} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{(n-1)\,s_x s_y}.$$

(d) The r values, computed between two given variables in different samples from the same population, lie dispersed forming a *sampling distribution* of r around the population correlation coefficient (ρ) because of their varying differences, called

sampling errors (s_e), from the latter ; for each sample r, $s_e = r - \rho$. The s_e of r may be estimated by working out the *standard error* (s_r) of the r.

(e) Sampling distributions of r are unimodal and *symmetrical* if the corresponding population ρ amounts to zero, but are progressively more and more *skewed* either positively or negatively, according respectively to the rising negative and positive values of ρ.

(f) The squared value (r^2) of product-moment r between two variables is often used as a measure of that proportion of variance of each of them, which depends on the variance of the other variable ; this is called the *coefficient of determination* (r^2). For example, 0.49 proportion of the variance of X would be determined by the variance of Y, and vice versa, if r_{xy} amounts to -0.70.

6.3.3. Computation of r from ungrouped data

Using the ungrouped scores of the variables X and Y, product-moment r is generally worked out between them with either of the following alternative formulae, each derived from the foregoing equation for the relation between r and $Cov(X,Y)$. (See Sub-section 6.3.2).

(a) *From sum of products and sums of squares :*

Where $\bar{X}$ and $\bar{Y}$ are the sample means of the respective variables, $(X - \bar{X})$ and $(Y - \bar{Y})$ are the differences between each score and its mean, $\Sigma(X - \bar{X})^2$ and $\Sigma(Y - \bar{Y})^2$ are the *sums of squares* of respective variables, and $\Sigma(X - \bar{X})(Y - \bar{Y})$ is the *sum of products* of $(X - \bar{X})$ and $(Y - \bar{Y})$ values of each case,

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \, \Sigma(Y - \bar{Y})^2}}.$$

(b) *From raw scores of variables :*

Where n is the number of cases in the sample or the sample size, ΣX and ΣY are the sums of scores of the respective variables, ΣX^2 and ΣY^2 are the sums of respective squared scores, and ΣXY is the sum of the products of X and Y scores of each case,

$$r = \frac{n\Sigma XY - \Sigma X \, \Sigma Y}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}.$$

6.3.4. Significance of computed r

You should be aware that a significance test has to be undertaken to find out the

probability *(P)* of the null hypothesis *(H_0)* being correct, because the r has been worked out using a sample. The H_0 proposes here that the computed r has resulted from mere chances associated with random sampling by laws of probability and is not significant or meaningful. Stated differently, the H_0 contends that population ρ amounts to zero only, that the random choice of a particular sample has led to the observed value of r, and that the difference between the observed value of r and the zero value of population ρ can be explained away by the sampling error of the computed r.

To work out the probability P of the H_0 being correct, the computed r is transformed into Student's t by using the *SE* (s_r) of the r ; the *df* of t is also worked out.

$$s_r = \sqrt{\frac{1-r^2}{n-2}}; \qquad t = \frac{r}{s_r}; \qquad df = n-2 .$$

The computed t is next compared with two-tail critical t scores with the same *df*. Only if the computed t either exceeds or equals the critical t for a particular significance level (α) not higher than 0.05, P is considered too low ($P \leq \alpha$), the H_0 is consequently rejected and the computed r is considered significant. On the contrary, if the critical t exceeds the computed t, P is considered too high ($P > \alpha$), the H_0 is consequently retained and the computed r has no significance.

Example 6.3.1.

Work out product-moment r to find whether or not there is a significant correlation between tracheal ventilation (X ml/minute) and O_2 consumption (Y ml/minute) using the following data of a sample of grasshoppers. ($\alpha = 0.01$.)

Animal :	1	2	3	4	5	6	7	8	9	10
X :	66.0	89.1	72.0	87.5	75.2	78.2	83.5	71.6	85.6	76.3
Y :	3.3	4.9	3.5	4.7	3.7	4.0	4.3	3.4	4.4	3.8

Critical t scores : $t_{0.01(9)} = 3.250$; $t_{0.01(8)} = 3.355$; $t_{0.01(18)} = 2.878$.

Solution :

(a) *Using the sums of squares :*

$$\bar{X} = \frac{\Sigma X}{n} = \frac{785.0}{10} = 78.5 \text{ ml.}\ \bar{Y} = \frac{\Sigma Y}{n} = \frac{40.0}{10} = 4.0 \text{ ml.}$$

Table 6.1 : Table for computing sum of products and sums of squares.

	X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
	66.0	3.3	–12.5	–0.7	156.25	0.49	+ 8.75
	89.1	4.9	+10.6	+0.9	112.36	0.81	+ 9.54
	72.0	3.5	–6.5	–0.5	42.25	0.25	+ 3.25
	87.5	4.7	+9.0	+0.7	81.00	0.49	+ 6.30
	75.2	3.7	–3.3	–0.3	10.89	0.09	+ 0.99
	78.2	4.0	–0.3	0	0.09	0	0
	83.5	4.3	+5.0	+0.3	25.00	0.09	+ 1.50
	71.6	3.4	–6.9	–0.6	47.61	0.36	+ 4.14
	85.6	4.4	+7.1	+0.4	50.41	0.16	+ 2.84
	76.3	3.8	–2.2	–0.2	4.84	0.04	+ 0.44
Σ	785.0	40.0	–	–	530.70	2.78	+ 37.75

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2\,\Sigma(Y-\bar{Y})^2}} = \frac{37.75}{\sqrt{530.70\times 2.78}} = +0.98.$$

(b) *Using raw scores (alternative method) :*

Table 6.2 . Table for computing r from raw scores.

	X	Y	X^2	Y^2	XY
	66.0	3.3	4356.00	10.89	217.80
	89.1	4.9	7938.81	24.01	436.59
	72.0	3.5	5184.00	12.25	252.00
	87.5	4.7	7656.25	22.09	411.25
	75.2	3.7	5655.04	13.69	278.24
	78.2	4.0	6115.24	16.00	312.80
	83.5	4.3	6972.25	18.49	359.05
	71.6	3.4	5126.56	11.56	243.44
	85.6	4.4	7327.36	19.36	376.64
	76.3	3.8	5821.69	14.44	289.94
Σ	785.0	40.0	62153.20	162.78	3177.75

$$r = \frac{n\Sigma XY - \Sigma X\Sigma Y}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2]\,[n\Sigma Y^2 - (\Sigma Y)^2]}} = \frac{10\times 3177.75 - 785.0\times 40.0}{\sqrt{[10\times 62153.20 - 785.0^2]\times[10\times 162.78 - 40.0^2]}}$$

$= + 0.98.$

(c) *Significance of computed r :*

$$s_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0.98^2}{10-2}} = 0.070. \qquad t = \frac{r}{s_r} = \frac{0.98}{0.070} = 14.000$$

$df = n-2 = 10-2 = 8.$ $\alpha = 0.01.$ Critical $t_{0.01(8)} = 3.355.$

As the computed t far exceeds the critical $t_{0.01}$, P is considered too low. So, the H_0 is rejected. There is a significant correlation between tracheal ventilation and O_2 consumption. ($P < 0.01$).

6.4 Kendall's rank correlation

Kendall's *rank correlation coefficient* or Kendall's *tau* (τ) is a *rank-based* simple linear correlation coefficient. It is a powerful nonparametric alternative to product-moment r when the more rigorous assumptions for the latter are not fulfilled. It ranges from –1.00 to +1.00 in value.

6.4.1. Assumptions for Kendall's tau

You may compute Kendall's tan for correlating either ordinal variables with their magnitudes expressed in ranks, or measurement variables after first changing their scores into ranks, provided the following assumptions are justifiable. (i) The variables should either be ordinal variables with the cases of the sample already graded into ranks, or be such other variables whose scores can be changed into ranks. (ii) Ranks or scores of the two variables should bear linear association. (iii) The pair of ranks or scores of each case should have occurred due to random sampling using laws of probability so that the sample may be considered as representing the population. (iv) No assumption need be made for the continuous nature of any of the variables. (v) No assumption is also necessary for the normality of distribution of scores or ranks of either variable in the population.

It follows from these assumptions that Kendall's tau cannot be used for nominal variables like sex, as ranks cannot be assigned for them to the cases of the sample. Moreover, being less powerful than the product-moment r, tau should be used only when assumptions for r are found not to be logical.

6.4.2. Computation of Kendall's tau

(a) In case of ordinal variables with their magnitudes already graded in ranks, those ranks may be used directly in computing tau. But for continuous or discontinuous measurement variables, their scores are first changed into ranks in ascending order; the ranks of two variables would thus form two separate series of ranks (Table 6.3).

In this procedure, two or more identical scores of a variable are given an identical *average rank* which is the mean of the ranks they would have got if they were successive nonidentical scores ; the next score following such a *tied set* of scores is given the rank it would have got if the immediately preceding score was not a member of a tied set and did not enjoy any average rank. This system of ranking of scores leads to two types of *inaccuracies* in all rank-dependent statistics, viz. (i) error owing to *varying differences* between the scores bearing successive ranks, and (ii) error arising from an *average rank* given to every score of a tied set instead of their separate individual ranks. However, you will presently find that in computing tau, the inaccuracy owing to average ranks is sought to be minimized.

(b) After ranking the scores of two variables X and Y into two separate series, the ranks of that variable, which is free from any tied set or average rank, are arranged in an ascending order along a column of a table ; against every rank of this column, the paired rank of the same case with respect to the other variable is entered in a second column alongside (Table 6.4).

If each or neither of the variables has any tied set of scores, the ranks of any of them may be arranged in the first column, pairing each with the rank of the same case in the other variable along the second column.

(c) Moving downwards along the second column of ranks from its top, each of its ranks is used in turn as a *pivotal rank* for comparing with every successive subsequent rank following that pivotal rank, counting each subsequent rank as +1, 0 or −1 according as it exceeds, equals or falls short of the pivotal one, Such counts of subsequent ranks are entered in a third column and totalled as ΣC, after using all the second-column ranks in turn as successive pivotal ranks (Table 6.4).

(d) *Correction terms* (ΣT_X and ΣT_Y) are worked out for countering the errors due to the use of average ranks for X and Y scores, respectively, using the number (t_X or t_Y) of cases in each tied set of the respective variables.

$$\Sigma T_X = \Sigma[t_X (t_X - 1)] \; ; \qquad \Sigma T_Y = \Sigma[t_Y (t_Y - 1)].$$

(e) Kendall's tau (τ) is then worked out as follows, using the sample size (n), ΣC, ΣT_X and ΣT_Y.

$$\tau = \frac{2\Sigma C}{\sqrt{[n(n-1) - \Sigma T_X][n(n-1) - \Sigma T_Y]}}.$$

If any of the variables is free from tied scores, its ΣT is taken as zero. So, if both the variables have no tied scores, the formula simplifies into ;

$$\tau = \frac{2\Sigma C}{n(n-1)}.$$

6.4.3. Significance of computed tau

Because the tau is computed using a sample, the H_0 proposes that it is not significant and has resulted from the use of a particular sample drawn by random sampling. To find the probability P of this H_0 being correct, for sample sizes exceeding 10, the computed tau is transformed into Student's t using its SE (s_τ) and the obtained t is compared with two-tail critical t_α values.

$$s_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}}\,; \qquad t = \frac{\tau}{s_\tau}; \qquad df = \infty$$

The computed t is significant, only if it either exceeds or equals the critical t value for a chosen significance level ($P \leq \alpha$).

Example 6.4.1

Work out Kendall's tau to find whether or not there is a significant correlation ($\alpha = 0.05$) between gill weight (X) and body weight (Y), using the following data of a sample of crabs.

Animal :	1	2	3	4	5	6	7	8	9	10	11
X gm :	0.15	0.11	0.19	0.30	0.13	0.20	0.25	0.30	0.15	0.22	0.15
Y gm :	11.20	8.10	13.20	14.50	8.45	12.20	14.55	14.00	11.25	9.50	11.45

Critical t values : $t_{0.05(10)}=2.228$; $t_{0.05(20)}=2.086$; $t_{0.05(\infty)}=1.960$.

Solution :

Table 6.3. Ranking of gill weights (X) and body weights (Y).

X	Ranks (R_X)	Y	Ranks (R_Y)
0.15	4	11.20	4
0.11	1	8.10	1
019	6	13.20	8
0.30	10.5	14.50	10
0.13	2	8.45	2
0.20	7	12.20	7
0.25	9	14.55	11
0.30	10.5	14.00	9
0.15	4	11.25	5
0.22	8	9.50	3
0.15	4	11.45	6

R_X and R_Y ranks are given in ascending orders to X and Y scores, respectively, in Table 6.3. As Y has no tied set of scores, R_Y ranks are entered in ascending order along the first column of Table 6.4, and the respective paired R_X ranks are entered

against them in the second column. Every R_X rank is taken in turn as the pivotal rank and compared with the subsequent R_X ranks successively, counting each of the latter as +1, 0 or –1, according as the subsequent rank exceeds, equals or is lower than the pivotal rank. The algebraic sum of these counts gives ΣC.

Table 6.4. Count of subsequent ranks.

R_Y	R_X	Count of subsequent ranks	Total (C)
1	1	+1+1+1+1+1+1+1+1+1+1	+10
2	2	+1+1+1+1+1+1+1+1+1	+9
3	8	–1–1–1–1–1+1+1+1	–2
4	4	0+0+1+1+1+1+1	+5
5	4	0+1+1+1+1+1	+5
6	4	+1+1+1+1+1	+5
7	7	–1+1+1+1	+2
8	6	+1+1+1	+3
9	10.5	0–1	–1
10	10.5	–1	–1
11	9		
Total			+35 (ΣC)

$n = 11$.

$\Sigma T_X = \Sigma[t_X\,(t_X-1)] = 3(3-1)+2(2-1) = 8$, because X has one set of three tied scores and one set of two tied scores.

$\Sigma T_Y = \Sigma[t_Y(t_Y-1)] = 0$, as Y has no tied set.

$$\tau = \frac{2\Sigma C}{\sqrt{[n(n-1)-\Sigma T_X]\,[n(n-1)-\Sigma T_Y]}} = \frac{2\times 35}{\sqrt{[11(11-1)-8]\,[11(11-1)-0]}} = +0.66.$$

$$s_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}} = \sqrt{\frac{2(2\times 11+5)}{9\times 11(11-1)}} = 0.234. \qquad t = \frac{\tau}{s_\tau} = \frac{0.66}{0.234} = 2.821. \quad df = \infty.$$

$\alpha = 0.05$. Critical $t_{0.05(\infty)} = 1.960$.

As the computed t exceeds the critical $t_{0.05}$, P is lower than 0.05, So, P is too low. The H_O is rejected. There is a significant correlation between gill weight and body weight. ($P < 0.05$.)

6.5 Partial correlation

In any system, there exist innumerable variables, many of which are associated with each other. So, the product-moment r between any two given variables lacks precision because it has resulted only partly from the direct correlation between them

while its remaining part has arisen from the association of those two variables with one or more other variables in common. So, to get a more correct measure of the direct correlation between two given variables, the effects of one or more other variables affecting them in common should be eliminated or held constant. *Partial r* is the correlation coefficient between two variables, eliminating or partialling out one or more variables associated with both of them. For example, the r_{12} worked out between blood glucose (X_1) and blood insulin (X_2) in a sample of mammals may suffer from the drawback that it has been partly derived from the association of blood cortisol (X_3) with both X_1 and X_2 ; so, for a more correct correlation between X_1 and X_2, a partial correlation coefficient ($r_{12.3}$) has to be worked out between those two, partialling out the variable X_3.

6.5.1. Types of partial correlation

Partial *r* values range from –1.00 to +1.00 and may be either *positive* or *negative*. Partial *r* may be either *linear* or *nonlinear* according to respectively linear and nonlinear associations. Again, partial *r* may belong to different *orders* according to the number of variables partialled out or held constant during its computation. For example, partial $r_{12.3}$ between blood glucose (X_1) and blood insulin (X_2), partialling out blood cortisol (X_3), is a *first-order partial correlation* ; in contrast, $r_{12.34}$ between glomerular filtration rate (X_1) and glomerular blood pressure (X_2) partialling out two other variables, viz., plasma protein osmotic pressure (X_3) and Bowman capsular fluid pressure (X_4), is a *second-order partial correlation*. Our discussion will be confined here to first-order linear partial correlations only. For the first-order partial $r_{12.3}$ for example, those components of X_1 and X_2 scores as are correlated with X_3 are eliminated, correlating only such respective remaining components of X_1 and X_2 as are not associated with X_3.

6.5.2. Assumptions for partial linear correlation

For the computation of partial linear correlations, following assumptions should be justifiable. (i) All the variables involved should be *continuous measurement variables*. (ii) All those variables should have their scores distributed in *nearly normal distributions* in the population without significant bilateral asymmetry. (iii) There should exist significant *linear correlations* between variables of each pair. (iv) Each pair of scores of every case in the sample should occur *at random* in accordance with laws of probability.

6.5.3. Computation of first-order linear partial *r*

Linear partial *r* of any order is basically worked out from the product-moment *r* values of each pair of variables involved. For the first-order partial $r_{12.3}$ between variables X_1 and X_2 partialling out the variable X_3, the product-moment r_{12}, r_{13} and

r_{23} values between the respective variables of three relevant pairs have to be used. Thus,

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{[1-r_{13}^2]\,[1-r_{23}^2]}}.$$

Similarly, for the first-order partial r values between variables of other pairs, eliminating the respective third variables,

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{[1-r_{12}^2]\,[1-r_{23}^2]}}; \quad r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{[1-r_{12}^2]\,[1-r_{13}^2]}}.$$

In each case, the H_O proposes that the computed partial r is not significant and has resulted from the use of a particular sample drawn at random by laws of probability. To find the probability P of the H_O being correct, the computed partial r is transformed into Student's t, using the SE of the former. For example, for $r_{12.3}$,

$$s_{r_{12.3}} = \sqrt{\frac{1-r_{12.3}^2}{n-3}}; \qquad t = \frac{r_{12.3}}{s_{r_{12.3}}}; \qquad df = n-3.$$

The computed t is then compared with the appropriate critical t value having the same df. The partial r is considered significant, only if the computed t either exceeds or equals the critical t_α for the chosen level of significance ($P \leq \alpha$).

Example 6.5.1.

Find whether or not there is a significant linear partial correlation between O_2 consumption (X_1 ml/minute) and tracheal ventilation volume (X_2 ml/minute) partialling out atmospheric O_2 tension (X_3 mm Hg), using the following data of a sample of 53 locusts. ($\alpha = 0.05$).

$$r_{12} = +0.75\ ;\ r_{13} = +0.35\ ;\ r_{23} = +0.25.$$

$$t_{0.05(52)} = 2.007\ ;\ t_{0.05(51)} = 2.008\ ;\ t_{0.05(50)} = 2.009.$$

Solution :

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{[1-r_{13}^2]\,[1-r_{23}^2]}} = \frac{0.75 - 0.35 \times 0.25}{\sqrt{[1-0.35^2]\,[1-0.25^2]}} = +0.73.$$

$$s_{r_{12.3}} = \sqrt{\frac{1-r_{12.3}^2}{n-3}} = \sqrt{\frac{1-0.73^2}{53-3}} = 0.097. \qquad t = \frac{r_{12.3}}{s_{r_{12.3}}} = \frac{0.73}{0.097} = 7.526.$$

$df = n-3 = 53-3 = 50$. Critical $t_{0.05(50)} = 2.009$.

The computed t exceeds the critical $t_{0.05}$; so, P is too low and the H_O is rejected. Hence, there is a significant linear partial correlation between X_1 and X_2, partialling out X_3. ($P < 0.05$.)

6.6 Multiple correlation

Multiple correlation consists of the quantitative assessment of the magnitude and direction of correlation between a given variable and the combination or weighted sum of two or more other variables. The single variable being correlated is called the *criterion* while the variables, whose combination is being correlated, are called the *predictors.* An example of multiple correlation *with three variables*—one criterion and two predictors— is the *multiple linear correlation coefficient* $R_{1.23}$ between wing length (X_1) and the combination or weighted sum of trunk length (X_2) and body weight (X_3) of a sample of insects, Higher multiple correlations *with more than three variables* (e.g., $R_{1.234}$) may also be worked out with more than two predictors. Multiple correlations may be either *positive* or *negative* according to the direction of correlation between the criterion and the weighted sum of predictors, and may range from –1.00 to +1.00. Multiple correlations may again be either *linear* or *nonlinear* according to the linear or nonlinear form of association between the criterion and the combination of predictors.

The squared value of multiple correlation (e.g., $R^2_{1.23}$) is called the *coefficient of multiple determination* and serves as an estimate of that proportion of the total variance of criterion, which is dependent on the combined contribution of all the predictors. On the contrary, the *coefficient of multiple non-determination* (e.g., $K^2_{1.23}$) serves as a measure of that proportion of the total variance of criterion, which is not determined by the combined contribution of the predictors and is given by : $K^2_{1.23} = 1 - R^2_{1.23}$.

6.6.1. Assumptions for multiple linear correlations

Multiple correlations can be computed with any number of variables, only when there are significant correlations between the variables of each pair. So, the assumptions to be justifiable for multiple linear correlations are closely similar to those for product-moment r between the variables in pairs. Thus, (i) the criterion as also each predictor should be *continuous measurement variables* with no gap in their scales of scores, (ii) scores of each variable should have a *normal or near-normal distribution* in the population without much bilateral asymmetry, (iii) there should exist significant *linear correlations* between the variables of each pair, and (iv) each pair of scores of every case should occur in the sample in accordance to the *laws of probability.*

6.6.2. Computation of multiple linear correlation with three variables

Multiple linear correlations are basically worked out from the product-moment r values of each pair of variables involved. It cannot be computed if the product-moment r is not significant for any of the pairs of variables. For working out the multiple linear correlation coefficient ($R_{1.23}$) between a criterion X_1 and the combination of two predictors (X_2 and X_3), the *beta coefficients* (β_2 and β_3) are first calculated from the product moment r values (r_{12}, r_{13} and r_{23}) between the respective variables; β_2 and β_3 serve as measures of the proportions of variance of the criterion, associated with the variances of respective predictors (X_2 and X_3).

$$\beta_2 = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2};\quad \beta_3 = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2};\quad R_{1.23} = \sqrt{\beta_2 r_{12} + \beta_3 r_{13}};$$

$$s_{R_{1.23}} = \frac{1}{\sqrt{n-3}};\qquad t = \frac{R_{1.23}}{s_{R_{1.23}}};\qquad df = n-3$$

To find the probability (P) of the H_O being correct, the Student's t worked out from $R_{1.23}$ is compared with critical t_α score for the chosen significance level (α).The computed $R_{1.23}$ is considered significant, only if the computed t is either higher than or equal to the critical t_α value ($P \leq \alpha$).

Example 6.6.1

Use the data of Example 6.5.1 to find whether or not there is a significant multiple linear correlation between O_2 consumption (X_1) and the weighted sum of tracheal ventilation (X_2) and atmospheric O_2 tension (X_3). ($\alpha = 0.05$.)

$t_{0.05(49)}$=2.010 ; $t_{0.05(50)}$=2.009 ; $t_{0.05(51)}$=2.008 ; $t_{0.05(52)}$=2.007.

Solution :

r_{12}= +0.75 ; r_{13}= +0.35 ; r_{23}= +0.25 ; n=53.

$$\beta_2 = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} = \frac{0.75 - 0.35 \times 0.25}{1 - 0.25^2} = 0.71.$$

$$\beta_3 = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} = \frac{0.35 - 0.75 \times 0.25}{1 - 0.25^2} = 0.17.$$

$$R_{1.23} = \sqrt{\beta_2 r_{12} + \beta_3 r_{13}} = \sqrt{0.71 \times 0.75 + 0.17 \times 0.35} = 0.77.$$

$$s_{R_{1.23}} = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{53-3}} = 0.141.\qquad t = \frac{R_{1.23}}{s_{R_{1.23}}} = \frac{0.77}{0.141} = 5.461.$$

$df = n-3 = 53-3 = 50.$ $\alpha = 0.05.$ Critical $t_{0.05(50)}$=2.009.

As the computed t exceeds the critical $t_{0.05}$, P is too low ($P < 0.05$). The H_O is rejected and the computed $R_{1.23}$ is significant.

6.7 Regression

Some variables cannot be measured directly or easily, with sufficient precision, or without errors. For predicting the very likely score of such a variable in any given individual or case, a statistical method of prediction, called *regression*, may be applied. The latter depends on the already known or measured scores of one or more variables correlated significantly with the variable to be predicted. In any regression, the variable to be predicted is called the dependent variable or *criterion* for that regression ; the variable(s) whose known or measured score(s) may form the basis of the prediction, should be called the independent variable(s) or *predictor(s)*. In every regression, there is a single criterion ; but there may be one or more predictors in a regression.

6.7.1. Types of regression

Regression may be broadly classified into simple and multiple regressions depending on the *number of predictors* used. In all types of regression, scores of a single variable would be predicted; but where the known or measured score of a single predictor is used in working out the regression, the latter is called a *simple regression*, while the scores of more than one predictor are used in predicting the score of a criterion in a *multiple regression*. It should be understood that in predicting the score of a criterion in any individual, the predictor scores of the same individual must be used ; moreover, there must exist significant simple correlations between the scores of the criterion and those of each predictor. An example of simple regression is the regression of the blood insulin concentraion (X_1) in a patient on his/her blood sugar concentraion (X_2), X_1 and X_2 being respectively the criterion and the predictor. A multiple regression of the O_2 consumption (X_1) may be worked out in a locust on the combination of scores of its tracheal ventilation (X_2) and the O_2 tension (X_3) in its inspired air— X_1 is the criterion here while X_2 and X_3 are two predictors.

Regressions may again be classified into *linear and nonlinear regressions*, depending on whether there is respectively a linear association or a nonlinear association between the criterion and the predictor. For example, if the criterion is linearly correlated with the predictor, the scores of the former are predicted by working out an equation for a straight line, depending on the linear association between the two. On the contrary, if the criterion has a nonlinear correlation with the predictor, scores of the criterion have to be predicted in terms of a curved line like a sigmoid or hyperbolic or exponential curve, according to the form of their association.

6.7.2. Models of regression

You have to choose from three alternative models of regression in predicting the scores of a criterion, according to the *nature of the predictor* used for the regression. (i) *Model I or fixed model regression* is chosen while using *fixed experimental treatments* as predictors that are prevented from random variations by planned and well-controlled applications of their different levels on the subjects ; this model of regression expresses both the interdependence and the cause-and-effect relation between the criterion and the predictor. Thus, a model I simple regression consists of the regression of O_2 consumption in a sample of dragon flies on the pre-determined levels of the insecticide rotenone applied on them. (ii) *Model II or random model regression* is worked out when the independent variables or predictors are *uncontrolled random variables* free to vary at random ; this model expresses the interdependence, but not the cause-and-effect relation of the criterion and the predictors. Thus, you would work out a model II simple regression of O_2 consumption in a sample of dragon flies on the naturally occurring uncontrolled levels of a pesticide in the environment. (iii) *Model III or mixed model regression* consists of only multiple regressions when *both random variables and fixed experimental treatments* are used as predictors. An example is the multiple regression of tracheal ventilation volume on the combination of atmospheric nitric oxide levels and administered levels of ephedrine.

6.7.3. Properties of regression

(a) Regression is an expression of the dependent variable or criterion as a *function* of the independent variable or predictor.

(b) It follows that a regression can be worked out only when there is a *significant correlation* between the criterion and the predictor, and may be linear or nonlinear according to the linear or nonlinear nature of that correlation.

(c) Because regression is based on correlation, it resembles the latter in being effective only within the confines of the *specific stratum or section of the population* from which a sample was used, and other conditions that prevailed during its work-out, but not beyond their limits.

(d) Regression predicts only a *probable score* of the criterion on a given score of the predictor in a case so that the predicted score ($\hat{Y}$) may differ from the actual score (Y) of the criterion in that case by varying amounts.

(e) So, for any chosen predictor score (X), the actual criterion scores (Y) of different individuals having that X score lie *dispersed* around the regression-based predicted $\hat{y}$ score in the form of a distribution.

(f) The magnitude of deviations of actual criterion scores (Y) from the regression-predicted criterion score ($\hat{Y}$) can be estimated by working out the *SE of estimate* (s_{YX}) of the criterion on the predictor.

(g) Between a pair of variables correlated with one another, regression can be worked out in *two ways*, viz., a regression of variable X as criterion on variable Y as predictor, and another regression of variable Y as criterion on variable X as predictor. However, it is sensible to compute the regression of that variable of the pair as criterion, whose direct measurement is less precise or/and more complicated than that of the other variable, while the latter is used as the predictor.

(h) A *regression equation* is worked out using a statistic called the *regression coefficient* (e.g., b_{YX} for the regression of Y on X) which is a measure of the average rate of change of criterion scores (Y) with each unit change in the scores (X) of the predictor.

6.8 Simple linear regression

This is the regression for predicting the probable score of a criterion (say, Y) on the measured, known or given score of a *single predictor* (say, X) where the two variables are *linearly correlated.* It may belong to *either model I or model II,* depending on whether the predictor is a fixed experimental treatment free from random errors, or a classification variable suffering from random errors. The regression of the criterion Y on the predictor X and that of the criterion X on the predictor Y require the working out of *two* respective and separate regression equations. For linear regression, each *regression equation* is an equation for a straight line (*regression line*) expressing the scores of the criterion as the *linear function* of the scores of the predictor. The *slope* of the regression line is given by the *regression coefficient* which is the measure of the average rate of change of criterion scores for unit changes in predictor scores. The regression coefficient is used in working out the *y-intercept* of the regression line, i.e., its point of intersection with the ordinate scale for criterion scores.

6.8.1. Assumptions for simple linear regression

Following assumptions should be justifiable for working out a simple linear regression of a criterion (Y) on a single predictor (X). (i) The criterion as well as the predictor should be *continuous measurement variable,* scores of both being quantitatively measurable and occurring even in infinitely small fractions of units. (ii) Scores of each variable should form a *normal or near-normal distribution* in the population from which the sample has been drawn. (iii) There should exist a significant *linear corelation* between the criterion and the predictor. (iv) The actual score of criterion for each individual should occur in the sample obeying the *laws of probability,* thus ensuring the representative nature of the sample with respect to the population. (v) The actual criterion scores (Y) of a large number of cases, having an identical

predictor score (X), should form a *normal distribution* around the criterion score ($\hat{Y}$) worked out by regression on that predictor score. (vi) If a model I regression is intended, the predictor should be a *fixed experimental treatment* free from random errors; for a model II regression, the predictor should be a *random or classification variable* beyond the control of the investigator and suffering from random errors.

6.8.2. Properties of simple linear regression

(a) Simple linear regression is possible only if there is a *significant product-moment r* between the variables intended to be the criterion and the predictor.

(b) If the predicted criterion scores (say, $\hat{Y}$) are plotted against their respective predictor scores (say, X), the plotted points should lie scattered around a *straight regression line* of Y on X, obeying the *principle of least squares.*

(c) The *slope* of the regression line is given by the *regression coefficient* of the criterion on the predictor; e.g., b_{YX} for the regression of Y on X, and b_{XY} for the regression of X on Y.

(d) *Magnitude* of the regression coefficient is a measure of the average rate of change of criterion scores with unit changes in predictor scores. Positive and negative *algebraic signs* of the regression coefficient correspond respectively to positive and negative correlations between the two variables and indicate respectively ascending and descending gradients of the regression line.

(e) The *y-intercept* of the regression line determines the general level of the line and is given, in case of the regression of Y on X, by a_{YX} worked out using the b_{YX}.

(f) For any given predictor score (say, X), the actual criterion scores (Y) of a large number of cases having that X score lie scattered in the form of a *normal distribution* around the regression-predicted $\hat{Y}$ score as the mean of that distribution.

(g) The linear regression of Y on X, and that of X on Y would yield two *separate regression lines* inersecting at the point corresponding to means ($\overline{X}$ and $\overline{Y}$) of the two variables.

(h) The *angle* between the intersecting regression lines depends on the magnitude of the product-moment r value (r_{YX} or r_{XY}) between the two variables, ranging from zero (when r is +/–1.00) to 90^0 (when r is 0.00).

(i) The product-moment r between the criterion and the predictor (r_{YX} or r_{XY}) is the *geometric mean* of the regression coefficients (b_{YX} and b_{XY}) of the two regression lines : r_{YX} or $r_{XY} = \sqrt{b_{YX} b_{XY}}$.

(j) Deviations of actually measured criterion scores (Y) from the predicted criterion

score ($\hat{Y}$) may be estimated by the *SE of estimate* (s_{YX} for $\hat{Y}$ on X) using the *SD of* the criterion scores. Thus, for the respective regression equations,

$$s_{YX} = s_Y\sqrt{1-r_{YX}^2} \; ; \quad s_{XY} = s_X\sqrt{1-r_{XY}^2} \; .$$

6.8.3. Computation of simple linear regression

1. *Regression of Y on X :*

(i) Computation of b_{YX} *using sum of products :*

$$\bar{X} = \frac{\Sigma X}{n}; \quad \bar{Y} = \frac{\Sigma Y}{n}; \quad b_{YX} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2}.$$

(ii) Computation of b_{YX} *using raw scores :*

$$b_{YX} = \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}.$$

(iii) Computation of a_{YX} and *regression equation :*

$$a_{YX} = \bar{Y} - b_{YX}\bar{X}; \quad \hat{Y} = a_{YX} + b_{YX}X.$$

2. *Regression of X on Y :*

(i) Computation of b_{XY} *using sum of products :*

$$\bar{X} = \frac{\Sigma X}{n}; \quad \bar{Y} = \frac{\Sigma Y}{n}; \quad b_{XY} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(Y-\bar{Y})^2}.$$

(ii) Computation of b_{XY} *using raw scores :*

$$b_{XY} = \frac{n\Sigma XY - \Sigma X\,\Sigma Y}{n\Sigma Y^2 - (\Sigma Y)^2}.$$

(iii) Computation of a_{XY} and *regression equation :*

$$a_{XY} = \bar{X} - b_{XY}\bar{Y}; \quad \hat{X} = a_{XY} + b_{XY}Y.$$

3. *Drawing of the regression line :*

Several predictor scores are chosen from within their range in the sample, for computing the corresponding criterion scores with the help of the regression equation. The computed criterion scores are plotted against the respective predictor scores on a graph paper and the plotted points are used in drawing the regression line. (See Example 6.8.1.)

Example 6.8.1.

Work out the simple linear regression of O_2 consumption (Y ml/minute) on tracheal ventilation (X ml/minute) using the data of Example 6.3.1.

Solution :

1. *Computation of b_{YX} from sum of products :*

From the data of Example 6.3.1, the scores of tracheal ventilation (X) and O_2 consumption (Y) are entered in Table 6.5 for further treatment.

Table 6.5. Computation of sum of products and sum of squares.

	X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})^2$	$(X - \bar{X})(Y - \bar{Y})$
	66.0	3.3	– 12.5	– 0.7	156.25	+ 8.75
	89.1	4.9	+ 10.6	+ 0.9	112.36	+ 9.54
	72.0	3.5	– 6.5	– 0.5	42.25	+ 3.25
	87.5	4.7	+ 9.0	+ 0.7	81.00	+ 6.30
	75.2	3.7	– 3.3	– 0.3	10.89	+ 0.99
	78.2	4.0	– 0.3	0	0.09	0
	83.5	4.3	+ 5.0	+ 0.3	25.00	+ 1.50
	71.6	3.4	– 6.9	– 0.6	47.61	+ 4.14
	85.6	4.4	+ 7.1	+ 0.4	50.41	+ 2.84
	76.3	3.8	– 2.2	– 0.2	4.84	+ 0.44
Σ	785.0	40.0	—	—	530.70	+ 37.75

$$\bar{X} = \frac{\Sigma X}{n} = \frac{785.0}{10} = 78.5. \qquad \bar{Y} = \frac{\Sigma Y}{n} = \frac{40.0}{10} = 4.0.$$

$$b_{YX} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(X - \bar{X})^2} = \frac{37.75}{530.70} = 0.071.$$

2. *Computation of b_{YX} from raw scores (alternative method) :*

From the data of Example 6.3.1, X and Y scores are entered in Table 6.6 for further treatment.

Table 6.6. Computation of b_{YX} from raw scores.

	X	Y	X^2	XY
	66.0	3.3	4356.00	217.80
	89.1	4.9	7938.81	436.59
	72.0	3.5	5184.00	252.00
	87.5	4.7	7656.25	411.25
	75.2	3.7	5655.04	278.24
	78.2	4.0	6115.24	312.80
	83.5	4.3	6972.25	359.05
	71.6	3.4	5126.56	243.44
	85.6	4.4	7327.36	376.64
	76.3	3.8	5821.69	289.94
Σ	785.0	40.0	62153.20	3177.75

$$\bar{X} = \frac{\Sigma X}{n} = \frac{785.0}{10} = 78.5. \qquad \bar{Y} = \frac{\Sigma Y}{n} = \frac{40.0}{10} = 4.0.$$

$$b_{YX} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} = \frac{10 \times 3177.75 - 785.0 \times 40.0}{10 \times 62153.20 - 785.0^2} = 0.071.$$

3. *Computation of a_{YX} and regression equation :*

$$a_{YX} = \bar{Y} - b_{YX}\,\bar{X} = 4.0 - 0.071 \times 78.5 = -1.57.$$

$$\hat{Y} = a_{YX} + b_{YX} X, \quad or \quad \hat{Y} = -1.57 + 0.071X.$$

4. *Drawing of regression line :*

Four X scores are chosen at random from within the range of X in the data and used in computing the respective $\hat{Y}$ scores.

(i) Where $X = 70$, $\hat{Y} = a_{YX} + b_{YX} X = -1.57 + 0.071 \times 70 = 3.4$.

(ii) Where $X = 75$, $\hat{Y} = a_{YX} + b_{YX} X = -1.57 + 0.071 \times 75 = 3.8$.

(iii) Where $X = 80$, $\hat{Y} = a_{YX} + b_{YX} X = -1.57 + 0.071 \times 80 = 4.1$.

(iv) Where $X = 85$, $\hat{Y} = a_{YX} + b_{YX} X = -1.57 + 0.071 \times 85 = 4.5$.

Each $\hat{Y}$ score is plotted against the corresponding X score on a graph paper and the plotted points are used to draw the regression line of Y on X (Fig 6.1).

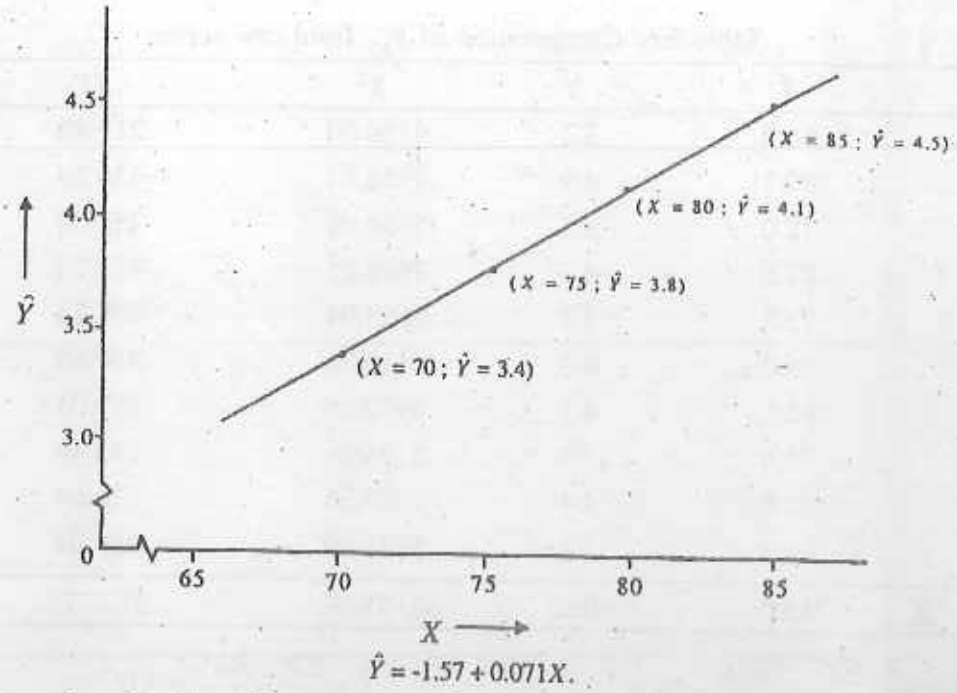


Fig 6.1. Linear regression line of O_2 consumption ($\hat{Y}$) on tracheal ventilation (X) for Example 6.8.1.

6.9 Multiple regression

Statistical prediction of the scores of a dependent variable or criterion, using the known or measured scores of two or more independent variables or predictors, is called *multiple regression.* It is basically the method of expressing the criterion as a function of the weighted sum of two or more predictors.

6.9.1. Types of multiple regression

Where the criterion has a significant linear correlation with each of the predictors, *multiple linear regression* is used in predicting its scores from the combination of scores of those predictors. On the contrary, *multiple nonlinear regressions* have to be used in predicting the scores of the criterion if the latter has significant hyperbolic, sigmoid, exponential or other nonlinear forms of association with the predictors.

Methods of multiple regressions also differ according to the *number* of independent variables/predictors used in the computation. Our discussion will be confined here to multiple linear regression with three variables, i.e., two predictors.

6.9.2. Models of multiple regression

Multiple regressions may be of three different models according to the nature of

predictors used. In a *model I or fixed model multiple regression*, all the predictors are fixed experimental treatments (e.g., levels of administered insulin) under the control of investigator and free from random errors; for example, the regression of blood glucose on combinations of different administered levels of insulin and cortisol. In a *model II or random model multiple regression*, all the predictors are beyond the control of investigator and liable to changes at random; for example, the regression of gill weights of fishes on their body weights and trunk lengths, both of the latter being random variables. A *model III or mixed model multiple regression* is worked out when both types of predictors are simultaneously applied in the experiment; for example, the regression of O_2 consumption by insects on atmospheric O_2 tension (random) and administered levels of dimercaprol (fixed).

6.9.3. Assumptions for multiple linear regression

Multiple linear regression can be worked out, only if the following conditions may be logically assumed to be fulfilled. (i) Predictors as well as criterion are continuous measurement variables without genuine gaps in their respective scales of values. (ii) Scores of every variable are almost normally distributed in the population without very distinct bilateral asymmetry. (iii) Actual score of the dependent variable for each case appears in the sample obeying the laws of probability. (iv) The criterion should have linear association with each of the predictors.

6.9.4. Computation of multiple linear regression

For working out the multiple linear regression of a criterion (X_1) on the combination of two predictors (X_2 and X_3), partial regression coefficients ($b_{12.3}$ and $b_{13.2}$) are computed using the beta coefficients (β_2 and β_3) and the standard deviations (s_1, s_2 and s_3) of all three variables. The beta coefficients have been described in Subsection 6.6.2. *Partial regression coefficients*, viz., $b_{12.3}$ and $b_{13.2}$, are the measures of slopes of regression lines of the criterion X_1 on the predictors X_2 and X_3, respectively, when the respective other predictors are eliminated or partialled out. The means of the variables and the partial regression coefficients are used in working out $a_{1.23}$ which is the *y-intercept* of the regression line. Where $\hat{X}_1$ is the predicted score of the criterion,

$$\beta_2 = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}; \qquad \beta_3 = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2};$$

$$b_{123} = \beta_2 \times \frac{s_1}{s_2}; \qquad b_{132} = \beta_3 \times \frac{s_1}{s_3}; \qquad a_{123} = \bar{X}_1 - b_{123}\bar{X}_2 - b_{132}\bar{X}_3;$$

$$\hat{X}_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3.$$

The *standard error of estimate* ($s_{1.23}$) serves as an estimate of the deviations of actual criterion scores (X_1) from the predicted $\hat{X}_1$ score, and is worked out using the *coefficient* of *multiple determination* which is the squared multiple linear correlation coefficient ($R^2_{1.23}$) between the criterion (X_1) and the combination of predictors (X_2 and X_3).

$$R^2_{1.23} = \beta_2 r_{12} + \beta_3 r_{13}; \qquad s_{1.23} = s_1\sqrt{1-R^2_{1.23}}.$$

Example 6.9.1.

Work out the multiple linear regression of glomerular filtration rate (X_1 ml/min) on glomerular blood pressure (X_2 mm Hg) and capsular fluid pressure (X_3 mm Hg), using the following data of a sample of 40 chimpanzees.

$\bar{X}_1$ = 120 ml/min; $\bar{X}_2$ = 58 mm Hg; $\bar{X}_3$=18 mm Hg;

s_1 = 21.5 ml; s_2 = 14.2 mm Hg; s_3= 3.5 mm Hg;

r_{12} = + 0.82; r_{13} = – 0.21; r_{23}= + 0.18.

Solution :

$$\beta_2 = \frac{r_{12} - r_{13}r_{23}}{1-r^2_{23}} = \frac{0.82-(-0.21)\times 0.18}{1-0.18^2} = 0.89.$$

$$\beta_3 = \frac{r_{13} - r_{12}r_{23}}{1-r^2_{23}} = \frac{-0.21-0.82\times 0.18}{1-0.18^2} = -0.37.$$

$$b_{12.3} = \beta_2 \times \frac{s_1}{s_2} = 0.89\times\frac{21.5}{14.2} = 1.35; \quad b_{13.2} = \beta_3 \times \frac{s_1}{s_3} = -0.37\times\frac{21.5}{3.5} = -2.27.$$

$$a_{1.23} = \bar{X}_1 - b_{12.3}\bar{X}_2 - b_{13.2}\bar{X}_3 = 120 - 1.35\times 58 - (-2.27)\times 18 = 82.56.$$

$$\hat{X}_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3, \text{ or } \hat{X}_1 = 82.56 + 1.35X_2 - 2.27X_3.$$

$$R^2_{1.23} = \beta_2 r_{12} + \beta_3 r_{13} = 0.89\times 0.82 - 0.37\times(-0.21) = 0.81$$

$$s_{1.23} = s_1\sqrt{1-R^2_{1.23}} = 21.5\sqrt{1-0.81} = 9.37.$$

6.10 Summary

Correlation determines quantitatively the strength and indicates the direction of relationship between two or more variables. It may be simple or multiple according as two or more variables are involved, linear or nonlinear according to the form of association between them, and positive or negative according as the correlated variables

change in the same or opposite directions. Correlation serves as a measure of association between the given variables within only that section of the population from which a sample has been used for its computation, and does not necessarily indicate any cause-effect relationship between the variables. Product-moment correlation coefficient measures the correlation between two continuous, normally distributed and linearly associated variables. Kendall's rank correlation coefficient is used for linear correlations between the ranks given to measurement and ordinal variables. Partial correlation is a measure of magnitude and direction of association between two given variables, eliminating the effects of one or more other variables correlated with them. Multiple correlation is the correlation between a single given variable and the combination of two or more other variables.

Regression is the statistical prediction of the probable score of a given variable called the criterion, on the basis of the measured scores of one or more other variables called the predictors, correlated with the criterion. Regression may be simple or multiple according as one or more predictors are used. Linear regression is worked out if the criterion is linearly correlated with the predictor or predictors. The criterion score predicted by regression is only a probable score and may differ from the actual criterion score. Simple linear regression involves the computation of a regression cofficient which is a measure of average rate of linear changes of criterion scores per unit change of predictor scores. Models of regression are determined by the natures of predictors, the latter being fixed experimental treatments in model I and random variables in model II. Multiple regression consists of the prediction of probable scores of a given criterion on the combination of two or more predictors. It is computed using partial regression coefficients, each of which is a measure of the rate of change of the criterion per unit change of one of the predictors when the effects of other predictors are partialled out.

6.11 Terminal questions

1. (a) Explain what you understand by simple regression and model I regression.
   (b) State the assumptions for simple linear regression.
   (c) Work out the simple linear regression of tracheal ventilation (Y ml) on O_2 consumption (X ml) using the following data of a sample of insects.

Animal :	1	2	3	4	5	6	7	8	9	10
X :	4.5	3.7	3.6	3.2	3.0	3.2	4.0	3.4	2.8	3.9
Y :	87	75	82	73	72	74	85	73	70	78

2. (a) State the assumptions for product-moment r.
   (b) Discuss the properties of product-moment correlation.

(c) Work out product-moment r using the data of Question 1(c) and find its significance, choosing the correct critical t value from below. ($\alpha = 0.05$.)

$t_{0.05(19)} = 2.093$; $t_{0.05(8)} = 2.306$; $t_{0.05(9)} = 2.262$.

3. (a) Explain what you mean by first-order and second-order partial correlations, citing examples.

(b) Discuss the assumptions for linear partial correlations.

(c) Find whether or not there is a significant linear partial correlation between gill weights (X_1 gm) and trunk lengths (X_2 cm), partialling out body weights (X_3 gm) in the following sample of 43 salmons using the r values between the respective variables and the correct critical t value from below. ($\alpha = 0.01$.)

$r_{12} = +0.55$; $r_{13} = +0.30$; $r_{23} = +0.28$.

$t_{0.01(42)} = 2.698$; $t_{0.01(41)} = 2.701$; $t_{0.01(40)} = 2.704$.

4. (a) Give a brief account of different models of regression.

(b) Describe the assumptions for multiple linear correlations.

(c) Use the data of question 3(c) to find whether or not there is a significant multiple linear correlation between gill weights (X_1) and the combination of trunk lengths (X_2) and body weights (X_3). ($\alpha = 0.01$.)

5. (a) Describe different types of regression.

(b) Discuss the properties of simple linear regression.

(c) Workout the multiple linear regression equation of variable X_1 on the combination of variables X_2 and X_3 in the following sample of 50 animals using their means, standard deviations and product-moment r values given below.

$\bar{X}_1 = 12.0$; $\bar{X}_2 = 8.5$; $\bar{X}_3 = 6.4$; $s_1 = 2.82$; $s_2 = 2.10$; $s_3 = 1.21$;

$r_{12} = +0.72$; $r_{13} = -0.21$; $r_{23} = -0.23$.

6. (a) Discuss where you would compute Kendall's tau, describing its assumptions.

(b) Describe the sources of inaccuracies in the computation of Kendall's tau and a method for countering one of them.

(c) Using the data of Question 1(c), work out Kendall's tau to find whether or not there is a significant correlation between O_2 consumption (X ml) and tracheal ventilation (Y ml), choosing the correct critical t value from below. ($\alpha = 0.01$).

$t_{0.01(10)}=3.169$; $t_{0.01(9)}=3.250$; $t_{0.01(20)}=2.845$; $t_{0.01(\infty)}=2.576$;

6.12 Answers

1. (a) See Sub-sections 6.7.1 and 6.7.2.
   (b) See Sub-section 6.8.1.
   (c) See Example 6.8.1.
2. (a) See Sub-section 6.3.1.
   (b) See Sub-section 6.3.2.
   (c) See Example 6.3.1.
3. (a) See Sub-section 6.5.1.
   (b) See Sub-section 6.5.2.
   (c) See Example 6.5.1.
4. (a) See Sub-section 6.7.2.
   (b) See Sub-section 6.6.1.
   (c) See Example 6.6.1.
5. (a) See Sub-section 6.7.1.
   (b) See Sub-section 6.8.2.
   (c) See Example 6.9.1.
6. (a) See Sub-section 6.4.1.
   (b) See Sub-section 6.4.2.
   (c) See Example 6.4.1.

Unit 7 ❑ PROBABILITY THEORY

Structure

7.1 Introduction

This unit will describe random occurrences of results of most scientific investigations. You will read about random events of diverse types in the discrete sample space of the experiment. Nature of independent events will be briefly presented. Probability will be defined and explained, and its theorems will be discussed in brief.

Objectives

After studying this unit, you should be able to do the following :

- define random experiments and their sample spaces,
- define and explain random events,
- classify and define different types of random events,
- explain what is meant by independent events,
- understand and define probability, and
- describe the theorems of probability.

7.2 Random experiments

Most experiments or scientific investigations can be repeated any number of times under the same conditions, but their results are neither predictable before the termination of the particular trial, nor identical in most of the cases inspite of apparently identical situations. Such experiments are called *random experiments*. Their basic

characteristics are as follows : (i) they may be *repeated any number of times* under the same situations and even with the same samples, (ii) all results or outcomes possible for such an experiment are known and together constitute its *sample space*, (iii) but the result is *not predictable* for any specific performance of the experiment so that (iv) *any of those predictable results* may be yielded by the experiment on any particular occasion, (v) repeated performances of the same experiment under apparently unchanged situations may consequently give results *mostly differing* from each other and coinciding only sometimes, and (vi) the sample space consists of elements called *sample points.*

If, say, you have to choose at random a single donor for a blood donation camp, the possible ABO blood group of a randomly chosen donor may be either A or B or O or AB group. This set of blood groups exhausts the possibility of choice of ABO blood groups. Such a set of possible outcomes, e.g., [A, B, C, D] in this case, is called the *sample space* for this experiment or sampling, and consists here of four possible elements or outcomes in a single choice. Each such element in a sample space is called a *sample point.*

In case one animal has to be chosen in a single chance from among animals of male (M) and female (F) sexes, the sample space for the single choice would consist of the set [M, F] with M and F as its sample points. In contrast, when choosing twice from those animals, each one (M or F) of the first choice may be succeeded by either of those two options in the second choice so that the chosen elements may constitute 2^2 or 4 sample points, viz., FF, FM, MF and MM, and this set of all four sample points forms the sample space for two successive choices.

A sample space, consisting of a set of all possible results or choices, may be *continuous* or *nondiscrete* if it consists of an infinite set of sample points that cannot be counted. On the contrary, a *discrete* or *discontinuous* sample space comprises either a finite set or an enumerable, though infinite, set of sample points.

7.3 Random events

An *event* consists of the occurrence, existence or happening of any case, incident, individual, phenomenon or score. It is called a *random event* if it either occurs or does not occur as an outcome of a random experiment or choice with discrete sample space. Any subject of a discrete sample space may constitute a random event in the relevant experiment. A single set of elements or a combination of such sets may occur as a random event. Occurrence of a random event should obey the *laws of probability*; moreover, in a random experiment or sampling, every event or individual should have an *equal probability of occurrence.*

Random events may be of various types according to the different manners of the probabilities of their occurrences. For example, an *elementary* or *simple event* is a single event of the sample space while a *compound event* comprises more than one element or sample point in the sample space. It should be understood that individual simple events as well as the entire sample space are considered as events. If the occurrence of either of two events is prevented by the occurrence of the other, they are known as *mutually exclusive events;* thus, A and B would be mutually exclusive of each other if the probability (P) of their simultaneous occurrence is zero, i.e., $P(A.B) = 0$, so that A and B would never occur together. For an event occurring at random in a random experiment, its *complementary event* is its non-occurrence in the latter. An *elementary event* is a single element of the sample space while a *compound event* consists of a collection of more than one single sample point or element in the sample space. If the occurrence of each of two events influences and changes the probability of occurrence of the other, they are called *mutually dependent events;* on the contrary, two events are *mutually independent* if the occurrence of each does not affect or alter the probability of occurrence of the other. If neither of two events has any probability of occurring in preference to the other so that either of them may precede the other in occurring in the outcome of a random experiment, they are called *equally likely events.* Where the outcome of a random experiment consists of a collection of a given number of random events, that collection is *exhaustive* if one or more or those events must occur in the result of the experiment. A *sure event* of a random experiment is the whole discrete sample space occurring with certainty in that experiment and has, therefore, a probability of 1.00 (i.e., 100% chance) of occurrence, while its empty subset with zero probability of occurrence is the *impossible or null event.*

7.4 Independent events

Two random events of a random experiment are *mutually independent* of one another if the occurrence of either of them does not affect or alter the probability of occurrence of the other. If A and B are independent events, the probability of occurrence of neither of them (A or B) would be enhanced or decreased, whether or not it is preceded or accompanied by the occurrence of the other. Thus, each of two independent events follows its own probability of occurrence in turning up in the experiment. Moreover, each such event can occur alone and also together or preceding or following the other one. The probability $P(A.B)$ of the simultaneous or successive occurrences of those two events is given by the multiplication theorem of probability. (See Subsection 7.6.2.)

$$P(A.B) = P(A) \times P(B).$$

Thus, the probability of simultaneous or successive occurrences of k number of independent events is given by :

$$P(A.\ B.\ C.....k) = P(A) \times P(B) \times P(C) \times .....\times P(k).$$

7.5 Probability

You have read in Sub-section 3.4.1. of Unit 3 about the *relative frequency* (f/n) of any event or score in a sample, and its computation by dividing the frequency (f) of that event or score with the total number (n) of the scores or events in the sample. In a simple manner, the relative frequency of a particular type of event or score in a very vast or nearly infinite total number of events or scores, may be taken as the *probability* (P) of that particular event or score. In other words, probability is the measure of random occurrence of an event among an infinite total number of events. You may thus understand that the total number of all events, forming the *entire sample space* and constituting the *sure event*, has a 100% certainty or a probability of 1.00 of occurrence; so, the probability of occurrence of every other event or element, being a part of the entire sample space, would amount to a specific fractional part of 1.00 while a *null event* with no probability of occurrence would have a zero probability. Thus, for any event other than the sure event, the probability of random occurrence would either exceed or equal zero : $P \geq 0$. This makes the scale of probability or relative frequency a continuous one in any probability distribution (see Sub-section 3.4.1). For example, if 8769 pomfret fishes of trunk length of 8.3 cm occur in a random sample of 57421 pomfret fishes, the probability of random occurrence of the given trunk length of 8.3 cm may be worked out as follows.

$$P = \frac{f}{n} = \frac{8769}{57421} = 0.15.$$

However, the relative frequency of an event may differ from its probability of random occurrence, if the total number of all events or trials is small. With the increase in the total number of events or cases, the relative frequency gets progressively closer to the probability. *Probability* (P) is, therefore, the *limiting relative frequency* of the event under consideration, when the sample size approaches or reaches infinity. For exmple, in choosing one animal from a population with a male : female ratio of 0.55 : 0.45, the possible alternatives are two only, viz., either a male or a female, so that the relative frequency of a male animal being chosen amounts to 1.00 in the first case and 0.00 in the second; but in choosing two animals either simultaneously or successively from the same population, there would be three alternative outcomes,

viz., one male and one female, or two males, or two females, so that the relative frequency of males varies from 0.50 in the first case, to 1.00 in the second case, and to 0.00 in the third case. This way the relative frequency of occurrence of males goes on changing with the number of choices, until choices are made an infinitely large number of times when the relative frequency would approach or reach the probability (P) of 0.55 expected from the population sex ratio.

7.6 Theorems of probability

You will next read about two basic theorems of probability.

7.6.1. Addition theorem

If there are k number of *mutually exclusive events* to choose from in one chance, so that the occurrence or choice of any one of them excludes the probability of occurrence of the others and more than one of them cannot occur simultaneously, then the addition theorem gives the probability of occurrence of any one event (e.g., A or B or C .....or k) out of the total number of exclusive events, as the sum obtained by adding the respective probabilities of their individual occurrences.

P(A or B or C or ..... k) = P(A) + P(B) + P(C) + ...+ $P(k)$.

For example, if a single rat has to be chosen or to occur at random in one chance out of a total of 15 rats (k = 15), the probability P(C) of random occurrence or choice of rat C separately and individually in one chance is given by $1/k$ and amounts to 1/15 or 0.0667; probabilities of similar individual occurrences of rats D and E would also amount to 1/15 or 0.0667 each. Because they are mutually exclusive, the occurrence of any of them in a single trial would prevent the chances of occurrences of the other two. In such a case, according to the addition theorem, the probability of either C or D or E occurring or getting chosen in one chance equals the sum of probabilities of random, separate and individual occurrence of each of those three.

P(C or D or E) = P(C) + P(D) + P(E) = 0.0667 + 0.0667 + 0.0667 = 0.20.

7.6.2. Multiplication theorem

If there are k number of *independent events* (A, B, C, ...k) to choose from, so that the choice or occurrence of any of the events does not affect or alter the probability of simultaneous or successive random occurrence of any other, the multiplication theorem then gives the probability of simultaneous or successive occurrences of a number of those independent events as the product of the respective individual probabilities of random occurrences of the latter. Thus, the multiplication theorem states that the probability of combined occurrences, simultaneous or successive, of

more than one independent event is the product of the probabilities of their individual occurrences. That way, the combined probability of occurrence of all the k number of independent events, either at the same time or after each other, would be given by :

$$P(\text{A. B. C.....}k) = P(\text{A}) \times P(\text{B}) \times P(\text{C}) \times \text{.....} \times P(k).$$

For exmple, if rats A, C and D have to be chosen successively at random out of a total of 15 rats ($k = 15$), the probability of separate individual choice of each of the three is given by $1/k$ which amounts to 1/15 or 0.0667. If the chosen rat is replaced in the group before the next choice to keep the total number k of the rats unchanged, the probability of choice of successive rats would remain unaffected to maintain their independent status as events. In such a case, the combined probability of successive choices of A, C and D would be given by :

$$P(\text{A.C.D}) = P(\text{A}) \times P(\text{C}) \times P(\text{D}) = 0.0667 \times 0.0667 \times 0.0667 = 0.0003.$$

7.7 Summary

Random experiments may be repeated any number of times, but their precise outcome is not predictable. A set of possible outcomes of an experiment constitutes a sample space and each element in that set is a sample point. Events may consist of an individual single element as well as the entire sample space. Random events occur at random under the laws of probability and belong to various types. Mutually independent events do not affect the probabilities of each other, while mutually exclusive events prevent the occurrences of each other so that they never occur together. Either of two equally likely events can precede the other in occurring.

Probability of an event is its limiting relative frequency when the total frequency of all events under consideration approaches infinity. The probability of occurrence of any one of several alternative exclusive events out of a total number of events is given by the addition theorem of probability as the sum of probabilities of individual occurrences of those alternative events. The multiplication theorem gives the probability of combined occurrence of a number of independent events as the product of the probabilities of their separate individual occurrences.

7.8 Terminal questions

1. (a) Give a brief account of random experiments, describing their basic characteristics.

   (b) Describe the addition theorem of probability and its application.

(c) Explain what are sample spaces. Define continuous and discrete sample spaces.

2. (a) Discuss random events and their different types.
   (b) Write briefly about independent events and the probability of their simultaneous or successive occurrences.
   (c) Describe the probability of random occurrence of any one of a number of mutually exclusive events with an example.
3. (a) Give a brief account of the concept of probability.
   (b) Discuss the addition and multiplication theorems of probability.
   (c) What do you understand by mutually exclusive and mutually independent events?

7.9 Answers

1. (a) See Section 7.2 .
   (b) See Sub-section 7.6.1.
   (c) See Section 7.2.
2. (a) See Section 7.3.
   (b) See Section 7.4 and Sub-section 7.6.2.
   (c) See Sub-section 7.6.1.
3. (a) See Section 7.5.
   (b) See Sub-sections 7.6.1 and 7.6.2.
   (c) See Sections 7.3, 7.4 and 7.6.

❑ APPENDIX ❑

Table 1. Some fractional areas of unit normal curve from its mean to different z scores.

z score :	1.94	1.95	1.96	1.97	1.98	1.99	2.00	2.01	2.02	2.03	2.04
Area :	.4738	.4744	.4750	.4756	.4761	.4767	.4772	.4778	.4783	.4788	.4793
z score :	2.05	2.06	2.07	2.08	2.09	2.10	2.51	2.52	2.53	2.54	2.55
Area :	.4798	.4803	.4808	.4812	.4817	.4821	.4940	.4941	.4943	.4945	.4946
z score :	2.56	2.57	2.58	2.59	2.60	2.96	2.97	2.98	2.99	3.00	4.00
Area :	.4948	.4949	.4951	.4952	.4953	.4985	.4985	.4986	.4986	.4987	.49997

Table 2. Some Two-tail Critical *t* values.

df :	11	12	13	14	15	16	17	18	19
α 0.05 :	2.201	2.179	2.160	2.145	2.131	2.120	2.110	2.101	2.093
α 0.01 :	3.106	3.055	3.012	2.977	2.947	2.921	2.898	2.878	2.861
df :	20	21	22	23	24	25	26	27	28
α 0.05 :	2.086	2.080	2.074	2.069	2.064	2.060	2.056	2.052	2.048
α 0.01 :	2.845	2.831	2.819	2.807	2.797	2.787	2.779	2.771	2.763
df :	29	30	40	∞					
α 0.05 :	2.045	2.042	2.021	1.960					
α 0.01 :	2.756	2.750	2.704	2.576					

Table 3. Some Critical chi square values.

df :	1	2	3	4	5	6	7	8	9	10
α 0.05 :	3.84	5.99	7.82	9.49	11.07	12.59	14.07	15.51	16.92	18.31
α 0.01 :	6.64	9.21	11.34	13.28	15.09	16.81	18.48	20.09	21.67	23.21

Table 4. Some Critical *F* values.

df :	1,8	1,9	1,10	1,11	1,12	1,13	1,14	1,15	1,16	1,17	1,18
α 0.05 :	5.32	5.12	4.96	4.84	4.75	4.67	4.60	4.54	4.49	4.45	4.41
α 0.01 :	11.26	10.56	10.04	9.65	9.33	9.07	8.86	8.68	8.53	8.40	8.28
df :	2,8	2,9	2,10	2,11	2,12	2,13	2,14	2,15	2,16	2,17	2,18
α 0.05 :	4.46	4.26	4.10	3.98	3.88	3.80	3.74	3.68	3.63	3.59	3.55
α 0.01 :	8.65	8.02	7.56	7.20	6.93	6.70	6.51	6.36	6.23	6.11	6.01

Acknowledgements :

Values in the above-mentioned tables have been quoted from the following publications. Table 1 : J. E. Wert, *Educational Statistics,* McGraw-Hill Book Co. Tables 2 and 3 : R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research,* Longman Group Ltd. Table 4 : G. W. Snedecor and W. G. Cochran, *Statistical Methods,* Iowa State University Press.

Group A(II)

Biotechnology

Unit 1 ❐ Principles of Assay of DNA, RNA, Western, Southern, Northern Blotting

Structure

1.1 Introduction

Once a specific DNA sequence has been isolated by cloning, the cloned DNA can be used as a probe to detect the presence and in some cases amounts, of complementary nucleic acids in complex mixtures such as total cellular DNA or RNA. These procedures depend on the exquisite specificity of nucleic acid hybridization. Related methods are used to locate DNA regions encoding specific mRNAs and transcription start sites.

1.2 Southern blotting

1.2.1 Technique of sourthern blotting

The technique of *Southern blotting* named after its originator Edward Southern, can identify specific restriction fragments in a complex mixture of restriction fragments. The DNA to be analyzed, such as the total DNA of an organism, is digested to completion with a restriction enzyme. For an organism with a complex genome, this digestion may generate millions of specific restriction fragments. The complex mixture of fragments is subjected to Gel Electrophoresis to separate the fragments according to size.

Even though all the fragments are not resolved by gel electrophoresis, an individual fragment that is complementary to a specific DNA clone can be detected. The restriction fragments present in the gel are denatured with alkali and transferred onto a nitrocellulose filter or nylon membrane by blotting.

For this purpose, gel is laid on top of a buffer saturated filter paper, placed on a solid support (eg., glass plate) with its two edges immersed in the buffer. A sheet of nitrocellulose membrane is placed on the top of the gel and a stack of many

papers (paper towels) on to of this membrane. A weight of about 0.5 Kg is placed on top of the paper towels. The buffer solution is drawn up by filter paper wick, and passes through the gel to the nitrocellulose membrane and finally to the paper towels. While passing through the gel, the buffer carries with it single stranded DNA, which binds on to the nitrocellulose membrane, when the buffer passes through it to the paper towels. After leaving this arrangement for a few hours or overnight, paper towels are removed and discarded. The nitrocellulose membrane with single stranded DNA bands blotted onto it, is baked at 80°C for 2-3 hours to fix the DNA permanently on the membrane. This procedure preserves the distribution of the fragments in the gel, creating a replica of the gel on the filter, The filter then is incubated under hybridization conditions with a specific radio labeled DNA probe usually generated from a cloned restriction fragment. The DNA restriction fragment that is complementary to the probe hybridizes, and its location on the filter can be revealed by autoradiography.

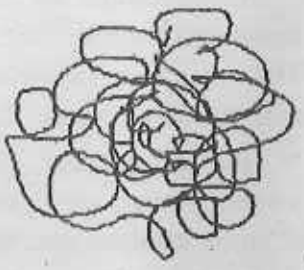
1.2.2 Significance

- ★ Southern blotting permits a comparison between the restriction map of DNA isolated directly from an organism and the restriction map of cloned DNA.
- ★ Southern blotting also is used to map restriction sites in genomic DNA next to the sequence of a cloned DNA fragment. This provides a rapid method of comparing the restriction maps of different individual organisms in the region surrounding a cloned fragment.
- ★ Deletion and insertion mutations are readily detected, as well as sequence differences in specific restriction sites.

1.2.3 Schematic representation of *southern blot*

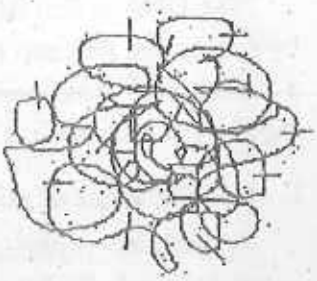
The Southern blot is used to detect the presence of a particular bit of DNA in a sample. The DNA detected can be a single gene, or it can be part of a larger piece of DNA such as a viral genome.

DNA is extracted from the cells and purified



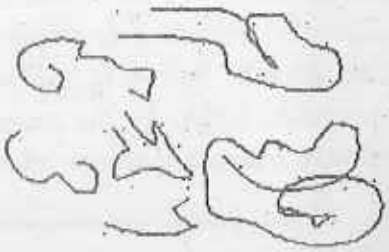
A. Restriction digest with restriction endonuclease

DNA is restricted (cut) with enzymes.



B. DNA fragments

In this example, a large piece of DNA is chopped into smaller pieces using a restriction enzyme.



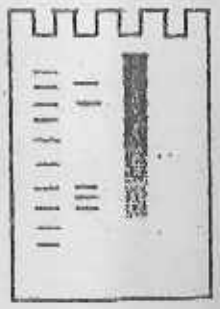
C. Loading the gel

The DNA is loaded into a well of the gel matrix.

- Lane 1 contains size standards (a mix of known DNA fragments)
- Lane 2 contains the restricted DNA
- Lane 3 contains unrestricted (whole) DNA

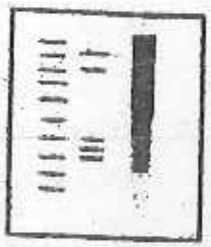
D. Running the gel

An electric current is passed through the gel and the DNA moves away from the negative electrode. The distance moved depends on the size of the DNA fragment. Standards are used to quantitate the size. Unrestricted, large DNA runs as a smear due to random shearing (breaking) of the DNA. The DNA can be visualized by staining first with a fluorescent dye and then lighting with UV.



E. Transfer of the DNA to a membrane

The DNA is first denatured (made single stranded–usually by raising the pH) and then transferred out of the gel and onto a membrane. The transfer can be done electrically or by capillary action with a high salt solution.



F. Development of the blot

A labelled probe specific for the gene in question is incubated with the blot. The blot is washed to remove non-specifically bound probe and then a development step allows visualization of the DNA that is bound. See Southern Up Close for a detailed description of this process. In this example the gene is found in the second largest fragment of the restricted DNA. In the unrestricted DNA it migrates more slowly because it is part of a larger molecule.

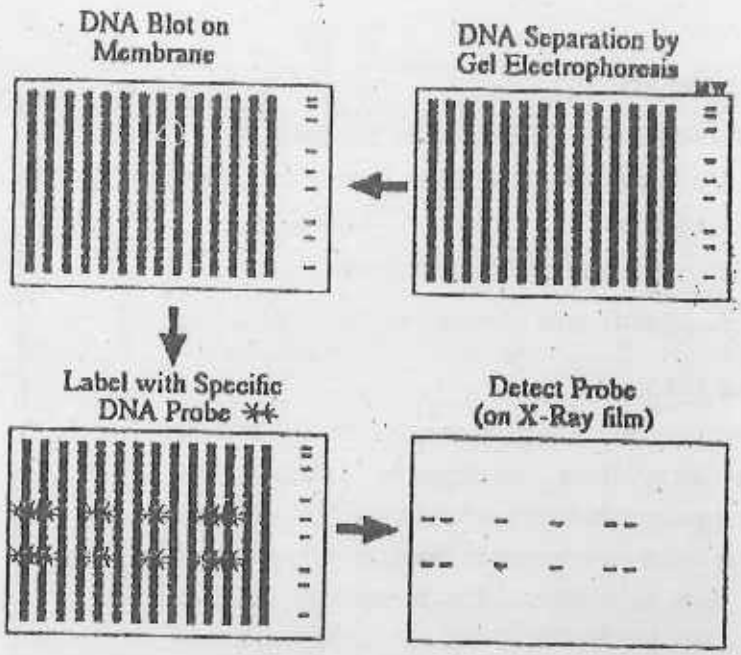


Fig. 1.1 : Schematic representation of southern blotting

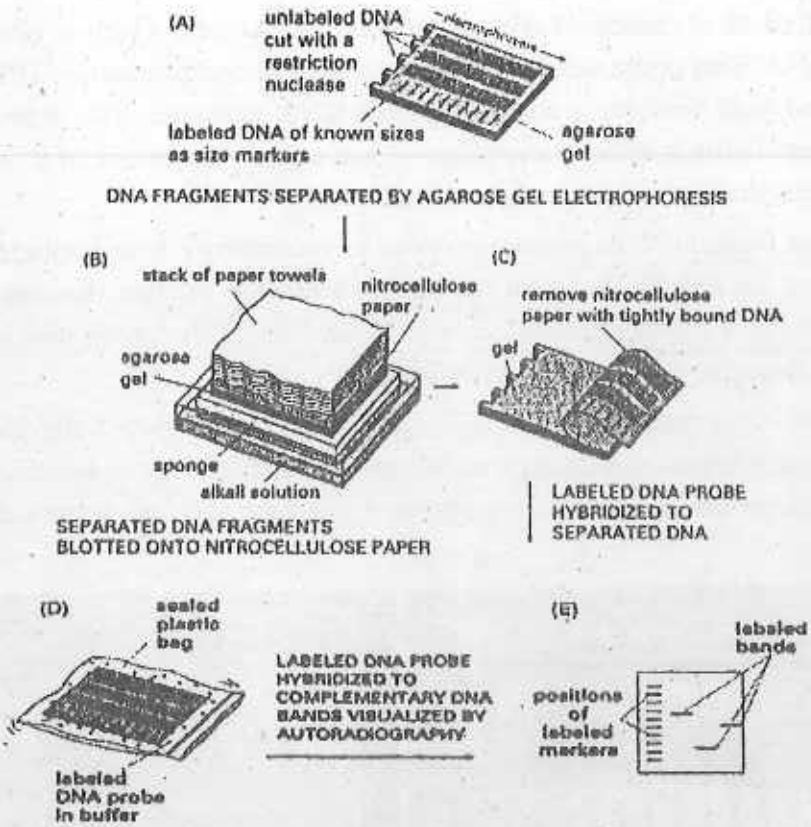


Fig. 1.2 : Process involved in sourthern blotting

1.2.4 Protocol

This is a brief overview protocol of how a southern blot (more formally called an DNA blot) is performed and what type of data you can obtain from one.

1) DNA (genomic or other source) is digested with a restriction enzyme and separated by gel electrophoresis, usually an agarose gel. Because there are so many different restriction fragments on the gel, it usually appears as a smear rather than discrete bands. The DNA is denatured into single strands by incubation with NaOH.
2) The DNA is transferred to a membrane which is a sheet of special blotting paper. The DNA fragements retain the same pattern of separation they had on the gel.

3) The blot is incubated with many copies of a probe which is single-stranded DNA. This probe will form base pairs with its complementary DNA sequence and bind to form a double-stranded DNA molecule. The probe cannot be seen but it is either radioactive or has an enzyme bound to it (e.g. alkaline phosphatase or horseradish peroxidase).

4) The location of the probe is revealed by incubating it with a colorless substrate that the attached enzyme converts to a colored product that can be seen or gives off light which will expose X-ray film. If the probe was labeled with radioactivity, it can expose X-ray film directly.

Below is an example of a real Southern blot used to detect the presence of a gene that was transformed into a mixed cell population. In this Southern blot, it is easy to determine which cells incorporated the gene and which ones did not.

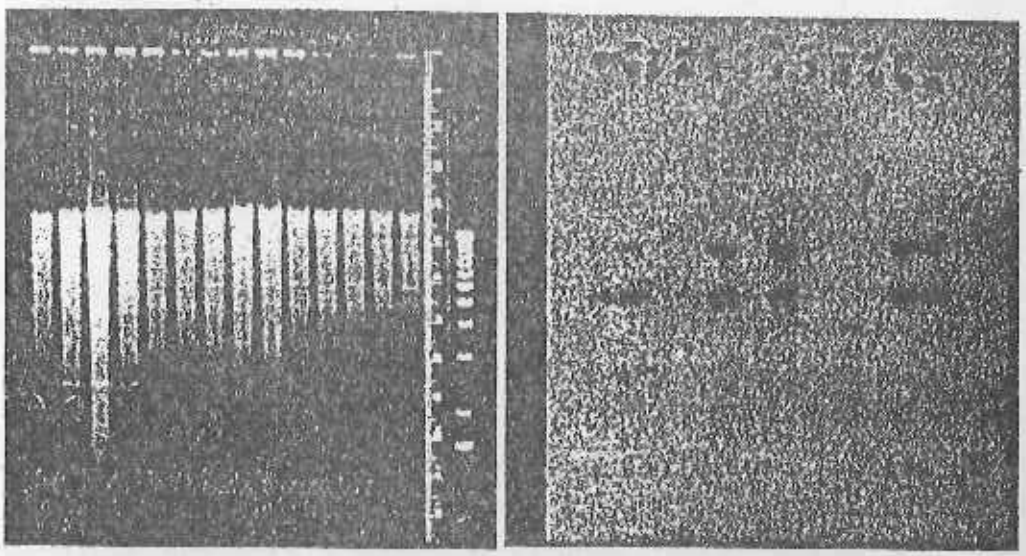


Fig. 1.3 : The figure on the left shows a photograph of a 0.7% agarose gel that has 14 different samples loaded on it (plus molecular weight marker in the far right lane and a glowing ruler used for analysis of the results). Each sample of DNA has been digested with the same restriction enzyme (EcoRI). Notice that the DNA does not appear as a series of discrete bands but rather as a smear. The DNA was transferred to nitrocellulose and then probed with a radioactive fragment of DNA that was derived from the transformed gene. The figure on the right is a copy of the X-ray film and reveals which strains contain the target DNA and which ones do not.

1.3 Northern blotting

1.3.1 Techinque of northern blotting

Initially the technique of Southern blotting used for DNA transfer from gel to the membrane, could not be used for blot-transfer of RNA. Instead mRNA bands from the gel were blot-transferred into a chemically reactive paper, prepared by diazotization of aminobenzyloxymethyl paper. The technique being related to Southern blotting was called *Northern blotting* and is used to detect a particular RNA in a mixture of RNAs. A RNA sample, often a total cellular RNA, is denatured by treatment with an agent (eg., formaldehyde) that prevents hydrogen bonding between base pairs, ensuring that all the RNA molecules have an unfolded, linear conformation. The individual RNAs are then separated according to the size by gel electrophoresis and transferred to a nitrocellulose filter to which the extended denatured RNAs adhere. The filter then is exposed to a labeled DNA probe and subjected to autoradiography. Because the amount of a specific RNA in a sample can be estimated from a Northern blot, the procedure is widely used to compare the amounts of a particular mRNA in cells under different conditions.

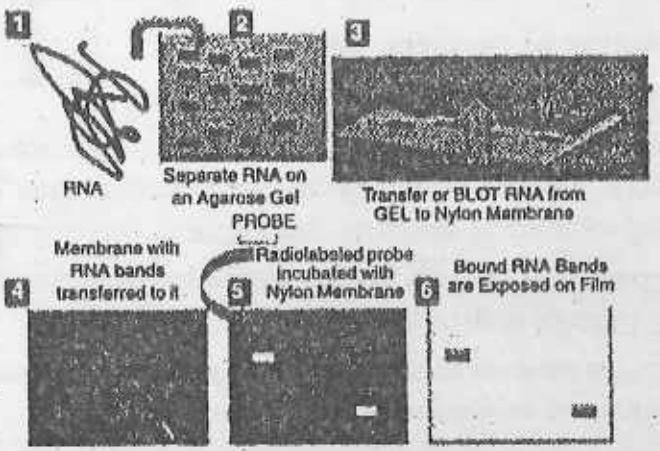


Fig. 1.4 : Steps involved in northern blotting

Similarly to Southern blotting, the hybridization probes may be DNA or RNA in northern blotting.

A variant of the procedure known as the reverse northern blot was occasionally (although, infrequently) used. In this procedure, the substrate nucleic acid (that is

affixed to the membrane) is a collection of isolated DNA fragments, and the probe is RNA extracted from a tissue and radioactively labelled.

1.3.2 Applications of the northern blot

Northern blots have been superceded in most areas by Real Time PCR and microarray approaches. It is not often used for clinical or diagnostic purposes.

The northern blot protocol and its variations are used however in molecular biology research to :

- ★ a gold-standard for the direct study of gene expression at the level of mRNA (messenger RNA transcripts).
- ★ detection of mRNA transcript size.
- ★ study RNA degradation.
- ★ study RNA splicing–can detect alternatively spliced transcripts.
- ★ study RNA half-life.
- ★ study IRES (internal ribosomal entry site)–to remove possibility of RNA digestion vs 2nd cistron translation.
- ★ often used to confirm and check transgenic / knockout mice (animals).

1.3.3 Disadvantages of northern blotting

The disadvantages of using northern blotting include :

- ★ Often radioactivity is used. This prevents ease of performing it, use and disposal. New methods of non-radioactive detection have been generated allowing non-radioactive detection. See Pierce.
- ★ The whole process of northern blotting takes a long time usually, from sample preparation through to detection.
- ★ If RNA samples are even slightly degraded by RNases, the quality of the data and quantitation of expression is quite negatively affected.
- ★ The standard northern blot method is relatively less sensitive than nuclease protection assays and RT-PCR. The sensitivity of northern blots may be increased with the use of nylon positively-charged membranes, use of a highly specific antisense probe.
- ★ Detection with multiple probes is a problem. Often, the membranes must be stripped before hybridization and detection with a second probe. This is a

problem as harsh conditions are required to strip off probes from the blot and is also time consuming. Also, there is a limit to the amount of times a blot may be stripped.

1.3.4 Advantages of northern blotting

The advantages of northern blots include :

- ★ It is a widely accepted and well regarded method
- ★ northern blotting is a straight-forward method
- ★ Often it is used as a confirmation or check
- ★ Often a gold-standard
- ★ it is a versatile protocol as it can allow the usage of many types of probes (vs Real time PCR) including: radiolabeled and non-radiolabeled, *in vitro* transcribed RNA and even oligonucleotides such as primers.
- ★ Sequences with even partial homology, unlike real time PCR or other methods can be used as hybridization probes (i.e sequence from different species for homology analysis, or even genomic fragments can be used).

1.4 Western blotting

The **western blot** (alternately, **immunoblot**) is a method to detect a specific protein in a given sample of tissue homogenate or extract. It uses gel electrophoresis to separate native or denatured proteins by the length of the polypeptide (denaturing conditions) or by the 3-D structure of the protein (native/ non-denaturing conditions). The proteins are then transferred to a membrane (typically nitrocellulose or PVDF), where they are probed (detected) using antibodies specific to the target protein. This method is used in the fields of molecular biology, biochemistry, immunogenetics and other molecular biology disciplines.The method originated from the laboratory of George Stark at Stanford. The name western blot was given to the technique by W. Neal Burnette.

1.4.1 Steps in a western blot

A. Tissue preparation

Samples may be taken from whole tissue or from cell culture. In most cases, solid tissues are first broken down mechanically using a blender (for larger sample volumes), using a homogenizer (smaller volumes), or by sonication. Cells may also

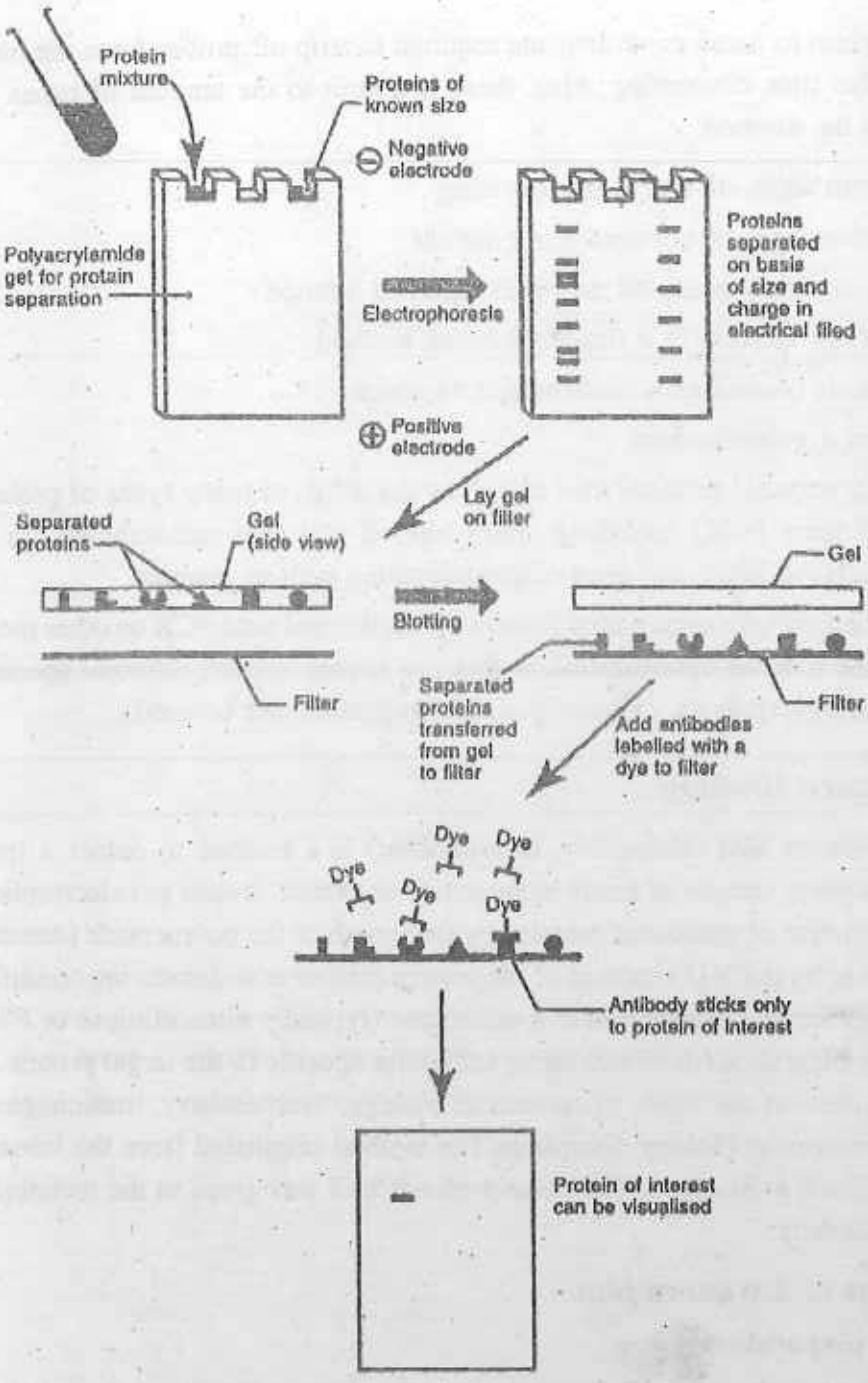


Fig. 1.5 : Gel-electrophoresis proeuct for western blotting

be broken open by one of the above mechanical methods. However, it should be noted that bacteria, virus or environmental samples can be the source of protein and thus western blotting is not restricted to cellular studies only. Assorted detergents, salts, and buffers may be employed to encourage lysis of cells and to solubilize proteins. Protease and phosphatase inhibitors are often added to prevent the digestion of the sample by its own enzymes. A combination of biochemical and mechanical techniques–including various types of filtration and centrifugation–can be used to separate different cell compartments and organelles.

B. Gel electrophoresis

The proteins of the sample are separated using gel electrophoresis. Separation of proteins may be by isoelectric point (pI), molecular weight, electric charge, or a combination of these factors. The nature of the separation depends on the treatment of the sample and the nature of the gel. By far the most common type of gel electrophoresis employs polyacrylamide gels and buffers loaded with sodium dodecyl sulfate (SDS). SDS-PAGE (SDS polyacrylamide gel electrophoresis) maintains polypeptides in a denatured state once they have been treated with strong reducing agents to remove secondary and tertiary structure (e.g. S-S disulfide bonds to SH and SH) and thus allows separation of proteins by their molecular weight. Sampled proteins become covered in the negatively charged SDS and move to the positively charged electrode through the acrylamide mesh of the gel. Smaller proteins migrate faster through this mesh and the proteins are thus separated according to size (usually measured in kilo Daltons, kD). The concentration of acrylamide determines the resolution of the gel–the greater the acrylamide concentration the better the resolution of lower molecular weight proteins. The lower the acrylamide concentration the better the resolution of higher molecular weight proteins. Proteins travel only in one dimension along the gel for most blots. Samples are loaded into wells in the gel. One lane is usually reserved for a marker or ladder, a commercially available mixture of proteins having defined molecular weights, typically stained so as to form visible, coloured bands. It is also possible to use a two-dimensional (2-D) gel which spreads the proteins from a single sample out in two dimensions. Proteins are separated according to isoelectric point (pH at which they have neutral net charge) in the first dimension, and according to their molecular weight in the second dimension.

C. Transfer

In order to make the proteins accessible to antibody detection, they are moved

from within the gel onto a membrane made of nitrocellulose or PVDF. The membrane is placed on top of the gel, and a stack of tissue papers placed on top of that. The entire stack is placed in a buffer solution which moves up the paper by capillary action, bringing the proteins with it. Another method for transferring the proteins is called electroblotting and uses an electric current to pull proteins from the gel into the PVDF or nitrocellulose membrane. The proteins move from within the gel onto the membrane while maintaining the organization they had within the gel. As a result of this "blotting" process, the proteins are exposed on a thin surface layer for detection. Both varieties of membrane are chosen for their non-specific protein binding properties (i.e. binds all proteins equally well). Protein binding is based upon hydrophobic interactions, as well as charged interactions between the membrane and protein. Nitrocellulose membranes are cheaper than PVDF, but are far more fragile and do not stand up well to repeated probings.The uniformity and overall effectiveness of transfer of protein from the gel to the membrane can be checked by staining the membrane with Coomassie or Ponceau S dyes. Coomassie is the more sensitive of the two, although Ponceau S's water solubility makes it easier to destain and probe the membrane as described below.

D. Blotting

Since the membrane has been chosen for its ability to bind protein, and both antibodies and the target are proteins, steps must be taken to prevent interactions between the membrane and the antibody used for detection of the target protein. Blocking of non-specific binding is achieved by placing the membrane in a dilute

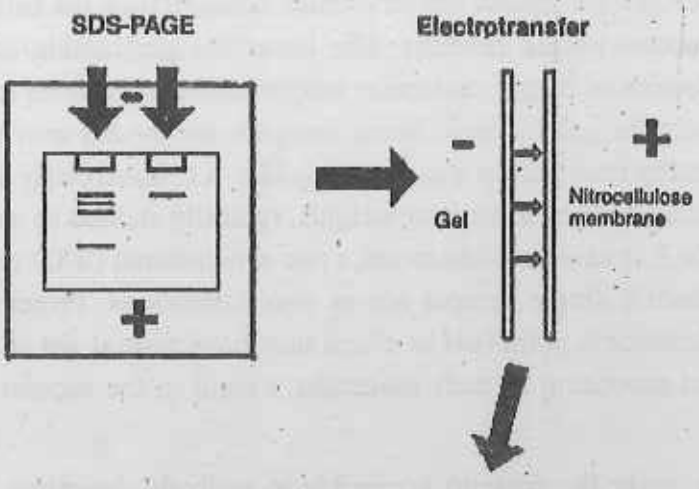


Fig. 1.6 : Immuno-staining of blot

solution of protein - typically Bovine serum albumin (BSA) with a minute percentage of detergent such as Tween 20. The protein in the dilute solution attaches to the membrane in all places where the target proteins have not attached. Thus, when the antibody is added, there is no room on the membrane for it to attach other than on the binding sites of the specific target protein. This reduces "noise" in the final product of the Western blot, leading to clearer results, and eliminates false positives.

E. Detection

During the detection process the membrane is "probed" for the protein of interest with a modified antibody which is linked to a reporter enzyme, which when exposed to an appropriate substrate drives a colourimetric reaction and produces a colour. For a variety of reasons, this traditionally takes place in a two-step process, although there are now one-step detection methods available for certain applications.

★ **Primary antibody**—Antibodies are generated when a host species or immune cell culture is exposed to the protein of interest (or a part thereof). Normally, this is part of the immune response, whereas here they are harvested and used as sensitive and specific detection tools that bind the protein directly.After blocking, a dilute solution of primary antibody (generally between 0.5 and 5 micrograms/ml) is incubated with the membrane under gentle agitation. Typically, the solution is comprised of buffered saline solution with a small percentage of detergent, and sometimes with powdered milk or BSA. The antibody solution and the membrane can be sealed and incubated together for anywhere from 30 minutes to overnight. It can also be incubated at different temperatures, with warmer temperatures being associated with more binding, both specific (to the target protein, the "signal") and non-specific ("noise").

★ **Secondary antibody**—After rinsing the membrane to remove unbound primary antibody, the membrane is exposed to another antibody, directed at a species-specific portion of the primary antibody. This is known as a secondary antibody, and due to its targeting properties, tends to be referred to as "anti-mouse," "anti-goat," etc. Antibodies come from animal sources (or animal sourced hybridoma cultures); an anti-mouse secondary will bind to just about any mouse-sourced primary antibody. This allows some cost savings by allowing an entirelab to share a single source of mass-produced antibody, and provides far more consistent results. The secondary antibody is usually linked to biotin or to a reporter enzyme such as alkaline phosphatase or horseradish peroxidase.

This means that several secondary antibodies will bind to one primary antibody and enhances the signal.

Most commonly, a **horseradish peroxidase**-linked secondary is used in conjunction with a chemiluminescent agent, and the reaction product produces **luminescence** in proportion to the amount of protein. A sensitive sheet of photographic film is placed against the membrane, and exposure to the light from the reaction creates an image of the antibodies bound to the blot.

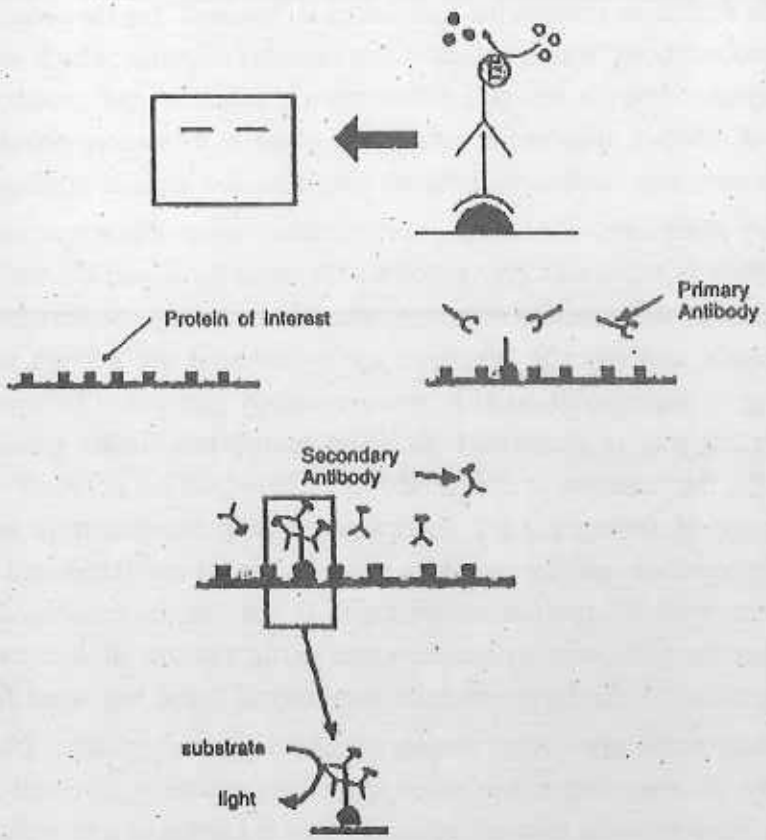


Fig. 1.7 : Schematic representation of western blot

F. Analysis

After the unbound probes are washed away, the western blot is ready for detection of the probes that are labeled and bound to the protein of interest. In practical terms, not all westerns reveal protein only at one band in a membrane. Size approximations

are taken by comparing the stained bands to that of the marker or ladder loaded during electrophoresis. The process is repeated for a structural protein, such as actin or tubulin, that should not change between samples. The amount of target protein is indexed to the structural protein to control between groups. This practice ensures correction for the amount of total protein on the membrane in case of errors or incomplete transfers.

1.4.2 Colorimetric detection

The colorimetric detection method depends on incubation of the western blot with a substrate that reacts with the reporter enzyme (such as peroxidase) that is bound to the secondary antibody. This converts the soluble dye into an insoluble form of a different color that precipitates next to the enzyme and thereby stains the membrane. Development of the blot is then stopped by washing away the soluble dye. Protein levels are evaluated through densitometry (how intense the stain is) or spectrophotometry.

Medical diagnostic applications

- The confirmatory HIV test employs a western blot to detect anti-HIV antibody in a human serum sample. Proteins from known HIV-infected cells are separated and blotted on a membrane as above. Then, the serum to be tested is applied in the primary antibody incubation step; free antibody is washed away, and a secondary anti-human antibody linked to an enzyme signal is added. The stained bands then indicate the proteins to which the patient's serum contains antibody.

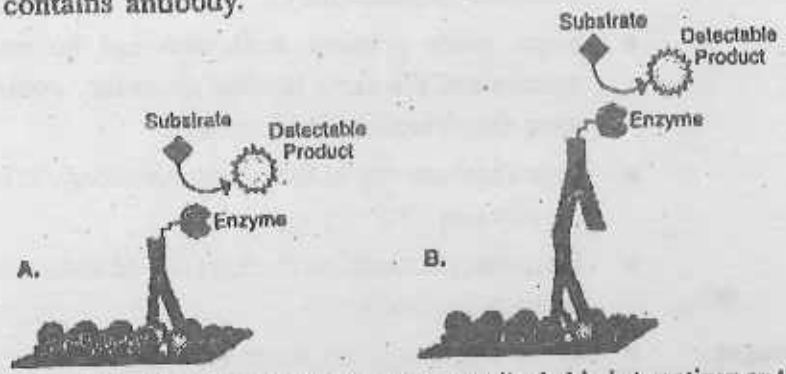


Fig. 1.8 : In the direct detection method, labeled primary antibody binds to antigen on the membrane and reacts with substrate,creating a detectable signal. B. In the indirect detection method, unlabeled primary antibody binds to the antigen. Then, a labeled secondary antibody binds to the primary antibody and reacts with the substrate.

- A western blot is also used as the definitive test for Bovine spongiform encephalopathy (BSE, commonly referred to as 'mad cow disease').
- Some forms of Lyme disease testing employ western blotting.

Table 1. Comparison of direct and indirect detection methods

	Direct detection
Advantages	• Quick methodology since only one antibody is used. • Cross-reactivity of secondary antibody is eliminated. • Double staining is easily achieved using different labels on primary antibodies from the same host.
Disadvantages	• Immunoreactivity of the primary antibody may be reduced as a result of labeling. • Labeling of every primary antibody is time consuming and expensive. • No flexibility in choice of primary antibody label from one experiment to another. • Little signal amplification.
	Indirect method
Advantages	• Sensitivity is increased because each primary antibody contains several epitopes that can be bound by the labeled secondary antibody, allowing for signal amplification. • A wide variety of labeled secondary antibodies are available commercially. • Since many primary antibodies can be made in one species and the same labeled secondary antibody can be used for detection, it is versatile. • Immunoreactivity of the primary antibody is not affected by labeling. • Different visualization markers can be used with the same primary antibody.
Disadvantages	• Cross-reactivity may occur with the secondary antibody, resulting in nonspecific staining. • An extra incubation step is required in the procedure.

Unit 2 ❐ Spectrophotometry, Colorimetry, Radioactivity, NMR & Raman Spectroscopy

Structure

2.1 Spectrophotometry and Colorimetry

2.1.1 Introduction

The interaction of electromagnetic radiation with matter is essentially a quantum phenomenon and dependent upon both the properties of radiation and the appropriate structural parts of the material involved. An understanding of the properties of electromagnetic radiation and its interaction with matter leads to a recognition of variety of types of spectra and consequently spectroscopic techniques and their application to the solution of biological problems.

Absorption spectrophotometry in the ultraviolet and visible region is considered to be one of the oldest physical methods used for quantitative analysis and structural elucidation. It should be kept in mind that IR and NMR techniques are mainly used for structural elucidation and qualitative analysis. On the other hand, UV-visible spectrophotometry is mainly used for quantitative analysis and serves as a useful auxiliary tool for structural elucidation.

Spectrophotometry is mainly concerned with the following regions of spectrum, ultraviolet, 200-400 nm; visible 400-800 nm. Colorimetry is concerned with the visible region of spectrum.

The instruments in use of for measuring the emission or absorption of radiant energy from substances are called by various names, such as photometers, colorimeters, spectrophotometers etc.

Photometer : This is an instrument, which measures the ratio, or some function of the two, of radiant power of two electromagnetic beams. This is an instrument

employing a filter to isolate a narrow wavelength region and a photocell or phototube to measure the intensity of radiation.

Spectrophotometer : This instrument measures the ratio, or a function of two, of the radiant power of two electromagnetic beams over a large wavelength region. In this instrument a monochromator is used instead of a filter. Spectrophotometer employs most sensitive detectors like phototube or photo-multipliers.

Colorimeter : Any instrument used for measuring absorption in the visible region is generally known as a colorimeter. In fact, some commercial filter photometers are called colorimeters.

2.1.2 Basic theory of spectrophotometry and colorimetry

When light is incident upon homogeneous medium a part of the incident light is reflected, a part is absorbed by the medium and the rest is allowed to transmit as such. If I_o, I_a, I_t, and Ir denote the incident, absorbed, transmitted and reflected light respectively, then : $I_o = I_a + I_t + I_r$. The value of I_r is very small and can be eliminated for air-glass interfaces. Under this condition one can write : $I_o = I_a + I_t$.

The laws governing absorption are generally known as Beer's law and Lambert's law. The Beer-Lambert law is a combination of the above laws, each dealing separately with the absorption of light, related to the concentration of the absorber (the substance responsible for absorbing the light) and the pathway or thickness of the layer (related to the absolute amount of absorber).

Lambert's law : *"When a beam of light is allowed to pass through a transparent medium, the rate of decrease of intensity with the thickness of medium is directly proportional to the intensity of light."*

Mathematically one can write $-[dI/dl] \propto I$; or $-[dI/dl] = kI$.....[1]

Where I = intensity of incident light of wavelength λ. l = thickness of the medium and k = proportionality constant.

On integration of the above equation we get, $\ln[I_o/I_t] = kl$

or $\log[I_o/I_t] = [k/2.303]t = Kl$.....[2].

Where I_o = Intensity of incident radiation (at l = 0); I_t = Intensity of transmitted radiation (l = 1). K = absorption coefficient.

Beer's law : This law states that *the amount of light absorbed by a material is proportional to the number of absorbing molecules, i.e. concentration of the absorbing solution.*

This can be mathematically expressed in the form of an equation similar to the one above. $\log[I_o/I_t]=[k'/2.303]c$.....[3]. Where k' = constant. C = concentration of the absorbing material.

Combining the equations [2] and [3] we get, $\log[I_o/I_t] = \varepsilon cl$.....[4]. Here K and k' merge to become a single constant. Equation 4 is the mathematical representation of Lambert-Beer's law. This combined law states that, *the amount of light absorbed (absorbance) is proportional to the concentration of the absorbing substance and to the thickness of the absorbing materai (path-length).*

The term $\log[I_o/I_t]$ is known as absorbance (A) or optical density (O.D.). ε is known as molar absorption coefficient or molar absorptivity or molar extinction coefficient. The term I_t/I_o is known as transmittance.

So the final conventional form of Lambert-Beer's law is $A = \varepsilon cl$.....[5]

If c is expressed in mol lit^{-1} and l in cm then the unit of ε is lit. mol^{-1} cm^{-1}. It should be remembered that the value of a substance depends on the wavelength of light used and on the medium in which the light absorbing species is present.

Note that the relationship between absorbance and concentration is linear. As concentration increase the absorbance also increases. This relationship allows one to convert an absorbance value into a concentration.

2.1.3 Instrumentation of a spectrophotometer

A spectrophotometer is an instrument designed to detect the amount of radiant light energy absorbed by molecules. To do this, the instrument must have five basic components : a light source; a prism or diffraction grating; an aperture or slit; a detector (a photoelectric tube); and a digital meter to display the output of the phototube. The arrangement of these parts is shown below.

When light is reflected from a diffraction grating, it is split into its component colors or wavelengths, which then diverge. Sections of the projected spectrum can be either blocked or allowed to pass through the slit so that only one wavelength will pass to the other sections of the spectrophotometer (The position of the grating is adjustable so that the region of the spectrum projected on the slit can be changed). Light that passes through the slit travels to the phototube, where it creates an electric current proportional to the number of photons striking the phototube. If a digital meter is attached to the phototube, the electric current output can be measured and recorded. The scale is usually calibrated in two ways: percent transmittance, which runs on a scale from 0 to 100; and absorbance, or optical density units, which runs from 0 to 2.

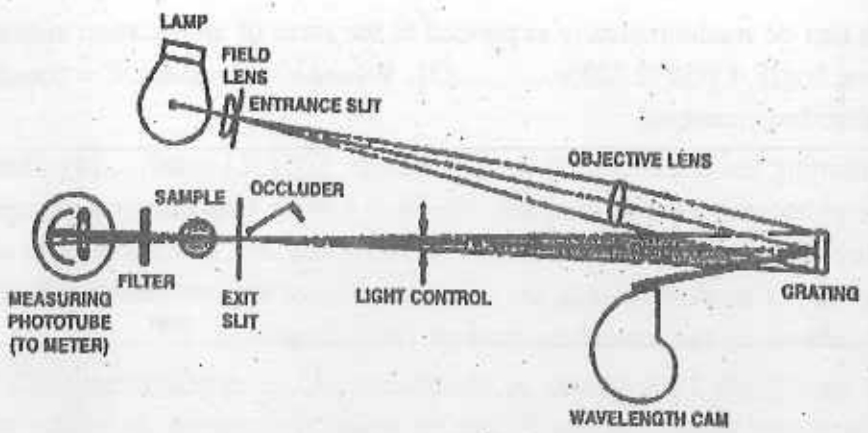
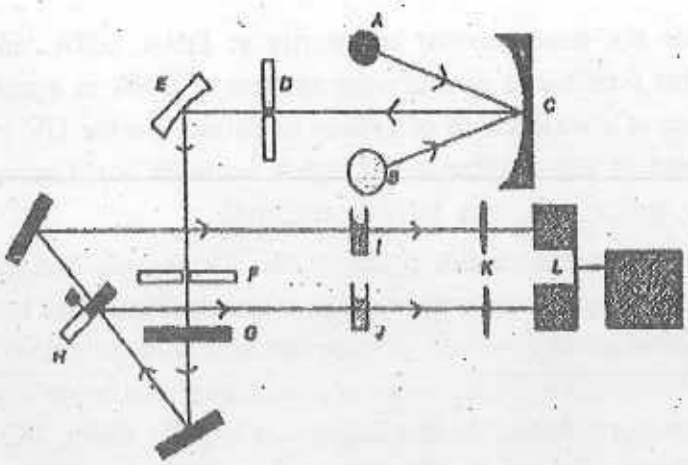


Fig. 2.1 Arrangement of different parts of a spectrophotometer

Before the light-absorbing properties of a solution can be measured, two adjustments on the spectrophotometer are necessary. First, the diffraction grating must be adjusted so that the desired wavelength of light passes through the slit. This is usually the wavelength of light that is most absorbed by the compound under consideration. Secondly, the output of the phototube must be adjusted or calibrated to correct the drift in the electronic circuits and dirt or contaminating the material in the light path between the source and the detector.

A modern version of a spectrophotometer (double beam) works as follows. The instrument is a simple device, with A and B being bulbs (A is a deuterium bulb, for UV range radiation, and B is a tungsten bulb for the visible range). When a device scans a range of frequencies it simply uses one bulb for visible, then switches to the other bulb for the UV. The beam is focused onto a mirror (C) that sends a beam through a slit (D) to make the beam narrow. The beam then hits the diffraction grating (E). This grating is a simple block that sends off different frequencies of light at different angles (like a prism). The required wavelength of radiation is sharpened through another slit (F), and then past through a filter (G) which removes any wide angle diffractive radiation. The beam is then bounced off of a mirror through a chopper mirror (H). This splits the beam equally between the blank ("reference") cell (I) and the sample cell (J). The collected beam is focussed through a lens (K) and collected at detectors (photomultiplier tubes or PMT, one for each beam, L). The information is computed at the workstation M and the spectrum is formulated.



A single beam device works much the same way, with slits, mirrors and diffraction gratings just as a double beam device does, but does not have a chopper mirror or two cell slots. In this case, the reference (blank) spectrum for the solvent is taken separately before the sample. The workstation stores the reference spectrum, and then deducts it from the sample spectrum when the sample spectrum is taken, to give the sample UV-Vis spectrum.

Because all solutions of chemical compounds absorb light of specific wavelengths, spectrophotometry can be useful in identifying compounds. Furthermore, because the amount of light absorbed is proportional to the concentration of a compound, spectrophotometry is also useful in determining concentrations.

2.1.4 Applications of UV spectrophotometry

Photometry being a very versatile technique, has diverse applications. UV-Vis spectroscopy may be used to identify classes of compounds in both the pure and in biological preparations. From the characteristics absorption spectra of the specific class of compounds it is possible to identify the substances in biological milieu.

In biological studies, Molecular biologists routinely work with DNA, RNA, and proteins and have devised some simple, fast spectrophotometric assays for these molecules. The purpose of this exercise is to use the UV absorbance of biological samples to obtain qualitative and quantitative information about those samples.

2.1.5 The UV absorbance spectra of nucleic acids and proteins

Most biological molecules do not intrinsically absorb light in the visible range, but they do absorb ultraviolet light. Biologists take advantage of UV absorbance to

quickly estimate the concentration and purity of DNA, RNA, and proteins in a sample. It is also possible to quantify the amount of DNA in a sample by looking at its absorbance at a wavelength of 260nm or 280nm (in the UV region). The UV method described in this chapter is not highly accurate but it is very widely used since it is easy, quick, and little DNA is required.

Proteins have two absorbance peaks in the UV region, one between 215-230 nm, where peptide bonds absorb, and another at about 280 nm due to light absorption by aromatic amino acids (tyrosine, tryptophan and phenylalanine). Certain of the subunits of nucleic acids (purines) have an absorbance maximum slightly below 260 nm while others (pyrimidines) have a maximum slightly above 260 nm. Therefore, although it is common to say that the absorbance peak of nucleic acids is 260 nm, in reality, the absorbance maxima of different fragments of DNA vary somewhat depending on their subunit composition.

Concentration measurements of nucleic acids and proteins in a sample

It is possible to determine the concentration of nucleic acids or proteins based on their absorbance at a wavelength of 260 nm or 280 nm respectively. A calibration curve using standards of known concentration can be constructed. For accurate results, the standard curve should be prepared using the protein of interest or DNA that is similar to that in the sample being measured. The linear range for DNA values is reported to be from about 5–50 µg/mL. Depending on the protein, UV analysis of proteins at 280 nm has a linear range from about 0.1–5 mg/mL.

Biologists commonly use a "short-cut" to roughly estimate the concentration of nucleic acid or protein in a sample based on the sample's absorbance at 260 or 280 nm. This short-cut method uses absorptivity constants. Recall that given an absorptivity constant, it is possible to by-pass the preparation of a standard curve by applying Beer's Law.

The absorptivity constant for a particular protein at 280 nm depends on its composition. Proteins that contain a higher percentage of aromatic amino acids have higher absorptivities at 280 nm than those with fewer. The absorptivity constant for a nucleic acid depends on its base composition and on whether it is single-stranded or double-stranded. Despite the fact that different proteins and nucleic acid fragments vary in their absorptivity, analysts commonly use "average" absorptivity constants to estimate the concentration of nucleic acid or protein in a sample.

Estimation of the purity of a nucleic acid preparation

It is possible to use UV spectrophotometry to estimate the purity of a solution of nucleic acids. This method involves measuring the absorbance of the solution at two wavelengths, usually 260 nm and 280 nm, and calculating the ratio of the two absorbances :

- ★ An A_{260}/A_{280} ratio of 2.0 is characteristic of pure RNA.
- ★ An A_{260}/A_{280} of 1.8 is characteristic of pure DNA.
- ★ A_{260}/A_{280} ratio of about 0.6 is characteristic of pure protein

Therefore, a ratio of 1.8 - 2.0 is desired when purifying nucleic acids. A ratio less than 1.7 means there is probably a contaminant in the solution, typically either protein or phenol.

Binding of small molecules to proteins

Whenever a substrate binds to a protein, it affects the polarity of the region or makes certain amino acids residues, hitherto accessible, unavailable. In doing so, it might quite frequently cause some spectral changes which may be measured. If these spectral changes are compared to the results obtained with other studies such as spectrophotometric titration, quite a bit of information about the structure of the active site might be obtained.

Binding of small molecules to nucleic acids

When a small molecule having chromophoric site bind to nucleic acid (by external stacking or intercalation of through groove of nucleic acids) the UV-Vis spectrum of the molecule gets perturbed. From the spectral change and absorbance values at a particular characteristic wavelength of the small molecule, different binding parameters like association constant, size of binding sites can be calculated.

Other than above applications, spectrophotmetry can be in enzyme assay, molecular weight determination, kinetics of a reaction, thermal stability studies of nucleic acids, proteins *etc.*

2.2 Isotopes and radioactivity

2.2.1 Isotopes

Elements are defined by the number of protons, positively charged subatomic particles, in an atom's nucleus. The number of protons in an atom's nucleus is

termed its *atomic number*. Isotopes of a given element carry different numbers of neutrons, or neutrally charged particles, in their nuclei. The sum of the number of neutrons and protons in an atom's nucleaus defines its approximate *atomic weight*. For example, all carbon atoms have six protons; isotopes of carbon can have 6, 7, or 8 neutrons.

2.2.2 Radioactivity

Radioactive isotopes (also called *radioisotopes*) have unstable nuclei. These isotopes disintegrate to form atoms with stable nuclei by the release of subatomic particles and *gamma rays* (akin to X-rays). The radioactive elements are referred to as parents atoms; the atoms they disintegrate to form are called daughter products.

Some isotopes release an *alpha particle* during nuclear disintegration; an alpha particle consists of two protons and two neutrons (equivalent to the nucleus of an atom of helium). Others release a *beta particle*, which is an electron, or negatively charged nuclear particle. Beta particles originate in the nucleus, presumably by breakdown of a neutron into its proton-electron components. Gamma rays are released during both types of radioactive decay.

When an isotope emits an alpha particle, the resultant daughter product has an atomic number two units less than its parent's atomic number, and an atomic weight four units less than its parent's atomic weight. When an isotope emits a beta particle, it decays to a daughter with an atomic number one unit greater and an essentially unchanged atomic weight.

Some isotopes decay and immediately produce a stable daughter product. For example, one-step decays to stable daughters are completed by the radiogenic isotopes ^{14}C (decaying to ^{14}N by the beta process), and ^{87}Rb (decaying to ^{87}Sr by the beta process). Others decay and produce unstable daughters, which then become the parent products of their own daughters. Unstable isotopes producing unstable daughters form a *radioactive decay chain*. For example, the ^{235}U decay chain eventually produces ^{207}Pb, a stable daughter.

Using empirical data, it is possible to statistically forecast what percentage of a radioisotope's population will decay over a given period of time. This has enabled workers to define a *half-life* for each radioisotope, the period required for one-half of the original parent population to decay to its stable daughter product. Each radioisotope has its own characteristic half-life.

Suppose that at its inception, a sample contains 100 units of a parent radioisotope. After one half-life has passed, there will remain 50 units of the parent isotope, and 50 units of the daughter product will have been produced. After another half-life, 25 units of the parent isotope will remain, and 75 units of the daughter product will have been produced. After another half-life, 12.5 units of the parent isotope will remain, and 87.5 units of the daughter product will have been produced. Through time, the number of parents constantly decreases while the number of daughters constantly increases. Theoretically, although the number of parents will become insignificantly small, there should never come a time when all of the parent population has decayed to daughters.

Knowing the value of a specific isotope's half-life, it is possible to determine the age of a geologic or archaeologic sample by evaluating the amount of parent and daughter isotopes in it. For example, given the half-life of U-235 is 7×10^8 y, suppose you have a rock sample containing minerals having 1 unit of ^{235}U and 3 units of ^{207}Pb. The sample must have originally contained 4 units of parent material, and 25% of the parent material (U-235) remains. Examination of the curve above shows that time equivalent to two half-lives have passed, or approximately 1.4 billion years.

2.2.2.1 Exponential decay law

The transition of a nucleus from one state to another is characterised by a "lifetime". This does not imply that each nucleus will "live" for that time (or approximately that time). It was shown very early on that decay is a *random* process: this gives rise to an exponential decay law.

The decay of (a particular state of) a nucleus is determined by one number, the decay constant λ. (We will see how this relates to a lifetime). The study of nuclear transition mechanisms leads to an understanding of the decay constant. In this section we will examine the consequences of this random process.

Experience has shown that decay is a random process. The probability of a nucleus decaying in a time interval dt is λdt. The probability is thus independent of time, it is independent of the age of a particular nucleus and is the same for all nuclei in the same state (i.e. decay is a Poisson process). As a result of this, we cannot predict when a particular nucleus will decay, we can only make predictions about ensembles.

Suppose that we have $N(t)$ nuclei in a particular state at time t. Then in a time interval dt a number $N(t)\lambda dt$ will decay, so $dN(t) = -\lambda N(t)dt$.

Integrating this gives : $N(t) = N(0).exp(-\lambda t)$: the famous *exponential decay law*.

A characteristic of this equation is that in a fixed time interval T, a fixed fraction $exp(-\lambda t)$ will decay. Thus we cannot talk of a lifetime of a nucleus, but only of what fraction will decay in a certain interval. This leads us to single out a particular time interval as interesting: we define the *half-life* as that time interval in which half of an initial sample will decay. From the equation $N(T_{1/2}) = N(0)/2$ we thus obtain, $T_{1/2} = ln2/\lambda = 0.693/\lambda$.

It is also useful to define a *mean life* τ in the usual statistical way :

$$\tau = \frac{\int t\, dN}{\int dN} = \frac{1}{\lambda} = 1.44 T_{1/2}$$

The definition of the half-life suggests a way in which the decay constant can be measured. However, determination of small amounts of rare heavy nuclides is a difficult process. A way out is obtained by looking rather at the *rate* of decay. We define the *activity* $A(t) = \lambda N(t)$. Thus for the case of exponential decay, $A(t) = \lambda N(t) = -dN(t)/dt$

From the exponential radioactive decay law we obtain $A(t) = A(0)exp(-\lambda t)$

Note that activity represents the number of decays per unit time interval. This is not necessarily equal to (minus) the rate of change of the number of atoms present—this is only true in the case of a single decay as above.

Thus we can determine the decay constant by measuring the rate of decay, i.e. by measuring the number of decays ("counts per second") in a certain time interval as a function of time.

These relationships can be used to determine the age of a geologic or archaeologic sample. Results of such studies are most effeective if enough time has passed to let a substantial amount of the daughter product grow (perhaps 10%), and are of limited use if morethan six half-lives have passed (because not enough of the parent material remains to study). Dating of archaeological samples is commonly conducted using C-14, which has a half-life of 5730 y. Dating of geologic samples is most often accomplished using K-40 (with a K-Ar half-life of 1.3×10^9 y), Rb-87 (with a half-life of 4.9×10^{10} y), U-235 (with a half-life of 7.0×10^8 y), and U-238 (with a half-life of 4.5×10^9 y).

Interpretation of data must take into consideration several factors that can yield inaccurate results. For example, metamorphic processes can "reset" radiometric clocks. If daughter products are noble gases - for example, Ar or Rn- loss of the daughter product can occur as the gases diffuse from minerals.

In other instances, a mineral can be created with a substantial amount of a daughter product already incorporated. For example, ^{40}K is commonly used to determine radiometric ages. About 90% of all ^{40}K undergoes a beta decay to produce a daughter of ^{40}Ca; ^{40}K can also undergo a process called *electron capture* to produce a daughter of ^{40}Ar. (In electron capture, a proton is transformed to a neutron.) Radiometric dating techniques focus on the ^{40}K-^{40}Ar system because Ca is a common constituent in many rock-forming minerals, and it is not possible to distinguish the Ca that was derived from decay of ^{40}K from the Ca that was originally in the sample.

2.2.2.2 Units of radioactivity

The unit of activity is the bequerel (Bq), defined by 1 Bq = 1 decay per second, or curie (Ci), where 1 Ci = 3.7×10^{10} decays per second.

2.2.2.3 Application of radioisotopes in biology

Radioisotopes are widely used in biological sciences. However, a brief description is given here.

1. ***Metabolic studies*** : Radioistopes have been used almost in each phase metabolic pathways : amino acid metabolism, photosynthesis, TCA cycle, protein biosynthsis, steroid metabolism *etc.* By the use of proper radio-isotope, isolation and identification of metabolites have been possible. Radioisotopes provides a good method of ascertaining turnover times for a particular compound. Using proper isotopes it has been shown that liver protein is turned over in 7-14 days, while collagen is turned over at a rate less than 10% per annum.
2. ***Pharmacological and clinical applications*** : Another field where the radioisotopes are widely used is in the drug development. Radioisotopes are widely used for diagnostic tests. The isotopes which are commonly used are ^{131}I, ^{108}Au, ^{32}P, ^{133}Xe, ^{99}Tc, ^{51}Cr *etc.* ^{51}Cr is used for determination of blood volume; ^{133}Xe is used for lung function test. Kidney function test using [^{133}I] iodohippuric acid are used in diagnoses of kidney infection, kidney blockage,

or imbalance of function between two kidneys. Radioisotopes are very useful in treating cancer. ^{60}Co has been to regress many tumors. Yittrium-90 has been used for cancer therapy. ^{238}Pu is used very successfully to operate a pacemaker.

3. ***Radioimmunoassay*** : One of the most significant advances in biochemical techniques in recent years has been the development of radioimmunoassay. The radioimmunological determination of hormones or other biologically important substances that occur in very low concentrations is a very important aspect of radioisotope use.
4. ***Ecological studies*** : Radioisotopes are also used in ecology. Migratory patterns and behaviour patterns of many animals can be monitored by using radioisotopes. Another ecological application is in the examination of food chains where the primary producers can be made radioactive and the path of radioactivity followed throughout the resulting food chain.
5. ***Application in food technology*** : ^{60}Co or ^{137}Ce are used as strong γ-emitter in food industry for the sterilization of foods like milk, meat *etc.*

2.2.2.4 Shielding of radiation

To decrease dosease, precautions are taken to limit a worker's exposure in nuclear waste areas. Radiation in the form of alpha particles cannot pass through material merely as thick as a piece of paper. Beta particles cannot pass through metal. Gamma rays can be halted by lead shielding. So it is possible to wear protective clothing and to use respirators to limit exposure to radiation. As per international norms, the permissible occupational dose is 5 rems a year. The exposure any workers receive is monitored carefully and evaluated at regular periods.

2.3 NMR Spectroscopy

2.3.1 Introduction

Nuclear Magnetic Resonance spectroscopy is a very important and powerful tool for the chemist for the determination of molecular structure and it is theoretically complex analytical tool. In this part the basic theory behind the technique is to be discussed. It is important to remember that, NMR deals with the experiments on the nuclei of atoms, not the electrons. The chemical environment of specific nuclei is deduced from information obtained about the nuclei.

2.3.2 Theory

Nuclear spin and the splitting of energy levels in a magnetic field

The particles like electrons, protons and neutrons, can be imagined as spinning on their axes. In many atoms (such as ^{12}C) these spins are paired against each other, such that the nucleus of the atom has no overall spin. However, in some atoms (such as ^{1}H and ^{13}C) the nucleus does possess an overall spin. The rules for determining the net spin of a nucleus are as follows;

1. If the number of neutrons and the number of protons are both even, then the nucleus has NO spin.
2. If the number of neutrons plus the number of protons is odd, then the nucleus has a half-integer spin (i.e. 1/2, 3/2, 5/2)
3. If the number of neutrons and the number of protons are both odd, then the nucleus has an integer spin (i.e. 1, 2, 3)

The overall spin, I, is important. Quantum mechanics tells that a nucleus of spin I will have $2I + 1$ possible orientations. A nucleus with spin 1/2 will have 2 possible orientations. In the absence of an external magnetic field, these orientations are of equal energy. If a magnetic field is applied, then the energy levels split. Each level is given a *magnetic quantum number, m.*

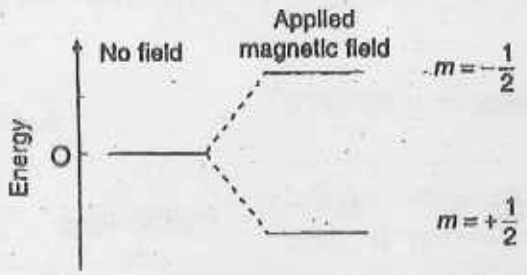


Fig. 2.2. Energy levels for a nucleus with spin quantum number 1/2

When the nucleus is in a magnetic field, the initial populations of the energy levels are determined by thermodynamics, as described by the Boltzmann distribution. This is very important, and it means that the lower energy level will contain slightly more nuclei than the higher level. It is possible to excite these nuclei into the higher level with electromagnetic radiation. The frequency of radiation needed is determined by the difference in energy between the energy levels.

Calculating transition energy

The nucleus has a positive charge and is spinning. This generates a small magnetic field. The nucleus therefore possesses a magnetic moment, which is proportional to its spin, I.

$$\mu = \frac{\gamma I h}{2\pi}$$

The constant γ, is called the *magnetogyric ratio* and is a fundamental nuclear constant which has a different value for every nucleus. h is Plancks constant.

The energy of a particular energy level is given by;

$$E = \frac{\gamma h}{2\pi} mB$$

Where B is the strength of the magnetic field at the nucleus.

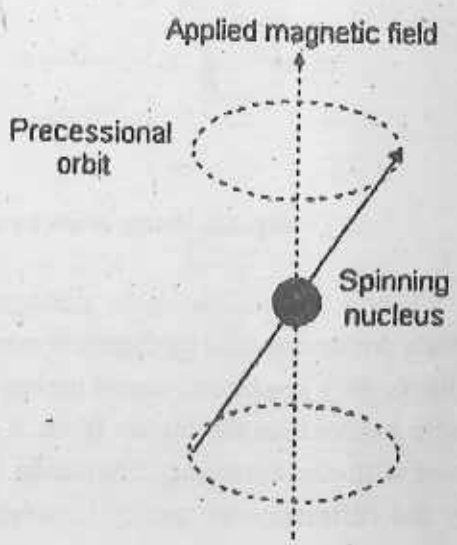
The difference in energy between levels (the transition energy) can be found from

$$\Delta E = \frac{\gamma h B}{2\pi}$$

This means that if the magnetic field, B, is increased, so is ΔE. It also means that if a nucleus has a relatively large magnetogyric ratio, then ΔE is correspondingly large.

The absorption of radiation by a nucleus in a magnetic field

In this part, "classical" view of the behaviour of the nucleus–that is, the behaviour of a charged particle in a magnetic field is to be considered. Here, a nucleus (of spin 1/2) is considered in a magnetic field. This nucleus is in the lower energy level (i.e. its magnetic moment does not oppose the applied field). The nucleus is spinning on its axis. In the presence of a magnetic field, this axis of rotation will *precess* around the magnetic field. The frequency of precession



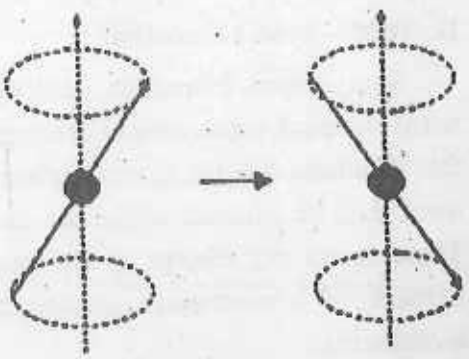
is termed the *Larmor frequency*, which is identical to the transition frequency.

The potential energy of the precessing nucleus is given by;

$E = -\mu B \cos\theta$ where θ is the angle between the direction of the applied field and the axis of nuclear rotation.

If energy is absorbed by the nucleus, then the angle of precession, θ, will change. For a nucleus of spin ½, absorption of radiation "flips" the magnetic moment so that it opposes the applied field (the higher energy state).

It is important to realise that only a small proportion of "target" nuclei are in the lower energy state (and can absorb radiation). There is the possibility that by exciting these nuclei, the populations of the higher and lower energy levels will become equal. If this occurs, then there will be no further absorption of radiation. The spin system is *saturated*. The possibility of saturation means that we must be aware of the relaxation processes which return nuclei to the lower energy state.



Relaxation processes

How do nuclei in the higher energy state return to the lower state (relaxation process)?

There are two major relaxation processes;

- Spin – lattice (longitudinal) relaxation
- Spin – spin (transverse) relaxation

A. Spin – lattice relaxation

Nuclei in an NMR experiment are in a sample. The sample in which the nuclei are held is called the *lattice*. Nuclei in the lattice are in vibrational and rotational motion, which creates a complex magnetic field. The magnetic field caused by motion of nuclei within the lattice is called the *lattice field*. This lattice field has many components. Some of these components will be equal in frequency and phase to the Larmor frequency of the nuclei of interest. These components of the lattice field can interact with nuclei in the higher energy state, and cause them to loose

energy (returning to the lower state). The energy that a nucleus losses increases the amount of vibration and rotation within the lattice (resulting in a tiny rise in the temperature of the sample).

The relaxation time, T_1 (the average lifetime of nuclei in the higher energy state) is dependant on the magnetogyric ratio of the nucleus and the mobility of the lattice. As mobility increases, the vibrational and rotational frequencies increase, making it more likely for a component of the lattice field to be able to interact with excited nuclei. However, at extremely high mobilities, the probability of a component of the lattice field being able to interact with excited nuclei decreases.

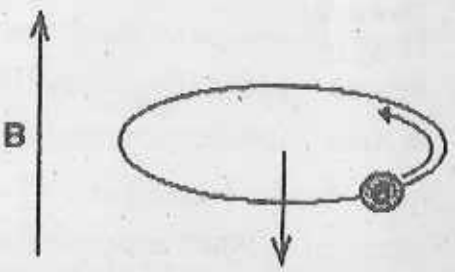
B. Spin – spin relaxation

Spin – spin relaxation describes the interaction between neighbouring nuclei with identical precessional frequencies but differing magnetic quantum states. In this situation, the nuclei can exchange quantum states; a nucleus in the lower energy level will be excited, while the excited nucleus relaxes to the lower energy state. There is no net change in the populations of the energy states, but the average lifetime of a nucleus in the excited state will decrease. This can result in line-broadening.

Chemical shift

The magnetic field at the nucleus is not equal to the applied magnetic field; electrons around the nucleus shield it from the applied field. The difference between the applied magnetic field and the field at the nucleus is termed the *nuclear shielding*.

Consider the s-electrons in a molecule. They have spherical symmetry and circulate in the applied field, producing a magnetic field which opposes the applied field. This means that the applied field strength must be increased for the nucleus to absorb at its transition frequency. This *upfield shift* is also termed *diamagnetic shift*.



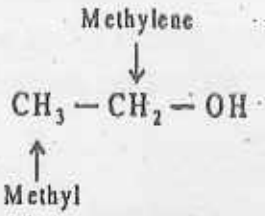
Magnetic field produced by circulating electron

Electrons in p-orbitals have no spherical symmetry. They produce comparatively large magnetic fields at the nucleus, which give a *low field shift*. This "deshielding" is termed *paramagnetic shift*.

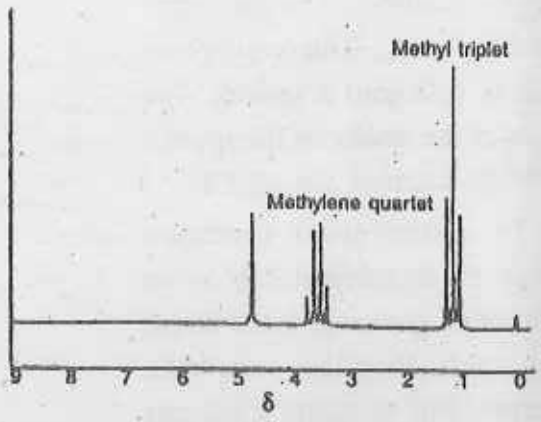
Chemically different hydrogens in an organic molecule do not experience the same magnetic field. Electrons shield the nucleus thereby reducing the effective magnetic field and requiring energy of a lower frequency to cause resonance. On the other hand, when electrons are withdrawn from a nucleus, the nucleus is deshielded and feels a stronger magnetic field requiring more energy (higher frequency) to cause resonance. Thus, NMR can provide information about a hydrogen's electronic environment. Generally, hydrogens bound to carbons attached to electron withdrawing groups tend to resonate at higher frequencies (more downfield (to the left) from TMS, tetramethylsilane, a common NMR standard). The position of where a particular hydrogen atom resonates relative to TMS is called its chemical shift. Learning where typical hydrogens resonate requires experience and study, but learning some common chemical shifts will provide you with a tremendous advantage at solving structural problems using NMR.

Spin – spin coupling

Consider the structure of ethanol ;

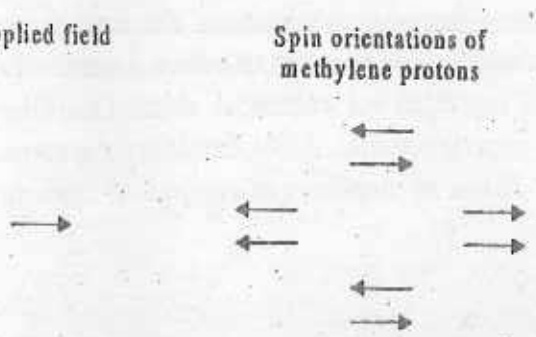


The ^1H NMR spectrum of ethanol (below) shows the methyl peak has been split into three peaks (a *triplet*) and the methylene peak has been split into four peaks (a *quartet*). This occurs because there is a small interaction (*coupling*) between the two groups of protons. The spacings between the peaks of the methyl triplet are equal to the spacings between the peaks of the methylene quartet. This spacing is measured in Hertz and is called the *coupling constant, J*.



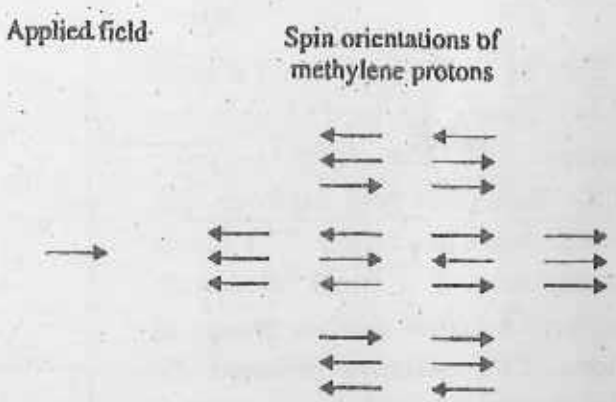
To see why the methyl peak is split into a triplet, let's look at the methylene protons. There are two of them, and each can have one of two possible orientations (aligned with or opposed against the applied field). This gives a total of four possible states;

In the first possible combination, spins are paired and opposed to the field. This has the effect of reducing the field experienced by the methyl protons; therefore a slightly higher field is needed to bring them to resonance, resulting in an upfield shift. Neither combination of spins opposed to each other has an effect on the methyl peak. The spins paired in the direction of the field produce a downfield shift. Hence, the methyl peak is split into three, with the ratio of areas 1 : 2 : 1.



Similarly, the effect of the methyl protons on the methylene protons is such that there are eight possible spin combinations for the three methyl protons;

Out of these eight groups, there are two groups of three magnetically equivalent combinations. The methylene peak is split into a quartet. The areas of the peaks in the quartet have the ration 1 : 3 : 3 : 1.



In a *first-order* spectrum (where the chemical shift between interacting groups is much larger than their coupling constant), interpretation of splitting patterns is quite straightforward;

- The multiplicity of a multiplet is given by the number of equivalent protons in neighbouring atom~ plus one, i.e. *the n + 1 rule.*

- Equivalent nuclei do not interact with each other. The three methyl protons in ethanol cause splitting of the neighbouring methylene protons; they do not cause splitting among themselves.
- The coupling constant is not dependant on the applied field. Multiplets can be easily distinguished from closely spaced chemical shift peaks.

2.3.3 Instrumentation

There are two general types of NMR instrument; continuous wave and Fourier transform. Early experiments were conducted with continuous wave (C.W.) instruments, and in 1970 the first Fourier transform (F.T.) instruments became available. This type now dominates the market.

Continuous wave NMR instruments

Continuous wave NMR spectrometers are similar in principle to optical spectrometers. The sample is held in a strong magnetic field, and the frequency of the source is slowly scanned (in some instruments, the source frequency is held constant, and the field is scanned).

Fourier transform NMR instruments

The magnitude of the energy changes involved in NMR spectroscopy are small. This means that sensitivity is a major limitation. One way to increase sensitivity would be to record many spectra, and then add them together; because noise is random, it adds as the square root of the number of spectra recorded. For example, if one hundred spectra of a compound were recorded and summed, then the noise would increase by a factor of ten, but the signal would increase in magnitude by a factor of one hundred, giving a large increase in sensitivity. However, if this is done using a continuous wave instrument, the time needed to collect the spectra is very large (one scan takes two to eight minutes).

In FT-NMR, all frequencies in a spectrum are irradiated simultaneously with a radio frequency pulse. Following the pulse, the nuclei return to thermal equilibrium. A time domain emission signal is recorded by the instrument as the nuclei relax. A frequency domain spectrum is obtained by Fourier transformation.

Solvents for NMR spectroscopy

NMR spectra are usually measured using solutions of the substance being

investigated. It is important that the solvent itself doesn't contain any simple hydrogen atoms, because they would produce confusing peaks in the spectrum.

There are two ways of avoiding this. One can use a solvent such as tetrachloromethane, CCl_4, which doesn't contain any hydrogen, or one can use a solvent in which any ordinary hydrogen atoms are replaced by its isotope, deuterium– for example, $CDCl_3$ instead of $CHCl_3$. All the NMR spectra used on this site involve $CDCl_3$ as the solvent.

Deuterium atoms have sufficiently different magnetic properties from ordinary hydrogen that they don't produce peaks in the area of the spectrum that we are looking at.

2.3.4 Applications of NMR

1. ***Structural diagnosis*** : The study of molecular structure, conformational changes and certain types of kinetic investigation is the main use of NMR in the biological field. The use of this technique in drug metabolism studies is of increasing importance. Thus structural information which relates the biological functions of the antibiotics such as valinomycin and gramicidin has been obtained from NMR studies. Structural information about small proteins such as some neurotoxins, various cytochromes, hen egg-white lysozyme and calcium binding proteins have been obtained using NMR spectroscopy. For proteins studies, however, the technique is combined with X-ray study for better information.
2. ***Study of dynamic characteristics of protein structure*** : Processes which modulate internuclear distances *e.g.* the internal motions of macromolecules or chemical exchange, affect the intrinsic properties of NMR resonance and make it possible to probe the dynamic aspect of molecular structure by this technique. Thus internal motions of proteins such as the opening of secondary structure, aromatic side chain rotation, segmental motion of the main chain, and the overall tumbling of protein have been studied using NMR spectroscopy. The proteins which have been studied for their dynamic characteristics include histones, cytochrome B5, prothrombin, plasminogen and chromogranin A.
3. ***Study of nucleic acids*** : Application of NMR to the nucleic acids includes not only the structural information regarding DNA or RNA, but also the investigations of interactions between various drugs and DNA and between

proteins and DNA. Sequence specific oligosaccharides have been obtained but work on intact glycoproteins is difficult due to problem in convoluting the data.

4. ***Membrane transport*** : Membrane transport either in *vivo* systems or using synthetic membranes has been studied using NMR. The technique has been used to study alanine and lactate transport in the human eryothrocyte by exploiting the difference in the magnetic susceptibility between the inside and out side of the cells. Very recently Na^+ transport in human erythrocyte has also been studied using NMR.
5. ***Studies on complex formation*** : Using NMR it is possible to detect very small conformational changes. Thus NMR has been used to study complex formation such as the binding of ligand to enzyme, a drug to DNA, an agonist to the receptor, or an antigen to an antibody.
6. ***Thermodynamic studies*** : In the general case of an equilibrium between two states, NMR can be used to measure the associated thermodynamic quantities. NMR has been used to measure the quantities such as binding constants, Enthalpies and entropies of binding, partition coefficients etc. Examples can be cited on folding and unfolding of proteins and t-RNAs, and the interaction between actinomycin D and deoxy-pGpC. This technique has also been used on the mechanism of action of proteases and on the mechanism of their inhibition by some other proteins.
7. ***Magnetic reseonance imaging*** : The applications described earlier may be extended into the clinical environment. Now-a-days physiological samples can be studied directly. In recent times an important outgrowth of the development of NMR is magnetic resonance imaging (MRI). MRI offer the analytical biochemist and the clinician a phenomenological variety of procedures. An MRI of the human body is created from the magnetic resonance of the protons of water. The instrument creates different magnetic field strengths B_o at different locations in the human body. In other words, unlike the NMR described here, MRI uses a non-homogeneous magnetic field B_o. So different water molecules will resonate at different B_1 frequencies depending on their physical location. Hence, the B_1 frequency indicates where the water molecules reside and the strength of the resonance signal indicates the abundance of water molecules at that location. The image from water molecules in tumors

can be enhanced because the excited states of these water molecules relax back to the ground state more rapidly than water molecules in non-tumor environments, and this difference in relaxation can be detected.

2.4 Raman Spectroscopy

2.4.1 Theory

The Raman effect and normal Raman scattering

When light is scattered from a molecule most photons are elastically scattered. The scattered photons have the same energy (frequency) and, therefore, wavelength, as the incident photons. However, a small fraction of light (approximately 1 in 10^7 photons) is scattered at optical frequencies different from, and usually lower than, the frequency of the incident photons. The process leading to this inelastic scatter is the termed the Raman effect. Raman scattering can occur with a change in vibrational, rotational or electronic energy of a molecule. Here we will use the term Raman effect to mean vibrational Raman effect only.

The difference in energy between the incident photon and the Raman scattered photon is equal to the energy of a vibration of the scattering molecule. A plot of intensity of scattered light versus energy difference is a Raman spectrum.

The scattering process

The Raman effect arises when a photon is incident on a molecule and interacts with the electric dipole of the molecule. It is a form of electronic (more accurately, vibronic) spectroscopy, although the spectrum contains vibrational frequencies. In classical terms, the interaction can be viewed as a perturbation of the molecule's

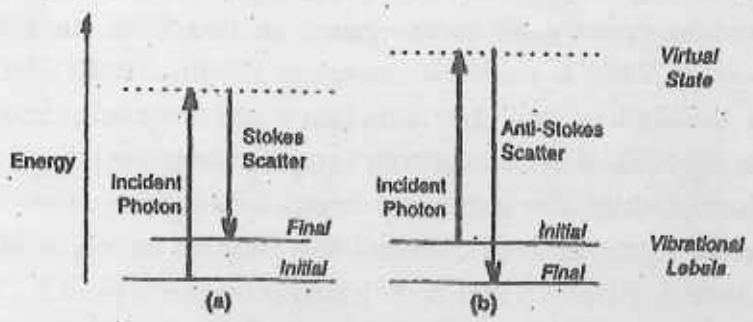


Fig. 2.3 Energy level diagram for Raman scattering; (a) Stokes Raman scattering (b) anti-Stokes Raman scattering.

electric field. In quantum mechanics the scattering is described as an excitation to a virtual state lower in energy than a real electronic transition with nearly coincident de-excitation and a change in vibrational energy. The scattering event occurs in 10^{-14} seconds or less. The virtual state description of scattering is shown in Figure 2.3(a)

The energy difference between the incident and scattered photons is represented by the arrows of different lengths in Figure 2.3a. Numerically, the energy difference between the initial and final vibrational levels, $\bar{\nu}$, or Raman shift in wave numbers (cm^{-1}), is calculated through equation 1

$$\bar{\nu} = \frac{1}{\lambda_{incident}} - \frac{1}{\lambda_{scattered}} \quad \text{..... [1]}$$

in which $\lambda_{incident}$ and $\lambda_{scattered}$ are the wavelengths (in cm) of the incident and Raman scattered photons, respectively. The vibrational energy is ultimately dissipated as heat. Because of the low intensity of Raman scattering, the heat dissipation does not cause a measurable temperature rise in a material.

At room temperature the thermal population of vibrational excited states is low, although not zero. Therefore, the initial state is the ground state, and the scattered photon will have lower energy (longer wavelength) than the exciting photon. This Stokes shifted scatter is what is usually observed in Raman spectroscopy. Figure 2.3a depicts Raman Stokes scattering.

A small fraction of the molecules are in vibrationally excited states. Raman scattering from vibrationally excited molecules leaves the molecule in the ground state. The scattered photon appears at higher energy, as shown in Figure 2.3b. This anti-Stokes-shifted Raman spectrum is always weaker than the Stokes-shifted spectrum, but at room temperature it is strong enough to be useful for vibrational frequencies less than about 1500 cm^{-1}. The Stokes and anti-Stokes spectra contain the same frequency information. The ratio of anti-Stokes to Stokes intensity at any vibrational frequency is a measure of temperature. Anti-Stokes Raman scattering is used for contactless thermometry. The anti-Stokes spectrum is also used when the Stokes spectrum is not directly observable, for example because of poor detector response or spectrograph efficiency.

Vibrational energies

The energy of a vibrational mode depends on molecular structure and

environment. Atomic mass, bond order, molecular substituents, molecular geometry and hydrogen bonding all effect the vibrational force constant which, in turn dictates the vibrational energy. For example, the stretching frequency of a phosphorus-phosphorus bond ranges from 460 to 610 to 775 cm^{-1} for the single, double and triple bonded moieties, respectively. Much effort has been devoted to estimation or measurement of force constants. For small molecules, and even for some extended structures such as peptides, reasonably accurate calculations of vibrational frequencies are possible with commercially available software.

Vibrational Raman spectroscopy is not limited to intramolecular vibrations. Crystal lattice vibrations and other motions of extended solids are Raman-active. Their spectra are important in such fields as polymers and semiconductors. In the gas phase, rotational structure is resolvable on vibrational transitions. The resulting vibration/rotation spectra are widely used to study combustion and gas phase reactions generally. Vibrational Raman spectroscopy in this broad sense is an extraordinarily versatile probe into a wide range of phenomena ranging across disciplines from physical biochemistry to materials science.

Raman selection rules and intensities

A simple classical electromagnetic field description of Raman spectroscopy can be used to explain many of the important features of Raman band intensities. The dipole moment, P, induced in a molecule by an external electric field, E, is proportional to the field as shown in equation 2.

$$P = \alpha E \quad \ldots\ldots\ldots [2]$$

The proportionality constant a is the polarizability of the molecule. The polarizability measures the ease with which the electron cloud around a molecule can be distorted. The induced dipole emits or scatters light at the optical frequency of the incident light wave.

Raman scattering occurs because a molecular vibration can change the polarizability. The change is described by the polarizability derivative, $\partial\alpha/\partial q$, where q is the normal coordinate of the vibration. The selection rule for a Raman-active vibration, that there be a change in polarizability during the vibration, is given in equation 3.

$$\frac{\partial \alpha}{\partial q} \neq 0 \quad \ldots\ldots\ldots [3]$$

The Raman selection rule is analogous to the more familiar selection rule for

an infrared-active vibration, which states that there must be a net change in permanent dipole moment during the vibration.

If a vibration does not greatly change the polarizability, then the polarizability derivative will be near zero, and the intensity of the Raman band will be low. The vibrations of a highly polar moiety, such as the O–H bond, are usually weak. An external electric field can not induce a large change in the dipole moment and stretching or bending the bond does not change this.

Resonance-Enhanced Raman scattering

Raman spectroscopy is conventionally performed with green, red or near-infrared lasers. The wavelengths are below the first electronic transitions of most molecules, as assumed by scattering theory. The situation changes if the wavelength of the exciting laser within the electronic spectrum of a molecule. In that case the intensity of some Raman-active vibrations increases by a factor of 10^2-10^4. This resonance enhancement or resonance Raman effect can be quite useful.

Metalloporphyrins, carotenoids and several other classes of biologically important molecules have strongly allowed electronic transitions in the visible. The spectrum of the chromophoric moiety is resonance enhanced and that of the surrounding protein matrix is not. This allows the physical biochemist to probe the chromophoric site (often the active site) without spectral interference from the surrounding protein. Resonance Raman spectroscopy is also a major probe of the chemistry of fullerenes, polydiacetylenes and other "exotic" molecules which strongly absorb in the visible. Although many more molecules absorb in the ultraviolet, the high cost of lasers and optics for this spectral region have limited UV resonance Raman spectroscopy to a small number of specialists.

Surface-Enhanced Raman scattering

The Raman scattering from a compound (or ion) adsorbed on or even within a few Angstroms of a structured metal surface can be 10^3-10^6X greater than in solution. This surface-enhanced Raman scattering is strongest on silver, but is observable on gold and copper as well. At practical excitation wavelengths, enhancement on other metals is unimportant. Surface-enhanced Raman scattering (SERS) arises from two mechanisms.

The first is an enhanced electromagnetic field produced at the surface of the metal. When the wavelength of the incident light is close to the plasma wavelength

of the metal, conduction electrons in the metal surface are excited into an extended surface electronic excited state called a surface plasmon resonance. Molecules adsorbed or in close proximity to the surface experience an exceptionally large electromagnetic field. Vibrational modes normal to the surface are most strongly enhanced.

The second mode of enhancement is by the formation of a charge-transfer complex between the surface and analyte molecule. The electronic transitions of many charge transfer complexes are in the visible, so that resonance enhancement occurs.

Molecules with lone pair electrons or pi clouds show the strongest SERS. The effect was first discovered with pyridine. Other aromatic nitrogen or oxygen containing compounds, such as aromatic amines or phenols, are strongly SERS active. The effect can also been seen with other electron-rich functionalities such as carboxylic acids.

2.4.2 Applications

The use of Raman spectroscopy is mainly in the biochemical research for intermediate-sized molecules such as drugs, metabolic intermediates and substrates. Examples are the identification of substances such as penicillins, small peptides and environmental pollutants. Raman spectroscopy is a rapid method for measuring certain contaminants in foodstuff. The principal advantage of Raman spectroscopy that water has a fairly weak Raman spectrum. Therefore biological samples are easily studied in aqueous solution. Use of this technique in study of photosynthesis and respiration in plants is valuable, particularly for CO_2 metabolism.

In the past, Raman spectroscopy was not often used for the studies of biopolymers because of various difficulties. In the use of normal visible light as a source, there is always contamination of the exciting light with the Raman spectrum. But, presently by the use of LASER this difficulty has been solved.

2.5 References

1. Hollas J.M. (2002) *Basic Atomic and Molecular Spectroscopy., Willey/RSC, London*
2. Pavia, D.L, Lampmann, N.G.M and Kriz G.S. (2001) *Introduction to Spectroscopy*, 3rd Edn., Harcourt, New York.

3. Slater, R..J. (2002) *Radiation in Biology- A Practical Approach*, 2nd Edn., Oxford University Press, Oxford.
4. Billington, D., Jayson, G.G. and Maltby, P.J. (1992) *Radioisotopes,* Bios Scientific, Oxford.
5. Evans, J.N.S. (1995) *Biomolecular NMR Spectroscopy*. Oxford University Press, Oxford.
6. *Principles and Techniques of Biochemistry and Molecular Biology,* (2005) Edited by Wilson K. and Walker J, 6th Edn, Cambridge University Press
7. Schrader, B. *Infrared and Raman Spectroscopy*; Schrader, B. ed., VCH Publishers Inc. : New York, 1995.
8. Myers, A.B., Mathies, R.A. *Biological Applications of Raman Spectroscopy: Volume 2:* John Wiley and Sons: New York, 1987.

Unit 3 ❐ Fluorescence microscopy, confocal microscopy, electron microscopy

Structure

3.1 Fluorescence microscopy

3.1.1 Introduction

Fluorescence is a member of the ubiquitous luminescence family of processes in which susceptible molecules emit light from electronically excited states created by either a physical (for example, absorption of light), mechanical (friction), or chemical mechanism. Generation of luminescence through excitation of a molecule by ultraviolet or visible light photons is a phenomenon termed photoluminescence, which is formally divided into two categories, fluorescence and phosphorescence, depending upon the electronic configuration of the excited state and the emission pathway. Fluorescence is the property of some atoms and molecules to absorb light at a particular wavelength and to subsequently emit light of longer wavelength after a brief interval, termed the fluorescence lifetime. The process of phosphorescence occurs in a manner similar to fluorescence, but with a much longer excited state lifetime.

Fluorescence illumination and observation is the most rapidly expanding microscopy technique employed today, both in the medical and biological sciences, a fact which has spurred the development of more sophisticated microscopes and numerous fluorescence accessories.

A fluorescence microscope is basically a conventional light microscope with added features and components that extend its capabilities.

- ★ A conventional microscope uses light to illuminate the sample and produce a magnified image of the sample.
- ★ A fluorescence microscope uses a much higher intensity light to illuminate the sample. This light excites fluorescence species in the sample, which then emit

light of a longer wavelength. A fluorescent microscope also produces a magnified image of the sample, but the image is based on the second light source, the light emanating from the fluorescent species, rather than from the light originally used to illuminate, and excite, the sample. In contrast to other modes of optical microscopy that are based on macroscopic specimen features, such as phase gradients, light absorption, and birefringence, fluorescence microscopy is capable of imaging the distribution of a single molecular species based solely on the properties of fluorescence emission. Thus, using fluorescence microscopy, the precise location of intracellular components labeled with specific fluorophores can be monitored, as well as their associated diffusion coefficients, transport characteristics, and interactions with other biomolecules. In addition, the dramatic response in fluorescence to localized environmental variables enables the investigation of pH, viscosity, refractive index, ionic concentrations, membrane potential, and solvent polarity in living cells and tissues.

3.1.2 Basic requirements of fluorescence microscope optics

Light source : In order to generate enough excitation light intensity to furnish secondary fluorescence emission capable of detection, powerful light sources are needed. These are usually either mercury or xenon arc (burner) lamps, which produce high-intensity illumination powerful enough to image faintly visible fluorescence specimens.

Nearly all fluorescence microscopes use the objective lens to perform two functions :

1. **Focus the illumination (excitation) light on the sample :** In order to excite fluorescent species in a sample, the optics of a fluorescent microscope must focus the illumination (excitation) light on the sample to a greater extent than is achieved using the simple condenser lens system found in the illumination light path of a conventional microscope.
2. **Collect the emitted fluorescence :** This type of excitation-emission configuration, in which both the excitation and emission light travel through the objective, is called epifluorescence. The key to the optics in an epifluorescence microscope is the separation of the illumination (excitation) light from the fluorescence emission emanating from the sample. In order to obtain either an image of the emission without excessive background illumination, or a measurement of the fluorescence emission without background "noise", the optical elements used to separate these two light components must be very efficient.

The dichroic mirror

In a fluorescence microscope, a dichroic mirror is used to separate the excitation and emission light paths. Within the objective, the excitation emission shares the same optics.

In a fluorescence microscope, the dichroic mirror separates the light paths.

- ★ The excitation light reflects off the surface of the dichroic mirror into the objective.
- ★ The fluorescence emission passes through the dichroic to the eyepiece or detection system.

The dichroic mirror's special reflective properties allow it to separate the two light paths Fig. 3.1. Each dichroic mirror has a set wavelength value, called the transition wavelength value, which is the wavelength of 50% transmission. The mirror reflects wavelengths of light below the transition wavelength value and transmits wavelengths above this value. This property accounts for the name given to this mirror (dichroic, two color). Ideally, the wavelength of the dichroic mirror is chosen to be between the wavelengths used for excitation and emission. The dichroic mirror is a key element of the fluorescence microscope, but it is not able to perform all of the required optical functions on its own. Typically, about 90% of the light at wavelengths below the transition wavelength value are reflected and about 90% of the light at wavelengths above this value are transmitted by the dichroic mirror. When the excitation light illuminates the sample, a small amount of excitation light is reflected off the optical elements within the objective and some excitation light is scattered back into the objective by the sample. Some of this "excitation" light is transmitted through the dichroic mirror along with the longer wavelength light emitted by the sample. This "contaminating" light would otherwise reach the detection system if it were not for another wavelength selective element in the fluorescence microscope : an emission filter.

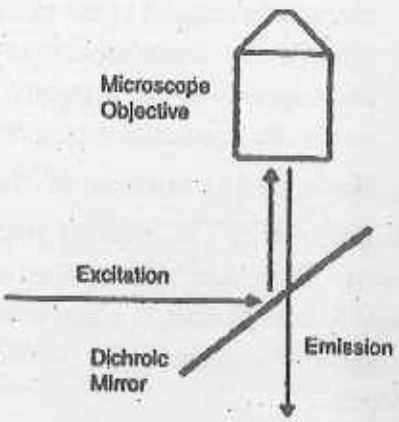


Fig. 3.1 The Dichroic mirror separates excitation and emission light paths.

Excitation and emission filters

Two filters are used along with the dichroic mirror :

- ★ **Excitation filter** : In order to select the excitation wavelength, an excitation filter is placed in the excitation path just prior to the dichroic mirror.
- ★ **Emission filter** : In order to more specifically select the emission wavelength of the light emitted from the sample and to remove traces of excitation light, an emission filter is placed beneath the dichroic mirror. In this position, the filter functions to both select the emission wavelength and to eliminate any trace of the wavelengths used for excitation.

These filters are usually a special type of filter referred to as an interference filter, because of the way in which it blocks the out of band transmission. Interference filters exhibit an extremely low transmission outside of their characteristic bandpass. Thus, they are very efficient in selecting the desired excitation and emission wavelengths.

The Filter cube

The dichroic mirror is mounted on an optical block commonly referred to as a filter cube. The excitation and emission filters are usually affixed to the filter cube. This cube provides a convenient means to change the dichroic mirror without direct handling of either the mirror or filters. Fig. 3.2 shows the light path through the filter cube in a fluorescent microscope.

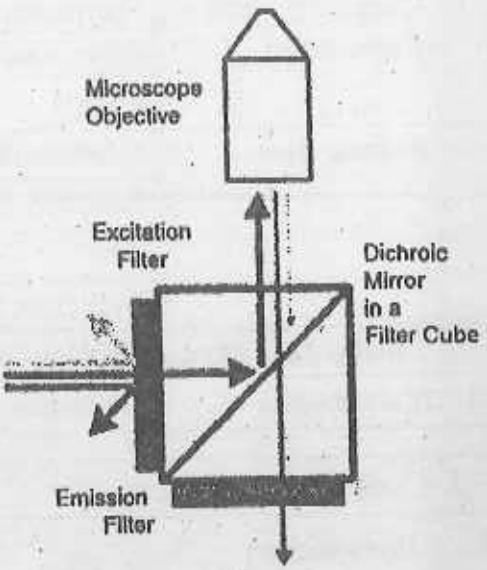


Fig. 3.2. Light path through the filter cube in a fluorescence microscope.

It is often the case that a specific combination of excitation filter, emission filter and dichroic mirror are needed to visualize and/or quantitate the fluorescence emission from a particular fluorescent species. In newer models of fluorescence microscopes, manufacturers have provided a means to change these optical elements in a convenient manner by arranging a set of four or more filter cubes in a circular (or linear) turret under the objective. With a turret arrangement, a specific filter cube can be selected in a manner similar to that of selecting a specific objective.

Fluorophores in fluorescence microscopy

Many fluorescent probes are constructed around synthetic aromatic organic chemicals designed to bind with a biological macromolecule. Fluorescent dyes are also useful in monitoring cellular integrity (live versus dead and apoptosis), endocytosis, exocytosis, membrane fluidity, protein trafficking, signal transduction, and enzymatic activity. In addition, fluorescent probes have been widely applied to genetic mapping and chromosome analysis in the field of molecular genetics. A list of some fluorophores give in the following table.

Dye		Excitation Maximum (nm)	Emission Maximum (nm)
1. Commonly used fluorophores	Fluorescein	496	518
	Texas Red	592	610
	Bodipy	503	511
2. Nuclear dyes	Acridine Orange	502	526
	Ethidium Bromide	510	595
	DAPI	359	461
	Propidium Iodide	536	617
3. Calcium indicators	Calcium Green	506	533
4. Mitochondria	Rhodamine 123	507	529

3.2 Confocal microscopy

3.2.1 Introduction

Now-a days, confocal microscopy is considered to be one of the most important advances ever achieved in optical microscopy. This microscopy provides several advantages over conventional widefield optical microscopy, including the ability to control depth of field, elimination or reduction of background information. The basic key to the confocal approach is the use of spatial filtering techniques to eliminate out-of-focus light or glare in specimens. The popularity of confocal microscopy in recent years, due in part to the relative ease with which extremely high-quality images can be obtained from specimens prepared for conventional fluorescence microscopy, and the growing number of applications in cell biology that rely on imaging both fixed and living cells and tissues.

Modern confocal microscopes can be considered as completely integrated electronic systems where the optical microscope plays a central role in a configuration that consists of one or more electronic detectors, a computer (for image display, processing, output, and storage), and several laser systems combined with wavelength selection devices and a beam scanning assembly. In most cases, integration between the various components is so thorough that the entire confocal microscope is often collectively referred to as a digital or video imaging system capable of producing electronic images. These microscopes are now being employed for routine investigations on molecules, cells, and living tissues that were not possible just a few years ago.

3.2.2 Principles of confocal microscopy

The confocal principle in epi-fluorescence laser scanning microscopy is diagrammatically presented in Fig. 3.3. Coherent light emitted by the laser system (excitation source) is used in confocal microscope. Laser is a good candidate for providing a point-like light soure. Unlike a mixture of wavelength, laser provides light at discreted band of wavelength and has very high intensity which are also advantages for fluorescence exitation. In practical, a point-like light source is achieved by using a laser light passing through a illumination pinhole. This point-like light source is directed to the specimen by a beam splitter to form a point-like illumination in the specimen. The point-illumination move or scan on the specimen by the help of a scanner. The reflected emission light from specimen's focal plane passes through the detecting pinhole and form point-like image on detector PMT (photon multiply tube). The significant amount of fluorescence emission that occurs at points above and below the objective focal plane is not confocal with the pinhole (termed Out-of-Focus Light Rays in Fig 3.3). Because only a small fraction of the out-of-focus fluorescence emission is delivered through the pinhole aperture, most of this extraneous light is not detected by the photomultiplier and does not contribute to the resulting image.

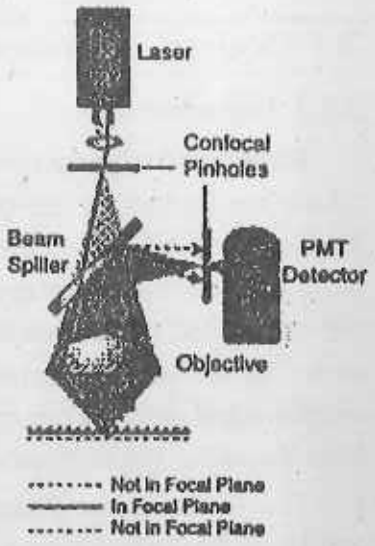


Fig. 3.3 Basic structure of Confocal microscope

3.2.3 Applications of confocal microscopy

The broad range of applications confocal microscopy includes a wide variety of studies in neuroanatomy and neurophysiology, as well as morphological studies of a wide spectrum of cells and tissues. In addition, the growing use of new fluorescent proteins is rapidly expanding the number of original research reports coupling these useful tools to modern microscopic investigations. Other applications include resonance energy transfer, stem cell research, photobleaching studies, lifetime imaging, multiphoton microscopy, total internal reflection, DNA hybridization, membrane and ion probes, bioluminescent proteins, and epitope tagging.

Taking advantage of the high scan speed of line scan in confocal scanner and fast sampling speed on the PMTs (in contrast, digital camera has to take image in a frame of n × n matrix, the frame readout speed is limited), confocal microscope can be used to monitor the highly dynamic intra cellular events such as calcium release, concentration change of calcium and other ions like K, Na, Mg, Ze, pH change in the cytoplasm of a cell.

3.3 Electron microscopy

3.3.1 Introduction

Electron Microscopes were developed due to the limitations of Light Microscopes which are limited by the physics of light to 500x or 1000x magnification and a resolution of 0.2 micrometers. Electron Microscopes are scientific instruments that use a beam of highly energetic electrons to examine objects on a very fine scale. In the early 1930's this theoretical limit had been reached and there was a scientific desire to see the fine details of the interior structures of organic cells (nucleus, mitochondria...etc.). This required 10,000x plus magnification which was just not possible using Light Microscopes.

The Transmission Electron Microscope (TEM) was developed by Max Knoll and Ernst Ruska in Germany in 1931. It was the first type of Electron Microscope to be developed and is patterned exactly on the Light Transmission Microscope except that a focused beam of electrons is used instead of light to "see through" the specimen.

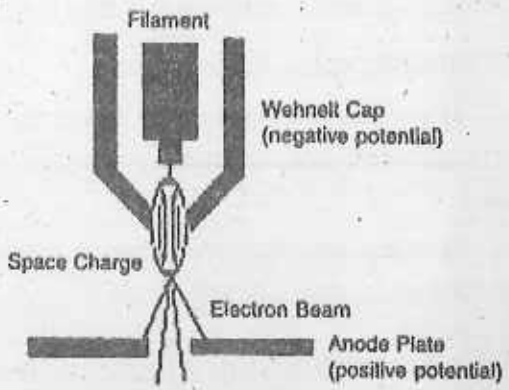
The first Scanning Electron Microscope (SEM) debuted in 1942 with the first commercial instruments around 1965. Its late development was due to the electronics involved in "scanning" the beam of electrons across the sample.

3.3.2 Principle

Electron Microscope (EM) function exactly in the same way as their optical counterparts. In EM a focused beam of electrons instead of light is used to "image" the specimen and gain information as to its structure and composition. The steps involved in all EMs are as follows :

A stream of electrons is formed (by the Electron Source) and accelerated toward the specimen using a positive electrical potential. All Electron Microscopes utilize an electron source of some kind with the majority using a Themionic Gun. A Thermionic Electron Gun functions in the following manner

1. A positive electrical potential is applied to the anode. The filament (cathode) is heated until a stream of electrons is produced and the electrons are then accelerated by the positive potential down the column.
2. A negative electrical potential is applied to the Whenelt Cap. As the electrons move toward the anode any ones emitted from the filament's side are repelled by the Whenelt Cap toward the optic axis.



3. A collection of electrons occurs in the space between the filament tip and Whenelt Cap. This collection is called a space charge. Those electrons at the bottom of the space charge (nearest to the anode) can exit the gun area through the small (< 1 mm) hole in the Whenelt Cap These electrons then move down the column to be later used in imaging and all electrons have similar energies (monchromatic).

   This electron stream obtained from electron gun is confined and focused using metal apertures (A thin disk or strip of metal (usually Pt) with a small circular through-hole. Apertures are used to restrict electron beams and filter out unwanted scattered electrons before image formation) and magnetic lenses into a thin, focused, monochromatic beam. This beam is focused onto the sample using a magnetic lens. The interactions occur inside the irradiated sample affect the electron beam and these interactions and effects are detected and transformed into an image in electron microscope.

An electron microscope can yield the following information :

3.3.3 Topography

The surface features of an object or "how it looks", its texture; direct relation between these features and materials properties (hardness, reflectivity...etc.)

3.3.4 Morphology

The shape and size of the particles making up the object; direct relation between these structures and materials properties (ductility, strength, reactivity...etc.)

3.3.5 Composition

The elements and compounds that the object is composed of and the relative amounts of them; direct relationship between composition and materials properties (melting point, reactivity, hardness...etc.)

Crystallographic Information

How the atoms are arranged in the object; direct relation between these arrangements and materials properties (conductivity, electrical properties, strength... etc.).

A more specific treatment of the workings of two different types of EMs are described in more detail :

3.4 Transmission electron microscope (TEM) : principle and application

A TEM works much like a slide projector. A projector shines a beam of light through the slide, as the light passes through it is affected by the structures and objects on the slide. These effects result in only certain parts of the light beam being transmitted through certain parts of the slide. This transmitted beam is then projected onto the viewing screen, forming an enlarged image of the slide.

TEMs work the same way except that they shine a beam of electrons (like the light) through the specimen(like the slide). Whatever part is transmitted is projected onto a phosphor screen for the user to see. A more technical explanation of a typical TEMs workings is as follows (Fig. 3.4) :

1. In TEM a stream of monochromatic electrons is produced from the electron gun (the "Virtual Source" at the top). This stream is focused to a small, thin, coherent beam by the use of condenser lenses 1 and 2. The first lens largely

determines the "spot size"; the general size range of the final spot that strikes the sample. The second lens actually changes the size of the spot on the sample; changing it from a wide dispersed spot to a pinpoint beam. The beam is restricted by the condenser aperture, knocking out high angle electrons.

2. The beam strikes the specimen and parts of it are transmitted and this transmitted portion is focused by the objective lens into an image. Optional Objective and Selected Area metal apertures can restrict the beam; the Objective aperture enhancing contrast by blocking out high-angle diffracted electrons, the Selected Area aperture enabling the user to examine the periodic diffraction of electrons by ordered arrangements of atoms in the sample.

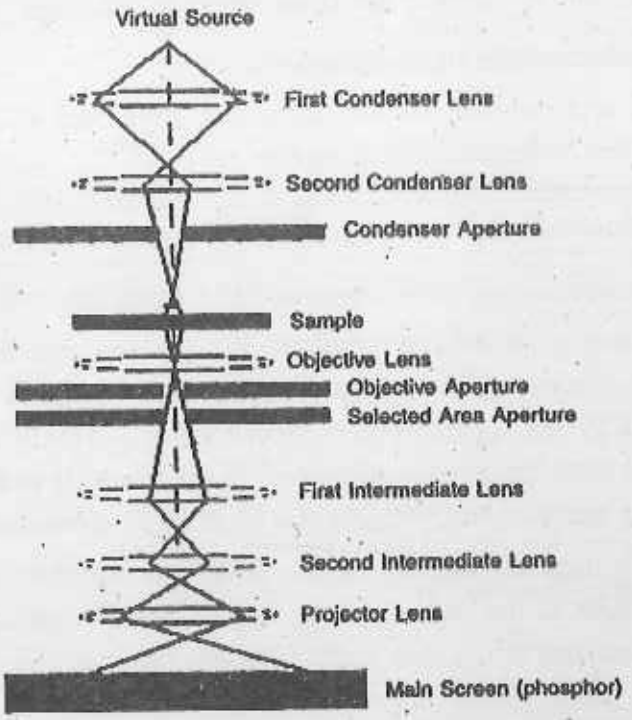


Fig. 3.4 : Technical explantion of a typical–TEM

3. The image is passed down the column through the intermediate and projector lenses, being enlarged all the way. The image strikes the phosphor image screen and light is generated, allowing the user to see the image. The darker areas of the image represent those areas of the sample that fewer electrons were transmitted through (they are thicker or denser). The lighter areas of the

image represent those areas of the sample that more electrons were transmitted through (they are thinner or less dense).

From the application of TEM the following information can be obtained. TEMs are patterned after Transmission Light Microscopes and will yield similar information.

Morphology

The size, shape and arrangement of the particles which make up the specimen as well as their relationship to each other on the scale of atomic diameters.

Crystallographic Information

The arrangement of atoms in the specimen and their degree of order, detection of atomic-scale defects in areas a few nanometers in diameter

Compositional Information (if so equipped)

The elements and compounds the sample is composed of and their relative ratios, in areas a few nanometers in diameter

3.5 Scanning electron microscope (SEM) : principle and application

A detailed explanation of how a typical SEM functions follows (Fig. 2.5) :

The "Virtual Source" at the top represents the electron gun, producing a stream of monochromatic electrons. The stream is condensed by the first condenser lens (usually controlled by the "coarse probe current knob"). This lens is used to both form the beam and limit the amount of current in the beam. It works in conjunction with the condenser aperture to eliminate the high-angle electrons from the beam.

1. The beam is then constricted by the condenser aperture (usually not user selectable), eliminating some high-angle electrons The second condenser lens forms the electrons into a thin, tight, coherent beam and is usually controlled by the "fine probe current knob".
2. A set of coils then "scan" or "sweep" the beam in a grid fashion (like a television), dwelling on points for a period of time determined by the scan speed (usually in the microsecond range)
3. The final lens, the Objective, focuses the scanning beam onto the part of the specimen desired. When the beam strikes the sample (and dwells for a few microseconds) interactions occur inside the sample and are detected with various instruments.

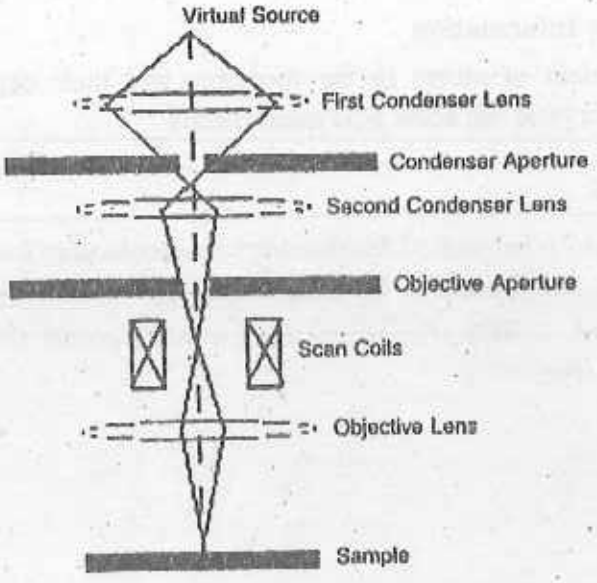


Fig. 3.5 : Explanation of a typical SEM

4. Before the beam moves to its next dwell point these instruments count the number of interactions and display a pixel on a CRT whose intensity is determined by this number (the more reactions the brighter the pixel).
5. This process is repeated until the grid scan is finished and then repeated, the entire pattern can be scanned 30 times per second.

SEMs are patterned after Reflecting Light Microscopes and yield similar information :

Topography

The surface features of an object or "how it looks", its texture; detectable features limited to a few manometers

Morphology

The shape, size and arrangement of the particles making up the object that are lying on the surface of the sample or have been exposed by grinding or chemical etching; detectable features limited to a few manometers

Composition

The elements and compounds the sample is composed of and their relative ratios, in areas – 1 micrometer in diameter

Crystallographic Information

The arrangement of atoms in the specimen and their degree of order; only useful on single-crystal particles >20 micrometers

3.6 References

1. *Principles and Techniques of Biochemistry and Molecular Biology,* (2005) Edited by Wilson K. and Walker J, 6th Edn, Cambridge University Press.
2. Abramowitz M. (2003) *Microscope Basics and Beyond.* Olympus of America Inc. Melville, Ny.

Unit 4 ❐ Biosensor-Nature & Application

Structure

4.1 Biosensor

A biosensor is an analytical device which converts a biological response into an electrical signal in order to quantify them. The term 'biosensor' is often used to cover sensor devices used in order to determine the concentration of substances and other parameters of biological interest even where they do not utilize a biological system directly. Biosensors make use of the specificity of biological processes, enzymes for their substrates and other ligands, antibodies for their antigens, lectins for carbohydrate and nucleic acid for their complementary sequences. A typical biosensor has a number of connected parts (Fig. 4.1)

A successful biosensor must possess at least some of the following beneficial features :

1. The biocatalyst must be highly specific for the purpose of the analyses, be stable under normal storage conditions and, except in the case of colorimetric enzyme strips and dipsticks show good stability over a large number of assays (i.e. much greater than 100).
2. The reaction should be as independent of such physical parameters as stirring, pH and temperature as is manageable. This would allow the analysis of samples with minimal pre-treatment. If the reaction involves cofactors or coenzymes these should, preferably, also be co-immobilised with the enzyme.
3. The response should be accurate, precise, reproducible and linear over the useful analytical range, without dilution or concentration. It should also be free from electrical noise.
4. If the biosensor is to be used for invasive monitoring in clinical situations, the probe must be tiny and biocompatible, having no toxic or antigenic effects. If it is to be used in fermenters it should be sterilisable. This is preferably performed by autoclaving but no biosensor enzymes can presently withstand such drastic wet-heat treatment. In either case, the biosensor should not be prone to fouling or proteolysis.
5. The complete biosensor should be cheap, small, portable and capable of being used by semi-skilled operators.

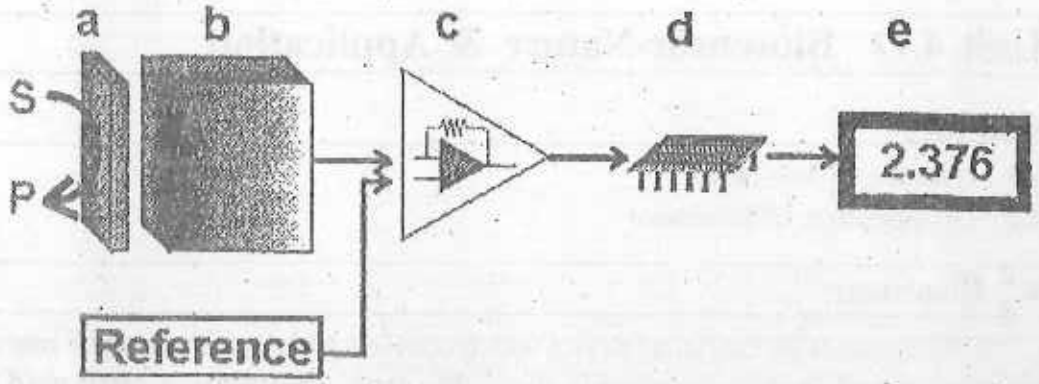


Fig. 4.1 : Schematic diagram showing the main components of a biosensor. The biocatalyst (a) converts the substrate to product. This reaction is determined by the transducer (b) which converts it to an electrical signal. The output from the transducer is amplified (c), processed (d) and displayed (e).

The key part of a biosensor is the transducer (shown as the 'black box' in Figure 4.1) which makes use of a physical change accompanying the reaction. In biosensor, a transducer is used to convert the biological events, be it a catalytic response or a binding reaction, in to a form where it can be read directly or further processed by a microprocessor. Based on transducer types it may be following types

1. The heat output (or absorbed) by the reaction **(calorimetric biosensors),**
2. Changes in the distribution of charges causing an electrical potential to be produced **(potentiometric biosensors),**
3. Movement of electrons produced in a redox reaction **(amperometric biosensors),**
4. Light output during the reaction or a light absorbance difference between the reactants and products **(optical biosensors),** or
5. Effects due to the mass of the reactants or products **(piezo-electric biosensors).**

Transducer	Examples
Electrochemical	Clark oxygen electrode, ion selective electrode
Optical	Photodiodes, Waveguide systems
Piezo electric	Quartz crystals, surface acoustic wave devices
Calorimetric	Thermistor, thermopile

4.2 Application of biosensor

There are three so-called 'generations' of biosensors; first generation biosensors where the normal product of the reaction diffuses to the transducer and causes the electrical response, second generation biosensors which involve specific 'mediators' between the reaction and the transducer in order to generate improved response, and third generation biosensors where the reaction itself causes the response and no product or mediator diffusion is directly involved.

The electrical signal from the transducer is often low and superimposed upon a relatively high and noisy (i.e. containing a high frequency signal component of an apparently random nature, due to electrical interference or generated within the electronic components of the transducer) baseline. The signal processing normally involves subtracting a 'reference' baseline signal, derived from a similar transducer without any biocatalytic membrane, from the sample signal, amplifying the resultant signal difference and electronically filtering (smoothing) out the unwanted signal noise. The relatively slow nature of the biosensor response considerably eases the problem of electrical noise filtration. The analogue signal produced at this stage may be output directly but is usually converted to a digital signal and passed to a microprocessor stage where the data is processed, converted to concentration units and output to a display device or data store.

The biological response of the biosensor is determined by the biocatalytic membrane which accomplishes the conversion of reactant to product. Immobilised enzymes possess a number of advantageous features which makes them particularly applicable for use in such systems. They may be reused, which ensures that the same catalytic activity is present for a series of analyses. This is an important factor in securing reproducible results and avoids the pitfalls associated with the replicate pipetting of free enzyme otherwise necessary in analytical protocols. Many enzymes are intrinsically stabilised by the immobilisation process, but even where this does not occur there is usually considerable apparent stabilisation. It is normal to use an excess of the enzyme within the immobilised sensor system. This gives a catalytic redundancy (i.e. $h \ll 1$) which is sufficient to ensure an increase in the apparent stabilisation of the immobilised enzyme. Even where there is some inactivation of the immobilised enzyme over a period of time, this inactivation is usually steady and predictable. Any activity decay is easily incorporated into an analytical scheme by regularly interpolating standards between the analyses of unknown samples. For

these reasons, many such immobilised enzyme systems are re-usable up to 10,000 times over a period of several months. Clearly, this results in a considerable saving in terms of the enzymes' cost relative to the analytical usage of free soluble enzymes.

When the reaction, occurring at the immobilised enzyme membrane of a biosensor, is limited by the rate of external diffusion, the reaction process will possess a number of valuable analytical assets. In particular, it will obey the relationship shown in equation. It follows that the biocatalyst gives a proportional change in reaction rate in response to the reactant (substrate) concentration over a substantial linear range, several times the intrinsic K_m. This is very useful as analyte concentrations are often approximately equal to the K_ms of their appropriate enzymes which is roughly 10 times more concentrated than can be normally determined, without dilution, by use of the free enzyme in solution. Also following from equation which is the independence of the reaction rate with respect to pH, ionic strength, temperature and inhibitors. This simply avoids the tricky problems often encountered due to the variability of real analytical samples (e.g, fermentation broth, blood and urine) and external conditions. Control of biosensor response by the external diffusion of the analyte can be encouraged by the use of permeable membranes between the enzyme and the bulk solution. The thickness of these can be varied with associated effects on the proportionality constant between the substrate concentration and the rate of reaction (i.e. increasing membrane thickness increases the unstirred layer (δ) which, in turn, decreases the proportionality constant, k_L, in equation. Even if total dependence on the external diffusional rate is not achieved (or achievable), any increase in the dependence of the reaction rate on external or internal diffusion will cause a reduction in the dependence on the pH, ionic strength, temperature and inhibitor concentrations.

Unit 5 ❐ Detection of Apoptotic cell

Structure

5.1 Comet assay

The Comet-assay is a single cell based technique that allows to detect and quantitate DNA damages. Therefore the assay uses nuclei embedded in agarose and exposed to an electric field. This technique was first described in 1984 by Johanson and Oesling and was later modified by Singh and coworkers in it's alkaline version to detect single strand DNA breaks.

Principle

The nuclear DNA of normal cells consists of few and very long molecules, but the nuclear DNA of cells undergoing apoptosis degrade in to many pieces of very small, small and medium sizes.

When whole cell electrophoresis is carried out, very long nuclear DNA molecules of normal cells exist in the form of a single large mass, but heterogeneous nuclear DNA fragments of apoptotic cells align in a smaller to bigger pattern, which nearly assumes the shape of a comet.

Experiment

a) Control or normal cells are mixed with a high quality soft agarose (cell : agarose = 1 : 10) and a suspension is obtained by adding PBS, pH 7.4 to the mixture

b) Similarly, late apoptotic cells (cells exposed to drug, for example Staurosporin 1μ mol for 6 hours are mixed with agarose and a suspension is obtained by adding PBS to the mixture

c) These two suspension are dropped on two groves over a Comet slide

d) Two electrodes are connected to the two ends of the slide and a week electric current (17V) is passed between the two electrodes for a while (10 min)

e) Thereafter the DNA of the cells are stained with SYBR green dye.or DAPI or EtBr (Ethidium Bromide)

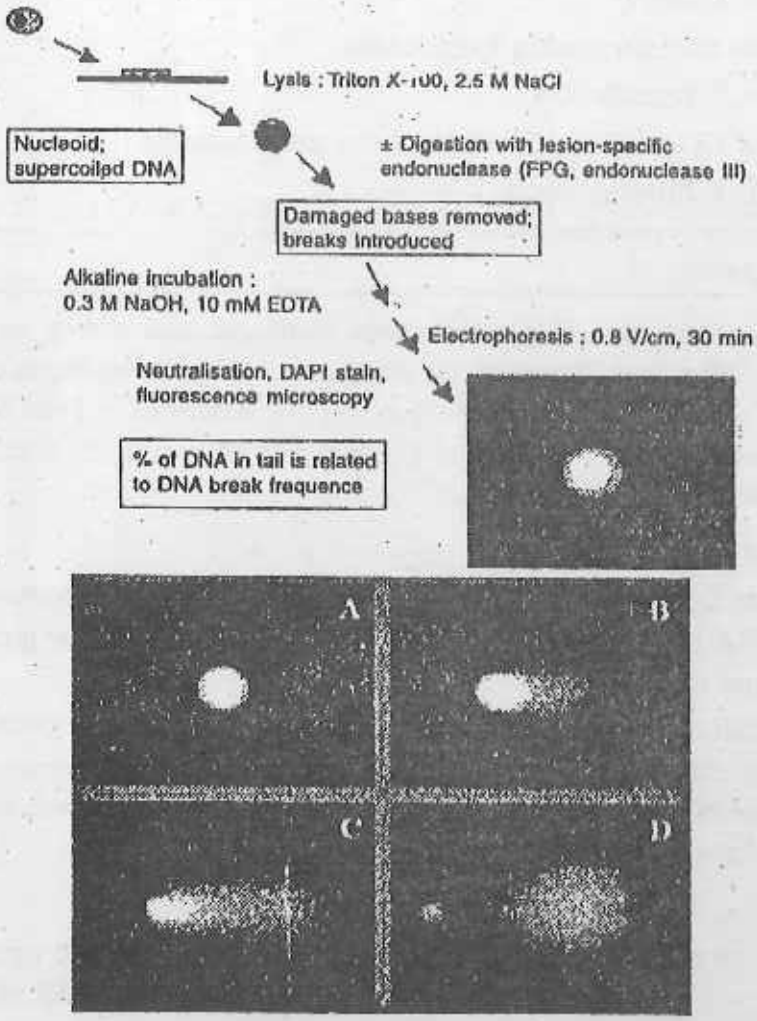


Fig. 5.1 : The comet assay – with enzyme modification

Observations

a) After the brief electrophoresis the nucleus of a control cell retains its rounded shape, as the very large DNA molecules having high molecular weights exist together as a single large mass.

b) On the other hand, the DNA fragments of an apoptotic cell having various sizes and very low, low and medium molecular weights tend to migrate towards the anode (+ ve electrode), the smallest DNA fragment shows the fastest mobility and the larger bigger fragments show a much slower mobility. Thus the orientation of very small, small and medium sized DNA fragments look like a comet and the nucleus tapers towards the anode.

Inferences

The study reveals that the nuclear DNA undergoes gradual fragmentation into very small small and medium sized molecules during apoptosis.

5.2 Nuclear lamin assay

The nuclear lamina is a dense, ~ 30 to 100 nanometers thick, fibrillar network composed of intermediate filaments made of lamin that lines the inner surface of the nuclear envelope in animal cells.

When the cell begins to divide, the nuclear lamina and envelope disappears. After the chromosomes have separated into two daughter cells, the lamina reforms around the chromosomes, and eventually combines back into the nuclear envelope. The Nuclear lamina, or the inner nuclear membrane (INM), is a scaffold-like network of protein filaments surrounding the nuclear periphery. This scaffold is made of mostly the type V intermediate filament proteins, lamin A/C and B, which together form a complex meshwork underneath the INM .The lamins are coiled-coil structures that contain a small N-terminal head followed by a rod-like domain (coiled-coil) and a C-terminal globular tail. Via these coiled-coil regions lamins can form parallel dimers, which in turn form polymers with other lamin dimers in an anti-parallel manner (head-to-tail). Although quite resistant to biochemical extraction, the nuclear lamina is nonetheless dynamic and can depolymerise

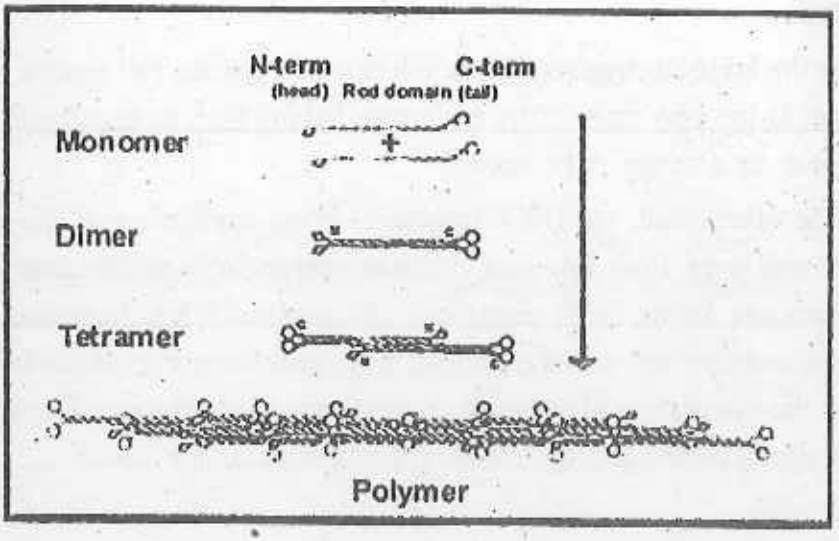


Fig. 5.2 : Formation of Lamin polymer

5.3 Lamin B assay

Principle

Lamin B is a filamentous protein which is essential for maintenance of integrity of nuclear membrane of the nuclei of eukaryotic cells. Lamin B of normal cells has a molecular weight of 70 KDa. In cells undergoing apoptosis nuclear membranes tend to lose integrity and this phenomenon is accompanied by degradation of laminB molecules in to < 50 KDa sized peptides that can no longer maintain the integrity of nuclear membrane.

Experiment

Nuclear membrane proteins are isolated from (i) control or normal cells (ii) cells in an early stage of apoptosis (cells exposed to 1μ mol concentration of staurosporin for 3 hours) and (iii) cells in an advanced stage of apoptosis (exposed to 1μ mol concentration of staurosporin for 6 hours). Staurosporin induced apoptotic changes in normal cells kept *in vitro*.

The isolated nuclear membrane proteins are subjected to PAGE (polyacrylamide gel electrophoresis) followed by Western Blot analysis using anti lamin B antibody (monoclonal). A lane of size marker protein is also used during the study.

Observations

i. In case of nuclear membrane proteins of control cells, anti Lamin B binding indicated the appearance of one distinct band on the Western Blot. The size markers indicate that the band has a molecular weight of 70 KDa.

ii. In case of nuclear proteins of early apoptotic cells, anti lamin B binding indicates the appearance of two discrete, narrow bands on the western blot. The size markers indicate that one of these bands has a molecular weight of 70KDa and the other has molecular weight of < 50 KDa.

iii. In case of nuclear membrane protein of late apoptotic cells, anti laminB binding indicates the appearance of one distinct band on the western blot. Its molecular weight as indicated by the size markers is < 50 KDa.

Inference

The study clearly indicates

i. Lamin B of control cells is a 70 KDA protein which appears as a distinct band on Western blot.

ii. In early apoptotic cells, all lamin B molecules are degraded in to <50 KDa sized peptides. Consequently, a distinct band appears near 50 KDa region on the Western blot but no band is formed at 70 KDa region

iii. In late apoptotic cells, all lamin B molecules are degraded in to < 50 KDa sized peptides. Consequently, a distinct band appears near 50 KDa region on the Western blot but no band is formed at 70 KDa

Control cells	Early apoptosis	Late apoptosis
70 KDa	70KDa 50KDa	50KDa

Fig. 5.3 : Bands of Lamin B of control & apoptotic cells on Western blot

5.4 Caspase activation

Caspases are a family of calcium-dependent cysteine proteases, which play essential roles in apoptosis (programmed cell death), necrosis and inflammation. Caspases are essential in cells for apoptosis, one of the main types of programmed

cell death in development and most other stages of adult life, and have been termed "executioner" proteins for their roles in the cell. The importance of caspases to apoptosis and programmed cell death was originally established by Robert Horvitz and colleagues

5.5 Caspases and apoptosis

The caspases are a family of proteins that are one of the main executors of the apoptotic process. They belong to a group of enzymes known as cysteine proteases and exist within the cell as inactive pro-forms or zymogens. These zymogens can be cleaved to form active enzymes following the induction of apoptosis. Induction of apoptosis via death receptors typically results in the activation of an initiator

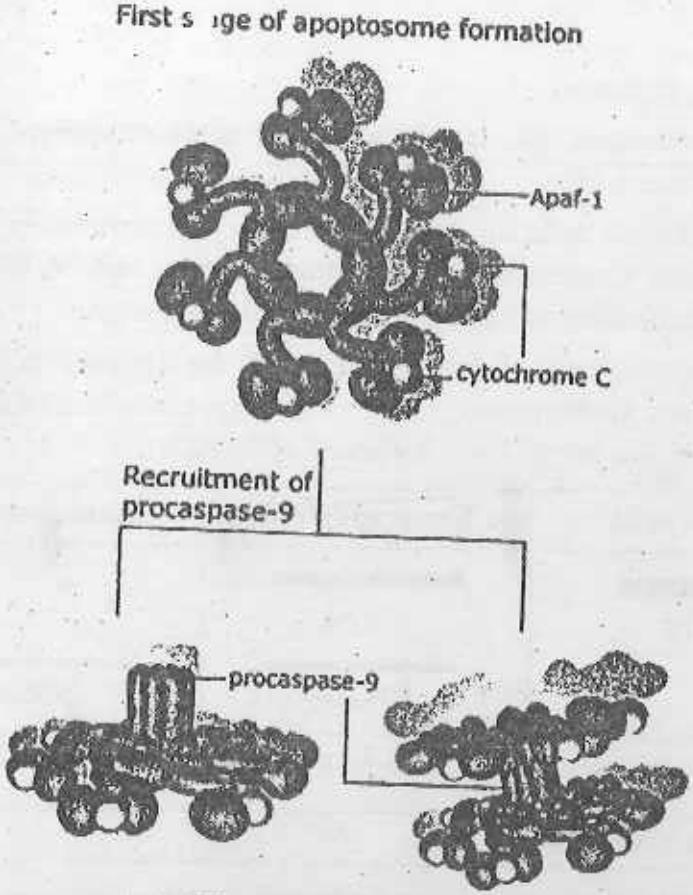


Fig. 5.4 : Caspase Activation

caspase such as caspase 8 or caspase 10. These caspases can then activate other caspases in a cascade. This cascade eventually leads to the activation of the effector caspases, such as caspase 3 and caspase 6. These caspases are responsible for the cleavage of the key cellular proteins, such as cytoskeletal proteins, that leads to the typical morphological changes observed in cells undergoing apoptosis.

5.6 The apoptosome

There are a number of other mechanisms, aside from activation of the death receptors, through which the caspase cascade can be activated. Granzyme B can be delivered into cells by cytotoxic T lymphocytes and is able to directly activate caspases 3, 7, 8 and 10. The mitochondria are also key regulators of the caspase cascade and apoptosis. Release of cytochrome C from mitochondria can lead to the activation of caspase 9, and then of caspase 3. This effect is mediated through the formation of an apoptosome, a multi-protein complex consisting of cytochrome C, Apaf-1, pro-caspase 9 and ATP. The formation of the apoptosome is illustrated below (Fig. 5.5b).

5.7 Caspases and chromatin breakdown

One of the hallmarks of apoptosis is the cleavage of chromosomal DNA into nucleosomal units. The caspases play an important role in this process by activating DNases, inhibiting DNA repair enzymes and breaking down structural proteins in the nucleus. This processes is illustrated below.

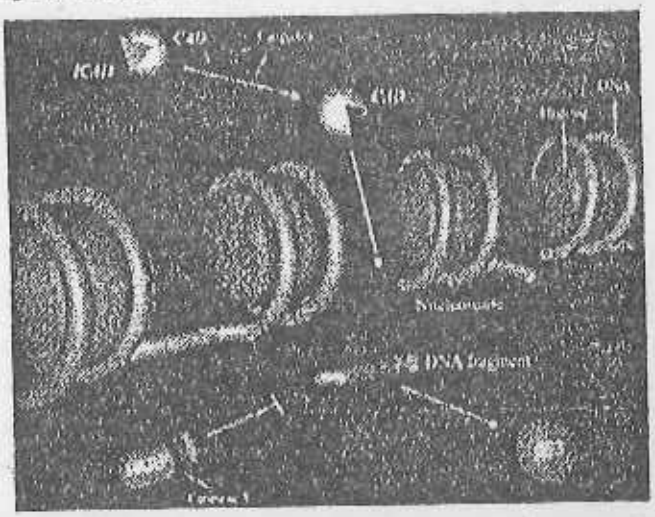


Fig. 5.5(a) : Figure showing the caspases and chromatin breakdown

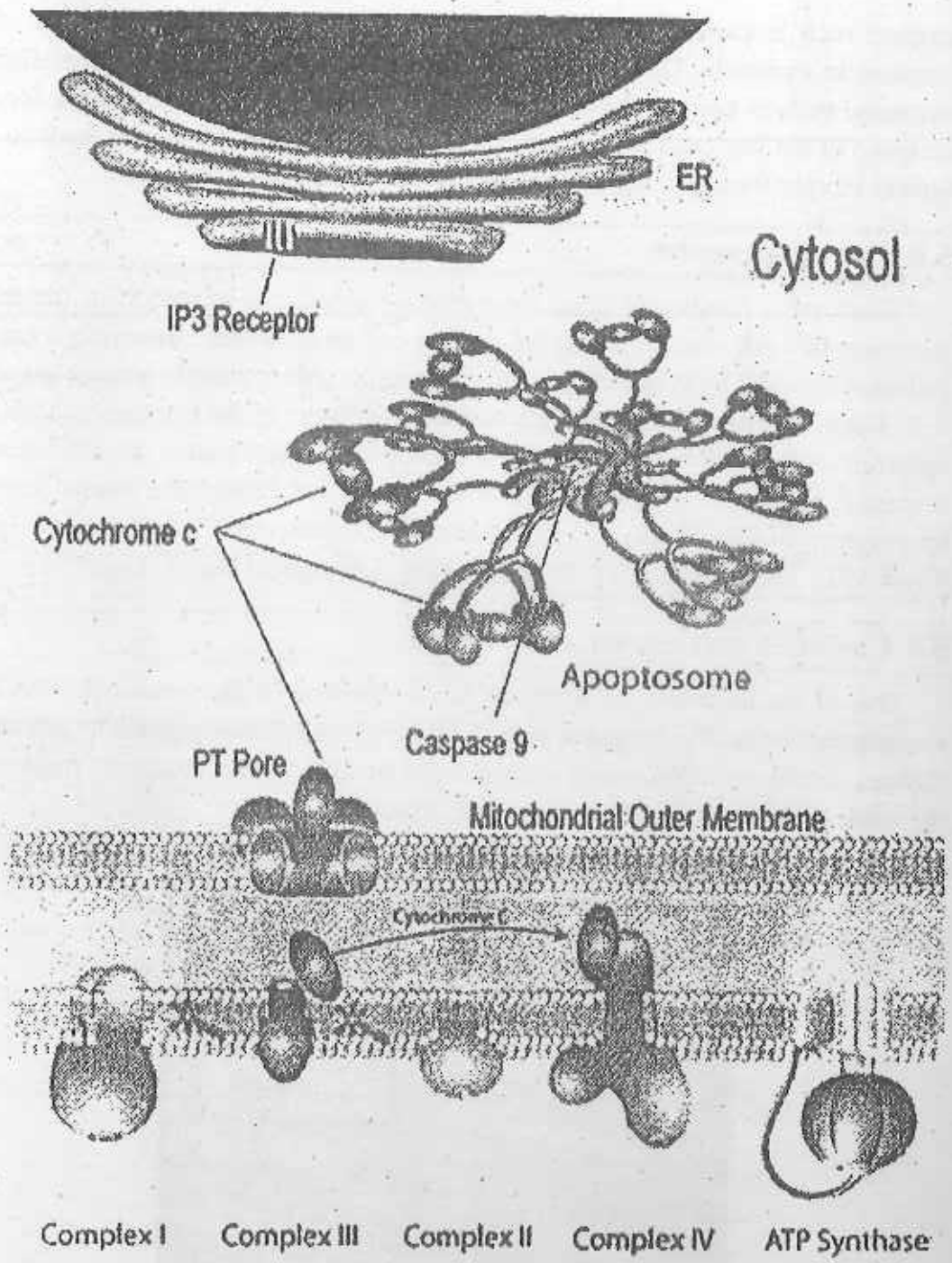


Fig. 5.5(b). Formation of Apoptosome

1) Inactivation of enzymes involved in DNA repair

The enzyme poly (ADP-ribose) polymerase, or PARP, was one of the first proteins identified as a substrate for caspases. PARP is involved in repair of DNA damage and functions by catalyzing the synthesis of poly (ADP-ribose) and by binding to DNA strand breaks and modifying nuclear proteins. The ability of PARP to repair DNA damage is prevented following cleavage of PARP by caspase-3.

2) Breakdown of structural nuclear proteins

Lamins are intra-nuclear proteins that maintain the shape of the nucleus and mediate interactions between chromatin and the nuclear membrane. Degradation of lamins by caspase 6 results in the chromatin condensation and nuclear fragmentation commonly observed in apoptotic cells.

3) Fragmentation of DNA

The fragmentation of DNA into nucleosomal units - as seen in DNA laddering assays–is caused by an enzyme known as CAD, or caspase activated DNase. Normally CAD exists as an inactive complex with ICAD (inhibitor of CAD). During apoptosis, ICAD is cleaved by caspases, such as caspase 3, to release CAD. Rapid fragmentation of the nuclear DNA follows.

5.8 Caspase 3 degradation

Principle

Procaspase 3 found in the cytosol of normal cells represents an inactive precursor of the enzymes caspase3. Procaspase 3 has a molecular weight of 32 KDa. With the onset of apoptosis of a cell, procaspase 3 breaks in two fragments, a 20 KDa pepteide and a 12 KDa peptide. The larger peptide functions as the active caspase 3, which initiates the cascade of apoptotic changes of cells.

Experiments

Cytosol fractions are isolated from (a) control or normal cells (b) cells in an early stage of apoptosis (ii) cells exposed to 1μ mol concentration of saturosporin for 3 hours and (c) cells in an advanced satge of apoptosis (exposed to 1μ mol concentration of staurosporin for 6 hours). Saturosporin induces apoptotic changes in normal cells kept *in vitro*.

The isolated cytosol fraction are subjected to PAGE, followed by Western Blot analysis using an anti procaspase 3 monoclonal antibody. A lane of size marker proteins is also used during the study.

Obeservation

i) In case of cytosol fraction of control cells, anti procaspase3 binding indicates the appearance of one distinct band on the Western Blot. The size markers indicates that the band has molecular weight of 32 KDa

ii) In case of cytosol fraction of early apoptotic cells, anti procaspase 3binding indicates the appearance of three distinct bands on the western blot. The size markers indicate that the bands has a molecular weight of 32 KDa, 20 KDa, 12 KDa.

iii) In case of cytosol fraction of late apoptotic cells, anti procaspase 3 binding indicates the appearance of two distinct bands on the western blot. As indicated by the size markers, one band has a molecular weight of 20 KDa and the other band has a molecular weight of 12 Kda

Inference

1. Procaspase 3 of control cells is a 32 KDa protein, which appears as a distinct band on the western blot
2. In early apoptotic cells, a part of procaspase 3 remain intact and appears as a band at 32KDa region on the western blot and another part degrades in to 20KDa peptides and a 12 KDa peptide, which appears as a narrow band and a faint band on the western blot.
3. In late apoptotic cells, all procaspase 3 molecules are degraded in to 20 KDa and 12Kda peptides. Consequently one band appears at 20KDa region and another band at 12Kda region on the western blot.

Control cells	Early apoptosis	Late apoptosis
32 KDa	32 KDa	
	20 KDa	20 KDa
	12 KDa	12 KDa

Fig. 5.6 : Bands of Procaspase 3 of normal & apoptotic cells on Western Blot

5.9 Assay of Cytochrome C release from mitochondria

Introduction

Cytochrome *c*, or cyt *c* is a small heme protein found loosely associated with the inner membrane of the mitochondrion. It is a soluble protein, unlike other cytochromes, and is an essential component of the electron transfer chain, where it carries one electron. It is capable of undergoing oxidation and reduction, but does not bind oxygen. It transfers electrons between Complexes III and IV. It belongs to cytochrome c family of proteins. Cytochrome c is a highly conserved protein across the spectrum of species, found in plants, animals, and many unicellular organisms. This, along with its small size (molecular weight about 12,000 daltons), makes it useful in studies of cladistics. Its primary structure consists of a chain of 100 amino acids.The cytochrome *c* molecule has been studied for the glimpse it gives into evolutionary biology. Both chickens and turkeys have the identical molecule (amino acid for amino acid) within their mitochondria, whereas ducks possess molecules differing by one amino acid. Similarly, both humans and chimpanzees have the identical molecule, while rhesus monkeys possess cytochromes differing by one amino acid.

Principle

When mitochondrial fraction and cytosol fraction are isolated from normal cells by differential centrifugation, the presence of an enzyme cytochrome C can be

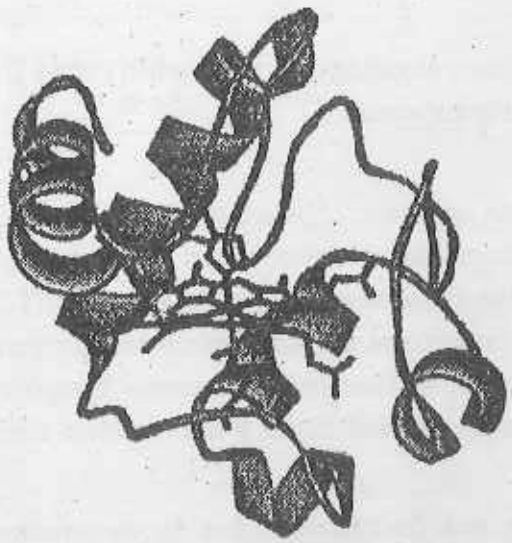


Fig. 5.7 : Cytochrome c, somatic

detected exclusively in the mitochondrial fraction. On the other hand when the aforesaid fraction are isolated from cells undergoing apoptosis a significant quantity of cytochrome C can be detected in the cytosol fraction and only a small quantity can be detected in the mitochondrial fraction. It is likely that cytochrome C after its redistribution from mitochondria to cytosol helps in caspase activation which in turn bring about apoptotic changes of cells.

Experminent I

Mitochondrial fraction and cytosol fraction are isolated by differential ultracentrifugation from (a) control or normal cells (b) late apoptotic cells (cells exposed to 1μ mol concentration of staurosporin for 6 hours). Saturosporin induces apoptotic changes. The isolated fraction were then subjected to polyacrylamide gel electrophoresis (PAGE) followed by Western Blot analysis using an anti cytochrome C antibody.

Observations

(a) Anti cytochrome C binding indicates the appearance of a single distinct band on the western blot in case of the mitochondrial fraction but not the cytosol fraction collected from normal cells (b) Again anti cytochrome C binding indicates the appearance of one narrow band and one distinct band on western blot in case of the mitochondrial fraction and cytosol fraction, respectively collected from the apoptotic cells.

Inference

The study reveals that cytochrome C gets redistributed from the mitochondria to cytosol of cells during apoptosis.

Experiments II

Control cells and late apoptotic cells are taken and each kind of cells is stained with two fluorescent compounds (i) mitotracker : a dye that gives red fluorescence and specifically stains the mitochondria of cells and (ii) FITC anticytochrome C : a conjugate of a green fluorescent dye fluorescein isothiocyante and a monoclonal antibody against cytochrome C, the whole fluorescent complex specifically binds to cytochrome C. All cells are studied under a fluorescence microscope.

Observations

(i) In control cells red fluorescence due to mitotracker is located on the mitochondrial surface and green fluorescence due to FITC anticytochrome C is

located in the interior of the mitochondria. (ii) In apoptotic cells, red fluorescence is located on the mitochondrial surface while green fluorescence is located principally in the cytosol instead of the mitochondrial interior.

Inference

The study reveals that cytochrome C gets redistributed from the mitochondria to the cytososl of cells during apoptosis. This is why the green fluorescent FITC anticytochrome C binds cytochrome C in the cytosol of apoptotic cells.

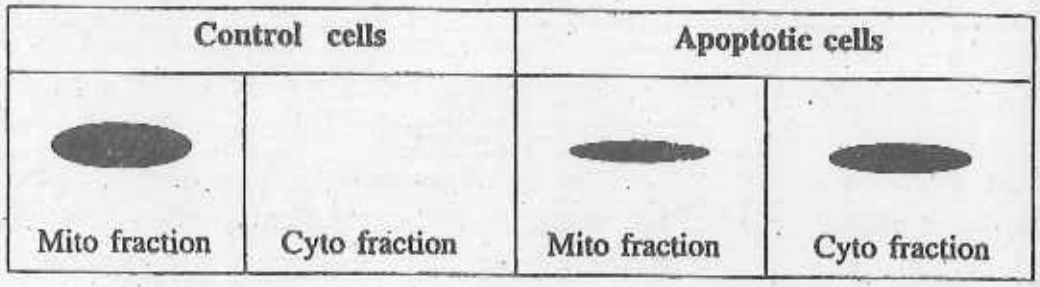


Fig. 5.8 : Cytochrome C bands of normal & apoptotic cells on western blot.

5.10 Annexin V binding assay

Introduction

Apoptosis is a cell death process characterized by morphological and biochemical features occurring at different stages. Once triggered, apoptosis proceeds with different kinetics depending on cell types and culminates with cell disruption and formation of apoptotic bodies. A critical stage of apoptosis involves the acquisition of surface changes by dying cells that eventually results in the recognition and the uptake of these cells by phagocytes. Different changes on the surface of apoptotic cells such as the expression of thrombospondin binding sites, loss of sialic acid residues and exposure of a phospholipid-like phosphatidylserine (PS) are the notable one. Phospholipids are asymmetrically distributed between inner and outer leaflets of the plasma membrane with phosphatidylcholine and sphingomyelin exposed on the external leaflet of the lipid bilayer, and phosphatidylserine predominantly observed on the inner surface facing the cytosol. Exposure of PS on the external surface of the cell membrane has been reported for activated platelets and senescent erythrocytes. Recently, it was shown that cells undergoing apoptosis break up the phospholipid asymmetry of their plasma membrane and expose PS which is translocated to the outer layer of the membrane. This occurs in the early phases of apoptotic cell death

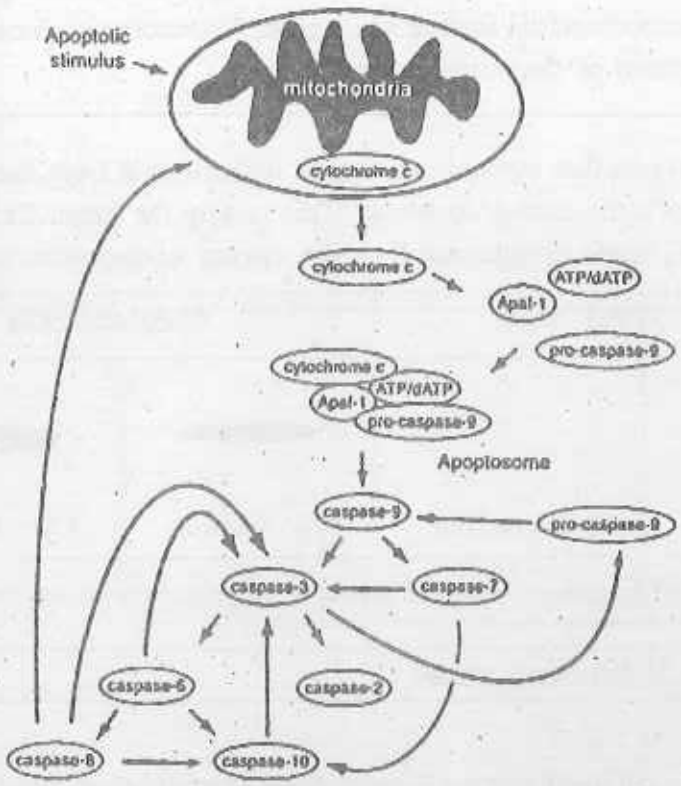


Fig. 5.9 : Apoptotic signal liberates cytochrome C from mitochondria

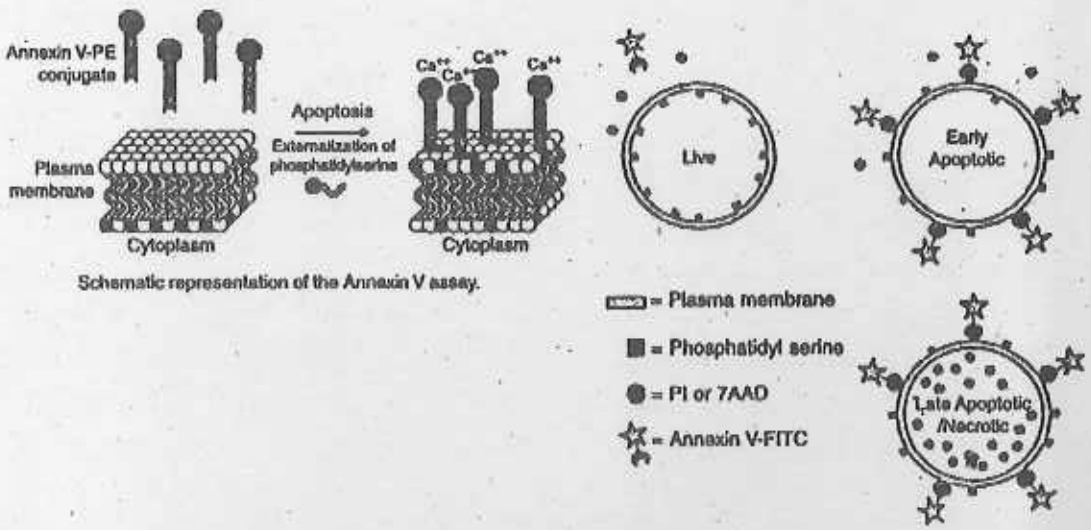
during which the cell membrane remains intact. This PS exposure may represent a hallmark (early and widespread) in detecting dying cells. **Annexin V,** belonging to a recently discovered family of proteins, the annexins, with anticoagulant properties has proven to be a useful tool in detecting apoptotic cells since it preferentially binds to negatively charged phospholipids like PS in the presence of Ca^{2+} and shows minimal binding to phosphatidylcholine and sphingomyeline. Changes in PS asymmetry, which is analyzed by measuring Annexin V binding to the cell membrane, were detected before morphological changes associated with apoptosis have occurred and before membrane integrity has been lost. By conjugating FITC to Annexin V it is possible to identify and quantitate apoptotic cells on a single-cell basis by flow cytometry. Staining cells simultaneously with FITC-Annexin V (green fluorescence) and the non-vital dye propidium iodide (red fluorescence.)

Principle

The plasma membrane of a eukaryotic cell contains two leaflets of lipid molecules. Different lipid molecules remains asymmetrically distributed in those two leaflets. In normal cells, phosphatidylcholine, sphingomyelin and glycolipid occur mainly in the outer leaflets but phosphatidylserine (PS) and phosphatidylethanolamine occur mainly in the inner leaflets while asmall amount of cholesterol occurs in either of the two leaflets. Amongst all these lipid molecules, only PS carries a net negative charge while others are electrically neutral at physiological PH. With the onset of apoptosis, PS molecules are redistributed from the inner to outer leaflets of lipid bilayer and their negatively charged heads get exposed to the cell exterior. A protein called annexin V which remains positively charged at a PH above 7 can bind to the head of a PS molecules by an electrostatic bond.

Experiment

Control or normal early apoptotic and late apoptotic cells (cells exposed to sataurosporin for 3 & 6 hr respectively) are taken out from respective cultures. Now a protein called annexin V which has previously been tagged with a dye FITC is added to each group of cells. After a while the cells of different groups are examined under a fluorescence microscope using a green filter.



Schematic representation of the Annexin V assay.

Fig. 5.10. Experiment showing annexin V binding assay

Observation

The outer surface of late apoptotic cells shows an intense green fluorescence and the early apoptotic cells show a pale green fluorescence surface, but the control cells do not show any fluorescence on their surface.

Inference

The experiment clearly indicates that FITC – annexin V conjugate does not bind to the cell membrane of control cells, but can bind with the cell membrane of apoptotic cells, especially the late apoptotic cells. Therefore it becomes evident that the PS molecules pass from the inner to the outer lipid leaflets of the cell membrane when a cell undergoes apoptosis.

Unit 6 ❐ Elementary Idea of Cryopreservation

Structure

6.1 Introduction

The word Cryopreservation is a combination of two words- Cryo + Preservation. The term *cryo* comes from the Greek word ***kruos***, which means frost and thus it stands for frosting or it can be expressed as cooling to frosting. So the technique that employs cooling to very low temperatures to preserve the biological specimen is known as cryopreservation. In Cryopreservation, cells or whole tissues are preserved by cooling to low sub-zero temperatures, such as (typically) 77 K or –196°C (the boiling point of liquid nitrogen). At these low temperatures, any biological activity, including the biochemical reactions that would lead to cell death, is effectively stopped.

Cryobiology succeeds in preserving living systems at cryogenic temperatures by limiting the destructive effects of low temperature during cooling and to let the

protective effects dominate, over virtually unlimited period of time. The limit of hazardous sub-zero temperature depends on the biological system being preserved and the method of preservation.

6.2 History

James Lovelock was one of the most important early workers on the theory of cryopreservation. Dr. Lovelock's work suggested that damage to red blood cells during freezing was due to osmotic stresses. Lovelock in early 1950s had also suggested that increasing salt concentrations in a cell as it dehydrates to lose water to the external ice might cause damages to the cell.

6.3 Natural cryopreservation

Tardigrades, commonly known as Water bears, can survive freezing at low temperatures by replacing most of their internal water with the sugar trehalose. Sugars and other solutes that do not easily crystallize have the effect of limiting the stresses that damage cell membranes. Trehalose is a sugar that does not readily crystallize. Mixtures of solutes can achieve similar effects.

6.4 Freezable tissues

In general, cryopreservation is easier for thin samples and small clumps of individual cells, because these can be cooled more quickly and so require lower doses of toxic cryoprotectants.

- Semen (which can be used successfully almost indefinitely after cryopreservation),
- Human eggs (oocytes)
- Human embryos (that are 2, 4 or 8 cells when frozen)
- Blood (special cells for transfusion, or stem cells)
- Tissue samples (eg. tumours and histological cross sections)

6.5 Basic principle

Internal energy of a physical system can be measured from the temperature of the system. This energy in fluid systems allows the molecular motions from one place to other within the fluid. Falling of temperature leads to the reduction of the energy level and consequently the molecular motions within the system. In case of pure water, falling of temperature below certain point, results in an abrupt reorganization of water molecules into organized solid lattice i.e. ice crystal. This is called ***freezing***. In other systems temperature reduction causes progressive slowing of molecular motion, and chemical reactions rates until the critical temperature is reached below which energy is insufficient for appreciable mobility of molecules. This point is known as ***glass transition temperature***. At this point, the system almost completely loses its fluidity and becomes solid liquid or vitrified (*vitrous* means glass). In this phase molecular rearrangements are practically stopped. At several degrees below the glass transition temperature, molecular motions are slow to extreme and the changes are nil for all practical purposes even for several hundred years.

6.6 Risks

Phenomena which can cause damage to cells during cryopreservation are :

a. Solution effects,
b. Extracellular ice formation,
c. Dehydration and
d. Intracellular ice formation.

a. Solution effects

Solution effects caused by concentration of solutes in non-frozen solution during freezing as solutes are excluded from the crystal structure of the ice. High concentrations can be very damaging.

b. Extracellular ice formation

When tissues are cooled slowly, water migrates out of cells and ice forms in the extracellular space. Too much extracellular ice can cause mechanical damage due to crushing.

c. Dehydration

The migration of water causing extracellular ice formation can also cause cellular dehydration. The associated stresses on the cell can cause damage directly.

d. Intracellular ice formation

While some organisms and tissue can tolerate some extracellular ice, any appreciable intracellular ice is almost always fatal to cells.

6.7 Prevention of risks

In so-called 'slow cooling' methods, the biological material is cooled at a range of cooling rates that are fast enough to prevent 'slow cooling damage' but are slow enough to allow sufficient dehydration of the cells to prevent intracellular ice formation. The dehydrated cells in the 'unfrozen fraction' that remains between the masses of ice will ultimately reach a stable glassy state, or 'vitrify'. In so-called vitrification methods, the water content is lowered before cooling by adding high concentrations of cryoprotective agents (CPA). Thus, no ice is formed at all, and the entire sample will vitrify. This allows fast cooling rates without risk of intracellular ice formation. The CPA concentration of vitrification solutions can be minimised by using very high cooling and thawing rates. By using extremely high cooling rates, vitrification is possible even in complete absence of CPAs.

Vitrification provides the benefits of cryopreservation without the damage due to ice crystal formation. In clinical cryropreservation, vitrification usually requires the addition of cryoprotectants prior to cooling. The cryoprotectants act like antifreeze; they lower the freezing temperature. They also increase the viscosity. Instead of crystalizing, the solution turns into an amorphous ice—i.e. it vitrifies.

Generally, two conditions required to allow vitrification are

(i) an increase in the viscosity and

(ii) a depression of the freezing temperature.

Many solutes do both, but larger molecules generally have larger effect, particularly on viscosity. Rapid cooling also promotes vitrification.

In artificial cryopreservation, the solute must penetrate the cell membrane in order to achieve increased viscosity and depressed freezing temperature inside the cell. Sugars do not readily permeate through the membrane. Those solutes that do, such as dimethyl sulfoxide, a common cryoprotectant, are often toxic in high

concentration. One of the difficult compromises faced in artificial cryopreservation is limiting the damage produced by the cryoprotectant itself.

6.8 Cryoprotectant

Cryoprotectants are the substances which are used to protect biological tissues from freezing damage (damage due to ice formation). Practically these are the chemicals that tend to reduce injury to the cells during freezing and thawing.

Arctic and Antarctic insects, fish, amphibians and reptiles create cryoprotectants in their bodies to minimize freezing damage during cold winter periods. Insects most often use sugars as cryoprotectants. Arctic frogs use glucose and Arctic salamanders create glycerol in their liver for use as cryoprotectant.

Conventional cryoprotectants are glycols, such as ethylene glycol, propylene glycol and glycerol. Ethylene glycol is commonly used as automobile antifreeze and propylene glycol has been used to reduce ice formation in ice cream. Dimethyl sulfoxide (DMSO) is also regarded as a conventional cryoprotectant. Glycerol and DMSO have been used for decades to reduce ice formation in gametes and embryos that are cold-preserved in liquid nitrogen.

Successful cryopreservation requires that viability be maintained after having remained in liquid nitrogen for an indefinite period. Too rapid freezing will form large intracellular ice crystals that will break the membranes on thawing. Slow freezing forms extracellular ice, if intracellular concentration of solutes is sufficiently too high to freeze. This condition can be achieved by addition of penetrating cryoprotectants. When extracellular ice is formed the water is withdrawn from the cell, as the vapor pressure density of ice is lower than that of largely aqueous cytoplasm. Water keeps on coming out till vapor pressure densities reach equilibrium. This process shrinks the cytoplasm and the plant and microbial cells can undergo temporary plasmolysis.

Non-penetrating cryoprotectants (for eg. polyethelene oxide (PEO), polyvinyl pyrrolidone (PVP)) are the alternative choice to ensure a smooth dehydration during freezing. Rapid dehydration will denature the proteins by the formation of localized high concentration that precipitates proteins. At a critical temperature corresponding to the freezing points for equilibrium mixture of solid and liquid, is called ***eutectic point*** for the cytoplasm. At this temperature, the cell contents can freeze without formation of ice crystals.

Common Cryoprotectants are :

- Dimethyl sulfoxide (DMSO)
- Ethylene glycol
- Glycerol
- Propylene glycol
- Sucrose
- Trehalose

Depending on their ability to permeate across the cell membrane, they are broadly classified into two groups. *Penetrating cryoprotectants* can move into the cell across the membrane but *non-penetrating cryoprotectants* can not.

A. Mechanism of action of penetrating cryoprotectants

Mechanism of action of penetrating cryoprotectant is based on their colligative properties i.e. the collective properties that a solution has in the presence of these compounds. Particularly it is the reduction in salt concentration at a given temperature that reduces the injury at that temperature. Thus an effective penetrating cryoprotectant should provide colligative properties in which the salt is buffered down to low temperature and the protectant should be freely permeable across the cell membrane to buffer the intracellular salt as well.

Examples : Ethylene Glycol, Propylene glycol, Glycerol, Dimethyl sulfoxide

B. Mechanism of action of non-penetrating cryoprotectants

Non-Penetrating Cryoprotectants act by dehydrating cell at high sub-freezing temperatures, and thus allow them to rapidly cool, before the slow cooling injury can cause extensive damage. These compounds are usually polymers that form extensive hydrogen bonds with water, reducing the effects of water to a much greater extent than could be predicted by their molar concentration.

Example : Hydroxy-ethyl- starch (HES), Polyvinyl pyrrolidone (PVP) and Polyethylene oxide (PEO).

Mixtures of cryoprotectants have less toxicity and are more effective than single-agent cryoprotectants. A mixture of formamide with DMSO, propylene glycol and a colloid was for many years the most effective of all artificially created cryoprotectants. Cryoprotectant mixtures have been used for vitrification.

Some cryoprotectants function by lowering a solution's or a material's Glass transition temperature. In this way, the cryoprotectants prevent actual freezing, and the solution maintains some flexibility in a glassy phase. Many cryoprotectants also function by forming hydrogen bonds with biological molecules as water molecules are displaced. Hydrogen bonding in aqueous solutions is important for proper protein and DNA function. Thus, as the cryoprotectant replaces the water molecules, the biological material retains its native physiological structure (and function), although they are no longer immersed in an aqueous environment. This preservation strategy is most often observed in anhydrobiosis.

6.9 Steps for cryopreservation of cells and microorganisms

There are techniques to cryopreserve most of isolated cell types, microorganisms, specimens of plant origin, and small embryos..

Preparation of cells

Usually a cryoprotective additive is required for significant survival after thawing. Glycerol and DMSO are most common.

The cryoprotective agent (CPA) is diluted to proper concentration in the medium for suspending the cells. CPA like DMSO and glycerol will serve as a solvent for sodium chloride, the molar concentration of this salt has to be kept at the required level in CPA solution. The molecular weight of DMSO is 78g/mol and density is 1.1g/mol; 7.1 ml. is used in 100 ml. to get one mol/litre.

The CPA must be allowed to equilibrate within the cell before the freezing is begun. CPA is added at physiological temperatures then the cells are left for about 15 minutes. With large cells or at low temperatures or both, the equilibration period of even of one hour may be required.

Microbes grown in aerobic conditions show greater recovery after cryopreservation than those grown in anaerobic conditions. The density of cells suspended in the growth medium with CPA should be 10^7 cells /ml or more. The density of tissue culture cells in the suspension is to be 10^5 to 10^7 cells/ml. The cell line must be free from contaminants before cryopreservation.

Freezing

Small plastic vials with screw caps or snap lids are used for freezing. Plastic straws are used for sperm cells. For these cells requiring very rapid warming flame sealed glass ampoules may be used.

Protozoa and most of other animal cells are cooled at controlled rates for better survival. A lot of animal cells have optimum cooling rate between 0.1 to 10°C/min. Generally practiced rate is 1°C.

Controlled cooling can be done by following methods.

1. Liquid nitrogen freezer
2. Controlled alcohol bath
3. Multistep cooling
4. Freezing containers.

Ice formation occurs in a sample of bulk solution by formation of ice nuclei. Once ice is formed, continuous crystallization releases the latent heat. This heat warms the sample to the melting point. The freezing apparatus is to be programmed to maintain the temperature after seeding so that the latent heat is fully dissipated. With the lowering of temperature the maintenance of cooling rate remains no more important. Generally the sample once having dropped to –60°C can directly be placed in liquid nitrogen.

Storage

The storage temperature affects the length of successful storage. These two are inversely proportional. Lower is the storage temperature longer is the safe storage. For liquid nitrogen (–196°C) the limiting factor is background radiation, which may cause accumulation of DNA alterations. A suitably shielded freezer gives indefinite storage. The cryoprserved stored material should be exposed to higher temperature to minimum.

Thawing

It should be as rapid as possible, with the temperature being kept limited to 37°C. The best and safest way of thawing (warming) the sample is to immerse it in a 37°C water bath with gentle agitation. In order to increase the rate of warming, the sample is frozen in a vial with high surface area to volume ratio (for e.g. straws). After thawing the cell recovery has to be measured.

6.10 Viability assays

It is the 'ability of the treated sample to exhibit a specific function(s) expressed as the proportion of the same function exhibited by the same sample before treatment or an identical sample'.

1. **Subjective Assay** : When as assay is to be established for the first time the statistically significant data of observed parameters are less important than the general impression of the results. In order to avoid the possible individual variations among observers and also the systematic errors, the objective assay should be made only when an assay has been established.
2. **Qualitative Assay** : Here the results are descriptive or more theoretical than being numerical or mathematical expressions. Such assays are useful if quality evaluation is desired and the quantity measurement is either difficult or not possible. Qualitative evaluation needs degree of consistency and an experienced expertise as for example in pathological examinations.
3. **Quantitative Assay** : This may be of two types.
   A. Direct : This is a sort of direct assay of dose response relationship regarding a specific response.
   B. Indirect : The response is quantitatively measured with respect to specific dosage or treatment given to the sample.
4. **Binary assay** : When two outcomes are feasible, that is known as binary assay. It also minimizes the experimental errors or bias of the observer, as the two outcomes may show interdependence.

6.11 Cryopreservation protocols

Cryopreservation of cells

1. Cultures to be cryopreserved should be healthy, free from contamination, and should be maintained in log phase growth for several days before freezing.
2. Grow attaching cell culture to late log phase, trypsinize and centrifuge. Only centrifugation is necessary for freezing suspension cells.

3. Resuspend cells in sterile serum-containing culture medium containing 10% v/v dimethyl sulfoxide (DMSO). Work should proceed quickly to minimize the length of time the cells are exposed to DMSO in the liquid state. The highest purity DMSO should be used.

4 Place the appropriate volume and cell number into cryopreservation ampules. Plastic or glass ampules may be used. However, plastic ampules with external silicon seals function best when kept above liquid nitrogen temperature (i.e. in the vapor phase). Immersion into the liquid nitrogen phase can result in liquid nitrogen entering the ampule and the contents of the ampule spraying out during the defrosting procedure. If storage in liquid nitrogen is preferred, plastic ampules with internal "o" rings perform satisfactorily. Glass ampules offer the best results due to the secure seal and the rapidity with which the ampule can be defrosted, thereby allowing for higher culture viability. However, they can be inconvenient to use due to the requirements of flame sealing.

5. Place the ampules in a controlled-rate freezer and cool at a rate of 1°C/minute. If a controlled-rate freezing apparatus is not available, adequate results can be obtained by :

   a. Placing the ampule inside a 1-inch foam-insulated box and keeping the box at –70°C for 12 hours.

   b. Cooling the ampules in the liquid nitrogen phase using a liquid nitrogen canister insert.

   c. Placing the ampules in an isopropanol bath that is subsequently cooled in a –70°C freezer.

   d. Placing the ampules directly into a –20°C freezer for several hours and then transferring to a –70°C for further cooling.

   e. Placing the ampules directly into a –70°C freezer.

   The last two methods (d and e) are not ideal since the culture viability can be affected and result in the loss of sensitive populations. These methods should only be used when no other options are available.

After freezing, the ampules should be transferred to a liquid nitrogen-fill storage vessel. Prolonged storage at temperatures above –135°C will result in decreased viabilities.

6.12 Cryopreservation of oocytes and embryos

Freezing

1. Specimens to be frozen are placed into labeled cryovials containing 0.3 ml of cryopreservative medium (1.5 M 1,2- propanediol in modified DPBS for pronuclear oocytes) at room temperature. Cryovial caps are securely tightened.
2. Cryovials are held at room temperature for 30 minutes. At the end of this equilibration period, cryovials are loaded onto freezing cans within the biological freezer.
3. a. Start point = room temperature
   b. Temperature drops within the unit at a rate of –1° C/min until –6.5°C is reached.
   c. Freezer is held at –6.5° C for an additional 5 minutes.
   d. Manual seeding is performed by touching forceps to outside the cryovials.
   e. Freezer is held at –6.5° C for an additional 5 minutes.
   f. Temperature drops within the unit at a rate of –0.5° C/ min. until –80° C is reached.
   g. Cryovials are plunged into liquid nitrogen, loaded onto storage canes, and placed into appropriate storage tanks.

Thawing

a. Start point = –100° C
b. Cryovials are removed from storage tanks and loaded onto thawing canes within biologic freezing unit.
c. Cryovials are maintained at –100 °C for 5 minutes before proceeding.
d. Temperature warms at a rate of + 8.0 °C/min until room temperature is reached.
e. Cryovials are held at room temperature for 5 minutes.
f. Specimens are removed from cryovials and placed in 1.0 M 1, 2-propanediol solution in DPBS for 5 minutes at room temperature.
g. Specimens are transferred to 0.5 M 1, 2-propanediol solution in DPBS for 5 minutes at room temperature.
h. Specimens are transferred to 0.0 M 1, 2-propanediol solution in DPBS for 5 minutes at room temperature.

i. Specimens are transferred to equilibrated Ham's F-10 culture medium containing 15% CO_2 until cleavage pattern is established.

j. Preembryo transfer is scheduled after preembryo viability and regular cleavage pattern is established, usually within 24 hours of thaw.

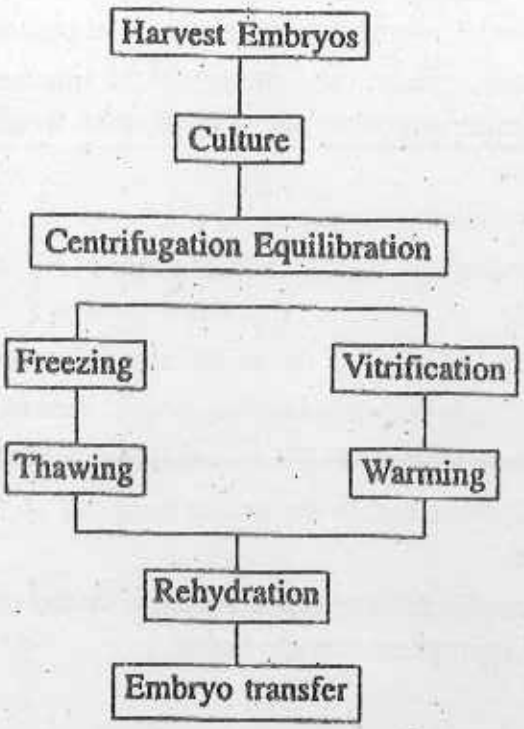


Fig. 6.1 Schematic representation of cryopreservation of embryo.

6.13 Cryopreservation of spermatozoa

Cryopreservation of spermatozoa by equilibrium cooling usually consists of the following steps;

1. Cells in suspension are placed in a solution of a cryoprotective agent (CPA)
2. The cells are cooled to temperatures near 0°C.
3. They are then cooled to subzero temperatures at a moderately low rate of about 5-10°C / min to an intermediate subzero temperature (about –75°C for spermatozoa) and then plunged into liquid nitrogen at –196°C for storage.
4. To restore their function, cryopreserved cells are warmed and thawed and the CPA removed.

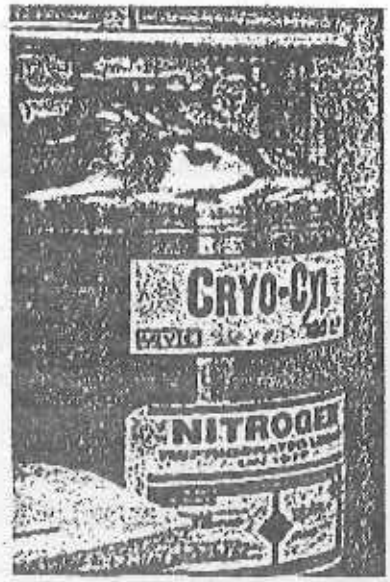


Fig. 6.2 A tank of *liquid nitrogen*, used to supply a cryogenic freezer (for storing laboratory samples at a temperature of about −150 degrees Celsius).

6.14 Cryopreservation of organs

There are three approaches for preservation of organs

1. Storage at hypothermic temperature that is above freezing but below 37°C.
2. Freezing and thawing and storage at cryogenic temperatures.
3. Storage at cryogenic temperatures without the formation of ice crystals during cooling or thawing.

A. Freezing of Organs

The conventional concept includes the application of techniques for preservation of cells with some alterations assuming that the organs are simply big cells. But there are many problems associated with this purpose because in an organ the different kind of cells are present, and each type has its own specific cryopreservation limits for optimal recovery. Extracellular ice can mechanically damage the organ, specially the blood vasculature. Osmotic movement of interstitial water may further cause mechanical stress. Mechanical fractures can occur in such frozen vitrious solids and the parts of organ may become separated. The attachments between cells and their basement membranes too can be disrupted.

B. Vitrification

It is the best process to preserve the whole organ. This process includes the conversion of aqueous solution into an amorphous solid. When a liquid is cooled down, it can be brought below its melting point without a change in phase. Formation of ice requires nucleation, in which ice nucleus as a cluster of few water molecules reaches the critical size. With lower temperatures, the critical size of nucleus becomes smaller and finally acquires the size of clusters within liquid. This solution can be cooled down to homogenous nucleation temperature, at which the probability of nucleation is 1 in a super cooled state, but below this crystallization begins.

Crystal growth requires molecular translation within the liquid. If a liquid is cooled at a fast rate the viscosity of liquid might increase to that level where molecular translation is much slow to allow nucleation and crystal growth as at this temperature the time required for nucleation becomes infinite i.e. the ice will practically be never formed. Thus obtained amorphous solid is stable below this temperature.

Vitrification is also possible by adding solutes, which form structures in water and ice will form only when these structures are broken down. The solute molecules interfere between water molecules that tend to cluster and so hydrogen-boniñg network necessary for ice formation is prevented from occurring.

C. Important aspects of cryopreservation of organs

Cooling injures the tissue, so there arise some problems in cryopreservation of organs. Those problems can be handled by applying following methods.

1. High concentration of vitrification solutions (VS) :
   a. The VS is injected through artery or vein of the organ under a computerized perfusion (infusion) system. The process is continually monitored.
   b. Glass forming chemical like propylene glycol may help to reduce concentration of VS.
   c. The atmospheric pressure is increased along with lowering of freezing point of the normal saline and the temperature of hypothermic storage is lowered. The time of VS perfusion is increased but the concentration is decreased by glass forming chemical.
2. The fracturing is prevented by removing the sample from its container after it has been cooled and by storing it at –135°C and in liquid nitrogen.
3. As rapid cooling is desired to prevent ice formation, the cooling is done with

suitable agitation and using perfusion, which is again not stationary to cool the organ.

4. During thawing, the rapid warming is done to prevent ice formation and this may be achieved by microwave heating.

Factors associated with cooling and cryopreservation that contribute to cellular injury and death in biological systems :

System	Type/ cause of damage
All	Intracellular ice formation, extracellular ice formation, apoptosis, toxicity, calcium imbalance, free radicals, ATP levels, general metabolism, fertilization failure, cleavage failure, intracellular pH, parthenogenetic activation, cleavage
Membrane	Rupture, leakage, fusion, microvilli, phase transition
Chromosomes	Loss/gain, polyspermy, polygyny (failure to extrude polar body), tetraploidy
DNA	Apoptosis, fusion, rearrangements
Cytoskeleton	Microtubules dissolve, actin
Proteins/enzymes	Dehydration, loss of function
Ultrastructure	Microvilli, mitochondria, vesicles, cortical granules, zona pellucida
Lipids	Free radicals?

(Ref : Cryopreservation of oocytes and ovarian tissue by Helen M. Picton, Roger G. Gosden, Stanley P.Leibo: Gamete source, manipulation and disposition)

6.15 Significance

Cryopreservation helps in protecting the specimen from any kind of deterioration.

Cryopreservation allows virtually indefinite storage of biological material without deterioration for at least several thousands of years.

Repeated cultures of cell lines may cause loss of biochemical functions or of morphogenetic properties due to genetic changes. Cryopreservation in liquid nitrogen is one of the methods that eliminate this problem in certain microbial cultures. Moreover, it is specially suited for eukaryotes.

Cryopreservation techniques enable the banking of cells to prevent any genetic drift of a culture line and also to renew a culture line in case that becomes

contaminated for being used in autologous or allogenic transplantation, for *in vitro* fertility treatment, animal husbandry and many other applications.

Cryopreservation of gametes (unfertilized oocytes and spermatozoa) and embryos is desirable to improve the efficiency of assisted reproduction. Stored oocytes could potentially be used for several purposes.

Applications are as follows :

1. Improve the efficiency of *In Vitro* Fertilization
2. Alternative to embryo freezing
3. Oocyte preservation for patients with ovarian hyperstimulation
4. Oocyte donation programme
5. The treatment of congenital infertility disorders
6. Prevent fertility loss through surgery
7. Treatment of premature ovarian failure

Unit 7 ❐ Molecular Separation Techniques

Structure

7.1 Chromatography

7.1.1. Introduction

The word *Chromatography* originates from two Greek words *khroma* (colour) and *graphia* (writing), and means coloured writing as the original experiments demonstrated the colour differences. Chromatography is one of the most effective techniques to separate and identify chemical compounds. Although the method was developed in 1906 by Russian botanist Mikhail Tswett, who separated plant pigments (hence the name), it was described in 1855 by Karl Runge, a German chemist who separated inorganic materials by paper chromatography. Chromatography became an important technique with the work of Archer Martin and R.L. M. Synge in 1944, who later received a Nobel Prize for developing the methodology of partition chromatography.

There are many kinds of chromatography–adsorption, partition, ion-exchange and molecular-sieve and many specialized techniques for using them- column, paper, thin-layer and gas chromatography.

7.1.2 Theoretical Basis of Chromatography

Liquid chromatography, a commonly used technique to separate mixtures of proteins, nucleic acids, and other molecules, is based on the principle that molecules dissolved in a solution will interact (bind and dissociate) with a solid surface. If the solution is allowed to flow across the surface, then molecules that interact frequently

with the surface will spend more time bound to the surface and thus move more slowly than molecules that interact infrequently with the surface. Liquid chromatography is performed in a tube packed tightly with spherical beads. The nature of these beads determines whether separation of proteins depends on differences in mass, charge or binding affinity. All types of chromatography are based on partition or distribution coefficient which is the expression of the way of the distribution of the compound in two immiscible phases, the mobile and the stationary ones. The coefficient value is expressed as :

$$\frac{\text{Concentration of phase I}}{\text{Concentration of phase II}} = K_d \text{ (the distribution constant)}$$

The stationary phase may be solid or a gel or an immobilized mixture of liquid and solid. The mobile phase may be a liquid of a gas as this phase flows through or on surface of the stationary phase. Depending on the nature of compounds to be separated, the two phases have to be chosen with regard to the method to be adopted.

7.2 Thin-Layer chromatography (TLC)

Thin layer chromatography (TLC) is a chromatography technique used to separate chemical compounds. It involves a *stationary phase* consisting of a thin layer of adsorbent material, usually silica gel plates (SiO_2) or aluminium oxide(Al_2O_3), or cellulose immobilised onto a flat, inert carrier sheet. A *liquid phase* consisting of the solution to be separated is then dissolved in an appropriate solvent and is drawn up the plate via capillary action, separating the experimental solution based on the polarity of the components of the compound in question.Unlike column chromatography where the elution is through the solid phase, here in Thin-Layer chromatography the samples move on the flat surface and the different fractions get distributed as separated from each other and remain adhered to the surface instead of being discharged out as fractional effluents. The distance travelled by each component is found out by applying a suitable colouring agent so that each component spot is visualized. The ratio of these distances is known as R_f value. For a given compound the R_f value is characteristic with a given solvent under controlled solute concentration, pH and temperature. The rate of migration of various substances being separated depends on their relative solubilities in the polar stationary and non-polar mobile phase. In single step of separation, a given solute is distributed between

the mobile and the stationary phases with reference to partition coefficient. Migration rate of a substance can be expressed as :

The R_f value—R_f is the retention factor, or how far up a plate the compound travels.

R_f = distance traveled by the substance/distance traveled by the solvent front.

Thin-Layer preparation and Chromatogram development:

Approximately 0.25 mm. thick layer of a solid material is spread on a glass or plastic plate and used as an ideal support for the fractional movement of components of an applied sample. If the thickness is below 0.2 mm the R_f value is affected. Various commercially available spreading materials can be carefully applied to form an evenly thin layer of required thickness. The atmosphere of the separation chamber must be fully saturated with solvent vapours to prevent the variations in R_f value. This is very quick method completing many separations in less than one hour. The separated spots are quite compact to allow detection of even lower concentrations. The separated compounds can be detected either by spray (Ninhydrin in case of amino acids) or a suitable agent or by scanning if the sample is radioactive labeled.

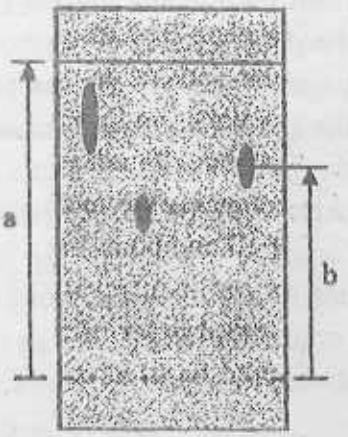


Fig. 7.1 Figure showing how Rf value is calculated

7.2.2. Plate preparation

TLC plates are made by mixing the *adsorbent*, such as silica gel G 20, with a small amount of inert binder like calcium sulfate (gypsum) and water. This mixture is spread as a thick slurry on an unreactive carrier sheet, usually glass, thick aluminum foil, or plastic, and the resultant plate is dried and *activated* by heating in an oven for thirty minutes at 110 °C. The thickness of the adsorbent layer is typically around 0.1–0.25 mm for analytical purposes and around 1–2 mm for preparative TLC.

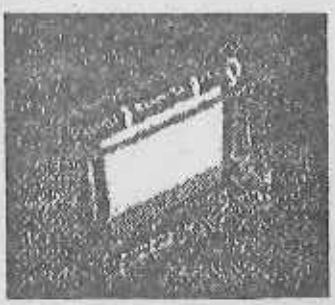


Fig. 7.2 A typical Silica G TLC plate in solvent

7.2.3 Technique

A small spot of solution containing the sample is applied to a plate, about one centimeter from the base.

The plate is then dipped into a suitable solvent, such as ethanol or water, and placed in a sealed container. The solvent moves up the plate by capillary action and meets the sample mixture, which is dissolved and is carried up the plate by the solvent. Different compounds in the sample mixture travel at different rates owing to differences in solubility in the solvent, and owing to differences in their attraction to the stationary phase.

Separation of compounds is based on the competition of the solute and the mobile phase for binding places on the stationary phase. For instance, if normal phase silica gel is used as the stationary phase it can be considered polar. Given two compounds which differ in polarity, the most polar compound has a stronger interaction with the silica and is therefore more capable to dispel the mobile phase from the binding places. Consequently, the less polar compound moves higher up the plate (resulting in a higher Rf value). If the mobile phase is changed to a more polar solvent or mixture of solvents, it is more capable of dispelling solutes from the silica binding places and all compounds on the TLC plate will move higher up the plate. Practically this means that if you use a mixture of ethyl acetate and heptane as the mobile phase, adding more ethyl acetate results in higher Rf values for all compounds on the TLC plate. Changing the polarity of the mobile phase will not result in reversed order of running of the compounds on the TLC plate. If a reversed order of running of the compounds is desired, an a polar stationary phase should be used, such as C18-functionalized silica.

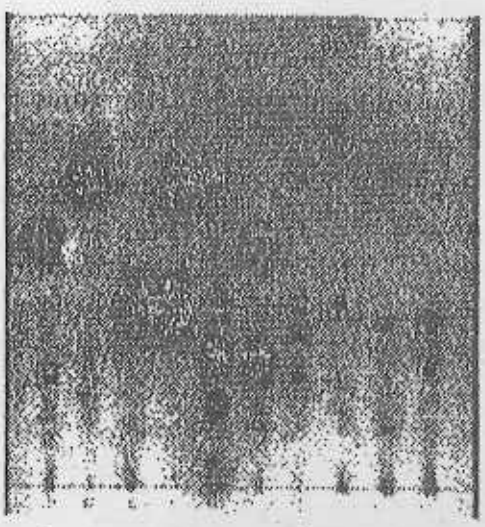


Fig. 7.3 Figure showing a typical chromatogram of ten essential oils, spot developed by vanillin

As the solvent moves past the spot that was applied, an equilibrium is established for each component of the mixture between the molecules of that component which are adsorbed on the solid and the molecules which are in solution. In principle, the components will differ in solubility and in the strength of their adsorption to the adsorbent and some components will be carried farther up the plate than others. When the solvent has reached the top of the plate, the plate is removed from the developing chamber, dried, and the separated components of the mixture are visualized. If the compounds are colored, visualization is straightforward. Usually the compounds are not colored, so a UV lamp is used to visualize the plates. (The plate itself contains a fluor which fluoresces everywhere except where an organic compound is on the plate.)

7.2.4 Analysis

As the chemicals being separated may be colorless, several methods exist to visualize the spots :

- ★ Often a small amount of a fluorescent compound, usually Manganese-activated Zinc Silicate, is added to the adsorbent that allows the visualization of spots under a blacklight (UV_{254}). The adsorbent layer will thus fluoresce light green by itself, but spots of analyte quench this fluorescence.
- ★ Iodine vapors are a general unspecific color reagent.
- ★ Specific color reagents exist like Ninhydrin into which the TLC plate is dipped or which are sprayed onto the plate

Once visible, the R_f value , or Retention factor, of each spot can be determined by dividing the distance traveled by the product by the total distance traveled by the solvent (the solvent front). These values depend on the solvent used, and the type of TLC plate, and are not physical constants. If the compounds are colored, they are easy to see with the naked eye. If not, a UV lamp is used.

Thus TLC is a simple, quick, and inexpensive procedure that gives the chemist a quick answer as to how many components are in a mixture. TLC is also used to support the identity of a compound in a mixture when the Rf of a compound is compared with the Rf of a known compound (preferably both run on the same TLC plate).

Applications

Its wide range of uses include

- assaying radiochemical purity of radiopharmaceuticals

- determination of the pigments a plant contains
- detection of pesticides or insecticides in food
- analysing the dye composition of fibers in forensics, or
- identifying compounds present in a given substance

7.3 Troubleshooting TLC

All of the above (including the procedure page) might sound like TLC is quite an easy procedure. But what about the first time you run a TLC, and see spots everywhere and blurred, streaked spots? As with any technique, with practice you get better. One thing you have to be careful Examples of common problems encountered in TLC :

- The compound runs as a streak rather than a spot

  The sample was overloaded. Run the TLC again after diluting your sample. Or, your sample might just contain many components, creating many spots which run together and appear as a streak. Perhaps, the experiment did not go as well as expected.

- The sample runs as a smear or a upward crescent.

  Compounds which possess strongly acidic or basic groups (amines or carboxylic acids) sometimes show up on a TLC plate with this behavior. Add a few drops of ammonium hydroxide (amines) or acetic acid (carboxylic acids) to the eluting solvent to obtain clearer plates.

- The sample runs as a downward crescent.

  Likely, the adsorbent was disturbed during the spotting, causing the crescent shape.

- The plate solvent front runs crookedly.

  Either the adsorbent has flaked off the sides of the plate or the sides of the plate are touching the sides of the container (or the paper used to saturate the container) as the plate develops. Crookedly run plates make it harder to measure Rf values accurately.

- Many, random spots are seen on the plate.

  Make sure that you do not accidentally drop any organic compound on the plate. If get a TLC plate and leave it laying on your workbench as you do the experiment, you might drop or splash an organic compound on the plate.

- No spots are seen on the plate.

  You might not have spotted enough compound, perhaps because the solution of the compound is too dilute. Try concentrating the solution, or, spot it several times in one place, allowing the solvent to dry between applications. Some compounds do not show up under UV light; try another method of visualizing the plate. Or, perhaps you do not have any compound because your experiment did not go as well as planned.

  If the solvent level in the developing jar is deeper than the origin (spotting line) of the TLC plate, the solvent will dissolve the compounds into the solvent reservoir instead of allowing them to move up the plate by capillary action. Thus, you will not see spots after the plate is developed.

- You see a blur of blue spots on the plate as it develops.

  Perhaps, you used an ink pen instead of a pencil to mark the origin?

7.4 Ion exchange chromatography

7.4.1. Introduction

Ion exchange chromatography is a type of adsorption chromatography in which retention of a solute due to its reversible electrostatic interaction with the oppositely charged groups on an ion exchanger. Hence, this technique is useful for separation of compounds which bear a net electric charge such as proteins, amino acids and nucleic acids etc. Ion exchangers are prepared from either certain synthetic resins which are insoluble porous organic molecules or naturally occurring biopolymers such as cellulose to which various groups known as fixed ions are covalently attached. These fixed ions are balanced by equal and oppositely charged ions from the solution referred to as counter ions. Depending upon the nature of the counter ions, these ion exchangers can be divided into two types :

1. Cation Exchanger : counter ions are cationic or positively charged ions. (CM Cellulose)
2. Anion Exchanger : Counter ions are negatively charged. (DEAE Cellulose)

Counter ions are mobile and can be easily exchanged by other similarly charged molecules in the sample. Nature of the resin matrix remains unchanged during this exchange process.

Some commonly used ion-exchangers

Type	Matrices	Functional group	Name of the Functional Group
Weak cation exchangers	Agarose Cellulose Dextran Polyacrylate	$-COO$ $-CH_2COO^-$	Carboxy Carboxymethyl
Strong cation exchangers	Cellulose Dextran Polystyrene	$-SO_3^-$ $-CH_2SO_3^-$ $-CH_2CH_2CH_2SO_3^-$	Sulpho Sulphomethyl Sulphopropyl
Weak anion exchangers	Agarose Cellulose Dextran Polystyrene	$-CH_2CH_2N^+H_3$ $-CH_2CH_2N^+(CH_2CH_3)_2$	Aminoethyl Diethylaminoethyl
Strong anion exchangers	Cellulose Dextran Polystyrene	$-CH_2N^+(CH_3)_3$ $-CH_2CH_2N^+(CH_2CH_3)_3$ $-CH_2N^+(CH_3)_2CH_2CH_2OH$ $-CH_2CH_2N^+(CH_2CH_3)_2CH_2CHCH_3OH$	Trimethylaminomethyl Triethylaminoethyl Dimethyl-2hydroxyethylaminomethyl Diethyl-2-hydroxypropylaminoethyl

Table : Different types of ion exchangers and their chemical nature

Generally resin based ion exchangers are used for separation of low molecular weight biomolecules and cellulosic ion exchangers are more suitable for isolation of macromolecules (for eg; proteins and nucleic acids).

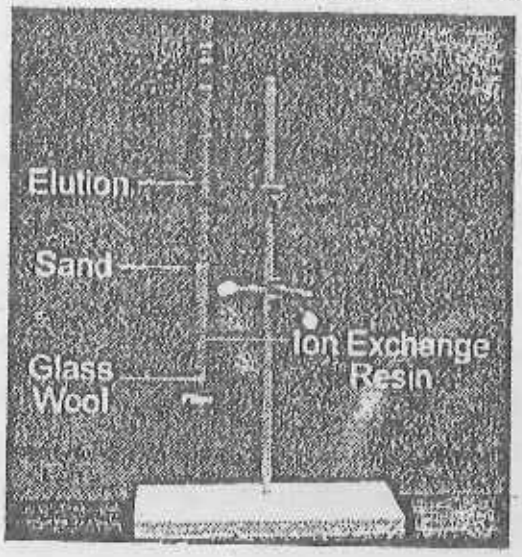


Fig. 7.4 An ion exchange column

7.4.2 Principle

Ion exchangers should be precycled to get an appropriate charge and their complete swelling. An anion exchanger is at first should be treated with alkali and then with acid and finally washed with water till it is neutral. On the other hand, cation exchanger is first given the acid treatment and then alkali treatment and finally washed with water till neutral. After precycling, the ion exchanger is packed into a column and is equilibrated with counter ion by passing 2-3 bed volumes of starting buffer of a required pH. In case of an anion exchanger, Tris-HCl buffer and for cation exchanger K^+ or Na^+ buffer is used. The process of ion exchange takes place in different steps. For example, the ion exchanger has positive fixed ions which are balanced by negatively charged counter ions. After the addition of sample containing mixture of compounds on the top of the exchanger bed, the solute molecules having charge similar to that of the counter ions (i.e. negative ions in this case) get exchanged with the counter ions and bind reversibly but strongly to the ion exchanger. Different compounds are bound with different strengths. Strength of binding depends upon the degree of the charge and the density of the charge (amount of charge per unit volume of the molecule) of the solute. The other solute particles (i.e. positively charged and neutral) have no affinity for the stationary phase and are washed down along with the starting buffer. The bound solute molecules are then released in succession by altering pH or ionic strength of the elution buffer. This sort of elution is known as gradient elution. (Fig. 7.6)

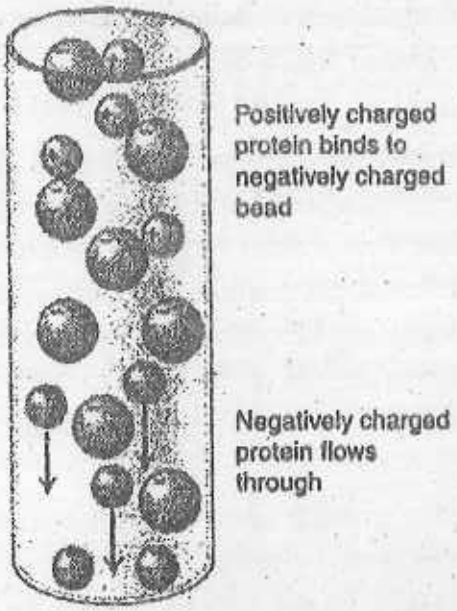


Fig. 7.6 Beads showing binding capability

Stepwise (discontinuous) or continuous gradient may be applied but continuous gradient is preferred because it gives better resolution and less tailing. Generally, with anion exchanger a buffer with increasing pH and increasing ionic strength but

with cation exchanger a buffer with decreasing pH and increasing ionic strength gradient is employed for desorption of bound substances. In both cases increasing ionic strength displaces the bound substances while change in pH on either side would bring the pH of the bound substance to isoelectric point where the substance will have no net charge and so it will get detached from the ion exchanger and will be eluted. The ion exchanger can be regenerated by again passing the starting buffer continuously. The eluted droplet was collected either manually or by using fraction collector. The protein content of the eluted amount was determined either by spectrophometer or by Folin Lowry method. (Figs. 7.7 & 7.8).

Gradient elution : This is of two types.

(i) Discontinuous gradient: This is also known as stepwise gradient. In this method, a fixed volume of the mobile phase of a particular pH, ionic strength or polarity is passed through the column. This is then followed by elution with fixed volume of mobile phase of different composition and this process of developing the column with definite volumes of mobile phase of changing composition is continued. Here, the polarity, the pH or the ionic strength of the eluent is changed in a stepwise manner and at every step, the change is rather abrupt.

(ii) Continuous gradient: In this method, the composition of the eluent is changed at a steady and constant rate. This can be done by using a gradient maker which is composed of two vessels of equal diameter with an outlet tube at their base. They are connected by a Teflon tube having an adjustable screw type pinch cock. The recipient chamber is connected to the column via teflon tubing having a pinch cock. In the beginning, both the pinch cocks are tightened and the two eluents of different pH, ionic strength or polarity are poured into each of the two vessels. The recipient chamber generally contains the starting buffer. The solution in the recipient chamber is continuously stirred using a magnetic stirrer. Both the pinch cocks are opened. As the eluent from the recipient vessel starts flowing into the column, the same volume of the eluent from the donor chamber enters into the recipient chamber where it is mixed by the stirrer. In this way, the pH, the ionic strength or polarity of the buffer in the recipient chamber changes continuously at a constant rate depending upon the composition of the eluent in the donor chamber. (Fig 7.5)

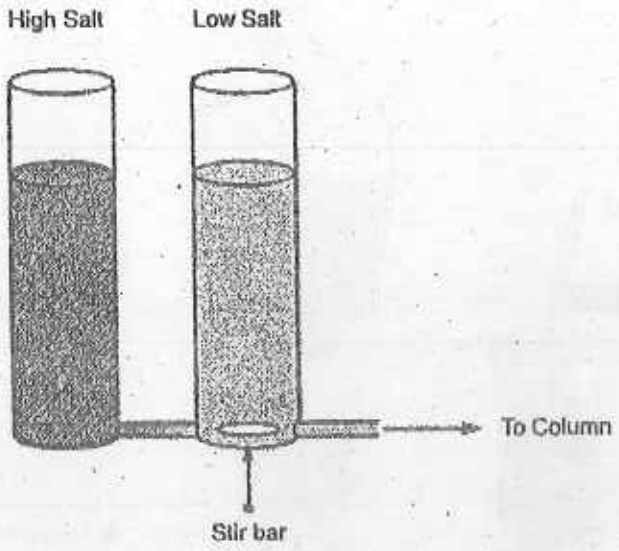


Fig. 7.5. Diagram of gradient maker.

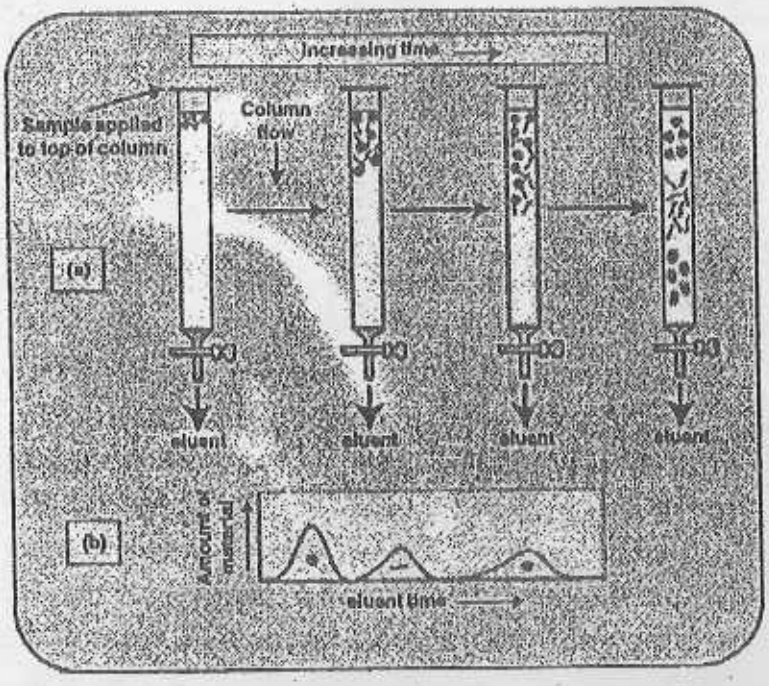


Fig. 7.7. Elution profile of Ion Exchange chromatographay

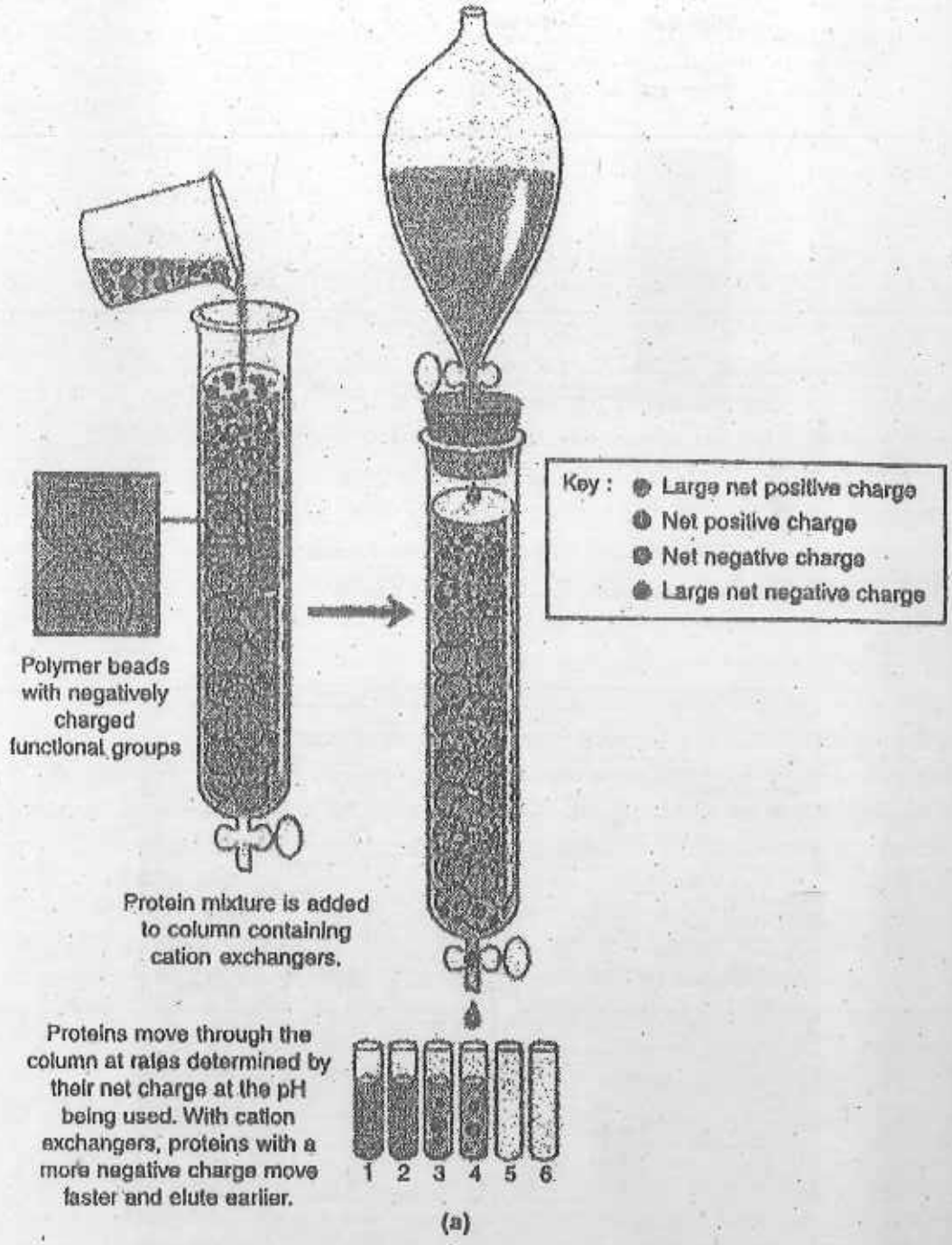
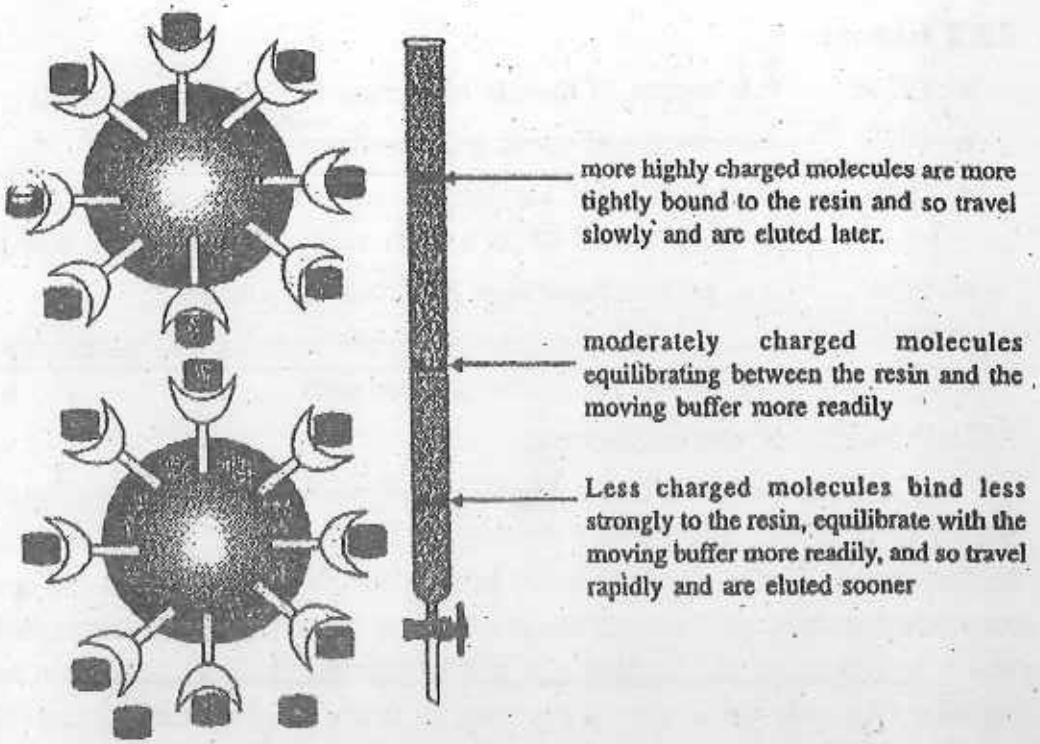


Fig. 7.8. Overall representation of ion exchange chromatography



7.5 SDS–PAGE (Sodium Dodecyl Sulphate–Polyacrylamide Gel Electrophoresis)

7.5.1 Introduction

Electrophoresis is the study of movement of charged particles in an electric field. This method pioneered by A. Tiselius in 1937 was a moving boundary electrophoresis where electrophoresis was carried out in buffer solution. But the demerits of this method were sample diffusion and mixing of sample components due to convection of currents. These limitations have been overcome in zone electrophoresis where a buffer saturated solid matrix is used as an electrophoretic medium in place of a buffer solution. The sample is applied onto the medium as a narrow zone or band and the molecules with different mobilities travel as a distinct zones which gradually separate from each other. This technique is widely used for separation and analysis of a large number of biomolecules such as amino acids, peptides, proteins, nucleotides and nucleic acids.

7.5.2 History

- 1930s – first reports of the use of sucrose for gel electrophoresis
- 1955 – introduction of starch gels, mediocre separation
- 1959 – introduction of acrylamide gels (Raymond and Weintraub); accurate control of parameters such as pore size and stability
- 1964 – disc gel electrophoresis (Ornstein and Davis)
- 1969 – introduction of denaturing agents especially SDS separation of protein subunit (Weber and Osborn)

7.5.3 Principle of electrophoresis

Any charged ion or molecule migrates when placed in an electrical field. The rate of migration of a compound depends on its net charge, size, shape and the applied current. This can be represented by the equation. The movement of a charged molecule in an electric field is often expressed in terms of electrophoretic mobility which is defined as the velocity per unit of electric field. Thus electrophoretic mobility of a molecule is directly proportional to the charge density (charge /mass ratio). Molecules with different charge/ mass ratio migrate under the electric field at different rates and hence get separated.

7.5.3 Polyacrylamide gel electrophoresis (PAGE)

Electrophoresis in polyacrylamide gel is the most widely used technique for the analysis and characterization of proteins and nucleic acids. Use of polyacrylamide gels has several advantages because it is chemically inert, gives superior resolution, is amendable to preparation of gels with wide range of pore sizes and that the gels are stable over a wide range of pH, temperature and ionic strength. Polyacrylamide gel is prepared by polymerization of acrylamide monomers into long chains and cross-linking these by bifunctional reagents like N,N'-methylene bisacrylamide. Polymerization is initiated by ammonium persulfate or riboflavin. N, N, N', N'-tetramethylene diamine (TEMED) catalyses formation of free radicals from persulfate which in turn initiate polymerization. In the case of riboflavin-TEMED system, light causes photodecomposition of riboflavin and production of necessary free radicals. Oxygen inhibits polymerization. Gel solutions have, therefore, to be degassed prior to the polymerization step. The pore size of polyacrylamide gels can be varied by changing total concentration of acrylamide and bis-acrylamide monomers in a

fixed volume of gelation solution. Gels ranging from 3 to 30% acrylamide concentration can be made and these can be used for separation of molecules of size upto $1x10^6$ Daltons. A gel with low percentage has larger pore size and is suitable for separation of high molecular weight compounds and a high percentage gel has smaller pore size and is used for separation of relatively low molecular weight compounds. Alternatively, gels of different pore sizes may be prepared by varying the relative concentrations of acrylamide and bis-acrylamide. Gradient gels with linear gradients of increasing acrylamide concentration give better resolution and are quite commonly used.

If the proteins are denatured and put into an electric field, they will all move towards the positive pole at the same rate, with no separation by size. So we need to put the proteins into an environment that will allow different sized proteins to move at different rates. The environment of choice is polyacrylamide, which is a polymer of acrylamide monomers. When this polymer is formed, it turns into a gel and we will use electricity to pull the proteins through the gel so the entire process is called polyacrylamide gel electrophoresis (PAGE). A polyacrylamide gel is not solid but is made of a laberynth of tunnels through a meshwork of fibers (Fig.7.9).

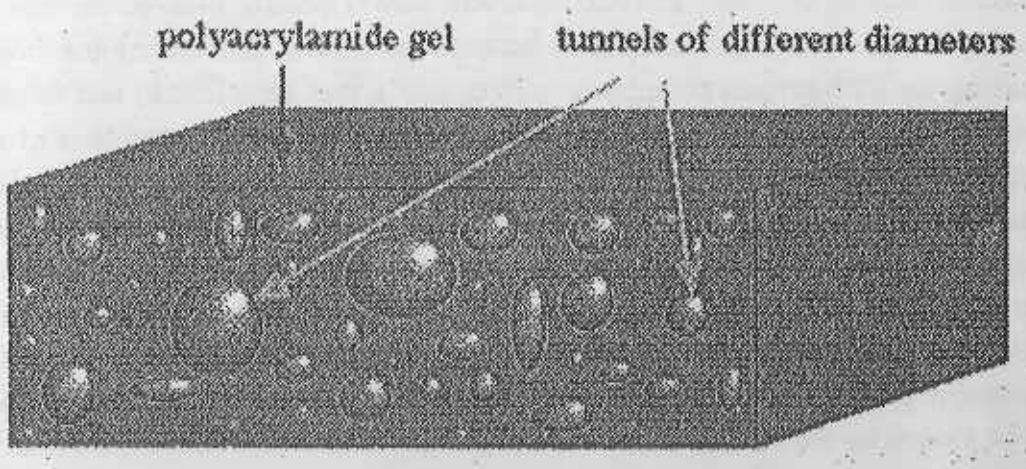


Fig. 7.9. This cartoon shows a slab of polyacrylamide (dark gray) with tunnels (different sized red rings with shading to depict depth) exposed on the edge. Notice that there are many different sizes of tunnels scattered randomly throughout the gel.

7.6 SDS-PAGE (Polyacrylamide gel Electrophoresis)

7.6.1 Principle

The widely used electrophoretic method, the SDS-PAGE was developed by Laemmli. This is a discontinuous system consisting of two continuous, but distinct gels: a resolving or separating (lower) gel and a stacking (upper) gel. The two gels are cast with different porosities, pH and ionic strength. In addition, different mobile ions are used in the gel and electrode buffers. The buffer discontinuity acts to concentrate large volume samples in the stacking gel, resulting in better resolution than is possible using the same sample volumes in the gels without stackers. Proteins once concentrated in the stacking gel, are separated in the resolving gel.The Laemmli SDS-PAGE system is made up of four components. From the top of the cell downward, these are the electrode buffer, the sample, the stacking gel, and the resolving gel. Samples prepared in low conductivity buffer (0.06 M Tris-Cl, pH 6.8) are loaded between the higher conductivity electrode (0.025 M Tris, 0.192 M Glycine, pH 8.3) and stacking gel (0.125 M Tris-Cl, pH 6.8) buffers. When power is applied, a voltage drop develops across the sample solution which drives the proteins into the stacking gel. Glycine ions from the electrode buffer follow the proteins into the stackinggel. A moving boundary region is rapidly formed with the highly mobile chloride ions in the front and the relatively slow glycinate ions in the rear. A localized high-voltage gradient forms between the leading and trailing ion fronts, causing the SDS-protein complexes to form into a thin zone (stack) and migrate between the chloride and glycinate phases. Within broad limits, regardless of the height of the applied sample, all SDS-proteins condense into a very narrow region and enter the resolving gel as a well-defined, thin zone of high protein density. The large pore stacking gel (4% T) does not retard the migration of most proteins and serves mainly as an anticonvective medium. At the interface of the stacking and resolving gels, the proteins experience a sharp increase in retardatio due to the restrictive pore size of the resolving gel. Once in the resolving gel, proteins continue to be slowed by the sieving of the matrix. The glycinate ions overtake the proteins, which then move in a space of uniform pH (ph 9.5) formed by Tris and glycine. Molecular sieving causes the SDS-polypeptide complexes to separate on the basis of their molecular weights.

Since we are trying to separate many different protein molecules of a variety of shapes and sizes, we first want to get them to be linear so that the proteins no longer have any secondary, tertiary or quaternary structure (i.e. we want them to have the same linear shape). Consider two proteins that are each 500 amino acids long but one is shaped like a closed umbrella while the other one looks like an open umbrella. If you tried to run down the street with both of these molecules under your arms, which one would be more likely to slow you down, even though they weigh exactly the same? This analogy helps point out that not only the mass, but also the shape of an object will determine how well it can move through and environment. So we need a way to convert all proteins to the same shape–we use SDS. (Fig. 7.10)

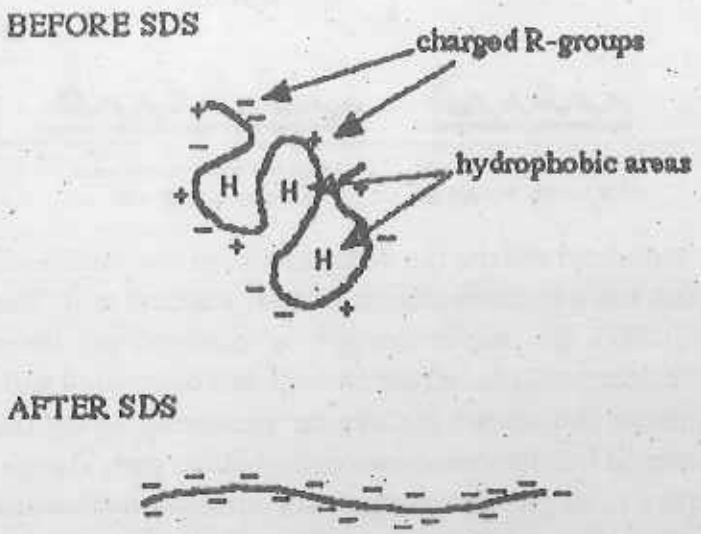
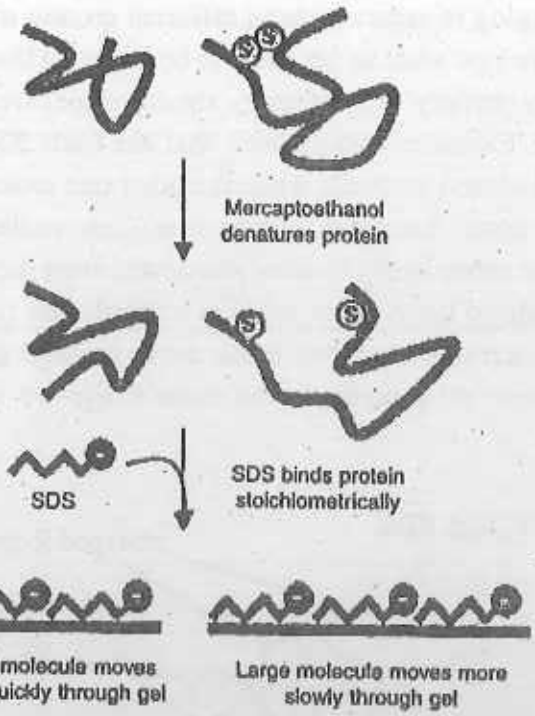


Fig. 7.10. This cartoon depicts what happens to a protein (pink line) when it is incubated with the denaturing detergent SDS. The top portion of the figure shows a protein with negative and positive charges due to the charged R-groups of the particular amino acids in the protein. The large H represents hydrophobic domains where non polar R-groups have collected in an attempt to get away from the polar water that surrounds the protein. The bottom portion shows that SDS can break up hydrophobic areas and coat proteins with many negative charges which overwhelms any positive charge in the protein due to positively charged R-groups. The resulting protein has been denatured by SDS (reduced to its primary structure) and as a result has been genearized.



SDS (sodium dodecyl sulfate) is a detergent (soap) that can dissolve hydrophobic molecules but also has a negative charge (sulfate) attached to it. Therefore, if a cell is incubated with SDS, the membranes will be dissolved and the proteins will be soluiblized by the detergent, plus all the proteins will be covered with many negative charges. So a protein that started out like the one shown in the top part of figure 7.10 will be converted into the one shown in the bottom part. The end result has two important features : 1) all proteins contain only primary structure and 2) all proteins have a large negative charge which means they will all migrate towards the positive pole when placed in an electric field. Now we are ready to focus on the second half – PAGE. The solution of proteins to be analyzed is first mixed with SDS, an anionic detergent which denatures secondary and non–disulfide–linked tertiary structures, and applies a negative charge to each protein in proportion to its mass. Without SDS, different proteins with similar molecular weights would migrate differently due to differences in mass charge ratio, as each protein has an isoelectric point and molecular weight particular to its primary structure. This is known as Native PAGE. Adding SDS solves this problem, as it binds to and unfolds the protein, giving a near uniform negative charge along the length of the polypeptide.SDS bind in a ratio of approximately 1.4 g SDS per 1.0 g protein (although binding ratios can vary from

1.1-2.2 g SDS/g protein), giving an approximately uniform mass:charge ratio for most proteins, so that the distance of migration through the gel can be assumed to be directly related to only the size of the protein. A tracking dye may be added to the protein solution to allow the experimenter to track the progress of the protein solution through the gel during the electrophoretic run. (Fig 7.11)

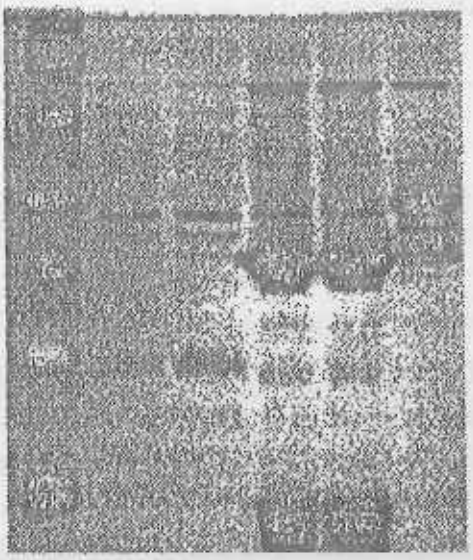


Fig. 7.11. Picure of an SDS-PAGE. The molecular marker is in the left lane

There are two layers of gel, namely stacking or spacer gel, and resolving or separating gel.

7.6.2 Stacking gel

The stacking gel is a large pore polyacrylamide gel (4%). This gel is prepared with Tris buffer pH 6.8 of about 2 pH units lower than that of electrophoresis buffer. Proteins are concentrated to several folds and a thin starting zone is achieved in a few minutes. This gel is cast over the resolving gel. The height of the stacking gel region was always maintained more than double the height and the volume of the sample to be applied.

7.6.3 Resolving gel

The resolving gel is a small pore polyacrylamide gel (3–30%). The Tris buffer used is of pH 8.8. In this gel, macro molecules separate according to their size. In

the present experiment, 8%, 10% and 12%. Resolving gel were used for separating different range of proteins. 8% gel for 24–205 kD proteins, 10% gel for 14-205 kD proteins and 12% gel for 14-66 kD proteins

7.6.4 Chemical ingredients

- **Tris** (tris (hydroxy methyl) aminomethane) ($C_4H_{11}NO_3$; mw : 121.14). It has been used as a buffer because it is an innocuous substance to most proteins. Its pKa is 8.3 at 20°C and reasonably a very satisfactory buffer in the pH range 7.0–9.0.
- **Glycine** (Amino Acetic Acid) ($C_2H_5NO_2$; mw : 75.07). Glycine has been used as the source of trailing ion or slow ion because its pKa is 9.69 and mobility of glycinate are such that the effective mobility can be set at a value below that of the slowest known proteins of net negative charge in the pH range. The minimum pH of this range is somewhere around 8.0.
- **Acrylamide** (C_3H_5NO; mw : 71.08). It is a white crystalline powder. While dissolving in water, auto polymerisation of acrylamide takes place. It is a slow spontaneous process by which acrylamide molecules join together by head on tail fashion. But in presence of free radicals generating system, acrylamide monomers are activated into a free-radical state. These activated monomers polymerise quickly and form long chain polymers. This kind of reaction is known as Vinyl addition polymerisation. A solution of these polymer chains becomes viscous but does not form a gel, because the chains simply slide over one another. Gel formation requires hooking various chains together. Acrylamide is a neurotoxin. It is also essential to store acrylamide in a cool dark and dry place to reduce autopolymerisation and hydrolysis.
- **Bisacrylamide** (N, N'-Methylenebisacrylamide) ($C_7H_{10}N_2O_2$; mw : 154.17). Bisacrylamide is the most frequently used cross linking agent for poly acrylamide gels. Chemically it is thought of having two-acrylamide molecules coupled head to head at their non-reactive ends. Bisacrylamide was preserved at 4°C.
- **Sodium Dodecyl Sulfate (SDS)** ($C_{12}H_{25}NaO_4S$; mw : 288.38). SDS is the most common dissociating agent used to denature native proteins to individual polypeptides. When a protein mixture is heated to 100 °C in presence of SDS, the detergent wraps around the polypeptide backbone. It binds to polypeptides

in a constant weight ratio of 1.4 g/g of polypeptide. In this process, the intrinsic charges of polypeptides becomes negligible when compared to the negative charges contributed by SDS. Thus polypeptides after treatment becomes a rod like structure possessing a uniform charge density, that is same net negative charge per unit length. Mobilities of these proteins will be a linear function of the logarithms of their molecular weights.

- **Ammonium persulfate** (APS) ($N_2H_8S_2O_8$; mw : 228.2). APS is an initiator for gel formation.
- **TEMED** (N, N, N′, N′-tetramethylethylenediamine) ($C_6H_{16}N_2$; mw : 116.21). Chemical polymerisation of acrylamide gel is used for SDS-PAGE. It can be initiated by ammonium persulfate and the quaternary amine, N, N, N′, N′-tetramethylethylenediamine (TEMED). The rate of polymerisation and the properties of the resulting gel depends on the concentration of APS and TEMED. Increasing the amount of APS and TEMED results in a decrease in the average polymer chain length, an increase in gel turbidity and a decrease in gel elasticity. Decreasing the amount of initiators shows the reverse effect. The lowest catalysts concentrations that will allow polymerisation in the optimal period of time should be used. APS and TEMED are used, approximately in equimolar concentrations in the range of 1 to 10 mM.

7.6.5 Chemicals for processing and visualization

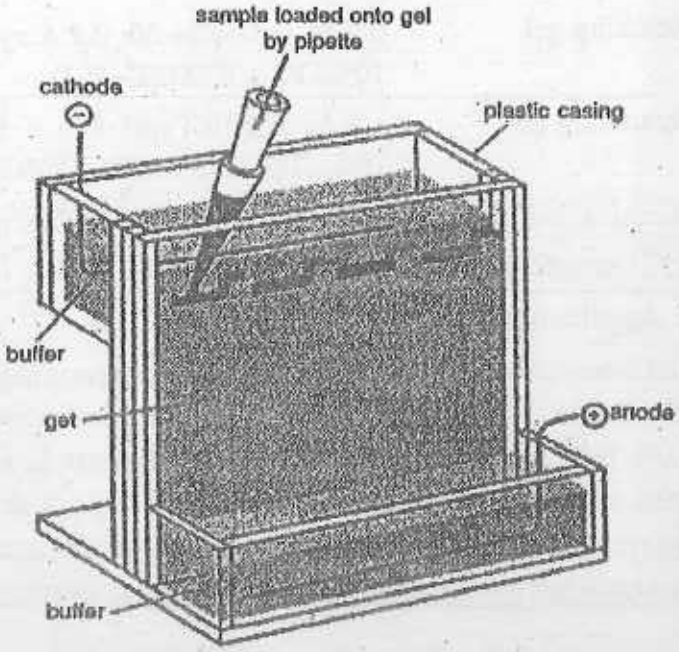
The following chemicals are used for processing of the gel and the protein samples visualized in it :

- **Bromophenol Blue** (BPB) (3′, 3″, 5′, 5″-Tetrabromophenolsulphonephthalein) ($C_{19}H_{10}Br_4O_5S$; mw : 669.99). BPB is the universal marker dye. Proteins and nucleic acids are mostly colourless. When they are subjected to electrophoresis, it is important to stop the run before they run off the gel. BPB is the most commonly employed tracking dye, because it is viable in alkali and neutral pH, it is a small molecule, it is ionisable and it is negatively charged above pH 4.6 and hence moves towards the cathode. Being a small molecule it moves ahead of most proteins and nucleic acids. As it reaches the cathodic end of the electrophoresis medium electrophoresis is stopped. It can bind with proteins weakly and give blue colour.

- **Glycerol** ($C_3H_8O_3$; mw : 92.09). It is a preservative and a weighing agent. Addition of glycerol (20-30 or 50%) is often recommended for the storage of enzymes. Glycerol maintains the protein solution at very low temperature, without freezing. It also helps to weigh down the sample into the wells without being spread while loading.
- **Coomassie Brilliant Blue** (CBB) ($C_{45}H_{44}N_3NaO_7S_2$; mw : 825.97). CBB is the most popular protein stain. It is an anionic dye, which binds with proteins non-specifically. The structure of CBB is predominantly non-polar. So is usually used (0.025%) in methanolic solution (40%) and Acetic Acid (7%). Proteins in the gel are fixed by acetic acid and simultaneously stained. The excess dye incorporated in the gel can be removed by destaining with the same solution but without the dye. The proteins are detected as blue bands on a clear background. As SDS is also anionic, it may interfere with staining process. Therefore, large volume of staining solution is recommended, approximately ten times the volume of the gel.
- **Butanol** ($C_4H_{10}O$; mw: 74.12). Water saturated butanol is used as an overlay solution on the resolving gel.
- **2-Mercaptoethanol** ($HS\text{-}CH_2CH_2OH$; mw : 78.13).

7.6.6. Apparatus

It consists of a power pack and an electrophoresis unit. The power pack supplies a stabilized direct current at controlled or required voltage and current output. The electrophoresis unit contains the electrodes, buffer reservoirs and gel casting assembly. Initially, electrophoresis in polyacrylamide gels was carried out in cylindrical rods of gel in glass tubes (~7 mm in diameter and 10 cm in length) which were then fixed in a specially designed apparatus. Now, slab gel, (0.75–1.5 mm thick), in which the gel is in the form of a rectangular or square flat slab, is more commonly used. Two types of slab gel apparatus are available–horizontal and vertical, the latter is usually more preferred. Slab gels have advantage over rod gels since a number of samples can be co- electrophoressed under identical conditions on a single gel so that band patterns can be directly compared. On the other hand only one sample can be loaded onto each gel rod. Also, the heat produced during electrophoresis is more easily dissipated by thin slab gels as compared to thick rod gels thus reducing distortion of bands due to heating (Fig. 7.12).



Buffers for SDS-Polyacrylamide gel electrophoresis

(a) Continuous system

Resolving gel buffer

0.1 M sodium phosphate, pH 7.2

Reservoir buffer

0.1 M sodium phosphate, pH 7.2

0.1% SDS

(b) Discontinuous system

Stacking gel buffer

0.125 M Tris-HCl, pH 6.8

Resolving gel buffer

0.375 m Tris-HCl, pH 8.8

Reservoir buffer

0.025 M Tris

0.192 M glycine

0.1% SDS

pH 8.3

7.6.7 Solutions for SDS-PAGE

Stacking gel : 0.5 M Tris-HCl+ 30: 0.8 Acryl: Bisacryl + 10% SDS + 10%APS + TEMED+H_2O

Resolving gel: 1.5 M Tris-HCl (pH 8.8) + 30: 0.8 Acryl: Bisacryl + 10% SDS + 10%APS + TEMED +H_2O

Running buffer: 0.025 M Tris + 0.192 M Glycine + 0.1 % SDS (pH 8.3)

2X Laemmli buffer : 50% glycerol + 10% SDS + 0.312 M tris-HCl.

7.6.8 Applications

Gel electrophoresis is used in forensics, molecular biology, genetics, microbiology and biochemistry. The results can be analyzed quantitatively by visualizing the gel with UV light and a gel imaging device. The image is recorded with a computer operated camera, and the intensity of the band or spot of interest is measured and compared against standard or markers loaded on the same gel. The measurement and analysis are mostly done with specialized software.

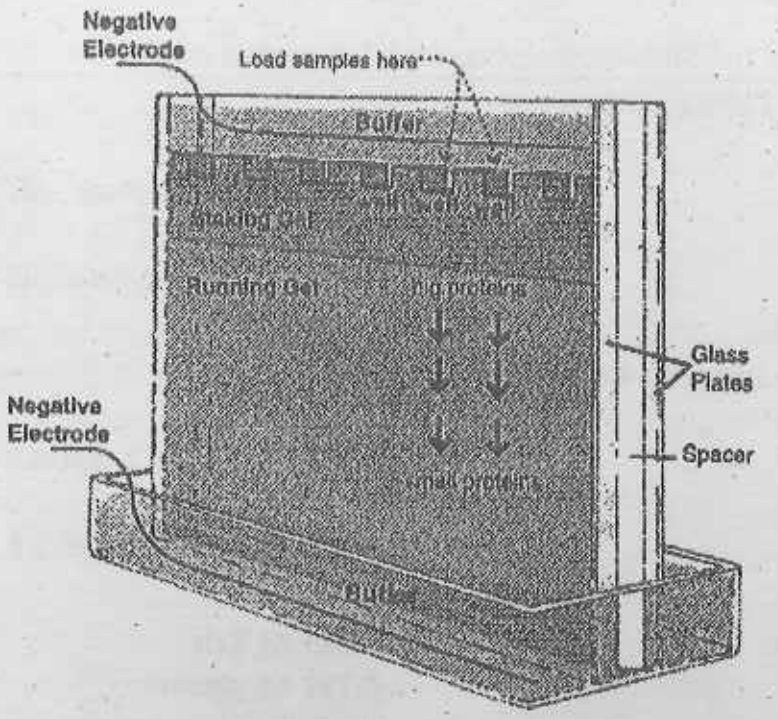


Fig. 7.12 An illustration of an apparatus used for SDS PAGE

Depending on the type of analysis being performed, other techniques are often implemented in conjunction with the results of gel electrophoresis, providing a wide range of field-specific applications.

7.7 Affinity Chromatography

7.7.1 Process of affinity chromatography

Affinity chromatography is a type of adsorption chromatography in which the substance to be isolated is specifically and reversibly bound to a complementary binding substance (ligand) immobilized on an insoluble chromatographic bed material (matrix). The other substances in the mixture remain unbound and are, therefore, washed away while the substance of interest (the bound substance) is subsequently recovered by displacement from the ligand either by specific (affinity) elution or by non-specific (change in pH or ionic concentration) elution. Affinity chromatography is based on utilization of specific biochemical property of a protein and not simply the physicochemical characters and so the separation is of high purification grade.

Some proteins have a characteristic property of binding tightly with certain molecules but not by forming covalent bonds. These proteins can be purified by affinity chromatography taking advantage of their binding affinity. A molecule called *ligand*, with which the concerned protein binds, is first covalently bound to an inert and porous matrix. The desired protein will be bind with this ligand when the sample solution will pass through the chromatographic material while other substances will wash through the column with buffer solution. An appropriate change in the elution conditions that will release the protein from chromatographic matrix will yield recovery of the protein of interest in highly purified form. Affinity chromatography is based on utilization of specific biochemical property of a protein and not simply the physicochemical characters and so the separation is of high purification grade.

For affinity chromatography matrix must be chemically inert so that it does not react with and alter the protein to be purified. It must have high porosity with large number of functional groups that may form covalent linkages with the ligand. Agarose, one of several such materials that have the said properties, has number of free hydroxyl groups and so is most widely used. There are number of commercially available materials having such properties, like Sepharose or CL- Sepharose. If the ligand contains a primary amino group which is not required for it's binding to the

concerned protein, then such a ligand can be linked with agarose in two steps :

1. Agarose is reacted with cynoen bromide(CNBr) and thus a stable and activated intermediate is formed, and
2. Ligand is allowed to react with activated agarose to form a covalently bound product.

Many proteins do not bind with the cyanogens -bromide coupled ligands. This is due to the steric interference with the agarose matrix. So, a spacer arm is almost always inserted between the matrix and the ligand so that the large molecules can reach the binding sites. If the spacer are too small the hindrances may continue and if too long the undesirable non-specific adsorption, specially of hydrophobic compounds, may occur. Practically a spacer arm of 2-10 carbon atoms proves to be optimum. An appropriate diamine (eg., diaminobutane dihydrochloride or hexamethyl diamine (cyanogens bromide) is coupled to agarose by amino alkylation. The resulting product is coupled with the ligand depending on functional groups of ligand. The ligand used in affinity chromatography for a given protein should have enough affinity to bind and thus to immobilize the protein in agarose gel but only to the extent that the protein must subsequently be released. If the ligand is a substrate for an enzyme that has to be separated, the chromatographic conditions must prevent the catalytic action of enzyme or otherwise that would destroy the ligand.

After the process of binding the protein with affinity chromatographic column and the washing to remove impurities is over, the protein has to be released from the column. There are two ways :

1. The column may be eluted with a solution containing a compound with higher affinity to the protein binding site that the bound ligand, so that ligand bound protein is set free by being substituted by the latter compound
2. The alteration in the solution conditions to end the protein-ligand complex stability may be brought by changing pH, ionic strength and also the temperature.

Care must be taken to protect the desired protein from any irreversible damages.

7.7.2 Uses and applications of affinity chromatography

It has been used to isolate a wide array of proteins ranging from enzymes to hormone receptors, and sometimes even complete cell. The proteins have also been

used the other way round e.g., insulin when covalently bound with agarose, is used to separate insulin receptors that have the affinity of temporarily binding with insulin. The hormone receptors are proteins attached to the surface of target cells and occur in very small amounts and hence there properties are almost unknown. Thus the efficiency and utility of chromatography for isolation of protein, which allows the separation of even difficult proteins in a single step as compared with earlier methods involving several steps with a yet purer and larger yield, is unquestionable.

A large number of biospecific ligands can be employed. Practically it is possible to select a ligand which displays absolute specificity and binds exclusively with one particular compound only. It is also possible to select a ligand which displays group specificity e.g. for the purification of enzymes a substrate analogue, or an inhibitor, a cofactor or an activator can be used as ligand while a specific receptor or a carrier protein can be used for the purification of vitamins and hormones. Antigen-antibody interactions can be exploited for the purification of either of these. Similarly for nucleic acid purification a complementary base sequence or histones and for lectins, cell surface receptors or polysaccharides can be successfully employed s ligands. But before selecting a ligand, two attributes should be kept in mind that

1. the ligand must exhibit specific and reversible binding with the substance to be purified and
2. it must have chemically reactive function groups which allow it to be attached to the matrix without destroying its binding activity with the substance of interest.

If several functional groups are available, the ligand should be coupled with the matrix via the group least likely to be involved in its specific interaction with the molecule to be isolated. The most common such groups are $-NH_2$, –COOH, –SH and –OH.

An ideal matrix to which the ligand is covalently bound must possess the following attributes :

(i) It must possesses suitable groups to which lignd can be covalently coupled. Many groups may be introduced into matrix in order to couple ligand. They may be nucleophilic as NH_2, SH, OH or electrophilic such as activated acid chlorides, carbonyls activated by carbodimide, isothiocyanate, or diazonium salts.

(ii) It must remain unchanged under the conditions of attachment of ligand.

(iii) During the binding of the macromolecule and its subsequent displacement from the ligand, it must retain its physical and chemical stability.

(iv) It must not exhibit nonspecific adsorption.

(v) It should have an open pore structure.

The most commonly used matrices are cross-linked dextrans (e.g. Sephacryl), agarose (e.g. Sepharose), polyacrylamide gel (Bio gel P), polysterene, cellulose, porous glass and silica. Sepharose is a bead-form of agarose gel which displays virtually all features required of a successful matrix for immobilizing biologically active molecules. The hydroxyl groups on the sugar residues can be easily derivatized for covalent attachment of a ligand.

Affinity chromatography occupies a unique place in separation technology since it is the only technique which enables purification of almost any biomolecule on the basis of its biological function. The principle of affinity chromatography has been extended to purify a large number of enzymes, other proteins including immunoglobulins, receptor proteins and nucleic acids. Poly (U) Sepharose 4B, poly (A) Sepharose 4B and immobilized single stranded DNA have been successfully used to isolate mRNA, viral mRNA and complementary RNA and DNA, respectively.

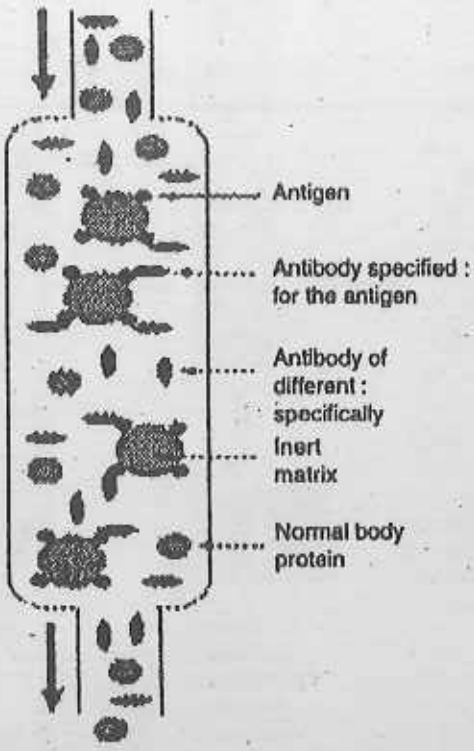
The technique of affinity chromatography has also been successfully employed for the separation of a mixture of cells into homogenous populations where it relies either on the antigenic properties of the cell surface or on the chemical nature of exposed carbohydrate residue on the cell surface or on a specific membrane receptor-ligand interactions.

Some useful application of affinity chromatography :

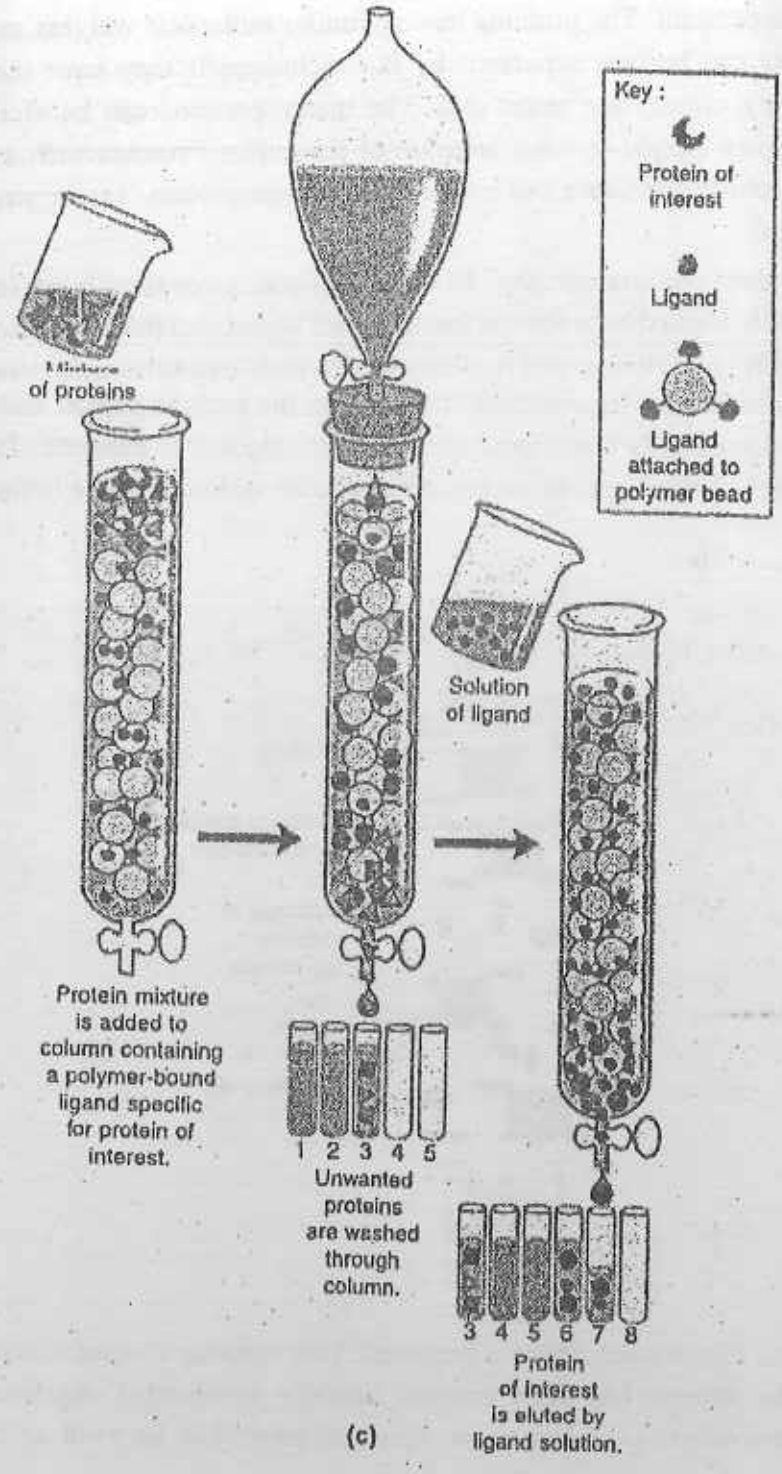
1. Dye ligand chromatography : A number of proteins including interferon, plasminogen and restriction endonucleases can be purified by this method. In this technique, immobilized cibacron blue (one of the triazine dyes) is used as ligand to purify proteins. The ionic group and the conjugated ring system of the ligand bind with catalytic or effector site of some proteins and so can be exploited to purify such proteins.
2. Metal affinity chromatography: Metals such as Zn^{2+}, Cu^{2+}, Cd^{2+}, Hg^{2+}, Co^{2+} and Ni^{2+} are immobilized by chelation using either iminodiacetic acid or tris (carboxymethyl) ethylene diamine. Binding of proteins with such metals is

pH dependent. The proteins having similar molecular weights and isoelectric points can be best separated by this technique if they have the differential binding affinity for metal ions. The bound proteins can be eluted either by changing the pH or ionic strength of the buffer. Proteins such as fibrinogen, superoxide dismutase and non-histone nuclear proteins can be purified by this method.

3. Covalent chromatography: In this technique, a covalent bond (e.g. disulfide bond) is formed between the immobilized ligand and the protein to be isolated. Ligands like thioprophl Sepharose and thiol Sepharose are used to purify proteins having large number of thiol groups such as papain and urease. The bound protein is then eluted either by ditiothreitol or cysteine. The larger the number of thiol groups on the protein to be isolated, better is the separation.



Step 1. An immunoadsorbent is prepared. This consists of a solid matrix to which the antigen has been coupled (usually covalently). Agarose, sephadex, derivatives of cellulose, or other polymers can be used as the matrix.


Key :
Protein of interest
Ligand
Ligand attached to polymer bead
of proteins
Solution of ligand
Protein mixture is added to column containing a polymer-bound ligand specific for protein of interest.
1 2 3 4 5
Unwanted proteins are washed through column.
3 4 5 6 7 8
Protein of interest is eluted by ligand solution.
(c)

Step 2. The serum is passed over the immunoadsorbent. As long as the capacity of the column is not exceeded, those antibodies in the mixture specific for the antigen (shown in red) will bind (noncovalently) and be retained. Antibodies of other specificities (green) and other serum proteins (yellow) will pass through unimpeded.

Step 3. **Elution.** A reagent is passed into the column to release the antibodies from the immunoadsorbent. Buffers containing a high concentration of salts and/or low pH are often used to disrupt the noncovalent interactions between antibodies and antigen. A denaturing agent, such as 8 M urea, will also break the interaction by altering the configuration of the antigen-binding site of the antibody molecule. Another, gentler, approach is to elute with a soluble form of the antigen. These compete with the immunoadsorbent for the antigen-binding sites of the antibodies and release the antibodies to the fluid phase.

Step 4. **Dialysis.** The eluate is then dialyzed against, for example, buffered saline in order to remove the reagent used for elution.

7.8 High Performance (Pressure) Liquid Chromatography

7.8.1 Principle

Separation in High Performance Liquid Chromatography or High Pressure Liquid Chromatography (HPLC) is obtained on the basis of partitioning, adsorption, ion exchange or molecular sieving phenomena. The conventional column chromatography suffers from two major drawbacks; namely it is generally time consuming process and quality of resolution is poor. Generally resolution of individual components can be improved by decreasing the particle size of stationary phase. But in conventional column chromatography, this is not feasible because the use of finer gel material will further lower the permeability of the column contributing to decreasing flow rate there by providing greater time for band broadening. The resistance to flow of mobile phase can be overcome by the use of high pressure. Recently, Stationary phases of smaller particle size which can withstand high pressures, have been developed, which has facilitated the development of a new chromatographic technique called HPLC, which gives faster and superior resolution with sharp and compact peaks. The basic HPLC equipment includes : (i) solvent reservoir (ii) pump (iii) damping device (iv) pressure gauge (v) sampling device (vi) column (vii) detector (viii) fraction collector and (ix) recorder.

7.8.2 Process

The pump delivers the solvent from the reservoir at a constant flow which is smoothed out by means of damping device. A safety valve is generally incorporated at a place where maximum pressure develops (normally after the pump). From the damping unit the mobile phase flows via sample injector into the column. The inlet pressure of the column is monitored with a manometer. After leaving the column, the sample components are monitored by the detector and their tracings drawn by the recorder.

7.8.3 Solvent reservoir

Solvents contain substantial amount of dissolvent air. Formation of air bubbles can seriously interfere with satisfactory separation by HPLC because the air bubbles affect the column efficiency and also the solute detection. Thus some of the conventional solvent reservoirs are equipped with degasifiers. In these the solvent reservoir is provided with a heater, temperature regulated magnetic stirrer or a condenser. Alternatively, solvent can also be degassed by heating, stirring, subjecting it to vacuum, ultrasonic vibrations or bubbling helium gas before pouring into the reservoir. The solvent used in HPLC must be of high purity as any traces of impurities or suspended material can seriously affect the column efficiency and can interfere with the detection system. Introduction of a 1-5 μm microfilter can be done prior to the pump to prevent any particulate impurities from entering into the column.

7.8.4 Pumps

In HPLC, the column is quite narrow and is packed with superfine particles. There is a high resistance to flow of solvent and high pressures are, therefore, required to achieve satisfactory and constant flow rates. Therefore, a good pumping system which delivers pulse free solvent flow up to 20 ml/min at pressures upto 300-400 atmospheres is one of the most important features in HPLC. All materials in the pump should be chemically resistant to all the solvents used in HPLC. Various pumping systems operate on the principle of either constant pressure or constant displacement.

Constant pressure pumps facilitate delivery of the solvent at a constant pressure. A gas at high pressure is introduced into the pump which forces, in turn the solvent from pump into he column. Constant pressure is maintained through out, which causes a decrease in permeability of the column with time which in turn results in

decreased flow rates. Such pumps do not compensate for this decrease in flow rate and so provide uniform and pulse less solvent flow.

The second types of pumps are the constant displacement pump. Such pumps displace a constant amount of the solvent from the pump into the column and so maintain a constant flow rate irrespective of the changing conditions within the column. These pumps produce small pulses of flow between two displacements and so pulse dampeners are usually introduced between the pump and the column to smoothen the flow and to minimize the pulsing effect. Two commonly used constant displacement pumps are—

(i) motor driven syringe type pump and

(ii) reciprocating pump which deliver a fixed and constant volume of the solvent onto the column at each stroke.

7.8.5 Sampling device

Sample can be introduced into the column either by a syringe injection through a septum of an injection port into eluent stream or by a sample loop from which it is swept into the column by the eluent. The sample is loaded directly on top of the column to avoid appreciable mixing of the sample with the eluent. In the syringe injection mode, the sample is injected with the help of a micro syringe directly onto the column bed. The system should not be under pressure while loading the sample. So, before applying the sample the pump is turned off and when the pressure is dropped near atmospheric pressure, the sample is introduced. After the sample has been injected, the pump is switched on again. This process is known as stop flow injection.

The other type is loop injection. In this method, sample is introduced with the help of a metal loop of fixed small volume. The loop is filled with the sample and by the appropriately adjusting the sample valve, the solvent from the pump is channeled through the loop. Sample is thus flushed by the solvent from the loop whose outlet opens directly at top of the column bed.

7.8.6 Columns

Stainless steel precision bored columns with an internal mirror finish are generally used for efficient packing. These straight columns of 15-50 cm length and 1-4 mm diameter can withstand very high pressures of upto 5.5×10^7 Pa and are relatively corrosion resistant. At the end of the column, homogenously porous plugs of stainless

steel or Teflon are used to retain the packing material and to ensure the uniform flow of the solvent through the column. At times, repeated application of impure samples such as urine, blood or crude cell extracts results in clogging and the loss of resolving power of the column. To prevent this, a short column of length of 1-2 cm and internal diameter equal to that of analytical column is generally introduced between the injector and the analytical column. The short column is called guard column and is packed with material with which analytical column is packed. The guard column retains the solid particles in the sample before enters the main column. The guard columns can be replaced at regular intervals.

7.8.7 Matrices and stationary phases

It should be remembered that packing material which serves as stationary phase or support for stationary phase should be pressure stable. Three forms of column packing materials are available based on nature of the rigid solid structure :

(i) Totally porous materials or microporous supports: In these supports the micropores ramify through particles which are generally 5-10îm in diameter.

(ii) Porous layer beads or pellicular supports: these are superficially porous supports where a thin, porous, active layer is coated onto a solid core such as impervious glass beads. The thickness of the porous layer is generally 1-3μm. The size of glass beads used is between 25-50μm.

(iii) Bonded pliases : The stationary phase is chemically bonded to an inert support such as silica. The type of particular stationary phase will depend on the separation principle.

7.8.8 Detectors

Detectors are the devices which continuously monitor changes in the composition of the eluent coming out of the column. Most commonly used detectors are refractive index detector, UV detector, electrochemical and flourometric detectors.

a. **Refractive index detector (RID)** : Refractive index (RI) of dilute solutions changes proportionately with solute concentration. This relationship is exploited for quantitative detection of solutes in the column eluate. The relationship between change of RI and solute concentration is only moderately dependent on the type of solute, making this a quite universal, but not very sensitive detection principle. RID can, therefore, be applied to general purpose. This detector suffers from many defects including low sensitivity, tendency to be

affected by temperature or flow speeds and incompatibility for being used in gradient elution unless chosen solvents are of identical RI. It measures the bulk RI of sample eluent system. Hence any substance whose RI differs sufficiently from that of eluent can be detected. To attain adequate sensitivity, the temperature of the eluent and measuring cell is held constant to ± 0.001°C. Variation in flow rates also interfere with response of differential refractometer. Hence very good damping is essential for the pumps producing pulsating flow.

b. **UV-VIS-absorption detectors** : Absorbance of a solution is proportional to the concentration of the absorbing solute, the light path length and the extinction coefficient (Lambert-Beer's law). Fixed wavelength detectors utilize lamps which emit light of a few discrete wavelengths. The most common of these lamps is the low pressure mercury lamp emitting over 90 percent of its light at 254 nm. Lower wavelength lamps are available such as zinc lamps (214 nm) and cadmium lamps (229nm). The combination of the lamp and a filter determines the fixed operating wavelength of the detector. Variable wavelength (V W) detectors use a light source with a continuous emission spectrum and a continuously adjustable (narrow) band filter, called monochromator. The most common light source for these detectors is the deuterium lamp whose usable emission spectrum ranges from about 190 nm to about 350 nm, with an intensity maximum between 220 and 240 nm. Above 300 nm, the output intensity is low, therefore, some VW detectors have an additional or optional tungsten lamp, which can be used at wavelength above 350 nm. Recently, specially designed VW detectors for HPLC have been introduced which allow automatic rapid change of wavelength setting within 1 to 2 seconds or less across their entire wavelength span, which typically ranges from about 190 nm to about 600 nm.

c. **Electrochemical detectors** : Methods used in HPLC based on electroanalysis can be classified as bulk property and solute property electrochemical detectors. Bulk property electrochemical detectors respond to a change in an electrochemical property of the bulk liquid flowing through the measuring cell, whereas solute property electrochemical detectors respond to a change in voltage (potentiometry) or current (voltammetry or coulometry) when an analyte passes through the cell. The potentiometric detectors have not been

commercialized. Voltammetric and coulometric detectors on the contrary are offered by various manufacturers.

d. **Fluorescence detectors** : The quantity of fluorescent emitted from excited molecules in dilute solutions is proportional to intensity of excitation source, illuminated volume of the sample solution, quantum efficiency of fluorescence of sample and the concentration of the solute to be detected. Ideally, fluorescence radiation, as a result of suitable excitation of the sample molecule, is measured against a dark background. Therefore, the main source of detector noise is dark current noise of the photodetector, which is mainly determined by temperature. As a rule for every 10°C increase in temperature the dark current noise of the photodiode gets doubled.

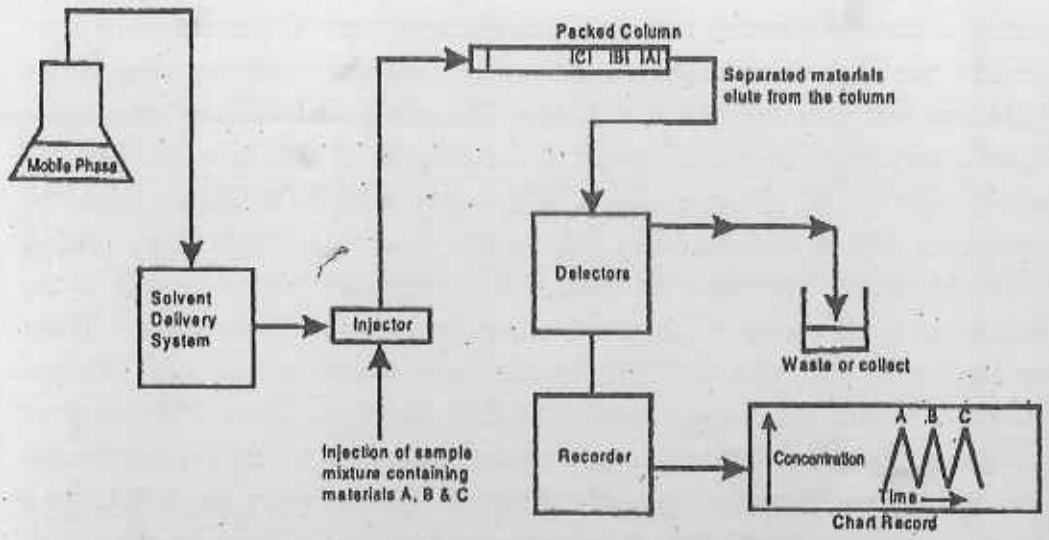


Fig. 7.13 Block diagram showing the components of an HPLC instrument

The sample to be analyzed is introduced in small volume to the stream of mobile phase and is retarded by specific chemical or physical interactions with the stationary phase as it traverses the length of the column. The amount of retardation depends on the nature of the analyte, stationary phase and mobile phase composition. The time at which a specific analyte elutes (comes out of the end of the column) is called the retention time and is considered a reasonably unique identifying characteristic of a given analyte. The use of pressure increases the linear velocity (speed) giving the components less time to diffuse within the column, leading to

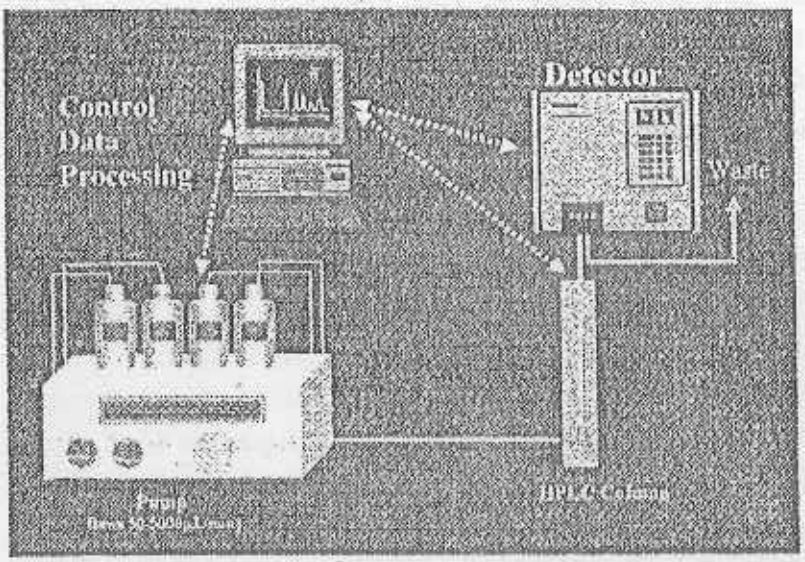


Fig. 7.14. An HPLC Setup

improved resolution in the resulting chromatogram. Common solvents used include any miscible combinations of water or various organic liquids (the most common are methanol and acetonitrile). Water may contain buffers or salts to assist in the separation of the analyte components, or compounds such as Trifluoroacetic acid which acts as an ion pairing agent.

7.8.9 Types of HPLC

Normal phase chromatography

Also known Normal phase HPLC (NP-HPLC) was the first kind of HPLC chemistry used, and separates analytes based on polarity. This method uses a polar stationary phase and a non-polar mobile phase, and is used when the analyte of interest is fairly polar in nature. The polar analyte associates with and is retained by the polar stationary phase. Adsorption strengths increase with increase in analyte polarity, and the interaction between the polar analyte and the polar stationary phase (relative to the mobile phase) increases the elution time. The interaction strength not only depends on the functional groups in the analyte molecule, but also on steric factors and structural isomers are often resolved from one another. Use of more polar solvents in the mobile phase will decrease the retention time of the analytes while more hydrophobic solvents tend to increase retention times. Particularly polar solvents in a mixture tend to deactivate the column by occupying the stationary

phase surface. This is somewhat particular to normal phase because it is most purely an adsorptive mechanism (the interactions are with a hard surface rather than a soft layer on a surface).

NP-HPLC had fallen out of favor in the 1970's with the development of reversed-phase HPLC because of a lack of reproducibility of retention times as water or organic solvents changed the hydration state of the silica or alumina chromatographic media. Recently it has become useful again with the development of HILIC bonded phases which utilize a partition mechanism which provides reproducibility. (HILIC : (Hydrophilic Interaction Chromatography or Hydrophilic Interaction LIquid Chromatography) is a version of normal phase liquid chromatography where the chromatographic mechanism for it as one of liquid-liquid partition chromatography.)

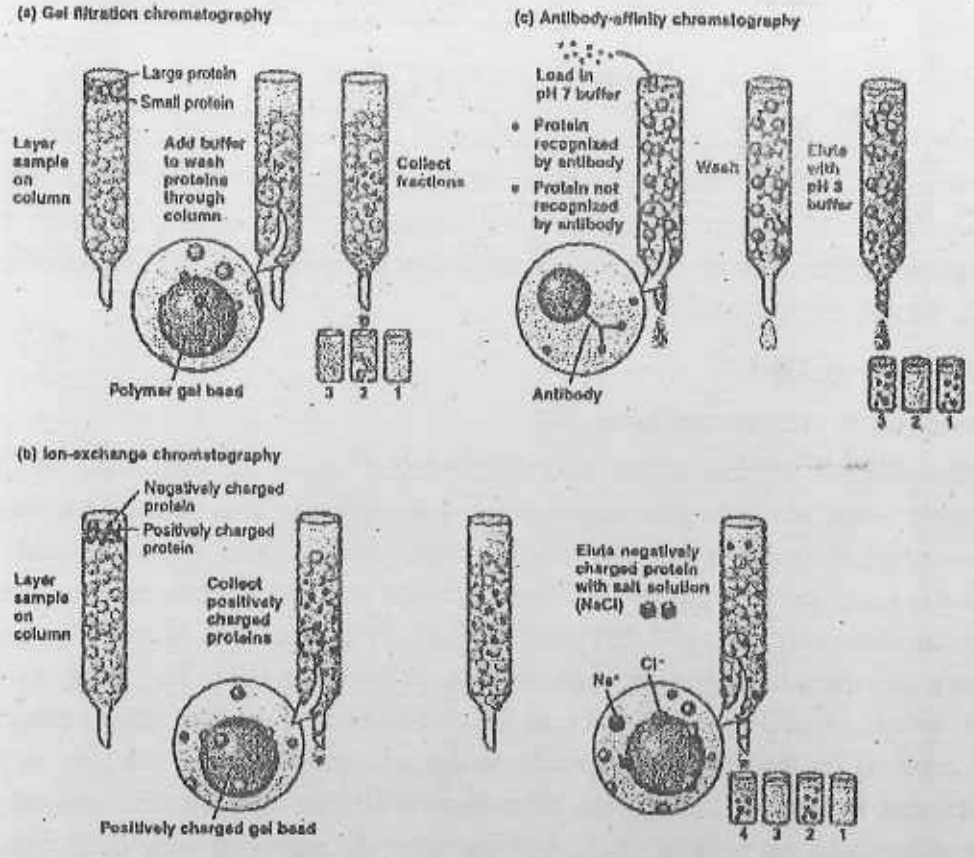


Fig. 7.15. Different chromatography at a glance

B. Reversed phase chromatography

Reversed phase HPLC (RP-HPLC or RPC) consists of a non-polar stationary phase and an aqueous, moderately polar mobile phase. One common stationary phase is a silica which has been treated with RMe_2SiCl, where R is a straight chain alkyl group such as $C_{18}H_{37}$ or C_8H_{17}. The retention time is therefore longer for molecules which are more non-polar in nature, allowing polar molecules to elute more readily. Retention Time (RT) is increased by the addition of polar solvent to the mobile phase and decreased by the addition of more hydrophobic solvent. Reversed phase chromatography (RPC) is so commonly used that it is not uncommon for it to be incorrectly referred to as "HPLC" without further specification. The pharmaceutical industry regularly employs RPC to qualify drugs before their release.RPC operates on the principle of hydrophobic interactions, which result from repulsive forces between a polar eluent, the relatively non-polar analyte, and the non-polar stationary phase. The binding of the analyte to the stationary phase is proportional to the contact surface area around the non-polar segment of the analyte molecule upon association with the ligand in the aqueous eluent. This solvophobic effect is dominated by the force of water for "cavity-reduction" around the analyte and the C18-chain versus the complex of both. The energy released in this process is proportional to the surface tension of the eluent (water : 73 erg/cm², methanol : 22 erg/cm²) and to the hydrophobic surface of the analyte and the ligand respectively. The retention can be decreased by adding less-polar solvent (MeOH, ACN) into the mobile phase to reduce the surface tension of water. Gradient elution uses this effect by automatically changing the polarity of the mobile phase during the course of the analysis.

Unit 8 ❐ Immunological Techniques

Structure

8.1 Immuno Gel Diffusion

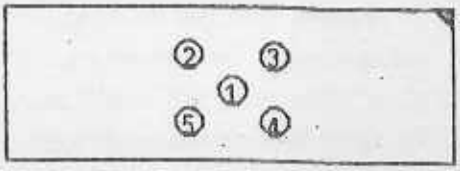
8.1.1 Introduction

Ouchterlony double immuno diffusion (or agar gel immunodiffusion) is a simple method which is still considered to be the gold standard for detection of ENAs (Extractable Nuclear Antigens). A gel plate is cut to form a series of holes in the gel. An extract of any cells (usually considered as a source of antigen) is placed in the centre well. The respective serum is then placed in one of the outer wells and the plate left for 48 hours to develop. During this time the antigens in the central well migrate in a radial fashion out of the centre well and the antibodies present in the serum migrate in a radial fashion out of the outer well. When these two meet the antibodies will bind to the antigens and form what is known as an immune complex. This immune complex precipitates in the gel to give a thin white line.

Precipitation occurs with most antigens because the antigen is multivalent i,e., has several antigenic determinants per molecule to which antibodies can bind (epitopes). Antibodies have at least two antigen binding sites, thus large aggregates or lattices of antigen and antibody are formed. Experimentally, an increasing amount of antigen is added to a constant amount of antibody in solution, initially at low antigen concentration, all of the antigen is contained in the precipitate. this is called the antibody-excess zone (Prozone phenomenon). As more antigen is added, the amount protein precipitated increases until the antigen/antibody molecules are at an

optimal ratio. This is known as the equivalence zone or equivalence point. When the amount of antigen in solution exceeds the amount of antibody, the amount of precipitation will decrease. This is known as the antigen excess zone.

The line will give a full identity (continuous line), partial identity (continuous line with a spur coming off one end–like a branch off the main line) or a non identity where the two lines cross completely.



Immuno Gel diffusion slide. Wells are made on agar coated slide. Antiserum are placed on well 1 and antigen(s) are placed on wells 2-5.

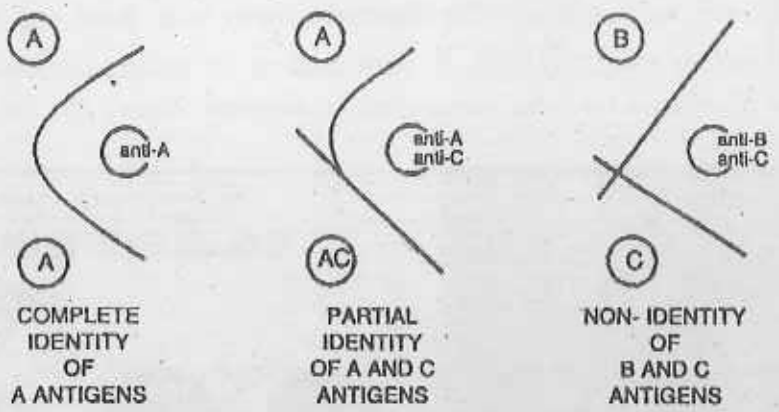


Fig. 8.1 Banding pattern for identity pattern analysis

The basis of this method lies in the fact that most antigens when they react with their antibodies are precipitated as the antigen antibody complex. If antigens and antibody are allowed to diffuse in opposite direction in a supporting medium they will form a precipitate at the zone of equivalence of antigen and antibodies forming a line. If more than two appropriately placed antigens are used and a single antiserum against both is used one of the following possibilities may occur.

1. The two precipitin lines merge in one another, showing the two antigens are identical (complete identity)
2. The two precipitin lines cross one another completely showing that the two antigens are totally different (complete non identity).
3. One of the lines crosses the other but the other does not cross the first

forming a spur, showing that there is some similarity between the two antigens though they are not identical (partial identity).

Procedure

1gm% solution of Agarose was prepared in normal saline, containing 0.01% sodium azide, by warming to 70°-80°C. The warm solution was poured on clean dry glass slide (4ml per slide) and allowed to cool and set to form a transparent gel. Wells of 3mm in diameter were cut in the gel. One of them was filled with antiserum the other with antigen so that the antigen wells were equidistant from the antiserum well. The plates were left for 48hr. for development of precipitin bands. The plates were then repeatedly washed with normal saline air dried for 48hrs, between filter paper stripes, stained with 1% Amido Black 10B for 10 minutes and finally destained within 7% acetic acid solution. The destained slides were dried and stored.

Ref. : Ouchterlony O. (1948). *In vitro* method for testing the toxin producing capacity of diphtheria bacteria, *Acta. Path. Microbiol. Scand*, 25, 186

Fig. 8.2 A typical Immunogeldiffusion plate

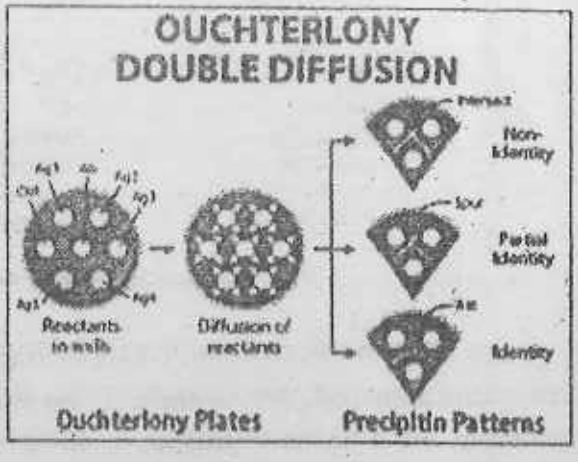


Fig. 8.3 Different precipitation pattern

8.2 Immunoelectrophoresis

This method uses two specific properties of proteins, their electrophoretic mobility and their reaction with their antibody to give a precipitate of the antigen antibody complex. Antigens are separated on an agarose gel and then allowed to react with rheir antibody to give a precipitin line as in immunogeldiffusion.

8.2.1 Procedure

1gm% solution of agarose was made in 0.05M Tris Glycine buffer pH 8.6 containing 0.01% sodium azide by warming to about to 70°-80°C on a water bath. The warm solution was poured on clean and dry slide (4ml per slide) so that a thick layer forms on the slide. The solution was then allowed to cool when it set to form a transparent gel. Two wells each 3 mm in diameter were punched 1 cm from the median line. The test samples were poured in to these wells. Electrophoresis was run at a current of 5 mA per slide for 2.5 hr. After electrophoresis.a 3 mm trough was cut along the median line and and respective antiserum was poured in to the trough. The slides were then left in a humid atmosphere in closed petridish for 48 hr for development of precipitin lines. The slides were then washed, dried, stained and destained as described in case of immunogeldiffusion.

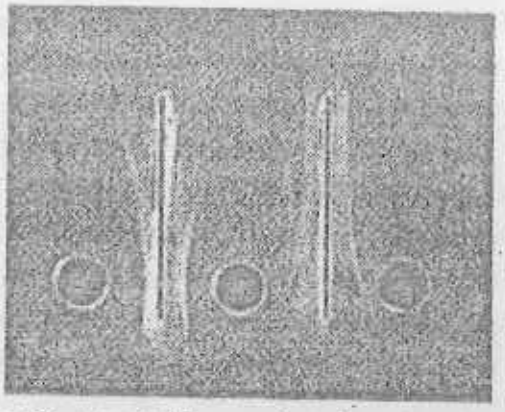


Fig. 8.4 A typical immunoelctrophoresis stained slide

Ref. Grabber P and Williums CR (1953) Methods for combined investigation of electrophoresis and immunochemical properties of a protein. *Biophysic Biochem Acta,* 10, 193

8.3 Rocket Immunoelectrophoresis (Laurell technique)

This is similar to single diffusion in agar in that the antisera is incorporated into the agar which is then poured onto a large plate to cool. Wells are cut along one end and filled with standard solutions of the antigen or with the unknown solution. An electric current is than established across the agar which drives the antigen in the direction of the current. As before, the antigen precipitates where its concentration is equal to that in the agar. The height of the "rocket" is proportional to the

concentration of antigen in the well, and the amount of antigen in the unknown well is determined by referring to the standard curve.

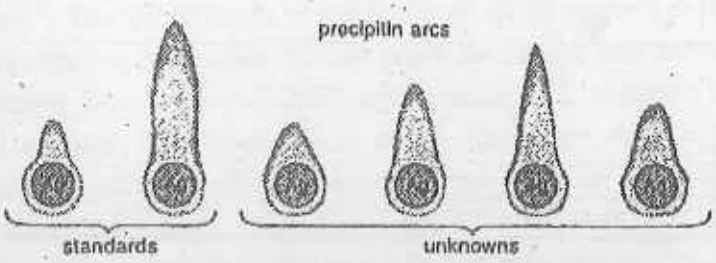


Fig. 8.5 Rocket Immunoelectrophoresis

8.4 Counter Current, Immunoelectrophoresis

There are situations in which there is a need for a rapid identification of a antigens (pathogen) but is not particularly concerned about the concentration (eg., identification of potential bacterial meningitis). Countercurrent electrophoresis is a qualitative technique that answers this purpose. The pH of solutions of antibody to the antigen of interest (e.g., *N. meningitis*) and the unknown antigen are adjusted so they have opposite charges (often about pH 8.2 since bacterial antigens are usually negatively charged at that pH and immunoglobulins are positively charged). They are placed in wells in agar and a current applied across the plate (above). A precipitin line at the equivalence point identifies the antigen as being from the pathogen.

Fig. 8.6. Counter current immunoelectrophoresis

8.5 Single Radial Immunodiffusion (SRID)

8.5.1 Introduction

It is a technique for quantitating soluble proteins that involves placing the solution to be measured into a well cut into an agar or agarose gel containing

antiserum specific for the protein. As the solution to be measured diffuses out of the well, it complexes with the antiserum and forms a ring, the size of which is proportional to the quantity of soluble protein in the well. Abbreviated SRID. Also known as Mancini method.

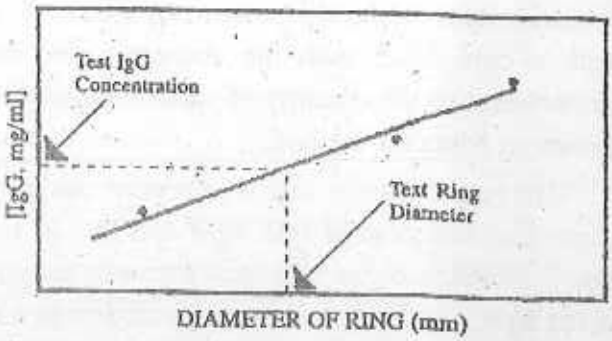
This is a technique that is in routine use in hospital labs to measure such things as the concentration of IgG, IgM and IgA in a patient's serum. A small amount of agar is melted and cooled to just above its set point. Antisera to human IgG is added to the agar which is mixed and poured onto a small plate, often about the size of a microscope slide. Small holes (1-2 mm diameter) are cut in the gel about 2 cm apart. The well (hole) is filled with serum or a known amount of IgG, and the plate is covered and allowed to stand for 24 hours (sometimes less). As the IgG diffuses out of the well, it reacts with the anti-IgG in the agar and precipitates. The precipitate is white and can be seen in the clear agar as a white circle around the well. (It is sometimes stained with a blue dye to make it even more visible.) The diameter of the circle is directly proportional to the concentration of IgG in the sample. A standard curve is made using known concentrations of IgG, and the concentration of the unknown is determined by comparison of the size of the circle with the standard curve.

8.5.2 Protocol

1. Align the Ab-agar slide so that the cut away corner of the gel is in the **upper right corner**. Use the micropipets to fill the wells. The liquid will fill the pipet by capillary action. The gel will draw the liquid out of the pipet.
2. Fill the first well on the LEFT almost level to the agar surface with the test serum. Do Not Overfill.

3. In the same manner, fill the remaining 3 wells, left to right with the 3 reference sera. Identify each well. Do not refill these wells! this test measures how much igg is present in the test serum and the volume must be the same for all of the wells in order for the results to be accurate.
4. Carefully place the agar slide back in the humidity box. They will be returned after 24 hours.
5. At the next lab period, use the millimeter ruler to measure the diameter of each precipitin ring.

6. Determine the concentration of the unknown test serum by drawing a Standard Curve–plotting the known concentrations of the reference sera versus their diffusion diameters on 2-cycle semi-log paper (see example).



7. The plot of the reference sera data should give a STRAIGHT LINE, and the concentration of the IgG in the test serum can be accurately determined.

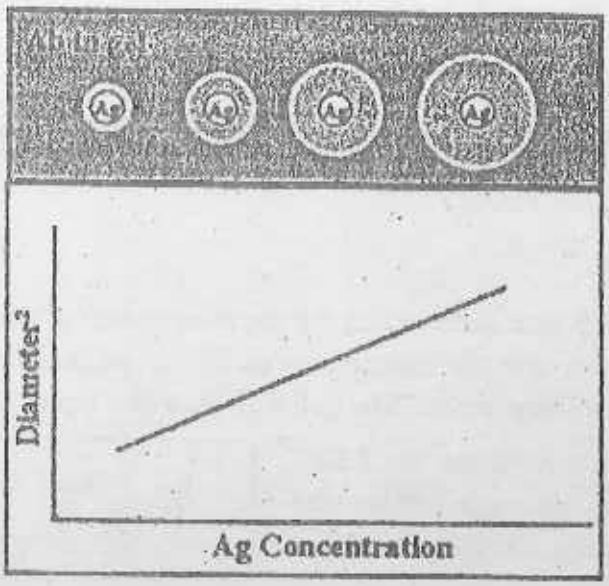


Fig. 8.7 Graphical representation of SRID

8.6 Agglutination reactions

The interaction between antibody and a particulate antigen results in visible clumping called ***agglutination***. The agglutination reaction is similar in principle to the precipitation reaction. Just as antibody excess inhibits precipitation reactions, an excess of antibody inhibits agglutination reactions; this inhibition, called the ***prozone***

effect, can be caused by several mechanisms. First, high levels of antibody increase the likelihood that a single antibody molecule will bind to two or more epitopes on a single particulate antigen rather than cross-linking epitopes on two or more particulate antigens. The prozone effect can also occur at high concentrations of antibodies that bind to the antigen but do not induce agglutination; these antibodies, called ***incomplete antibodies***, are often of the IgG class. At high concentrations of IgG, incomplete antibodies may occupy all of the antigenic sites, thus blocking access by IgM, which is a good agglutinator. The lack of agglutinating activity of an incomplete antibody may be due to restricted flexibility in the hinge region, making it difficult for the antibody to assume the required angle for optimal cross-linking of epitopes on two or more particulate antigens. Alternatively, the density of epitope distribution or the location of some epitopes in deep pockets of a particulate antigen may make it difficult for antibodies, specific for these epitopes, to agglutinate certain particulate antigens.

8.6.1 Hemagglutination

Agglutination reactions are routinely performed to type red blood cells (RBCs). In typing for the ABO antigens, RBCs are mixed on a slide with antisera to the A and B blood-group antigens. If the antigen is present on the cells, they agglutinate, forming a visible clump on the slide. Determination of which antigens are present

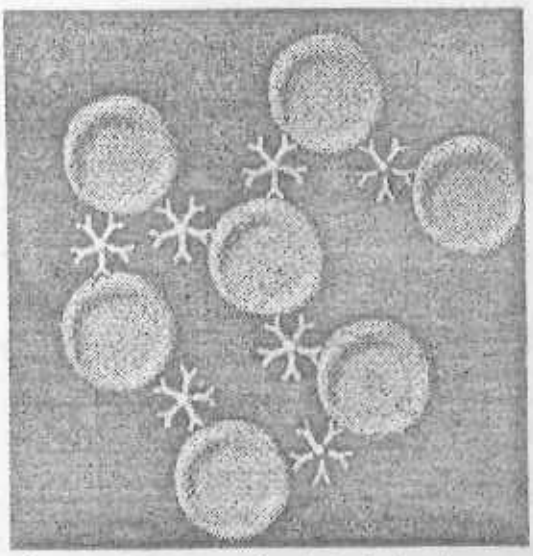


Fig. 8.8 IgM mediated Agglutination

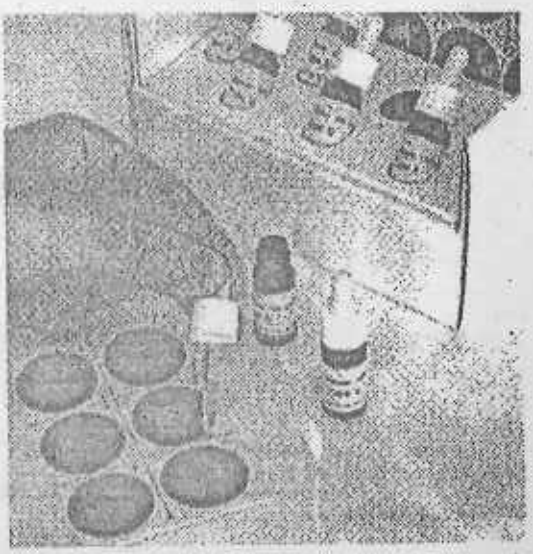


Fig. 8.9 Blood Group Testing

on donor and recipient RBCs is the basis for matching blood types for transfusions. At neutral pH, red blood cells are surrounded by a negative ion cloud that makes the cells repel one another ; this repulsive force is called ***zeta potential***. Because of its size and pentameric nature, IgM can overcome the zeta potential and cross-link red blood cells, leading to agglutination.The smaller size and bivalency of IgG makes it less able to overcome the zeta potential. For this reason, IgM is more effective than IgG in agglutinating red blood cells. Antibodies to some RBC antigens (e.g; the Rh antigen) are of the IgG class exclusively. In order to agglutinate Rh^+ red blood cells with anti-Rh antibody, the zeta potential must be reduced. This is commonly done by placing the red blood cells in serum, which has a high net negative charge that reduces the effect of the negative ion cloud surrounding the red cells, thus allowing anti-Rh antibody to agglutinate Rh^+ cells.

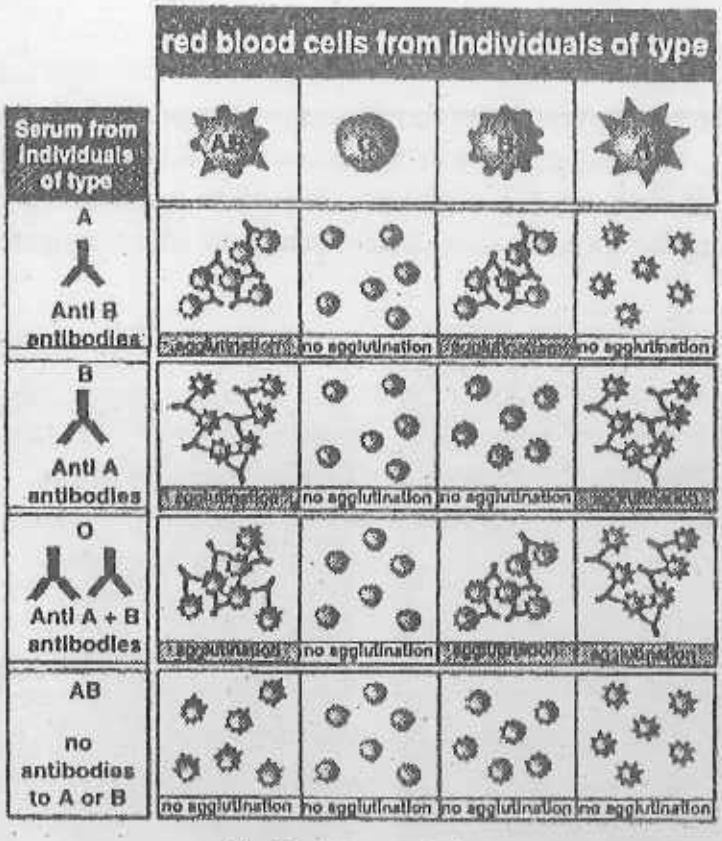


Fig. 8.10 Hemagglutination

8.6.2 Bacterial agglutination

A bacterial infection often elicits the production of serum antibodies specific for surface antigens of the bacterial cells. The presence of such antibodies can be detected by bacterial agglutination reactions. Serum from a patient thought to be infected with a given bacterium is serially diluted in a series of tubes to which the bacteria is added. The last tube showing visible agglutination will reflect the serum antibody titer of the patient. The agglutination titer is defined as the reciprocal of the last serum dilution that elicits a positive agglutination reaction. For example, if serial two fold dilutions of serum are prepared and if the dilution of 1/640 shows agglutination but the dilution of 1/1280 does not, then the agglutination titer of the patient's serum is 640. For some bacteria high-titer serum can be diluted up to 1/50,000 and still show agglutination.

The agglutination titer of an antiserum can be used to diagnose a bacterial infection. Patients with typhoid fever, for example, show a significant rise in the agglutination titer to *Salmonella typhi*. Agglutination reactions also provide a way to type bacteria. For instance, different species of the bacterium *Salmonella* can be distinguished by agglutination reactions with a panel of typing antisera.

8.6.3 Passive agglutination

The sensitivity and simplicity of agglutination reactions can be extended to soluble antigens by the technique of *passive agglutination*. In this technique, a soluble antigen is mixed with red blood cells that have been treated with tannic acid or chromium chloride, both of which promote adsorption of the antigen to the surface of the cells. Serum containing antibody is serially diluted into microtiter plate wells, and the antigen-coated red blood cells are added to each well; agglutination is assessed by the size of characteristic spread pattern of agglutinated red blood cells on the bottom of the well.

Passive hemagglutination is far more sensitive than precipitin reactions and can detect antibody concentrations as low as 0.001μg/ml. the sensitivities of precipitation and hemagglutination can be compared by testing an antiserum to hen ovalbumin in a tube-precipitation reaction and in passive hemagglutination with ovalbumin-coated red blood cells. Dilution of the antiserum by 1:5 results in loss of precipitation ability, whereas the antiserum still functions in passive agglutination out to a dilution of 1 : 10,000. Antigen can also be coupled to particles of latex or the mineral colloid bentonite.

8.6.4 Agglutination inhibition

A modification of the agglutination assay, called agglutination inhibition, is a highly sensitive assay to detect small quantities of an antigen. One type of pregnancy test uses latex particles coated with human chorionic gonadotropin (HCG) and antibody to HCG. The addition of urine from a pregnant woman, which contains HCG, inhibits agglutination of the latex particles, and so the absence of agglutination indicates pregnancy. Agglutination inhibition can also be used to determine if an individual is using certain types of illegal drugs such as cocaine or heroin. A urine or blood sample containing the suspected drug is first incubated with antibody specific for the drug. Then red blood cells or other particles coated with the drug are added. If the red blood cells are not agglutinated by the antibody, then it suggests that the individual may have been using the illicit drug. One problem with these tests is that some legal drugs have chemical structures similar to those of illicit drugs, and these legal drugs may crossreact with the antibody giving a false positive reaction. For this reason a positive reaction must be confirmed by a nonimmunologic method.

Agglutination inhibition is also widely used in clinical laboratories to determine if an individual has been exposed to certain types of viruses that cause agglutination of red blood cells. If an individual's serum contains specific antiviral antibodies, then the antibodies will bind to the virus and interfere with hemagglutination by the virus. The reciprocal of the last serum dilution to show inhibition of rubella hemagglutination

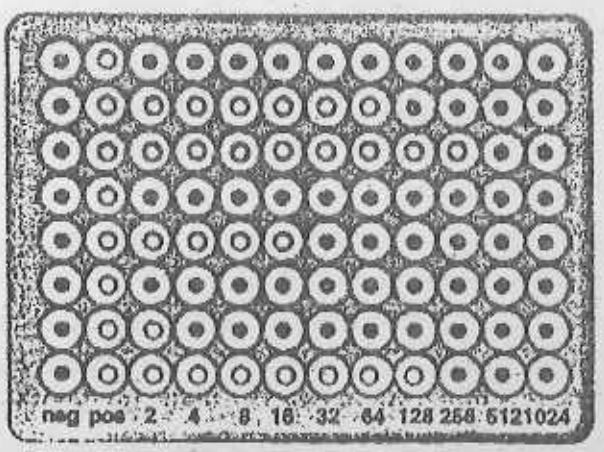


Fig. 8.11 : A typical IHA Plate showing RBC agglutination

is the titer of the serum. A titer greater than 10 (1 : 10 dilution) indicates that a woman is immune to rubella, whereas a titer of less than 10 is indicative of a lack of immunity and the need of immunization with the rubella vaccine.

8.7 Enzyme Linked Immunosorbant Assay (ELISA)

8.7.1 Process of ELISA

Enzyme-Linked ImmunoSorbent Assay, or ELISA, is a biochemical technique used mainly in immunology to detect the presence of an antibody or an antigen in a sample. The ELISA has been used as a diagnostic tool in medicine and plant pathology, as well as a quality control check in various industries. In simple terms, in ELISA an unknown amount of antigen is affixed to a surface, and then a specific antibody is washed over the surface so that it can bind to the antigen. This antibody is linked to an enzyme, and in the final step a substance is added that the enzyme can convert to some detectable signal. Thus in the case of fluorescence ELISA, when light is shone upon the sample, any antigen/antibody complexes will fluoresce so that the amount of antigen in the sample can be measured.

Performing an ELISA involves at least one antibody with specificity for a particular antigen. The sample with an unknown amount of antigen is immobilized on a solid support (usually a polystyrene microtiter plate) either non-specifically (via adsorption to the surface) or specifically (via capture by another antibody specific to the same antigen, in a "sandwich" ELISA). After the antigen is immobilized the detection antibody is added, forming a complex with the antigen. The detection antibody can be covalently linked to an enzyme, or can itself be detected by a secondary antibody which is linked to an enzyme through bioconjugation. Between each step the plate is typically washed with a mild detergent solution to remove any proteins or antibodies that are not specifically bound. After the final wash step the plate is developed by adding an enzymatic substrate to produce a visible signal, which indicates the quantity of antigen in the sample. Older ELISAs utilize chromogenic substrates, though newer assays employ fluorogenic substrates with much higher sensitivity.

8.7.2 Types

A. ***Indirect ELISA*** : The steps of the general, "indirect," ELISA for determining serum antibody concentrations are :

1. Apply a sample of known antigen of known concentration to a surface, often the well of a microtiter plate. The antigen is fixed to the surface to render it

immobile. Simple adsorption of the protein to the plastic surface is usually sufficient. These samples of known antigen concentrations will constitute a standard curve used to calculate antigen concentrations of unknown samples. Note that the antigen itself may be an antibody.

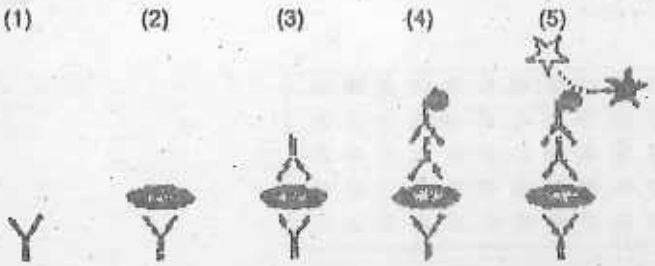
2. The plate wells or other surface are then coated with serum samples of unknown antigen concentration, diluted into the same buffer used for the antigen standards. Since antigen immobilization in this step is due to non-specific adsorption, it is important for the total protein concentration to be similar to that of the antigen standards.
3. A concentrated solution of non-interacting protein, such as Bovine Serum Albumin (BSA) or casein, is added to all plate wells. This step is known as blocking, because the serum proteins block non-specific adsorption of other proteins to the plate.
4. The plate is washed, and a detection antibody specific to the antigen of interest is applied to all plate wells. This antibody will only bind to immobilized antigen on the well surface, not to other serum proteins or the blocking proteins.
5. The plate is washed to remove any unbound detection antibody. After this wash, only the antibody-antigen complexes remain attached to the well.
6. Secondary antibodies, which will bind to any remaining detection antibodies, are added to the wells. These secondary antibodies are conjugated to the substrate-specific enzyme. This step may be skipped if the detection antibody is conjugated to an enzyme.
7. Wash the plate, so that excess unbound enzyme-antibody conjugates are removed.
8. Apply a substrate which is converted by the enzyme to elicit a chromogenic or fluorogenic or electrochemical signal.
9. View/quantify the result using a spectrophotometer, spectrofluorometer, or other optical/electrochemical device.

The enzyme acts as an amplifier; even if only few enzyme-linked antibodies remain bound, the enzyme molecules will produce many signal molecules. A major disadvantage of the indirect ELISA is that the method of antigen immobilization is non-specific; any proteins in the sample will stick to the microtiter plate well, so small

concentrations of analyte in serum must compete with other serum proteins when binding to the well surface. The sandwich ELISA provides a solution to this problem.

ELISA may be run in a qualitative or quantitative format. Qualitative results provide a simple positive or negative result for a sample. The cutoff between positive and negative is determined by the analyst and may be statistical. Two or three times the standard deviation is often used to distinguish positive and negative samples. In quantitative ELISA, the optical density or fluorescent units of the sample is interpolated into a standard curve, which is typically a serial dilution of the target.

B. Sandwich ELISA : In sandwich ELISA (1) Plate is coated with a capture antibody; (2) sample is added, and any antigen present binds to capture antibody; (3) detecting antibody is added, and binds to antigen; (4) enzyme-linked secondary



antibody is added, and binds to detecting antibody; (5) substrate is added, and is converted by enzyme to detectable form.

A less-common variant of this technique, called "sandwich" ELISA, is used to detect sample antigen. The steps are as follows:

1. Prepare a surface to which a known quantity of capture antibody is bound.
2. Block any non specific binding sites on the surface.
3. Apply the antigen-containing sample to the plate.
4. Wash the plate, so that unbound antigen is removed.
5. Apply primary antibodies that bind specifically to the antigen.
6. Apply enzyme-linked secondary antibodies which are specific to the primary antibodies.
7. Wash the plate, so that the unbound antibody-enzyme conjugates are removed.
8. Apply a chemical which is converted by the enzyme into a color or fluorescent or electrochemical signal.

9. Measure the absorbance or fluorescence or electrochemical signal (e.g., current) of the plate wells to determine the presence and quantity of antigen.

The image to the right includes the use of a secondary antibody conjugated to an enzyme, though technically this is not necessary if the primary antibody is conjugated to an enzyme. However, use of a secondary-antibody conjugate avoids the expensive process of creating enzyme-linked antibodies for every antigen one might want to detect. By using an enzyme-linked antibody that binds the Fc region of other antibodies, this same enzyme-linked antibody can be used in a variety of situations. The major advantage of a sandwich ELISA is the ability to use crude or impure samples and still selectively bind any antigen that may be present. Without the first layer of "capture" antibody, any proteins in the sample (including serum proteins) may competitively adsorb to the plate surface, lowering the quantity of antigen immobilized.

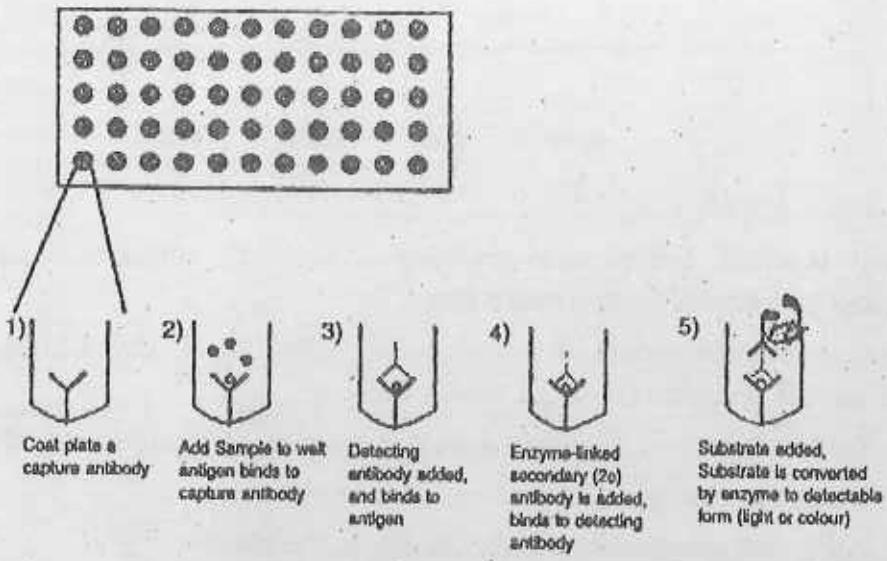


Fig. 8.12 : Sandwich ELISA assay method

C. ***Competitive ELISA*** : A third use of ELISA is through competitive binding. The steps for this ELISA are somewhat different than the first two examples:

1. Unlabeled antibody is incubated in the presence of its antigen.
2. These bound antibody/antigen complexes are then added to an antigen coated well.

3. The plate is washed, so that unbound antibody is removed. (The more antigen in the sample, the less antibody will be able to bind to the antigen in the well, hence "competition.")
4. The secondary antibody, specific to the primary antibody is added. This second antibody is coupled to the enzyme.

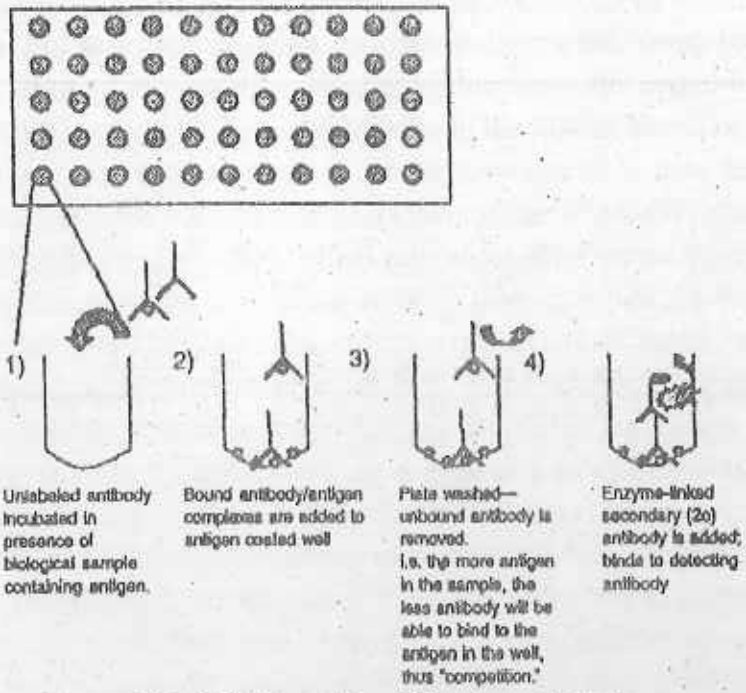


Fig. 8.13 Competitive Elisa Assya Method

In competitive ELISA, the higher the original antigen concentration, the weaker the eventual signal.

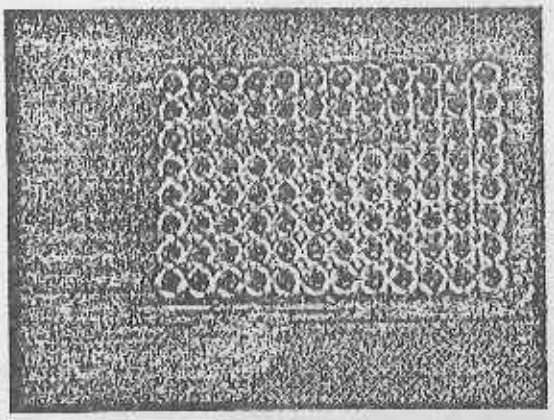


Fig. 8.14 A 96-well microtiter plate being used for ELISA.

5. A substrate is added, and remaining enzymes elicit a chromogenic or fluorescent signal.

The purpose of an ELISA is to determine if a particular protein is present in a sample and if so, how much. There are two main variations on this method: you can determine how much antibody is in a sample, or you can determine how much protein is bound by an antibody. The distinction is whether you are trying to quantify an antibody or some other protein. In this example, we will use an ELISA to determine how much of a particular antibody is present in an individuals blood. ELISAs are performed in 96-well plates which permits high throughput results. The bottom of each well is coated with a protein to which will bind the antibody you want to measure. Whole blood is allowed to clot and the cells are centrifuged out to obtain the clear serum with antibodies (called primary antibodies). The serum is incubated in a well, and each well contains a different serum (see figure below). A positive control serum and a negative control serum would be included among the 96 samples being tested.After some time, the serum is removed and weakly adherent antibodies are washed off with a series of buffer rinses. To detect the bound antibodies, a secondary antibody is added to each well. The secondary antibody would bind to all human antibodies and is typically produced in a rodent. Attached to the secondary antibody is an enzyme such as peroxidase or alkaline phosphatase. These enzymes can metabolize colorless substrates (sometimes called chromagens) into colored products. After an incubation period, the secondary antibody solution is removed and loosely adherent ones are washed off as before. The final step is the addition the enzyme substrate and the production of colored product in wells with secondary antibodies bound.When the enzyme reaction is complete, the entire plate is placed into a plate reader and the optical density (i.e. the amount of colored product) is determined for each well. The amount of color produced is proportional to the amount of primary antibody bound to the proteins on the bottom of the wells.

8.8 Radio Immunoassay

8.8.1 Introduction

A highly sensitive and specific assay method that uses the competition between radiolabeled and unlabeled substances in an antigen-antibody reaction to determine the concentration of the unlabeled substance; it can be used to determine antibody concentrations or to determine the concentration of any substance against which

specific antibody can be produced. (Abbreviated RIA). RIA a scientific method used to tes' antigens (for example, hormone levels in the blood) without the need to use a bioassay. It involves mixing known quantities of radioactive antigen (frequently labeled with gamma-radioactive isotopes of iodine attached to tyrosine) with antibody to that antigen, then adding unlabeled or "cold" antigen and measuring the amount of labeled antigen displaced. Radioimmunoassay (RIA) is a sensitive method for measuring very small amounts of a substance in the blood. Radioactive versions of a substance, or isotopes of the substance, are mixed with antibodies and inserted in a sample of the patient's blood. The same non-radioactive substance in the blood takes the place of the isotope in the antibodies, thus leaving the radioactive substance free.The amount of free isotope is then measured to see how much of the original substance was in the blood. This isotopic measuring method was developed in 1959 by two Americans, biophysicist Rosalyn Yalow (1921-) and physician Solomon A. Berson (1918-1972).Yalow and Berson developed the first radioisotopic technique to study blood volume and iodine metabolism. They later adapted the method to study how the body uses hormones, particularly insulin, which regulates sugar levels in the blood. The researchers p. ɔved that Type II (adult onset) diabetes is caused by the inefficient use of insulin. Previously, it was thought that diabetes was caused only by a lack of insulin.In 1959 Yalɔw and Berson perfected their measurement technique and named it radioimmunoassay (RIA). RIA is extremely sensitive. It can measure one trillionth of a gram of material per milliliter of blood. Because of the small sample required for measurement, RIA quickly became a standard laboratory tool.

As an example of how this technique works, let's apply it to insulin. To measure insulin, the first step is to mix known amounts of radioisotope-tagged insulin and antibodies. These combine chemically. Next, a small amount of the patient's blood is added. The insulin contained in the blood displaces some of the tagged insulin. The free-tagged insulin is then measured with isotope detectors and the patient's insulin level is calculated.

RIA has many uses, including narcotics (drug) detection, blood bank screening for the hepatitis (a highly contagious condition) virus, early cancer detection, measurement of growth hormone levels, tracking of the leukemia virus, diagnosis and treatment of peptic ulcers, and research with brain chemicals called neurotransmitters.

8.8.2 Protocol

Two following basic techniques are used :

1. Competitive RIA
2. Excess reagent RIA

8.8.3 Competitive RIA

This technique is of common use. It is performed as follows :

A constant amount of antibody (known) to the antigen to be assayed is added to a series of tubes. A small amount of Radio labelled antigen plus increasing amounts of unlabelled antigen are then added to the tubes. After a period of incubation , the antibody bound (labeled & unlabelled) antigen is separated from the free or unbound antigen by one of the several methods such as precipitation by 50% saturated ammonium sulfate, or by anti-globulin. The amount of radioactivity of precipitate (bound labelled antigen) and of supernatant (free labelled antigen) are determined.A graph is plotted to shoe relationship of amount of unlabelled antigen added to the ratio of bound& free labelled antigen. From this curve the concentration of unknown antigen in test sample can be determined by determining the ratio of bound & free labelled antigen in test procedure performed by adding test sample to the tube containing antibody & labelled antigen in amounts used for preparing graph.This technique has been used to measure levels of many antigens, drugs, hormones etc in human body fluids.

8.8.4 Excess reagent RIA

In excess reagent technique the antibody is usually labelled and used in excess concentration. Analyte is measured by determining the radioactivity in the conjugate, Which is separated as described in competitive RIA.

8.8.5 Solid phase RIA

Polystyrene microlitre plates or beads are used as solid phase. Wells or beads are coated with specific antibodies to the antigen to be assayed. These are treated with specific antibodies to the antigen to be assayed. These are treated with test sample and then with radio labelled antibodies in known but excess amount. The absorbed radioactivity is measured & compared with similar activity in similar test with negative control. samples with residual counts 2.1 times or greater than that on negative control are considered positive. Similarly known antibodies can be detected by using antigen coated wells

8.8.6 Detection of antibodies by solid phase RIA

A. Competitive binding RIA

The solid phase is coated with known antigen & then treated with mixture of test sample & known amount of labelled antibody. The radioactivity of bound labelled antibody is determined. It is inversely proportional to the amount of antibody in the test sample.

B. Blocking or Inhibition RIA

In this method the test sample is first incubated with known quantity of antigen, then the mixture is transferred to solid phase coated with antibodies & treated with labelled antibodies. The amount of labelled antibodies will be less if the test sample contains less antibodies compared with negative control a reduction in counts of 40% or more is considered as positive.

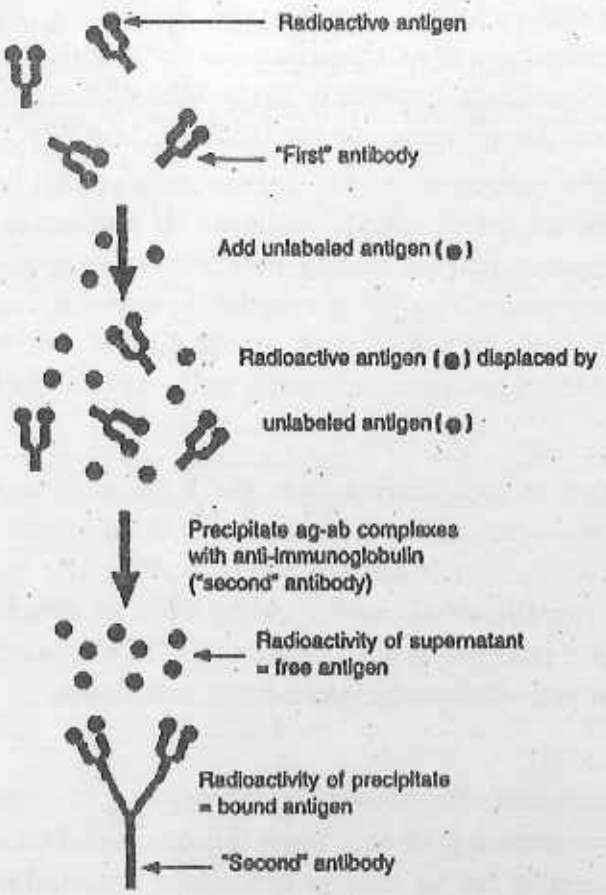


Fig. 8.15 Process of blocking or inhibition of RIA

Unit 9 ❐ Fluorochrome and Immunofluorescence

Structure

9.1 Introduction

A **fluorophore**, in analogy to a chromophore, is a component of a molecule which causes a molecule to be fluorescent. It is a functional group in a molecule which will absorb energy of a specific wavelength and re-emit energy at a different (but equally specific) wavelength. The amount and wavelength of the emitted energy depend on both the fluorophore and the chemical environment of the fluorophore. This technology has particular importance in the field of biochemistry and protein studies, eg. in immunofluorescence and immunohistochemistry.

It was shown by Albert coon in 1944 that antibodies could be with fluorescent dye or fluorochrome to detect specific antigens. If antibodies are tagged with fluorochromes, immune complexes bearing these lebelled antibodies can be detected by fluorescence study using light of appropriate wavelength. Antibody molecules bound to specific antigens in cells and tissue can similarly be visualized. In this immunofluorescence technique some commonly used fluorochromes are given bellow.

Fluorescein

A fluorescing dye, an acid fluorochrome, that is the most widely used label for immunofluorescence procedure. The fluorescence of this molecule is very high, and excitation occurs at 494 nm and emission at 521. Sodium salt of this compound is used in solution to reveal corneal lesions and as a test of circulation in the retina and extremities. The isothiocyanate derivative (FITC) is used for labeling of immunoglobulins in various immunofluorescence techniques.

Rhodamine

This is another organic dye, absorbs in the yellow green range (515 nm) and emits a deep red fluorescence (546 nm). Since this compound emits at alonger wave length than fluorescein, it can be used in two colour immunofluorescence assay.

Phycoerythrin is an efficient absorber of light and a brilliant emitter of red fluorescence, stimulating its use as a label for immunofluorescence.

Eosin is a fluorescent red dye resulting from the action of bromine on fluorescein. It can be used to stain cytoplasm, collagen and muscle fibers for examination under the microscope.

There are many fluorescein derivatives, for example fluorescein isothiocyanate, often abbreviated as FITC. FITC is the original fluorescein molecule functionalized with an isothiocyanate group (-N=C=S), replacing a hydrogen atom on the bottom ring of the structure. This derivative is reactive towards amine groups on proteins inside cells. A succinimidyl-ester functional group attached to the fluorescein core, creating NHS-fluorescein, forms another common amine reactive derivative.

A newer generation of fluorochromes such as the Alexa Fluors and the DyLight Fluors are generally more photostable, brighter, and less pH-sensitive than other standard dyes of comparable excitation and emission.

9.2 Flow Cytometry

Flow cytometry is a method for quantitating components or structural features of cells primarily by optical means. It is a technique for counting, examining and sorting microscopic particles suspended in a stream of fluid. Flow cytometry allows simultaneous multiparametric analysis of the physical and/or chemical characteristics of single cells flowing through an optical/electronic detection apparatus. Although it makes measurements on one cell at a time, it can process thousands of cells in a few seconds. Since different cell types can be distinguished by quantitating structural features, flow cytometry can be used to count cells of different types in a mixture.

9.2.1 Principles of flow cytometry

Flow cytometers involve sophisticated fluidics, laser optics, electronic detectors, analog to digital converters, and computers. In flow cytometry analysis a beam of light (usually laser light) of a single frequency is directed onto a hydrodynamically focused stream of monodisperse suspension of cells. Each suspended cell, passing through the beam, scatters the light in some way, and fluorescence chemicals in the cell may be excited into emitting light at a lower frequency than the light source. A number of detectors are aimed at the point where the stream passes through the light beam; one in line with the light beam (Forward Scatter or FSC) and several perpendicular to it (Side Scatter (SSC) and one or more fluorescence detectors).

This combination of scattered and fluorescence light is picked up by the detectors in flow cytometry. By analyzing fluctuations in brightness at each detector it is possible to deduce various facts about the physical and chemical structure of each individual particle.

Light scatter alone is often quite useful. It is commonly used to exclude dead cells, cell aggregates, and cell debris from the fluorescence data. It is sufficient to distinguish lymphocytes from monocytes from granulocytes in blood leukocyte samples. Fluorescence intensities are typically measured at several different wavelengths simultaneously for each cell. Fluorescent probes are used to report the quantities of specific components of the cells.

9.2.2 Applications

Following parameters can be monitored using flow cytometry :

1. volume and morphological complexity of cells
2. cell pigments
3. DNA (cell cycle analysis, cell kinetics, proliferation etc.)
4. chromosome analysis and sorting (library construction, chromosome paint)
5. proteins, cell surface antigens (CD markers)
6. intracellular antigens (various cytokines, secondary mediators etc.)
7. nuclear antigens, enzymatic activity
8. pH, intracellular ionized calcium, magnesium, membrane potential, membrane fluidity
9. apoptosis, cell viability, monitoring electropermeabilization of cells
10. characterising multi-drug resistance (MDR) in cancer cells
11. various combinations (DNA / surface antigens etc.)

The most numerous flow cytometers are those used for complete blood cell counts in clinical laboratories, these do not employ fluorescence. More versatile research instruments employ fluorescence, hence may be distinguished as flow cytofluorometers. Ploidy and cell cycle analysis of cancers is the major diagnostic use. Lymphomas and leukemias are intensively studied for surface markers of diagnostic and prognostic value. Although less expensive alternative technologies are under development, until the present time, flow cytometry has been the method of choice for monitoring CD4 lymphocyte levels in the blood of AIDS patients.

9.3 The Fluorescence-activated cell sorting-FACS

9.3.1 FACS basics

The acronym FACS stands for Fluorescence-Activated Cell Sorting (aka. Fluorescence-Assisted Cell Sorting). It is a type of flow cytometry in which a heterogeneous population of suspended cells are characterized and separated based upon the intensity of fluorescence they emit while passing single file through an illuminated volume. FACS was the first flow cytometric technology.

9.3.2 Fluorescence-activated cell sorting principles

The process begins by placing the cells labeled with a fluorescent dye into a flask and forcing the cells to enter a small nozzle one at a time (fig. 9.1). The dye is coupled to a monoclonal antibody and binds to those cells coated with the antigen for which the antibody is specific. The cells travel down the nozzle which is vibrated at an optimal frequency (at some 40,000 cycles per second) to produce 40,000 discrete droplets each second at fixed distance from the nozzle. As the cells flow down the stream of liquid, they are scanned by a laser (blue light in figure 9.1). Just before the stream breaks into droplets the flow passes through a fluorescence measuring station where the fluorescence character of interest of each cell is measured. As each labeled cell passes through the beam, its resulting fluorescence is detected by a photocell. Some of the laser light is scattered (red cone emanating from the red cell) by the cells and this is used to count the cells. This scattered light can also be used to measure the size of the cells. An electrical charging ring is placed just at the point

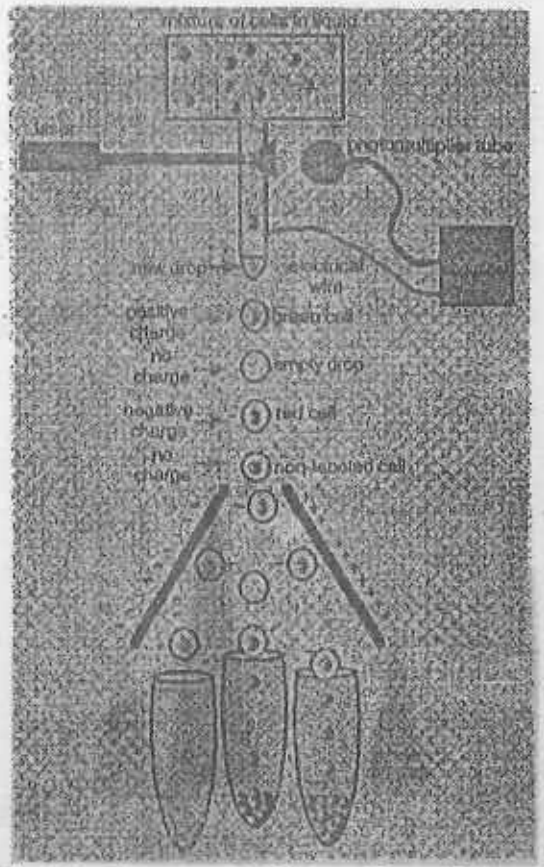


Fig. 9.1. Diagram of FACS machine cells have been fluorescent tagged with either red of green antibodies, though not every cell expresses the epitope and therefore some are not tagged either color.

where the stream breaks into droplets. A charge is placed on the ring based on the immediately prior fluorescence intensity measurement and the opposite charge is trapped on the droplet as it breaks from the stream. The droplets retain this charge as they pass between a pair of charged metal plates. The positively-charged drops are attracted to the negatively-charged plate and vice versa. Uncharged droplets (those that contain no cell or a cell that fails to meet the desired criteria of fluorescence and size) pass straight into a third container and are later discarded. This apparatus can sort as many as 300,000 cells per minute. The cells are not damaged by the process. In fact, because the machine can be set to ignore droplets containing dead cells, the percent viability of the sorted cells can be higher than that in the original suspension.

The separation of a subpopulation of cells can be done by tagging those of interest with an antibody linked to a fluorescent dye. The antibody is bound to a protein that is uniquely expressed in the cells of interest. The laser light excites the dye which emits a color of light that is detected by the photomultiplier tube, or light detector. By collecting the information from the light (scatter and fluorescence) a computer can determine which cells are to be separated and collected.

9.3.3 Quantifying FACS data

FACS data collected by the computer can be displayed in two different ways. What we want to know is how many cells of each color were sorted. In the first example (Fig. 9.2), we see the intensity of the green or red fluorescence is plotted on the X-axis and the number of cells with each level of flourescence is plotted on the Y-axis. In this example, there were twice as many red cells sorted as green or unlabeled cells, but the level of light was greater from the green cells than the red cells. This method is best if all cells are either green, red or unlabeled and no cells are labeled both colors.

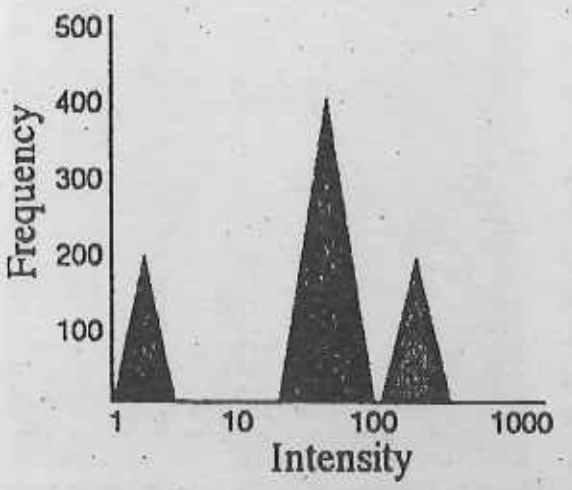


Fig. 9.2 Quantifying FACS data. This graph shows the number of cells (X-axis) and level of fluorescence emitted (Y-axis) by the labeled cells.

Common FACS applications

FACS can be used to distinguish living versus dead cells, different types of cells (i.e. WBC's), isolate cells for Cloning or PCR, separating sperm for gender selection, or for identifying apoptotic cells.

9.3.5 Separating living from dead cells by FACS

Dead cells are permeable to ethidium monoazide (EMA), a chemical that can enter the cell and bind to the DNA under a fluorescent lamp. As the FACS is run, cells containing EMA (dead) are separated out.

9.4 References

1. Shapiro, H.M. (2003) *Practical Flowcytometry*, 4th Edn.John Willey and Sons, New York.
2. Spector D.L. and Goldman R.D (2004)*Live Cell imaging: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
3. *Principles and Techniques of Biochemistry and Molecular Biology*, (2005) Edited by Wilson K. and Walker J, 6th Edn, Cambridge University Press

Unit 10 ❐ Elementary Idea of Bioinformatics

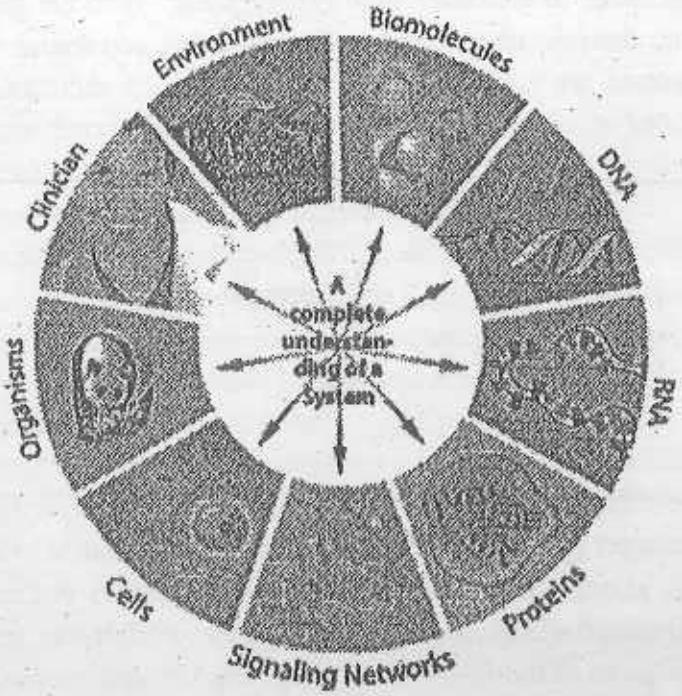
Structure

1.1 Introduction

Bioinformatics is an emerging discipline and has applications in frontline areas of biotechnology like proteomics, genomics analysis, drug design, gene therapy, diagnostics, crop improvement, biochemical process etc. Bioinformatics has become a frontiline applied science and is of vital importance to study new biology which is widely recognized as the new scientific endeavour of the twenty first century. The growth in full genomic sequencing, structural genomics, proteomics, micro-array etc. will be very slow without application of bioinformatics.

Bioinformatics and computational biology thus involve the use of techniques including applied mathematics, informatics, statistics, computer science, artificial intelligence, chemistry, and biochemistry to solve biological problems usually on the molecular level. Research in computational biology often overlaps with systems

biology. Major research efforts in the field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, and the modeling of evolution.



The terms bioinformatics and computational biology are often used interchangeably. However bioinformatics more properly refers to the creation and advancement of algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data. Computational biology, on the other hand, refers to hypothesis-driven investigation of a specific biological problem using computers, carried out with experimental or simulated data, with the primary goal of discovery and the advancement of biological knowledge. Put more simply, bioinformatics is concerned with the information while computational biology is concerned with the hypotheses.

10.2 Avenues

Sequence analysis—Sequence alignment and Sequence database

Since the Phage ϕ-X174 was sequenced in 1977, the DNA sequences of hundreds of organisms have been decoded and stored in databases. The information is analyzed

to determine genes that encode polypeptides, as well as regulatory sequences. A comparison of genes within a species or between different species can show similarities between protein functions, or relations between species (the use of molecular systematics to construct phylogenetic trees). With the growing amount of data, it long ago became impractical to analyze DNA sequences manually. Today, computer programs are used to search the genome of thousands of organisms, containing billions of nucleotides. These programs would compensate for mutations (exchanged, deleted or inserted bases) in the DNA sequence, in order to identify sequences that are related, but not identical. A variant of this sequence alignment is used in the sequencing process itself. The so-called shotgun sequencing technique (which was used, for example, by The Institute for Genomic Research to sequence the first bacterial genome, *Haemophilus influenzae*) does not give a sequential list of nucleotides, but instead the sequences of thousands of small DNA fragments (each about 600-800 nucleotides long). The ends of these fragments overlap and, when aligned in the right way, make up the complete genome. Shotgun sequencing yields sequence data quickly, but the task of assembling the fragments can be quite complicated for larger genomes. Another aspect of bioinformatics in sequence analysis is the automatic search for genes and regulatory sequences within a genome. Not all of the nucleotides within a genome are genes. Within the genome of higher organisms, large parts of the DNA do not serve any obvious purpose. This so-called junk DNA may, however, contain unrecognized functional elements. Bioinformatics helps to bridge the gap between genome and proteome projects—for example, in the use of DNA sequences for protein identification.

Genome annotation

In the context of genomics, annotation is the process of marking the genes and other biological features in a DNA sequence. The first genome annotation software system was designed in 1995 by Dr. Owen White, who was part of the team that sequenced and analyzed the first genome of a free-living organism to be decoded, the bacterium *Haemophilus influenzae*. Dr. White built a software system to find the genes (places in the DNA sequence that encode a protein), the transfer RNA, and other features, and to make initial assignments of function to those genes. Most current genome annotation systems work similarly, but the programs available for analysis of genomic DNA are constantly changing and improving.

10.3 Computational evolutionary biology

Evolutionary biology is the study of the origin and descent of species, as well as their change over time. Informatics has assisted evolutionary biologists in several key ways; it has enabled researchers to :

- ★ trace the evolution of a large number of organisms by measuring changes in their DNA, rather than through physical taxonomy or physiological observations alone,
- ★ more recently, compare entire genomes, which permits the study of more complex evolutionary events, such as gene duplication, lateral gene transfer, and the prediction of factors important in bacterial speciation,
- ★ build complex computational models of populations to predict the outcome of the system over time
- ★ track and share information on an increasingly large number of species and organisms

The area of research within computer science that uses genetic algorithms is sometimes confused with computational evolutionary biology, but the two areas are unrelated.

10.4 Measuring biodiversity

Biodiversity of an ecosystem might be defined as the total genomic complement of a particular environment, from all of the species present, whether it is a biofilm in an abandoned mine, a drop of sea water, a scoop of soil, or the entire biosphere of the planet Earth. Databases are used to collect the species names, descriptions, distributions, genetic information, status and size of populations, habitat needs, and how each organism interacts with other species. Specialized software programs are used to find, visualize, and analyze the information, and most importantly, communicate it to other people. Computer simulations model such things as population dynamics, or calculate the cumulative genetic health of a breeding pool (in agriculture) or endangered population (in conservation). One very exciting potential of this field is that entire DNA sequences, or genomes of endangered species can be preserved, allowing the results of Nature's genetic experiment to be remembered *in silico*, and possibly reused in the future, even if that species is eventually lost.

10.5 Analysis of gene expression

The expression of many genes can be determined by measuring mRNA levels with multiple techniques including microarrays, expressed cDNA sequence tag (EST) sequencing, serial analysis of gene expression (SAGE) tag sequencing, massively parallel signature sequencing (MPSS), or various applications of multiplexed in-situ hybridization. All of these techniques are extremely noise-prone and/or subject to bias in the biological measurement, and a major research area in computational biology involves developing statistical tools to separate signal from noise in high-throughput gene expression studies. Such studies are often used to determine the genes implicated in a disorder: one might compare microarray data from cancerous epithelial cells to data from non-cancerous cells to determine the transcripts that are up-regulated and down-regulated in a particular population of cancer cells.

10.6 Analysis of regulation

Regulation is the complex orchestration of events starting with an extracellular signal such as a hormone and leading to an increase or decrease in the activity of one or more proteins. Bioinformatics techniques have been applied to explore various steps in this process. For example, promoter analysis involves the identification and study of sequence motifs in the DNA surrounding the coding region of a gene. These motifs influence the extent to which that region is transcribed into mRNA. Expression data can be used to infer gene regulation: one might compare microarray data from a wide variety of states of an organism to form hypotheses about the genes involved in each state. In a single-cell organism, one might compare stages of the cell cycle, along with various stress conditions (heat shock, starvation, etc.). One can then apply clustering algorithms to that expression data to determine which genes are co-expressed. For example, the upstream regions (promoters) of co-expressed genes can be searched for over-represented regulatory elements.

10.7 Analysis of protein expression

Protein microarrays and high throughput (HT) mass spectrometry (MS) can provide a snapshot of the proteins present in a biological sample. Bioinformatics is very much involved in making sense of protein microarray and HT MS data; the former approach faces similar problems as with microarrays targeted at mRNA, the latter involves the problem of matching large amounts of mass data against predicted

masses from protein sequence databases, and the complicated statistical analysis of samples where multiple, but incomplete peptides from each protein are detected.

10.8 Analysis of mutations in cancer

In cancer, the genomes of affected cells are rearranged in complex or even unpredictable ways. Massive sequencing efforts are used to identify previously unknown point mutations in a variety of genes in cancer. Bioinformaticians continue to produce specialized automated systems to manage the sheer volume of sequence data produced, and they create new algorithms and software to compare the sequencing results to the growing collection of human genome sequences and germ line polymorphisms. New physical detection technology are employed, such as oligonucleotide microarrays to identify chromosomal gains and losses (called comparative genomic hybridization), and single nucleotide polymorphism arrays to detect known *point mutations*. These detection methods simultaneously measure several hundred thousand sites throughout the genome, and when used in high-throughput to measure thousands of samples, generate terabytes of data per experiment. Again the massive amounts and new types of data generate new opportunities for bioinformaticians. The data is often found to contain considerable variability, or noise, and thus Hidden Markov model and change-point analysis methods are being developed to infer real copy number changes. Another type of data that requires novel informatics development is the analysis of lesions found to be recurrent across many tumors.

10.9 Prediction of protein structure

Protein structure prediction is another important application of bioinformatics. The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it. In the vast majority of cases, this primary structure uniquely determines a structure in its native environment. (Of course, there are exceptions, such as the bovine spongiform encephalopathy– aka Mad Cow Disease - prion.) Knowledge of this structure is vital in understanding the function of the protein. For lack of better terms, structural information is usually classified as one of *secondary*, *tertiary* and *quaternary* structure. A viable general solution to such predictions remains an open problem. As of now, most efforts have been directed towards heuristics that work most of the time.One of the key ideas in bioinformatics is the notion of homology. In the genomic branch of bioinformatics,

homology is used to predict the function of a gene: if the sequence of gene *A*, whose function is known, is homologous to the sequence of gene *B*, whose function is unknown, one could infer that B may share A's function. In the structural branch of bioinformatics, homology is used to determine which parts of a protein are important in structure formation and interaction with other proteins. In a technique called homology modeling, this information is used to predict the structure of a protein once the structure of a homologous protein is known. This currently remains the only way to predict protein structures reliably. One example of this is the similar protein homology between hemoglobin in humans and the hemoglobin in legumes (leghemoglobin). Both serve the same purpose of transporting oxygen in the organism. Though both of these proteins have completely different amino acid sequences, their protein structures are virtually identical, which reflects their near identical purposes.

10.10 Comparative genomics

The core of comparative genome analysis is the establishment of the correspondence between genes (orthology analysis) or other genomic features in different organisms. It is these inter genomic maps that make it possible to trace the evolutionary processes responsible for the divergence of two genomes. A multitude of evolutionary events acting at various organizational levels shape genome evolution. At the lowest level, point mutations affect individual nucleotides. At a higher level, large chromosomal segments undergo duplication, lateral transfer, inversion, transposition, deletion and insertion. Ultimately, whole genomes are involved in processes of hybridization, polyploidization and endosymbiosis, often leading to rapid speciation. The complexity of genome evolution poses many exciting challenges to developers of mathematical models and algorithms, who have recourse to a spectra of algorithmic, statistical and mathematical techniques, ranging from exact, heuristics, fixed parameter and approximation algorithms for problems based on parsimony models to Markov Chain Monte Carlo algorithms for Bayesian analysis of problems based on probabilistic models.Many of these studies are based on the homology detection and protein families computation.

10.11 Modeling biological systems

Systems biology involves the use of computer simulations of cellular subsystems (such as the networks of metabolites and enzymes which comprise metabolism,

signal transduction pathways and gene regulatory networks) to both analyze and visualize the complex connections of these cellular processes. Artificial life or virtual evolution attempts to understand evolutionary processes via the computer simulation of simple (artificial) life forms.

10.12 High-throughput image analysis

Computational technologies are used to accelerate or fully automate the processing, quantification and analysis of large amounts of high-information-content biomedical imagery. Modern image analysis systems augment an observer's ability to make measurements from a large or complex set of images, by improving accuracy, objectivity, or speed. A fully developed analysis system may completely replace the observer. Although these systems are not unique to biomedical imagery, biomedical imaging is becoming more important for both diagnostics and research. Some examples are :

- ★ High-throughput and high-fidelity quantification and sub-cellular localization (high-content screening, cytohistopathology)
- ★ Morphometrics
- ★ Clinical image analysis and visualization
- ★ Determining the real-time air-flow patterns in breathing lungs of living animals
- ★ Quantifying occlusion size in real-time imagery from the development of and recovery during arterial injury
- ★ Making behavioral observations from extended video recordings of laboratory animals.
- ★ Infrared measurements for metabolic activity determination.

10.13 Protein-protein docking

In the last two decades, tens of thousands of protein three-dimensional structures have been determined by X-ray crystallography and Protein nuclear magnetic resonance spectroscopy (protein NMR). One central question for the biological scientist is whether it is practical to predict possible protein-protein interactions only based on these 3D shapes, without doing protein-protein interaction experiments. A variety of methods have been developed to tackle the Protein-protein docking problem, though it seems that there is still much place to work on in this field.

10.14 Advantage of bioinformatics

The publication of Atlas of protein sequences and structures By M Dayhoff *et al.* in 1965 paved the way for the rapid growth of protein databases. Protein sequence databases consist of protein sequences and information about protein stored. Some of the major universal databases are listed in table 1

Search Engine	Content	url
Swiss-Prot	Expertly curated protein sequence database	www. expasy.org/sprot
PIR-PSD	Protein information resource	www.pir.georgetown.edu
ExProt	Sequences of proteins with experimentally verified fnction	www. cmbi.kun.nl/ ExProt
NCBI	Protein database all protein sequences	www.ncbi.nlm.nih.gov/entrez
RefSeq	NCBI reference sequence database	www. ncbi.nlm. nih.gov/RefSeq
MIPS	Mammalian protein protein interaction database	www. mips.gsf.de
UniProt	Universal protein knowledge database	www. uniprot.org

10.15 Application / Case study

Venomous animal produces a myriad of important pharmacological components. The individual components or venoms (toxins) are used in ion channel and receptor studies, drug discovery and formation of insecticides. The toxin data are scattered across public databases which provide sequence and structural descriptions, but very limited functional annotation. The exponential growth of newly identified toxin data has created a need for better data management. Venom informatics is a systematic bioinformatics approach in which classified consolidated and cleaned venom data are stored in to repositories and integrated with advanced bioinformatics tools for the analysis of structure and function of toxins.

Information describing venom toxin sequences and the features, structures and function of venom toxins is scattered across multiple sources. Primary databases

such as GenBank or SWISS- PROT contain only basic sequence information. Each entry contains the toxin sequence, its name, taxonomy of the source organism, a list of features and references.

For example all snake venom phospholipase A_2 ($svPLA_2$) toxin entries in the Gene Bank and SWISS PROT databases. The structural database PDB contains 3D information of approximately 200 venom toxins (as of Nov 2002). To our knowledge only three specialized venom databases are currently available as major resources for the study of venom toxins. The databases contain entries collected from different sources cleaned organized, analyzed and classified according to their structure function relationship. The SCORPION database of nearly 300 entries of scorpion toxin sequence are annotated and classified according to their structural and functional properties. The MOLLUSK data base contains more than 450 peptides from the cone snail venoms where each entry has a unique field to facilitate comparison of conotoxin entries.

These entire search data base commonly used bioinformatics methods for analyzing venom toxin data for

- ★ Phylogentic analysis.
- ★ Multiple sequence alignments,
- ★ 3D structure analysis.
- ★ Homology modeling.
- ★ Evolutionary interpretation
- ★ Genomic interpretation

Hence application of bioinformatics is important in the field of toxinology because venom-toxins are functionally diverse but belong to a limited number of structural families, they are ideal for application of data mining techniques for discovery of previously unknown relationship among data. Thus venom informatics lessons will be useful for study of diverse types of active peptides.

Notes

মানুষের জ্ঞান ও ভাবকে বইয়ের মধ্যে সঞ্চিত করিবার যে একটা প্রচুর সুবিধা আছে, সে কথা কেহই অস্বীকার করিতে পারে না। কিন্তু সেই সুবিধার দ্বারা মনের স্বাভাবিক শক্তিকে একেবারে আচ্ছন্ন করিয়া ফেলিলে বুদ্ধিকে বাবু করিয়া তোলা হয়।

— রবীন্দ্রনাথ ঠাকুর

ভারতের একটা mission আছে, একটা গৌরবময় ভবিষ্যৎ আছে, সেই ভবিষ্যৎ ভারতের উত্তরাধিকারী আমরাই। নূতন ভারতের মুক্তির ইতিহাস আমরাই রচনা করছি এবং করব। এই বিশ্বাস আছে বলেই আমরা সব দুঃখ কষ্ট সহ্য করতে পারি, অন্ধকারময় বর্তমানকে অগ্রাহ্য করতে পারি, বাস্তবের নিষ্ঠুর সত্যগুলি আদর্শের কঠিন আঘাতে ধূলিসাৎ করতে পারি।

— সুভাষচন্দ্র বসু

Any system of education which ignores Indian conditions, requirements, history and sociology is too unscientific to commend itself to any rational support.

— Subhas Chandra Bose

Price : Rs. 225.00

(Not for Sale)

Published by Netaji Subhas Open University, DD-26, Sector-1, Salt Lake, Kolkata-700 064 & printed at Printtech, 15A, Ambika Mukherjee Road, Kolkata-700 056

NETAJI SUBHAS OPEN UNIVERSITY

STUDY MATERIAL

POST GRADUATE
ZOOLOGY

Paper - 6

Group : B

Immunology and
Microbiology

PREFACE

In the curricular structure introduced by this University for students of Post-Graduate Degree Programme, the opportunity to pursue Post-Graduate course in any subject introduced by this University is equally available to all learners. Instead of being guided by any presumption about ability level, it would perhaps stand to reason if receptivity of a learner is judged in the course of the learning process. That would be entirely in keeping with the objectives of open education which does not believe in artificial differentiation.

Keeping this in view, study materials of the Post-Graduate level in different subjects are being prepared on the basis of a well laid-out syllabus. The course structure combines the best elements in the approved syllabi of Central and State Universities in respective subjects. It has been so designed as to be upgradable with the addition of new information as well as results of fresh thinking and analysis.

The accepted methodology of distance education has been followed in the preparation of these study materials. Co-operation in every form of experienced scholars is indispensable for a work of this kind. We, therefore, owe an enormous debt of gratitude to everyone whose tireless efforts went into the writing, editing and devising of proper lay-out of the materials. Practically speaking, their role amounts to an involvement in 'invisible teaching'. For, whoever makes use of these study materials would virtually derive the benefit of learning under their collective care without each being seen by the other.

The more a learner would seriously pursue these study materials, the easier it will be for him or her to reach out to larger horizons of a subject. Care has also been taken to make the language lucid and presentation attractive so that they may be rated as quality self-learning materials. If anything remains still obscure or difficult to follow, arrangements are there to come to terms with them through the counselling sessions regularly available at the network of study centres set up by the University.

Needless to add, a great deal of these efforts is still experimental–in fact, pioneering in certain areas. Naturally, there is every possibility of some lapse or deficiency here and there. However, these do admit of rectification and further improvement in due course. On the whole, therefore, these study materials are expected to evoke wider appreciation the more they receive serious attention of all concerned.

Professor (Dr.) Subha Sankar Sarkar
Vice-Chancellor

Fifth Reprint : January, 2020

**Printed in accordance with the regulations of the
Distance Education Bureau of the University Grants Commission.**

POST GRADUATE ZOOLOGY

[M.Sc]

PAPER : GROUP
PGZO - 6 : B

Writer & Editor
Prof. Amal Bhattacharya

Netaji Subhas Open University

PGZO–6
Quantitative Biology & Biotechnology, Immunology & Microbiology

Group - B
Immunology & Microbiology

Unit 1 ❐ Overview of Immune System, Components of Immunity, Innate and Adaptive Immunity

Structure

1.1 Overview of immune system

Immunity is the state of protection from infectious disease. The latin term 'immunis' meaning exempt is the main source of english word immunity. In proper sense this response is the defensive reactivity to a specific molecular configuration that develops following contact with it in vertebrate system. This special reactivity is the resistance to second infection or allergic response. The specific molecules released by the foreign substance are either pathogen or allergen. Immunity has both specific and non specific components. The non specific component, innate immunity is a set of disease resistant mechanism that are spontaneous and not specific to a particular pathogen. Phagocytic cells like macrophages play an important role in many aspects of innate immunity. In contrast, the specific components of adaptive immunity display high degree of specificity and the concept of memory for secondary response. The major cellular agents of adaptive immunity are lymphocytes and the antibodies and certain other molecules.

Historical concept : The belief of immunity in human population developed in course of time. Common human belief in society and scientist's observations equally contributed to the development of this discipline during early period. Jenner (1798) tried inoculation method in human body. Pasteur (1881) observed attenuation and gave vaccine idea (derived from latin word vacca meaning cow). In twentieth century scientists detected humoral immunity in blood serum, cellular immunity in body lymphocyte cell and phagocytosis by macrophage. Further biological procedures identified B and T lymphocyte, structure of antibody and major histocompatibility complex region of cell etc.

1.2 Components of immunity

There are two types of immunity e.g. innate and adaptive or acquired immunity. As because the adaptive immunity needs some time to work properly, innate immunity provides the first line of defence during critical period just at the beginning of infection at the primary level. The innate and adaptive immunity do not operate independently of each other, they function as a highly interactive and cooperative system, producing a total synchronized attack against the pathogen.

1.3 Theory of immunity

Erlich's selective theory (1900) and Horowitz's instructional theory were not considered valid. The clonal selection of Burnett (1950) only maintains the concept

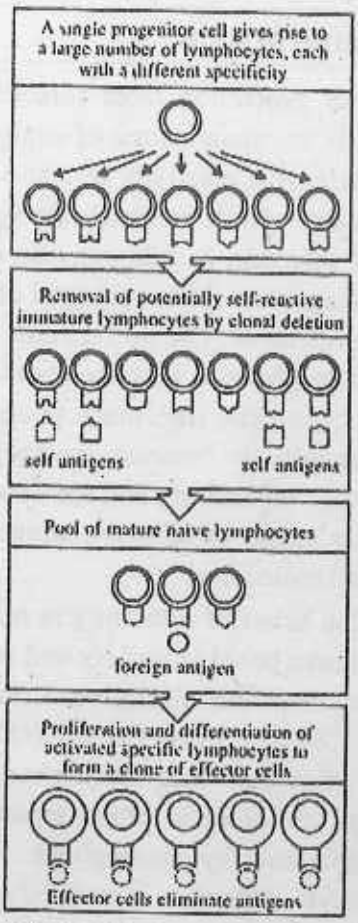


Fig. 1.1 Clonal selection theory

of acquired immunity. According to this theory an individual lymphocyte expresses unique membrance receptor which is specific for a dintinct antigen. This unique receptor of antigen binds to antibody or T lymphocytes and activates the lymphocyte cell to proliferate in a clone.

1.4 Features of innate and specific (adaptive) immunity

The following table will show the features of innate and specific (adaptive) immunity.

	Features	Innate	Specific
Characteristics	(1) Specificity for microbe	Relatively low	High
	(2) Diversity	Limited	Large
	(3) Specialization	Relatively stereotype	Highly specialized
	(4) Memory	No	Yes
Components	(1) Physical and chemical barriers	Skin mucosal epithalia, antimicrobal chemicals (e.g. defensins)	Cutaneous and mucosal immune sepheres; secreted antibodies
	(2) Blood proteins	Complement	Antibodies
	(3) Cells	Phagocytes (macrophages, neutrophils) natural killer cells	Lymphocytes

1.4.1 Innate (non-specific) immunity

Type	Mechanism
Anatomic barriers	
Skin	Mechanical barries retards entry of microbes. Acidic environment (pH 3-5) retards growth of microbes.
Mucous membranes	Normal flora compete with microbes for attachment sites and nutrients. Mucous entraps foreign microorganisms. Cilia propels microorgnisms out of the body.
Physiologic barriers	
Temperature	Normal body temperature inhibits growth of some pathogens. Fever response inhibits growth of some pathogens.

Type	Mechanism
Low pH	Acidity of stomach contents kills most of ingested microorganisms.
Chemical mediators	Lysozyme cleaves bacterial cell wall. Interferon induces antiviral state in uninfected cells. Complement lyses microorganisms or faciliates phagocytosis.
Phagocytic/endocytic barriers	Various cells internalize (endocytose) and breakdown foreign macromolecules. Specialised cells (blood monocytes, neutrophils, tissue macrophages) internalize (phagocytose), kill and digest whole microorganisms.
Inflammtory barriers	Tissue damage and infection induce leakage of vascular fluid, containing serum proteins with antimicrobial activity and influx of phagocytic cells into the affected area.

1.4.2 Adaptive Immunity : Adaptive immunity is capable of recognizing and selectively eliminating specific foreign microorganisms and molecules (antigens). Unlike innate immune responses, adaptive immune response reactions are specific antigenic challenges and display four attributes:

(1) Antigenic specificity
(2) Diversity
(3) Immunologic memory
(4) Self / non-self recognition

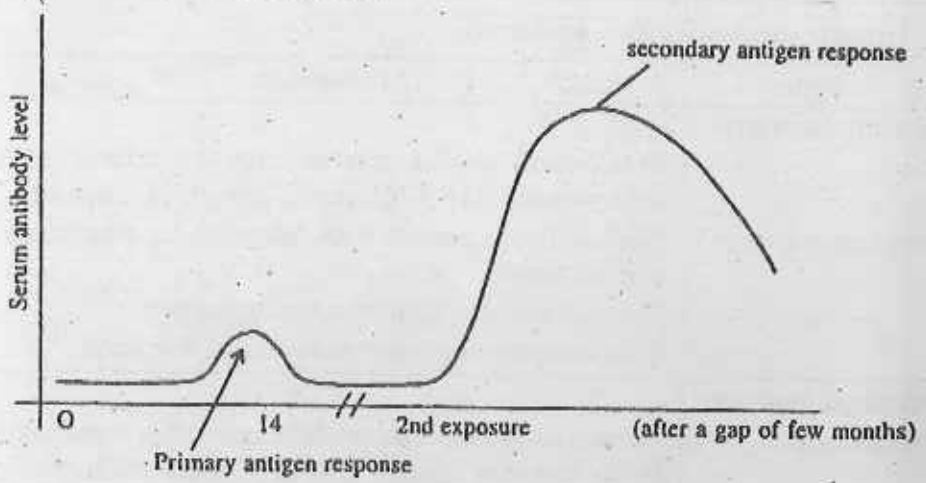


Fig. 1.2. Primary and secondary response serum level

An effective immune response involves two major groups of cells e.g. lymphocytes and antigen presenting cells (APC). The initial interaction of a naive lymphocyte with an antigen generates a primary response and the response is kept in memory so that in second contact of the host with the same antigen will induce a very rapid and hightened secondary response and the peak occurs in less time.

I. B Lymphocytes :

(1) They mature within bone marrow, however in bird it is processed by bursa fabricius. Bone marrow stem cell is the precursor of all blood cell. The mature B lymphocyte expresses a **unique antigen binding receptor** on its membrane.

(2) The B cell receptor is a membrane bound antibody molecule.

(3) On binding of antigen to a naive B cell, the cell divides rapidly; its progeny differentiate into **memory B cells and effector B Cells called plasma cells.**

(4) Memory B-cells continue to express same BCR as their parent naive B cell.

(5) Plasma cells produce antibody in a form that can be secreted.

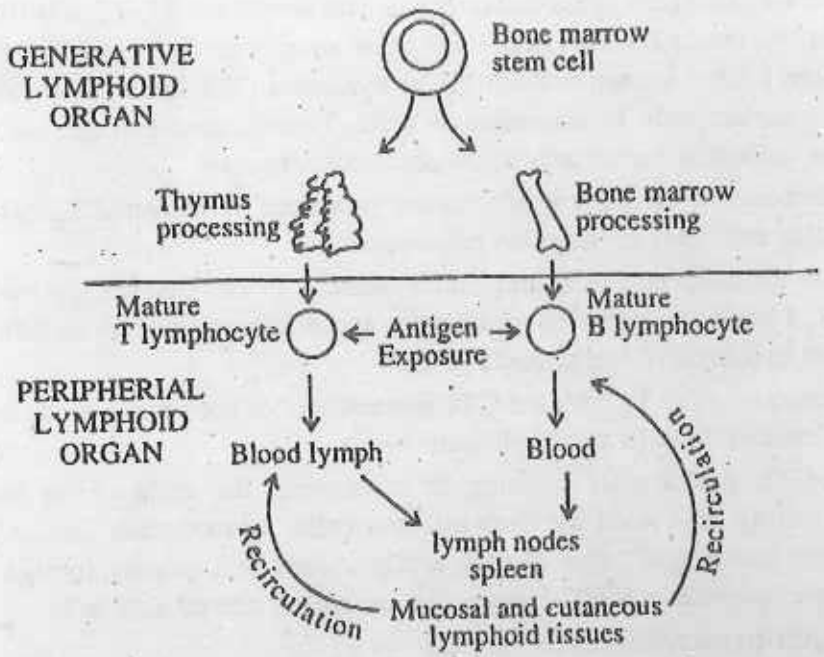


Fig. 1.3. ***Maturation of lymphocytes*** : Mature lymphocytes develop from bone marrow stem cell in generative (primary) lymphoid organ and immune response to foreign antigen develops in the peripheral lymphoid tissue.

II. T lymphosytes :

(1) T lymphocytes arise in bonemarrow but migrate to thymus for maturing.

(2) T cell expresses a unique antigen binding receptor (TCR) on its membrane. They can recognise only antigen that is bound to cell membrane proteins called major histocompatibility complex (MHC).

(3) There are two well defined subpopulations of T cells: T helper (T_H) and T-cytotoxic (Tc) cells. Although a third type of T cell called T supressor (Ts) has also been postulated.

(4) T helper and T cytotoxic cells can be distinguished from one another by the presence of either CD4 or CD8 membrane glycoproteins on their surface. CD stands for cluster of differentiation and refers to a molecule recognised by a cluster of monoclonal antibodies that can be used to identify the lineage or stage of differentiation of lymphocyte and thus to distinguish one class of lymphocytes from another.

(5) T cells displaying CD4 generally function as T_H cells, whereas those displaying CD8 function as Tc cells. Both of them perform intracellular killing mechanism of pathogen.

(6) After a T_H cell recognises and interacts with an antigen-MHC class II molecule complex, the cell is activated–it becomes an effector cell that secretes various growth factors known collectively as cytokines. The secreted cytokines play an important role in activating B cells, Tc cell, macrophages and various other cells that participate in the immune response.

(7) Differences in the pattern of cytokine produced by activated T_H cells result in different types of immune responses.

(8) Under the influence of T_H-derived cytokine, a Tc cell that recognizes antigen-MHC I molecule complex proliferates and differentiates into an effector cell celled cytotoxic T lymphocyte (CTL).

(9) In contrast to the T_H cell, the CTL generally does not secrete many cytokines and instead exhibits cytotoxic activity.

(10) The CTL has a vital function in monitoring the cells of the body and eliminating any, such as virus-infected cells, tumour cells and cells of a foreign tissue graft, that display antigen, cells that display foreign antigen complexed with a MHC I molecule are called altered self cells.

1.4.3 Antigen-presenting cells (APC)

(1) Certain specialized cells, which includes macrophage, B lymphocytes and dendritic cells are distinguished by 2 properties : i) they express class II

MHC molecules on their membranes and ii) they are able to deliver a co-stimulatory signal that is necessary for T_H cell activation.

(2) APCs first internalize antigen either by phagocytosis or by endocytosis and then display a part of that antigen bound to a class II MHC molecule on their membrane.

1.4.4 Interaction of lymphyocytes and APC

The APC (mainly macrophage) takes up the antigen and present it to T or B cell. T cells can act independant of B Cell. B cells require co-stimulation from T cell. T cells are activated on stimulation by APC and antigen attached to MHC. Complement activation process of the serum protein also form an effector arm of the humoral response.

1.5 Selective questions

(1) Define immunity. Explain the differences between innate and acquired immunity.

(2) Comment on primary and secondary immune response.

(3) Mention the maturation process of lymphocytes.

(4) State the role of antigen presenting cell (APC), B lymphocyte and T lymphocyte.

1.6 Selected readings

(1) Coico, R., Sunshine, G, and Benjamini, E. 2003. **Immunology** 5th Edn. Wiley-Liss, New Jersy.

(2) Janeway, C, A., Travers, P., Walpart, M. and Capra, J.D. 1999. **Immuno Biology**, 4th edn. Current Biology Publications, New York.

(3) Roitt Ivan 1994. **Essential Immunology** 8th edn. Blackwell Scientific Publications, London.

Unit 2 ❐ Cells and Organs of the Immune System

Structure

2.1 Introduction

Our immune system in body is controlled by white blood cells. The lymphocyte performs the role of acquired immune function and possess the major attributes of acquired immunological principles.

Embryonic Stem Cell is the precursor of all different types of cells that form the basis of all organ formations. Haematopoietic Stem Cells (HSC) generate blood cells (WBC and RBC). The process begins in embryonic sac during 1st week of development. In the 3rd month of pregnancy the HSC migrate from the yolksac to the foetal liver and then to the spleen. Later on after 6 months bone marrow is the precursor for differentiation (Fig. 2.1).

HSC give rise to i) lymphoid progenitor cell–B and T lymphocytes, natural killer cell. ii) Myeloid proganitor cell–RBC and WBC cells. The WBC are neutrophil, eosinophil, basophil, monocyte, mast cell and platelets (Fig. 2.2) The lymphon is a collective term for the primary and secondary lymphoid organs and their interconnecting blood vassels and lymphatics (Fig. 2.3). The blood cells enter circulation after differentiating in bone marrow. Various growth factors are involved during differentiation. These are : i) A group of acidic glycoprotein for colony stimulating factor (CSF) which are all cytokine material.The CSF's are Multilineage CSF, Interleukin-3, Macrophage colony stimulating factor (M-CSF), granulocyte CSF (G-CSF); granulocyte-macrophage CSF (GM-CSF). Another important haematopoietic cytokine detected by the method is a glycoprotein called erythropoietin (EPO). The expression of a progenitor cell to a particular differentiation pathway is controlled by a specific cytokine. It has been shown that regulation of haematopoiesis is controlled by certain gene like GATA-2.The steadystate regulation of haematopoiesis is controlled by factors like i) control of the level and types of cytokine produced by bonemarrow stromal cells, ii) release of cytokins by activated T cells and macrophage cells iii) regulation of cytokine production by stem cells

and progenitor cells iv) Apoptosis process of regulated cell death.

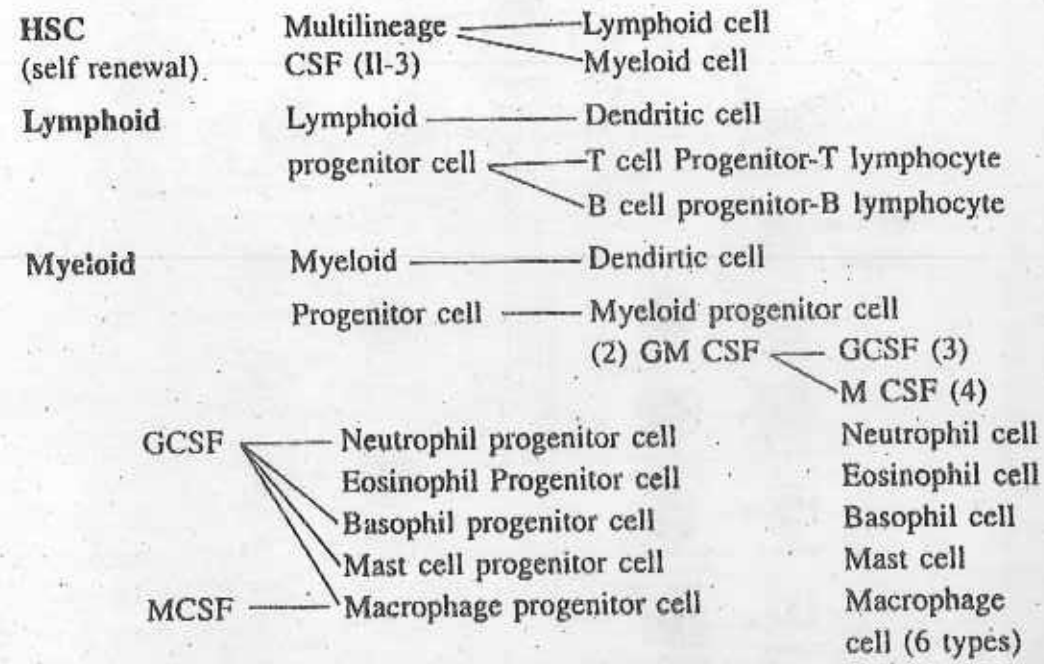


Chart : showing the process of immune cell formation

Abbreviations :

HSC	-	Haematopoietic stem cell
CSF	-	Colony stimulating factor
GMCSF	-	Granulocyte monocyte colony stimulating factor
GCSF	-	Granulocyte colony stimulating factor
M CSF	-	Monocyte colony stimulating factor
1-4	-	`Nos mark the colony stimulating factors

2.2 Cells of the immune system

Lymphoid cell : The lymphocytes constitute 20-40% of the WBC population. They are B and T lymphocytes and Natural Killer Cell. Natural Killer (Null) Cells have no surface markers like B and T lymphocytes. B and T lymphocytes (Naive stage) are 8 to 10 μm in diameter with large nucleus having dense heterochromatin. The cells have thin layer of cytoplasm with mitochondrion, ribosome and lysosome. B lymphocyte's receptor is a membrane bound antibody molecule. It is processed by bonemarrow or bursa in bird. In presence of antigen the naive B lymphocyte differentiates into memory B cell, effector B cell (Plasma cell). The mature B cell, has some markers on its surface like i) B 220 (a form of CD 45) ii) CRI (CD 35)

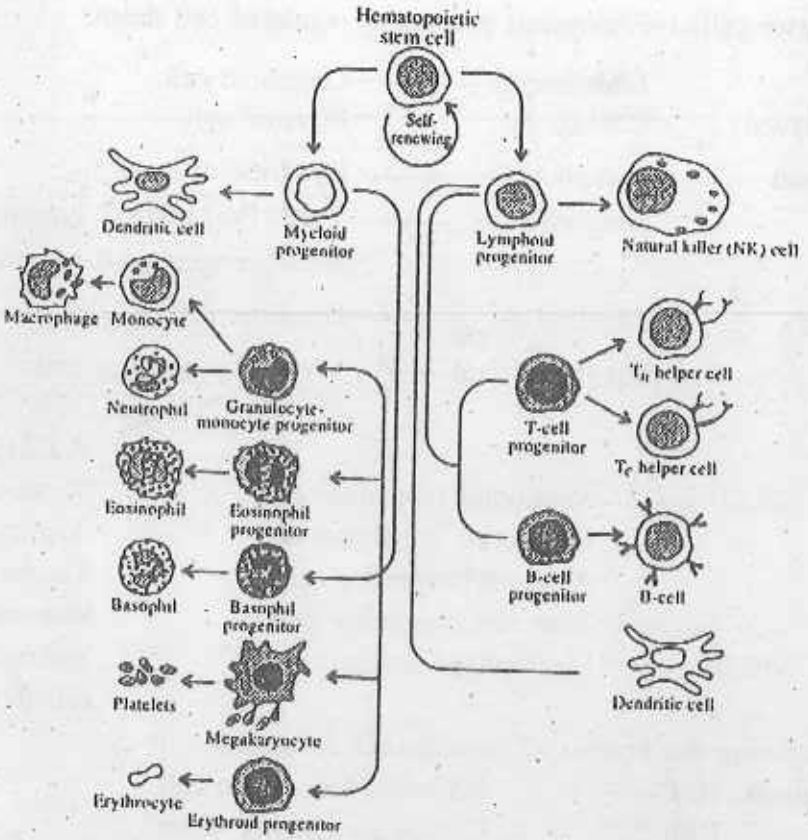


Fig. 2.1 Haematopoiesis—an outline of the formation of myeloid and lymphoid cells

and CR II (CD 21) receptors for complement factors iii) FcR II (CD 32) receptor for IgG (during ADCC) iv) B7-1 (CD 80) and B7-2 (CD 86) for interacting with CD 28 and CTLA-4 on the surface of T cell v) CD 40 for interaction with T helper cell.

B cells are activated by T helper cytokine factors and are differentiated to plasma cell and memory cell line. All clonal progeny from a given B cell secrete antibody molecule with same specificity. Plasma cells are teminally differentiated cell.

T lymphocytes are processed in thymus. they have membrane receptor for antign like B cell but they do not recognise a free antigen. It recognises antigen only when it is bound to a self molecule encoded by genes. So the antigen must be displayed together with MHC molecule on the surface of antigen presenting cell, virus infected cell, cancer cell and grafts. T lymphocytes are of two types mainly as T helper and T cytotoxic cell with receptor molecule CD4 and CD8 respectively.

Null cells are lymphocytes in peripheral blood which do not express the membrane molecule and receptors. They have no specificity like B and T lymphocyte and no memory cells. Most members of the population are large, granular lymphocyte and are natural killer (NK) cells. They constitute about 5% to 10% of the lymphocytes in human peripheral blood. They act against tumour cell and virally infected cell by releasing IFN-γ. NK cell kills virally infected cell along with T cytotoxic cell in a process called antibody-dependent cell mediated cytotoxicity where it is attached to antibody molecule.

Myeloid GM Cell : Monocyte (or macrophage in tissue)–It is a mononuclear cell in blood circulation and in tissue it becomes enlarged (5 to 10 times) and called macrophages, with phagocytic and killing mechanism. A number of antimicrobial and cytotoxic substances are produced by activated macrophage and can destroy phagocytosed microganisms. In an oxygen dependent killing mechanism activated phagocytes produce a number of reactive oxygen intermediate and reactive nitrogen intermediates that have potent microbial activity. Activated macrophage release lysozyme, Tumour necrosis factor and various hydrolytic enzymes which are oxygen independent killing mechanism. Macrophages in tissues are called alveolar cell (in lung), Histiocytes (in connective tissue). Kupffer cells (in liver), Mesengial cell (in kidney), Microglial cell (in Brain) and Osteoclasts (in bone).

Granulocyte cells are neutrophils, eosinophil, basophil and mast cells.

Neutrophil : Neutrophils are with multilobed nucleus and granulated cytoplasm. It stains both in acid and basic dyes. They are chemotactic and phagocytic. Movement of circulating neutrophils into tissue is called extravasation. Few chemotactic factors are accumulted at infiltration site during movement of neutrophil. These chemotactic factors are complement factors, blood clotting chemicals, cytokines released by activated T helper and macrophage cells. Neutrophils are phagocytic. The lytic enzymes and antimicrobial substances are contained in primary and secondary granules. Like macrophase they have also oxygendependent and independent pathways to generate antimicrobial substances.

Eosinophils : Eosinophils are with bilobed nucleus and stain in acid dye eosin. They act against parasite infection. They are also phagocytic in nature sometimes.

Basophils : Basophils have lobed nucleus and bind basic dyes. They express high affinity for $F_C\varepsilon$ receptor and can be triggered by antigen-binding to IgE and thereby mediate immediate hypersensitivity reaction to antigen. This character is also shared by mast cell. Basophils enter tissue only when they are recruited into inflammatory site. They have structural and functional similarities to mast cells.

Mast cells : They are like basophil and act against allergy. They are found

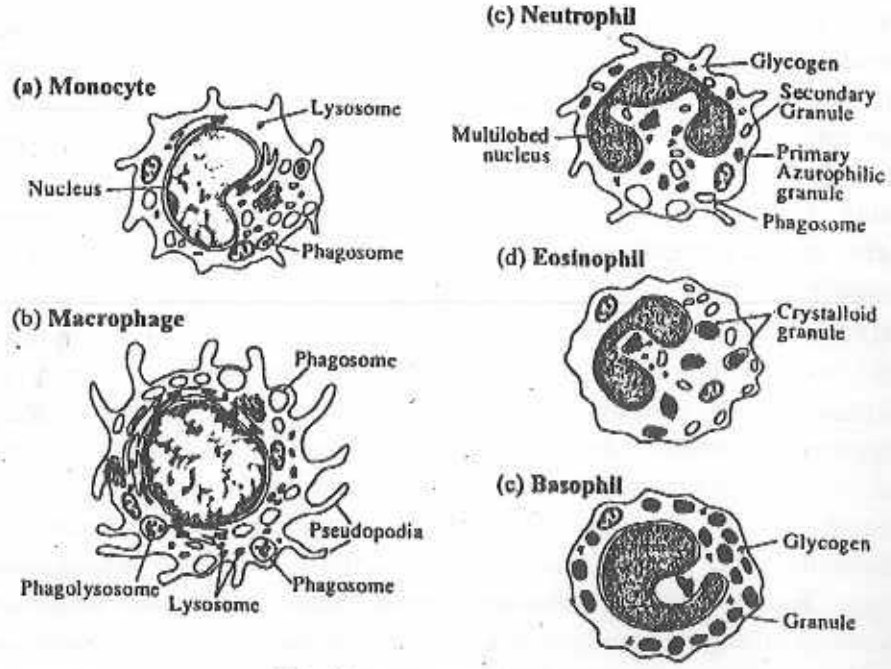


Fig. 2.2. Diagram of granulocyte cells

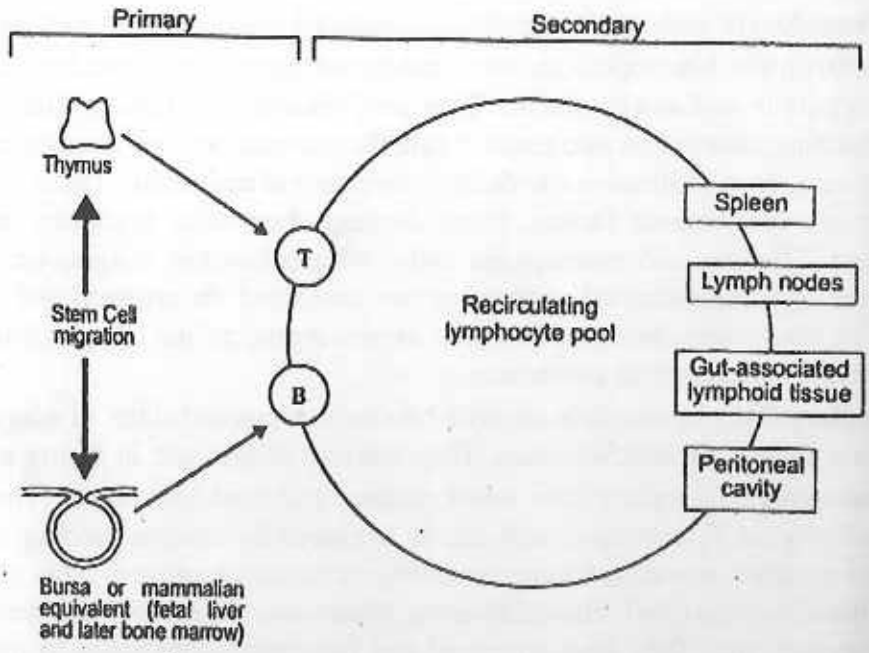


Fig. 2.3. The Lymphoid system and its relation with organs involved

throughout the body, predominantly located near blood vessels, nerves, lymphoid organs and beneath epithelium. They may be round, oval or spindle like with round nuclear. The cytoplasm contains membrane bound granules and lipid bodies. Mast cells found in the mucosa of gastrointestinal tract have chondroitin sulphate as their major granule proteoglycan where histamine production is minor. Whereas mast cells found in lungs and serosa of body tissue contain heparin as their major granule proteoglycan and histamine is produced in large quantity.

Dendritic cell process and present antigen to T_H cell. They are classified by their location as Langerhans cell in epidermis (skin) and mucous membrane, interstitial dendritic cell in heart, lung, liver, kidney, GI tract; intendigitating dendritic cell present in T cell areas of secondary lymphoid tissue or thymic medulla and follicular dendritic cells in cell rich region.

2.3 Organs of the immune system

Mature lymphocytes develop from bonemarrow stem cell in the generative (primary) lymphoid organ like Bone marrow, Bursa of fabricius and Thymus. Immune responses to foreign antigens occur in the peripheral lymphoid organs like lymph nodes, spleen, mucosal and cutaneous lymphoid tissues (secondary and tertiary lymphoid organs). In both T and B lymphocytes development, a selection process eliminates immature lymphocytes that react with self antigen, in addition, thymocytes that do not recognise self MHC molecules are also eliminated.

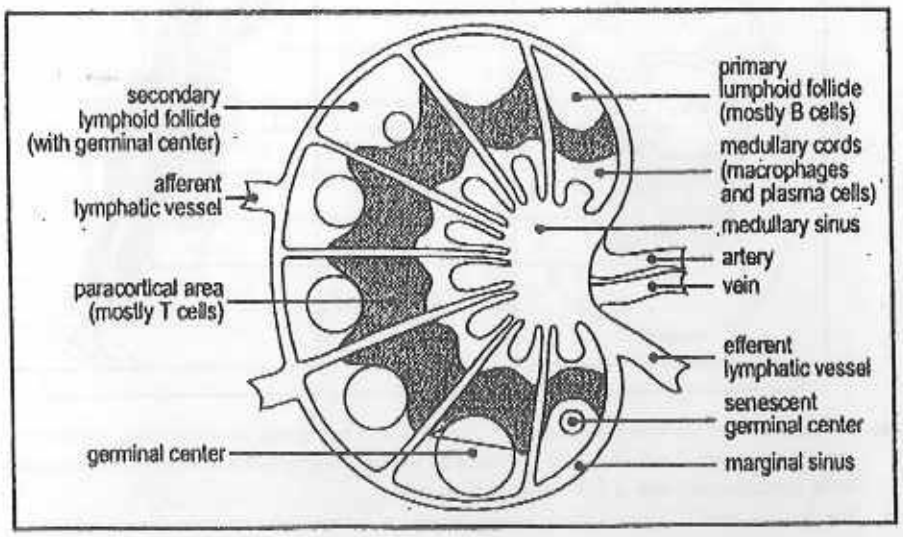


Fig. 2.4a. Organization of lymphnode. The node consists of an outermost cortex and an inner medulla.

In secondary lymphoid organs lymphocytes interact with antigen and undergo clonal proliferation and differentiation into effector cells. Lymph nodes trap antigen from regional tissue space and spleen collects antigens from blood. Lymph node and spleen are encapsulated organ (Fig. 2.4a, b). Less organized lymphoid tissues are in

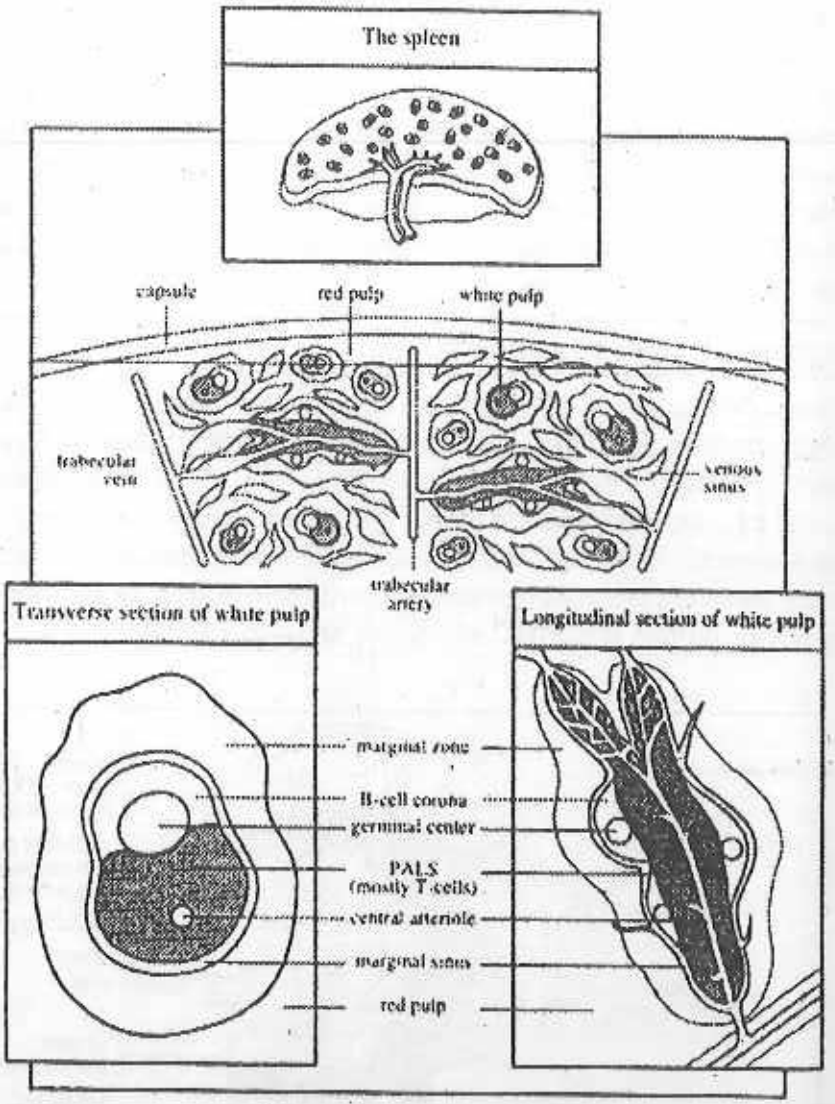


Fig. 2.4b Organization of spleen (top diagram) The spleen is composed of white pulp, rich in lymphoid cells, and red pulp, which contains many sinuses as well as large quantities of erythrocytes and macrophage, some lymphocytes and a few other cells types.
The bottom two diagrams show enlargements of a transverse section (lower left) and longitudinal section (lower right) of white pulp.

mucous membrane and are encapsulated. These include loose clusters of lymphoid follicles in the intestinal lamina propria and Peyer's patches found on the walls of the intestine. Cutaneous associated lymphoid tissue constitute the tertiary lymphoid organ (Fig. 2.5a, b and 2.6)

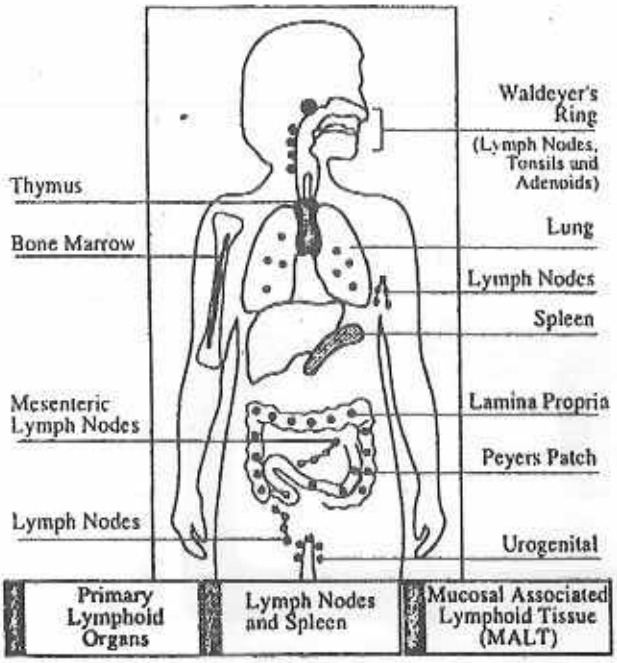


Fig. 2.5a The distribution of major lymphoid organs and tissues throughout the body.

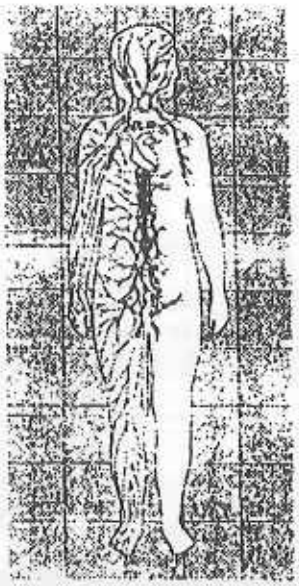


Fig. 2.5b The network of lymph nodes and lymphatics lymph nodes occur at junctions of the draining lymphatics. The lymph finally collects in the thoracic duct and hence returns to the blood stream via the left subclavian vein

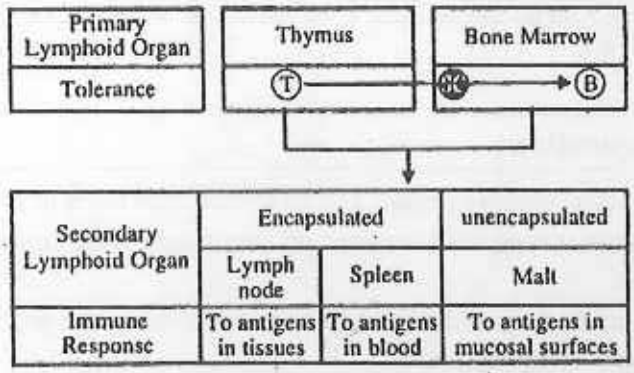


Fig 2.6. The functional organization of lymphoid tissue. T–Thymus; B–Bone marrow; SC–Stem cell; MALT–Mucous associated lymphoid tissue

Communication between these tissues and the rest of the body is maintained by a pool of recirculating lymphocytes which pass from the blood into the lymph node, spleen and other tissue and back to the blood by major lymphiatic channel as thoracic duct. (Fig. 2.7)

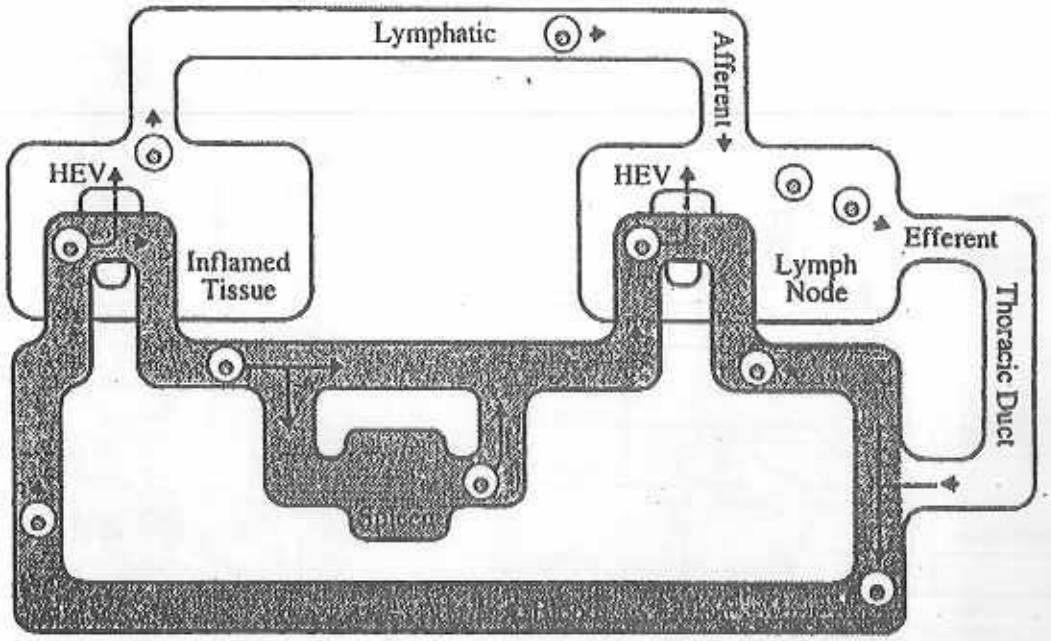


Fig. 2.7 The traffic of lymphocytes through out the body.
(HEV–Highwalled endothelium of the postcapillary venule)
Lymphocytes in blood circulation enter tissue and lymph nodes passing through HEV and leave via the draining lymphatics, the efficient lymphatics finally emerge from the node and enter thoracic duct which returns the lymphocytes to the blood stream. In the spleen lymphocytes enter the lymphoid area (while pulp) from the arterioles pass to the sinusoids of the erythroid area (red pulp) and leave by the splenic vein.

2.4 Selective questions

(1) List the primary lymphoid organs and state their functions in immune response.

(2) List the secondary lymphoid organs and summarize their functions in immune response.

(3) Give an outline diagram of Haematopoiesis and explain the process of formation of myeloid and lymphoid cells.

(4) Draw and describe different granulocyte cells.

2.5 Selected readings

(1) Janeway,C.A., Travers, P., Walport, M. and Capra J.D 1999. **Immuno Biology.** 4th edn. Current Biology Publications, New York

(2) Kuby J. 1999 Immunology 4th edn. Elsevier Publication, New York

(3) Roitt Ivan 1994. Essential Immunology 8th edn. Blackwell Scientific Publication, London.

Unit 3 ❐ Antigenicity and Immunogenicity, Immunogen properties, Adjuvant, Epitope, Hapten

Structure

3.1 Antigenicity and immunogenicity

Substances capable of evoking a specific immune response is generally called antigen. However we have to consider many points to determine if an antigen is also immunogen. **Immunogenicity** is the ability of an antigen to induce a humoral or cellular response. Whereas **Antigenicity** is the ability of the antigen simply to interact specifically with free antibody or with antigen-binding receptors on lymphocytes but fail to induce immune response B and T lymphocytes recognize small sites called antigenic determinats or epitopes on a complex immunogen. So this difference is mainly at functional level. The binding of antigen with B and T lymphocyte is highly specific; the immune components are capable of recognizing various physicochemical aspects of the component. There are several laws of physical chemistry controlling the phenamenon. These are Vender Waals forces, electrostatic interactions, hydrophobic interactions and hydrogen bonds function.

3.2 Immunogen properties

Factors that influence immunogenicity are mainly foreignness, molecular size, chemical composition of immunogen.

Foreignness : The immunogen must be a non-self molecule. The common experimental antigen bovine serum albumin (BSA) is not immunogenic when injected in cow but is strongly immunogenic when injected into rabbit. As a rule the more foreign the substance it becomes more immunogenic. In certain exceptional cases, as for example the body develops immune response against corneal tissue and sperm which are isolated from the immune system.

Molecular weight : The immunogen must have minimum molecular weight: small compounds with a molecular weight below 1000 Da (e.g. penicillin, progesterone, aspirin) are not immunogenic. Normally the chemicals of molecular weight above 6,000 Da are immunogenic (e.g. albumin, tetanus toxid).

Chemical complexity : Co-polymers of several amino acids are highly immunogenic. Protein structure of primary, secondary, tertiary and quaternary nature also have different nature during immunogenicity.

Degradability : The enzymatic degradation of protein antigen by antigen presenting cell is an important factor during binding with T cell. Carbohydrates are not processed and they cannot activate T cell although they can activate B cells.

3.3 Further requirements for immunogenicity

The genetic make up (genotype) of the organism play an important role in determining whether a given molecule will stimulate the immune response. So also in case of many parasitic infections the genotype of particular mice strain show either resistance or susceptibility during infection period.

The dosage and route of administration of antigen play a significant role in immunogenicity. Antigen administered via subcutaneously generally becomes very effective.

3.3.1 Adjuvant : Adjuvants (from latin adjuvare, to help) are substances that when mixed with an immunogen enhance the immunogenicity. They prolong antigen persistence, enhance co-stimulatory signal, induce granuloma formation, stimulate lymphocyte proliferation non specifically. Adjuvants contain microbial components (e.g. mycobacterial extracts). Other standard adjuvants are aluminium hydroxide and aluminium phosphate.

Table 3.1 Common adjuvants and their mode of action

Adjuvant	Composition	Mechanism of action
1) Alluminium hydroxide or aluminium phosphate (alum)	Aluminium hydroxide	Increases antigen uptake by APC and delayed release of antigen
2) Alum with a mycobacterial derived dipeptide	Aluminium Hydroxide gel with muramyl dipeptide	Enhanced antigen acceptance by APC, delayed release of antigen induction; Co-stimulation function of APC induced.

Adjuvant	Composition	Mechanism of action
3) Alum with *Bordetella pertusis*	Aluminium hydroxide gel with killed B pertusis	Increased uptake of antigen, delayed release of antigen and also costimulation function of APC induced.
4) Freund's complete adjuvant	oil in water with killed mycobacteria	Increased uptake of antigen by APC, delayed release of antigen. Co-stimulation function of APC induced.
5) Freund's incomplete adjuvant	Oil in water	Increased uptake of antigen by APC, delayed release of antigen.
6) Immune stimulatory complex	Open cage like structures with cholesterol and saponin.	Delivery of antigen to cytosol to induce cytotoxic T cell response

3.3.2 Epitopes

Lymphocytes recognize distinct sites on the macromolecule of antigen called epitopes or antigenic determinants. Epitopes are the active region of an immunogen that bind to lymphocytes. (Fig. 3.1).

Epitopes recognized by B and T cells

Protein antigens usually contain both sequential or non-sequential aminoacids.

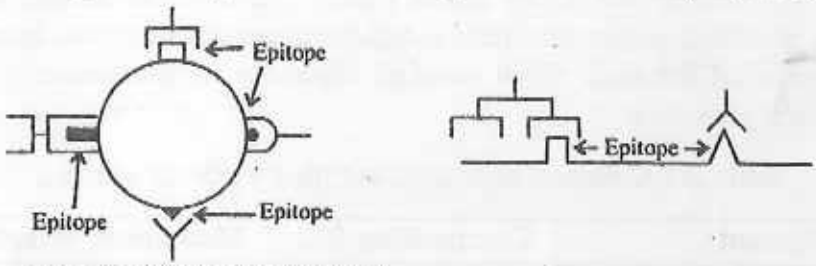


Fig. 3.1a. Different type of epitopes

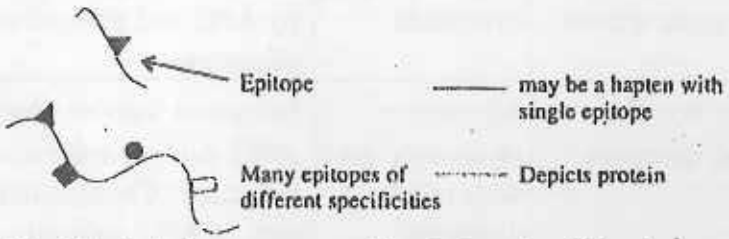


Fig. 3.1b. Antigenic structures containing single and multiple epitopes

Epitopes may be composed of sequential contiguous residues arranged in a sequence along the polypeptide chains or non-sequential residues from segments of the chain brought together by the folded confirmation of an antigen (Fig. 3.3). Normally the molecules at bends of α-helical regions are sequential (ex-sperm whale myoglobin) (Fig. 3.2). Spermwhale myoglobin also contains several non-sequential epitopes (or conformational types). Due to folding in tertiary protein structure this binding sites come close.

Table 3.2. Comparison of antigen recognition by T cells and B cells

Characteristics	B cells	T cells
1) Interaction with antigen	binary complex of immunoglobulin and antigen	Ternary complex T cell receptor, antigen and MHC molecule
2) Binding with soluble antigen	Yes	No
3) MHC Role	No	Yes
4) Chemical nature	Protein, Polysaccharide and Lipid	mostly protein
5) Property	Sequential or non sequential amino acid	Linear peptide by process-ed antigen and MHC molecule

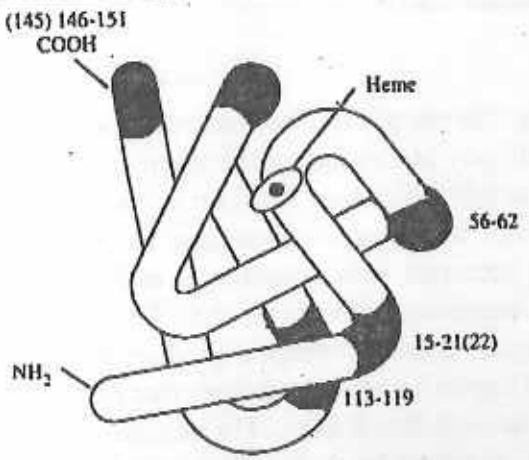


Fig. 3.2. Showing sperm whale myoglobin containing 5 linear B cell epitopes (black colour)

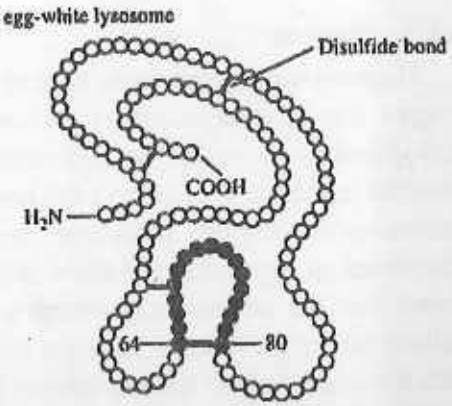


Fig. 3.3. Showing amino acid residues (circles) which forms a non-sequential epitope 'loop' (black colour) resulting from a disulfide bond between residues 64 and 80.

T cell epitope

Antigenic peptides recognized by T cells form trimolecular complexes with a T cell receptor and MHC molecule (Fig. 3.4).

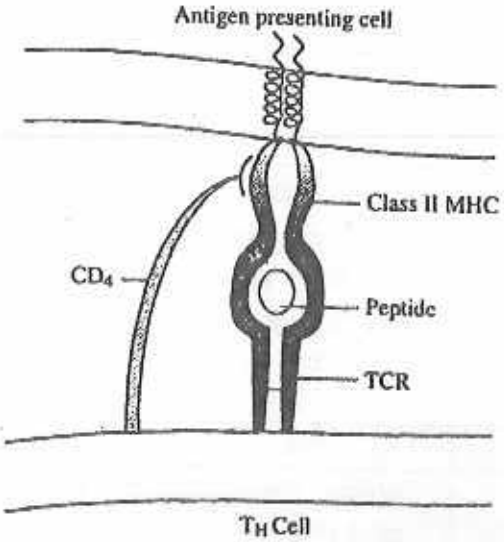
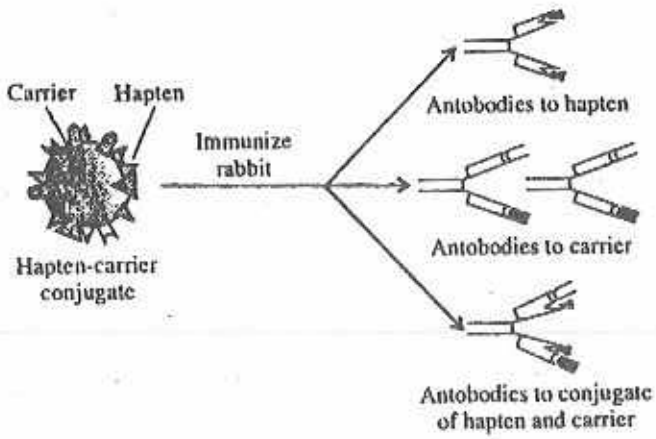


Fig. 3.4 Diagram of the ternary complex formed between a T cell receptor on a T_H cell, an antigen and a class II MHC mol.

Antigens that are recognized by T cells have two distinct interaction sites; an agrotope which interacts with a MHC molecule and an epitope which interacts with the T cell receptor

3.3.3 Hapten

Hapten (from greek word hapten meaning "no grasp") is a low molecular weight antigen that is recognised by preformed antibody but is not immunogenic unless conjugated to a 'carrier' or high molecular weight substance. The carrier molecule provides epitopes recognized by helper T cell. If carrier is conjugated to a non-immunogenic hapten molecule the latter becomes immunogenic. Landsteiner performed an experiment to show the role of hapten and carrier (fig. 3.5). The figure shows that the animal immunized with a hapten-carrier conjugate produced three distinct sets of antibodies. One set comprised hapten specific antibodies that reacted with the same hapten on any carrier as well as with free hapten. The second set of antibodies was specific for the carrier protein, as shown by their ability to bind both the hapten-modified and unmodified carrier protein. Lastly antibodies reacted only with specific conjugate of hapten and carrier used for immunization. This binding of haptens by antihapten antibodies is significant in defining the precision of antigen



Injection with	Antibodies formed
Hapten (DNP)	None
Protein carrier (BSA)	Anti-BSA
Hapten -carrier conjugate (DNP-BSA)	Anti-DNP (major) Anti-BSA (minor) Anti-DNP/BSA (minor)

Fig. 3.5. Landsteiner experiment to explain the role of hapten

binding by antibody molecules. Anti-hapten antibodies are important in medical science as they mediate allergic reactions (Type-1 Hypersensitivity reaction) to penicillin and other compounds which react with host protein to form a coupled hapten that can stimulate an antibody response.

Many biological substances like drugs, peptide hormones and steroid hormones can function as hapten. Conjugate of these haptens with large protein carrier can be used to produce hapten-specific antibodies. These antibodies are useful clinically for measuring the presence of various chemicals in our body.

A hapten-carrier conjugate with hapten (DNP-Dinitrophenol) chemically linked to a large protein carrier (BSA–Bovine serum albumin) is shown in Fig. 3.5. Immunization with DNP alone evokes no anti-DNP antibodies, but immunization with DNP-BSA combined elicits three predominant, indicating that in this reaction hapten is the immuno-deminant epitope is a hapten carrier conjugate.

3.4 Selective questions

(1) Define— Antigen, immunogen and show the difference between them.

(2) What is hapten and show its difference from carrier. Describe the Landsteiner's experiment on hapten-carrier conjugate.

(3) Define Adjuvaut. Discuss the role of different types of adjuvants.

(4) Explain the difference between sequential and conformational determinants with diagram.

3.5 Selected readings

(1) Abbas, A.K. and Lichtman, A.H. 2003. Cellular and Molecular Immunology, 5th edn. Elsevier Science, USA.

(2) Kuby J. 1999 Immunology 4th edn. Elsevier Publication, New York.

Unit 4 ❐ Complement System, MAC Mediated Lysis

Structure

4.1 Introduction

The complement system a component of the innate immune mechanism, comprises a group of more than 30 serum and cell surface proteins that interact with other immune system molecules in a highly regulated manner to provide many of the effector functions of humoral immunity and inflammation.

The name of the complement system derived from experiments performed by Charles Bordet shortly after the discovery of antibodies. He demonstrated if fresh serum containing an antibacterial antibody was added to the bacteria at a physiologic temperature (37^0C), the bacteria were lysed. If however the serum was heated to 56^0C or more, it lost its lytic property. As antibody is heat stable, so the heat labile factor that assists the reaction was named by Bordet as "complements" that enhances lytic function of antibodies.

4.2 Function of complement system

i) Certain activated complement components mediate **cytolysis** by polymerizing on cell surfaces to form pores by disrupting the integrity of the phospholipid bilayer in the membranes of these cells. Hence the microbes can be killed by **osmoticlysis.**

ii) **Opsonization** of foreign organisms or particles occurs by binding of complement proteins to their surfaces. These complement proteins are called **opsonins.** Phagocytic leukocytes express specific receptors for these opsonins. In this way, opsonins promote phagocytosis of particles or organisms.

iii) **Activation of inflammation** occurs in response to the generation of certain proteolytic fragments of complement proteins. These complement-derived peptides act on several targets. They activate mast cells, causing reactions that resemble immediate hypersensitivity; in extreme cases, this reaction can mimic

anaphylaxis and these complement fragments are sometimes called **anaphylotoxins.** Other targets of complement derived peptides include vascular endothelium, smooth muscle, and inflammatory leucocytes.

iv) The complement system promotes **solubilization and phagocytic clearance of immune complexes,** thereby minimizing the damage caused when immune complexes that are formed in the circulation deposit in tissues or vessel walls.

v) The complement system plays a significant role in **promoting humoral immune responses** by aiding in antigen presentation to B cells in germinal centres and by lowering the threshold of sensitivity of B cellls activation by antigens. These functions are mediated by receptors for complement fragments expressed on follicular dendritic cells and on B-lymphocytes. (Fig 4.1)

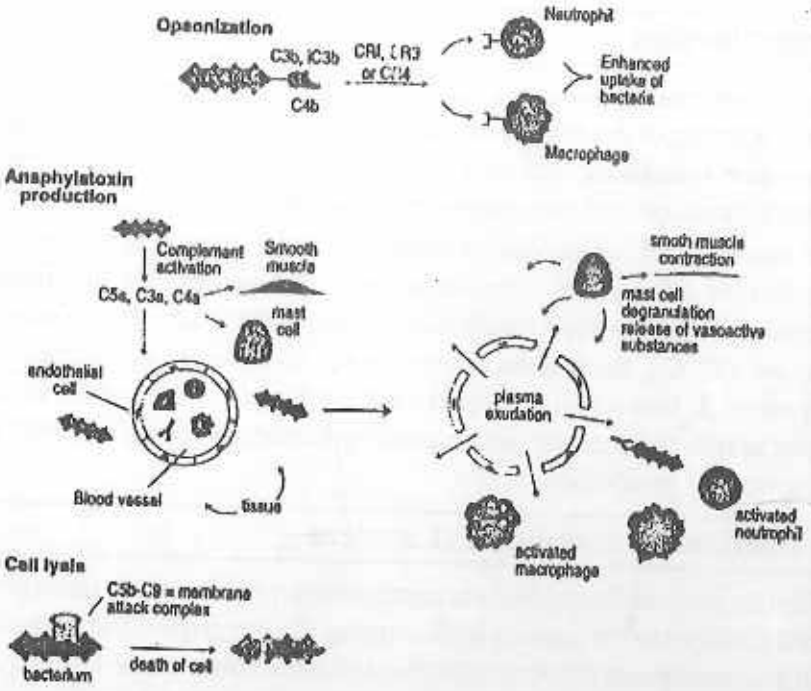


Fig. 4.1. Major functions of complement (1) : production of opsonins and anaphylatoxins, and cell lysis.

4.3 Properties of complement system

(1) The complement systems amplifies the response to microbes by means of an enzymatic cascade.

Proteolytic cascades allow for tremendous amplification, since each individual

enzyme molecule activated at one step can generate multiple activated enzymes or activated fragments, at the next step.

(2) Two converging pathways of complement activation co-exist, each one initiated by a specific set of stimuli, but both pathways share homologous molecules with similar functions. The pathways are :

a) alternative pathway, (Table 4.2 and Table 4.4)

b) classical pathway (Table 4.1 and 4.4)

Both pathways converge to a common final pathway that generates an assembly of proteins with cytolytic activity called the **membrane attack complex.** (Table 4.3).

(3) Although the components of the proteolytic cascades of the complement systems are present as soluble serum proteins, they are inactive or have only a low level of spontaneous activation in the circulation. Two mechanisms ensure that **initiation of complement cascades occurs only at certain sites and stay localized where it will be most useful.**

First, some complement components are only activated by binding to certain types of molecules that are present on the surfaces of infectious organisms but not on normal host cells.

Second, the requirement for immune complex for classical pathway initiation focuses complement activation to sites where specific antibodies bind to foreign antigens.

(4) The various biologic functions of the complement system are mediated in two general ways, First, complexes of complement proteins become directly bound to, and influence the fate of, microbes or immune complexes. Secondly, soluble fragments of complement proteins are generated during activation. These fragments diffuse from the sites where they are generated, bind to specific receptors or other nearby cells and thereby activate effector functions of those cells.

(5) The complement system is highly regulated by several soluble and cell membrane associated proteins that inhibit complement activation at multiple steps (Table 4.5). These regulatory mechanisms have two main functions. Firstly they limit or stop complement activation after the system has appropriately performed its functions. Secondly they prevent abnormal or constitutive complement activation in the absence of microbes and antibodies. Thus these regulatory mechanisms in effect enable the complement system to distinguish self from non-self and thereby prevent damage to normal tissues.

Classical Pathway :

Table 4.1 : Classical complement pathway : Proteins that participate in formation of C5 convertase

Component	Active protein/ split product	Immunologic function
C1 (6A+6B+ 6C+2r+2s)	C1q	Binds to Fc region of IgM/IgG
	C1r	serine protease : enzymatically activates C1S.
	C1s	Serine protease : enzymatically activates C4 & C2.
C4 ($\alpha + \beta + \gamma$)	C4a	Peptide mediator of inflammation (anaphylatoxin)
	C4b	Binds C2 forming complex that is cleaved by C1s to yield C4b2a.
C2 (α)	C2a	serine protease : C4b2a acts as C3 converstase
	C2b	? (unknown function)
C3 ($\alpha + \beta$)	C3a	Peptide mediator of inflammation (anaphylatoxin)
	C3b	Binds to C4b2a to form C4b2a3b (C5 convertase); major opsonin.

Alternative Pathway :

Table 4.2 : Alternative complement pathway : Proteins that participate in formation of C5 converstase

Component	Active protein / Split product	Immunologic function
C3 ($\alpha + \beta$)	C3a	Peptide mediator of inflammation (anaphylatoxin)
	C3b	Binds factor B, forming complex that is cleaved by factor D to yeield C3bBb
Factor B (α)	Ba	? (unknown function)
	Bb	serine potease : (C3bBb acts as C3 covertase, which generatee (C3B Bb 3b) (C5 convertase).
Factor D (α)	D	Serine protease : cleaves factor B that is bound to C3b to form C3 convertase.
	C3	
Properdin (α_4)		Binds to and stabilizes C3bBb (C3 convertase).

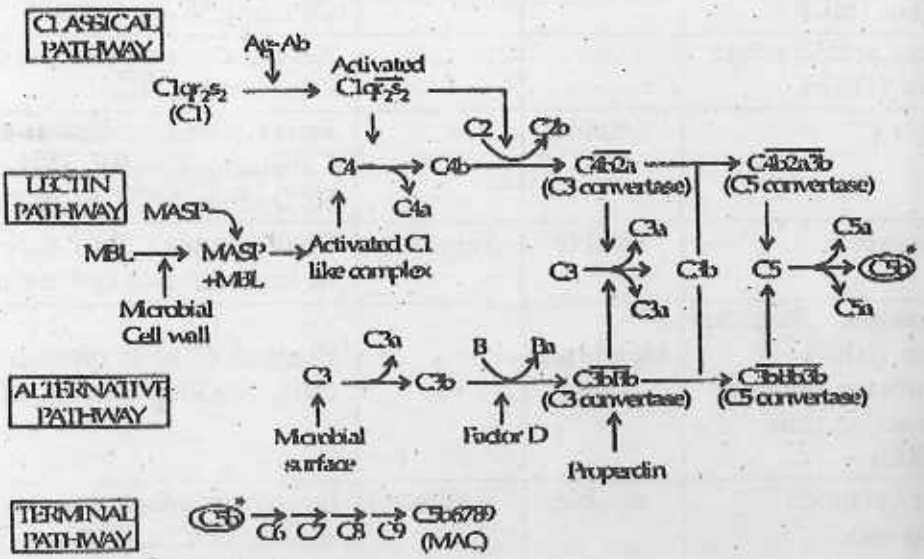
Terminal Complement Pathway :

Table 4.3 : Terminal complement pathway: Proteins involved in the formation of the Membrane Attack Complex (MAC)

Component	Active protein / Split product	Immunologic function
C5 ($\alpha + \beta$)	C5a	Peptide mediator of chemotaxis and inflammation (anaphylatoxin)
	C5b	Binds C6 to initiate formation of MAC
C6 (α)	C6	Bind to C5b to form C5b6; binds C7
C7	C7 (α)	After an amphophilic transition of C5b67, the resulting complex inserts into the lipid bilayer.
C8 ($\alpha + \beta + \gamma$)	C8	C5b678 binds multiple C9 molecules, initiating their polymerization
C9	C9	polymerizes to complete formation of MAC pore.

Combined Pathway :

Table 4.4 : Combined Pathway



In all the tables shown in text the complexes with eazymatic activity are designated by a bar over the number or symbol like C $\overline{4b2a}$, $\overline{C3bBb}$.

4.4 Regulation of complement system

Table. 4.5 Proteins that regulate complement system

Protein	Type	Pathway affected	Immunologic function
C1 inhibitor (C1 inh)	soluble	classical	Serine protease inhibitor ; Causes $C1r_2s_2$ to dissociate from C1q
C4b binding Protein (C4b BP)	,,	classical & lectin	Blocks formation of C3 convertase by binding C4b ; cofactor for cleavage of C4b by factor I
Factor H	,,	Alternative	Blocks formation of C3 convertase by binding C3b; co-factor for cleavage of C3b by factor I
Complement receptor (CRI) type I membrane co-factor protein (MCP)	Membrane-bound	all	Block formation of C3 convertase by binding C4b or C3b; co-factor for cleavage of C4b or C3b by factor I
Decay accelerating factor (DAF)	,,	,,	Accelerates dissociation of $\overline{C4b2a}$ and $\overline{C3bBb}$
Factor I	soluble	,,	Serine protease; cleaves C4b & C3b using C4bBP, CRI, factor H, DAF or MCF as co-factor
S-protein	soluble	Terminal	Binds soluble C5b67 & prevents its insertion into cell membrane
Homologus Restriction factor (HRF) Membrane Inhibitor of reactive lysis (MIRL)	Membrane bound	,,	Binds to C5b678 on autologous cells, blocking binding of C9
Anaphylatoxin inactivator	soluble	Effector	Inactivates anaphylatoxin activity of C3a, C4a and C5a by carboxy peptidase N removal of C-terminal Arginine

4.5 MCA mediate lysis

The terminal pathway of complement activation polymerize to form pores in the membrane of the pathogen (Table 4.3). The membrane attack complex start with action of C5 convertase on C5. C5 is cleaved to form C5b and C5a. Subsequently C5b the initiator of the complex binds to C6 and the C5b, 6 complex then binds to C7. The hydrophobic domain of C7 acts on lipid bilayer of the pathogen.C567 is bound with C8 and C9. The complex protein of all these components induces polymerization into the annular or ringstructure on the membrane and is called membrane-attack complex. The process leads to complete lysis of the pathogen membrane (fig. 4.2). The final step in formation of the complex is the terminal molecule C9 binding and polymerization. C9 is a perforin like chemical and can move into the pore created in membrane by hydrophilic and amphiphilic property. The free flow of MAC chemical through the pore destroys the osmotic balance of the cell and damage it completely.

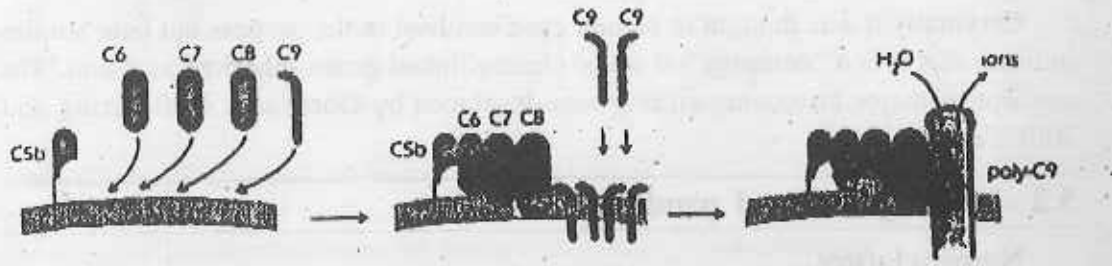


Fig. 4.2. Formation of the membrane attack complex. Late stage complement components C5b-C9 bind sequentially to form a complex on the cell surface. Multiple C9 components bind to this complex and polymerize to form poly-C9, creating a channel that disrupts the cell membrane.

4.6. Selective questions

(1) Mention the functions of the complement system.
(2) With a schematic diagram show the relation between classical and alternative complement pathways. Mention the role of different factors (components) of the complement.
(3) Describe the membrane–attack complex process involving three major pathways.

4.7 Selected readings

(1) Abbas, A.K. and Lichtman, A.H. 2003 **Cellular and Molecular Immunology,** 5th edn. Elsevier Science, USA.
(2) Janeway, C.A., Travers, P., Walport, M and Capra J.D. 1999 **Immuno Biology** 4th edn. current Biology Publication, New York.

Unit 5 ❐ Structure of MHC (elemenatary idea)

Structure

5.1 Introduction

The term major histocompatibility complex is arrived at from research oriented for acceptance or rejection of tissues-literally **histocompatibility**-transplanted between different members of the same species.

Originally it was thought to be one gene involved in the process but later studies indicate that it is a "complex"- a set of closely linked genes inherited as a unit. The concept of major histocompatibility was developed by Gorer and Snell during mid 20th Century.

5.2 MHC genes and products

Nomenclature

Human chorosome 6 contain human MHC, known as HLA (human leucocyte antigen). Names of other species for example are BOLA for bovine system, SLA for swine. The name of mouse MHC is H-2 locus located on chromosome 17.

Two major sets of MHC genes, known as **MHC class I** and **MHC class II** and their cell-surface expressed products are involved in T cell responses. The three independent genes that code for human class I MHC molecule are called HLA-A, HLA-B and HLA-C. The MHC class II molecules are obtained from 3 sets of genes --HLA-DP, HLA-DQ and HLA-DR.

Each MHC Class II subregion contains an A & B gene that code for a chain, α or β respectively of a two chain MHC class II molecule. Thus HLA-DPA gene codes for DPα of the DP molecule and HLA -DPB gene codes for the other chain DP-β of the HLA-DP molecule.

Mouse MHC Class I molecules are coded by K, D & L gene H2 has 2 MHC class II regions-rather than 3 in humans known as I-A, I-E that code for I-A α β and I-E α β molecules respectively.

Pattern of MHC molecules expression in different cells

MHC Class I molecules are expressed at varying levels on almost every nucleated cell in the body. MHC II molecules have a more limited distribution than class I molecule. They are present on Antigen presenting cell (APC).

Expression of MHC-I & MHC II molecules are coordinate, in that all molecules of each class can be expressed on the cell surface at the same time. They are however under distinct control.

5.3 Structure of MHC molecules

The simplified structure of MHC I and MHC II are shown in Fig. 5

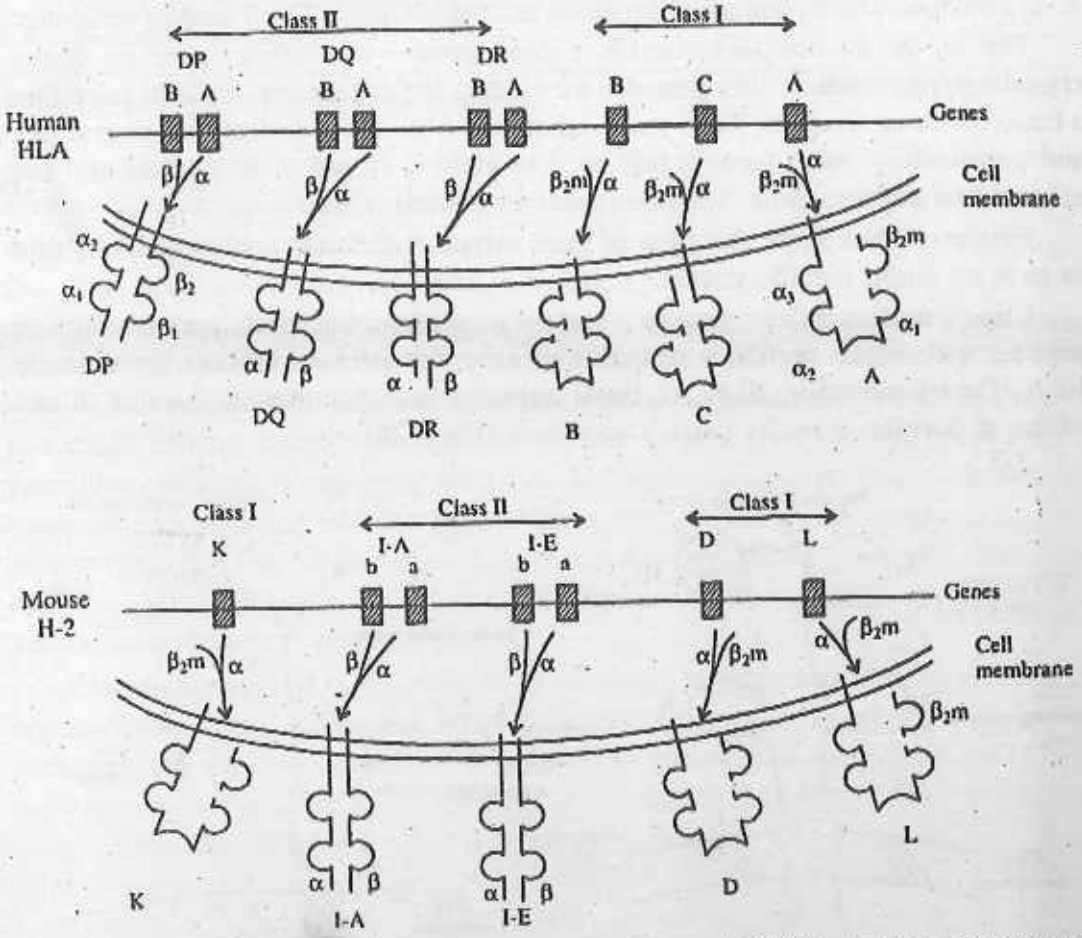


Fig.5.1 a and b Simplified depiction of the human (A) and mouse (B) MHC, showing regions & genes coding for polymorphic MHC Class I and Class II molecules. β_2m = β_2 microglobulin, encoded outside MHC.

5.3.1 Structure of MHC Class I molecules

Each Class I-gene codes for a transmembrane glycoprotein of approximate molecular weight 43 K Da, which is referred to as α, or heavy chain. It has 3 extra cellular domains α_1, α_2, α_3. It is expressed at the surface of a cell in covalent association with a small invariant polypeptide called β_2 microglobulin (β_2 m; molecular weight 12 K Da). β_2 m is coded by another chromosome and has a domain analogous to a single Ig domain.

At the cell surface, MHC I & β_2 m appear as a 4-domain molecule with α_3 & β_2m justaposed closest to the membrane.

The sequence differences between different Class-I molecules is restricted to α_1 & α_2 domains. The α_3 domain is invariant and binds CD8, a T-cell surface molecule.

The α_1 & α_2 domains contain a deep groove or cleft as seen by X-ray crystallographic studies. This groove in the binding site of peptides. the cleft resembles a basket with an irregular floor, made up of amino acid in a β plated sheet structure and surrounding walls form α helices. The cleft is closed at both ends and can accomodate peptides with 8-9 aminoacids in a linear way.

Further studies show that floor of each variant is different and thus can be said to have an allele specific pocket.

Class I molecules can bind to a variety of peptides but binds preferentially to peptides with certain **motifs.** Such motifs are known as **anchor residues.** For example HLA Class-I molecule HLA-A2 binds peptides with peptides at position 2 and valine at position 9 in the peptide sequence. (Fig 5.2b).

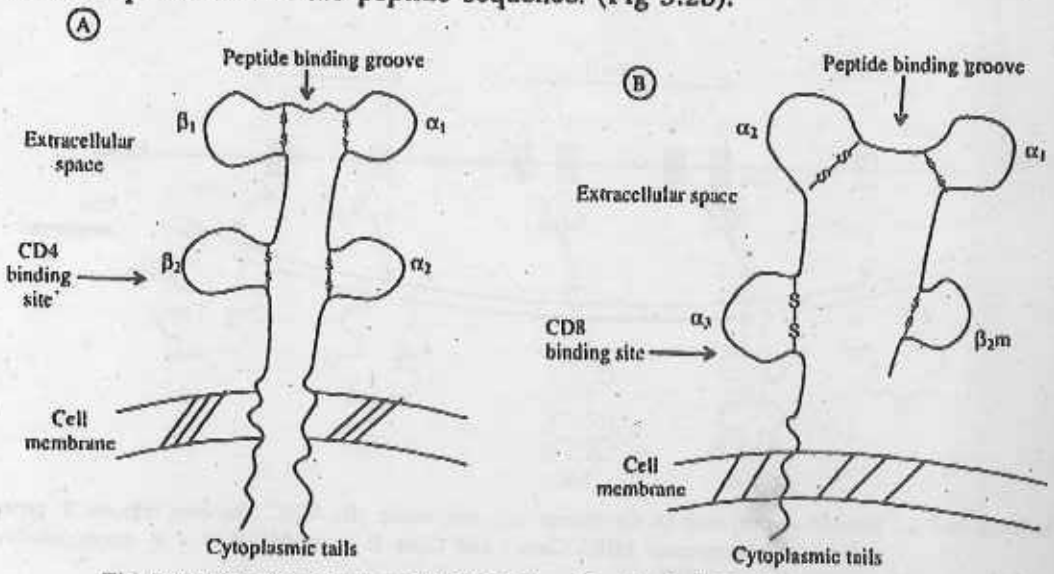


Fig. 5.2A & B. Depictions of (A) MHC Class II molecule (B) MHC Class-I molecule.

5.3.2 Structure of MHC class II molecule

MHC Clas-II α & β genes code for chains of approximate molecular weight 35,000 Da respectively. Each chain is a transmembrane glycoprotein molecule with cytoplasmic tails and extracellular Ig-like domains, the domains referred to as α_1, α_2 and $\beta_1$and β_2. It is made up of variable or polymorphic regions α_1 and β_2 and invariant or nonpolymorphic regions α_2 and β_2. The T-cell molecule CD4 binds to the invariant region.

The peptide binding groove of MHC II molecule in formed by interaction between domains of different chains, the α_1 and β_1 domain. The floor consists of 8 β-pleated sheets with each α_1 and β_1 contributing 4 each. The groove is open at both ends, allowing larger peptides to bind. The MHC Class-I binds peptide varying in length from 12 to approximately 17 aminoacids in a linear array.

Peptide binding to MHC Class II also exhibit motifs. Because the length of peptides are variable the motif is generally seen in the central region of the peptide, the region that fits inside the MHC class-II binding groove. (Fig. 5.2A).

5.4 Function of MHC molecules

MHC Class I : (1) Associated with antigenic peptides of infecting pathogens produced within the host cell.

(2) Vesicular transport of MHC-I peptide complex to the infected cells membrane.

(3) Interaction with TCR of Tc cells; co-induction by CD8 molecule.

(4) Activation of Tc cell and destruction of infected cell.

MHC-Class II : (1) Vescicular transport towards endolysome.

(2) Fragmentation of foreign peptides by lysomal enzymes present in endolysome; peptides derives from phagocytosed pathogen.

(3) Association of MHC-Class II with foreign peptide.

(4) vescicular transport to cell membrane and expression/ presentation of foreign peptide MHC II complex.

(5) Association with TCR of T_H cells; co-induction of T_H cells with CD4 molucules.

(6) Secretion of cytokines by TH cells, recruitment of macrophages, NK cells & Tc cells to infection site and formation of their effector population.

5.5 Selective questions

(1) Draw and describe class I MHC and class II MHC molecule.
(2) Write notes on :
   (a) Features of class I MHC and class II MHC
   (b) Function of two types of MHC molecules.
   (c) Pattern of MHC molecule expression in different cells.

5.6 Selected readings

Kuby, J. 1999 Immunology 4th edn. Elsevier publication, New york.

Unit 6 ❐ Structural Diversity of Immunoglobulin

Structure

6.1 Introduction

Immunoglobulins function as antibodies, the antigen binding proteins that are present on the B cell membrane and also are secreted by plasma cell. The class of an immunoglobulin molecule is determined by its heavy chain (isotype). They are classified as IgM, IgG, IgD, IgE and IgA. The antigenic specificity of each B cell is determined by the membrane bound antigen binding receptor of antibody expressed by the cell. The antibody on a B lymphoblast can recognize epitopes on macromolecules with incredible precision. Protein antigens that differ by only a single aminoacid often can be distinguished from each other.

The amazing feature of the vertebrate immune system is its ability to respond to an apparently unlimited number of foreign antigens. It has been estimated at the present situation that an individual can build up B (and T) cells with different antigenic specificities upto range of 10^{15} to 10^{18} times. The genome of the individual contain only less number of genes e.g 30,000 to 40,000 approximately. This diversity is generated due to participation of multiple germ like gene segments generating antibody along with some special arrangements and procedures. It may also be noted here that this maturation process and also incase of T lymphocyte it is accompanied with another selection process by which self reactive antibodies and T cells are all eliminated. This process develops **tolerance** or nonresponsiveness towards body cell antigen.

6.2 Points of diversity

Multiple gene segments code for antibody structure. Light chains are composed of Kappa (κ) and Lamda (λ) chain and heavy chains are determined by H chains. The structure of the constant region determines whether κ or λ arrangement will be taken in light chain. The κ or λ light chains and the heavy chains are encoded by separate multigene families situated on different chromosome. The arrangement of genes in Mouse are in 6, 16, 12 and in Human are 2, 22 and 14 respectively for K,

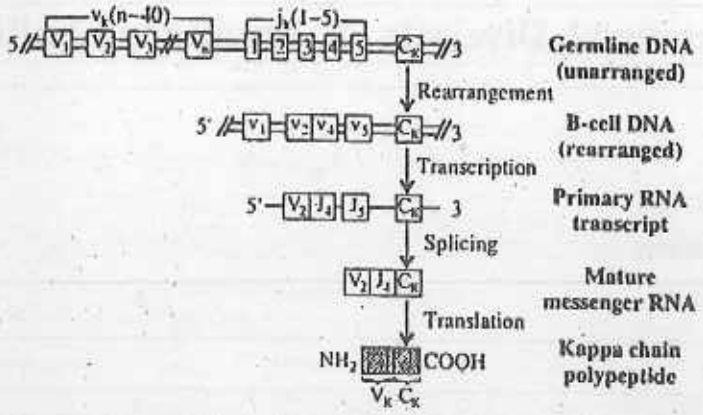


Fig. 6.1. The genetic events leading to the synthesis of a kappa light chain.

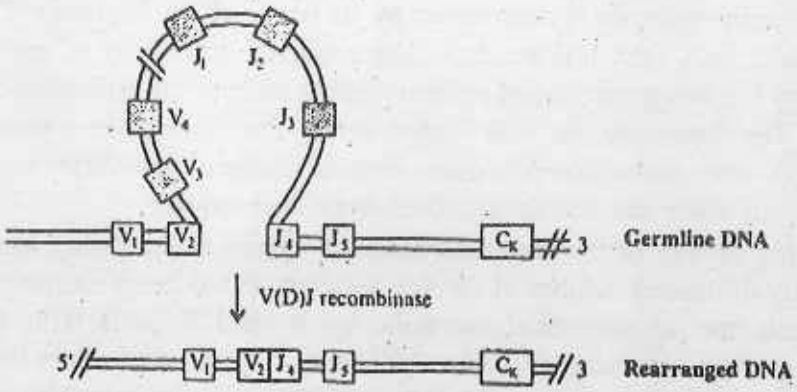


Fig. 6.2. Rearrangement of DNA coding for a kappa light chain

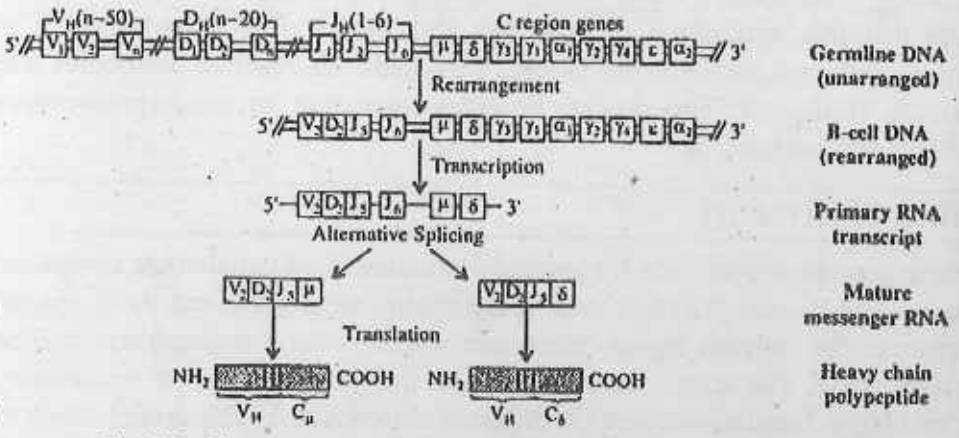


Fig. 6.3. The genetic events leading to the synthesis of a human heavy chain

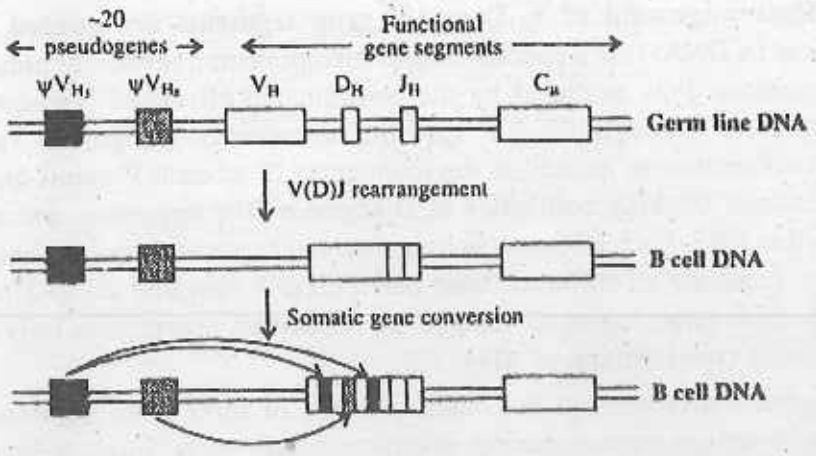


Fig. 6.4. Somatic gene conversion generates diversity in Ig genes of several species. The figure illustrates the phenomenon in the chicken Ig heavy-chain locus; short sequences of DNA from one or more pseudogenes (3 and 8 in the figure) are copied into the rearranged B-cell VDJ unit.

λ and H chains. In germ line DNA there are multigene families--gene segments (exons) which are separated by noncoding introns (silent). The segment arrangement of light chains are VJC segment and in heavy chain VDJ for variable region and C for constant region (Fig. 6.1, 6.2, 6.3, 6.4). The process of anitbody structure formation is same as in any other peptide chain formation during protein synthesis like; germline DNA–B cell DNA–Primary transcript–Transcription–primary transcript splicing–m RNA formation–Translation–protein (Peptide) chain formation according to leader sequence.

The rearrangement of Ig gene and its subsequent diversified structure is achieved by several means. (1) Compete V regions are generated by the somatic recombination of separate gene segments as shown in diagram. (2) V region genes are present in multiple copies. According to Janeway *et al* (2005) the segments for light and heavy chains are as in Table 6.1.

Number of functional gene segments in human immunoglobulin

Table. 6.1 A glimpse of different segment

Segment	Light chain		Heavy chain
	K	λ	H
Variable (V)	40	30	65
Diversity (D)	0	0	27
Joining (J)	5	4	6

(3) **Rearrangement of V, D and J gene segments are guided by flanking sequences in DNA.** It is a special kind of recombination involving non-homologous gene segments. It is mediated by the co-ordinated effects of lymphocyte specific recombinases. The lymphocyte specific recombinas recognizes specific DNA recognition sequences located in the intervening 3′ of each V exons and 5′ of each J segment and flanking both sides of D segment. The sequences are one turn RSS and two turn RSS. Each RSS contains a conserved heptamer sequence and a conserved nonamer sequence of different base pair (12bp 1 turn and 23bp-2 turn). During rearrangement gene segments flanked by a one-turn spacer join only to segments flanked by a two turn spacer. (Fig. 6.5).

The gene rearrangement that combines two or three gene segments to form a complete V-region exon generates diversity in two ways. First, there are multiple different copies of each type of gene segment, and different combinations of gene segments can be used in different rearrangement procedures. This is called **combinatorial diversity.** Secondly **Junctional diversity** is introduced at the joints between the different gene segments as a result of addition and subtraction of nucleotides by the recombination process. A third source of diversity is also combinatorial, arising from the many possible different combinations of heavy and light chain V-regions that pair to form that antigen binding site in the immunoglobulin molecule. The two means of generating combinatorial diversity alone could give rise to more than 10^6 different antibody molecules. Those two combinatorial diversity along with junctional diversity can lead to total receptors of 10^{11} different types that form the repertoire of receptors of B cell. Finally somatic hypermutation introduces point mutation into the rearranged V-region genes of activated B cells creating further diversity.

The steps can be elaborated in more detail as mentioned below.

The ultimate antibody diversity is generated by four main process three of which are consequences of the process of recombination and the last one is a mutational process.

(1) **Inherited gene segments are used in different combinations–** The V,D,J gene segments have multiple copies for combination. For human K light chain there are approximately 40 functional V_k gene segments and 5 J_k gene segments and so there is scope for 200 different V_k regions. For λ light chain there are approximately 30 functional V_λ gene segments and 4 J_λ gene segment, yielding 120 possible V_λ regions. So in all, 320 different light chains can be made. Similarly for heavy chains in human there are 65 functional V_H gene segments, approximately 27 D_H gene segments and 6 J_H gene segments, and thus around 11,000 different possible V_H

region (65×27×6 is round 11,000). The ability to create different combinations of a small number of gene segments is known as **combinatorial diversity**. The gene segments which do not take part in encoding protein are called Pseudogenes.

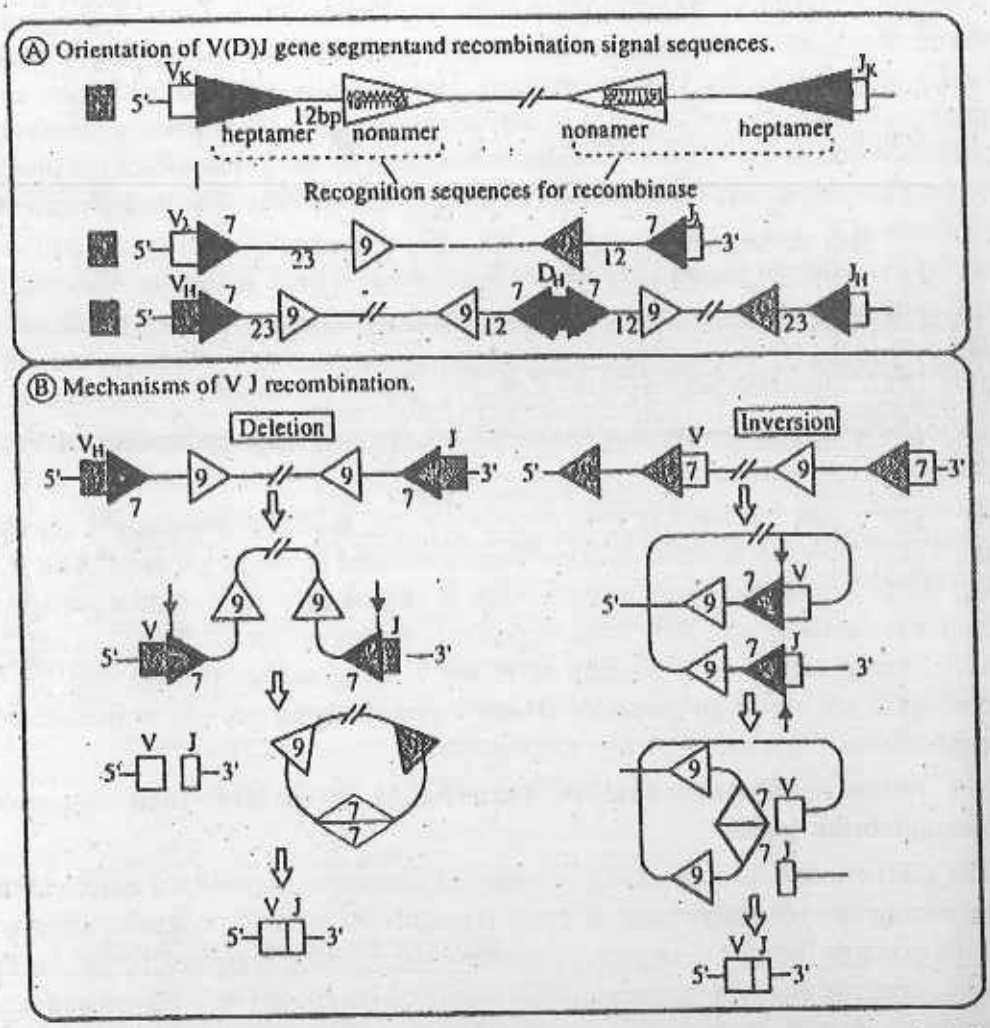


Fig. 6.5. V(D)J recombination. The DNA sequences and mechanisms involved in recombination in the Ig gene loci are depicted. The same sequences and mechanisms apply to recombinations in the TCR loci.
A. Conserved heptamer (7 bp) and nonamer (9 bp) sequence, separated by 12 or 23-bp spacers are located adjacent to V and J exons (for κ and λ loci) or V, D, and J exons (n the H chain locus). The V(D)J recombination recognizes these recombination signal sequenses and brings the exons together.
B. Recombination of V and J exons may occur by deletion of intervening DNA and ligation of the V and J segments (*left panel*) or, if the V gene is in the opposite orientation, by inversion of the DNA followed by ligation of adjacent gene segments (*right panel*). Arrows indicate the sites where germline sequences are cleaved and later rejoined.

(2) Variable addition and subtraction of nucleotides at the junctions between the gene segments encoding the V region contributes to diversity in the third hypervariable region (CDR3)

The third region of hypervariable light and heavy chain falls at the junction between the V gene segment and the J gene segment and in the heavy chain is partially encoded by the D gene segment. The diversity of the third hypervariable segment. The diversity of the third hypervariable region is significantly increased by the addition and deletion of nucleotides at two steps in the formation of the junctions between gene segments. This is known as **junctional diversity**. The added nucleotides are known as P- nucleotides and N-nucleotides. (Fig. 6.6). P-nucleotides sequences make up palindromic sequences added to the end of the gene segments, N-nucleotides are so called because they are non-template-encoded. They are added by an enzyme called terminal deoxy-nucleotidyl transferase (TdT) to single stranded ends of the coding DNA after hairpin, cleavage (see Fig. 6.6).

(3) Specialized enzymes are required for somatic recombination of V gene segments

Two genes that stimulate Ig gene recombination, called recombination activating genes 1 & 2 (RAG-1 & RAG-2) have been identified in the pre B cells. The RAG-1 and RAG-2 proteins form a dimer that is responsible for lymphocyte-specific recombinase activity and they have important properties like (i) They are cell-type specific, being acting only in cells of B and T lymphocytes lineage and (ii) The recombinase are active in imnature B and T lymphocytes but not in mature cells. (Fig.6.6).

(4) Somatic hypermutation introduces diversity into expressed immunoglobulin genes

The mechanisms for generating diversity as discussed in previous points all take place during the rearrangement of gene segments in the initial development of B cells in primary lymphoid organs. However in secondary lymphoid organ there is an additional mechanism that generates diversity throughout the V region and which operates on B cells in secondary lymphoid organs after functional antibody genes have been assembled. This process is known as **somatic hypermutation** and it introduces point mutations into the V region of the rearranged heavy and light chain genes at a very high rate giving rise to mutant immunoglobulin molecules on the surface of the B cell. The antibody diversity in adult individual is mainly derived from somatic hypernutations acquired during life time. This system of heritable and acquired components of diversity are present in mammal. Whereas Birds do not use somatic recombination or hypermutation to maintain diversity, but create their

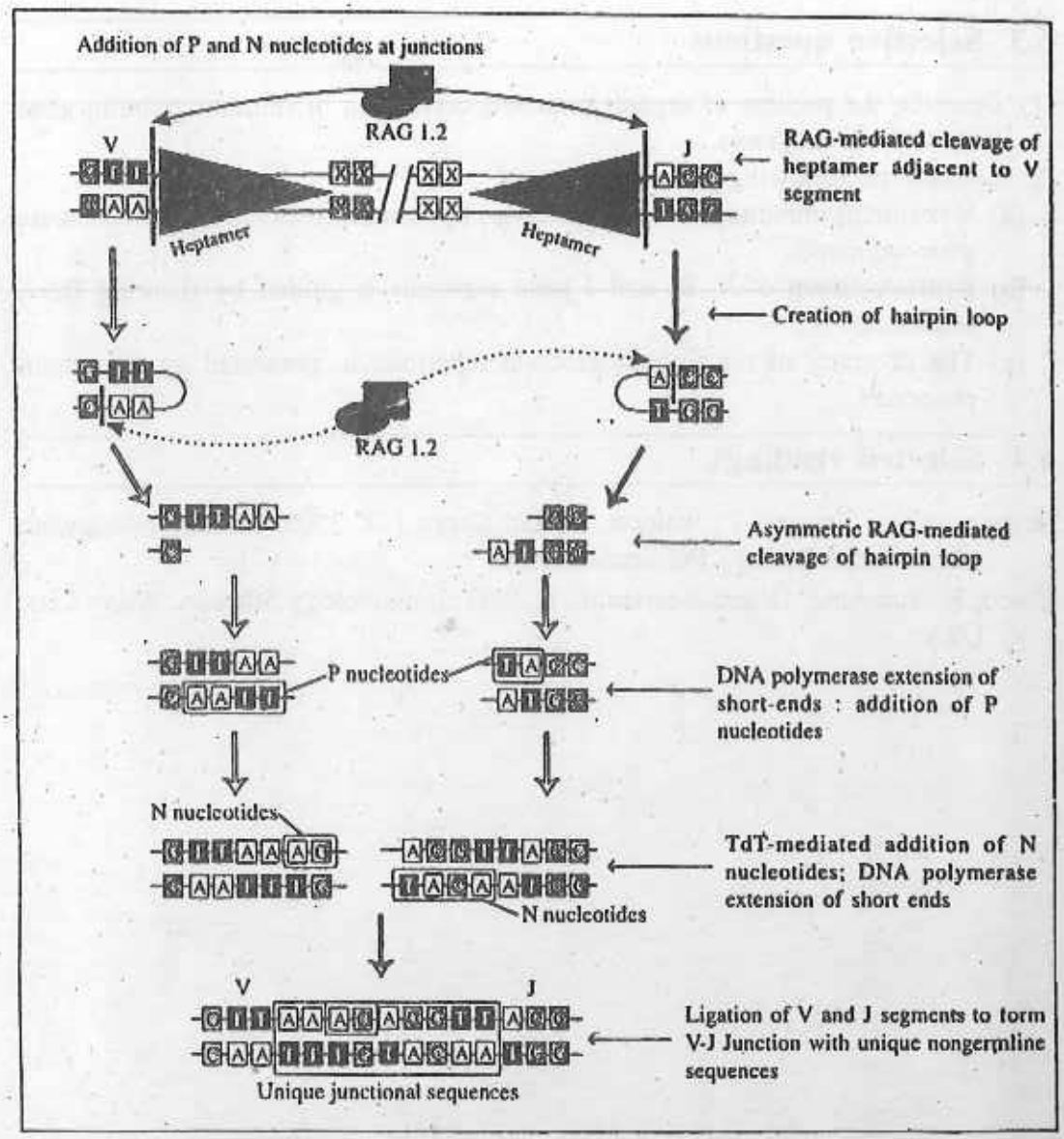


Fig. 6.6. Nucleotides (P sequences) may be added to broken DNA ends to repair asymmetric breaks. Other nucleotides (N regions) may be added to the sites of VD, VJ, or DJ junctions by the action of the enzyme terminal deoxyribonucleotidyl transferase (TdT), generating new sequences that are not present in the germline.

antibody pool gene conversion from germline pseudogenes. So at the end it can be presumed that the combination of all these sources of diversity creates a vast resource of antibody specificity from a limited number of genes.

6.3 Selective questions

(1) Describe the process of organization and expression of immunoglobulin gene with suitable diagrams.

(2) Explain the following points :

(a) V region of immunoglobulin is generated by somatic recombination of separate gene segments.

(b) Rearrangement of V, D, and J gene segments is guided by flanking DNA sequences.

(c) The diversity of the immunoglobulin repertoire is generated by four main processes.

6.4 Selected readings

Janeway, C.A., Travers, P., walport, M and Capra J.D. 2005. Immunobiology,6th edn..Current Biology Publication, N.Y.

Coico, R. Sunshine, G and Benjamini, E 2003. Immunology 5th edn. Wiley Liss. USA

Unit 7 ❐ Hypersensitivity

Structure

7.1 Introduction

Immune response to antigen or allergen with hightened or inappropriate nature upon reexposure is called Hypersensitivity. The reactions are broadly considered as of four types.

7.2 Type I hypersensitive reaction (IgE-mediated hypersensitivity)

It is induced by certain type of antigens, referred to as allergens. The term allergen refers specifically to non-parasitic antigens capable of stimulating type I responses in allergic individuals.

The IgE class of antibody binds with high affinity to Fc receptor on the surface of tissue mast cells and blood basophils and are said to be sensitized. A later exposure to the same allergen crosslinks the membrane-bound IgE on sensitized mast cells & basophils, causing degranulation of these cell. The pharmacologically active mediators released from the granules act on the surrounding tissues. The principal effects are-vasodilation and smooth muscle contraction. It may be either systemic or localized, depending on the extent of mediator release.

7.2.1 Common allergens of this type of reaction.

1) **Protein**
   Foreign serum
   vaccines
2) **Plant pollens**
   Rye grass
   Ragwood
   Timothy grass
   Birch trees
3) **Drugs**
   Penicillin
   Sulfonamides
   Local anesthetics
   Salicylates
4) **Food**
   Nuts
   Sea foods
   Eggs
   Pea, Beans & Milk
5) **Insect products**
   Bee venom
   Wasp venom
   Ant Venom
   Cockroach calyx
   Dust mites
6) Mold spores
7) Animal hair & dander

7.2.2 Components of Type I reactions

(1) Allergens (2) Reaginic Antibody (IgE) (3) Mast Cells & Basophils

The reaginic activity of IgE depends on its ability to bind to a receptor specific for the Fc region of the ε chain e.g. FcεRI and Fcε R II.

Mast cells and Basophils express FcεRI, the high affinity IgE receptor. This helps to bind IgE despite its low concentration (1×10^{-7}/M).

7.2.3 Structure of FcεRI (Fig. 7.1a)

(1) FcεRI contains 4 polypeptide chains : an 'α' & a 'β' chain and 2 identical disulphide-linked 'γ' chain.

(2) The external region of the 'α' chain contains 290 aa domains that exhibit homology with Ig domain. The domains interact with C_H3/C_H3 & C_H4/C_H4.

(3) The β chain spans the plasma membrane four times and is thought to link the α, chain to the γ homodimer.

(4) The 2 γ chain are **disulfide linked** & extend a considerable distance into the cytoplasm.

(5) Each γ chain has a conserved sequence in its cytosolic domain known as an **immunoreceptor hyrosine-based activation motif** (ITAM).

(6) Allergen mediated crosslinkage of the bound IgE results in aggregation of the FcεRI receptor & rapid tyrosine phosphorylation, which initiates the process of mast-cell degranulation.

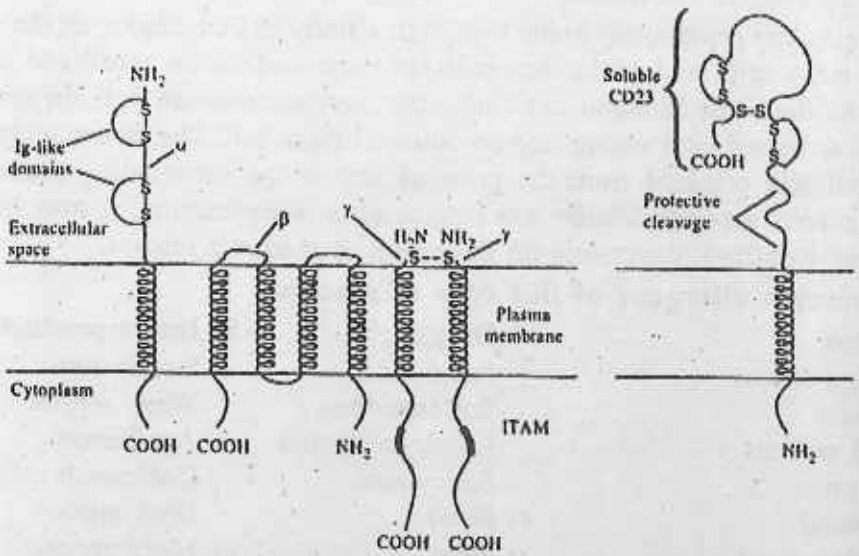


Fig. 7.1. Schematic diagrams of the high affinity FcεRI and low affinity FcεRII receptors that bind the Fc region of IgE. Fig. 7.1(a) shows active ITAM motif 7.1(b) shows NH_2 terminal towards cell internal space and COOH terminal towards extracellular space.

7.2.4 Structure of Fc ε RII/CD 23) (Fig. 7.1b)

(1) It has a lower affinity for IgE & specific for C_H3/C_H3 domain of the IgE

(2) Consists of a single polypeptide with a large extracellular domain, a single transmembrane domain & a short cytoplasmic tail.

(3) An important character is that the c-terminus of the polypeptide is extracellular & the N-terminus is cytosolic.

(4) A soluble form of $F_C\varepsilon RII$ (CD 23) may be generaed by autoproteolysis of the membrane receptor.

(5) IG-E crosslinking Fc ε RII has been shown to activate B-cells, alveolar macrophages & eosinophils; the soluble form has been shown to enhance Ig-E production by B cells.

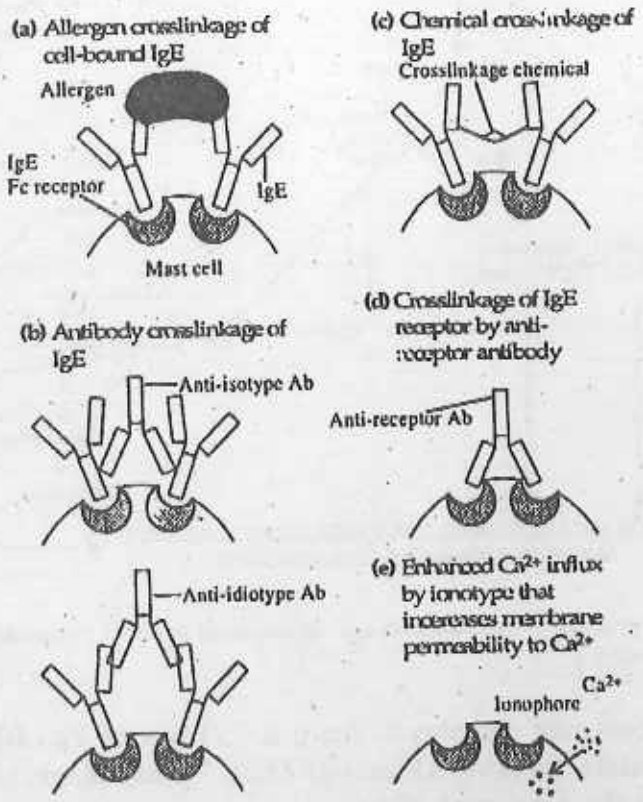


Fig. 7.2. Schematic diagrams of the mechanisms that can trigger degranulation of mast cells. Note that mechanims (b) and (c) not require allergen; mechanisms (d) and (e) require neither allergen *nor* IgE and mechanism (e) does not even require receptor crosslinkage.

7.2.5 Mechanism of Ig-E mediated degranulation

(1) **Receptor crosslinkage :** IgE mediated degranulation begins when an allergen crosslinks Ig-E that is bound to the Fc receptor on the surface of a mast cell or basophil. (fig. 7.2) Crosslinking can be mediated by (a) polyvalent allergen; (b) anti-isolypic or antiidiotypic antibody (c) Crosslinking Chemicals, (d) Antirecptor antibody, (e) influx of Ca^{+2} ions.

Note : The importance of crosslinkage is indicated by the inability of monovalent allergens, which cannot crosslink the fixed Ig-E, to trigger degranulation.

(2) **Intracellular events :** The events are shown in Chart 7.1

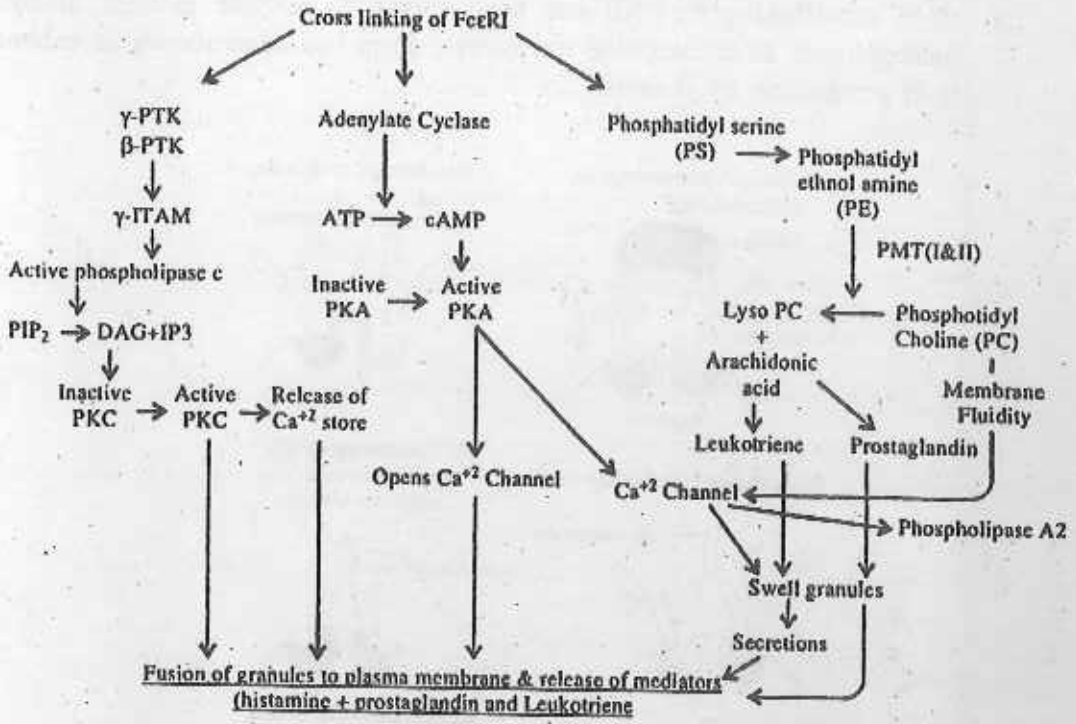


Chart 7.1 Fusion of granules to plasma membrane & release of mediaors (histamine + prostaglandin and Leukotriene)

(i) The cytoplasmic domains of the β & γ chains of **Fc ε RI are associated with protein tyrosine kinases (PTKs), Crosslinkage** of FcεRI receptor **activates** the associated PTKs.

(ii) **This results in phosphorylation of tyrosines with the ITAMs of the γ as well as phospholipase C.**

(iii) Phospholipase C converts phosphatidylinositol 4-5 bisphosphate (PIP_2) into diacylglycerol (DAG)& inositol triphosphate (IP_3)

(iv) DAG activate protein kinase C (PKC); IP_3 mobilizes intra cellular Ca^{+2} stores; PKC & Ca^{+2} is necessary for microtubulor assembly & fusion of granules with plasma membrane.

(v) Crosslinkage of FcεRI activates an enzyme that converts Phosphatidyl serine (PS) to Phosphatidyl ethanolanine (PE). PE is methylated to Phosphatidyl chohine (PC) by enzymes Phospholipid methyl transferase I (PMTI) & PMTII. Accumulation of PC cause an increase in membrane fluidly & faciliated the formation Ca^{+2} channels.

(vi) Influx of Ca^{+2} activates Phospholipale A2; this promotes breakdown of PC to lysophosphatidyl choline (lysoPC) & arachidonic acid.

(vii) Arachidonic acid is converted into potent mediators; the leukotrienes & prostaglandin D_2.

(viii) Crosslinking also activates adenylate cyclase, leading to increase in c-AMP.

(ix) c-AMP dependant protein kinases are thought to phosphorylate the granule memrane proteins, changing its pemeability to water & Ca^{+2}.

(x) The consequent swelling facilitates fusion with the plasma membrane & release of mediators.

7.2.6 Mediators of Type-I reactions

The clinical manifestations of type I hypersensitive reactions are related to the biological effects of the mediators released during mast cell or basophil degranulation.

The mediators can be classified as either primary or secondary. The primary mediators are produced before degranulation and stored in the granules. The secondary mediators either are synthesized after target cell activation or are released by the breakdown of membrane phospholipids during the degranulation process.

Leukotrienes and Prostaglandins :

They are not formed until the mast cell degranulates and enzymaic breakdown of phospholipids in the plasma membrane takes place. Their effects are more pronounced and longer lasting than that of histamine.

Leukotrienes mediate broncho constriction, increased vascular permeability and mucous production.

Prostaglandin D2 causes bronchoconstriction.

Cytokines :

Human mast cells secrete IL-4, IL-5, IL-6 and TNF-α. These cytokines alter the

local environment, eventually leading to recruitment of inflammatory cells such as neutrophils and eosinophils. IL-4 increases IgE production by B cells. IL-5 is important for recruitment of eosinophils. TNF-Y contributes to shock in systemic anaphylaxis.

Histamine : Formed by decarboxylation of amino acid histidine; major component of mast cell granules; about 10% of granule weight.

There are 3 types of histamine receptors H_1, H_2 and H_3; they have different tissue distribution and function.

Binding of histamine to H1 induces contraction of smooth muscles of intestine & branches; increased permeability of venules; increased mucous secretion by goblet cells.

Interaction with H2 receptors increases vasopermeability and dilation and stimulate exocrine glands.

Remark : Histamine binding to H2 receptros on mast cell and basophil supresses degranulation; thus has a negative feedback on mediator release.

7.2.7 Consequences

(A) Systemic Anaphylaxis :

It is a shock like and often fatal state whose onset occurs within minute of type-I hypersensitive reactions.

On injection of antigen to a sensitized guineapig, its respiration is laboured, blood pressure drops. As smooth muscle of GI tract and bladder contracts, the animal defaecates and urinates. Finally bronchial constriction results in death by asphyxiation within 5 minutes of injection.

All these events stem from systemic vasodilation and smooth muscle contraction brought on by mediators released.

Epinephrine counteracts the effect of mediators by relaxing the smooth muscles and reducing vascular permeability.

Localized Anaphylaxis (Atopy)

In it, the reaction is limited to a specific target tissue, often involving epithetial surface at the site of allergen entry. The tendency to manifest localized anaphylactic reaction is inherited and is called **Atopy.**

Allergic Rhinitis

This results from the reaction of airborne allergens with sensitized mast cells in the conjunctival and nasal mucosa to induce the release of pharmalogically active mediator from mast cells. The mediator caused localized vasodiltion and increased capillary permeability.

Symptoms : watery exudation of the conjunctival, nasal mucosa, as well as sneezing & coughing.

Asthma

It is triggred by degranulation of mast cells with release of mediators, but instead of occuring in nasal mucosa, the reaction develop in **lower respiratory tract.**

The resulting contraction of the bronchial smooth muscles leads to bronchoconstriction. Airway edema, *mucous* secretion, and inflammation contribute to the bronchial constriction and to airway obstruction.

The asthmatic response can be divided into early and late responses. Early response occurs within minutes of allergen exposure & primarily involve histamine, leukotrienes & prostaglandins. The effect lead to brochoconstriction, vasodilation and buildup of mucous.The late response occurs hours later and involves IL-4, IL-5, Il-16 and TNF-α, eosinophil chemoactie factor (ECF), and platelet acivating factor (PAF). The overall effects are endothelial cell adhesion, and recruitment of eosinophils and neurophil in bronchial tissue.

Food allergies

Allergens crosslinking IgE on mast cells along the upper or lower GI tract can induce localized smooth muscle contraction and vasodilation.

Mast cell degranulation along the gut increase permeability of mucus membranes, so that allergens can enter blood stream. Various symptoms can develop depending on site of deposit of allergen. Symptoms include, asthma, **wheel and flare** reaction.

Atopic dermatitis : (Allergic eczema)

In allergic individuals, serum IgE levels are often elevated. The individuals develop skin eruptions that are erythromatous and filled with pus. The allergic individual shows an elevated TH_2 cells response and increased number of eosinophils.

7.3 Type II Antibody mediated cytoloxic hypersensitivity

(1) Antibody can mediat cell destruction by activating complement system to create pores in the membrane of foreign cells.

(2) Antibody can also mediate destruction by ADCC (antibody dependent cell-mediated cytotoxicity). In this process, cytotoxic cells with Fc receptors bind to the Fc region of antibodies on target cells and promote killing of cells.

(3) Antibody bound to a foreign cell can also serve as an opsonin, enabling phagocytic cells with Fc or C3b receptors to find and phagocytose the

antibody coated cell.

Examples : i) Transfusion reactions

ii) Hemolytic disease of newborn (erythroblastosis foetalis)

iii) Drug induced haemolytic anemia

7.4 Type III immune complex mediated hypersensitivity

They develop when immune complexes activate the complement systems array of immune effector molecules. C3a, C4a, C5a complement split products are anaphyltoxins and cause localized mast cell degranulation and consequent increase in local vascular permeability. They also function as chemotactic factors of nontrophils.

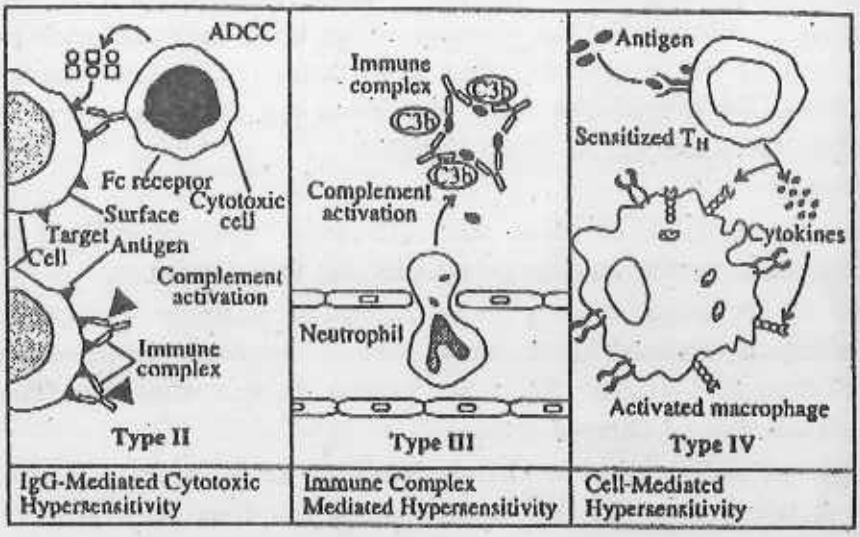


Fig. 7.3. Outlines drawings three types of Hypersensitive reactions (Type II, III and IV)

The lytic enzymes released from neutrophils as they attempt to clear the immune complex destroys localized host tissue.

Activation of complement induce agregation of platelets, resulting in release of clotting factors and lead to formation of microthrombi.e.g. Insect bite, meningitis, malaria, rheumatoid arthritis,

7.5 Type IV (DTH) : delayed type hypersensitivity (DTH)

Activation of DTH of T lymphocytes causes release of cytokine like IL-2, IFN-γ macrophage inhibiting factor (MIF) and TNF-β.

The overall effect is to draw macrophage in the area of action and activate them. Activated macrophage release enzymes and local tissue destruction results. Time lapse for reaction is typically 48-72 hours and hence named DTH e.g. *Mycrobacterium leprae* infection; *Mycobacterium tuberculosis* infecion

7.6 Selective questions

1. Mention the IgE-mediated common allergic reactions. State the syndrome, common allergeus involved and response.
2. Allergic responses can be divided into immediate and latephase responses. Explain all different types of allergic responses involved in both the cases.
3. What is delayed type hypersensitive reaction. Explain with proper diagram.
4. What are the Type II and Type III hypersensitive reaction. Explain both the reactions.

7.7 Selected Readings

Kuby, J. 1999 Immumology, Elsevier Publication, New York

Unit 8 ❐ Elementary Concept of Invertebrate Immunity

Structure

8.1 Introduction

The defence mechanism and its evolution is connected with the stepwise modification of unicellular organism to assending phyla of the invertebrate kingdom. In invertebrate phyla different types of cells, certain organs and body fluids are primarily involved in non-specific (innate) type of immune response. The major difference of this non-specific immunity from adaptive immunity is the i) absence of lymphoid organ and tissue ii) absence of immunoglobin molecule and no memory or genetic basis for synthesis of defence organs. However on the otherhand they are capable of distinguishing between self and non-self. Invertebrates possess an extremely effective physicochemical barriers as their first line of defence. Coelenterates, annelids, mollusks and tunicates has thick layer of mucus that surrounds their body, entraps and kills potential pathogens.Tough tests in coelentenates, molluscs, echinoderms and arthropods form barrier to invasion. The invaders after overcoming barriers are exposed to a range of interacting cellular and partly humoral defence reactions like :

i) Blood clotting / co-agulation and wound healing
ii) Phagocytosis
iii) Encapsulation responses
iv) Natural and inducible antimicrobial factors.

These reactions consider non-self factor and have receptor molecules role in blood cell surface. Recent observations on the evolution of immune systems supports the idea of a well developed defence mechanism in coelomate groups. Toll-like receptors is considered as a method of pathogen recognition system. The attached diagram (fig 8.3) shows the relation of *Drosphila* and mammalian Toll signalling

pathway. Toll receptor in ***Drosophila*** acts in host defense mechanism. This type of receptor has also been found in mouse.

8.2 Cellular component of immunity in invertebrates (Fig. 8.1)

Porifera

Cells-like choanocytes, pinacocytes and archeocytes are all able to internalize various materials. The archeocytes represent the major cells type in defense system

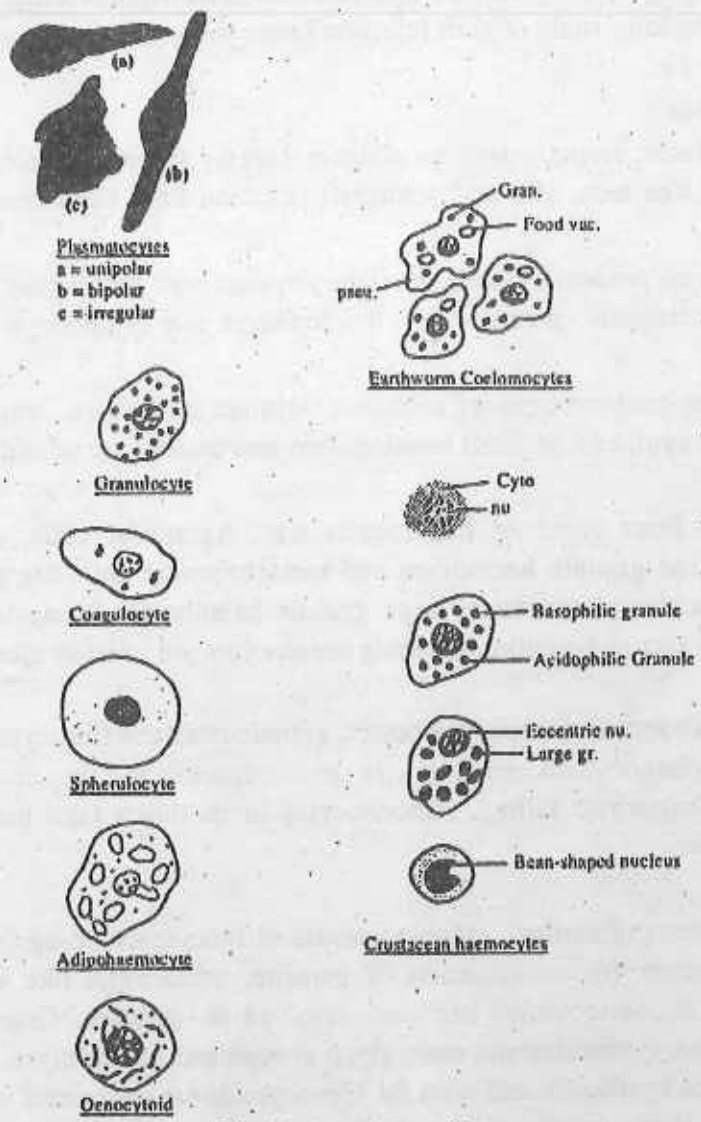


Fig. 8.1. Invertebrate cells

in sponges. The sponges also exhibit some degree of lectin formation, wound healing, histocompatibility and cytotoxic reaction. Phagocytosis is the major defense reaction and the archeoegtes are the effector cells assisted by pinococytes, coelencytes and spherulous cells.

Coelenterate

Phagocytosis, wound repair, allograft rejection has been shown by coelenterates. The regeneration power in colenterate is assumed to be most powerful type in whole invertebrate kingdom. study of graft rejection factor shows allogeneic and xenogenieic recognition power.

Platyhelminthes

These triblasitc animals perform phagocytosis by amoeboid neoblast. Common graft rejection like auto, allo and xenograft rejection have been observed.

Nemertea

Ribbonworms possess a closed circulatory system with blood cells like basophilic granulocytes, nerophilic granulocytes, macrophages and lymphocyte like cells.

Annelida

Earthworms-coelomocytes of coelomic fluid act in defense, while haemocytes are involved in synthesis of fresh haemoglobin and catabolism of old haemoglobin.

Arthropoda

Crustacea- Four types of haemocytes e.g., Agranular cells, small granule haemocytes, large granule haemocyte and metachromatic cells are present. Small granule haemocytes phagocytise, large granule haemocytes encapsulate and their 76KDa protein acts as opsonin. Agranulr haemocytes are used in clotting of blood.

Molluses

In molluscs haemocytes are halinocytes, granulocytes and fibrocytes. Halinocytes are engaged in phagocytosis, granulocytes secrete opsonin and they secrete lysosonal enzymes for phagocytic killing. Amoebocytes in molluscs take part in humoral defence reaction.

Insects

Plasmatocytes (agranular), granulocytes do phagocytosis, co-agulocytes release chemotactic factors for co-agulation of parasite, other cells like spherulocytes, adipohaemocytes, oenocytoids are also involved in defence. Granulocytes and spherulocytes can synthesize and store glyco protein and mucopolysacharides while adipohaemocytes synthesize and store fat. Oenocytoids are concerned with darkening of cuticle after larval moults and co-agulocytes help in blood coagulation.

8.3 Major functional components of the invertebrate immune mechanism

i) Opsonin in phagocytosis

It is a process in which foreign particles or micorbes get coated by certain blood borne substances known as opsonin. In invertebrale haemaglutinin acts as opsonin. (Fig. 8.2)

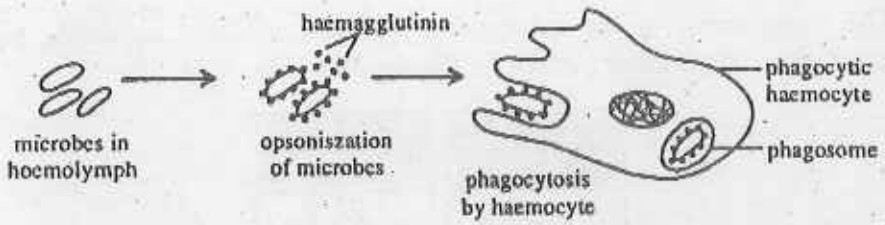


Fig. 8.2. A model of opsonization in invertebrates

ii) Nodule formation in mollusca

It is a kind of defence reaction of certain gastropods and bivalve against invading parasite. It is a process of formation of a fibrous capsule around large sized parasites and leading to death by asphyxiation. The process is much similar but not identical to another process called encapsulation as observed in insects and crustaeans. In encapsulation process blood cells or haemocytes form capsule whereas in nodule the capsule is formed by fibres of fibroblast cells.

iii) Cytotoxic reaction in invertebrates

Neither thymocytes or T-lymphocytes are present in invertebrate. However graft-implantation experiments indicate that some haemocytes of invertebrates have cytotoxic effect over foreign or non-self cells. This observation has been made in annelida and mollusca. Allograft rejection has been observed in starfish (echinodermata).

iv) Phagocytosis by insect haemocytes

Phgocytosis occurs in three steps-a) recognition of pathogen by receptor moleculs b) ingestion of pathogen and c) final disposal.

Opsonization in haemocytic phagocytosis enhance the recognition and phagocytosis process.

v) Encapsulation in insects and in crustacea

Granulocytes and coagulocytes act in the process of capsule fomation while

plasmatocytes play the important role in producing a capsule in insect. Granulocytes, the largegranule haematocytes of decapod crustaceans can encapsulate metazoan parasites and fungal hyphae.

vi) **Components of a complement system of vertebrate in echinoderm**

The alternative pathwy of complement with factors-like factor B, factor D are present in echinoderm.

A lectin pathway of complement activation is present in higher invertebrates like urochordates.

vii) The components of the mammation Toll-like receptor signalling pathway has similar functional pattern in ***Drosophila*** (Fig. 8.3).

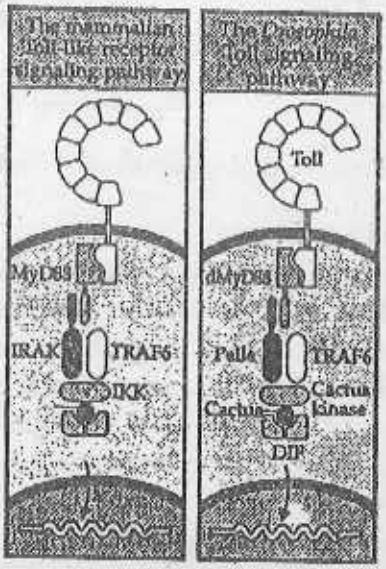


Fig. 8.3. The components of the mammalian toll-like receptors signalling pathway has similar functional pattern as in *Drosoptila,* different abbreviation stand for chemicals in pathways like My D88 or dMy D88 is an adaptor protein, IRAK, TRAFE, IKB, IKK and NFKB are components of signalling pathway.

The mid step in both signalling pathway occurs via the interaction of death domains, between MyD88 and IRAK is mammal and between d MyD88 and Pelk in *Drosophila.* Both IRAK and Pelle are serive xinase. Simlarly the end point reactions involving TRAF6, IKB, IKK in mammal and cactus and cactus kinase are homologous in nature (Adapted from Janeway, *et al,* 2005).

8.4 Naturally occuring humoral type defense process in invertebrate

There is no immunoglobulin in invertebrate, but the fluid in body possess a range of humoral defence factors. Agglutinins, lysozyme, lysin, non-lysozyme bacteriocidins, lysosomal enzymes and immobilization factor are few examples to

name. Components parallel to vertebrate complement system has also been reported. Phagocytes bear c3b-like receptors and can enhance phagocytosis. In insects many antibacterial proteins can be induced within few hours of antigen injection. One such factor, a ceropin, called P4 or haemolin has been obseved to have homology with certain immunoglobulin. Cytokine like factors are found in protozoan pheromore Er-1, similar to IL-2 in functional aspect. Similarly molecules like IFN-γ, IFN-β and TNF have also been isolated and characterised from annelids, echinoderms and tunicates.

8.5 Selective questions

1. Draw and describe the role of different cells in invertebrates involved in defence function.
2. Compare the toll like receptors of *Drosophila* with mammalian system.

8.6 Selected readings

Sima, P and Vetvicka, V. 1993. Evolution of immune reactions. Critical Review in immunology CRC Pross 13(2) : 83-114.

Janeway C., Travers. P, Walpart, M. and shlomchik, M.J. 2005 Immunobiology 6th edn., Churchill livingstone, London, 823 pp.

Unit 9 ❐ Epidemiology of Microbe Related Diseases

Structure

9.1 Introduction

Epidemiology is difined as the overall situation of the particular disease and related informations like nature, distribuion, cause transfer procedure, preventive, measure and management of the disease. It is also known as the "natural history of disease" or can be called "the human face of ecology".

Surveillance, a term closely linked with the study of epidemiology has three main elements—

1) Systemic collection of pertinent data

2) The orderly consolidation and evaluation of the data

3) The prompt dissemination of the findings, especially to those who can take appropriate action.

9.2 Factors of eipdemiology

The infection process is actually a dynamic state involving three main factors : the microorganism, the host and the environment.

The micro-organisms

The concept of **Virulence** is an important point to understand the epidemiological consequence of microbe. It is difined as the degree of pathogenicity of an infectious agent indicated by fatality rates and / or its ability to invade and damage the tissues of the host. The degree of virulence depends on **invasiveness**, the capacity of organism to spread widely through the body, and **toxigenicity** ,the toxin producing property of the organism.

A second variable is the **dosage** of the organism and this is clearly related to the virulence. A small number of organisms of high virulence is usually sufficient to cause disease in a susceptible person, whereas if the organism is of low verulence it often fails to cause disease.

Portal of entry is the third variable as many orgnisms have a predilection for a particulr tissue or organ. ***Salmonella typhi*** usually causes typhoid only when it enters the human body through the mouth in food or water.

The Host

Host reaction to a microorganism depends on the ability to resist infection. The individual may not possess sufficient resistance against a particular pathogen to prevent contraction of infection when exposed to the organism. Alternatively, the individual may possess specific protective antibodies or cellular immunity as a result of previous infection or immunization.

However, immunity is relative and may be overwhelmed by an excessive dose of the infectious agent or if the person is infected via an unusual portal of entry.

It may also be impaired by immunosupressive drug therapy, concurrent disease, or the aging process.

The three main factors of infection are shown in Fig. 9.1.

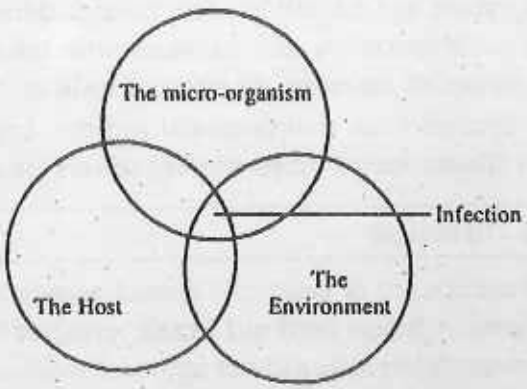


Fig. 9.1. The 3 main factors involved in the infectious process.

The Environment

The environment plays a major role in the occurrence, spread and control of infection. The virtual dissappearance of relapsing fever, plague and cholera, and rarity of indigenous typhoid fever and the relative infrequency of tuberculosis and bacillary dysentery from UK and USA are all indications of improvements which have taken place in environmental conditions. The decrease in overcrowding, and infestation together with the demand for cleaner water supplies, and better sanitation have been of paramount importance in producing these dramatic improvements.

9.3 The spread of infection

The infection spreads in an epidemiological pattern.

These are—

i) It spreads directly from one person to another. Clinical symptoms are easily detectatable and healthy carriers are not a feature. Ex.-Measles.

ii) The infection spreads through healthy carrier. Typhoid, paratyphoid , diptheria are examples where healthy individuals may harbour the bacilli responsile for such disease.

iii) Infections in which persons harbour the organism before the onset of clinical illness.

Streptococcus pneumonie may not cause any harm until an event such as skull fracture allows the transfer of the bacterian from middle ear to cerebrospinal space where it can potentially cause meningitis.

iv) Infection is derived from animal sources as for example leptospirosis, Q fever, anthrax, rabies and brucellosis are diseases derived from animals by zoonosis. These are spread by direct contact with the animal concerned or indirectly by means of ingestion of infected milk or bone products.

v) Infections are derived from environmental sources. Legionnaires disease is an example of illness spread from cooling towers or airconditioners.

9.4 Outbreaks of infection

It is defined as the occurrence of cases of a disease associated in time or locaion among a group of persons. A **house hold out break** involves two or more persons resident in the same private household and not apparently connected with any other case of outbreak, A general outbreak involves 2 or more persons who are not confined to one private household.

9.4.1 Patterns of outbreak

There are 3 main patterns of outbreak—

(i) **The explosive outbreak** : It is characterized by occurrence of a large proposition of cases in a relatively short period of time; there is a sharp rice and fall in the number of infected persons. It is also known as **common source** or **point source** outbreak. This pattern is often associated with food & water contamination (Fig. 9.2)

(ii) **Person to person spread** : Outbreaks caused by infections which are spread from person to person have a more protected course taking longer than explosive outbreaks to build up & to subside.

Diseases such as dysentery, hepatitis type A and gastro-enteritis which are usually spread by the faecal-oral route often follow the pattern of spread. (Fig. 9.3).

(iii) **Explosive outbreaks with subsequent person-to-person spread** : This pattern is often apparent when there is contamination of a common water or food source and the initial cases subsequently infect their contacts. Thus, the pattern of the outbreak in a combination of that process with an explosive outbreak, but followed by a slower decline. (Fig. 9.4)

9.4.2 Analysis of outbreak

The fundamental pieces of information which should be sought whenever an outbreak occurs are as follows :

(1) WHO gets infected?
(2) WHERE were those became infected?
(3) WHEN did the infection occur?
(4) WHAT was the common factor?
(5) HOW did those involved become infected?
(6) WHY did the infection occur?

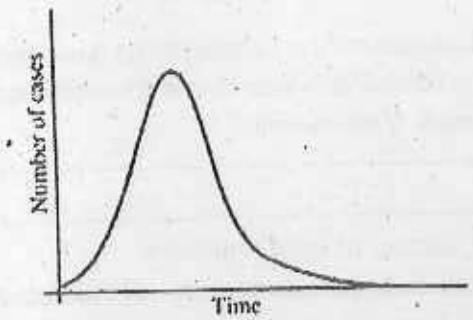


Fig. 9.2. Epidemic curve apparent when there is an explosive outbreak.

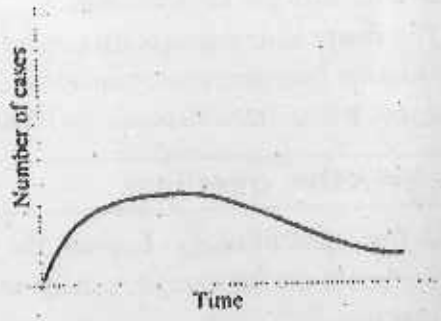


Fig. 9.3. Epidemic curve apparent when there is person-to-person spread of infection.

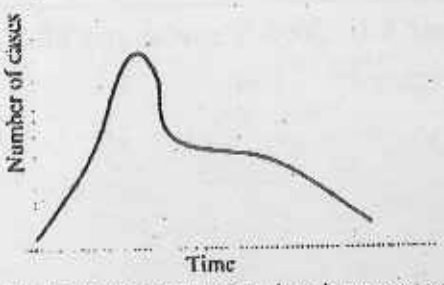


Fig. 9.4. Epidemic curve apparent when there is person-to person spread subsequent to a common source outbreak.

9.4.3 Investigation of outbreaks

In the investigation of outbreaks it is important to have a standardized approach to the various steps involved. Such an approach might have the following as a basis :

1. Verify the diagnosis.
2. Establish the existence of an outbreak.
3. Establish the extent of an outbreak.
4. Identify common charecteristics of experiences of the affected persons.
5. Investigate the source and vehicle of infection.
6. Analyse the findings.
7. Construct an hypothesis.

9.4.4 Control of outbreaks

The investigation of an outbreak should be carried out as swiftly as possible so that adequate control measures can be started without delay. Knowledge of the **source of infection** and **route of transmission** and the **persons at risk** should allow appropriate acion to be taken in order to achieve success.

The **source of infection** may be—(1) Human cases or carriers. (ii) Animal cases or carriers, (iii) the environment.

The **route of transmission** can be—(1) Direct or indirect contact, (ii) Air-borne transmission,(iii) percutaneous transmission (iv) Food & water-borne transmission, (v) Insect-borne transmission, (vi) Transplacental transmission.

9.5 Selective questions

1. Define epidemiology. Explain the role of factors of epidemiology.
2. Comment on the spread, pattern of outbreak and control measures of microbial infection.

9.6 Selected readings

Abbas, A.K. and Lichtman, A.H. 2003. Cellular and Molecular Immunology. 5th edn. Elsevier Science, USA.

Unit 10 ❐ Host Microbe Interaction–Immune Response to Protozoa, Bacteria and Virus

Structure

10.1 Introduction

The host microbe interaction is an integrated mechanism where a series of co-ordinated events must be overcome by pathogen to establish the infection in a susceptible host. The barriers to be overcome are—

i) Epithelial surface of skin, respriratory surface, gut.

ii) Mucous membrane of gut.

iii) lower pH value of stomach and upper intestine

iv) gastric enzymes.

v) antibacterial peptides.

On overcoming these physical barriers the pathogen face complement system and NK cells which are members of innate immunity along with its most important effector cell the phagocyte.

Even if this system is breached the pathogen has to deal with humoral and cell mediated immune system (acquired immunity) of the host. Potential outcomes of the interaction between a host and a microbe has been shown in a diagrammatic way below (Table 10.1)

10.2 Viral infections

Specific immune effector mechanism and nonspecific defence mechanism act in coordination to eliminate an infectiong virus. Virus, on the other hand, try to overcome one a more of these mechanisms & establish infection. The final outcome of host-virus, interaction depends on the effectivity of the hosts defensive mechanism and resistance offered by the virus. (Fig. 10.1 and 10.2)

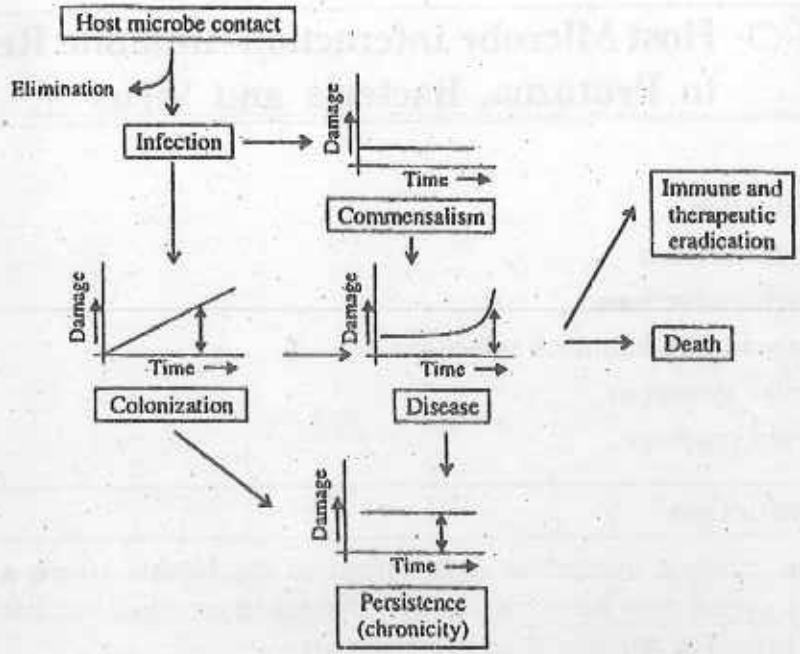


Table 10.1 Model of host microbe contact (Adapted from Casadevall and pirofski, 2000)

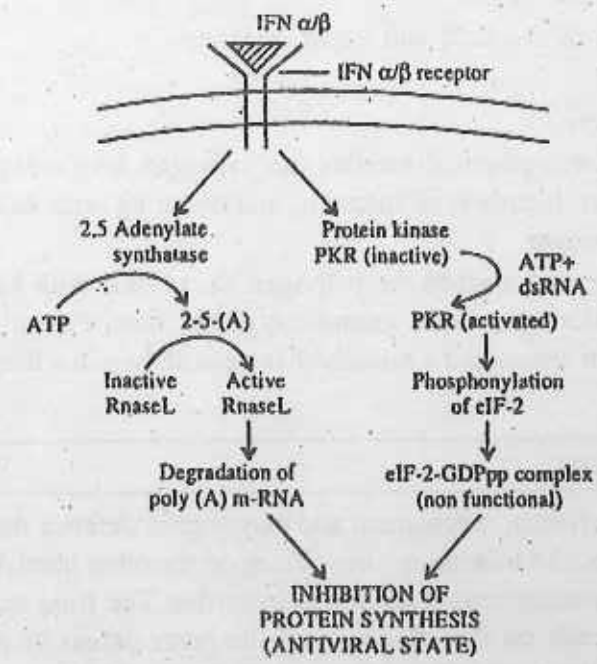


Fig.10.1 Induction of antiviral state by interferon α/β (IFNα/β)

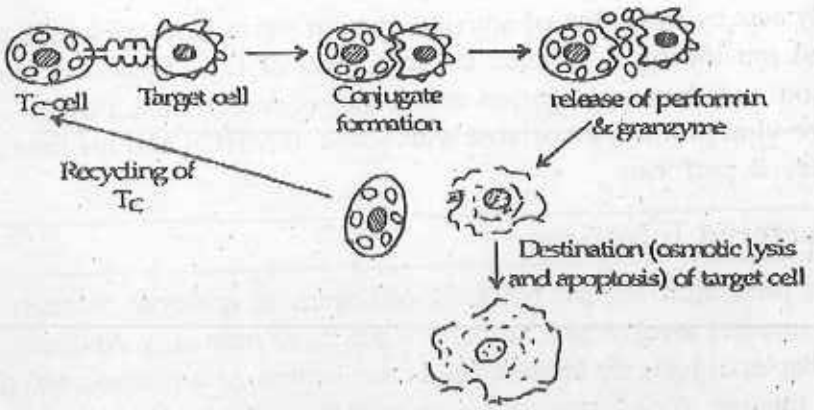


Fig. 10.2 Tc cell mediated killing of target cell.

Innate responses against viral infection is mediated by induction of type I interferous and by NK cells. Viral infection induce infected cells to produce and secrete IFN-α & IFN-β. On binding of IFN-α / IFN-β to their respective receptors on non-infected surrounding host cells, a state, better known as "antiviral state" occurs. The receptors of IFN-α / IFN-β transduce this signal to the cell interior via JAK/STAT signalling pathway. This in turn induces the expression of certain genes.

One of these genes code for an enzyme 2'-5' Adenylate synthetase, which in turn activates a ribonuclease, RNase L, that degrades viral DsRNA. Another function of IFN-α / IFN-β receptor is activation of a specific protein Kinase, PKR, which inactivates el F-2 a translation initiation factor. Viral protein synthesis is blocked and hence no new virus particles can be produced.

On binding of IFN-α/IFN-β to their specific receptor on NK cells, the NK cells are activated & they can effectively lyse & kill virally infected host cells.

Virus enter host cell by certain surface molecules which at as receptors for host cell surface molecules. Antibodies are particularly effective in protection against viral infection if they are localized at the site of entry of the virus into the body. Blocking of viral receptors renders them incapable of entering host cells and makes them easy target for phagocytic killing. The antibodies also acts as opsonins and help in viral clearance.

However, when infection has occured, the viral DNA is generally integrated in Host DNA and under such conditions, the antibodies become helpless. It is at this point that CD4+ TH1 cell & CD8+ Tc cells act in unison to alleviate viral infection. T_H1 cells secrete IL-2, IFNγ & TNF as defence against infection.

IFNγ acts by induction of antiviral state in surroundin cells cells by method as described previously. IL-2 helps in recruitment of CTL precesors into an effector population and their aggregation at the site of infection. CTLs or Tc cells then recognize viral proteins associated with host cell MHC-I and mediate its killing by granofmes & perforins.

10.3 Bacterial infections

Host protection against bacterial pathogens is achieved through a variety of mechanisms that involve both humoral and cellular immunity. Antibacterial defenses include bacterial lysis via antibody and complement, opsonization and phagocytosis, with elimination of phagocytosis bacteria by the liver, spleen and other components of the reticulo endothetial system.

Bacterial pathogens can be distinguished as intracellulor or extracellular pathogens. The defense mechanism of host vary along with the abode of the bacterial pathogen.

However antibodies to play an important role in negeting the toxins--endotoxin (bacterial cell wall component) & exotoxin (secretory product). By acting against exotoxins the anibodies act to reduce the pahtogenecity and damage caused by bacteria as most exotoxins interfere with normal physiological mechanisms of host cell. When targetted against endotoxins, the antibodies act as opsonins and help in their phagocytonis by NK cells and macrophages.

The complex peptidoglycan coat of gram positive bacteria contain teichoic and molecules which are highly immunogenic and form their major antigenic determinant. However their presence prevent lysis by complement activation. Gram negative bacteria on the other hand are susceptible to lysis by complement activation. The lipopolysacchride coat (LPS) act as immunogen as well as inducers to lectin binding pathway of complement activation.

Bacteria residing inside host cell, however escape the above described process of destruction/elimination. They can be eliminated by host, only if T_H & Tc acts in unision, i.e. by cellular immunity. MHC proteins along with viral anigens, activate T_H cells which release lytokines thereby recruiting phagocytic as well as Tc cells as effectors of clearance.

Phagocytic cells kill by injection, & then by acting upon endosomes with lysosomal enzymes. Tc cells utilize perform & granzyme pathway. The performin secreted by Tc cells attach to the membrane of the infected cell and form channel through which water can enter the injected cell and cause osmotic lysis. "Granzymes" a set of DNA degrading enzymes secreted by Tc cells also enter infected host cell

via peforin channel. They degrde the host cell DNA into fragments and thus induce regulated cell death.

10.4 Protozoan and helminth infection

As in the case of bacterial infection the protozoal infection may be divided into two blood groups–intracellular infection and extra cellular type.

In case of extracellular protozoans, the innate immunity mediated by complement system (Lectin and alternate pathway) and NK cells form the first line of defence. the specific antibodies against protozoan forms the humoral branch of immunity. They reader protection by activating classical complement pathway as well ad by acting as opsonius and phagocytic function by macrophage.

In case of intracellular protozoan infection, cellular immunity mediated by TH and Tc cells play an important role along with the immune reaction as in case of bacteria. The TH cells by secreting cytokines attract and activate Tc, NK cells & macrophages. Macrophages ingest protozoa infected with cells and destroy them. Tc cells interact with host MHC coupled with protozoan antigen and lyse infected cells by secretion of perforia and granzymes. Helminth parasites are large and they are blood parasite or tissue parasite in host. Some kind of immunity to reinfection is the general rule but existing worms are not always destroyed. This property of host system is a state called concomitant immunity.

Adult or even larval worms are too large to be destroyed by antibody with or without complement or by phagocytic cells. The normal effector mechanism is antibody-dependent cell mediated cytotoxicity (ADCC) in which the worms become coated with IgG or IgE antibody which binds eosinophils and other cells that actually destroy the parasite. As in chronic protozoa infections the immune responses are multifactorial and vary according to stage of development of the invading worm. Helminths are all capable of evading immune responses of the host and contribute much to the pathology of the infection in host's body.

Parasites (Protozoa and Helminth) can evade the immune response of the host by various methods. They become intracellular and avoid antibody like *Leishmania* spp., *Toxoplasma gondii* and *Trypanosoma cruzi*. Parasites like african trypanosoma (*T. brucei*) and malarial parasites (*Plasmodium*), nematode *Trichinella* show antigenic variation of their surface glycoprotein antigen. The non-specific way of evading the immune response is by interfering with the operation and control of its delicate balance. The mechanisms are directed towards damaging the normal functioning of host immune mechanism by immunodepression, polyclonal B cell activation etc.

10.5 Selective questions

1. Explain the hypothesis or potential outcomes of the interaction between a host and a microbe as proposed by Casadevall and Pirofski.
2. Elaborate the immunity to (a) virus (b) bacteria (c) protozoa and helminth parasites.
3. Explain briefly the mechanisms by which pathgens evade the immune response.

10.6 Selected readings

1. Abbas, A.K. and Lichtman, A.H. 2003. Cellular and Molecular Immunology. 5th edn. Elsevier Science, USA.
2. Casadevall, A and Pirofski, L 2000 Host pathogen miteractions; basic concepts of microbial commensalism, colonization, infection and disease. Infect Immun 68 : 6511.