NETAJI SUBHAS OPEN UNIVERSITY

STUDY MATERIAL

POST GRADUATE
ZOOLOGY

Paper : 10
Group : B

Laboratory Course
(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 Subjects 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 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

Second Edition : August, 2015

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

Post Graduate Zoology
[M. Sc.]

PAPER : PGZO 10
GROUP : B

	Writer	Editor
Units 1-4 :	Dr. Kamales Kr. Misra	Prof. Buddhadeb Manna

Professor (Dr.) Debesh Roy
Registrar

Netaji Subhas
Open University

PGZO-10
Laboratory Course

Group
B
Quantitative Biology & Biotechnology

Unit 1 □ Correlation, regression, ANOVA

Structure

1.1 Paired Sample t-test

One of the most fundamental concepts in research is the concept of **correlation.** If two variables are correlated, this means that research worker can use information about one variable to predict the values of the other variable.

A paired sample t-test is used to determine whether there is a significant difference between the average values of the same measurement made fewer than two different conditions. Both measurements are made on each unit in a sample, and the test is based on the paired differences between these two values. The usual null hypothesis is that the difference in the mean values is zero. For example, the yield of two strains of barley is measured in successive years in twenty different plots of agricultural land (the units) to investigate whether one crop gives a significantly greater yield than the other, on average.

The null hypothesis for the paired sample t-test is

$H0: d = \mu 1 - \mu 2 = 0$ where, d is the mean value of the difference.

This null hypothesis is tested against one of the following alternative hypotheses, depending on the question posed:

$H1: d = 0$
$H1: d > 0$
$H1: d < 0$

The paired sample t-test is a more powerful alternative to a two-sample procedure, such as the two-sample t-test, but can only be used when we have matched samples.

1.2 Correlation Coefficient

A correlation coefficient is a number between -1 and 1, which measures the degree to which two variables are linearly related. If there is perfect linear relationship with positive slope between the two variables, we have a correlation coefficient of 1; if there is positive correlation, whenever one variable has a high (low) value, so does the other. If there is a perfect linear relationship with negative slope between the two variables, we have a correlation coefficient of -1; if there is negative correlation, whenever one variable has a high (low) value; the other has a low (high) value. A correlation coefficient of 0 means that there is no linear relationship between the variables.

The mathematical formula for computing r is:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

where n is the number of pairs of data.

The value of r is such that $-1 \le r \le +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

- *Positive correlation*: If x and y have a strong positive linear correlation, r is close to +1. An r-**value** of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increases, values for y also increase.
- *Negative correlation*: If x and y have a strong negative linear correlation, r is close to -1. An r-**value** of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.
- *No correlation*: If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables.
- Note that r is a dimensionless quantity; that is, it does not depend on the units employed.
- A *Perfect correlation* of ± 1 occurs only when the data points all lie exactly on a straight line. If $r = +1$, the slope of this line is positive. If $r = -1$, the slope of this line is negative.
- A correlation greater than 0.8 is generally described as *strong*, whereas a correlation less than 0.5 is generally described as *weak*. These values can vary based

upon the "type" of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

Correlation is a measure of association between two variables. The variables are not designated as dependent or independent. The two most popular correlation coefficients are: Spearman's correlation coefficient r and Pearson's product-moment correlation coefficient.

1.3 Spearman Rank Correlation Coefficient

The Spearman rank correlation coefficient is one example of a correlation coefficient. It is usually calculated on occasions when it is not convenient, economic, or even possible to give actual values to variables, but only to assign a rank order to instances of each variable. It may also be a better indicator that a relationship exists between two variables when the relationship is non-linear.

Commonly used procedures, based on the Pearson's Product Moment Correlation Coefficient, for making inferences about the population correlation coefficient make the implicit assumption that the two variables are jointly normally distributed. When this assumption is not justified, a non-parametric measure such as the Spearman Rank Correlation Coefficient might be more appropriate.

1.4 Pearson's Product Moment Correlation Coefficient

Pearson's product moment correlation coefficient, usually denoted by r, is one example of a correlation coefficient. It is a measure of the linear association between two variables that have been measured on interval or ratio scales, such as the relationship between height in inches and weight in pounds. However, it can be misleadingly small when there is a relationship between the variables but it is a non-linear one.

There are procedures, based on r, for making inferences about the population correlation coefficient. However, these make the implicit assumption that the two variables are jointly normally distributed. When this assumption is not justified, a non-parametric measure such as the Spearman Rank Correlation Coefficient might be more appropriate.

Worked examples :

1. The length and weight of 7 groups of lizards of a species are given below:

Length in cm	11.7	13.9	15.5	17.8	18.5	19.2	21.0
Weight in g	7.10	12.42	15.35	23.20	28.45	32.25	39.84

Serial No	Length (X)	Weight (Y)	x	y	x^2	y^2	xy
1	11.7	7.10	-5.0	-15.55	26.01	241.8	79.3
2	13.9	12.42	-2.9	-10.23	8.41	104.6	23.36
3	15.5	15.35	-1.3	-7.3	1.69	53.2	9.49
4	17.8	23.20	+1	+0.55	1.0	0.30	0.55
5	18.5	28.45	+1.7	+5.8	2.89	33.64	9.86
6	19.2	32.25	+3	+9.6	9.0	92.16	28.8
7	21	39.84	+4.2	+17.19	17.64	295.49	72.2
N=7	$\sum X=117.6$	$\sum Y=158.6$			$\sum x^2$ $=66.64$	$\sum y^2$ $=821.9$	$\sum xy$ $=223.56$

Sum up of all values i.e. $\sum x^2$, $\sum y^2$, $\sum x.y$ and then put value in the formula to obtain '*r*'

$$\overline{X} = \frac{117.6}{7} = 16.8;\ \overline{Y} = \frac{158.6}{7} = 22.65$$

Using the above formula the result is calculated as -

$$r = \frac{\sum x.y}{\sqrt{\sum x^2.\sum y^2}} = \frac{223.56}{\sqrt{66.64 \times 821.19}} = \frac{223.56}{\sqrt{54724.1}} = \frac{223.56}{233.93} = 0.96.$$

Conclusion : There is a strong positive correlation between the length and weight of body of the lizard species. Calculated value of *r* is very high; therefore, both variables are highly correlated.

1. Analysis of another example

Correlation between reading and spelling of following data using computational formula

Student	Reading (X)	Spelling (Y)	X^2	Y^2	XY
1	3	11	9	121	33
2	7	1	49	1	7
3	2	19	4	361	38
4	9	5	81	25	45
5	8	17	64	289	136
6	4	3	16	9	12
7	1	15	1	225	15
8	10	9	100	81	90
9	6	15	36	225	90
10	5	8	25	64	40
Sum	55	103	385	1401	506

If we plug each of these sums into the raw score formula we can calculate the correlation coefficient.

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{(10)(506) - (55)(103)}{\sqrt{(10)(385) - (55)^2}\sqrt{(10)(1401) - (103)^2}}$$

$$= \frac{5060 - 5665}{\sqrt{3850 - 3025}\sqrt{14010 - 10609}} = \frac{-605}{\sqrt{825}\sqrt{3401}}$$

$$\frac{-605}{(28.723)(58.318)} = \frac{-605}{1675.0679} = -0.36$$

Conclusion: The correlation obtained is -.36, showing that there is a small negative correlation between reading and spelling. The correlation coefficient is a number that can range from -1 (perfect negative correlation) through 0 (no correlation) to 1 (perfect positive correlation).

Exercise :

1. Marks of 10 students in Zoology and Statistics are given below:

Zoology (X):	32	38	48	43	40	22	41	69	35	64
Statistics (Y):	30	31	38	43	33	11	27	76	40	59

Calculate product-moment correlation coefficient and interpret the result.

2. Number of ponds (X) in five villages and number of fishes (Y) in the pord are as follows:

X :	17	17	18	19	19	20	21	22	23
Y :	230	210	290	230	330	320	360	340	320

Find out the rank correlation and interpret the result.

1.5 Regression Equation

A regression equation allows us to express the relationship between two (or more) variables algebraically. It indicates the nature of the relationship between two (or more) variables. In particular, it indicates the extent to which you can predict some variables by knowing others, or the extent to which some are associated with others.

A linear regression equation is usually written

$$Y = a + bX + e$$

Where,

Y is the dependent variable
a is the intercept
b is the slope or regression coefficient
X is the independent variable (or covariate)
e is the error term

The equation will specify the average magnitude of the expected change in Y given a change in X. The regression equation is often represented on a scatter plot by a regression line. The constant 'a' and 'b' can be obtained by the following formula:

$$a = \overline{Y} - b.\overline{X}$$

$$b = \frac{\sum x.y}{\sum x^2} = \frac{\sum X.Y - \frac{\sum X.\sum Y}{N}}{\sum X^2 - \frac{(\sum X)^2}{N}} \quad \text{where, } x = X - \overline{X}\ ; y = Y - \overline{Y}.$$

Procedure of the test

1. Plot a graph between two variables taking independent variable on X-axis and dependent variable on Y-axis. Find out the values of 'a' and 'b'.

   For drawing the line of best fit (regression line) find out any two values of *y* associated with corresponding *x* by using the equation y = a + bx.

2. Plot these two obtained values on the graph.
3. Make a straight line intersecting through these two points to get regression line.

   A linear regression line has an equation of the form ***Y* = *a* + *bX***, where ***X*** is the explanatory variable and ***Y*** is the dependent variable. The slope of the line is ***b***, and ***a*** is the intercept (the value of *y* when *x* = 0).

Worked examples

In an experiment data, recorded on two parameters such as length of a species of fish and number of ova per fish, is given as follows. Obtain the two regression equations.

Length of fish (X)	18	25	25	32	35	20	30	13	30	30	37	40	20	25	27	40	15	23	35	23
No. of ova (Y)	20	25	33	35	40	26	30	15	25	37	43	42	23	28	33	45	20	20	33	30

$$\sum X = 543 \text{ and } \sum Y = 603, \sum X^2 = 15923, \sum Y^2 = 19563$$

$$\sum X.Y = 17532, \ \overline{X} = 27.15, \overline{Y} = 30{,}15, r = +0.908$$

$$\text{Solution} \quad b = \frac{\sum x.y}{\sum x^2} = \frac{\sum X.Y - \frac{\sum X.\sum Y}{N}}{\sum X^2 - \frac{(\sum X)^2}{N}} = \frac{17572 - \frac{543 \times 603}{20}}{15923 - \frac{(543)^2}{20}} = \frac{1160.55}{1180.55} = 0.98$$

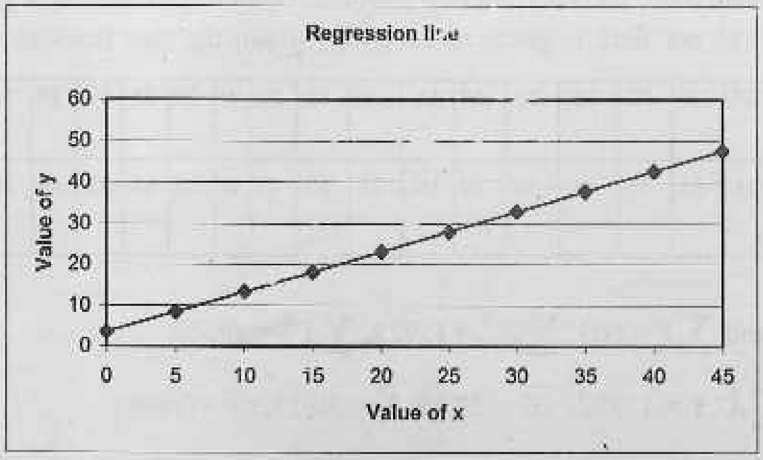
$$a = \overline{Y} - b.\overline{X} = 30.15 - 0.98 \times 27.15 = 30.15 - 26.61 = 3.54$$

$$Yx = a + bx = 3.54 + 0.98x$$

Estimate the value of y after putting the x values in the equation ($Yx = 3.54 + 0.98x$)

$x = 0 \quad Yx = 3.54$
$x = 5 \quad Yx = 8.44$
$x = 10 \quad Yx = 13.44$
$x = 15 \quad Yx = 18.24$
$x = 20 \quad Yx = 23.14$
$x = 25 \quad Yx = 28.04$
$x = 30 \quad Yx = 32.94$
$x = 35 \quad Yx = 37.84$
$x = 40 \quad Yx = 42.74$
$x = 45 \quad Yx = 47.64$

If these values are plotted on a graph, there will be a straight line, which is called the estimated regression line.



1.6 Difference between correlation and linear regression

- Correlation and linear regression are not the same. Correlation quantifies the degree to which two variables are related. Correlation does not find a best-fit line (that is regression). We simply are computing a correlation coefficient (r) that tells us how much one variable tends to change when the other one does.
- With correlation you don't have to think about cause and effect. You simply quantify ow well two variables relate to each other. With regression, us do have to think about cause and effect, as the regression line is determined as the best way to predict Y from X.

- With correlation, it doesn't matter which of the two variables we call "X" and which we call "Y". We'll get the same correlation coefficient if we swap the two. With linear regression, the decision of which variable we call "X" and which we call "Y" matters a lot, as we'll get a different best-fit line if we swap the two. The line that best predicts Y from X is not the same as the line that predicts X from Y.
- Correlation is almost always used when we measure both variables. It rarely is appropriate when one variable is something we experimentally manipulate. With linear regression, the X variable is often something we experimentally manipulate (time, concentration...) and the Y variable is something we measure.

1.7 Analysis of variance (ANOVA)

In statistics, **analysis of variance** (ANOVA) is a collection of statistical models, and their associated procedures, in which the observed variance is partitioned into components due to different explanatory variables. In its simplest form ANOVA gives a statistical test of whether the means of several groups are all equal, and therefore generalizes Student's two-sample *t*-test to more than two groups.

One-way analysis of variance (ANOVA) tests allow us to determine if one given factor, such as drug treatment, has a significant effect on gene expression behavior across *any* of the groups under study. A significant p-value resulting from a 1-way ANOVA test would indicate that a gene is differentially expressed in at least one of the groups analyzed. If there are more than two groups being analyzed, however, the 1-way ANOVA does not specifically indicate which pair of groups exhibits statistical differences. Post Hoc tests can be applied in this specific situation to determine which specific pair/pairs are differentially expressed. A One-Way Analysis of Variance is a way to test the equality of three or more means at one time by using variances.

$$SS(T) = \sum \left(x - \bar{X}_{GM}\right)^2 \qquad SS(B) = \sum n\left(\bar{x} - \bar{X}_{GM}\right)^2$$

Assumptions

- The populations from which the samples were obtained must be normally or approximately normally distributed.
- The samples must be independent.
- The variances of the populations must be equal.

Hypotheses

The null hypothesis will be that all population means are equal; the alternative hypothesis is that at least one mean is different. In the following, lower case letters apply to the individual samples and capital letters apply to the entire set collectively. That is, n is one of many sample sizes, but N is the total sample size.

Problem: Susan Sound predicts that students will learn most effectively with a constant background sound, as opposed to an unpredictable sound or no sound at all. She randomly divides twenty-four students into three groups of eight. All students study a passage of text for 30 minutes. Those in group 1 study with background sound at a constant volume in the background. Those in group 2 studies with noise that changes volume periodically. Those in group 3 studies with no sound at all. After studying, all students take a 10-point multiple-choice test over the material. Their scores follow:

group	test scores							
1) constant sound	7	4	6	8	6	6	2	9
2) random sound	5	5	3	4	4	7	2	2
3) no sound	2	4	7	1	2	1	5	5

x_1	x_1^2	x_2	x_2^2	x_3	x_3^2
7	49	5	25	2	4
4	16	5	25	4	16
6	36	3	9	7	49
8	64	4	16	1	1
6	36	4	16	2	4
6	36	7	49	1	1
2	4	2	4	5	25
9	81	2	4	5	25
$\Sigma x_1 = 48$ $(\Sigma x_1)^2 = 2304$	$\Sigma x_1^2 = 322$	$\Sigma x_2 = 32$ $(\Sigma x_2)^2 = 1024$	$\Sigma x_2^2 = 148$	$\Sigma x_3 = 27$ $(\Sigma x_3)^2 = 729$	$\Sigma x_3^2 = 125$
$M_1 = 6$		$M_2 = 4$		$M_3 = 3.375$	

$$SS_{total} = (322 + 148 + 125) - \frac{(48 + 32 + 27)^2}{24}$$

$= 595 - 477.04$ **or** $SS_{total} = 117.96$

$$\bar{X}_{GM} = \frac{\sum n\bar{x}}{\sum n}$$

$$SS_{among} = \left[\frac{2304}{8} + \frac{1024}{8} + \frac{729}{8}\right] - 477.04$$

$$= 507.13 - 477.04$$

$$SSamong = 30.08$$

$$\bar{X}_{GM} = \frac{\sum x}{N}$$

$$SSwithin = 117.96 - 30.08 = 87.88$$

Source	SS	df	MS	F
Among	30.08	2	15.04	3.59
Within	87.88	21	4.18	

*(according to the F sig/probability table with df = (2,21) F must be at least 3.4668 to reach $p < .05$, so F score **is** statistically significant)

Interpretation: Susan conclude that her hypothesis **may** be supported. The means are as she predicted, in that the constant music group has the highest score. However, the significant *F* only indicates that at least two means are significantly different from one another, but she can't know which specific mean pairs significantly differ until she conducts a post-hoc analysis.

Between Group Variation

The variation due to the interaction between the samples is denoted SS(B) for Sum of Squares Between groups. If the sample means are close to each other (and therefore the Grand Mean) this will be small. There are k samples involved with one data value for each sample (the sample mean), so there are k-1 degrees of freedom.

The variance due to the interaction between the samples is denoted by MS(B) for Mean Square Between groups. This is the between group variation divided by its degrees of freedom. It is also denoted by s_b^2

Within Group Variation

The variation due to differences within individual samples, denoted by SS(W) for

Sum of Squares Within groups. Each sample is considered independently, no interaction between samples is involved. The degree of freedom is equal to the sum of the individual degrees of freedom for each sample. Since each sample has degrees of freedom equal to one less than their sample sizes, and there are k samples, the total degrees of freedom is k less than the total sample size: df = N - k.

The variance due to the differences within individual samples is denoted by MS(W) for Mean Square Within groups. This is the within group variation divided by its degrees of freedom. It is also denoted by s_w^2

It is the weighted average of the variances (weighted with the degrees of freedom).

Unit 2 □ Gel electrophoresis of serum protein

Structure

2.1 **Background and Purpose**
2.2 **Sample Preparation**
2.3 **Electrophoresis**
2.4 **Tatal protein staining of electrophorefic gels**

2.1 Background and Purpose

Gel electrophoresis is a useful method to separate and/or identify proteins and nucleic acids. In SDS-polyacrylamide gel electrophoresis (SDS-PAGE), proteins are separated largely on the basis of polypeptide length, and so their molecular weight can also be estimated. SDS does however denature the protein, so activity stains cannot be used to identify particular enzymes. Described below is the protocol for preparing and using Laemmli discontinuous gels. In this system, two sequential gels are actually used; the top gel, called the stacking gel, is slightly acidic (pH 6.8) and has a low (5.5%) acrylamide concentration to make a porous gel. Under these conditions proteins separate poorly but form thin, sharply defined bands separation is better. The lower gel, called the separating, or resolving gel, is more basic (pH 8.8), and has a higher polyacrylamide content (normally, 12%), which causes the gel to have narrower channels or pores. As a protein, concentrated into sharp bands by the stacking gel, travels through the separating gel, the narrower pores have a sieving effect, allowing smaller proteins to travel more easily and hence rapidly, than larger proteins.

SDS-Polyacrylamide Gel Electrophoresis Method

Standard SDS-polyacrylamide gel electrophoresis (Laemmli)—gel preparation. Volumes given are sufficient for small (8 cm X 10 cm X 1.5 mm) gel format (10 ml of monomer). Scale up volumes as needed.

1. Pour the Separating gel

Set up your gel apparatus, prepare separating gel monomer. Add TEMED just prior to pouring gel (I "pour" the gels using a Pasteur pipet and a rubber bulb). Allow to polymerize before adding stacking gel by overlaying gently with water or n-butanol. With higher % gels, one can immediately pour the stacking gel on the unpolymerized separating gel. Be careful not to mix the two layers.

Separating Gels, in 0.375 M Tris, pH 8.8

7%, 10%, 12%, 15%

distilled H_2O 5.1 ml, 4.1 ml, 3.4 ml, 2.4 ml

1.5 M Tris-HCl, pH 8.8 - 2.5 ml

20% (w/v) SDS, 0.05 ml

Acrylamide/Bis-acrylamide (30%/0.8% w/v) 2.3 ml, 3.3 ml, 4.0 ml, 5.0 ml

10% (w/v) APS ammonium persulfate 0.05 ml

TEMED 0.005 ml

Total 10.005 ml monomer

2. *Pour the Stacking gel*

After the separating gel has polymerized, decant the overlay, prepare the stacking monomer, add the TEMED, and pour. Insert the comb and allow to polymerize completely before running.

Stacking Gels, 4.0% gel, 0.125 M Tris, pH 6.8

distilled H_2O 3.075 ml

0.5 M Tris-HCl, pH 6.8 - 1.25 ml

20% (w/v) SDS 0.025 ml

Acrylamide/Bis-acrylamide (30%/0.8% w/v) 0.67 ml

10% (w/v) ammonium persulfate 0.025 ml

TEMED 0.005 ml

Total Stack monomer 5.05 ml

For best results:

(i) Make ammonium persulfate solution fresh daily.

(ii) Degas solutions before adding TEMED for 15 min at room temperature.

3. *Running the gel*

Usually gels are run at constant current, 25-50 mA, depending on gel size. Here's the recipe for 5X SDS-PAGE running buffer. Dilute to 1X before use.

5X Running Buffer, pH 8.3 (1 liter)

Tris Base 15 g

Glycine 72 g

SDS 5 g

distilled water to 1 liter

Store at room temperature until use.

4. Laemmli Sample buffer

Dilute samples at least 1:4 with sample buffer, heat at 95°C for 4 minutes prior to loading.

Sample Buffer (8 ml)

Distilled water 4.0 ml

0.5 M Tris-HCl 1.0 ml

Glycerol 0.8 ml

10% SDS 1.6 ml

beta-mercaptoethanol 0.4 ml

0.05% (w/v) bromophenol blue 0.2 ml

2.2 Sample Preparation

Serum sample preparation

Blood serum should be collected from *vena cava* of laboratory white rat and allowed to clot for 2 hours at room temperature. The clotted material was removed by centrifugation at 3000 rpm for 15 min. Hemolytic material was not observed. The sera obtained from the blood samples were frozen immediately without any further treatment in liquid nitrogen and stored at -80°C until further analysis. [The protein concentration of serum was determined with the Bradford protein assay (Bio-Rad Protein Assay Dye Reagent Concentrate, Bio-Rad), using bovine gamma globulin as the standard.] The protein concentration ranged from 80 to 90 μg/μl for wild type mouse serum samples.

For best results, all samples were kept in identical, low ionic strength buffers.

1. Mix 50 μL of each sample with an equal volume of one of the denaturing buffers below.

2. Heat in a boiling water bath for one minute. In most cases, brief boiling (1 -2 min) improves denaturation, but it may also cause the protein to precipitate.

Denaturing Buffers (DB)

Components	DB I .	DB II
Tris.HCl	0.25 M	0.0125 M
SDS	2% (w/v)	2% (w/v)
β-Mercaptoethanol	2% (v/v)	5% (v/v)
Urea	8 M.	0
Glycerol	0	20%(v/v)
Bromophenol blue	0.001% (w/v)	0.001% (w/v)
pH	6.2	6.8

2.3 Electrophoresis

1. Remove the comb and clamp the gel to the electrophoretic apparatus.
2. Fill the top electrolyte compartment with running buffer.
3. Check for leaks from the top into the bottom compartment. If there are no leaks, fill the bottom compartment.
4. With a plastic Pasteur pipette, thoroughly rinse each well in the stacking gel with running buffer.
5. Apply the sample by using a micropipette to carefully add up to ~25 μL of protein in DB1 or DB2 to the bottom of a well. The volume and protein concentration of the sample should be sufficient to give at least 10 μg of each protein. If possible, avoid using the end wells.
6. Apply 15 μL of the molecular weight standards to one or two wells, preferably in an asymmetric position, to allow the front and back of the gel to be identified later.
7. Carefully record the contents of each well.
8. Replace the cover of the electrophoretic cell, with the (+) symbol on the cover connected to the (+) on the cell, so that the anode (+) is the bottom electrode.
9. Check the electrical connections on the cell to ensure that solution is not in contact with either banana plug, and connect the anode to the (+) terminal on the power supply, and the cathode to the negative terminal. (Notice the convention inversion for electrodes: + is the anode, and - the cathode).
10. Apply 15 mA/gel until the proteins are well into the stacking gel, then 35 mA/gel until the tracking dye reaches the bottom of the gel (about 45 minutes in this system).

11. Always turn down the power and unplug the wires from the power supply before removing the cover.

It is often useful to apply different sample volumes to several wells, so at least one lane has bands that are detectable but not overloaded.

10X Running Buffer (Laemmli electrolyte buffer)

Components	Concentration	g/1L
Glycine	1.92 M	144
Tris base	0.25 M	36.3
SDS	1%	10
Dilute 10-fold before use. Replace if the final pH is not within 0.1 pH units of pH 8.3.		

2.4 Total protein staining of electrophoretic gels

Gel staining can modify the electrophoretic properties of the proteins and so may interfere with protein transfer during Western blotting. For this reason, it generally is not advisable to stain a gel that is to be used for a Western blot. It is useful, however, to stain the gel after performing a Western blot, to ensure that the protein has successfully transferred from the gel to the membrane. [Once you've performed the Western blot, follow the protocol below to do the total protein staining of your gel].

Protein Staining Solutions

Staining Solution - Dissolve 20 mg of CPTS in 1 L of 6 mM HCl. This solution is stable at room temperature.

Wash Solution - 6 mM HCl in 20% (v/v) methanol (0.5 mL conc. HCl in 799.5 mL deionized water, 200 mL methanol). This solution is stable forever at room temperature.

Protein Staining Procedure

1. Following Western blotting, drain excess buffer from the gel and rinse in wash solution to remove SDS and fix (immobilize) the proteins.
2. Rock the gel in the wash solution for 15 minutes, then remove and discard the solution.

3. Add enough staining solution to cover the gel. Stain for an hour, or until adequately stained.

4. Remove the staining solution and replace with 100 mL wash solution and 0.1 g DEAE-cellulose.

5. Swirl the wash over the gel by rocking the (covered) container for several minutes/hours, or until excess stain is removed and unstained areas are completely clear.

6. Photograph, interpret visually, or quantitate using appropriate densometric equipment.

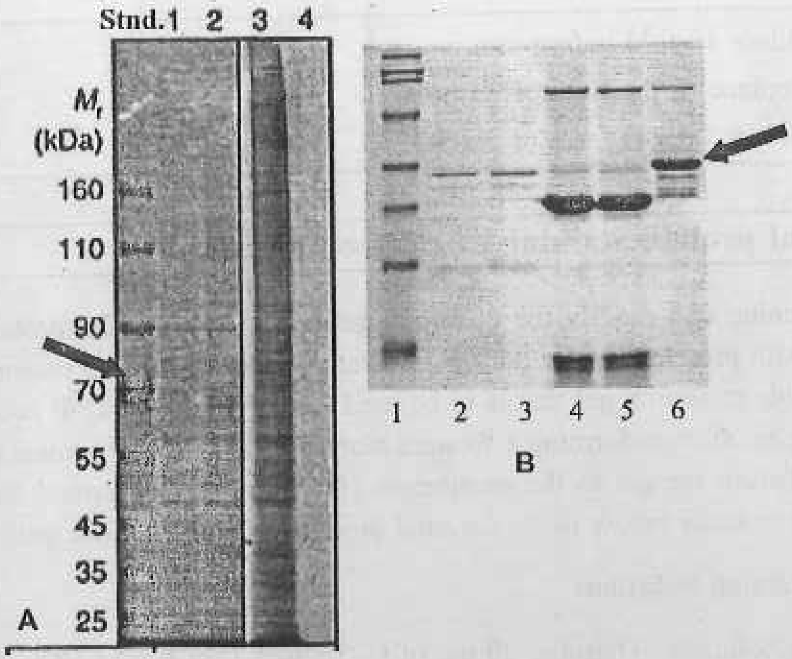


Fig. 1 : Two typical photographs of SDS-PAGE. Electrophoretic bands (dark black transverse lines, **arrows**) of samples are always compared with standard bands of known molecular weight as shown in figure A. In figure A, lanes 1, 2, 3 and 4 are showing banding pattern of unknown samples and first lane (Stnd = standard) shows the distribution of bands of known molecular weight proteins. The molecular weights (kDa) of the standard samples are given beside each band. The molecular weight of the different fractions of the unknown samples can be easily calculated in reference to the standard molecular weight. In figure B, lane 1 shows bands of standard proteins, lanes 2, 3, 4, 5 and 6 are experimental protein samples. Comparison is made between the bands of the standard and new bands of the samples (not present in the standard) characterizes the unknown protein sample. The molecular weight of the unknown protein sample can be deduced.

Unit 3 ❑ Analysis and interpretation of Southern, Northern and Western blotting from gel photograph

Structure

3.1 Introduction

3.2 Northern Blot

3.3 Southern Blot

3.4 Western Blot

3.5 Comparison of Blotting Methods

3.1 Introduction

- A southern blot is a method used to detect specific DNA sequences in complex DNA samples.

– It is a combination of several molecular biology techniques:

  - Restriction enzyme analysis
  - Agarose gel electrophoresis
  - Hybridization analysis

– After electrophoresis, DNA molecules are transferred from the agarose gel onto a filter membrane for probe hybridization.

  - A northern blot is almost identical to a Southern blot, but it involves the detection of RNA instead of DNA.

3.2 Northern Blot

Practice and analysis of northern blot

1. This is used to detect and identify RNA molecules with a specific base sequence. The sample RNA molecules are initially separated by gel electrophoresis.

2. The blotting is done on overlaying nitrocellulose paper on the gel. This transfers the single stranded RNA molecules from the gel to the corresponding position on the nitrocellulose paper, which binds tenaciously to single stranded RNA.

3. The nitrocellulose paper then incubated with ^{32}P labelled single stranded DNA or RNA for *hybridization*. These single stranded nucleic acids are used as ***probe*** and having a base sequence complementary to that of the searched DNA.

4. After hybridization unbound probes are washed and the nitrocellulose paper is *auto-radiographed* to locate the position of the searched RNA.

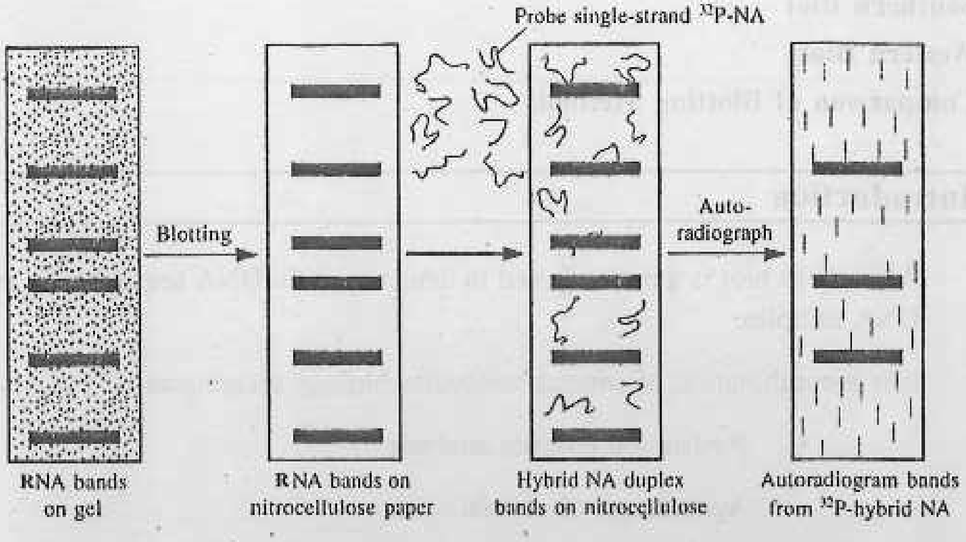


Fig 2 : Derection of **RAN** with specitic base-sequence by **Northern** transfer.

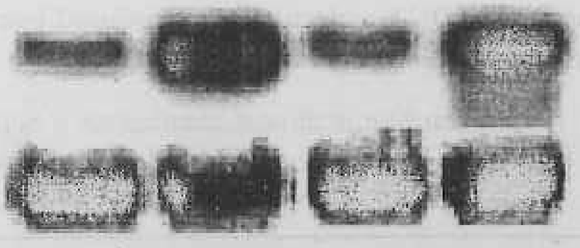


Fig 3 : Photograph of Northern Blot (RNA)

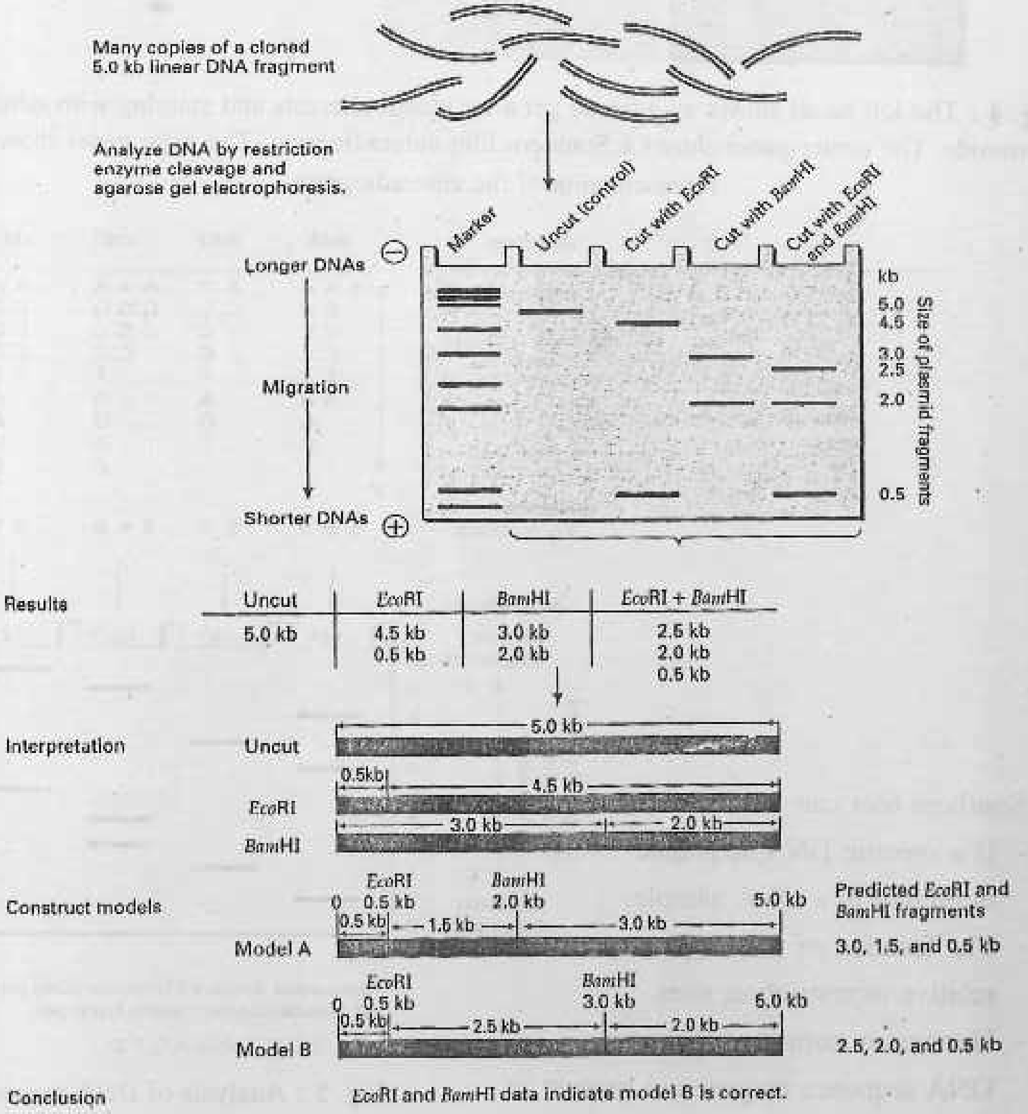
3.3 Southern Blot

Practice and analysis of Southern blot

1. Double stranded DNA molecules are first separated from each other by gel electrophoresis and then denatured into single stranded DNA by soaking the gel slab in 0.5% NaOH solution.
2. The blotting is done on overlaying nitrocellulose paper on the gel. This transfers

the single stranded DNA molecules from the gel to the corresponding position on the nitrocellulose paper, which binds tenaciously to single stranded DNA but not to the double stranded DNA.

3. The nitrocellulose paper then incubated with ^{32}P labelled single stranded DNA or RNA for *hybridization*. These single stranded nucleic acids are used as ***probe*** and having a base sequence complementary to that of the searched DNA.
4. After hybridization unbound probes are washed and the nitrocellulose paper is *auto-radiographed* to locate the position of the searched DNA.



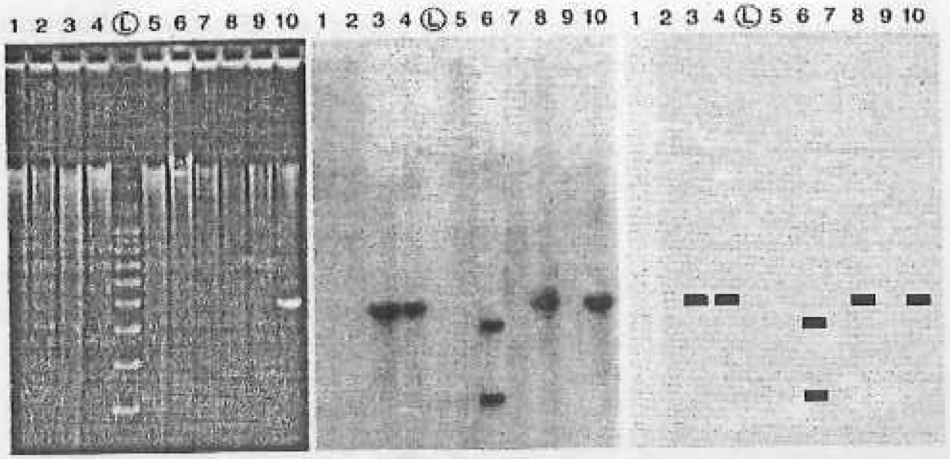


Fig. 4 : The left panel shows an agarose gel after electrophoresis and staining with ethidium bromide. The center panel shows a Southern blot autoradiogram. The right panel shows a representation of the autoradiogram.

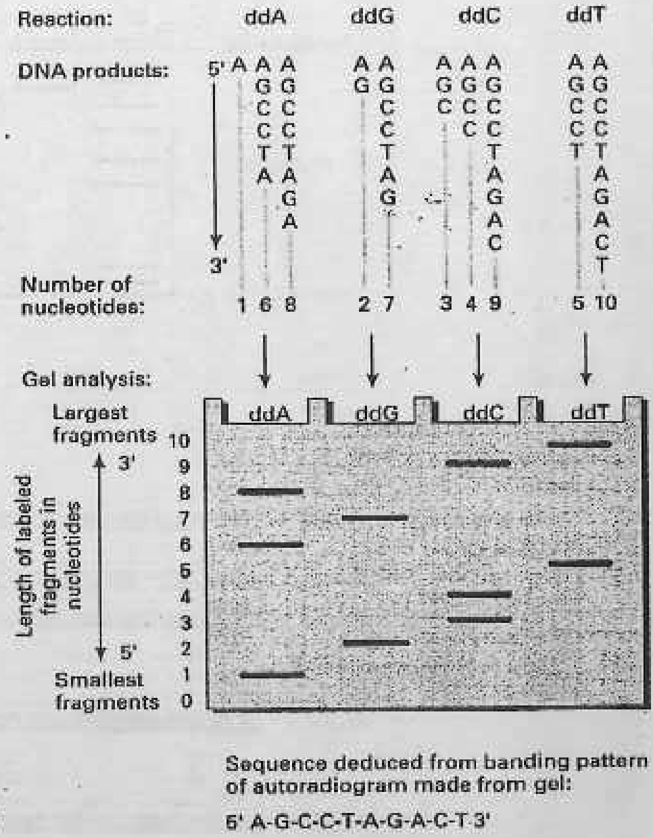


Fig. 5 : Analysis of DNA sequence from southern blot bands.

Southern blot can tell us:-

- If a specific DNA sequence is present in a DNA sample.
- The location of that sequence relative to restriction sites.
- How many copies of that spe DNA sequence or gene are present in the sample.

3.4 Western Blot

1. This technique is used for protein detection and estimation of a specific protein using antibody specific for it.
2. After gel electrophoresis of the sample, the blotting is done on overlaying nitrocellulose paper on the gel.
3. The nitrocellulose paper then incubated with rabbit antiserum containing radioisotope ^{125}I labelled antibodies raised against the specific protein being searched for *hybridization*. These are used as ***probe***.
4. After hybridization unbound probes are washed and the nitrocellulose paper is *auto-radiographed* to locate and estimate the searched protein.
   - Technique for protein detection.
   - probe = antibodies specific to the target protein
     - Linked to an enzyme

   If any antibody is bound by the protein of interest, a colured product will result when it is incubated with the substrate.

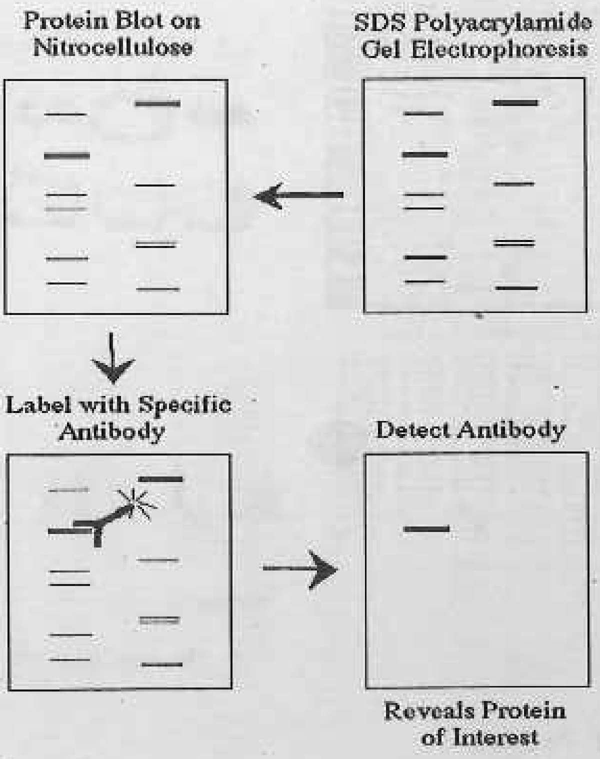
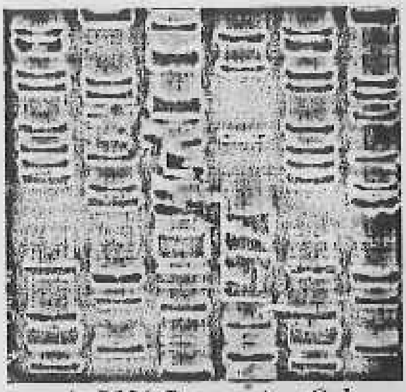


Fig. 6 : Steps for detection of protein using antibody in western blot technique.

Technique used in western blot shows that after blotting in nitrocellulose paper the desired protein is targeted by the specific antibody. Finally the antibody labeled protein will give a color in enzymatic reaction.

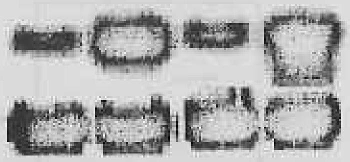
A composite image of four different types of electrophoresis.

Western Blot (Protein)

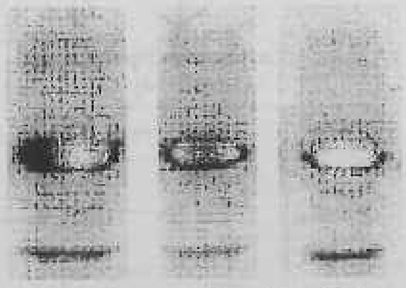


1. DNA Sequencing Gel

2. Northern Blot (RNA)



3. Protein Gel



4. Western Blot (Protein)

Fig. 7 : A Composite image of our different types of Blot techniques

3.5 Comparison of Blotting Methods

	Southern	**Northern**	**Western**	**Southwestern**
What is separated	DNA cut with restriction enzymes	Denatured RNA	Protein denatured with SDS	Characterizes DNA binding proteins
Probe	Radioactive gene X DNA	Radioactive gene X DNA	Antibody against protein X, labeled with enzyme or radioactivity	Labelled DNA probes
What do you learn	Restriction map of gene X chromosome	How much gene X mRNA is present? How long is gene X mRNA	How much protein X is present. How large is protein X.	Identify expression of specific DNA binding proteins

Unit 4 ☐ Data (protein and gene) Bank analysis

Structure

4.1 The Protein Data Bank (PDB)

4.2 Technologies for Whole Proteome Analysis

4.1 The Protein Data Bank (PDB)

It is a repository for the 3-D structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography or NMR spectroscopy and submitted by biologists and biochemists from around the world, can be accessed at no charge on the internet. The PDB is overseen by an organization called the Worldwide Protein Data Bank (wwPDB).

The PDB is a key resource in areas of structural biology, such as structural genomics. Most major scientific journals, and some funding agencies, such as the NIH in the USA, now require scientists to submit their structure data to the PDB. If the contents of the PDB are thought of as primary data, then there are hundreds of derived (i.e., secondary) databases that categorize the data differently. For example, both SCOP (Structaral Classification of Proteins)* and CATH** categorize structures according to type of structure and assumed evolutionary relations; GO*** categorize structures based on genes.

---

* **SCOP** database is a largely manual classification of protein structural domains based on similarities of their amino acid sequences and three-dimensional structures.
SCOP utilizes four levels of hierarchic structural classification :
1. **class** : general "structural architecture" of the domain
2. **fold** : similar arrangement of regular secondary structures but without evidence of evoutionary relatedness
3. **superfamily** : sufficient structural and functional similarity to infer a divergent evolutionary relationship but not necessarily detectable sequence homology
4. **Family** : some sequence similarity can be detected.

** The name **CATH** is an acronym of the four main levels in the classificiation

The four main levels of the CATH hierarchy are as follows :

Level	**Description**
1. Class,	the overall secondary-structure content of the domain
2. Architenture	a large-scale grouping of topologics which share particular structural features
3. Topology	high structural similarity but no evidence of homology. Equivalent to fold in SCOP
4. Homologous superfamily	indicative of a demonstrable evolutionary relationship. Equivalent to the superfamily lelve of SCOP.

CATH defines four classes : mostly-alpha, mostly-beta, alpha and beta, few secondary structures.

*** The **Gene Ontology**, or **GO**, is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species. The aims of the Gene Ontology project are threefold ; firstly, to maintain and further develop its controlled vocabulary of gene and gene product attributes ; secondly, to annotate genes and gene products, and assimilate and desseminate annotation data ; and thirdly, to provide tools to facilitate access to all aspects of the data provided by the Gene Ontology project.

The Protein Data Bank (PDB) was established at Brookhaven National Laboratories (BNL) in 1971 as an archive for biological macromolecular crystal structures. In the beginning the archive held seven structures, and with each year a handful more were deposited. In the 1980s the number of deposited structures began to increase dramatically. This was due to the improved technology for all aspects of the crystallographic process, the addition of structures determined by nuclear magnetic resonance (NMR) methods, and changes in the community views about data sharing. By the early 1990s the majority of journals required a PDB accession code and at least one funding agency (National Institute of General Medical Sciences) adopted the guidelines published by the International Union of Crystallography (IUCr) requiring data deposition for all structures.

The mode of access to PDB data has changed over the years as a result of improved technology, notably the availability of the WWW replacing distribution solely via magnetic media. Further, the need to analyze diverse data sets required the development of modern data management systems.

Initial use of the PDB had been limited to a small group of experts involved in structural research. Today depositors to the PDB have varying expertise in the techniques of X-ray crystal structure determination, NMR, cryoelectron microscopy and theoretical modeling. Users are a very diverse group of researchers in biology, chemistry and computer science, educators, and students at all levels.

The tremendous influx of data soon to be fueled by the structural genomics initiative, and the increased recognition of the value of the data toward understanding biological function, demand new ways to collect, organize and distribute the data.

In October 1998, the management of the PDB became the responsibility of the Research Collaborator for Structural Bioinformatics (RCSB). In general terms, the vision of the RCSB is to create a resource based on the most modern technology that facilitates the use and analysis of structural data and thus creates an enabling resource for biological research. Here, the current procedures for data deposition, data processing and data distribution of PDB data by the RCSB are described. In addition, the issues of data uniformity are addressed.

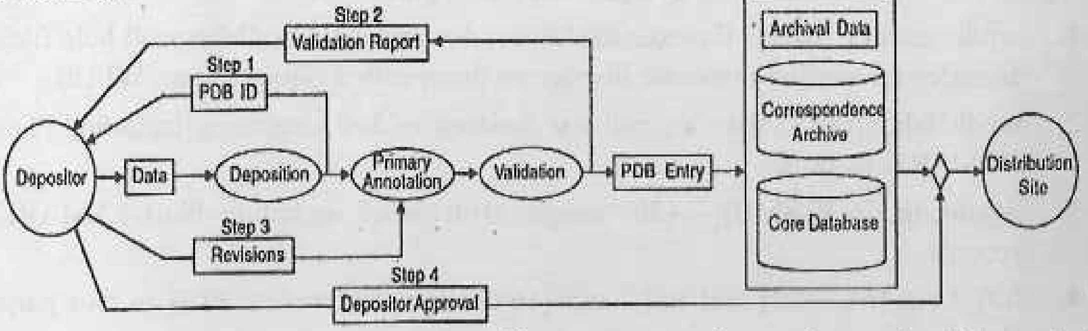


Fig. 8 : The steps in PDB data processing. Ellipses represent actions and rectangles define content

The Worldwide Protein Data Bank (wwPDB) consists of organizations that act as deposition, data processing and distribution centres for PDB data. The founding members are RCSB PDB (Research Collaboratory for Structural Bioinformatics Protein Data Bank). (USA), PDBe (Europe) and PDBj (Japan). The BMRB (Biological Magnitic Resonance Data Bank) (USA) group joined the wwPDB in 2006. The mission of the wwPDB is to maintain a single Protein Data Bank Archive of macromolecular structural data that is freely and publicly available to the global community. This site provides information about services provided by the individual member organizations and about projects undertaken by the wwPDB.

A newly standardized and enhanced version of the entire PDB archive is available at *ftp://ftp.wwpdb.org*. Users who maintain local copies of the wwPDB FTP will have to download the entire archive. Scripts to help in this process are available at *www.wwpdb.org/downloads.html*.

These data reflect the wwPDB's continuing commitment to providing accurate and detailed data to users worldwide. This release includes improvements and enhancements to the data, including details about the chemistry of the polymer and the ligands bound to it, biological assemblies, and binding sites of ligands and metal ions. An overview (PDF) is provided at the wwPDB website.

What is WPDB?

The PDB through Microsoft Windows, or WPDB for short, is a Microsoft Windows 3.1 Windows95 and Windows NT (client and server) based program to interrogate the 3-dimensional structure of biological macromolecules as found in the Protein Data Bank (PDB) using query and display tools like those shown above.

How to Get WPDB

WPDB is available via anonymous ftp from ftp.sdsc.edu in the directory */pub/sdsc/biology/WPDB*. The distribution is organized into 6 parts:

1. wpdbbin.zip [.7MB] - the executables and documentation (a Microsoft help file). Includes raswin the molecule display program called directly from WPDB.
2. wpdb100r.zip [2.1MB] - a small test database of 100 structures, including PDB REMARK records.
3. wpdb420r.zip [9.8MB] - 420 "unique structures," including PDB REMARK records
4. full_1.zip [68.9MB] and full_2.zip [16.6MB] the complete PDB in two parts (both required).

5. wpdbps.zip [.5MB] - program manual in color Postscript.
6. install - installation script (DOS)

   Only one of 2-4 is required. All files are compressed using pkzip. The program pkunzip.exe (runs under DOS) is available in the distribution directory if needed.

What is GenBank?

GenBank® is the genetic sequence database of National Institute of Health, USA (NIH), provides an annotated collection of all publicly available DNA sequences {*Nucleic Acids Research*, 2008 Jan;36(Database issue):D25-30}. There are approximately 85,759,586,764 bases in 82,853,685 sequence records in the traditional GenBank divisions and 108,635,736,141 bases in 27,439,206 sequence records in the WGS division as of February 2008.

The complete release notes for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA Data Bank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

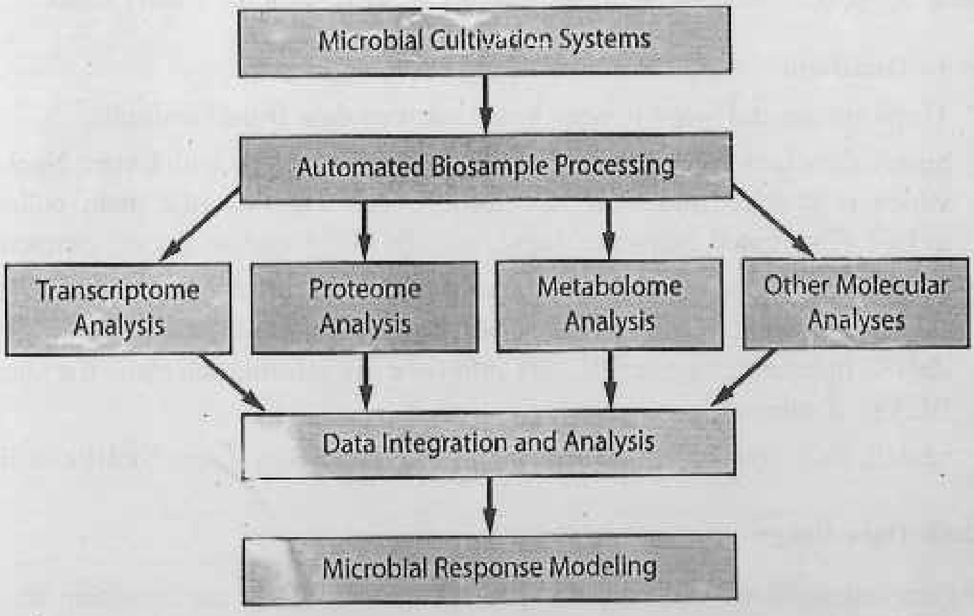
- Search GenBank for sequence identifiers and annotations with Entrez Nucleotide, which is divided into three divisions: CoreNucleotide (the main collection), dbEST (Expressed Sequence Tags), and dbGSS (Genome Survey Sequences).
- Search and align GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently; see BLAST info for more information about the numerous BLAST databases.
- Search, link, and download sequences programatically using NCBI e-utilities.

GenBank Data Usage

The GenBank database is designed to provide and encourage access within the scientific community to the most up to date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.

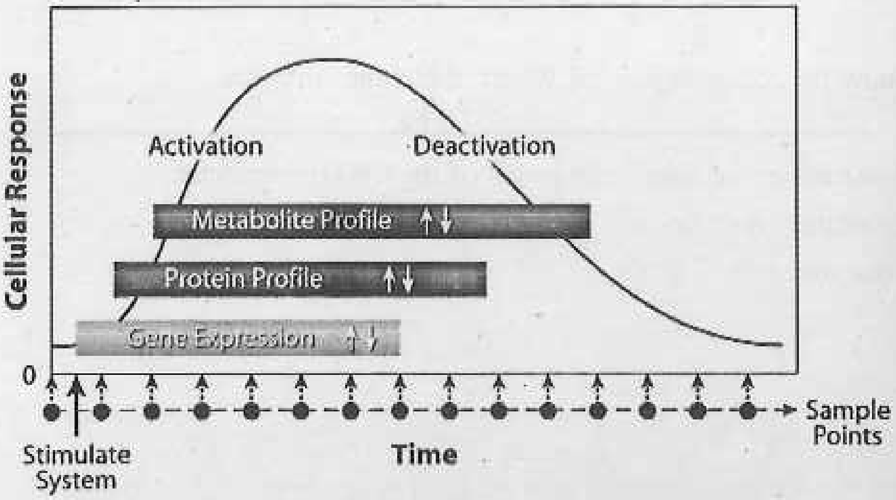
4.2 Technologies for Whole Proteome Analysis

Understanding a microbe's protein-expression profile under various environmental conditions will serve as a basis for identifying individual protein function and will provide the first step toward understanding the complex network of processes conducted by a microbe. Insight into a microbe's expression profile is derived from global analysis of mRNA, protein, and metabolite and other molecular abundance. Characterizing a microbe's expressed protein collection is important in deciphering the function of proteins and molecular machines and the principles and processes by which the genome regulates machine assembly and function and the resultant cellular function. This is not a trivial feat. A microbe typically expresses hundreds of distinct proteins at a time, and the abundance of individual proteins may differ by a factor of a million. Technologies emerging only recently have the potential to measure successfully all proteins across this broad dynamic range. The whole Proteome analysis core capabilities are showing below :



Measuring the time dependence of molecular concentrations—RNAs, proteins, and metabolites—is needed to explore the causal link between genome sequence and cellular function.

Generally, a microbial cell responds to a stimulus by expressing a range of mRNAs translated into a coordinated set of proteins. Measuring RNA expression (transcriptomics) will provide insight into which genes are expressed under a specific set of conditions and thus the full set of processes that are initiated for coordinated molecular response. An even-greater challenge will be detection of precursor regulatory proteins or signaling molecules that start the forward progression of a metabolic process. An example is master regulator molecules that simultaneously control the transcription of many genes. When activated and functioning, proteins expressed by RNA will yield metabolic products. Each organism has a unique biochemical profile, and measuring the cell's collection of metabolites, "metabolomics," is one of the best and most direct methods for determining the cell's biochemical and physiological status. Each of the molecular species' distinct temporal behaviours and their interrelationships must be understood.



To accurately establish causality between measured gene, protein, and metabolite events, sampling strategies must cover the full characteristic time scales of all three variables. Little is known about the time scale of gene, protein, and metabolite responses to specific biological stimuli or how response durations vary among genes and species.

High-capacity computation is needed to integrate all the data from *transcriptomics*, proteomics, and *metabolomics* with additional information obtained from other experimentation and modeling and simulation. These data will be combined to understand and predict microbial responses to different intracellular and environmental stimuli. Petabytes of data generated from all these different measurements will require a substantial investment in computational tools for reducing and analyzing massive data sets and integrating diverse data types.

Unit 5 ☐ Suggestive Questions

1. Find out the correlation of the data provided. Comment on the result.
2. Calculate the data set provided for ANOVA. Comment on the result.

   Or,

   Write the procedure for regression analysis and draw a regression graph from the data provided.
3. Write the procedure for serum sample preparation for gel electrophoresis.

   Or,

   Tabulate the components and quantities of denaturing buffers (DB I, DB II)
4. Write down the staining procedure for protein in Western blot gel.
5. With the help of word diagram show the steps in PDB data processing.

   Or,

   Show the technologies for Whole Proteome Analysis.

   Or,

   Write down the four main levels of the CATH hierarchy.
6. Laboratory note book.
7. Viva voce

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. 150.00

(Not for Sale to the Student of NSOU)

Published by Netaji Subhas Open University, DD-26, Sector-I, Salt Lake, Kolkata - 700064 & Printed at Gita Printers, 51A, Jhamapukur Lane, Kolkata-700 009.