

INDIRA GANDHI NATIONAL TRIBAL UNIVERSITY

AMARKANTAK, M.P. 484886



SUBJECT- Biotechnology

TITLE- Computational Biology

M. Sc Biotechnology 2nd Semester

Reference Notes

Dr. Parikipandla Sridevi

Assistant Professor

Dept. of Biotechnology

Faculty of Science

Indira Gandhi National Tribal University

Amarkantak, MP, India

Pin : 484887

Mob No: +919630036673, +919407331673

Email Id: psridevi@igntu.ac.in, devi.shri45@gmail.com

psridevi.igntu@gmail.com

M. Sc Biotechnology 2nd Semester

Paper name: Computational Biology

Reference Notes

Unit I

Principles and practice of statistical methods in biological research; Samples and Populations; Probability distributions- addition and multiplication theorems, Baye's theorem, Binomial, Poisson, and Normal distribution.

Contents

Introduction

How to determine the appropriate statistical test

Probability

Understanding terminology and symbol in probability

Probability theorem

Bayes' theorem

Normal distribution

Binomial theorem

Poisson distribution

Introduction

The science of statistics is essentially a branch of Applied Mathematics, and may be regarded as mathematics applied to observational data. As in other mathematical studies the same formula is equally relevant to widely different groups of subject matter. Consequently the unity of the different applications has usually been overlooked, the more naturally because the development of the underlying mathematical theory has been much neglected. We shall therefore consider the subject matter of statistics under three different aspects, and then show in more mathematical language that the same types of problems arise in every case. Statistics may be regarded as (i.) the study of populations, (ii.) as the study of variation, (iii.) as the study of methods of the reduction of data. The idea of a population is to be applied not only to living, or even to material, individuals. If an observation, such as a simple measurement, be repeated indefinitely, the aggregate of the results is a population of measurements. Such populations are the particular field of study of the Theory of Errors, one of the oldest and most fruitful lines of statistical investigation. Just as a single observation may be regarded as an individual, and its repetition as generating a population, so the entire result of an extensive experiment may be regarded as but one of a population of such experiments. The salutary habit of repeating important experiments, or of carrying out original observations in replicate, shows a tacit appreciation of the fact that the object of our study is not the individual result,

but the population of possibilities of which we do our best to make our experiments representative.

The calculation of means and probable errors shows a deliberate attempt to learn something about that population. This is a purely practical need which the science of statistics is able to some extent to meet. In some cases at any rate it is possible to give the whole of the relevant information by means of one or a few values. In all cases, perhaps, it is possible to reduce to a simple numerical form the main issues which the investigator has in view, in so far as the data are competent to throw light on such issues. The number of independent facts supplied by the data is usually far greater than the number of facts sought, and in consequence much of the information supplied by anybody of actual data is irrelevant. It is the object of the statistical processes employed in the reduction of data to exclude this irrelevant information, and to isolate the whole of the relevant information contained in the data.

This is a purely practical need which the science of statistics is able to some extent to meet. In some cases at any rate it is possible to give the whole of the relevant information by means of one or a few values. In all cases, perhaps, it is possible to reduce to a simple numerical form the main issues which the investigator has in view, in so far as the data are competent to throw light on such issues. The number of independent facts supplied by the data is usually far greater than the number of facts sought, and in consequence much of the information supplied by anybody of actual data is irrelevant.

How to determine the appropriate statistical test: Follow the steps

1. Specify the biological question you are asking.
2. Put the question in the form of a biological null hypothesis and alternate hypothesis.
3. Put the question in the form of a statistical null hypothesis and alternate hypothesis.
4. Determine which variables are relevant to the question.
5. Determine what kind of variable each one is.
6. Design an experiment that controls or randomizes the confounding variables.
7. Based on the number of variables, the kinds of variables, the expected fit to the parametric assumptions, and the hypothesis to be tested, choose the best statistical test to use.
8. If possible, do a power analysis to determine a good sample size for the experiment.
9. Do the experiment.
10. Examine the data to see if it meets the assumptions of the statistical test you chose (primarily normality and homoscedasticity for tests of measurement variables). If it doesn't, choose a more appropriate test.
11. Apply the statistical test you chose, and interpret the results.
12. Communicate your results effectively, usually with a graph or table.

1. One biological question is "Do the amino acid polymorphisms at the Pgm locus have an effect on glycogen content?" The biological question is usually something about biological processes, often in the form "Does changing X cause a change in Y?" You might want to know whether a drug changes blood pressure; whether soil pH affects the growth of blueberry bushes; or whether protein Rab10 mediates membrane transport to cilia.

2. The biological null hypothesis is “Different amino acid sequences do not affect the biochemical properties of PGM, so glycogen content is not affected by PGM sequence.” The biological alternative hypothesis is “Different amino acid sequences do affect the biochemical properties of PGM, so glycogen content is affected by PGM sequence.” By thinking about the biological null and alternative hypotheses, you are making sure that your experiment will give different results for different answers to your biological question.

3. The statistical null hypothesis is “Flies with different sequences of the PGM enzyme have the same average glycogen content.” The alternate hypothesis is “Flies with different sequences of PGM have different average glycogen contents.” While the biological null and alternative hypotheses are about biological processes, the statistical null and alternative hypotheses are all about the numbers; in this case, the glycogen contents are either the same or different. Testing your statistical null hypothesis is the main subject of this handbook, and it should give you a clear answer; you will either reject or accept that statistical null. Whether rejecting a statistical null hypothesis is enough evidence to answer your biological question can be a more difficult, more subjective decision; there may be other possible explanations for your results, and you as an expert in your specialized area of biology will have to consider how plausible they are.

4. The two relevant variables in the Verrelli and Eanes experiment are glycogen content and PGM sequence.

5. Glycogen content is a measurement variable, something that you record as a number that could have many possible values. The sequence of PGM that a fly has (V-V, V-L, A-V or A-L) is a nominal variable, something with a small number of possible values (four, in this case) that you usually record as a word.

6. Other variables that might be important, such as age and where in a vial the fly pupated, were either controlled (flies of all the same age were used) or randomized (flies were taken randomly from the vials without regard to where they pupated). It also would have been possible to observe the confounding variables; for example, Verrelli and Eanes could have used flies of different ages, and then used a statistical technique that adjusted for the age. This would have made the analysis more complicated to perform and more difficult to explain, and while it might have turned up something interesting about age and glycogen content, it would not have helped address the main biological question about PGM genotype and glycogen content.

7. Because the goal is to compare the means of one measurement variable among groups classified by one nominal variable, and there are more than two categories, the appropriate statistical test is a one-way anova. Once you know what variables you’re analysing and what type they are, the number of possible statistical tests is usually limited to one or two.

8. A power analysis would have required an estimate of the standard deviation of glycogen content, which probably could have been found in the published literature, and a number for the effect size (the variation in glycogen content among genotypes that the experimenters wanted to detect). In this experiment, any difference in glycogen content among genotypes would be interesting, so the experimenters just used as many flies as was practical in the time available.

9. The experiment was done: glycogen content was measured in flies with different PGM sequences.

10. The anova assumes that the measurement variable, glycogen content, is normal (the distribution fits the bell-shaped normal curve) and homoscedastic (the variances in glycogen content of the different PGM sequences are equal), and inspecting histograms of the data shows that the data fit these assumptions. If the data hadn't met the assumptions of anova, the Kruskal–Wallis test or Welch's test might have been better.

11. The one-way anova was done, using a spreadsheet, web page, or computer program, and the result of the anova is a Pvalue less than 0.05. The interpretation is that flies with some PGM sequences have different average glycogen content than flies with other sequences of PGM.

12. The results could be summarized in a table, but a more effective way to communicate them is with a graph:

Probability

Although estimating probabilities is a fundamental part of statistics, you will rarely have to do the calculations yourself. It's worth knowing a couple of simple rules about adding and multiplying probabilities.

Introduction- The basic idea of a statistical test is to identify a null hypothesis, collect some data, and then estimate the probability of getting the observed data if the null hypothesis were true. If the probability of getting a result like the observed one is low under the null hypothesis, you conclude that the null hypothesis is probably not true. It is therefore useful to know a little about probability.

One way to think about probability is as the proportion of individuals in a population that have a particular characteristic. The probability of sampling a particular kind of individual is equal to the proportion of that kind of individual in the population.

For example, in fall 2013 there were 22,166 students at the University of Delaware, and 3,679 of them were graduate students. If you sampled a single student at random, the probability that they would be a grad student would be $3,679 / 22,166$, or 0.166. In other words, 16.6% of students were grad students, so if you'd picked one student at random, the probability that they were a grad student would have been 16.6%. When dealing with probabilities in biology, and you are often working with theoretical expectations, not population samples.

For example, in a genetic cross of two individual *Drosophila melanogaster* that are heterozygous at the vestigial locus, Mendel's theory predicts that the probability of an offspring individual being a recessive homozygote (having teeny-tiny wings) is one-fourth, or 0.25. This is equivalent to saying that one-fourth of a population of offspring will have tiny wings.

Understand and use the terminology of probability

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An **experiment** is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **chance experiment**. Flipping one fair coin twice is an example of an experiment.

A result of an experiment is called an outcome. The sample space of an experiment is the set of all possible outcomes. Three ways to represent a sample space are: to list the possible outcomes, to create a tree diagram, or to create a Venn diagram.

The uppercase letter S is used to denote the sample space. For example, if you flip one fair coin, $S = \{H, T\}$ where H = heads and T = tails are the outcomes.

An **event** is any combination of outcomes. Upper case letters like A and B represent events. For example, if the experiment is to flip one fair coin, event A might be getting at most one head. The probability of an event A is written $P(A)$.

The probability of any outcome is the long-term relative frequency of that outcome. Probabilities are between zero and one, inclusive (that is, zero and one and all numbers between these values). $P(A) = 0$ means the event A can never happen. $P(A) = 1$ means the event A always happens. $P(A) = 0.5$ means the event A is equally likely to occur or not to occur.

For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative frequency of heads approaches 0.5 (the probability of heads).

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a **fair**, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head (H) and a Tail (T) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event A when all outcomes in the sample space are equally likely, count the number of outcomes for event A and divide by the total number of outcomes in the sample space. For example, if you toss a fair dime and a fair nickel, the sample space is $\{HH, TH, HT, TT\}$ where T = tails and H = heads. The sample space has four outcomes. A = getting one head. There are two outcomes that meet this condition $\{HT, TH\}$ so $P(A) = 2/4 = 0.5$.

Suppose you roll one fair six-sided die, with the numbers $\{1, 2, 3, 4, 5, 6\}$ on its faces. Let event E = rolling a number that is at least five. There are two outcomes $\{5, 6\}$. $P(E) = 2/6$ as the number of repetitions grows larger and larger.

This important characteristic of probability experiments is known as the **law of large numbers** which states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability.

Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word **empirical** is often used instead of the word observed.)

Understanding Terminology and Symbols

It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any

Example

The sample space S is the whole numbers starting at one and less than 20.

1. $S =$ _____ Let event $A =$ the even numbers and event $B =$ numbers greater than 13.
2. $A =$ _____, $B =$ _____
3. $P(A) =$ _____, $P(B) =$ _____
4. $A \text{ AND } B =$ _____, $A \text{ OR } B =$ _____
5. $P(A \text{ AND } B) =$ _____, $P(A \text{ OR } B) =$ _____
6. $A' =$ _____, $P(A') =$ _____
7. $P(A) + P(A') =$ _____
8. $P(A|B) =$ _____, $P(B|A) =$ _____; are the probabilities equal?

Solution:

1. $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$
2. $A = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}$, $B = \{14, 15, 16, 17, 18, 19\}$
3. $P(A)=9/19, P(B)=6/19$
- $A \text{ AND } B = \{14, 16, 18\}$, $A \text{ OR } B = 2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19\}$
- $P(A \text{ AND } B)=3/19, P(A \text{ OR } B)=12/19$
- $A'=1, 3, 5, 7, 9, 11, 13, 15, 17, 19; P(A'right)=10/19$
- $P(A)+P(A')=1(9/19+10/19)=1$
- $P(A|B) = P(A \text{ AND } B)/P(B)=3/6, P(B|A)=P(A \text{ AND } B)/P(A)=3/9, \text{ No}$

Probability = the number of ways of achieving success. the total number of possible outcomes. For example, the **probability** of flipping a coin and it being heads is $\frac{1}{2}$, because there is 1 way of getting a head and the total number of possible outcomes is 2 (a head or tail). We write $P(\text{heads}) = \frac{1}{2}$.

Multiplying probabilities You could take a semester-long course on mathematical probability, but most biologists just need to know a few basic principles. You calculate the probability that an individual has one value of a nominal variable and another value of a second nominal variable by multiplying the probabilities of each value together. For example, if the probability that a *Drosophila* in a cross has vestigial wings is one-fourth, and the probability that it has legs where its antennae should be is three-fourths, the probability that it has vestigial wings and leg-antennae is one-fourth times three-fourths, or 0.25×0.75 , or 0.1875.

This estimate assumes that the two values are independent, meaning that the probability of one value is not affected by the other value. In this case, independence would require that the two genetic loci were on different chromosomes, among other things.

The **theorem** states that the **probability** of the simultaneous occurrence of two events that are independent is given by the product of their individual **probabilities**.

Adding probabilities The probability that an individual has one value or another, mutually exclusive, value is found by adding the probabilities of each value together. “Mutually exclusive” means that one individual could not have both values.

For example, if the probability that a flower in a genetic cross is red is one-fourth, the probability that it is pink is one-half, and the probability that it is white is one-fourth, then the probability that it is red or pink is one-fourth plus one-half, or three-fourths.

Addition theorem on probability: If A and B are any two events then the **probability** of happening of at least one of the events is defined as $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

More complicated situations- When calculating the probability that an individual has one value or another, and the two values are not mutually exclusive, it is important to break things down into combinations that are mutually exclusive.

For example, let’s say you wanted to estimate the probability that a fly from the cross above had vestigial wings or leg-antennae.

You could calculate the probability for each of the four kinds of flies: normal wings/normal antennae ($0.75 \times 0.25 = 0.1875$), normal wings/leg-antennae ($0.75 \times 0.75 = 0.5625$), vestigial wings/normal antennae ($0.25 \times 0.25 = 0.0625$), and vestigial wings/leg-antennae ($0.25 \times 0.75 = 0.1875$).

Then, since the last three kinds of flies are the ones with vestigial wings or leg-antennae, you’d add those probabilities up ($0.5625 + 0.0625 + 0.1875 = 0.8125$).

In probability theory and statistics, Bayes' theorem (alternatively Bayes' law or Bayes' rule) describes the probability of an event, based on prior knowledge of conditions that might be related to the event. When applied, the probabilities involved in Bayes' theorem may have different probability interpretations.

Bayes' theorem (alternatively **Bayes' law** or **Bayes' rule**) describes the probability of an event, based on prior knowledge of conditions that might be related to the event. For example, if the risk of developing health problems is known to increase with age, Bayes' theorem allows the risk to an individual of a known age to be assessed more accurately than simply assuming that the individual is typical of the population as a whole.

One of the many applications of Bayes' theorem is Bayesian inference, a particular approach to statistical inference. When applied, the probabilities involved in Bayes' theorem may have different probability interpretations.

With Bayesian probability interpretation, the theorem expresses how a degree of belief, expressed as a probability, should rationally change to account for the availability of related evidence. Bayesian inference is fundamental to Bayesian state.

The concept of conditional probabilities introduced in Elementary Statistics. We noted that the conditional probability of an event is a probability obtained with the additional information that some other event has already occurred. We used $P(B|A)$ to denote the conditional probability of event B occurring, given that event A has already occurred. The following formula was provided for finding $P(B|A)$

The Gallup organization randomly selects an adult American for a survey about credit card usage. Use subjective probabilities to estimate the following

- a. What is the probability that the selected subject is a male?
- b. After selecting a subject, it is later learned that this person was smoking a cigar during the interview. What is the probability that the selected subject is a male?
- c. Which of the preceding two results is a prior probability?
Which is a posterior probability?

Solution

a. Roughly half of 1 Americans are males, so we estimate the probability of selecting a male subject to be 0.5. Denoting a male by M, we can express this probability as follows: $P(M) = 0.5$.

b. Although some women smoke cigars, the vast majority of cigar smokers are males. A reasonable guess is that 85% of cigar smokers are males. Based on this additional subsequent information that the survey respondent was smoking a cigar, we estimate the probability of this person being a male as 0.85. Denoting a male by M and denoting a cigar smoker by C, we can express this result as follows: $P(M | C) = 0.85$. c. In part (a), the value of 0.5 is the initial probability, so we refer to it as the prior probability. Because the probability of 0.85 in part (b) is a revised probability based on the additional information that the survey subject was smoking a cigar, this value of 0.85 is referred to a posterior probability.

Normal Distribution

- Applied to single variable continuous data e.g. heights of plants, weights of lambs, lengths of time
- Used to calculate the probability of occurrences less than, more than, between given values e.g. “the probability that the plants will be less than 70mm”, “the probability that the lambs will be heavier than 70kg”, “the probability that the time taken will be between 10 and 12 minutes”

- Standard Normal tables give probabilities—you will need to be familiar with the Normal table and know how to use it. First need to calculate how many standard deviations above (or below) the mean a particular value is, calculate the value of the “standard score” or “Z-

score". Use the following formula to convert a raw data value, X , to a standard score, Z : $Z = \frac{X - \mu}{\sigma}$. Suppose a particular population has $\mu = 17$ and $\sigma = 2$.

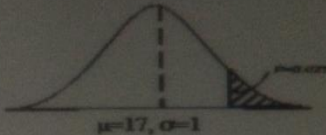
Find the probability of a randomly selected value being greater than

6. The Z score corresponding to $X = 6$ is $Z = -1.9$. ($Z = 1$ means that the value $X = 6$ is 1 standard deviation above the mean.)

Now use standard normal tables to find $P(Z > 1) = 0.6587$ (more about this later).
 Process:
 o Draw a diagram and label with given values i.e. (μ, σ) , (X) , and (Z) .
 o Shade area required as per question.
 o Convert raw score (X) to standard score (Z) using formula.
 o Use tables to find probability: eg $P(Z < 1) = 0.7413$.
 o Adjust this result to required probability.

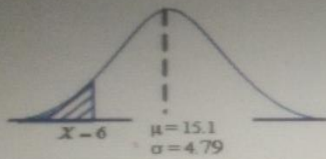
Normal Distribution Solutions

1. (This is an inverse problem)
 $\mu = 17, \sigma = 1$



The z value that cuts off the upper 2.5% of the standard normal is 1.96. Now find the value for potassium which is 1.96 SDs above the mean for a healthy person.
 Start with $z = \frac{x - \mu}{\sigma}$ i.e. $1.96 = \frac{x - 17}{1.0} \Rightarrow x = 1.96 \times 1.0 + 17 = 18.96$
 That is $X = 18.96$ mg/100 ml.

2.



$$Z = \frac{6 - 15.1}{4.79} = -1.9$$

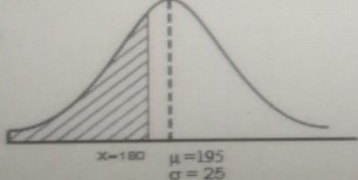
$$P(Z < -1.9) = 0.5 - P(0 < Z < 1.9)$$

$$= 0.5 - 0.4713$$

$$= 0.0287$$

(a) That is the proportion of wool with a crimp of 6 or less is 2.87%.
 (b) Yes, this is satisfactory since 2.87% is less than the stated 7% of the wool.

3. (a)



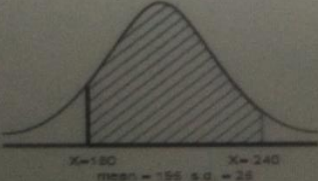
$$x = 180 \Rightarrow z = \frac{180 - 195}{25} = -0.6$$

$$P(z < -0.6) = 0.5 - P(0 < z < 0.6) = 0.5 - 0.2257 = 0.2743$$

i.e. probability of a runner taking less than 3 hours (180 mins) is 0.2743

(b) $p = 0.08 \Rightarrow Z = -1.41$
 Hence $-1.41 = \frac{X - 195}{25}$
 $\Rightarrow X = -1.41 \times 25 + 195$
 $= 159.75$ or 160 mins

(c)



$$X = 180 \Rightarrow Z = \frac{180 - 195}{25} = -0.6$$

$$P(Z < -0.6) = 0.2743$$

$$X = 240 \Rightarrow Z = \frac{240 - 195}{25} = 1.8$$

$$P(Z < 1.8) = 0.9641$$

$$\Rightarrow P(-0.6 < Z < 1.8) = 0.9641 - 0.2743 = 0.6898$$

i.e. proportion of runners taking between 3 and 4 hours (180 and 240 minutes) is approximately 70%.

Practice (Normal Distribution)

1 Potassium blood levels in healthy humans are normally distributed with a mean of 17.0 mg/100 ml, and standard deviation of 1.0 mg/100 ml. Elevated levels of potassium indicate an electrolyte balance problem, such as may be caused by Addison's disease. However, a test for potassium level should not cause too many "false positives". What level of potassium should we use so that only 2.5 % of healthy individuals are classified as "abnormally high"?

2. For a particular type of wool the number of 'crimps per 10cm' follows a normal distribution with mean 15.

1 and standard deviation 4.79.

(a) What proportion of wool would have a 'crimp per 10 cm' measurement of 6 or less?

(b) If more than 7% of the wool has a 'crimp per 10 cm' measurement of 6 or less, then the wool is unsatisfactory for a particular processing. Is the wool satisfactory for this processing?

3. The finish times for marathon runners during a race are normally distributed with a mean of 195 minutes and a standard deviation of 25 minutes.

a) What is the probability that a runner will complete the marathon within 3 hours?

b) Calculate to the nearest minute, the time by which the first 8% runners have completed the marathon.

c) What proportion of the runners will complete the marathon between 3 hours and 4 hours?

4. The download time of a resource web page is normally distributed with a mean of 6.5 seconds and a standard deviation of 2.3 seconds.

a) What proportion of page downloads take less than 5 seconds?

b) What is the probability that the download time will be between 4 and 10 seconds?

c) How many seconds will it take for 35% of the downloads to be completed?

Binomial Distribution

• Applied to single variable discrete data where results are the numbers of "successful outcomes" in a given scenario. e.g.: no. of times the lights are red in 20 sets of traffic lights, no. of students with green eyes in a class of 40, no. of plants with diseased leaves from a sample of 50 plants

• Used to calculate the probability of occurrences exactly, less than, more than, between given values e.g. the "probability that the number of red lights will be exactly 5" "probability that the number of green eyed students will be less than 7" "probability that the no. of diseased plants will be more than 10"

Parameters, statistics and symbols involved are: population parameter symbol sample statistic symbol probability of success π p sample size N n

• Other symbols : X , the number of successful outcomes wanted r x n C or C_r ; the number of ways in which x "successes" can be chosen from sample size n . Then C key on your calculator can be used directly in the formula.

Binomial Distribution Solutions

1. (a) $n = 20, p = 0.05, X < 1 \Rightarrow X = 0$
 $P(X = 0) = {}^{20}C_0 0.05^0 \times 0.95^{20} = 0.3585$
- (b) $n = 20, p = 0.05, X \leq 1 \Rightarrow X = 0, 1$
 $P(X = 0) = 0.3584$
 $P(X = 1) = {}^{20}C_1 0.05^1 \times 0.95^{19} = 0.0.3774$
 $\Rightarrow P(X = 0, 1) = 0.3585 + 0.3774 = 0.7359$
- (c) $n = 20, p = 0.05, X \leq 2 \Rightarrow X = 0, 1, 2$
 $P(X = 0) = 0.3584$
 $P(X = 1) = 0.0.3774$
 $P(X = 2) = {}^{20}C_2 0.05^2 \times 0.95^{18} = 0.0.0.1887$
 $\Rightarrow P(X = 0, 1, 2) = 0.3585 + 0.3774 + 0.1887 = 0.9246$

2. (a) $p = 0.05, n = 5, X = 2$
 $\Rightarrow P(X = 2) = {}^5C_2 \times 0.05^2 \times 0.95^3 = 0.0214$ i.e. $\approx 2\%$
- (b) $P(2 \text{ in two years}) = {}^2C_2 \times 0.05^2 \times 0.95^0 = 0.0025$ i.e. $\approx 0.25\%$
- (c) $P(\text{at least once}) = P(X \geq 1)$
 $= 1 - P(X = 0) = 1 - {}^4C_0 0.05^0 0.95^4$
 $= 0.1854$
3. (a) $p = 0.2, n = 15, X = 2$
 $\Rightarrow P(X = 2) = {}^{15}C_2 \times 0.2^2 \times 0.8^{13} = 0.2309$ i.e. $\approx 23\%$
- (b) $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0)$
 $= 1 - {}^{15}C_0 \times 0.2^0 \times 0.8^{15} = 0.9648$ i.e. $\approx 97\%$

Practice(Binomial Distribution)1 Executives in the New Zealand Forestry Industry claim that only 5% of all old sawmills sites contain soil residuals of dioxin (an additive previously used for anti-sap-stain treatment in wood) higher than the recommended level.

If Environment Canterbury randomly selects 20 old saw mill sites for inspection, assuming that the executive claim is correct :a) Calculate the probability that less than 1 site exceeds the recommended level of dioxin.

b) Calculate the probability that less than or equal to 1 site exceed the recommended level of dioxin.

c) Calculate the probability that at most (i.e., maximum of) 2 sites exceed the recommended level of dioxin.

2 Inland Revenue audits 5% of all companies every year.

The companies selected for auditing in any one year are independent of the previous year's selection.

a) What is the probability that the company 'Ross Waste Disposal' will be selected for auditing exactly twice in the next 5 years?

b) What is the probability that the company will be audited exactly twice in the next 2 years?

c) What is the exact probability that this company will be audited at least once in the next 4 years?

3 The probability that a driver must stop at any one traffic light coming to Lincoln University is 0.2. There are 15 sets of traffic lights on the journey. a) What is the probability that a student must stop at exactly 2 of the 15 sets of traffic lights?

b) What is the probability that a student will be stopped at 1 or more of the 15 sets of traffic lights?

Poisson distribution

This is often known as the distribution of rare events. Firstly, a Poisson process is where DISCRETE events occur in a CONTINUOUS, but finite interval of time or space. The following conditions must apply:

♣ For a small interval the probability of the event occurring is proportional to the size of the interval.

♣ The probability of more than one occurrence in the small interval is negligible (i.e. they are rare events).

Events must not occur simultaneously

♣ Each occurrence must be independent of others and must be at random.

♣ The events are often defects, accidents or unusual natural happenings, such as earthquakes, where in theory there is no upper limit on the number of events. The interval is on some continuous measurement such as time, length or area.

Practice (Poisson Distribution)

1. A radioactive source emits 4 particles on average during a five-second period.

a) Calculate the probability that it emits 3 particles during a 5-second period.

b) Calculate the probability that it emits at least one particle during a 5-second period.

c) During a ten-second period, what is the probability that 6 particles are emitted?

2. The number of typing mistakes made by a secretary has a Poisson distribution. The mistakes are made independently at an average rate of 1.65 per page. Find the probability that a three-page letter contains no mistakes.

3. A 5-litre bucket of water is taken from a swamp. The water contains 75 mosquito larvae. A 200mL flask of water is taken from the bucket for further analysis. What is a) the expected

number of larvae in the flask? b) the probability that the flask contains at least one mosquito larva?!) ($e^{-\lambda} \frac{\lambda^x}{x!}$ for $x = 0, 1, 2, 3, \dots$)

94. If the light bulbs in a house fail according to a Poisson law, and over the last 15 weeks there have been 5 failures, find the probability that there will not be more than one failure next week

Reference-

1. JOHN. H. McDONALD (hand book of biological statistics)
2. GEORGE CASELLA, R.L BERGER (statistical inference), Solution manual for statistical inference.
3. QMET201
4. Web..lumenlearning.com/introstats1>chapter,the terminology.

“Data Presentation – Types of Data, Methods of Data Representation”

- Introduction
- Types of data
- Data representation
- Methods of data representation
- Conclusion
- References

INTRODUCTION

- Any observation collected in respect of any characteristic or event is called data.
- Data Representation refers to the form in which data is stored, processed, and transmitted.
- Whenever the data is collected for some project, it is usually in the ‘raw’ form and not in an organized way. Descriptive statistics deals with sorting this raw data by putting it into a table or by presenting it in an appropriate chart or summarizing it numerically.
- An important consideration in sorting the raw data is the type of variable concerned.
- The data from some variables are best described with a table, some with a chart, and some with both. However, a numeric summary is more appropriate for some types of variable.
- Raw data carry/convey little meaning, when it is considered alone.
- The data is minimized, processed/analyzed and then presented systematically. So that it is converted into Information.
- It is important to note that, data that is not converted into information is of little value for evaluation and planning and cannot be used by those who are involved in decision making.

Types of Data

- To give a holistic picture of classification data can be divided into two types:
 - i. Quantitative data (numerical) and
 - ii. Qualitative data (descriptive, categorical/frequency count).

Quantitative data

The data that can be expressed in numbers/figures is called quantitative data.

It has two types:

(a) Discrete: Discrete variables can take only certain values and none in between e. g. number of patients in a hospital census may be 178 or 179, but it cannot be in between these two, similarly the number of syringes used in a clinic in one day or number of children in a family. It is expressed in whole number.

(b) Continuous: Continuous variables may take any value (typically between certain limits). For example age (25.5 years), weight (70.5 kg), height (1.5 meter), hemoglobin (12.5 gm.), blood pressure (135/95). It can be expressed in decimals.

Qualitative Data

- Also called descriptive/ categorical data/ frequency count.
- When the data are arranged in categories on the basis of their quality and there is gap between two values, it is called qualitative data, e.g. name, religion, marital status, socioeconomic status, awareness.
- Qualitative data cannot be expressed in numerical forms.
 - Types of qualitative data:
 - ❖ **Nominal data:** Nominal scale data are divided into categories that are only distinguished by their name and labels and cannot be classified one above another e.g. Race, name, sex, name of country, name of crops, type of blood.
 - In this type of data there is no implication of order or ratio.
 - Nominal data that falls into two groups are called dichotomous data e.g. male/ female, black/white, rural/ urban.
 - ❖ **Ordinal data:** When the categorical data can be placed in meaningful order on the basis of their quality, it is known as ordinal data. In this the exact difference between the two groups cannot be estimated e.g. pain categorized as mild, moderate and severe. Similarly scoring of students categorized as A (70% and above), B (60-69 %), C (50-59 %). In this the exact difference between the students placed in grade A and B cannot be estimated.
 - ❖ **Interval:** Interval scale data are like ordinal data in that they can be placed in a meaningful order.
 - In addition they have meaningful intervals between items, which are usually measured quantities. For example on the Celsius scale the difference between 100° C and 90° C is the same as the difference between 50° C and 40° C.

However because interval scales do not have an absolute zero, ratio of scores are not meaningful e.g. 100° C is not twice as hot as 50° C, because 0° C does not indicate a complete absence of heat.

Ratio: A ratio scale has the same properties as an interval scale; however, because it has an absolute zero, meaningful ratios do exist.

Most biomedical variables form a ratio scale e.g. weight in grams or pounds, time in seconds or days, blood pressure in millimeters of mercury, and pulse rate in beats per minute are all ratio scale data.

Data Representation

Principals of data representation:

- (a) To arrange the data in such a way that it should create interest in the reader's mind at the first sight.
- (b) To present the information in a compact and concise form without losing important details.
- (c) To present the data in a simple form so as to draw the conclusion directly by viewing at the data.
- (d) To present it in such a way that it can help in further statistical analysis.

Methods of Data representation

Data may be presented by 3 methods:

- Textual,
 - Tabular or
 - Graphical.
1. **Textual:** the data gathered and presented in paragraph form. It is a combination of texts and figures.
 2. **Tabular:** method of presenting data using the statistical table. A systematic organization of data in columns and rows.
 - Tabulation is the first step before the data is used for analysis or interpretation. Frequency distribution tables presents data in a relatively compact form, ready to use but certain information may be lost. The data can be reduced to manageable form using frequency tables.
 - Tables are the devices that are used to present the data in a simple form.
 - It is probably the first step before the data is used for analysis or interpretation. General principals of designing tables –
 - a) The tables should be numbered e.g. table 1, table 2 etc.
 - b) A title must be given to each table, which should be brief and self-explanatory.
 - c) The headings of columns or rows should be clear and concise.

- d) The data must be presented according to size or importance chronologically, alphabetically, or geographically.
- e) If percentages or averages are to be compared, they should be placed as close as possible.
- f) No table should be too large.
- g) Most of the people find a vertical arrangement better than a horizontal one because, it is easier to scan the data from top to bottom than from left to right.
- h) Foot notes may be given, where necessary, providing explanatory notes or additional information.

Types of tables:

- 1) Simple tables: Measurements of single set are presented
- 2) Complex tables: Measurements of multiple sets are presented
- 3) Graphical:

For quantitative data

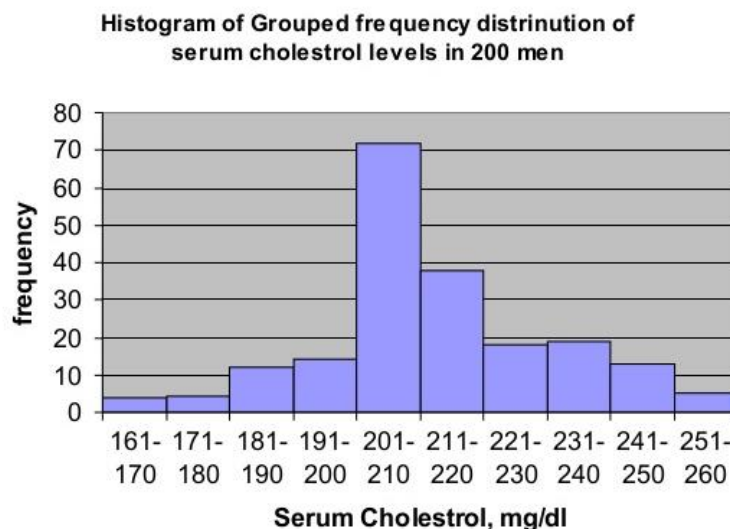
- 1. Histogram,
- 2. Frequency polygon,
- 3. Frequency curve,
- 4. Line chart,
- 5. Normal distribution curve,
- 6. Cumulative distribution curve
- 7. Scatter diagram.

For qualitative data

- 1. Bar chart,
- 2. Pictogram,
- 3. Pie chart and
- 4. Map diagram

Histogram:

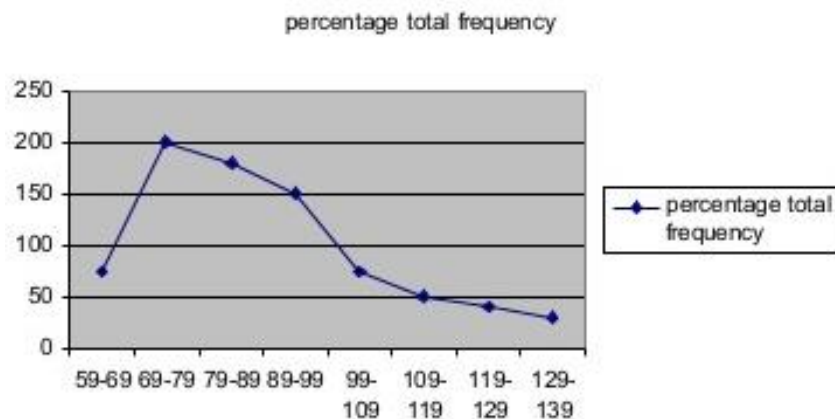
- Used for Quantitative, Continuous, Variables.
- It is used to present variables which have no gaps e.g. age, weight, height, blood pressure, blood sugar etc.
- It consists of a series of blocks. The class intervals are given along horizontal axis and the frequency along the vertical axis.



Frequency Polygon:

- Frequency polygon is an area diagram of frequency distribution over a histogram.

- It is a linear representation of a frequency table and histogram, obtained by joining the mid points of the histogram blocks. Frequency is plotted at the central point of a group.



Frequency Curve:

- A frequency-curve is a smooth curve for which the total area is taken to be unity. It is a limiting form of a histogram or frequency polygon. The frequency-curve for a distribution can be obtained by drawing a smooth and free hand curve through the mid-points of the upper sides of the rectangles forming the histogram.

Line Chart:

- A line graph, also known as a line chart, is a type of chart used to visualize the value of something over time. For example, a finance department may plot the change in the amount of cash the company has on hand over time. The line graph consists of a horizontal x-axis and a vertical y-axis.

Normal distribution curve:

- A normal distribution is sometimes informally called a bell curve. However, many other distributions are bell-shaped (such as the Cauchy, Student's *t*, and logistic distributions).

Cumulative distribution curve:

- The cumulative distribution function gives the cumulative value from negative infinity up to a random variable *X* and is defined by the following notation: $F(x) = P(X \leq x)$. Standard normal distribution showing standard deviations.

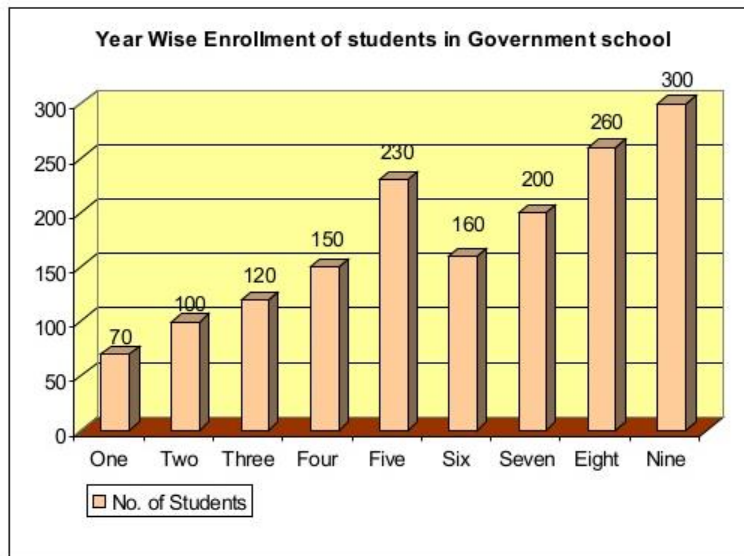
Scatter diagram:

- Scatter diagrams show the relationship between the two variables e.g. a positive correlation/ association between the intake of fat and sugar in the average diets of 41 countries.
- If the dots cluster round a straight line, it shows evidence of a relationship of a linear nature.
- If there is no such cluster, it is probable that there is no relationship between the variables.

Bar chart:

- The data presented is categorical.
- Data is presented in the form of rectangular bar of equal breadth.
- Each bar represent one variant /attribute. Suitable scale should be indicated and scale starts from zero. The width of the bar and the gaps between the bars should be equal throughout.
- The length of the bar is proportional to the magnitude/ frequency of the variable. The bars may be vertical or horizontal.

Bar charts



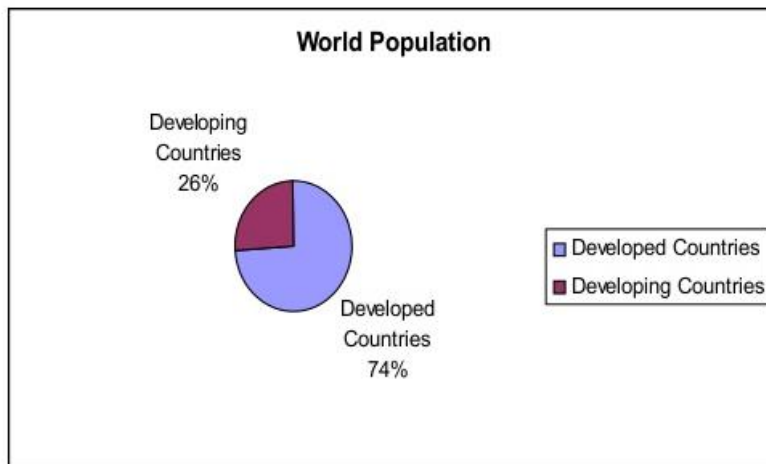
Pictogram:

- Popular method of presenting data to those who cannot understand orthodox charts.
- Small pictures or symbols are used to present the data, e.g. a picture of a doctor to represent the population physician.
- Fraction of the picture can be used to represent numbers smaller than the value of whole symbol.

Pie chart:

- Most common way of presenting data
- The value of each category is divided by the total values and then multiplied by 360 and then each category is allocated the respective angle to present the proportion it has.
- It is often necessary to indicate percentages in the segment as it may not be sometimes very easy virtually, to compare the areas of segments.

Pie Charts



Conclusions

- Text, tables, and graphs are effective communication media that present and convey data and information. They aid readers in understanding the content of research, sustain their interest, and effectively present large quantities of complex information.
- As journal editors and reviewers will scan through these presentations before reading the entire text, their importance cannot be disregarded. For this reason, authors must pay as close attention to selecting appropriate methods of data presentation as when they were collecting data of good quality and analyzing them.
- In addition, having a well-established understanding of different methods of data presentation and their appropriate use will enable one to develop the ability to recognize and interpret inappropriately presented data or data presented in such a way that it deceives readers' eyes.

References

- TEXTBOOK OF COMPUTER APPLICATIONS AND BIostatISTICS
- Huff D. How to Lie with Statistics. London: Penguin Books; 1991. pp. 1–124.
- <https://www.slideshare.net/ahsanshafi90/data-presentation-2-15572325>

Unit II

Measures of central tendency- Mean, Median, Mode; Measures of dispersion- Range, Mean deviation and Coefficient of variation, Standard deviation, Standard error; Correlation and regression;

Contents

Central Tendency Or Average (Mean (with type explained), Median, Mode)

Dispersion and Measures of Dispersion (with its types)

Range

Mean deviation

Coefficient of variation

Standard deviation

Standard error

Correlation

Regression

Central Tendency or Average

(Introduction)

Central tendency is defined as “the statistical measure that identifies a single value as representative of an entire distribution”. It aims to provide an accurate description of the entire data. It is the single value that is most typical/representative of the collected data. A measure of central tendency is a summary statistic that represents the centre point or typical value of a dataset. These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution. The tendency of data to cluster around a middle value. In statistics, the three most common measures of central tendency are the mean, median, and mode. Each of these measures calculates the location of the central point using a different method.

The central tendency is one of the most quintessential concepts in statistics. Although it does not provide information regarding the individual values in the dataset, it delivers a comprehensive summary of the whole dataset.

Definition of Average

A number expressing the central or typical value in a set of data, in particular the mode, median, or (most commonly) the mean, which is calculated by dividing the sum of the values in the set by their number.

Importance of Average

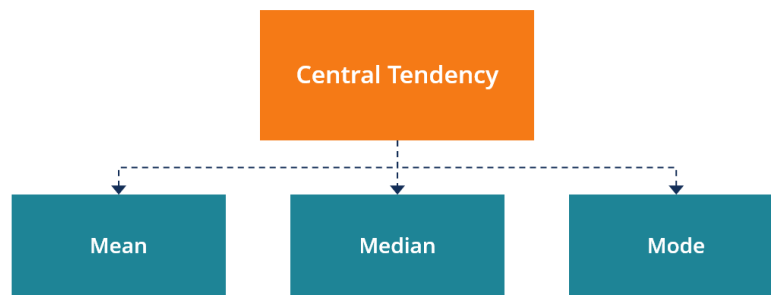
- Summarise a large amount of data into a single value; and
- Indicate that there is some variability around this single value within the original data.

Characteristics of Averages

G.U. Yule has suggested following characteristics

1. Clear definition
2. Unbiased
3. Not Ambiguous
4. Representative Of The Series
5. Unaffected
6. Sampling
6. Absolute number

Measure/Types of Central Tendency



Mean

The statistical mean refers to the mean or average that is used to derive the central tendency of the data in question. It is determined by adding all the data points in a population and then dividing the total by the number of points. The resulting number is known as the mean or the average.

Mean is the most commonly used measure of central tendency. There are different types of mean, viz.

Arithmetic mean,

weighted mean,

geometric mean (GM)

and harmonic mean (HM).

If mentioned without an adjective (as mean), it generally refers to the arithmetic mean.

Arithmetic mean

Arithmetic mean (or, simply, “mean”) is nothing but the average. It is computed by adding all the values in the data set divided by the number of observations in it. If we have the raw data, mean is given by the formula

$$\text{Mean } \bar{X} = \frac{\sum X}{n}$$

Where, \sum (the uppercase Greek letter sigma), X refers to summation, refers to the individual value and n is the number of observations in the sample (sample size).

$$\text{Mean } \bar{X} = \frac{\sum fX}{n}$$

Where, f is the frequency and X is the midpoint of the class interval and n is the number of observations

Merits of Arithmetic Mean

1 Certainty: Arithmetic mean is rigidly defined. So, its value is always definite and certain. Mean can never be biased.

2 Simplicity: Arithmetic mean is easy to calculate and simple to understand.

3 Stability: Arithmetic mean is a relatively stable measure. It is least affected by fluctuations of sampling. It remains the same if samples are drawn on random basis.

4 Based on all the observations of the series: Arithmetic mean is based on all the observations of a series

Therefore, it is the most representative measure.

5 Suitable for algebraic treatment: Arithmetic mean is capable of further algebraic treatment. Because of this attribute, arithmetic mean is extensively used in statistical analysis.

Demerits of Arithmetic Mean

1 Effect of extreme values: Since arithmetic mean is the average of all the values of a series, it is greatly affected by extreme fluctuations. Thus, it is not a true representative value of all the items of the series.

2 Problem in case of incomplete data: Arithmetic mean cannot be calculated unless all the items of the are known series

3 Mean value may not figure in the series: Arithmetic mean value sometimes does not appear in the series. For example, the arithmetic mean of 4, 8, 15 and 21 is 12 but it is not present in the series.

4 Misleading conclusions: Arithmetic mean sometimes provides misleading conclusions.

Types of Arithmetic Mean

Simple Arithmetic Mean

Weighted Arithmetic Mean

Combined Arithmetic Mean

Geometric Mean

The geometric mean is an average that is useful for sets of positive numbers that are interpreted according to their product and not their sum (as is the case with the arithmetic mean); e.g., rates of growth

Or defined as the arithmetic mean of the values taken on a log scale. It is also expressed as the nth root of the product of an observation.

$$\text{Geometric mean (GM)} = \sqrt[n]{(x_1)(x_2) \dots (x_n)}$$

$$\text{Log (GM)} = \frac{\Sigma(\log x)}{n}$$

GM is an appropriate measure when values change exponentially and in case of skewed distribution that can be made symmetrical by a log transformation. GM is more commonly used in microbiological and serological research. One important disadvantage of GM is that it cannot be used if any of the values are zero or negative

Merits of Geometric Mean

It has following merits:

1 It is based on all observations. Arithmetic mean has a bias for higher values whereas Geometric mean has bias for smaller observations.

2 It is not affected much by fluctuations of sampling. It is useful in averaging ratios, percentage rate of increase and decrease between two persons.

3 GM is used when the numbers reflect population counts that are extremely variable, Le., populations of mosquitoes, bacteria, etc.

Demerits of Geometric Mean

It has following demerits:

1 Geometric mean is a mathematical character. It is not easy to understand or to calculate for non-mathematical persons.

2 In any observation (x,x) is zero, Geometric mean would be zero and if any one of the observations is negative, the Geometric mean becomes imaginary.

Harmonic mean

It is the reciprocal of the arithmetic mean of the observations.

$$\text{Harmonic mean (HM)} = \frac{1}{\frac{\Sigma(1/x)}{n}} = \frac{n}{\Sigma(1/x)}$$

Alternatively, the reciprocal of HM is the mean of reciprocals of individual observations.

$$\frac{1}{HM} = \frac{\Sigma(1/x)}{n}$$

HM is appropriate in situations where the reciprocals of values are more useful. HM is used when we want to determine the average sample size of a number of groups, each of which has a different sample size.

Merits of harmonic Mean

It has following merits;

- 1 It is based on all observations.
- 2 It is not affected much by fluctuations of sampling.
- 3 As reciprocal values are involved, it gives greater weightage to smaller observations.

Demerits of Harmonic Mean

It has following demerits:

- 1 It is difficult to understand and calculate for biologists.
- 2 Its value cannot be obtained if any one of the observations is zero.

Median (Me)

If the values of a variable are arranged in ascending or descending order of magnitude, (i.e., ranked) the median is the value that divides the whole data into two equal parts one part having all values smaller than the median value and other part having all the values greater than the median value.

In other words, 50% of the observations will be smaller than the median while the other 50% will be larger than the median.

Thus, median is the value of the middle observation or the mean value of two middle observations. So that half of the data points fall above it and half below it. As a matter of fact, median is called the positional average. Median is calculated differently for ungrouped and grouped data.

If there is an odd number of numbers, the middle one is picked. For example, consider the list of numbers

1, 3, 3, 6, 7, 8, 9

This list contains seven numbers. The median is the fourth of them, which is 6.

If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of the two middle values For example, in the data set

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

the median is the mean of the middle two numbers: this is $(4+5)/2$ which is {4.5} (In more technical terms, this interprets the median as the fully trimmed mid-range). With this

convention, the median can be described in a caseless formula, as follows:

$$\text{median}(a) = \frac{a \left[\frac{l+1}{2} \right] + a \left[\frac{l+1}{2} \right]}{2}$$

Merits of Median

- 1 It is rigidly defined.
- 2 Median is easy to understand and calculate,
3. Median is not affected by extreme observations and is very useful in the case of skewed distribution

Demerits of Median

The demerits of median are:

1. Median cannot be determined in the case of even number of observations. We merely estimate it as the arithmetic mean of the two middle terms.
2. Median is relatively less stable than mean, particularly for small samples since it is affected more by fluctuations of sampling.

Mode (MO OR Mo)

Mode is the most frequently occurring value in a data. It means that for a given data, mode may or may not exist For example, let's observe mode for the following 3 sets of data:

- (a) 10,10,9,8,5, 4, 12, 10 : One mode i.e., 10
- (b)10,10,9,9,12,15,5 Two modes i.e., 10 and 9
- (c) 4,6,7,15,12, 13, 10 : No mode.

We note that first set of data (a) has single mode 10, the second set of data (b) has two modes 10 and 9 and the third set of data (c) has no mode. Set (b) has 10 and 9 as modes because they both occur 2 times and they occur more often than other values.

Graphic Representation of Mode in Frequency Distribution Data

In terms of frequency distribution, mode is the variable at which the curve peaks. Thus, based on the number of peaks in the curve the frequency distribution may be of following three types

1. Unimodal Frequency Distribution: A distribution data having one mode is called unimodal frequency distribution.
2. Bimodal Frequency Distribution: The frequency distribution having two peaks is called bimodal frequency distribution.
3. Multimodal Frequency Distribution: The frequency distribution with more than two peaks is called the multimodal frequency distribution.
4. Antimode: In U-shaped distribution the low point at the middle of the distribution is known as an antimode.

In case of individual series, we just have to inspect the item that occurs most frequently in the distribution. Further, this item is the mode of the series.

Mode for Discrete Series

In discrete series, we have values of items with their corresponding frequencies. In essence, here the value of the item with the highest frequency will be the mode for the distribution.

Mode for Frequency Distribution

for frequency distribution, the method for mode calculation is somewhat different. Here we have to find a modal class. The modal class is the one with the highest frequency value. The class just before the modal class is called the pre-modal class. Whereas, the class just after the modal class is known as the post-modal class. Lastly, the following formula is applied for calculation of mode

$$\text{Mode} = l + h [(f_1 - f_0) / (2f_1 - f_0 - f_2)]$$

Here, l = The lower limit of the modal class

f_1 = Frequency corresponding to the modal class,

f_2 = Frequency corresponding to the post-modal class,

and f_0 = Frequency corresponding to the pre-modal class

Merits of Mode

Mode has following merits:

- Mode is easy to calculate and understand.
- Mode can be calculated from a grouped frequency distribution with open-end classes.

Demerits of Mode

The demerits of mode are:

- Mode is not rigidly defined. It is ill defined if the maximum frequency is repeated or occurs in the very beginning or at the end of the distribution.

Dispersion and Measures of Dispersion

Introduction

Measurements of central tendency or average provide only single representative value for a set of observations in frequency distribution. These do not enable us to draw a full picture of the set. For better information, detailed scatteredness of data is essential. The degree or extent of scatteredness of data around an average or mean is called dispersion or deviation.

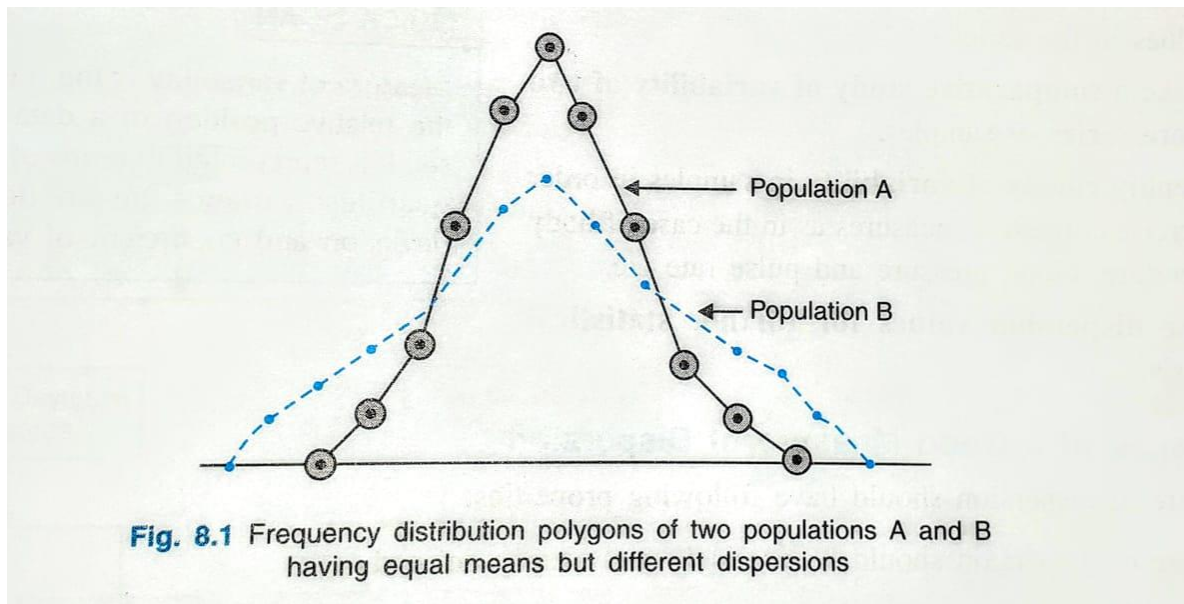
Illustration

The following three sets of data are not identical. Their means are the same but range is different.

Series	1. 60, 60, 60, 60, 60	Mean = 60	Range=60
Series	2. 30, 50, 85, 75, 60	Mean = 60	Range=30-85
Series	3. 10, 60, 90, 90, 50	Mean = 60	Range=10-90

The mean (\bar{X}) for all the three sets is 60 but the observations in each set are different. This variation in data is described as measure of dispersion.

In first set of data there is no dispersion because all the observations are the same. In second dispersion is evident because all the observations are different. The dispersion is small when the values of these observations are close together, i.e., show little variation. The dispersion is great when these values are widely spread out. For example, this Fig. shows the frequency polygons of two populations having equal means but population B is more variable than population A and is more spread out.



Definition of Dispersion

A measure of dispersion reflects how closely the data clusters around the measure of central tendency. It represents the deviation of value of individual observation on either side of the central value in a set of data.

example

Following is the data collected by a poultry breeder about the number of eggs laid in three poultry farms A, B, C.

TABLE 8.1 No. of Eggs laid in Three Poultry Farms A, B and C

Sl. No.	Poultry A Daily Egg Production	Poultry B Daily Egg Production	Poultry C Daily Egg Production
1	4,000	4,050	3,900
2	4,000	4,025	2,100
3	4,000	3,950	1,200
4	4,000	3,835	800
5	4,000	4,140	12,000
Total	20,000	20,000	20,000

$\bar{X} = \frac{20,000}{5} = 4,000$

In the above example, the mean egg production of A, B and C is the same. The mean does not show the fluctuation or variation in the number of eggs produced daily by B and C poultry. The daily egg production by poultry A is the same. It does not show variability. The variability in the egg production of poultry farm B is less than in poultry farm C. In B, the value spreads out between 3,835 and 4,140, whereas in case of poultry farm C, the value is spread out between 800 and 12,000. It means dispersion is more in poultry farm C's egg production than in case of poultry farm B.

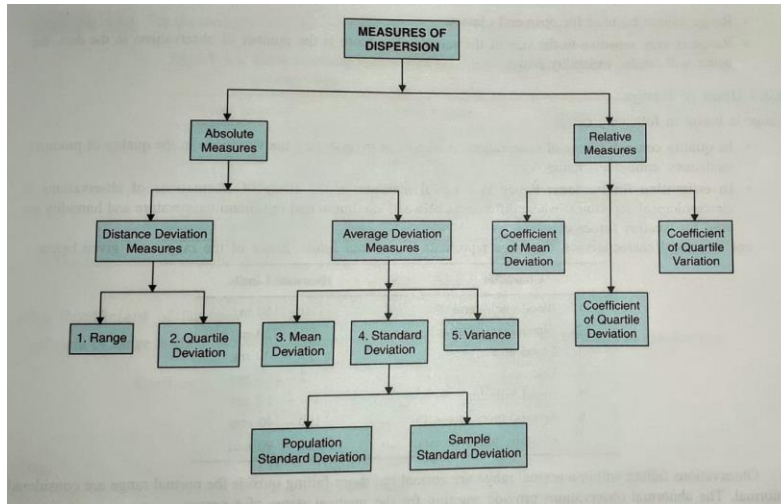
Importance of Dispersion

The measure of variance or dispersion is an important tool in biostatistical studies because biological phenomena are more variable than physical and chemical phenomena. Individual variations are found in haemoglobin percentage, in the number of RBCs and WBCs and even

the cure rate with the same drug varies in different patients of the same age and sex. The major objectives of measure of dispersion are:

1. To judge the reliability of measures of central tendency.
2. To obtain correct picture of distribution or dispersion of values in the series.
3. To make a comparative study of variability of two or more series or samples.
4. To identify causes of variability in samples in order to exercise corrective measures as in the case of body temperature, blood pressure and pulse rate etc.
5. To use dispersion values for further statistical analysis.

Types of Measures of Dispersion Measures of dispersion are of two types:



Absolute Measures of Dispersion

- Absolute measures of dispersion are expressed in the same unit in which observations are given.

- These measures are useful for comparing variation in two or more distributions where units of measurement are the same.

- Absolute measurements are of the following types:

A. Distance deviation measures

B. Average deviation measures

1. Distance Deviation Measures or Measures of Limits: Distance measures use distance of spread between two values in the data set. This distance becomes a measure of variability or measure of dispersion. The larger is the distance between two values the greater is the variability.

2. Average Deviation Measures: These are the average of deviation determined from the measure of central tendency. They are used more commonly for measuring variability or dispersion. Mean deviation, standard deviation and variance are such measures of dispersion.

Relative Measures of Dispersion

These are expressed as the ratio or percentage of or the coefficient of the absolute measure of dispersion. Therefore, relative measures of dispersion are also called coefficient of dispersion. These are pure-unit less numbers. Relative measures are used for comparing variability in two or more distributions having different units of measurements.

RANGE

Definition of Range-One way of measuring variation in a set of values is to compute the range. The range of distribution is the difference between the largest or highest and the

smallest or lowest values in a set of observations. It gives us some idea of the amount of variability present in the data or the spread of the data.

Range can be defined as maximum value minus the minimum value in a data. Range = Largest value in the series of data - Smallest value of that series.

$$R = L-S \text{ or } R = H-L$$

Merits of Range

It is easy to calculate, and easy to understand.

- It is useful in frequency distributions where only two extreme observations are considered.
- It is extensively used in statistical quality control.
- Its units are the same as the units of the variable being measured.

Uses of Range

Range is useful in following cases:

- **In quality control:** Range of observations is important in analysing the variations in the quality of produce medicines, antibiotics, tonics, etc.
- **In estimating fluctuations:** Range is a useful measure in the study of fluctuations of observations in meteorological department where differences between maximum and minimum temperature and humidity used for weather forecasts.

For biological characteristics, the range represents the normal limits. Some of the ranges are given below:

Character	Normal Limits
1. Blood cholesterol	120 – 150 mg
2. Blood sugar (fasting)	80 – 120 mg
3. Blood urea	15 – 40 mg
4. Uric acid	2 – 4 mg
5. Blood bilirubin	0.2 – 1.2 mg
6. Systolic blood pressure	100 – 140 mm
7. Diastolic blood pressure	80 – 90 mm

Observations falling within a normal range are normal but those falling outside the normal range are considered abnormal. The abnormal observations provide warning for the medical status of a person.

MEAN DEVIATION

Every value of the variable differs from the sample mean by some specific amount which is called its deviation. This deviation (d) of an observation (X) is given by the equation:

$$d = X - \bar{X}$$
$$\text{Median deviation (MD)} = \frac{\sum |X - \bar{X}|}{N} \text{ or } \frac{\sum |X|}{N}$$

Definition of Mean Deviation

Mean deviation is an average mean of the deviations of values from central value or central tendency. The central tendency can be arithmetic mean, mode or median. Mean deviation can

be defined as the mean of all the deviations in a given set of data obtained from an average. In case the deviation is greater than the mean, the deviation is positive, but if it is less than the mean, the deviation is negative. The deviation is calculated by adding the deviation of individual observations from their arithmetic mean without their regard to the sign and dividing the sum by the number of observations. The mean deviation can be calculated about any of the three averages, i.e., mean (M), median (Me) or mode (Mo).

Merits of Mean Deviation

- It is easy to understand and calculate
- It is not affected much by the extreme values.

Demerits of Mean Deviation

- It cannot be computed for distributions with open end classes.
- It is less reliable because algebraic signs are ignored.

Uses of Mean Deviation -Mean deviation and its coefficient are used in study and research in the field of medicine, microbiology, pharmacology, economic, social sciences and business.

STANDARD DEVIATION (SD)

Definition of Standard Deviation-Standard deviation of a series is the positive square root of the arithmetic mean of the squares of deviations of the various items from the arithmetic mean of the series. It is also called root mean square deviation. It is represented by Greek symbol (s) and in short form by SD. It represents the extent to which individual values differ from the average or mean.

Merits of Standard Deviation

- It summarises in one figure the deviation of a large distribution from mean.
- It is most reliable and dependable measure of dispersion.

Demerits of Standard Deviation

- (i) It gives more weightage to extreme values and less to the values that are closer to mean.
- (ii) The process of squaring deviations and then taking square root involves lengthy calculation. Hence, its calculation is not easy.

COEFFICIENT OF VARIABILITY OR COEFFICIENT OF VARIANCE

Definition-It is measurement of relative dispersion or relative variability. It gives an idea about the extent to which varieties of a character in two or more different series are scattered around the central value. Therefore, two or more distributions having same central values can be compared directly with the help of various measures of dispersion.

When we calculate standard deviation of two populations having different mean, the standard deviation alone does not give us correct comparative idea about the extent of variability in two populations. If this standard deviation is expressed as percentage of mean, such a parameter/statistics provides a comparative idea of the extent of variability.

STANDARD ERROR (SE)

Definition- It is a statistical term that measures the accuracy which a sample represents a population. It means SE measures chance deviation and not an error or mistake. Theoretically, the deviation of a sample mean from actual mean of a population is the standard error because sample mean and population is expected to be zero. But in biological experiments

standard error is never zero.

$$\text{Standard error (SE)} = \frac{\text{Standard deviation } (\sigma)}{\sqrt{\text{Sample size } (n)}}$$

i.e.,

$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$

Ex. Suppose in a poultry farm table (Population) of hen is 1500. The daily average egg laying capacity of all the hens is 105. Of these, 200 hens (sample) were selected randomly from the population. Per day average number of egg laying by the samples (200 hens) comes to 100. The difference between observed mean (105) and expected mean (100) comes to (105 - 100) = 5. This non-chance difference in the two means is called standard error.

CORRELATION (Relationship between Continuous Bivariate Variables)

Introduction

The statistical methods studied so far, like measures of central tendency, averages, percentages, measures of dispersion and skewness, etc., are associated with single variate only. In many cases, two continuous characters are measured in the same group of persons or variations of the same variable are studied in different groups of persons to find out how a change in the value of one variable affects or influences the value of other variable. For example, we may like to study relationship between height and weight, blood pressure and age, food type and weight gain, intensity of stimulus and reaction intensity or between price and demand, wages and price index, etc. or we may measure tallness or intelligence among the siblings. The study of such relationships or associations between two variables is called statistical relationship. Special methods have been developed to determine the existence of statistical relationship from bivariate data. These are correlation analysis and regression analysis.

CORRELATION ANALYSIS

Definition According to Croxton and Cowden correlation is the degree of association or strength of relationship between two qualitatively measured or continuous variables. According to Connor if two or more variables (X and Y) vary in sympathy and the movements in one tend to be accompanied by corresponding movements (variations) in the other. then these two variables (X and Y) are said to be correlated. In case Y remains unaffected by change in X, both X and Y are uncorrelated. The concept of correlation analysis and term correlation originated with Galton in 1888.

Correlation can be of following types.

Depending on its extent and direction, the correlation between two variables may be of following

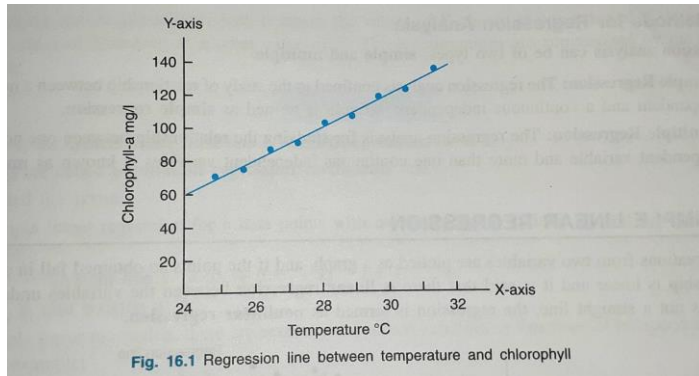
A Positive and Negative Correlations

B Linear and Non-Linear Correlations

C Simple, Partial and Multiple Correlations

REGRESSION OR REGRESSION ANALYSIS

Definition Regression is a statistical study for determining the strength of relationship or association between a dependent variable (Y) and an independent variable (X) for a given population. The relationship is expressed either as an equation for a line or a curve and is described as regression line or regression curve. The steepness and direction of regression lines are usually represented by numbers. These numbers are called regression coefficients.



Main Features of Regression Analysis

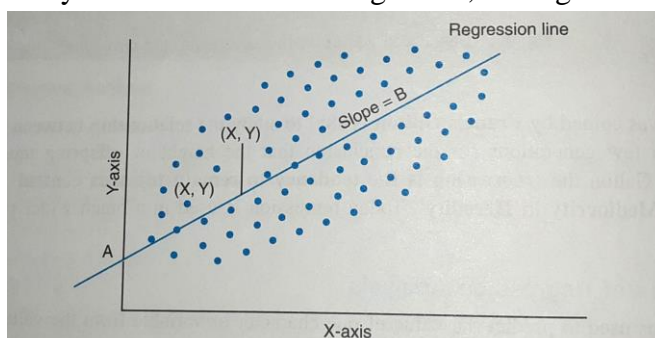
The regression analysis is used to predict the value of one character or variable from the value of the other character or variable. According to this, the two variables can be identified as:

- (a) **Dependent Variable:** The variable whose value is influenced or is to be predicted, is called a dependent variable and is represented by Y.
- (b) **Independent Variable:** The variable which influences the values is called an independent

Applications of Regression

1. The regression analysis is used to find out the measures of error present during the use of regression line for prediction. For this, standard error (SE) of estimate is calculated.
2. From the value of coefficient of correlation, one can find the degree of association between two variables.

SIMPLE LINEAR REGRESSION-When observations from two variables are plotted as a graph, and if the points so obtained fall in a straight line, the under study. How the relationship is linear and it is said that there is linear regression between the variables under study if the line is not a straight line, the regression is termed as nonlinear regression.



MULTIPLE REGRESSION ANALYSIS-Multiple regression analysis is a powerful technique used for predicting the unknown value of a dependent variable from the known values of two or more independent variables. The dependent variable is also called criterion or indicator and independent variable as predictor

NONLINEAR REGRESSION-Nonlinear regression is a form of regression analysis that helps in describing nonlinear relationships in the dependent (response) variable and independent (predictor) variables. The interaction of dependent variable may be with one or more independent variables.

References

Book on Biostatistics (3rd Edition) by Veer Bala Rastogi

Book on Fundamentals of Biostatistics 8th Edition by Bernard Rosner

INTRODUCTION TO BIOSTATISTICS SECOND EDITION by Robert R. Sokal and F. James Rohlf

Statistical inference- Hypothesis testing, Significance level, Test of significance for large and small samples; Parametric tests; Non parametric tests; Experimental design, Use of biostatistic softwares SPSS and sigma plot.

Contents

Statistical interference

Hypothesis testing

Significance level

Test of significance

Parametric test

Non parametric tests

SPSS

Sigma plot

Introduction:

Computational biology is an interdisciplinary field that develops and applies computational methods to analyse large collections of biological data, such as genetic sequences, cell population or protein sample to make new predictions or discovery new biology. The computational methods used include analytical methods, mathematical modelling and simulation

It is the interface between computer and biology. In other words it is the application for information technology in the study of biology.

It is used for analyst of data method related to genomics proteomics metabolomics and other biological aspects

It serves as bridge between observations (data) in diverse biologically related discipline and derivations of understanding about how the systems or processed functions and subsequently the applications.

In this field using of different methods for analysing date that include statistical interference and different types of applications.

Statistical interference: The main objective of sampling is to draw conclusions about the unknown population from the information provided by a sample. This is called statistical inference.

Statistical inference may be of two kinds: parameter estimation and Hypothesis testing.

Parameter estimation: Parameter estimation is concerned with obtaining numerical values of the parameter from a sample

Eg: A company may be interested in estimating the share of the population who are aware of its product.

HYPOTHESIS TESTING: Hypothesis testing is an essential procedure in statistics. A **hypothesis test** evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.

HYPOTHESIS: Is beginning with some assumptions about true value of the population parameters.

ESTIMATION: in estimation we use some formula in which substitutes the values of parameters of sample in order to obtain the numerical value of sample.

❖ Estimation done after hypothesis.

Test hypothesis contain six major steps to obtain the conclusion about the sample that include,

1. Hypotheses

- i. State in order:
- ii. Research Hypothesis
- iii. Null Hypothesis
- iv. Alternate Hypothesis (We should take the above hypothesis according to our interests)

2. Assumptions include:

- i. **Measurement level of data,**
- ii. **Distributions underlying the data,**
- iii. **Knowledge or lack of about population characteristics**
- iv. **Sample size and method,**
- v. **Sample characteristics necessary for applying the test statistic,**
- vi. **Level of significance for testing**
- vii. **Test statistics (or Confidence Interval Structure)**
- viii. **Structure to be used to test significance levels or set of confidence intervals (be sure to include the equations & notation)**
- ix. **Special conditions to be met by statistic**
- x. **Rejection region (or Probability Statement)**
- xi. **Expected measure of the test statistic as generated from tables or critical value for a confidence interval.**

3. Calculations (Annotated Spreadsheet)

4. Conclusions

Types of test hypothesis: Normality: tests for normal distribution in a population sample.

T-test: tests for a Student's t-distribution – ie, in a normally distributed population where standard deviation is unknown and sample size is comparatively small. Paired t-tests compare two samples.

Chi-Square Test for Independence: tests for an association of significance between two categorical variables in a population sample. Typically used with random sampling.

Homogeneity of Variance (HOV): tests the similarity of dispersion parameters in two or more population samples.

Analysis of Variance (ANOVA): tests for and analyzes differences between the means in several groups. Often used similarly to a t-test, but for more than two groups.

Mood's Median: compares the medians of two or more population samples.

Welch's T-test: tests for equality of means between two population samples. Also known as Welch's unequal variances t-test.

Kruskal-Wallis H Test: compares two or more groups with an independent variable, based on a dependent variable. Also known as one-way ANOVA on ranks.

Box-Cox Power Transformation: transforms a data set into normal distributions

Application of test hypothesis: used to detect the possibility of the sample(YES/NO).

Null hypothesis: is the first step in testing hypothesis

Set up such that it conveys a meaning that there exist no difference between the different samples.

Eg: null hypothesis – the mean pulse rate among the two groups are same or there is no significance difference between their pulse rate.

Significance value: P- Value is function of observed sample result (STASTIC) that use for testing a statistical hypothesis.

It is probability of null hypothesis being true, it can accept or reject null hypothesis based on P-value.

Practically, p- value $<0.05(5\%)$ is considered significance.

$P=0.05$ implies

It may go wrong 5 out of 100 by rejecting null hypothesis.

We can attribute significance with 95% confidence.

Why test of significance done : to assist the administration and clinicians in making decision.

- 1) THE DIFFERENCE IS REAL.
- 2) HAS IT HAPPEN BY CHANCE

CLASSIFICATION OF TEST SIGNIFICANCE

For Qualitative data:-

1. Standard error of difference between 2 proportions (SE_{p1-p2})
2. Chi-square test or X^2

For Quantitative data:-

1. Unpaired (student) 't' test
2. Paired 't' test
3. ANOVA

• By using various tests of significance we either: –Reject the Null Hypothesis (or) –Accept the Null Hypothesis

• Rejecting null hypothesis → difference is significant.

• Accepting null hypothesis → difference is not significant.

Statistical tests are intended to decide whether a hypothesis about distribution of one or more populations or samples should be rejected or accepted.

Statistical Tests are parametric tests and Non – Parametric tests

PARAMETRIC TEST: is a statistical test that makes assumptions about the parameters of the population distribution(s) from which one's data is drawn.

APPLICATIONS

- Used for Quantitative data.
- Used for continuous variables.
- Used when data are measured on approximate interval or ratio scales of measurement.
- Data should follow normal distribution.

PARAMETRIC tests: 1) Analysis of Variance (ANOVA) is a collection of statistical models used to analyze the differences between group means or variances

- Compares multiple groups at one time
- Developed by R.A.Fisher

Anova are two types 1) One way ANOVA

2) Two way ANOVA

One way ANOVA: Compares two or more unmatched groups when data are categorized in one factor Ex: 1. Comparing a control group with three different doses of aspirin

2. Comparing the productivity of three or more employees based on working hours in a company

Two way ANOVA:

• Used to determine the effect of two nominal predictor variables on a continuous outcome variable.

• It analyses the effect of the independent variables on the expected outcome along with their relationship to the outcome itself. Ex: Comparing the employee productivity based on the working hours and working conditions.

Assumptions of ANOVA:

- The samples are independent and selected randomly.
- Parent population from which samples are taken is of normal distribution.

- Various treatment and environmental effects are additive in nature.
 - The experimental errors are distributed normally with mean zero and variance σ^2 .
- 2) Z-Test: Z-test is a statistical test where normal distribution is applied and is basically used for dealing with problems relating to large samples when the frequency is greater than or equal to 30.

• It is used when population standard deviation is known.

Assumptions: • Population is normally distributed

- The sample is drawn at random

Conditions:

- Population standard deviation σ is known
- Size of the sample is large (say $n > 30$)

3) Student's t-test: • Developed by Prof. W.S. Gossett

• A t-test compares the difference between two means of different groups to determine whether the difference is statistically significant.

Types of t- tests:

1) One Sample t-test

Assumptions:

- Population is normally distributed
- Sample is drawn from the population and it should be random
- We should know the population mean

Conditions:

- Population standard deviation is not known
- Size of the sample is small (< 30)

Two sample t-test:

• Used when the two independent random samples come from the normal populations having unknown or same variance.

Assumptions:

1. Populations are distributed normally
2. Samples are drawn independently and at random

Conditions:

1. Standard deviations in the populations are same and not known
2. Size of the sample is small

4) Paired t-test: Used when measurements are taken from the same subject before and after some manipulation or treatment.

Ex: To determine the significance of a difference in blood pressure before and after administration of an experimental pressure substance.

Assumptions: 1. Populations are distributed normally

2. Samples are drawn independently and at random

Conditions: 1. Samples are related with each other

2. Sizes of the samples are small and equal

3. Standard deviations in the populations are equal and not known

Non parametric tests: Also known as distribution-free tests because they are based on fewer assumptions (e.g., they do not assume that the outcome is approximately normally distributed).

Non-parametric Methods: • Sign Test

- Wilcoxon Signed-Rank Test
- Mann-Whitney-Wilcoxon Test
- Kruskal-Wallis Test

1) Sign Test: A common application of the sign test involves using a sample of n potential customers to identify a preference for one of two brands of a product.

The objective is to determine whether there is a difference in preference between the two items being compared.

To record the preference data, we use a plus sign if the individual prefers one brand and a minus sign if the individual prefers the other brand.

Because the data are recorded as plus and minus signs, this test is called the sign test.

2) Wilcoxon Signed-Rank Test: • The Wilcoxon test is used when we are unwilling to make assumptions about the form of the underlying population probability distributions,

- but we want compare paired samples.
- Analogous to the dependent t-test we are interested in the difference
- in two measurements taken from each person.
- The rank sum of the positive (T+) and negative (T-) differences are calculated, the smallest of these is used as the test statistic to test the hypothesis.
- Two assumptions underlie the use of this technique. 1. The paired data are selected randomly.
- 2. The underlying distributions are symmetrical.

SPSS: Statistical Package for the Social Sciences

- ◆ SPSS is a comprehensive and flexible statistical analysis and data management solution.
- ◆ SPSS is a computer program used for survey authoring and deployment, data mining, text analytics, statistical analysis, and collaboration and deployment.

SPSS can take data from almost any type of file and use them to generate tabulated reports, charts, and plots of distributions and trends, descriptive statistics, and conduct complex statistical analyses.

- ◆ SPSS is among the most widely used programs for statistical analysis in social science.

Features of SPSS: ◆ It is easy to learn and use.

- ◆ It includes a full range of data. Management system and editing tools.
- ◆ It provides in-depth statistical capabilities.
- ◆ It offers complete plotting, reporting and presentation features.

Getting data into SPSS: ◆ Creating new SPSS data files

- ◆ Opening existing SPSS system files
- ◆ Importing data from an ASCII file
- ◆ Importing data from other file formats

Entering Data: ◆ The data editor offers a simple and efficient spreadsheet like facility for entering data and browsing the working data file.

- ◆ This window displays the content of the data file.
- ◆ One can create new data files or modify existing ones.
- ◆ One can have only one data file open at a time.
- ◆ This editor provides two views of the data

Data view: ◆ Displays the actual data values or defined value labels.

Variable view: ◆ Displays variable definition information, including defined variable and value labels, data type, etc.,

Saving Data: ◆ We need to save it and give it a name. The default extension name for saving files is '.save'.

◆ Ex. SSPS.sav

◆ Also we can able to retrieving already saved file

BASIC STEPS IN DATA ANALYSIS:

a) Get Your Data into SPSS: We can open a previously saved SPSS data file, read a spreadsheet, database, or text data file, or enter directly in the data editor.

b) Select a Procedure: Select a procedure from the menus to calculate statistics or to create a chart.

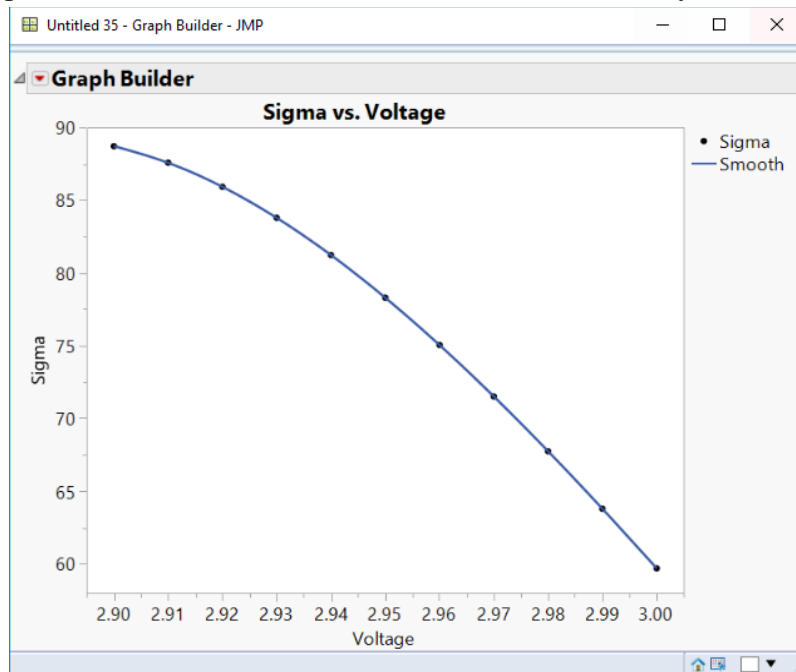
c) Select the Variable for the Analysis: Variables in the data file are displayed in a dialog box for the procedure

d) Run the Procedure: Results are displayed in the viewer.

Results can be obtained in the form of linear, chart, bar, diagrams.

Sigma plot: Sigma Plot is proprietary software used for scientific graphing and data analysis. It works on Microsoft windows.

The software can read multiple formats, such as Microsoft excels spreadsheets, and can also perform mathematical transforms and statistical analysis.



This is example of sigma plot which drawn on MS excel spreadsheet.

Unit III

Bioinformatics basics; Application and research; Present global bioinformatics scenario. Databases- characteristic of bioinformatics databases, Sequence databases- nucleotide and protein sequence databases. Tools- Need for tools, data mining tools, data submission tools e.g. nucleotide submission tools and protein sequence submission tools; Data analysis tools- nucleotide sequence analysis and protein sequence analysis tools e.g. BLAST & FASTA.

CONTENTS

1. Introduction to the bioinformatics
2. Applications of bioinformatics
3. Research and trends in bioinformatics
4. Tools used in the bioinformatics
5. Need for the tools
6. Conclusion

INTRODUCTION TO THE FIELD OF BIOINFORMATICS

Bioinformatics is the collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied to molecular genetics and genomics. The term bioinformatics was coined by Hwa Lim in the late 1980s, and popularized in the 1990s through its association with the human genome project. It is a young field that is still defining itself. The term commonly refers to computational work in genomics, the many other new ‘omics’ that are sprouting, and neighboring research areas. The precise boundaries of the field are elusive, but as a general rule the closer a research area is to genomics, the more likely its computational aspects will be labeled bioinformatics. To muddy the waters further, some people distinguish bioinformatics from computational biology, using the latter to denote computational work whose agenda is clearly biological; it is often difficult to draw this distinction in practice, as much of what is done in the field is interdisciplinary and combines computational and biological expertise. Bioinformatics is both an engineering art and a science. It encompasses the development of new computational methods and the application of those methods to solve biological problems. It also has a large service component in which computational resources, such as databases, are operated for the benefit of the research community. Bioinformatics is a broad field that has a central role in many areas of biological research. These include genomics and, more specifically, genomic sequencing and mapping, genome annotation, and comparisons of multiple genomes. Bioinformatics is also essential in transcriptomics — the study of transcribed sequences, both full-length cDNAs and expressed sequence tags (ESTs) — and the analysis of gene expression data typically measured using DNA microarrays or some form of sample sequencing. It is also crucial in proteomics for the analysis of protein sequences (e.g. to determine functional motifs), for the study of protein abundance (typically measured using two-dimensional gels or mass spectrometry), and the determination of protein structure either empirically or computationally. Bioinformatics is key in the analysis of protein–protein interactions and molecular pathways (the ‘interactome’) and in systematic studies of gene

regulation (the 'regulome'). It also plays a vital role in genetics, both in the discovery of new molecular genetic markers, such as single nucleotide polymorphisms, and the use of these and other markers to dissect the genetic basis of disease and other phenotypes. Bioinformatics is also essential in studies of evolution and phylogeny. Bioinformatics touches a wide range of biological research areas. Although there are scientific commonalities across these areas, there are also major differences that affect bioinformatics methods. For example, the problems of analyzing genomic sequences, transcribed sequences and protein sequences are similar in that they can all be described mathematically as sequences of letters, and all are subject to mutational pressures (hence diverge in predictable ways across taxa). Beyond this, however, the methods and issues are quite different. With genomic sequences, a key issue is gene prediction. With transcribed sequences, a key issue is clustering of redundant sequences to coalesce all sequences that belong to the same gene. For protein sequences, key issues are discovering functional motifs that are conserved across evolution, and the use of these motifs to functionally classify novel sequences. The term bioinformatics is the combination of biology and information developed in wake of the generation of the amino-acid sequences of proteins and nucleotide sequences of DNA. Bioinformatics term was given by Kem Helper in 1970. In 1962, Zuckerkandle and Paulings proposed amino acid sequences and used to study the evolutionary relationships among organisms. This proposed a new field of study called Molecular Evolution where a number of molecules contribute for analysis of amino acids sequences of functionally related proteins. The first comprehensive collection of amino acid sequences was compiled in the Atlas of Protein Sequence and Structure by National Biomedical Research Foundation. The collection was edited by Margaret.O.Dayhoff from 1965 to 1978. Dayhoof and co-workers also contributed to comparison of amino acid by developing computer Softwares. The European Molecular Biology Laboratory established their data library in 1980 to collect organized and distribute nucleotide sequence data and related information. The function is now performed by European Bioinformatics Institute (EBI). During early 1980s The National Center for Bioinformatics Information (NCBI) was established in USA. The National Bionuclear Research Foundation established protein information resource in 1984. The management, analysis and rapid accumulation of sequenced data requires new computer softwares and statistical methods. As a result variety of methods and tools have been developed cubicle greatly facilitate management, utilization and dissemination of biological information. By 1977, a method for sequencing DNA was discovered and the first genetic engineering company, Genetech was founded. With the help of the Joseph Sambrook who refined DNA electrophoresis using agarose gel along with Herbert Boyer and Stanley Cohen who invented DNA cloning. By 1981, 579 human genes had been mapped and insitu hybridization had become a standard method for mapping. Marvin Carruthers and Leory Hood made a huge leap in bioinformatics when they invented a method for automated DNA sequencing. In 1988, the Human Genome Organization (HUGO) was founded. This is an international organization of scientists involved in Human Genome Project. In 1989, the first complete genome map was published of the bacteria *Haemophilus influenza*.

Bioinformatics was fuelled by the need to create huge databases, such as GenBank and EMBL and DNA Database of Japan to store and compare the DNA sequence data erupting from the human genome and other genome sequencing projects. Today, bioinformatics

embraces protein structure analysis, gene and protein functional information, data from patients, pre-clinical and clinical trials, and the metabolic pathways of numerous species. The field is further broadened by the need for computation at several steps along the path from data generation to biological interpretation. A hallmark of omic biology is the use of powerful laboratory technologies to generate large, systematically collected datasets. The people who produce the data are experts in these technologies, whereas those who derive biological knowledge from the data are experts in specific biological problems — particular diseases, pathways, and so on — and make discoveries by combining data obtained using many different methods. A key challenge of bioinformatics is to bridge the considerable gap between technical data production and its use by scientists for biological discovery. Bioinformatics approaches in the search for natural products are a combination of molecular and chemical techniques. Important criteria of molecular approaches include phylogenetic resolution and potential to a large-scale screening. Application of comparative genome sequence analysis is essential for a better understanding of the genetic and epigenetic components of different bacterial taxa. With the increased numbers of fully sequenced microbial genomes, including those of well-known bacterial producers of natural products, it has become clear that the genomic and metabolic capacity of these microorganisms is much higher than initially anticipated. This is due to the discovery of ‘silent’ or ‘cryptic’ secondary metabolite gene clusters that encode the production of additional, unidentified compounds.

Recent examination of massive sequencing (metagenomics) approaches to analyze the composition of bacterial communities of complex milieu including sea water, provide an abundant source of molecular sequence data for analysis. These data are useful in comparative genome analyses to identify genes directly involved, for example, in nonribosomal peptide synthesis (peptide synthetase), modifying enzymes, or other genes coding the production of certain natural products. Often, the complete set of specific genes involved in the synthesis of a particular natural product is contained in a single operon. For example, as the presence of conserved sequence motifs and a modular organization of nonribosomal peptide synthetases often assembled into single bacterial operons, a specific sequence search algorithm can be developed to screen public database resources. This enables a detailed analysis of evolutionary, structural, and functional aspect of natural products production based on the comparison of molecular sequences, molecular modeling, and simulation. For example, the situation of genomic colinearity of modular synthetase components might also facilitate the identification of the molecular components of natural products production as well as the reconstruction of natural products synthesis pathways. This will permit to clarify the details of natural production systems and may allow the simulation of these pathways to explore possible strategies for the optimization or engineering of natural product production systems.

Despite the enormous flexibility of genomes, the corresponding metabolic synthesis networks follow specific inherent rules that are responsible for their rigidity.³⁹ Evolutionary designed strategies are ideally suited to utilize this genomic flexibility to adapt desired phenotypes to balance the metabolic network states required for optimal performance. The identification of genes involved in the metabolic synthesis of natural products by genome sequence analysis can be complemented by the analyses and modeling of natural products

production. Bioinformatics tools for the construction of metabolic networks from genome sequence (e.g., Pathway Tools developed by Karp and coworkers at the bioinformatics research group at SRI International (<http://www.sri.com>) and information from the literature can be used to infer and describe natural products synthesis pathways and analyze the production machinery of bacterial producers. It is generally recognized, particularly in systems responsible for the synthesis of diverse antibiotics, that, for example, nonribosomal peptide synthesis occurs within a molecular complex composed of modules or subunits grouping peptide synthetase modules and associated enzymatic activities. Bioinformatics has become a buzzword in the post-genomic era. However, the discipline is not new. This chapter introduces the three scientists whose initiatives 50 years ago led to the birth of the science of bioinformatics, and briefly discusses their contributions. Although at the time it was not called bioinformatics, the application of computers in protein-sequence analysis and tracing protein evolution was the rudimentary form of contemporary bioinformatics. These three scientists were Margaret Dayhoff, Richard Eck, and Robert Ledley. Of these, Margaret Dayhoff's contributions stood out the most and she is often credited as being the pioneer in bioinformatics for her varied contributions, including developing the first amino-acid substitution matrix for studying protein evolution.

APPLICATIONS OF BIOINFORMATICS:-

The branch of bioinformatics marks the advances in the biological sciences and biotechnology which are largely benefitted by the bioinformatics. The best example is the sequencing of the human genome in record time which would not have been possible without the help of bioinformatics. Some of the applications of bioinformatics are as follows:

1. Molecular medicine

- The human genome will have profound effects on the fields of biomedical research and clinical medicine.
- The completion of the human genome and the use of bioinformatics tools means that we can search for the genes directly associated with different diseases and begin to understand the molecular basis of these diseases more clearly.
- This new knowledge of the molecular mechanisms of disease will enable better treatments, cures and even preventative tests to be developed.

2. Personalized medicine

- Clinical medicine will become more personalized with the development of the field of pharmacogenomics.
- This is the study of how an individual's genetic inheritance affects the body's response to drugs.
- Today, doctors have to use trial and error to find the best drug to treat a particular patient as those with the same clinical symptoms can show a wide range of responses to the same treatment.
- In the future, doctors will be able to analyze a patient's genetic profile and prescribe the best available drug therapy and dosage from the beginning.

3. Preventative medicine:

With the specific details of the genetic mechanisms of diseases being unraveled, the development of diagnostic tests to measure a person's susceptibility to different diseases may become a distinct reality.

4. Gene therapy

- In the not too distant future with the use of bioinformatics tool, the potential for using genes themselves to treat disease may become a reality.
- Gene therapy is the approach used to treat, cure or even prevent disease by changing the expression of a person's genes.

5. Drug development

- At present all drugs on the market target only about 500 proteins.
- With an improved understanding of disease mechanisms and using computational tools to identify and validate new drug targets, more specific medicines that act on the cause, not merely the symptoms, of the disease can be developed.
- These highly specific drugs promise to have fewer side effects than many of today's medicines.

6. Microbial genome applications

- The arrival of the complete genome sequences and their potential to provide a greater insight into the microbial world and its capacities could have broad and far reaching implications for environment, health, energy and industrial applications.
- For these reasons, in 1994, the US Department of Energy (DOE) initiated the MGP (Microbial Genome Project) to sequence genomes of bacteria useful in energy production, environmental cleanup, industrial processing and toxic waste reduction.
- By studying the genetic material of these organisms, scientists can begin to understand these microbes at a very fundamental level and isolate the genes that give them their unique abilities to survive under extreme conditions.

7. Waste cleanup

- *Deinococcus radiodurans* is known as the world's toughest bacteria and it is the most radiation resistant organism known.
- Scientists are interested in this organism because of its potential usefulness in cleaning up waste sites that contain radiation and toxic chemicals.

8. Climate change Studies

- Increasing levels of carbon dioxide emission, mainly through the expanding use of fossil fuels for energy, are thought to contribute to global climate change.
- Recently, the DOE (Department of Energy, USA) launched a program to decrease atmospheric carbon dioxide levels.
- One method of doing so is to study the genomes of microbes that use carbon dioxide as their sole carbon source.

9. Alternative energy sources

- Scientists are studying the genome of the microbe *Chlorobium tepidum* which has an unusual capacity for generating energy from light

10. Biotechnology

- The archaeon *Archaeoglobus fulgidus* and the bacterium *Thermotoga maritima* have potential for practical applications in industry and government-funded environmental remediation.

- These microorganisms thrive in water temperatures above the boiling point and therefore may provide the DOE, the Department of Defence, and private companies with heat-stable enzymes suitable for use in industrial processes
- Other industrially useful microbes include, *Corynebacterium glutamicum* which is of high industrial interest as a research object because it is used by the chemical industry for the biotechnological production of the amino acid lysine.
- The substance is employed as a source of protein in animal nutrition.
- Biotechnologically produced lysine is added to feed concentrates as a source of protein, and is an alternative to soybeans or meat and bonemeal.
- *Lactococcus lactis* is one of the most important micro-organisms involved in the dairy industry.
- Researchers anticipate that understanding the physiology and genetic make-up of this bacterium will prove invaluable for food manufacturers as well as the pharmaceutical industry, which is exploring the capacity of *lactis* to serve as a vehicle for delivering drugs.

11. Antibiotic resistance

- Scientists have been examining the genome of *Enterococcus faecalis*-a leading cause of bacterial infection among hospital patients.
- They have discovered a virulence region made up of a number of antibiotic-resistant genes that may contribute to the bacterium's transformation from a harmless gut bacteria to a menacing invader.
- The discovery of the region, known as a pathogenicity island, could provide useful markers for detecting pathogenic strains and help to establish controls to prevent the spread of infection in wards.

12. Forensic analysis of microbes

- Scientists used their genomic tools to help distinguish between the strain of *Bacillus anthracis* that was used in the summer of 2001 terrorist attack in Florida with that of closely related anthrax strains.

13. The reality of bioweapon creation

- Scientists have recently built the virus poliomyelitis using entirely artificial means.
- They did this using genomic data available on the Internet and materials from a mail-order chemical supply.
- The research was financed by the US Department of Defence as part of a biowarfare response program to prove to the world the reality of bioweapons.
- The researchers also hope their work will discourage officials from ever relaxing programs of immunisation.
- This project has been met with very mixed feelings.

14. Evolutionary studies

- The sequencing of genomes from all three domains of life, eukaryota, bacteria and archaea means that evolutionary studies can be performed in a quest to determine the tree of life and the last universal common ancestor.

15. Crop improvement

- Comparative genetics of the plant genomes has shown that the organisation of their genes has remained more conserved over evolutionary time than was previously believed.
- These findings suggest that information obtained from the model crop systems can be used to suggest improvements to other food crops.
- At present the complete genomes of *Arabidopsis thaliana* (water cress) and *Oryza sativa* (rice) are available.

16. Insect resistance

- Genes from *Bacillus thuringiensis* that can control a number of serious pests have been successfully transferred to cotton, maize and potatoes.
- This new ability of the plants to resist insect attack means that the amount of insecticides being used can be reduced and hence the nutritional quality of the crops is increased.

17. Improve nutritional quality

- Scientists have recently succeeded in transferring genes into rice to increase levels of Vitamin A, iron and other micronutrients.
- This work could have a profound impact in reducing occurrences of blindness and anaemia caused by deficiencies in Vitamin A and iron respectively.
- Scientists have inserted a gene from yeast into the tomato, and the result is a plant whose fruit stays longer on the vine and has an extended shelf life.

18. Development of Drought resistance varieties

- Progress has been made in developing cereal varieties that have a greater tolerance for soil alkalinity, free aluminium and iron toxicities.
- These varieties will allow agriculture to succeed in poorer soil areas, thus adding more land to the global production base.
- Research is also in progress to produce crop varieties capable of tolerating reduced water conditions.

19. Veterinary Science

- Sequencing projects of many farm animals including cows, pigs and sheep are now well under way in the hope that a better understanding of the biology of these organisms will have huge impacts for improving the production and health of livestock and ultimately have benefits for human nutrition.

RESEARCH AND TRENDS IN BIOINFORMATICS

Bioinformatics uses advances in the area of computer science, information science, computer and information technology, communication technology to solve complex problems in life sciences and particularly in biotechnology. Data capture, data warehousing and data mining have become major issues for biotechnologists and biological scientists due to sudden growth in quantitative data in biology such as complete genomes of biological species including human genome, protein sequences, protein 3-D structures, metabolic pathways databases, cell line & hybridoma information, biodiversity related information. Advancements in information technology, particularly the Internet, are being used to gather and access ever-increasing information in biology and biotechnology. Functional genomics, proteomics, discovery of new drugs and vaccines, molecular diagnostic kits and pharmacogenomics are

some of the areas in which bioinformatics has become an integral part of Research & Development. The knowledge of multimedia databases, tools to carry out data analysis and modeling of molecules and biological systems on computer workstations as well as in a network environment has become essential for any student of Bioinformatics. Bioinformatics, the multidisciplinary area, has grown so much that one divides it into molecular bioinformatics, organalle bioinformatics and species bioinformatics. Issues related to biodiversity and environment, cloning of higher animals such as Dolly and Polly, tissue culture and cloning of plants have brought out that Bioinformatics is not only a support branch of science but is also a subject that directs future course of research in biotechnology and life sciences. The importance and usefulness of Bioinformatics is realized in last few years by many industries. Therefore, large Bioinformatics R & D divisions are being established in many pharmaceutical companies, biotechnology companies and even in other conventional industry dealing with biological. Bioinformatics is thus rated as number one career in the field of biosciences.

The potential of Bioinformatics in the identification of useful genes leading to the development of new gene products, drug discovery and drug development has led to a paradigm shift in biology and biotechnology-these fields are becoming more & more computationally intensive. The new paradigm, now emerging, is that all the genes will be known "in the sense of being resident in database available electronically", and the starting point of biological investigation will be theoretical and a scientist will begin with a theoretical conjecture and only then turning to experiment to follow or test the hypothesis. With a much deep understanding of the biological processes at the molecular level, the Bioinformatics scientist have developed new techniques to analyze genes on an industrial scale resulting in a new area of science known as 'Genomics'.

The shift from gene biology has resulted in the development of strategies-from lab techniques to computer programs to analyze whole batch of genes at once. Genomics is revolutionizing drug development, gene therapy, and our entire approach to health care and human medicine.

The genomic discoveries are getting translated in to practical biomedical results through Bioinformatics applications. Work on proteomics and genomics will continue using highly sophisticated software tools and data networks that can carry multimedia databases. Thus, the research will be in the development of multimedia databases in various areas of life sciences and biotechnology. There will be an urgent need for development of software tools for data mining, analysis and modeling, and downstream processing. Security of data, data transfer and data compression, auto checks on data accuracy and correctness will also be major research area of bioinformatics. The use of virtual Reality in drug design, metabolic pathway design, and unicellular organism design, paving the way to design and modification of muticellular organisms, will be the challenges which Bioinformatics scientist and specialist have to tackle. It has now been universally recognized that Bioinformatics is the key to the new grand data-intensive molecular biology that will take us into 21 century.

Bioinformatics - Industry Overview

The Bioinformatics industry has grown to keep up with the information explosion, growing at 25-50% a year. In 2000, the US market Research company estimated that the value of the Bioinformatics industry would touch \$2 billion. Now it s demand for individuals capable of

doing bioinformatics is soaring. Industry's demand for scientists with skills in Bioinformatics far exceeds the supply of qualified specialists in the field, Seems likely that this figure will be reached within the coming year. Therefore, companies are developing methods of spotting potential Bioinformatics experts and then training them on the job.

Bioinformatics and computational biology

Bioinformatics and computational biology each maintain close interactions with life sciences to realize their full potential. Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful. Computational biology uses mathematical and computational approaches to address theoretical and experimental questions in biology. Although bioinformatics and computational biology are distinct, there is also significant overlap and activity at their interface.

Biocomputing

Biocomputing is often used as a catch-all term covering all this area at the intersection of Biology and Computation. Although many other terms are used to name the same area. We can distinguish in to (non-disjoint) sub-fields:

Computational Biology

Computational Biology is application of core technology of computer science (eg. algorithms, artificial intelligence, databases etc) to problems arising from biology. Computational biology is particularly exciting today because the problems are large enough to motivate the efficient algorithms and moreover the demand of biology on computational science is increasing.

The most pressing tasks in bioinformatics involve the analysis of sequence information. Computational Biology is the name given to this process, and it involves the following:

- Finding the genes in the DNA sequences of various organisms
- Developing methods to predict the structure and/or function of newly discovered proteins and structural RNA sequences.
- Clustering protein sequences into families of related sequences and the development of protein models.
- Aligning similar proteins and generating phylogenetic trees to examine evolutionary relationships.

Machine Learning in bioinformatics

Presently a large list of bioinformatics tools and softwares are available which are based on machine learning. The twin of Bioinformatics, called Computational Biology have emerged largely into development of softwares and application using machine learning and deep learning techniques for biological image data analysis. Recently Google's Deep Learning library called TensorFlow was shown how it can be used in computational biology. Application of machine learning and deep learning in biology need to be explored further for building AI's which can be used for disease diagnosis and prediction.

According to the Science Daily news, biologist are increasingly turning into Data Scientist as Bioinformatics Data Scientist or Genomic Data Scientist. The market of bioinformatics and career needs in bioinformatics is increasing each year. It is predicted that in the near future, there will be a huge need for people having bioinformatics skills. Bioinformatics has become an essential interdisciplinary science for life science and biomedical sciences. However there is a huge demand in education and training in bioinformatics.

TOOLS USED IN BIOINFORMATICS

- 1. BLAST-** Basic Logic Alignment Search Tool (BLAST) is a family of user-friendly sequence search tool. It find region of local similarity between sequences. Identification of homologous for a particular sequence allows the prediction potential of potential formation and in modeling of 3-D structure. A local alignment finds the optimal alignment between sub regions or local regions specified sequences. A local alignment tool is used to find sequence motives, domains, etc. in the databases that are homologous to the submitted sequence motive domains, etc. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. This helps in the identification of unknown or new sequences, which can be valuable for research projects. BLAST can be utilized to infer fuctional and evolutionary relationships between sequences as well as help identify members of gene families. There are several programs in BLAST like BLAST-P, BLAST-X, BLAST-N, t BLAST-N, t BLASTx, Subject Databases, Sequence input, Parameters to adjust. Each of them serves as a specific purpose.
- 2. FASTA-** It is the DNA and protein sequence software package described David J. Lipman and William R. It is a pronounced technique which contain programs for protein and DNA translation. It is a heuristic for finding significant matches matches between a query string q and a database string d. FASTA's general strategy is to find the most significant diagonals in the dot-plot or dynamic programming matrix. The performance of the algorithm is influenced by a word-size parameter k, usually 6 for DNA and 2 for amino acids. FASTA is available online and hence is easy to use and easily available. The major focus of the package is of calculation of accurate similarity so that the biologist can judge whether an alignment has occurred by chance or it can be used to infer homology.
- 3. ENTREZ-** It is one of the most popular search engine present at NCBI, USA which searches bibliographic citations and biological data from a variety of reliable databases. This includes the Swiss-Prot, PDB, Gene Bank, ENBL, etc. ENTREZ offers a variety of criteria for search and is highly versatile and capable search tool. It can be used to search variety of information.
- 4. LOCUS LINK-** It contains information about genes including their official names. In addition, it allows one to search gene homology for a given gene and to obtain information about these genes.

- 5. PROSITE-** It is a collection of functional sites sequences pattern found in many proteins. It has many search tools for matching patterns and motifs.
- 6. Clustal W-** It is a fully automatic program for global multiple alignment of DNA and protein sequences. The alignment is progressive and considers the sequence redundancy. Trees can also be calculated from multiple alignments. The program has some adjustable parameters with reasonable defaults. The important tools in studying sequences are multiple alignments of protein sequences. They provide identification of conserved sequence regions. This is very useful in designing experiments to test and modify the function of specific proteins, in predicting the function and structure of proteins, and in identifying new members of protein families. The basic information they provide is identification of conserved sequence regions. This is very useful in designing experiments to test and modify the function of specific proteins, in predicting the function and structure of proteins, and in identifying new members of protein families.
- 7. RASMOL-** It aims to display, teach and generate publication quality images. It is a molecular graphics program intended for the visualization of proteins, nucleic acid and small molecules.

NEED FOR THE BIOINFORMATICS TOOLS

Bioinformatics is a comparatively young discipline and has progressed very fast in the last few years. It has made it possible to test our hypotheses virtually and therefore allows to take a better and an informed decision before launching costly experimentations. Although, more and more tools for analysing genomes, proteomes, predicting structures, rational drug designing and molecular simulations are being developed; none of them is 'perfect'. Therefore, the hunt for finding a better package for solving the given problems will continue.

1. Gene Identification and Sequence Analyses Sequence analyses refer to the understanding of different features of a biomolecule like nucleic acid or protein, which give to it its unique function. First, the sequences of corresponding molecule(s) are retrieved from public databases. After refinement, if needed, they are subjected to various tools that enable prediction of their features related to their function, structure, evolutionary history or identification of homologues with a great accuracy For example, data retrieval tools such as Entrez of PubMed allows one to search and retrieve data from a wide range of data domains. Similarly, pattern discovery tools such as Expression Profiler, Gene Quiz allow researchers to search out different patterns in the given data. Another set of tools is dedicated to carry out sequence comparison. These tools such as BLAST (Basic Local Alignment Search Tool), ClustalW enable one to compare gene or protein sequences to study their evolutionary history or origin. The data visualization tools such as Jalview, GeneView, TreeView, Genes-Graph allow researchers to view data in graphic representation. These tools use advanced mathematical modelling and statistical inferences such as dynamic programming, Hidden Markov Model (HMM), Regression analysis, Artificial Neural Network (ANN), Clustering and Sequence Mining to analyse the given sequence.

2. Phylogenetic Analyses- are procedures used to reconstruct the evolutionary relationship among a group of related molecules or organisms, to predict certain features of a molecule with unknown functions, to track gene flow, and to determine genetic relatedness. This all

could be represented on a genealogic tree or tree of life. The underlying principle of phylogeny is to group living organisms according to the degree of similarity: greater the similarity, closer the organisms would appear on a tree. A phylogenetic comparative analysis is widely used to control for the lack of statistical independence among species. Phylogenetic tools are commonly used to test various evolutionary hypotheses and have become indispensable for functional genomics, particularly when the functions of a gene are not known. For example, prior to the expression of an algal membrane protein, plastid terminal oxidase 1 (PTOX1), in tobacco chloroplasts, authors conducted a phylogenetic analysis to construct the evolutionary history and determine essential features of that particular polypeptide. The phylogenetic analysis revealed that the *Chlamydomonas reinhardtii* PTOX1 (Cr-PTOX1) has typical signatures of higher plant PTOX such as iron-binding sites, a conserved exon and various blocks of amino acids to act as plastoquinol terminal oxidase. Using phylogenetic analysis to study the evolutionary history of respiratory mechanisms in the deep-sea bacterium *Shewanella piezotolerans* WP3. The phylogenetic analyses coupled with reverse genetic studies revealed that out of two nitrate reductases, NAP- α and NAP- β , the hallmark of the genus *Shewanella*, the NAP- β evolved long before NAP- α molecules.

3. Sequence Databases- Biological sequence database refers to a vast collection of information about biological molecules such as nucleic acids, proteins and polymers, each molecule to be identified by a unique key. The stored information is not only important for future use but also serves as a tool for primary sequence analyses. Databases contain a variety of information; and therefore are classified into Primary, Secondary, or Composite databases, depending upon the information stored in them. For example, the data in a primary database is obtained through experimentation such as yeast-two hybrid assay, affinity chromatography, XRD or NMR approaches such as related to sequence or structure. SWISS-PROT, UniProt and PIR, GenBank, EMBL, DDBJ and the Protein Databank PDB are examples of primary databases. A secondary database contains information that is derived from the analysis of data stored in primary databases like conserved sequences, active sites of a protein family or conserved secondary motifs of protein molecules. Examples of secondary databases include SCOP, CATH, PROSITE eMOTIF. Consequently, the primary databases are of archival nature while secondary databases are termed as curated databases. A composite database contains information derived from different primary sources. Examples of composite databases include NRDB (nonredundant database), which contains data obtained from GenBank (CDS translations), PDB, SWISS-PROT, PIR, and PRF. Similarly, the INSD (International Nucleotide Sequence Database) is another example of composite database, which is collection of nucleic acid sequences from EMBL, GenBank, and DDBJ. The UniProt (universal protein sequence database) represents another example, which is also a collation of sequences derived from various other databases PIRPSD, Swiss-Prot, and TrEMBL. Similarly, wwPDB (worldwide PDB) is a composite of 3D structures in the RCSB (Research Collaboratory for Structural Bioinformatics), PDB, MSD, and PDBj

4. Genome Sequence Databases- The GenBank, built by the NCBI, is a vast collection of genome sequences of over 250,000 species. The data from GenBank can be accessed through the NCBI's integrated retrieval system, Entrez, while the literature is accessible via PubMed. Each sequence carries information about the literature, bibliography, organism, and a set of various other features, which include coding regions, promoters, untranslated regions,

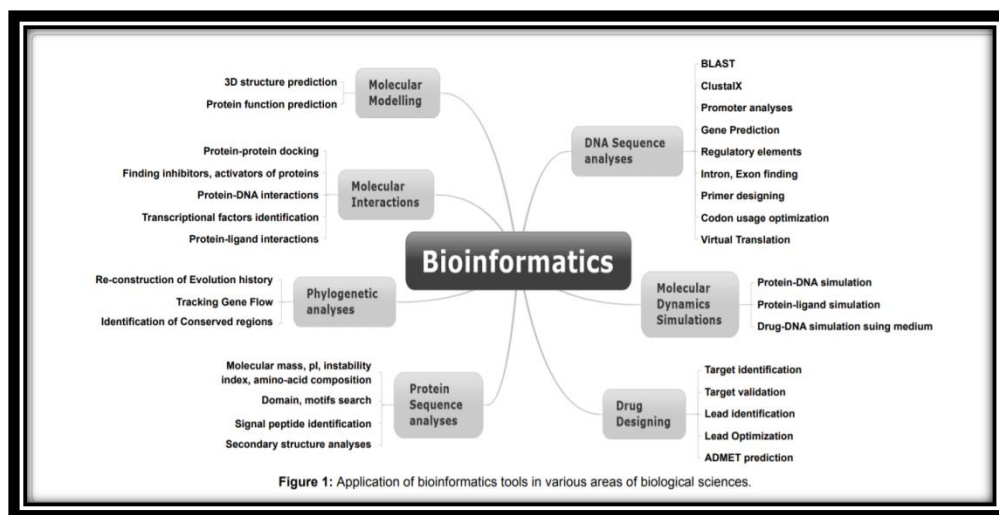
terminators, exons, introns, repeat regions, and translations. The sequence information stored in GenBank is obtained through submission both by the individual laboratories as well as by large-scale genome sequencing projects. Similarly, the Xenbase is an updated resource of genomic and biological data on the frogs including *Xenopus laevis* and *Xenopus tropicalis*, where *Xenopus* spp. are considered as model providing new knowledge in the field of developmental biology which may be exploited to modelling and simulation studies of the human diseases. The *Saccharomyces* Genome Database (SGD) contains comprehensive information of the yeast (*Saccharomyces cerevisiae*) and also provides bioinformatics tools to explore and analyse the data available in SGD.

5. Protein Sequence Databases- The most significant protein sequence databases include SWISSPROT (Swiss Protein) Databank [50], TrEMBL (translation of DNA sequences in EMBL), UniProt (Universal Protein Resource) [33], PIR (Protein Information Resource) and wwPDB (worldwide Protein DataBank). The SWISS-PROT represents one of the comprehensive protein sequence databases. The SWISS-PROT provides information of its entries, which has been generated both experimental as well computational studies. It also provides links to several other data sources such as GenBank, EMBL, DDBJ, PDB and various other secondary protein databases namely domains, posttranslational modifications, species-specific data collections. The protein information in SWISS-PROT mainly concentrates on model organisms and human. The TrEMBL by contrast provides information on proteins from all organisms. Similarly, the PIR is another comprehensive collection of protein sequences. It provides user several attractive features for example to search for a protein molecule via an 'interactive text search' and to perform various web-based analyses such as sequence alignment, matching of peptide molecules and peptide mass calculations. The UniProt is one of the comprehensive collections of protein sequence resources, which are open to free access. The UniProt database emerged by combining SWISS-PROT, PIR and TrEMBL collections. It provides all sorts of protein information ranging from sequence to function. The worldwide Protein Data Bank (wwPDB) has been exclusively designed to archive each single 3D structure of protein molecules to become freely available to the scientific community. The databank now contains over 83,000 experimentally generated structures. The PDB also constantly develops tools for the users to provide better access to the data

6. Miscellaneous Databases- The Rfam database contains comprehensive information about RNA molecules and their various features like secondary structures and gene expression modulating elements. The Rfam databases are hosted by the Wellcome Trust Sanger Institute and it is similar to the Pfam database for annotating protein families. As there are number of curated databases available, one of such databases is IntAct, which contains data on protein interactions. All data manually curated by MINT (Molecular INTERaction database) curators has been shifted onto the IntAct database at EMBL-EBI and have been merged with the existing IntAct data collections. MINT is another database that stores information about protein-protein interactions derived from already published data in literature. Curated databases for information on complex metabolic pathways have also been built. For example, the Reactome is one such curated database that represents a range of diverse human processes ranging from metabolism to signal transduction. The Reactome is an open source platform, which is freely available to be used and redistributed [57]. The Transporters Classification

Database (TCDB) is a collection of membrane transporters [58]. It uses an internationally approved Transport Classification (TC) system for the classification of protein, which is similar to that of Enzyme Commission (EC). However, it also has some differences from EC system; it provides functional and phylogenetic information as well, for example. The information of more than 600 families of transporters is available in this database. A TC number to sequenced homologues of unknown function is assigned only if it belongs to rare or under-represented family. Various subunits are represented by 'S' followed by a number such as S1, S2, S3 and so on. Whereas the proteins which act as accessory transporters as well as those whose characterization is not complete yet are represented by number 8 and 9, respectively. Similarly, the Carbohydrate-Active enzyme Database (CAZy) contains comprehensive information about carbohydrate-modifying enzymes and other information relevant to them. The enzymes are classified into distinct families on the basis of amino acid similarities in their sequences or the presence of various catalytic domains. The databases about the structure, classification and ontology of the lipid molecules have been discussed in detail elsewhere

7. Drug Designing- Drug discovery is a process by which new drug molecules are discovered or designed to cure different diseases. Before the advent bioinformatics tools, scientists used chemistry, pharmacology and clinical sciences to discover new compounds. However, the traditional process is quite slow and expensive as well. The market pressure to find new drugs in a short period with minimum risks has fuelled the interest in alternative ways of designing drugs such as bioinformatics. Bioinformatics has greatly facilitated this complex process and is playing a vital role in advancing the process of drug discovery/designing, since it is faster to analyse molecules on computer as compared to experimental approaches. In fact, a completely new and dedicated field known as Computer Aided Drug Design (CADD) has come into existence to discover novel drug molecules. The whole process of discovering and designing new drug molecules is quite complicated and is quite challenging. The entire process can be divided into four different steps: identification of drug target, validation of target, lead identification, and lead optimization. In this section, we will briefly discuss how bioinformatics is useful in discovering new drugs. Since drug molecules always act on a target to deliver therapeutic benefit to the patient. The target is a small key biomolecule that allows the drug molecule to produce a desired effect on metabolic or signalling pathway pertinent to the disease under study without interfering the normal functioning of the cell. Therefore, the very first step in the drug designing process is to identify a target involved in that disease. This demands a full knowledge of metabolic processes in normal as well as diseased conditions. The sequencing of human genome provided over 30,000 genes to researchers to include them in their search for new drug targets. Since then the number of potential drug targets is increasing day-by-day. Understanding how a gene functions is indeed a key to choose a gene as a target. A number of databases have been developed to facilitate the search of new drug targets.



CONCLUSION

Bioinformatics is the collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied to molecular genetics and genomics. The term bioinformatics was coined by Hwa Lim in the late 1980s, and popularized in the 1990s through its association with the human genome project. which computational resources, such as databases, are operated for the benefit of the research community. Bioinformatics is a broad field that has a central role in many areas of biological research. These include genomics and, more specifically, genomic sequencing and mapping, genome annotation, and comparisons of multiple genomes. Bioinformatics is also essential in transcriptomics — the study of transcribed sequences, both full-length cDNAs and expressed sequence tags (ESTs) — and the analysis of gene expression data typically measured using DNA microarrays or some form of sample sequencing. It is also crucial in proteomics for the analysis of protein sequences (e.g. to determine functional motifs), for the study of protein abundance (typically measured using two-dimensional gels or mass spectrometry), and the determination of protein structure either empirically or computationally. Bioinformatics is key in the analysis of protein–protein interactions and molecular pathways (the ‘interactome’) and in systematic studies of gene regulation (the ‘regulome’). It also plays a vital role in genetics, both in the discovery of new molecular genetic markers, such as single nucleotide polymorphisms, and the use of these and other markers to dissect the genetic basis of disease and other phenotypes. Bioinformatics is a comparatively young discipline and has progressed very fast in the last few years. It has made it possible to test our hypotheses virtually and therefore allows to take a better and an informed decision before launching costly experimentations. Although, more and more tools for analysing genomes, proteomes, predicting structures, rational drug designing and molecular simulations are being developed; none of them is ‘perfect’.

REFERENCES

- www.wikipedi.org
- www.researchgate.com
- www.sciendirect.com

Unit V

Prediction tools- multiple sequence alignment, phylogenetic tree, gene prediction, protein structure & functions prediction. Modeling tools: 2D and 3D protein modeling.

CONTENTS

- Introduction
- Multiple sequence alignment

Alignment methods

- Phylogenetic tree: applications and limitations
- References

INTRODUCTION

Sequence alignment

- In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.¹ Aligned sequences of nucleotide or amino acids residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns. Sequence alignments are also used for non-biological sequences, such as calculating the distance cost between strings in a natural language or in financial data.

Multiple sequence alignment

- A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. In many cases, the input set of query sequences are assumed to have an evolutionary relationship by which they share a linkage and are descended from a common ancestor. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins. Visual depictions of the alignment as in the image at right illustrate mutation events such as point mutations (single amino acid or nucleotide changes) that appear as differing characters in a single alignment column, and insertion or deletion mutations (indels or gaps) that appear as hyphens in one or more of the sequences in the alignment. Multiple sequence alignment is often used to assess sequence conservation of protein domains, tertiary and secondary structures, and even individual amino acids or nucleotides.
- Multiple sequence alignment also refers to the process of aligning such a sequence set. Because three or more sequences of biologically relevant length can be difficult and are almost always time-consuming to align by hand, computational algorithms are used to produce and analyze the alignments. MSAs require more sophisticated methodologies than pairwise alignment because they are more computationally complex. Most multiple sequence alignment programs use heuristic methods rather than global optimization because identifying the optimal alignment between more than a few sequences of moderate length is prohibitively computationally expensive.

- Multiple sequence alignment is an extension of pairwise alignment to incorporate more than two sequences at a time. Multiple alignment methods try to align all of the sequences in a given query set.

Pairwise Sequence Alignment vs Multiple Sequence Alignment

- **Pairwise Sequence Alignment** is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

```

Q K E S G P S S S Y C
| | | | |
V Q Q E S G L V R T T C

```

Multiple Sequence Alignment(MSA) is the alignment of three or more biological sequences of similar length. From the output of MSA applications, homology can be inferred and the evolutionary relationship between the sequences studied.

```

Q K E S G P S S S Y C
| | | | |
V Q Q E S G L V R T T C
| | | | |
V Q K E S L L V R S T C

```

ALIGNMENT METHODS

- There are various alignment methods used within multiple sequence to maximize scores and correctness of alignments. Each is usually based on a certain heuristic with an insight into the evolutionary process. Most try to replicate evolution to get the most realistic alignment possible to best predict relations between sequences.

Progressive alignment construction

- The most widely used approach to multiple sequence alignments uses a heuristic search known as progressive technique (also known as the hierarchical or tree method) developed by Da-Fei Feng and Doolittle in 1987. Progressive alignment builds up a final MSA by combining pairwise alignments beginning with the most similar pair and progressing to the most distantly related. All progressive alignment methods require two stages: a first stage in which the relationships between the sequences are represented as a tree, called a *guide tree*, and a second step in which the MSA is built by adding the sequences sequentially to the growing MSA according to the guide tree. The initial *guide tree* is determined by an efficient clustering method such as neighbor-joining or UPGMA, and may use distances based on the number of identical two-letter sub-sequences (as in FASTA rather than a dynamic programming alignment).
- Progressive alignments are not guaranteed to be globally optimal. The primary problem is that when errors are made at any stage in growing the MSA, these errors are then propagated through to the final result. Performance is also particularly bad when all of the sequences in the set are rather distantly related. Most modern progressive methods modify their scoring function with a secondary weighting

function that assigns scaling factors to individual members of the query set in a nonlinear fashion based on their phylogenetic distance from their nearest neighbours . This corrects for non-random selection of the sequences given to the alignment program.

- Progressive alignment methods are efficient enough to implement on a large scale for many (100s to 1000s) sequences. Progressive alignment services are commonly available on publicly accessible web servers so users need not locally install the applications of interest. The most popular progressive alignment method has been the Clustalfamily, especially the weighted variant ClustalW to which access is provided by a large number of web portals including GenomeNet, EBI, and EMBNet . Different portals or implementations can vary in user interface and make different parameters accessible to the user. ClustalW is used extensively for phylogenetic tree construction, in spite of the author's explicit warnings that unedited alignments should not be used in such studies and as input for protein structure prediction by homology modeling. Current version of Clustal family is ClustalW2. EMBL-EBI announced that CLustalW2 will be expired in August 2015. They recommend Clustal Omega which performs based on seeded guide trees and HMM profile-profile techniques for protein alignments. They offer different MSA tools for progressive DNA alignments. One of them is MAFFT (Multiple Alignment using Fast Fourier Transform).

T-Coffee

- Another common progressive alignment method called T-Coffee is slower than Clustal and its derivatives but generally produces more accurate alignments for distantly related sequence sets. T-Coffee calculates pairwise alignments by combining the direct alignment of the pair with indirect alignments that aligns each sequence of the pair to a third sequence. It uses the output from Clustal as well as another local alignment program LALIGN, which finds multiple regions of local alignment between two sequences. The resulting alignment and phylogenetic tree are used as a guide to produce new and more accurate weighting factors.
- Because progressive methods are heuristics that are not guaranteed to converge to a global optimum, alignment quality can be difficult to evaluate and their true biological significance can be obscure. A semi-progressive method that improves alignment quality and does not use a lossy heuristic while still running in polynomial time has been implemented in the program PSAlign.

Iterative methods

- A set of methods to produce MSAs while reducing the errors inherent in progressive methods are classified as "iterative" because they work similarly to progressive methods but repeatedly realign the initial sequences as well as adding new sequences to the growing MSA. One reason progressive methods are so strongly dependent on a high-quality initial alignment is the fact that these alignments are always incorporated into the final result — that is, once a sequence has been aligned into the MSA, its alignment is not considered further. This approximation improves efficiency at the cost of accuracy. By contrast, iterative methods can return to previously calculated pairwise alignments or sub-MSAs incorporating subsets of the query sequence as a

means of optimizing a general objective function such as finding a high-quality alignment score.

- A variety of subtly different iteration methods have been implemented and made available in software packages; reviews and comparisons have been useful but generally refrain from choosing a "best" technique. The software package PRRN/PRRP uses a hill-climbing algorithm to optimize its MSA alignment score and iteratively corrects both alignment weights and locally divergent or "gappy" regions of the growing MSA. PRRP performs best when refining an alignment previously constructed by a faster method.
- Another iterative program, DIALIGN, takes an unusual approach of focusing narrowly on local alignments between sub-segments or sequence motifs without introducing a gap penalty. The alignment of individual motifs is then achieved with a matrix representation similar to a dot-matrix plot in a pairwise alignment. An alternative method that uses fast local alignments as anchor points or "seeds" for a slower global-alignment procedure is implemented in the CHAOS/DIALIGN suite.
- A third popular iteration-based method called MUSCLE(multiple sequence alignment by log-expectation) improves on progressive methods with a more accurate distance measure to assess the relatedness of two sequences.

Consensus methods

- Consensus methods attempt to find the optimal multiple sequence alignment given multiple different alignments of the same set of sequences. There are two commonly used consensus methods, M-COFFEE and MergeAlign.M-COFFEE uses multiple sequence alignments generated by seven different methods to generate consensus alignments. MergeAlign.M is capable of generating consensus alignments from any number of input alignments generated using different models of sequence evolution or different methods of multiple sequence alignment. The default option for Merge Align is to infer a consensus alignment using alignments generated using 91 different models of protein sequence evolution.

Hidden Markov models

- Hidden Markov models are probabilistic models that can assign likelihoods to all possible combinations of gaps, matches, and mismatches to determine the most likely MSA or set of possible MSAs. HMMs can produce a single highest-scoring output but can also generate a family of possible alignments that can then be evaluated for biological significance. HMMs can produce both global and local alignments. Although HMM-based methods have been developed relatively recently, they offer significant improvements in computational speed, especially for sequences that contain overlapping regions
- Typical HMM-based methods work by representing an MSA as a form of directed acyclic graph known as a partial-order graph, which consists of a series of nodes representing possible entries in the columns of an MSA. In this representation a column that is absolutely conserved (that is, that all the sequences in the MSA share a particular character at a particular position) is coded as a single node with as many outgoing connections as there are possible characters in the next column of the alignment. In the terms of a typical hidden Markov model, the observed states are the

individual alignment columns and the "hidden" states represent the presumed ancestral sequence from which the sequences in the query set are hypothesized to have descended. An efficient search variant of the dynamic programming method, known as the Viterbi algorithm, is generally used to successively align the growing MSA to the next sequence in the query set to produce a new MSA. This is distinct from progressive alignment methods because the alignment of prior sequences is updated at each new sequence addition. However, like progressive methods, this technique can be influenced by the order in which the sequences in the query set are integrated into the alignment, especially when the sequences are distantly related.

- Several software programs are available in which variants of HMM-based methods have been implemented and which are noted for their scalability and efficiency, although properly using an HMM method is more complex than using more common progressive methods. The simplest is POA (Partial-Order Alignment); a similar but more generalized method is implemented in the packages SAM (Sequence Alignment and Modeling System) and HMMER. SAM has been used as a source of alignments for protein structure prediction to participate in the CASP structure prediction experiment and to develop a database of predicted proteins in the yeast species *S. cerevisiae*. HHsearch is a software package for the detection of remotely related protein sequences based on the pairwise comparison of HMMs. A server running HHsearch HHpred was by far the fastest of the 10 best automatic structure prediction servers in the CASP7 and CASP8 structure prediction competitions.

What is MSA

- Comparison of many (i.e., >2) sequences
- local or global



Phylogeny-aware method

- Most multiple sequence alignment methods try to minimize the number of insertions/deletions (gaps) and, as a consequence, produce compact alignments. This causes several problems if the sequences to be aligned contain non-homologous regions, if gaps are informative in a phylogeny analysis. These problems are common in newly produced sequences that are poorly annotated and may contain frame-shifts, wrong domains or non-homologous spliced exons. The first such method was developed in 2005 by Löytynoja and Goldman. The same authors released a software package called *PRANK* in 2008. *PRANK* improves alignments when insertions are present. Nevertheless, it runs slowly compared to progressive and/or iterative methods which have been developed for several years.
- In 2012, two new phylogeny-aware tools appeared. One is called *PAGAN* that was developed by the same team as *PRANK*. The other is *ProGraphMSA* developed by Szalkowski. Both software packages were developed independently but share common features, notably the use of graph algorithms to improve the recognition of non-homologous regions, and an improvement in code making these software faster than *PRANK*.

Motif finding

- Motif finding, also known as profile analysis, is a method of locating sequence motifs in global MSAs that is both a means of producing a better MSA and a means of producing a scoring matrix for use in searching other sequences for similar motifs. A variety of methods for isolating the motifs have been developed, but all are based on identifying short highly conserved patterns within the larger alignment and constructing a matrix similar to a substitution matrix that reflects the amino acid or nucleotide composition of each position in the putative motif. The alignment can then be refined using these matrices. In standard profile analysis, the matrix includes entries for each possible character as well as entries for gaps. Alternatively, statistical pattern-finding algorithms can identify motifs as a precursor to an MSA rather than as a derivation. In many cases when the query set contains only a small number of sequences or contains only highly related sequences, pseudocounts are added to normalize the distribution reflected in the scoring matrix. In particular, this corrects zero-probability entries in the matrix to values that are small but nonzero.
- Blocks analysis is a method of motif finding that restricts motifs to ungapped regions in the alignment. Blocks can be generated from an MSA or they can be extracted from unaligned sequences using a precalculated set of common motifs previously generated from known gene families. Block scoring generally relies on the spacing of high-frequency characters rather than on the calculation of an explicit substitution matrix. The *BLOCKS*server provides an interactive method to locate such motifs in unaligned sequences.
- Statistical pattern-matching has been implemented using both the expectation-maximization algorithm and the Gibbs sampler. One of the most common motif-finding tools, known as *MEME*, uses expectation maximization and hidden Markov

methods to generate motifs that are then used as search tools by its companion MAST in the combined suite MEME/MAST.

Non-coding multiple sequence alignment

- Non-coding DNA regions, especially TFBSs, are rather more conserved and not necessarily evolutionarily related, and may have converged from non-common ancestors. Thus, the assumptions used to align protein sequences and DNA coding regions are inherently different from those that hold for TFBS sequences. Although it is meaningful to align DNA coding regions for homologous sequences using mutation operators, alignment of binding site sequences for the same transcription factor cannot rely on evolutionary related mutation operations. Similarly, the evolutionary operator of point mutations can be used to define an edit distance for coding sequences, but this has little meaning for TFBS sequences because any sequence variation has to maintain a certain level of specificity for the binding site to function. This becomes specifically important when trying to align known TFBS sequences to build supervised models to predict unknown locations of the same TFBS. Hence, Multiple Sequence Alignment methods need to adjust the underlying evolutionary hypothesis and the operators used as in the work published incorporating neighbouring base thermodynamic information to align the binding sites searching for the lowest thermodynamic alignment conserving specificity of the binding site, EDNA

VISUALIZATION AND QUALITY CONTROL ALIGNMENT

- The necessary use of heuristics for multiple alignment means that for an arbitrary set of proteins, there is always a good chance that an alignment will contain errors. For example, an evaluation of several leading alignment programs using the BAliBase benchmark found that at least 24% of all pairs of aligned amino acids were incorrectly aligned. These errors can arise because of unique insertions into one or more regions of sequences, or through some more complex evolutionary process leading to proteins that do not align easily by sequence alone. As the number of sequence and their divergence increases many more errors will be made simply because of the heuristic nature of MSA algorithms. Multiple sequence alignment viewers enable alignments to be visually reviewed, often by inspecting the quality of alignment for annotated functional sites on two or more sequences. Many also enable the alignment to be edited to correct these (usually minor) errors, in order to obtain an optimal 'curated' alignment suitable for use in phylogenetic analysis or comparative modeling.
- However, as the number of sequences increases and especially in genome-wide studies that involve many MSAs it is impossible to manually curate all alignments. Furthermore, manual curation is subjective. And finally, even the best expert cannot confidently align the more ambiguous cases of highly diverged sequences. In such cases it is common practice to use automatic procedures to exclude unreliably aligned regions from the MSA. For the purpose of phylogeny reconstruction. The Gblocks program is widely used to remove alignment blocks suspect of low quality, according to various cutoffs on the number of gapped sequences in alignment columns. However, these criteria may excessively filter out regions with insertion/deletion events that may still be aligned reliably, and these regions might be desirable for other

purposes such as detection of positive selection. A few alignment algorithms output site-specific scores that allow the selection of high-confidence regions. Such a service was first offered by the SOAP program, which tests the robustness of each column to perturbation in the parameters of the popular alignment program CLUSTALW. The T-Coffee program uses a library of alignments in the construction of the final MSA, and its output MSA is colored according to confidence scores that reflect the agreement between different alignments in the library regarding each aligned residue. Its extension, TCS: (Transitive Consistency Score), uses T-Coffee libraries of pairwise alignments to evaluate any third party MSA. Pairwise projections can be produced using fast or slow methods, thus allowing a trade-off between speed and accuracy. Another alignment program that can output an MSA with confidence scores is FSA, which uses a statistical model that allows calculation of the uncertainty in the alignment. The HoT (Heads-Or-Tails) score can be used as a measure of site-specific alignment uncertainty due to the existence of multiple co-optimal solutions.

PHYLOGENETIC USE

- Multiple sequence alignments can be used to create a phylogenetic tree. This is made possible by two reasons. The first is because functional domains that are known in annotated sequences can be used for alignment in non-annotated sequences. The other is that conserved regions known to be functionally important can be found. This makes it possible for multiple sequence alignments to be used to analyze and find evolutionary relationships through homology between sequences. Point mutations and insertion or deletion events (called indels) can be detected.
- Multiple sequence alignments can also be used to identify functionally important sites, such as binding sites, active sites, or sites corresponding to other key functions, by locating conserved domains. When looking at multiple sequence alignments, it is useful to consider different aspects of the sequences when comparing sequences. These aspects include identity, similarity, and homology. Identity means that the sequences have identical residues at their respective positions. On the other hand, similarity has to do with the sequences being compared having similar residues quantitatively. For example, in terms of nucleotide sequences, pyrimidines are considered similar to each other, as are purines. Similarity ultimately leads to homology, in that the more similar sequences are, the closer they are to being homologous. This similarity in sequences can then go on to help find common ancestry.

PHYLOGENETIC TREE

- A phylogenetic tree or evolutionary tree is a branching diagram or "tree" showing the evolutionary relationships among various biological species or other entities their phylogeny based upon similarities and differences in their physical or genetic characteristics. All life on Earth is part of a single phylogenetic tree, indicating common ancestry.
- In a *rooted* phylogenetic tree, each node with descendants represents the inferred most recent common ancestor of those descendants, and the edge lengths in some trees may be interpreted as time estimates. Each node is called a taxonomic unit. Internal nodes are generally called hypothetical taxonomic units, as they cannot be directly observed.

Trees are useful in fields of biology such as bioinformatics, systematics, and phylogenetics. *Unrooted* trees illustrate only the relatedness of the leaf nodes and do not require the ancestral root to be known or inferred.

HISTORY

- The idea of a "tree of life" arose from ancient notions of a ladder-like progression from lower into higher forms of life (such as in the Great Chain of Being). Early representations of "branching" phylogenetic trees include a "paleontological chart" showing the geological relationships among plants and animals in the book *Elementary Geology*, by Edward Hitchcock (first edition: 1840).
- Charles Darwin (1859) also produced one of the first illustrations and crucially popularized the notion of an evolutionary "tree" in his seminal book *The Origin of Species*. Over a century later, evolutionary biologists still use tree diagrams to depict evolution because such diagrams effectively convey the concept that speciation occurs through the adaptive and semirandom splitting of lineages. Over time, species classification has become less static and more dynamic.
- The term *phylogenetic*, or *phylogeny*, derives from the two ancient greek words meaning "race, lineage", and (*g nesis*), meaning "origin, source".

PROPERTIES

1. Rooted tree.

- A rooted phylogenetic tree (see two graphics at top) is a directed tree with a unique node — the root — corresponding to the (usually imputed) most recent common ancestor of all the entities at the leaves of the tree. The root node does not have a parent node, but serves as the parent of all other nodes in the tree. The root is therefore a node of degree 2 while other internal nodes have a minimum degree of 3 (where "degree" here refers to the total number of incoming and outgoing edges).
- The most common method for rooting trees is the use of an uncontroversial outgroup close enough to allow inference from trait data or molecular sequencing, but far enough to be a clear outgroup.

2. Unrooted tree

- Unrooted trees illustrate the relatedness of the leaf nodes without making assumptions about ancestry. They do not require the ancestral root to be known or inferred. Unrooted trees can always be generated from rooted ones by simply omitting the root. By contrast, inferring the root of an unrooted tree requires some means of identifying ancestry. This is normally done by including an outgroup in the input data so that the root is necessarily between the outgroup and the rest of the taxa in the tree, or by introducing additional assumptions about the relative rates of evolution on each branch, such as an application of the molecular clock hypothesis

3. Bifurcating versus multifurcating

- Both rooted and Unrooted trees can be either bifurcating or multifurcating. A rooted bifurcating tree has exactly two descendants arising from each interior node (that is, it forms a binary tree), and an unrooted bifurcating tree takes the form of an unrooted

binary tree, a free tree with exactly three neighbors at each internal node. In contrast, a rooted multifurcating tree may have more than two children at some nodes and an unrooted multifurcating tree may have more than three neighbors at some nodes.

4. Labeled versus unlabeled

- Both rooted and unrooted trees can be either labeled or unlabeled. A labeled tree has specific values assigned to its leaves, while an unlabeled tree, sometimes called a tree shape, defines a topology only.

SPECIAL TREE TYPES

- Dendrogram

A dendrogram is a general name for a tree, whether phylogenetic or not, and hence also for the diagrammatic representation of a phylogenetic tree.

- Cladogram

A cladogram only represents a branching pattern; i.e., its branch lengths do not represent time or relative amount of character change, and its internal nodes do not represent ancestors.

- Phylogram

A phylogram is a phylogenetic tree that has branch lengths proportional to the amount of character change.

- Chronogram

A chronogram of Lepidoptera. In this phylogenetic tree type, branch lengths are proportional to geological time. A chronogram is a phylogenetic tree that explicitly represents time through its branch lengths.

- Dahlgrenogram

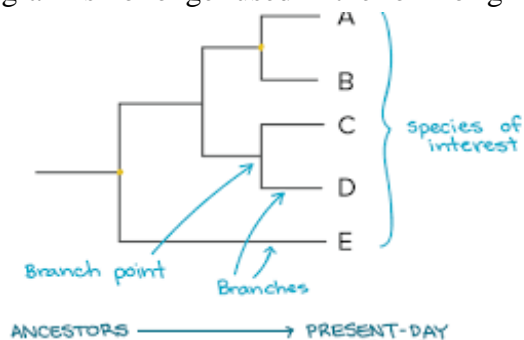
A Dahlgrenogram is a diagram representing a cross section of a phylogenetic tree

- Phylogenetic network

A phylogenetic network is not strictly speaking a tree, but rather a more general graph, or a directed acyclic graph in the case of rooted networks. They are used to overcome some of the limitations inherent to trees.

- Spindle diagram

A spindle diagram, or bubble diagram, is often called a romerogram, after its popularisation by the American palaeontologist Alfred Romer. It represents taxonomic diversity (horizontal width) against geological time (vertical axis) in order to reflect the variation of abundance of various taxa through time. However, a spindle diagram is not an evolutionary tree: the taxonomic spindles obscure the actual relationships of the parent taxon to the daughter taxon and have the disadvantage of involving the paraphyly of the parental group. This type of diagram is no longer used in the form originally proposed.



APPLICATIONS

- The inference of phylogenies with computational methods has many important applications in medical and biological research, such as drug discovery and conservation biology.
- Phylogenetic trees have already witnessed applications in numerous practical domains, such as in conservation biology, epidemiology, forensics, gene function prediction and drug development
- Other applications of phylogenies include multiple sequence alignment, protein structure prediction, gene and protein function and drug design.
- The computation of the tree of life containing representatives of all living beings on earth is considered to be one of the grand challenges in bioinformatics.

LIMITATIONS TO THE USE OF TREES

- It is important to remember that trees do have limitations. For example, trees are meant to provide insight a research question and not intended to represent an species history.
- Several factors like gene transfer, may affect the output placed into a tree.
- All knowledge of limitations related to DNA degradation over time must be considered, especially in the case of evolutionary trees aimed at ancient or extinct organisms.

REFERENCES

https://en.wikipedia.org/wiki/Multiple_sequence_alignment

https://en.wikipedia.org/wiki/Phylogenetic_tree

<https://www.slideshare.net/FaisalHussain23/phylogenetic-tree-types-and-applicantion-75067233>

Gene prediction and protein structure and function prediction

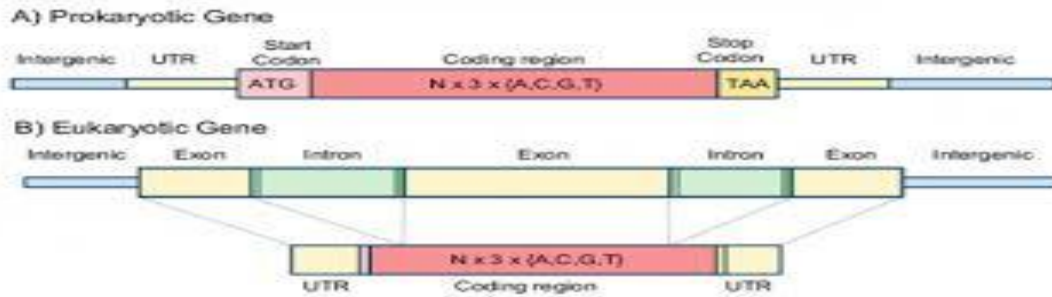
CONTENTS

- Introduction
- Gene prediction
- Protein structure and function prediction
 - 2d and 3d prediction
- References

INTRODUCTION

GENE PREDICTION

- In computational biology **gene prediction** or **gene finding** refers to the process of identifying the regions of genomic DNA that encode genes. This includes protein-coding genes as well as RNA genes, but may also include prediction of other functional elements such as regulatory regions. Gene finding is one of the first and most important steps in understanding the genome of a species once it has been sequenced.
- In its earliest days, "gene finding" was based on painstaking experimentation on living cells and organisms. Statistical analysis of the rates of homologous recombination of several different genes could determine their order on a certain chromosome, and information from many such experiments could be combined to create a genetic map specifying the rough location of known genes relative to each other. Today, with comprehensive genome sequence and powerful computational resources at the disposal of the research community, gene finding has been redefined as a largely computational problem.
- Determining that a sequence is functional should be distinguished from determining the function of the gene or its product. Predicting the function of a gene and confirming that the gene prediction is accurate still demands *in vivo* experimentation through gene knockout and other assays, although frontiers of bioinformatics research are making it increasingly possible to predict the function of a gene based on its sequence alone.
- Gene prediction is one of the key steps in genome annotation, following sequence assembly the filtering of non-coding regions and repeat masking.
- Gene prediction is closely related to the so-called 'target search problem' investigating how DNA binding proteins locate specific binding sites within the genomes. Many aspects of structural gene prediction are based on current understanding of underlying biochemical processes in the cell such as gene transcription, translation, protein-protein interactions and regulation process, which are subject of active research in the various omics fields such as transcriptomics, proteomics, metabolomics and more generally structural and functional genomics.



AB initio methods

- Ab Initio gene prediction is an intrinsic method based on gene content and signal detection. Because of the inherent expense and difficulty in obtaining extrinsic evidence for many genes, it is also necessary to resort to *ab initio* gene finding, in which the genomic DNA sequence alone is systematically searched for certain tell-tale signs of protein-coding genes. These signs can be broadly categorized as either *signals*, specific sequences that indicate the presence of a gene nearby, or *content*, statistical properties of the protein-coding sequence itself. *Ab initio* gene finding might be more accurately characterized as *gene prediction*, since extrinsic evidence is generally required to conclusively establish that a putative gene is functional.
- In the genomes of prokaryotes, genes have specific and relatively well-understood promoter sequences (signals), such as the Pribnow box and transcription factor binding sites, which are easy to systematically identify. Also, the sequence coding for a protein occurs as one contiguous open reading frame (ORF), which is typically many hundred or thousands of base pairs long. The statistics of stop codons are such that even finding an open reading frame of this length is a fairly informative sign. (Since 3 of the 64 possible codons in the genetic code are stop codons, one would expect a stop codon approximately every 20–25 codons, or 60–75 base pairs, in a random sequence.) Furthermore, protein-coding DNA has certain periodicities and other statistical properties that are easy to detect in sequence of this length. These characteristics make prokaryotic gene finding relatively straightforward, and well-designed systems are able to achieve high levels of accuracy.
- *Ab initio* gene finding in eukaryotes, especially complex organisms like humans, is considerably more challenging for several reasons. First, the promoter and other

regulatory signals in these genomes are more complex and less well-understood than in prokaryotes, making them more difficult to reliably recognize. Two classic examples of signals identified by eukaryotic gene finders are CpG islands and binding sites for a poly(A) tail.

- Second, splicing mechanisms employed by eukaryotic cells mean that a particular protein-coding sequence in the genome is divided into several parts (exons), separated by non-coding sequences (introns). (Splice sites are themselves another signal that eukaryotic gene finders are often designed to identify.) A typical protein-coding gene in humans might be divided into a dozen exons, each less than two hundred base pairs in length, and some as short as twenty to thirty. It is therefore much more difficult to detect periodicities and other known content properties of protein-coding DNA in eukaryotes.
- Advanced gene finders for both prokaryotic and eukaryotic genomes typically use complex probabilistic models, such as hidden Markov models (HMMs) to combine information from a variety of different signal and content measurements. The GLIMMER system is a widely used and highly accurate gene finder for prokaryotes. GeneMark is another popular approach. Eukaryotic *ab initio* gene finders, by comparison, have achieved only limited success; notable examples are the GENSCAN and geneid programs. The SNAP gene finder is HMM-based like Genscan, and attempts to be more adaptable to different organisms, addressing problems related to using a gene finder on a genome sequence that it was not trained against. A few recent approaches like mSplicer, CONTRAST, or mGene also use machine learning techniques like support vector machines for successful gene prediction. They build a discriminative model using hidden Markov support vector machines or conditional random fields to learn an accurate gene prediction scoring function.
- *Ab Initio* methods have been benchmarked, with some approaching 100% sensitivity, however as the sensitivity increases, accuracy suffers as a result of increased false positives.

Other signals

- Among the derived signals used for prediction are statistics resulting from the sub-sequence statistics like k-mer statistics, Isochore (genetics) or Compositional domain GC composition/uniformity/entropy, sequence and frame length, Intron/Exon/Donor/Acceptor/Promoter and Ribosomal binding site vocabulary, Fractal dimension, Fourier transform of a pseudo-number-coded DNA, Z-curve parameters and certain run features.
- It has been suggested that signals other than those directly detectable in sequences may improve gene prediction. For example, the role of secondary structure in the identification of regulatory motifs has been reported. In addition, it has been suggested that RNA secondary structure prediction helps splice site prediction.

Neural networks

- Artificial neural networks are computational models that excel at machine learning and pattern recognition. Neural networks must be trained with example data before being able to generalise for experimental data, and tested against benchmark data. Neural networks are able to come up with approximate solutions to problems that are hard to solve algorithmically, provided there is sufficient training data. When applied to gene prediction, neural networks can be used alongside other *ab initio* methods to predict or identify biological features such as splice sites. One approach involves using a sliding window, which traverses the sequence data in an overlapping manner. The output at each position is a score based on whether the network thinks the window contains a donor splice site or an acceptor splice site. Larger windows offer more accuracy but also require more computational power. A neural network is an example of a signal sensor as its goal is to identify a functional site in the genome.

COMPARITIVE GENOME APPROACHES

- As the entire genomes of many different species are sequenced, a promising direction in current research on gene finding is a comparative genomics approach.
- This is based on the principle that the forces of natural selection cause genes and other functional elements to undergo mutation at a slower rate than the rest of the genome, since mutations in functional elements are more likely to negatively impact the organism than mutations elsewhere. Genes can thus be detected by comparing the genomes of related species to detect this evolutionary pressure for conservation. This approach was first applied to the mouse and human genomes, using programs such as SLAM, SGP and TWINSCAN/N-SCAN and CONTRAST.

Multiple informants

- TWINSCAN examined only human-mouse synteny to look for orthologous genes. Programs such as N-SCAN and CONTRAST allowed the incorporation of alignments from multiple organisms, or in the case of N-SCAN, a single alternate organism from the target. The use of multiple informants can lead to significant improvements in accuracy.
- CONTRAST is composed of two elements. The first is a smaller classifier, identifying donor splice sites and acceptor splice sites as well as start and stop codons. The second element involves constructing a full model using machine learning. Breaking the problem into two means that smaller targeted data sets can be used to train the classifiers, and that classifier can operate independently and be trained with smaller windows. The full model can use the independent classifier, and not have to waste computational time or model complexity re-classifying intron-exon boundaries. The paper in which CONTRAST is introduced proposes that their method (and those of TWINSCAN, etc.) be classified as *de novo* gene assembly, using alternate genomes, and identifying it as distinct from *ab initio*, which uses a target 'informant' genomes.

- Comparative gene finding can also be used to project high quality annotations from one genome to another. Notable examples include Projector, GeneWise, GeneMapper and GeMoMa. Such techniques now play a central role in the annotation of all genomes.

PSEUDOGENE PREDICTION

- Pseudogenes are close relatives of genes, sharing very high sequence homology, but being unable to code for the same protein product. Whilst once relegated as byproducts of gene sequencing, increasingly, as regulatory roles are being uncovered, they are becoming predictive targets in their own right. Pseudogene prediction utilises existing sequence similarity and ab initio methods, whilst adding additional filtering and methods of identifying pseudogene characteristics.
- Sequence similarity methods can be customised for pseudogene prediction using additional filtering to find candidate pseudogenes. This could use disablement detection, which looks for nonsense or frameshift mutations that would truncate or collapse an otherwise functional coding sequence. Additionally, translating DNA into proteins sequences can be more effective than just straight DNA homology.
- Content sensors can be filtered according to the differences in statistical properties between pseudogenes and genes, such as a reduced count of CpG islands in pseudogenes, or the differences in G-C content between pseudogenes and their neighbours. Signal sensors also can be honed to pseudogenes, looking for the absence of introns or polyadenine tails.

PROTEIN STRUCTURE AND FUNCTION PREDICTION

- **Protein structure prediction** is the inference of the three-dimensional structure of a protein from its amino acid sequence—that is, the prediction of its folding and its secondary and tertiary structure from its primary structure. Structure prediction is fundamentally different from the inverse problem of protein design. Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes). Every two years, the performance of current methods is assessed in the CASP experiment (Critical Assessment of Techniques for Protein Structure Prediction). A continuous evaluation of protein structure prediction web servers is performed by the community project CAMEO3D.

PROTEIN STRUCTURE

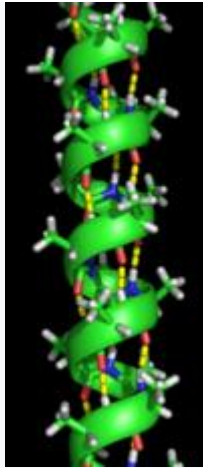
- Proteins are chains of amino acids joined together by peptide bonds. Many conformations of this chain are possible due to the rotation of the chain about each C α atom. It is these conformational changes that are responsible for differences in the three dimensional structure of proteins. Each amino acid in the chain is polar, i.e. it has separated positive and negative charged regions with a free carbonyl group, which can act as hydrogen bond acceptor and an NH group, which can act as hydrogen bond

donor. These groups can therefore interact in the protein structure. The 20 amino acids can be classified according to the chemistry of the side chain which also plays an important structural role. Glycine takes on a special position, as it has the smallest side chain, only one hydrogen atom, and therefore can increase the local flexibility in the protein structure. Cysteine on the other hand can react with another cysteine residue and thereby form a cross link stabilizing the whole structure.

- The protein structure can be considered as a sequence of secondary structure elements, such as α helices and β sheets, which together constitute the overall three-dimensional configuration of the protein chain. In these secondary structures regular patterns of H bonds are formed between neighboring amino acids, and the amino acids have similar Φ and Ψ angles.
- The formation of these structures neutralizes the polar groups on each amino acid. The secondary structures are tightly packed in the protein core in a hydrophobic environment. Each amino acid side group has a limited volume to occupy and a limited number of possible interactions with other nearby side chains, a situation that must be taken into account in molecular modeling and alignments.

α Helix

- The α helix is the most abundant type of secondary structure in proteins. The α helix has 3.6 amino acids per turn with an H bond formed between every fourth residue; the average length is 10 amino acids (3 turns) or 10 Å but varies from 5 to 40 (1.5 to 11 turns). The alignment of the H bonds creates a dipole moment for the helix with a resulting partial positive charge at the amino end of the helix. Because this region has free NH₂ groups, it will interact with negatively charged groups such as phosphates. The most common location of α helices is at the surface of protein cores, where they provide an interface with the aqueous environment. The inner-facing side of the helix tends to have hydrophobic amino acids and the outer-facing side hydrophilic amino acids. Thus, every third or fourth amino acids along the chain will tend to be hydrophobic, a pattern that can be quite readily detected. In the leucine zipper motif, a repeating pattern of leucines on the facing sides of two adjacent helices is highly predictive of the motif. A helical-wheel plot can be used to show this repeated pattern. Other α helices buried in the protein core or in cellular membranes have a higher and more regular distribution of hydrophobic amino acids, and are highly predictive of such structures. Helices exposed on the surface have a lower proportion of hydrophobic amino acids. Amino acid content can be predictive of an α -helical region. Regions richer in alanine (A), glutamic acid (E), leucine (L), and methionine (M) and poorer in proline (P), glycine (G), tyrosine (Y), and serine (S) tend to form an α helix. Proline destabilizes or breaks an α helix but can be present in longer helices, forming a bend.



-
- An alpha-helix with hydrogen bonds (yellow dots)

β-sheet

- β sheets are formed by H bonds between an average of 5–10 consecutive amino acids in one portion of the chain with another 5–10 farther down the chain. The interacting regions may be adjacent, with a short loop in between, or far apart, with other structures in between. Every chain may run in the same direction to form a parallel sheet, every other chain may run in the reverse chemical direction to form an anti parallel sheet, or the chains may be parallel and anti parallel to form a mixed sheet. The pattern of H bonding is different in the parallel and anti-parallel configurations. Each amino acid in the interior strands of the sheet forms two H bonds with neighboring amino acids, whereas each amino acid on the outside strands forms only one bond with an interior strand. Looking across the sheet at right angles to the strands, more distant strands are rotated slightly counter clockwise to form a left-handed twist. The C α atoms alternate above and below the sheet in a pleated structure, and the R side groups of the amino acids alternate above and below the pleats. The Φ and Ψ angles of the amino acids in sheets vary considerably in one region of the Ramachandran plot. It is more difficult to predict the location of β sheets than of α helices. The situation improves somewhat when the amino acid variation in multiple sequence alignments is taken into account.

Loop

- Loops are regions of a protein chain that are 1) between α helices and β sheets, 2) of various lengths and three-dimensional configurations, and 3) on the surface of the structure. Hairpin loops that represent a complete turn in the polypeptide chain joining two antiparallel β strands may be as short as two amino acids in length. Loops interact with the surrounding aqueous environment and other proteins. Because amino acids in loops are not constrained by space and environment as are amino acids in the core region, and do not have an effect on the arrangement of secondary structures in the core, more substitutions, insertions, and deletions may occur. Thus, in a sequence alignment, the presence of these features may be an indication of a loop. The positions of introns in genomic DNA sometimes correspond to the locations of loops in the encoded protein. Loops also tend to have charged and polar amino acids and are

frequently a component of active sites. A detailed examination of loop structures has shown that they fall into distinct families.

Coils

- A region of secondary structure that is not an α helix, a β sheet, or a recognizable turn is commonly referred to as a coil.

PROTEIN STRUCTURE

- **Primary structure**
 - the linear amino acid sequence of a protein, which chemically is a polypeptide chain composed of amino acids joined by peptide bonds..
- **Secondary structure**
 - the interactions that occur between the C, O, and NH groups on amino acids in a polypeptide chain to form α -helices, β -sheets, turns, loops, and other forms, and that facilitate the folding into a three-dimensional structure
- **Quaternary structure**
 - the three-dimensional configuration of a protein molecule comprising several independent polypeptide chains.

TYPES OF PROTEIN STRUCTURE PREDICTIONS

- PREDICTION in 1D
 - -secondary structure
 - -solvent accessibility
 - -transmembrane helices
- PREDICTION in 2D
 - -inter residue/strand contacts
- PREDICTION in 3D
 - -homology modelling
 - -fold recognition
 - -ab initio prediction

Energy- and fragment-based methods

- *Ab initio*- or *de novo*- protein modelling methods seek to build three-dimensional protein models "from scratch", i.e., based on physical principles rather than (directly) on previously solved structures. There are many possible procedures that either attempt to mimic protein folding or apply some stochastic method to search possible solutions (i.e., global optimization of a suitable energy function). These procedures tend to require vast computational resources, and have thus only been carried out for tiny proteins. To predict protein structure *de novo* for larger proteins will require better algorithms and larger computational resources like those afforded by either powerful supercomputers. Although these computational barriers are vast, the potential benefits of structural genomics (by predicted or experimental methods) make *ab initio* structure prediction an active research field.

Comparative protein modeling

- Comparative protein modelling uses previously solved structures as starting points, or templates. This is effective because it appears that although the number of actual proteins is vast, there is a limited set of tertiary structural motifs to which most proteins belong. It has been suggested that there are only around 2,000 distinct protein folds in nature, though there are many millions of different proteins.
- These methods may also be split into two groups:

Homology modeling

Homology modelling is based on the reasonable assumption that two homologous proteins will share very similar structures. Because a protein's fold is more evolutionarily conserved than its amino acid sequence, a target sequence can be modeled with reasonable accuracy on a very distantly related template, provided that the relationship between target and template can be discerned through sequence alignment. It has been suggested that the primary bottleneck in comparative modelling arises from difficulties in alignment rather than from errors in structure prediction given a known-good alignment. Unsurprisingly, homology modelling is most accurate when the target and template have similar sequences.

Protein threading

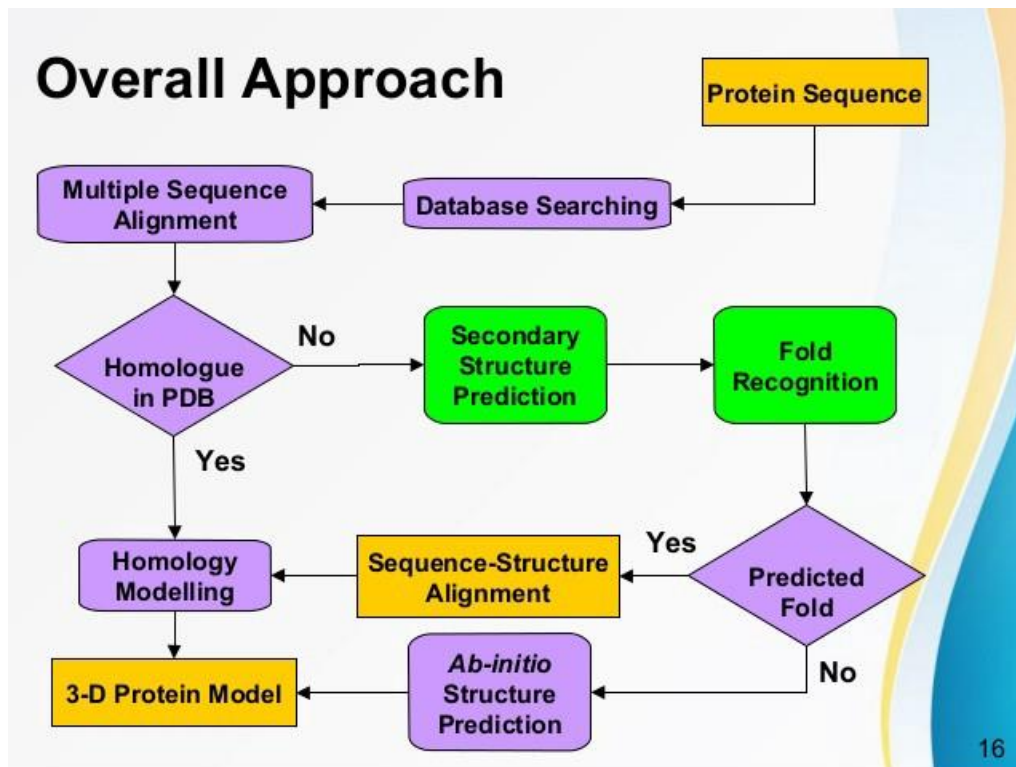
- Protein threading scans the amino acid sequence of an unknown structure against a database of solved structures. In each case, a scoring function is used to assess the compatibility of the sequence to the structure, thus yielding possible three-dimensional models. This type of method is also known as **3D-1D fold recognition** due to its compatibility analysis between three-dimensional structures and linear protein sequences. This method has also given rise to methods performing an **inverse folding search** by evaluating the compatibility of a given structure with a large database of sequences, thus predicting which sequences have the potential to produce a given fold.

Side-chain geometry prediction

- Accurate packing of the amino acid side chains represents a separate problem in protein structure prediction. Methods that specifically address the problem of predicting side-chain geometry include dead-end elimination and the self-consistent mean field methods. The side chain conformations with low energy are usually determined on the rigid polypeptide backbone and using a set of discrete side chain conformations known as "rotamers." The methods attempt to identify the set of rotamers that minimize the model's overall energy.
- These methods use rotamer libraries, which are collections of favorable conformations for each residue type in proteins. Rotamer libraries may contain information about the conformation, its frequency, and the standard deviations about

mean dihedral angles, which can be used in sampling. Rotamer libraries are derived from structural bioinformatics or other statistical analysis of side-chain conformations in known experimental structures of proteins, such as by clustering the observed conformations for tetrahedral carbons near the staggered (60° , 180° , -60°) values.

- Rotamer libraries can be backbone-independent, secondary-structure-dependent, or backbone-dependent. Backbone-independent rotamer libraries make no reference to backbone conformation, and are calculated from all available side chains of a certain type. Secondary-structure-dependent libraries present different dihedral angles and/or rotamer frequencies for α -helix, β -sheet, or coil secondary structures. Backbone-dependent rotamer libraries present conformations and/or frequencies dependent on the local backbone conformation as defined by the backbone dihedral angles and, regardless of secondary structure.
- The modern versions of these libraries as used in most software are presented as multidimensional distributions of probability or frequency, where the peaks correspond to the dihedral-angle conformations considered as individual rotamers in the lists. Some versions are based on very carefully curated data and are used primarily for structure validation, while others emphasize relative frequencies in much larger data sets and are the form used primarily for structure prediction, such as the Dunbrack rotamer libraries.
- Side-chain packing methods are most useful for analyzing the protein's hydrophobic core, where side chains are more closely packed; they have more difficulty addressing the looser constraints and higher flexibility of surface residues, which often occupy multiple rotamer conformations rather than just one.



Prediction in 2D

Inter-residue contacts

:Prediction problem is a hard one, but the stakes are high. Given all inter-residue contacts or distances 3D structure can be reconstructed by distance geometry or molecular dynamics. This is used for the determination of 3D structures by nuclear magnetic resonance (NMR) spectroscopy which produces experimental data of distances between protons. Can inter-residue contacts be predicted? Obviously, some fraction of these contacts can be: helices and strands can be assigned based on hydrogen-bonding pattern between residues. Thus, a successful prediction of secondary structure implies a successful prediction of some fraction of all the contacts. However, contacts predicted from secondary structure assignment are short-ranged, i.e., between residues nearby in sequence. For a successful application of distance geometry, long-range contacts have to be predicted, i.e., contacts between residues far apart in sequence. A few methods have been proposed for the prediction of long-range inter-residue contacts. Two questions surround such methods: first, how accurate are these prediction methods on average; and second, are all important contacts predicted?

- *Correlated mutations can imply spatial proximity.* In sequence alignments, some pairs of positions appear to co-vary in a physico-chemically plausible manner, i.e., a 'loss of function' point mutation is often rescued by an additional mutation that compensates for the change. One hypothesis is that compensations would be most effective in maintaining a structural motif if the mutated residues were spatial neighbours. Attempts have been made to quantify such a hypothesis and to use it for contact predictions. In general, prediction accuracy is rather poor, with a direct trade-off between predicting enough contacts, and predicting only correct ones, e.g., taking 5% of the best-predicted long-range contacts (sequence separation above 10 residues) the accuracy prediction is about 50% (A. Valencia, priv. communication).
- *Distinction between different models, no prediction of 3D, yet.* Analysing correlated mutations is only one way to predict long-range inter-residue contacts. Other methods use statistics, mean-force potentials, or neural networks. So far none of the methods appears to find a path between the Scylla of missing too many true contacts and the Charibdis of predicting too many false contacts. However, some of the methods provide sufficient information to distinguish between alternative models of 3D structure (Valencia, manuscript in preparation). The ambitious goal to predict long-range inter-residue contacts sufficiently accurately will hopefully continue to attract intellectual resources.

Inter-strand contacts

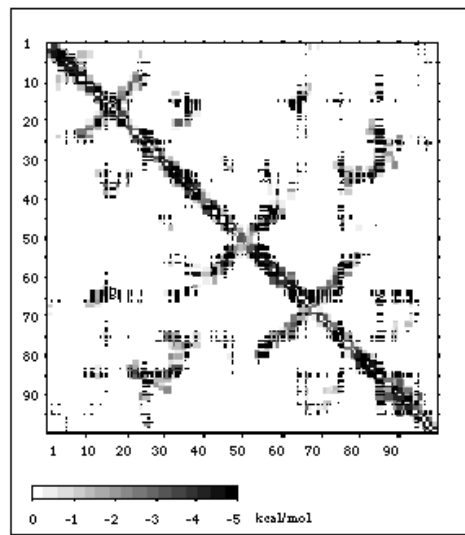
- *Simplifying the contact prediction problem.* One simplification of the problem to predict inter-residue contacts focuses on predicting the contacts between residues in adjacent strands. Such an attempt is motivated by the hope that such interactions are

more specific than are sequence-distant (long-range) contacts in general, and hence are easier to predict.

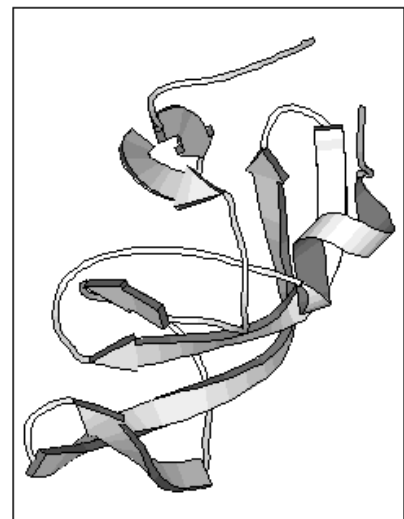
- *Identifying the correct b-strand alignment.* The only method published for predicting inter-strand contacts is based on potentials of mean-force similar to those used in the evaluation of strand-strand threading. Propensities are compiled by database counts for $2 \times 2 \times 2$ classes (parallel/anti-parallel, H-bonded/not H-bonded, N-/C-terminal). Each of the eight classes is divided further into five sub-classes in the following way. Suppose the two strand residues at positions i and j are in close in space. Then the following five residue pairs are counted in separate tables: $i/j-2$, $i/j-1$, i/j , $i/j+1$, $i/j+2$. Such pseudo-potentials identify the correct b-strand alignment in 35-45% of the cases.
- *Using evolutionary information to predict inter-strand contacts.* Even if the locations of strands in the sequence are known exactly, the pseudo-potentials cannot predict the correct inter-strand contacts in most cases. However, when using multiple alignment information, the signal-to-noise ratio increases such that inter-strand contacts have been predicted correctly for most of the strands inspected in some test cases. For the purpose of reliable contact prediction, this result is inadequate, especially as the locations of the strands are not known precisely. Can the pseudo-potentials handle errors resulting from incorrect prediction of strands? Various test examples using predictions by PHD sec as input to the strand pseudo-potentials indicate that the accuracy in predicting inter-strand contacts drops (T Hubbard, unpublished), but in some cases is still high enough to be useful for approximate modelling of 3D structure.

P	PP	P	128	110
Q	QQQY		175	97
I	FFQVI	E	70	60
T	SSIVR	E	77	69
L	LLSTL	E	120	14
W	WWQED	E	238	81
Q	RQQAQ	E	169	97
R	RRRPQ		200	63
P	PPPPP		24	49
L	VVTKF	E	71	59
V	VVLII	E	14	0
T	TTKEK	E	74	69
I	AALIV	E	0	0
K	HVKKF	E	90	73
I	ILLVI	E	4	0
G	EENGG		46	41
G	GGGTG		62	53
Q	QKRRR		68	71
L	PPLWW	E	118	59
K	VVFKY	E	31	73
E	EESKK	E	124	95
A	VVGLG	E	1	0
L	LLLLL	E	29	0
L	LLLVV	E	24	0
D	DDDDD	E	49	59
T	TTTTT		72	51
G	GGGGG		62	30
A	AAAAA		17	0
D	DDDDD		102	79
D	DDAKE		69	59
T	SSTTV		1	69
V	IIVIV	E	14	0
L	VVIVL	E	0	0

1D



2D



3D

Prediction in 3D

Homology modelling

- *Basic concept.* An analysis of PDB reveals that all protein pairs with more than 30% pairwise sequence identity have homologous 3D structures, i.e., the essential fold of the two proteins is identical, details such as additional loop regions - regions not in helices or strands - may vary. Structure is more conserved than is sequence. This is the pillar for the success of homology modelling. The principal idea is to model the structure of U (protein of unknown structure) based on the template of a sequence homologue of known structure. Consequently, the precondition for homology modelling is that a sequence homologue of known structure is found in PDB. Since homology modelling is currently the only theoretical means to successfully predict 3D structure, this has two implications. First, homology modelling is applicable to 'only' one quarter of the known protein sequences. Second, as the template of a homologue is required, no unique 3D structure can yet be predicted, i.e., no structure that has no similarity to any experimentally determined 3D structure.

High level of sequence identity: atomic resolution. The basic assumption of homology modelling is that U and T have identical backbones (main chain C). The task is to correctly place the side chains of U into the backbone of T . For very high levels of sequence identity between U and T (ideally differing by one residue only), side chains can be 'grown' during molecular dynamics simulations. For slightly lower levels (still of high sequence similarity), side chains are built based on similar environments in known structures. Rotamer libraries (libraries containing all side-chain orientations observed in known structures) are used in the following way. Rotamer distributions are extracted from a database of non-redundant sequences. Fragments of seven (helix, strand) or five residues (other) are compiled. Fragments of the same length are successively shifted through the backbone of U . For modelling the side chains of U only those fragments from the rotamer library are accepted which have the same amino acid in the centre as U , and for which the local backbone is similar to that around the evaluated position). Over the whole range of sequence identity between U and T for which homology modelling is applicable, the accuracy of the model drops with decreasing similarity. For levels of at least 60% sequence identity, the resulting models are quite accurate, for even higher values, the models are as accurate as is experimental structure determination. The limiting factor is the computation time required. How accurate is homology modelling for lower levels of sequence identity?

- *Low level of sequence identity: loop regions sometimes correct.* With decreasing sequence identity between the known structure H and the query protein U the number of loops that have to be inserted to align the two grows. An accurate modelling of loop regions, however, implies solving the structure prediction problem. The problem is simplified in two ways. First, loop regions are often relatively short and can thus be

simulated by molecular dynamics (note the CPU time required for molecular dynamics simulations grows exponentially with the number of residues of the polypeptide to be modelled). Second, the ends of the loop regions are fixed by the backbone of the template structure. Various methods are employed to model loop regions. The best have the orientation of the loop regions correct in some cases. This illustrates the current limitations of molecular dynamics: not even short loop regions can be predicted from sequence. Furthermore, for experimental structure refinement (use of molecular dynamics to improve consistency, and accuracy of experimental data) molecular dynamics is successfully applied to find a better solution when starting from an almost correct structure. However, for homology modelling, molecular dynamics refinement usually reduces prediction accuracy. Below about 40% sequence identity the accuracy of the sequence alignment used as basis for homology modelling becomes an additional problem. Nevertheless, even down to levels of 25-30% sequence identity, homology modelling produces coarse-grained models for the overall fold of proteins of unknown structure.

Remote homology modelling (threading)

- *Basic concept.* As noted in the previous section, naturally evolved sequences with more than 30% pairwise sequence identity have homologous 3D structures. Are all others non-homologous? Not at all. In the current PDB database there are thousands of pairs of structurally homologous pairs of proteins with less than 25% pairwise sequence identity (remote homologues). Actually, most similar protein structures are such remote homologues (Rost, in press). If a correct alignment between U (sequence of unknown structure) and a remote homologue T is given, one could build the 3D structure of U by (remote) homology modelling based on the template of T . A successful remote homology modelling must solve three different tasks. (1) The remote homologue (T) has to be detected. U and T have to be correctly aligned. The homology modelling procedure has to be tailored to the harder problem of extremely low sequence identity (with many loop regions to be modelled). Most methods developed so far have been primarily addressed to solve the first, and the second problem. The basic idea is to thread the sequence of U into the known structure of T and to evaluate the fitness of sequence for structure by some kind of environment-based or knowledge-based potential. Threading is in some respects a harder problem than is the prediction of 3D structure. However, the stakes are high: solving the threading problem could enable the prediction of thousands of protein structures. Indeed, threading has evolved to become one of the most active fields in the arena of protein structure prediction.
- *Variety of threading techniques.* The optimism generated by one of the first papers on threading published in the 90s has boosted attempts to develop threading methods. The principle idea has been to use structural propensities of amino acids (such as preferences for secondary structure formation, hydrophobicity, and polarity), and to then assess whether or not a given sequence with its structural preferences fits into the structural environment of a given structure. A principally different approach has been

pushed by Manfred Sippl. The idea is to use the rich knowledge deposited in the database of protein structures (PDB) by extracting mean-force potentials. Such potentials monitor the observed distances between residue pairs of particular amino acids, with a particular sequence separation (number of residues between the two). Until 1995, most threading methods have used mean-force-potentials. A more recent generation of threading methods is based on 1D prediction: first 1D structure (secondary structure and solvent accessibility) is predicted for a sequence of unknown structure, then the 1D structure is extracted for a library of known structures, and finally the observed and the predicted 1D structure strings are aligned by typical dynamic programming algorithms. Has all this effort achieved to crack the hard nut threading?

- *Remote homologues can often be detected.* First the good news: since the different mean-force-potentials which have been proposed capture different aspects of protein structure, the correct remote homologue is likely to be found by at least one of them. Now the bad news: so far, no single method has been able to detect the correct remote homologue for more than half of all test cases. For the methods which have been rigorously evaluated using large test sets, the correct remote homologue is detected in less than 40% of all cases. However, this performance is clearly superior to that of traditional sequence alignments at this low level of sequence identity. Furthermore, the success of the last Asilomar experiment on structure prediction (forthcoming issue of Proteins) suggests that the likelihood to detect the correct remote homologue is reasonably high when the choice is refined by experts.
- *3D prediction by threading is still not reliable.* Detecting the remote homology is only the first of the three obstacles. It appears that the second obstacle (correct alignment between U and T) is much more difficult and, unfortunately, there is no general solution so far. Thus the final step, building a 3D model, usually fails since the modelling procedures available today cannot correct the mistakes in the alignments. Although the last Asilomar experiment on structure prediction (forthcoming issue of Proteins) suggested that major improvements have been accomplished over the last two years, there are still very few publications to date which report accurate 3D predictions from threading methods. Currently, the successful use of threading methods requires sceptical, expert user intervention to spot wrong hits and false alignments. It is still possible that threading method will become the most successful structure prediction method, but a lot of detailed work lies ahead.

Conclusions

- Native 3D structures of proteins are encoded by a linear sequence of amino acid residues. To predict 3D structure from sequence is a task challenging enough to have occupied a generation of researchers. Have they finally succeeded in their goal? The bad news is: no, we still cannot predict structure for any sequence. The good news are: we have come closer, and growing databases facilitate the task.
- *Prediction in 3D: theory bridges the sequence-structure gap.* The only source for new, unique protein structures (structures for which no homologue exists in the database) are experiments. However, given the amount of time needed to determine a

protein structure experimentally, more non-unique structures can be predicted at atomic resolution by homology modelling in a month than have been determined by experiment over the last three decades. Homology derived models are frequently accurate at the level of atomic resolution. Unfortunately, most models typically have considerable co-ordinate errors in loop regions. Coarse-grained homology derived models are available for almost one third of the sequences deposited in the SWISS-PROT database. Threading techniques could increase this ratio considerably by finding more distant homologues. However, for large scale sequence analyses threading techniques are not yet reliable.

- *Predictions in 2D: so far of limited success.* The prediction accuracy of chain-distant inter-residue contacts is so far relatively limited. Analysis of correlated mutations can be used to distinguish between alternative models (e.g. for threading techniques). The prediction of inter-strand contacts appears to be useful in some cases. An accurate method for the automatic prediction of contacts between residues not close in sequence remains to be developed

REFERENCES

- https://en.wikipedia.org/wiki/Gene_prediction
- https://en.wikipedia.org/wiki/Protein_structure_prediction
- https://www.rostlab.org/papers/1998_encyclopedia/paper.html

Texts/References

01. Casella G. and Berger R. L., Statistical Inference (The Wadsworth and Brooks/Cole Statistics/Probability Series) b, Brooks/Cole Pub Company.
02. Grant G. R., Ewens W.J. , Statistical Methods in Bioinformatics: An Introduction. Springer Verlag.
03. Jagota A. Data Analysis and Classification for Bioinformatics, Bioinformatics By The Bay Press.
04. Spiegel M. R., Schiller J.J., Srinivasan R. A. , A. Srinivasan Schaum's Outline of Probability and Statistics. McGraw-Hill Trade.
05. Cynthia Gibas, Per Jambeck, "Developing Bioinformatics Computer Skills", O'Reilly Media, Inc., 2001.
06. David Edwards, Jason Eric Stajich, David Hansen, "Bioinformatics: Tools and Applications", Springer, 2009.
- 07 David W Mount, "Bioinformatics: Sequence and genome analysis", Cold spring harbor laboratory press, 2nd edition, 2004.
08. Stan Tsai C., "Biomacromolecules: Introduction to Structure, Function and Informatics", John Wiley & Sons, 2007.
09. Attwood T K, D J Parry-Smith, "Introduction to Bioinformatics", Pearson Education, 2005.
10. Parag Rastogi, "Bioinformatics Methods And Applications: Genomics Proteomics And Drug Discovery", PHI Learning Pvt. Ltd., 3rd edition, 2008

-----The End---