S.Y.B.Sc. Life Sciences **Paper III SEM III 2019**

QP code 53022 **Date 30/03/2019**

Answer Key

Q. 1 Do as directed:

Q. 1 A Define/Explain:

1. Entrez - Entrez Global Query is an integrated search and retrieval system that provides access to all databases simultaneously with a single query string and user interface. Entrez can efficiently retrieve related sequences, structures, and references. The Entrez system can provide views of gene and protein sequences and chromosome maps. Some textbooks are also available online through the Entrez system.

2. Backward/reverse reading frame - DNA encodes protein sequence by a series of three-nucleotide codons. Any given sequence of DNA can therefore be read in six different ways: Three reading frames in one direction (starting at different nucleotides) and three in the opposite direction, which are called as reverse reading frame.

3. Bioperl - BioPerl is a collection of Perl modules that facilitate the development of Perl scripts for bioinformatics applications. It has played an integral role in the Human Genome Project.

4. FTP - The File Transfer Protocol is a standard network protocol used for the transfer of computer files between a client and server on a computer network. FTP is built on a client-server model architecture using separate control and data connections between the client and the server.

5. Molecular modeling - Molecular modelling is a collection of (computer based) techniques for deriving, representing and manipulating the structures and reactions ofmolecules, and those properties that are dependent on these three dimensional structures.

6. Functional Genomics - Functional genomics is a field of molecular biology that attempts to make use of the vast wealth of data given by genomic and transcriptomic projects (such as genome sequencing projects and RNA sequencing) to describe gene (and protein) functions and interactions.

7. Stop codon- In the genetic code, a stop codon is a nucleotide triplet within messenger RNA that signals a termination of translation into proteins. Proteins are based on polypeptides, which are unique sequences of amino acids.

Q. 1 B Match the Columns:

a. - v
b. - iv
c. - vi
d. - viii
e. - vii
f. - ii
g. - iii

Q.1 C True or False **(06)**

1. False

2. True

3. False

4. False

5. True

6. True

Q. 2 A. Answer any one of the following: **(10)**

1. **Explain any three processes that bring in the change due to Evolution.**

Evolution is the process by which modern organisms have descended from ancient ancestors. Evolution is responsible for both the remarkable similarities and diversity in species.

Fundamental to the process is genetic variation upon which selective forces can act in order for evolution to occur.

- Descent and the genetic differences that are heritable and passed on to the next generation;
- Mutation, migration (gene flow), genetic drift, and natural selection as mechanisms of change;
- The importance of genetic variation
- The random nature of genetic drift and the effects of a reduction in genetic variation;
- Variation, differential reproduction, and heredity result in evolution by natural selection
- Coevolution between different species can affect each other's evolution
-

1. **Genetic Drift: Founders Effect**

The founder effect is a special case of genetic drift, occurring when a small group in a population splinters off from the original population and forms a new one. The new colony may have less genetic variation than the original population, and through the random sampling of alleles during reproduction of subsequent generations, continue rapidly towards fixation. This consequence of inbreeding makes the colony more vulnerable to extinction.

2. **Mutation**

Mutation can be defined as a change in the DNA sequence within a gene or chromosome of a living organism. Many mutations are neutral, i.e. they can neither harm nor benefit, but can also be deleterious or beneficial. Deleterious mutations can affect the phenotype and in turn, reduce the fitness of an organism and increase the susceptibility to several illnesses and disorders. On the other hand, beneficial mutations can lead to the reproductive success and adaptability of an organism to its environment. These beneficial mutations can be spread and fixed in the population due to natural selection processes if they help individuals in the population to reach sexual maturity and to successfully reproduce. Mutations are, undoubtedly, a source of genetic variation and serve as a raw material for evolution to act. Germ line

mutations occur in gametes (eggs or sperm cells) and can be pass on to offspring, whereas somatic mutations occur in non-reproductive cells and are not pass on to the following generation. Those mutations that occur in germ line are the most important to large-scale evolution because they can be transmitted to offspring.

3. **Migration, or Gene Flow:**

Alleles can flow from one population to another when animals migrate and begin to interbreed in new localities, or when there is deliberate crossing of breeds or subpopulations within breeds. High rates of gene flow can reduce the genetic differentiation between the two groups, increasing homogeneity. For this reason, gene flow has been thought to constrain speciation by combining the gene pools of the groups, thus preventing the development of differences in genetic variation that would have led to full speciation. In some cases migration may also result in the addition of novel genetic variants to the gene pool of a species or population.

2. **Explain any two processes that alter the Hardy Weinberg Equilibrium and bring about a change in the Allelic frequencie**

The Hardy–Weinberg principle, also known as the Hardy–Weinberg equilibrium, model, theorem, or law, states that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences. These influences include genetic drift, mate choice, assortative mating, natural selection, sexual selection, mutation, gene flow, meiotic drive, genetic hitchhiking, population bottleneck, founder effect and inbreeding.

A population must be large enough that chance occurrences cannot significantly change allelic frequencies significantly. To better understand this point, consider the random flipping of a fair coin. The coin is as likely to land on heads as it is on tails. If a coin is flipped 1000 times, it is likely to land on heads almost exactly 50% of the time. However, as you may know from experience, if the same coin is flipped only ten times, it is much less likely that it will land on heads 5 times. The same holds true for allele distributions in populations. Large populations are unlikely to be affected by chance changes in allele frequencies because those chance changes are very small in relation to the total number of allele copies. But in small populations with fewer copies of alleles, chance can greatly alter allele frequencies. In small populations, a change in allelic frequencies and phenotypes based on random occurrences is called genetic drift.

No Mutation

In order for allelic frequencies to remain constant, there must be no change in the number of copies of an allele due to mutation. This condition can be met in two ways. A population can experience little or no mutation. Alternatively, it can experience balanced mutation. Balanced mutation occurs when the rate at which copies of a given allele are lost to mutation equals the rate at which new copies are created by mutation.

No Immigration or Emigration

For allelic frequencies to remain constant in a population, individuals must not move in and out of that population. Whenever an individual enters or exits a population, it takes copies of alleles with it, changing the overall frequency of those alleles in the population.

Random Mating

In order for all alleles to have an equal chance of being passed down to the next generation, mating within the population must be random. Non-random mating can give an advantage to certain alleles, allowing them to be passed down to more offspring than other alleles, increasing their relative frequency in the population. The processes of natural selection, since

they usually select for individuals with greatest fitness for a given environment, usually work against random mating: the most fit organisms are most likely to mate.
Random Reproductive Success
Just as mating must be random, the survival of offspring to reproductive age, or reproductive success, must also be random. Again, natural selection usually works against such randomness.

Q. 2 B. Answer any two of the following: (10)

1. Any two types of Natural Selection processes

a) Stabilizing Selection
Most traits in the animal kingdom can be described by a bell curve, in terms of their distribution. Most animals of a certain species tend to show the same trait or feature, of relatively the same size. There are always some exceptions of larger or smaller traits in certain individuals, but generally most individuals sit somewhere in the middle. Stabilizing selection is a form of natural selection that screens against the outliers, or exceptions to the trait. The screen prevents those animals from reproducing as much as the "normal" or more regular individuals. Because of this bias, more babies are born that are "normal" and less outliers are seen in each consecutive generation. It is in this way that species can become very distinct from other species, yet all members of a species will look exactly alike.
Example of Stabilizing Selection
Because actual life is really complex, all of the following examples will be simplified hypothetical versions. For stabilizing selection, imagine a population of mice that lives in the woods. Some of the mice are black, some are white, and some are grey. If the mice had no predators, and no other forces acting on the color of their coat, it would have no reason to change and would only change randomly in response to certain mutations in the DNA. However, that is not the case with these mice. They have lots of predators. In the daytime, the mice are preyed on by foxes and house cats from a nearby village. In the nighttime, the owls and other predators scour the dark for dinner. Either way, the mice are in a tough position. But not all the mice face the same risk at all times. In the daytime, the black mice are much easier to spot, and predators eat more black mice. At night, the white mice stand out like little lights to the owls, and many more white mice are eaten. The grey mice are the only ones who survive more both during the day and at night. By the next generation, there will be many less black and white mice to reproduce. The genes that produce the grey mice will be replicated more than the others, creating many more grey mice in the next generation.After a few generations of a strong selective pressure, the entire population could be grey. It depends entirely on the genetic makeup of the trait, but in some cases a single trait is selected for and the rest are lost from the population. In other cases, the black and white traits could just become rarely seen traits. Retaining the traits can be an advantage when the predators change. For instance, if all the owls and night predators disappeared, it would be more beneficial to be black. The black mice would then take off, and become more frequent in the population.

b) Disruptive Selection Process:

Disruptive selection is a type of natural selection that selects *against* the average individual in a population. The makeup of this type of population would show phenotypes (individuals with groups of traits) of both extremes but have very few individuals in the middle. Disruptive selection is the rarest of the three types of natural selection and can lead to the

deviation in a species line.Basically, it comes down to the individuals in the group who get to mate—who survive best. They are the ones who have traits on the extreme ends of the spectrum. The individual with just middle-of-the-road characteristics is not as successful at survival and/or breeding to further pass on "average" genes. In contrast, a population functions in *stabilizing selection* mode when the intermediate individuals are the most populous. Disruptive selection occurs in times of change, such as habitat change or change in resources availability.Disruptive selection can be influenced by human interaction. Environmental pollution can drive disruptive selection to choose different colorings in animals for survival.

Disruptive Selection Examples: Color- in regards to camouflage, serves as a useful example in many different kinds of species, because those individuals that can hide from predators the most effectively will live the longest. If an environment has extremes, those who don't blend into either will be eaten the most quickly, whether they're moths, oysters, toads, birds or another animal. Eg: Industrial Melanism- Peppered moths

2. Any one of factor changing Allelic frequency.

Genetic Isolation
Genetic isolation refers to the separation of a potentially interbreeding population of animals into two or more groups that do not exchange genes. While reproductive isolation does not in itself change gene frequencies, it is often a precondition to such changes. Genetic isolation can occur as a result of geographic barriers, as was the case in the early development of the Channel Island breeds of cattle, the Jersey and Guernsey .
Genetic Drift
Random Drift consists of random fluctuations in the frequency of appearance of a gene, usually, in a small population. The process may cause gene variants to disappear completely, thereby reducing genetic variability. In contrast to natural selection, environmental or adaptive pressures do not drive changes due to genetic drift. The effect of genetic drift is larger in small populations and smaller in large populations.Genetic drift describes random fluctuations in the numbers of gene variants in a population. Genetic drift takes place when the occurrence of variant forms of a gene, called alleles, increases and decreases by chance over time. These variations in the presence of alleles are measured as changes in allele frequencies. Typically, genetic drift occurs in small populations, where infrequently occurring alleles face a greater chance of being lost. Once it begins, genetic drift will continue until the involved allele is either lost by a population or until it is the only allele present in a population at a particular locus. Both possibilities decrease the genetic diversity of a population. Genetic drift is common after population bottlenecks, which are events that drastically decrease the size of a population. In these cases, genetic drift can result in the loss of rare alleles and decrease the gene pool. Genetic drift can cause a new population to be genetically distinct from its original population, which has led to the hypothesis that genetic drift plays a role in the evolution of new species.
Or explain Mutation/Migration or Gene Flow

3. Compare Natural and Artificial selection:

Natural selection is a nature-made selection, and artificial selection is a man-made selection. The **main difference** between natural and artificial selection is that **natural selection produces a great biological diversity whereas artificial selection produces varieties of organisms such as improved crops and livestock.** Artificial selection, which is also called **selective breeding**, is mainly used in domestic populations. It is mainly used to maintain only beneficial traits over generations. However, natural selection only allows the favorable traits for the environment to be inherited by successive generations.

NATURAL SELECTION
VERSUS
ARTIFICIAL SELECTION

NATURAL SELECTION	ARTIFICIAL SELECTION
Natural selection is the process whereby organisms better adapted to their environment tend to survive and produce more offspring	Artificial selection is the process by which animals and plants are chosen by the breeder to produce desirable and inheritable characters in the successive generations
Nature-made selection process	Man-made selection process
Produces a huge biological diversity	Produces organisms with selected traits
Occurs in natural populations	Mainly occurs in domestic populations
Only allows favorable characters to be inherited over the successive generations	Allows only selected traits to be inherited over successive generations
A slow process	A rapid process
Facilitates evolution through generating biological diversity	Does not facilitate evolution
Examples: Selection of long-necked giraffes, and change in size and shape of beaks of birds upon the available food	Examples: Breeding of small dogs such as Chihuahua, and cattle which can produce more milk

4.Analogous structures as evidence of evolution.

There are many types of evidence for evolution, including studies in the molecular biology field (like DNA) and also in the developmental biology field. However, the most commonly used types of evidence for evolution are anatomical comparisons between species. While homologous structures show how similar species have changed from their ancient ancestors, analogous structures show how different species have evolved to become more similar.

Speciation

Speciation is the change over time of one species into a new species. So why would different species become more similar? Usually, the cause of convergent evolution is similar selection pressures in the environment. In other words, the environments in which the two different species live are similar and those species need to fill the same niche in different areas around the world. Since natural selection works in the same way in these types of environments, the same types of adaptations are favorable and those individuals with those favorable adaptations survive long enough to pass down their genes to their offspring. This continues until only individuals with favorable adaptations are left in the population.

Sometimes, these types of adaptations can change the structure of the individual. Body parts can be gained, lost, or rearranged depending on whether or not their function is the same as the original function of that part. This can lead to analogous structures in different species that occupy the same type of niche and environment in different locations.

Taxonomy

When Carolus Linnaeus first began classifying and naming species with taxonomy, he often grouped similar looking species into similar groups. This led to incorrect groupings when compared to actual evolutionary origins of the species. Just because species look or behave the same does not mean they are closely related.

Analogous structures do not have to have the same evolutionary path. One analogous structure may have come into existence long ago, while the analogous match on another species may be relatively new. They may go through different developmental and functional stages before they are fully alike. Analogous structures are not necessarily evidencing that two species came from a common ancestor. It is actually more likely they came from two separate branches of the phylogenetic tree and may not be closely related at all.

Examples of Analogous Structures

The eye of a human is very similar in structure to the eye of the octopus. In fact, the octopus eye is superior to the human eye in that it does not have a "blind spot". Structurally, that is really the only difference between the eyes. However, the octopus and the human are not closely related and reside far away from each other on the phylogenetic tree of life.Wings are a popular adaptation for many animals. Bats, birds, insects, and pterosaurall had wings. A bat is more closely related to a human than a bird or insect based on homologous structures. Even though all these species have wings and can fly, they are very different in other ways. They just all happen to fill the flying niche in their locations.Sharks and dolphins look very similar in their appearance due to color, placement of their fins, and overall body shape. However, sharks are fish and dolphins are mammals. This means that dolphins are more closely related to rats than they are sharks on the evolutionary scale. Other types of evolutionary evidence, like DNA similarities, have proven this.It takes more than looks to determine which species are closely related and which have evolved from different ancestors to become more similar through their analogous structures. However, analogous structures themselves are evidence for the theory of natural selection and the accumulation of adaptations over time.

Q.3 A) Solve any one of the following: **(10)**

1.

Heights of father in inches	65	66	67	68	69	70	71
Ht. of sons in gms	67	68	66	69	72	72	69

X	dx (68)	dx^2	Y	dy (69)	dy^2	dx.dy
65	-3	9	67	-2	4	6
66	-2	4	68	-1	1	-2
67	-1	1	66	-3	9	3
68	0	0	69	0	0	0
69	1	1	72	3	9	3
70	2	4	72	3	9	6
71	3	9	69	0	0	0
$\sum X=476$	$\sum dx=0$	$\sum dX^2=$ 28	$\sum Y=483$	$\sum dy=0$	$\sum dy^2=32$	$\sum dx.dy=16$

$r= \sum dx.dy / \sqrt{\sum dX^2 . \sum dy^2}$

$= 16/ \sqrt{28.32} =0.53$

2. From the following data obtain regression equation X on Y:

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

What would be X when Y=10.5?

For equation X on Y

For calculation of b

Therefore $b=N\sum x.y-\sum X.\sum y / N\sum y^2-(\sum y)^2$

X	X^2	Y	Y^2	X.Y
1	1	9	81	9
2	4	8	64	16
3	9	10	100	30
4	16	12	144	48
5	25	11	121	55
6	36	13	169	78

7	49	14	196	98
8	64	16	256	128
9	81	15	225	135
$\sum X=45$	$\sum X^2=285$	$\sum y=108$	$\sum y^2=1356$	$\sum x.y=597$

Therefore $b= (9 \times 597)- (45 \times 108) / (9 \times 1356)- (108)^2$

$= 513 / 540$

$= 0.95$

Also $\sum x= nA + b\sum y$

$45=9A+0.95 \times 108$

$A=-6.4$

Equation of X on Y

$X= a+bY$

For $X= 6.4 + 0.95 \times 10.5$

Therefore $X= 16.37$

Q.3 B) Answer any two of the following: (10)

1. P (one bulb defective) = $e^{-m} X m^r / r!$

Where $m = n.p = 100 X 5/100$

i.e. $m= 5$

P (one bulb defective) = $0.007 X 5^1 / 1$

$= 0.035$

2. P(one economist) = ${}^3C_1 / {}^{10}C_4$

P(one engineer)= ${}^4C_1 / {}^{10}C_4$

P(one statistician) = ${}^2C_1 / {}^{10}C_4$

P(one doctor) = ${}^1C_1 / {}^{10}C_4$

P(one of each together) = ${}^3C_1 / {}^{10}C_4 + {}^4C_1 / {}^{10}C_4 + {}^2C_1 / {}^{10}C_4 + {}^1C_1 / {}^{10}C_4$

$= 24/210 = 4/35$

3. There are 366 days in a leap year, will have 52 weeks and 2 days over

The 2 days can have Sunday-Monday, Monday-Tuesday, Tuesday-Wednesday, Wednesday-Thursday, Thursday-Friday, Friday-Saturday and Saturday-Sunday

Theses are 7 possible combinations

Therefore Probability= 2/7

4. Let X be the normal variate showing measurements.

Its mean $\mu = 65.5$ and $\sigma=6.2$

When X=54.8 $Z= X-\mu / \sigma$

54.8-65.5 /6.2

-1.72

When X=68.8, Z=68.8-65.5 / 6.2

o.53

Area under the curve -1.72 to 0.53 = area under 0 to 0.53 + 0 to -1.72

i.e. (0.5-0.09) + (0.5-0.3407) =0.191

%age of measurements between 54.8 to 65.5 are 19.1 %

Q. 4 A. Discuss any one of the following: (10)

1. Protein Data bases:

In the field of bioinformatics, a sequence database is a type of biological databasethat is composed of a large collection of computerized ("digital") nucleic acidsequences, protein sequences, or other polymer sequences stored on a computer.A Protein sequence database is a database usually secondary or composite database which contains information on proteins w.r.t their nucleotide sequences, translated protein sequences, molecular information, 3-D structures, interactions and functions. The UniProt database, ExPaSy database, PDB is an example of protein databases.A protein structure database is a database that is modeled around the various experimentally determined protein structures. The aim of most protein structure databases is to organize and annotate the protein structures, providing the biological community access to the experimental data in a useful way. Data included in protein structure databases often includes three-dimensional coordinates as well as experimental information, such as unit cell dimensions and angles for x-ray crystallography determined structures. Though most instances, in this case either proteins or a specific structure determinations of a protein, also contain sequence information and some databases even provide means for performing sequence based queries, the primary attribute of a structure database is structural information, whereas sequence databases focus on sequence information, and contain no structural information for the majority of entries. Protein structure databases are critical for many efforts in computational biology such as structure based drug design, both in developing the computational methods used and in providing a large experimental dataset used by some methods to provide insights about the function of a protein. PDB TM, Swiss-MODEL SCOP are some of the example of Protein structure Databases. (Students can explain any one).

Protein Data Bank (**PDB**) is a database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography, NMR spectroscopy, or, increasingly, cryo-electron microscopy, and submitted by biologists and biochemists from around the world, are freely accessible on the Internet via the websites of its member organisations (PDBe, PDBj, and RCSB). The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB. The PDB is a key in areas of structural biology, such as structural genomics. Most major scientific journals, and some funding

agencies, now require scientists to submit their structure data to the PDB. Many other databases use protein structures deposited in the PDB. For example, SCOP and CATH classify protein structures, while PDBsum provides a graphic overview of PDB entries using information from other sources, such as Gene ontology.PDB archive the three-dimensional structures of not only proteins but also all biologically important molecules, such as nucleic acid fragments, RNA molecules, large peptides such as antibiotic gramicidin and complexes of protein and nucleic acids. The database holds data derived from mainly three sources. Structure determined by X-ray crystallography form the large majority of the entries.This is followed by structures arrived at by NMR experiments. There are also structures obtained by molecular modelling. The data in the PDB is organized as flat files, one to a structure, which usually means that each file contain one molecule, or one molecular complex.

2. Process of ORF Finding:

In molecular genetics, an open reading frame (ORF) is the part of a reading frame that has the potential to be translated. An ORF is a continuous stretch of codons that contain a start codon (usually AUG) and a stop codon (usually UAA, UAG or UGA). An ATG codon within the ORF (not necessarily the first) may indicate where translation starts. The transcription termination site is located after the ORF, beyond the translation stop codon. If transcription were to cease before the stop codon, an incomplete protein would be made during translation. In eukaryotic genes with multiple exons, ORFs span intron/exon regions, which may be spliced together after transcription of the ORF to yield the final mRNA for protein translation.

A **reading frame** refers to one of three possible ways of reading a nucleotide sequence.
Let's say we have a stretch of 15 DNA base pairs:

```
acttagccgggacta
```

We can start translating, or reading, the DNA from the first letter, 'a,' which would be referred to as the first reading frame.Or we can start reading from the second letter, 'c,' which is the second reading frame.
Or we can start reading from the third letter, 't,' which is the third reading frame.The reading frame affects which protein is made. In the example below, the upper case letters represent amino acids that are coded by the three letters above and to the left of them.

```
reading frame:             123
                           |||
acttacccgggacta
first reading frame          T  Y  P  G  L
second reading frame          L  T  R  D
third reading frame            L  P  G  T
```

The illustration above shows three reading frames. However, there are actually **six reading frames**: three on the positive strand, and three (which are read in the reverse direction) on the negative strand. Six Frame Translation allows the researcher to find the most likely /probable ORF that can be translated and converted into a protein, functional unit of DNA coding sequence.

One common use of open reading frames (ORFs) is as one piece of evidence to

assist in gene prediction. Long ORFs are often used, along with other evidence, to initially identify candidate protein-coding regions or functional RNA-coding regions in a DNA sequence. The presence of an ORF does not necessarily mean that the region is always translated. For example, in a randomly generated DNA sequence with an equal percentage of each nucleotide, a stop-codon would be expected once every 21 codons. A simple gene prediction algorithm for prokaryotes might look for a start codon followed by an open reading frame that is long enough to encode a typical protein, where the codon usage of that region matches the frequency characteristic for the given organism's coding regions. By itself even a long open reading frame is not conclusive evidence for the presence of a gene. On the other hand, it has been proven that some short ORFs (sORFs) that lack the classical hallmarks of protein-coding genes (both from ncRNAs and mRNAs) can produce functional peptides. 5'NTR of about 50% of mammal mRNAs are known to contain one or several sORFs. 64–75% of experimentally found translation initiation sites of sORFs are conserved in the genomes of human and mouse and may indicate that these elements have function.

Q. 4 B. Describe any Two of the following: (10)

1. Role of NCBI in developing Bioinformatics:

NCBI and its goals, objectives and service provided

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information. It was established as the National Center for Biotechnology Information (NCBI) on November 4, 1988, as a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH).As a national resource for molecular biology information, NCBI's mission is to develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease. More specifically, the NCBI has been charged with creating automated systems for storing and analyzing knowledge about molecular biology, biochemistry, and genetics; facilitating the use of such databases and software by the research and medical community; coordinating efforts to gather biotechnology information both nationally and internationally; and performing research into advanced methods of computer-based information processing for analyzing the structure and function of biologically important molecules. To carry out its diverse responsibilities, NCBI:

- conducts research on fundamental biomedical problems at the molecular level using mathematical and computational methods
- maintains collaborations with several NIH institutes, academia, industry, and other governmental agencies
- fosters scientific communication by sponsoring meetings, workshops, and lecture series
- supports training on basic and applied research in computational biology for postdoctoral fellows through the NIH Intramural Research Program
- engages members of the international scientific community in informatics research and training through the Scientific Visitors Program

- develops, distributes, supports, and coordinates access to a variety of databases and software for the scientific and medical communities
- develops and promotes standards for databases, data deposition and exchange, and biological nomenclature

2. Programming Languages in Bioinformatics

Perl: Flexible, by a global repository (CPAN), thus it is small install new modules. It has Bio per, one of the first biological unit repositories that upsurge the usability from, for instance, change setups to do the phylogenetic investigation. There are several biological software those usages Perl such as GBrowse.

Python: Influential, flexible, plus easy to use, Python is a perfect language for constructing software tools plus applications for life science study and development. Bioinformatics Programming Using Python is faultless for anybody involved with bioinformatics investigators, support staff, students, as well as software developers fascinated in writing bioinformatics apps. By using Python, one can write code in the additional language, and otherwise have no programming experience. It's an outstanding self-instruction tool, in addition to a convenient reference while facing the challenges of real-life programming jobs. Clear language. Typically there is only one method to program with Python. "The correct one". It is simple plus stable. Biopython does not have as numerous modules as Bioperl, but they work. The Python language was intended to be as simple plus accessible as likely, without giving up any of the power required to develop stylish applications. Python's clean, steady syntax leaves it free from the delicacies and nuances that could make additional languages hard to learn and programs written in those languages hard to comprehend. Python's vibrant nature adds to its convenience. For instance, Python does not require to declare variables beforehand you use them, and the similar variable can mention to objects of different kinds over the course of its presence. Python can be moreover be used interactively.

R: It is a statistical programming language, thus it opens a world of analysis, from t-test toward PCA plus clustering. If you are going to do RNAseq study may be vital if you don't want to use paid software since 75% of RNAseq statistical sets are from Bioconductor (biological software repository for R). For instance, CummeRbund is an R package toward analyzing the outcomes from Cufflinks (a program toward calculating expression for RNAseq trials). It has the central repository (CRAN) thus install packages is easy. Graphs, graphs are just great with R. It has R-studio that is atmosphere software to usage R in a Matlab fashion. In brief, for persons who want toward add bioinformatics toward their toolbox, emphasis on education R first. R is free plus open source programming language that yields attractive graphics. R language is extensively used among the numerical community plus more lately in the data science as well as machine learning communal as well. Because of this fact, this has hire freelancers in recent ages as a stage for displaying plus delivering operative and applied BI.

C and C++ : C and C++ are excessive for making wonderful enhanced command-line tool similar aligners plus variant-callers, however you would have much calmer time education Python first in addition to then going toward the high-performance language for a specific

problematic in the future, meanwhile they are firmer toward learn, fussier, and take lots of additional code toward do the similar thing.

Ruby: Ruby is the hot language currently, for good cause largely owing to the control of Ruby on Rail for creating database-driven web apps like blogs otherwise twitter. Ruby, however, is not excessive for bioinformatics since it lacks communal support in term of package that R plus Python have

PHP and JavaScript: Java Script plus PHP are excessive language for web application; however, bioinformatics web application must never remain your first job. You might make computational technique in Python otherwise R plus then late create it into a web app; however that is not a mission for a novice. HTML plus CSS incidentally are not programming language, however actually mark up plus styling language that you would usage accompanied by JavaScript plus PHP for the web app sometime.

Java language: Java is widespread language that maximum persons have perceived of. In bioinformatics, distinguished instance is genome browser IGV.

SQL: Microsoft SQL would help as data warehouses for instances that we would through BI Tools. Microsoft SQL is comparatively simple toward install plus set up also this is free toward download. Moreover, there are instance database that configure flawlessly with it, for instance, Adventure Works database.

3.Three Applications in Bioinformatics in Human Health and medicine:

A) Genomics- Structural/Functional/Comparative

a. Structural genomics – physical mapping and predictions of 3-d structure of genome aided by sequence analysis, EST sequencing and linkage analysis

b. Functional genomics – gene expression analysis, prediction of gene function and establishment of gene libraries, transcription, translation and protein- protein interactions

c. Comparative genomics – finding sequence similarities and homologies between genomic sequences and their effect on the gene expression leading to evolution, establishing phylogenetic relationship among sequences and structures(molecular taxonomy)

B) Proteomics – Study of Proteoms. Using the genome sequences, identification of proteins, their structures, functions, interactions and modifications leading to change in the original function

C) Metabolomics – Structure and functions of various metabolic components, their original and modified pathways resulting in new molecules

f. Structure-Function Prediction: Elucidating function of a molecule based on its structure

D) Molecular Modeling and Molecular Dynamics – To predict structure from function, Simulating metabolism from the biochemical functions of an organism

h. Gene Therapy – Reversing disease phenotype by identifying the structure of the altered gene responsible for a hereditary disease

Any other application listed to be explained in detail with suitable example.

5. Discuss the Role of Internet in Bioinformatics:

At the turn of the millennium, two young technologies can be singled out which have a major impact on science, industry, and society: recombinant DNA and information technology. As they combine in the field of bioinformatics, they are transforming the pharmaceutical, agrochemical, and food industries and, as a consequence, university education. Much of today's information in the life sciences is generated by collaborative efforts at different locations worldwide, and effective communication is essential for success. Thus, the huge amount of data generated by large-scale genome sequencing activities, e.g., the human genome project, depends heavily on computing and telecommunications and stimulates further efforts in this area. When we talk about sources of biological information and computers for providing it, we can not ignore the role and impact of information superhighway i.e., Internet. The Internet (contraction of interconnected network) is the global system of interconnected computer networks that use the Internet protocol suite (TCP/IP) to link devices worldwide. Internet is the most potential tool of this information age and it is serving as a platform for Bioinformatics tool. It provides the opportunity to search that information, which was available only by reaching to the information centre. Areas of Services The Internet provides various facilities for Bioinformatics, such as; Bioinformatics research, Courses, Resources, Biological databases, Construction tools, Software resources, WWW search tools, Courses of Bioinformatics Advanced topics in Bioinformatics, Scientific databases, Electronic journals, Asking queries from the librarian in online manner, News events and activities such as; announcement for Bioinformatics interest group, meetings on federated databases, molecular biosciences and technology seminars.

In information technology, the World Wide Web (WWW) has become the dominant global communication network. It is based on the Internet, which has served already for more than 20 years as a communication resource among scientists. But only when the hypertext transfer protocol (HTTP) was introduced in 1990 did communication via the Internet became sufficiently easy and inexpensive to allow its general use. Moreover, HTTP is hardware-independent and thus accessible even through inexpensive personal computers which are connected directly to the Internet or via a modem to an Internet provider. This development has stimulated all kinds of commercial activities, and the number of Internet hosts and Internet web sites has reached nearly 40 and 4 million. At present, the number of web sites doubles every year, 100 million people worldwide are estimated to be active Internet users, and business on the order of USD 8 billion is done via the Internet. It is expected that within two more years the number of active users might increase tenfold to reach 1 billion, a dramatic increase driven mainly by the populous Asian nations, and that Internet-based sales will account for USD 300 billion or 1% of all global sales within only four years. Among the initiatives to enhance its quality and speed up transfer of large volumes of data, the Internet2 project is the most ambitious. The Internet2 serves exclusively for scientific purposes and "facilitate and coordinate the development, deployment, operation, and technology transfer of advanced, network-based applications and network services. Internet based Bioinformatics accentuates Higher education and accelerate the availability of new services and applications on the Internet. Even now in the era of Internet commerce, many thousands of WWW sites

are devoted to the global science network. In fact, many recent discoveries and developments, particularly in the life sciences, would be unthinkable without the Internet.

Q.5 Write short notes on any four of the following: (20)

3. Addition Theorem and Multiplication Theorem of Probability

The addition theorem states that 2 events A and B are mutually exclusive, the probability of occurrence of either A or B is the sum of their individual probability of A and B.

P(A or B)= P(A) + P(B) for mutually exclusive events

The multiplication theorem states that if 2 events A and B are independent the probability that they both will occur is equal to the product of their individual probability A and B

P(A and B) = P(A) X P(B)

4. Standard Normal curve

The standard normal curve is a bell shape and the top of the bell is directly above the mean

The curve is symmetrical about the line x=μ and x ranges from -∞ to +∞

Mean, median and mode coincide at x=μ as the distribution is symmetrical

X axis is asymptote to the curve

The points of inflexion of the curve are x= μ+sigma, x= μ-sigma

The total area under the normal curve is equal to unity and the %age distribution of area under the curve is 68% falls between one sigma +/- μ, 95.5% area falls between 2sigma +/-μ and about 99.7% area falls under 3 sigma +/-μ

This curve is unimodal

Mathematical equation can be determined if μ and sigma are known

5. SGD and its features:

The ***Saccharomyces*** Genome Database (SGD) provides comprehensive integrated biological information for the budding yeast ***Saccharomyces cerevisiae*** along with search and analysis tools to explore these data, enabling the discovery of functional relationships between sequence and gene products in fungi and higher organisms.

-Tools available for Genome study
-Genome structure, Gene list, Restriction Maps, Blasts, Protein sequences, Phylogenetic analysis
-Research in SGD
-Any other relevant information found in the database

6. Difference between a GenBank entry and FASTA format:

‘GenBank’ is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the ftp site. The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for previous GenBank releases are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. An annotated sample GenBank record for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

FASTA Format:

In bioinformatics and biochemistry, the **FASTA format** is a text-based format for representing either nucleotide sequences or amino acid (protein) sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences. The format originates from the FASTA software package, but has now become a near universal standard in the field of bioinformatics. The simplicity of FASTA format makes it easy to manipulate and parse sequences using text-processing tools and scripting languages like the R programming language, Python, Ruby, and Perl.

Eg. >HSBGPG Human gene for bone gla protein (BGP)

GGCAGATTCCCCCTAGACCCGCCCGCACCATGGTCAGGCATGCCCCTCCTCATCG
CTGGGCACAGCCCAGAGGGTATAAACAGTGCTGGAGGCTGGCGGGGCAGGCCA
A sequence record in a FASTA format consists of a single-line description (sequence name), followed by line(s) of sequence data. The first character of the description line is a greater-than (">") symbol.