JEPPIAAR ENGINEERING COLLEGE

JEPPIAAR NAGAR, CHENNAI – 119

DEPARTMENT OF BIOTECHNOLOGY

QUESTION BANK

ON

BT 6701– BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

REGULATION - 2013

IV YEAR & VII SEMESTER

BATCH: (2016-2020)

VISION OF THE INSTITUTION

- To build Jeppiaar Engineering College as an institution of academic excellence in technological and management education to become a world class University

MISSION OF THE INSTITUTION

- To excel in teaching and **learning, research and innovation** by promoting the principles of scientific analysis and creative thinking.
- To participate in the production, **development and dissemination of knowledge** and interact with **national and international communities.**
- To equip students with **values, ethics and life skills n**eeded to enrich their lives and enable them to meaningfully contribute to the **progress of society.**
- To prepare students for **higher studies and lifelong learning,** enrich them with the **practical and entrepreneurial skills** necessary to excel as future professionals and contribute to **Nation's economy**

PROGRAM OUTCOMES (PO)	
PO 1	**Engineering knowledge**: **Apply** the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
PO 2	**Problem analysis**: **Identify, formulate, review** research literature, and **analyze** complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
PO 3	**Design/development of solutions**: **Design solutions** for complex engineering problems and **design** system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations
PO 4	**Conduct investigations of complex problems**: **Use** research-based knowledge and research methods including **design** of experiments, **analysis** and **interpretation** of data, and **synthesis** of the information to provide valid conclusions.
PO 5	**Modern tool usage**: **Create**, **select**, and **apply** appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
PO 6	**The engineer and society**: **Apply** reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
PO 7	**Environment and sustainability**: **Understand** the impact of the professional engineering solutions in societal and environmental contexts, and **demonstrate** the knowledge of, and need for sustainable development.
PO 8	**Ethics**: **Apply** ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
PO 9	**Individual and team work**: **Function effectively** as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
PO 10	**Communication**: **Communicate effectively** on complex engineering activities with the engineering community and with society at large, such as, being able **to comprehend** and **write** effective reports and **design** documentation, **make** effective presentations, and **give** and **receive** clear instructions.
PO 11	**Project management and finance**: **Demonstrate** knowledge and understanding of the engineering and management principles and **apply** these to one's own work, as a member and leader in a team, to **manage** projects and in multidisciplinary environments.
PO 12	**Life-long learning**: **Recognize** the need for, and have the preparation and ability to **engage** in independent and life-long learning in the broadest context of technological change.

VISION OF THE DEPARTMENT	
To pursue excellence in producing bioengineers coupled with research attributes.	
MISSION OF THE DEPARTMENT	
M1	To **impart** quality education and **transform** technical knowledge into career opportunities.
M2	To **establish** a bridge between the **program and society** by **fostering** technical education.
M3	To **generate societal conscious technocrats** towards **community development**
M4	To **facilitate higher studies** and research in order to have an **effective career /** entrepreneurship.
PROGRAM EDUCATIONAL OBJECTIVES (PEOS)	
PEO - 1	To impart knowledge and produce competent graduates in the field of biotechnology
PEO - 2	To inculcate professional attributes and ability to integrate engineering issues to broader social contexts.
PEO - 3	To connect the program and community by fostering technical education.
PEO - 4	To provide a wide technical exposure to work in an interdisciplinary environment
PEO - 5	To prepare the students to have a professional career and motivation towards higher education.
PROGRAM SPECIFIC OUTCOMES (PSOS)	
PSO 1	**Professional Skills:** This programme will provide students with a solid foundation in the field of Biological Sciences and Chemical engineering enabling them to work on engineering platforms and applications in Biotechnology as per the requirement of Industries, and facilitating the students to pursue higher studies
PSO 2	**Problem-solving skills:** This programme will assist the students to acquire fundamental and problem solving knowledge on subjects relevant to Biotechnology thereby encouraging them to understand emerging and advanced concepts in modern biology
PSO 3	**Successful Career and Entrepreneurship:** Graduates of the program will have a strong successful career and entrepreneurial ability with the blend of inputs from basic science, engineering and technology, thereby enabling them to translate the technology and tools in various industries and/or institutes

BT6701 BIOINFORMATICS AND COMPUTATIONAL BIOLOGY LT P C 3 0 0 3

OBJECTIVES:

- To improve the programming skills of the student
- To let the students know the recent evolution in biological science.

UNIT I **9**

Introduction to Operating systems, Linux commands, File transfer protocols ftp and telnet, Introduction to Bioinformatics and Computational Biology, Biological sequences, Biological databases, Genome specific databases, Data file formats, Data life cycle, Database management system models, Basics of Structured Query Language (SQL).

UNIT II **9**

Sequence Analysis, Pairwise alignment, Dynamic programming algorithms for computing edit distance, string similarity, shotgun DNA sequencing, end space free alignment. Multiple sequence alignment, Algorithms for Multiple sequence alignment, Generating motifs and profiles, Local and Global alignment, Needleman and Wunsch algorithm, Smith Waterman algorithm, BLAST, PSIBLAST and PHIBLAST algorithms.

UNIT III **8**

Introduction to phylogenetics, Distance based trees UPGMA trees, Molecular clock theory, Ultrametric trees, Parsimonious trees, Neighbour joining trees, trees based on morphological traits, Bootstrapping. Protein Secondary structure and tertiary structure prediction methods, Homology modeling, *abinitio* approaches, Threading, Critical Assessment of Structure Prediction, Structural genomics.

UNIT IV **11**

Machine learning techniques: Artificial Neural Networks in protein secondary structure prediction, Hidden Markov Models for gene finding, Decision trees, Support Vector Machines. Introduction to Systems Biology and Synthetic Biology, Microarray analysis, DNA computing, Bioinformatics approaches for drug discovery, Applications of informatics techniques in genomics and proteomics: Assembling the genome, STS content mapping for clone contigs, Functional annotation, Peptide mass fingerprinting.

UNIT V **8**

Basics of PERL programming for Bioinformatics: Datatypes: scalars and collections, operators, Program control flow constructs, Library Functions: String specific functions, User defined functions, File handling.

TOTAL : 45 PERIODS

OUTCOMES:

Upon completion of this course, students will be able to

- Develop bioinformatics tools with programming skills.
- Apply computational based solutions for biological perspectives.
- Pursue higher education in this field.
- Practice life-long learning of applied biological science.

TEXT BOOKS:

- Lesk, A. K., "Introduction to Bioinformatics" 4th Edition, Oxford University Press, 2013
- Dan Gusfield, "Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology" Cambridge University Press, 1997.
- Durbin, R., Eddy, S., Krogh, A., and Mitchison, G., "Biological Sequence Analysis Probabilistic Models of proteins and nucleic acids" Cambridge, UK: Cambridge University Press, 1998.
- Mount, D.W., "Bioinformatics Sequence and Genome Analysis" 2nd Edition, Cold Spring Harbor Laboratory Press, 2004
- Tindall, J., "Beginning Perl for Bioinformatics: An introduction to Perl for Biologists" 1st Edition, O'Reilly Media, 2001

REFERENCE:

- Baldi, P. and Brunak, S., "Bioinformatics: The Machine Learning Approach" 2nd Edition, MIT Press, 2001.

CO NO	COURSE OUTCOME
C401.1	The students will have the ability to **Develop** bioinformatics **tools** with programming skills.
C401.2	The students will have the ability to **Apply** computational based **solutions** for biological perspectives.
C401.3	The students will have **the ability to understand**, **explain** and **perform** phylogenetic analysis and be **able to predict** the structure of proteins
C401.4	The students will have the **ability to learn** the AI and neural networking
C401.5	The students will have the **ability to understand, execute** the programs to solve biological issues by using PERL

SUBJECT : BIOINFORMATICS & COMPUTATIONAL BIOLOGY
SEMESTER : VII
YEAR : IV
REGULATION : R 2013
COURSE CODE : BT 6701

S.NO	TOPICS	CHAPTER	PAGE NO.
	UNIT I		
I	Introduction to Operating systems		
	Linux commands		
	File transfer protocols ftp and telnet		
	Introduction to Bioinformatics and Computational Biology		
	Biological sequences	3	108-153 (T1)
	Biological databases	4	108-154 (T1)
	Genome specific databases	5	108-155 (T1)
	Data file formats	3	106(T1)
	Data life cycle	3	106(T1)
	Database management system models		
	Basics of Structured Query Language (SQL)		
2	**UNIT II**		
	Sequence Analysis	4	154-188 (T1)
	Pairwise alignment	4	154-188 (T1)
	Dynamic programming algorithms for computing edit distance	4	154-188 (T1)
	string similarity	3	53-112 (T4)
	shotgun DNA sequencing	3	53-112 (T4)
	end space free alignment	3	53-112 (T4)
	Multiple sequence alignment	3	53-112 (T4)
	Algorithms for Multiple sequence alignment	4	154-188 (T1)
	Generating motifs and profiles	3	53-112 (T4)
	Local and Global alignment	4	154-188 (T1)
	Needleman and Wunsch algorithm, Smith Waterman algorithm	4	154-188 (T1)
	BLAST, PSIBLAST and PHIBLAST algorithms	4	154-188 (T1)
	UNIT III		
3	Introduction to Phylogenetics	4	189-206 (T1)
	Distance based methods UPGMA	4	189-206 (T1)
	Molecular clock theory	4	189-206 (T1)
	Ultrametric trees	4	189-206 (T1)
	Parsimonous trees	4	189-206 (T1)
	Neighbouring joining trees	4	189-206 (T1)
	Trees based on morphological traits	4	189-206 (T1)
	Bootstrapping.	4	189-206 (T1)
	Protein Secondary structure and tertiary structure prediction methods	5	225-227 (T1)
	Homology modeling, abinitio approaches, Threading	5	233-239 (T1)
	Critical Assessment of Structure Prediction, Structural genomics.	1	51(T1)

SUBJECT : BIOINFORMATICS & COMPUTATIONAL BIOLOGY
SEMESTER : VII
YEAR : IV
REGULATION : R 2013
COURSE CODE : BT 6701

S.NO	TOPICS	CHAPTER	PAGE NO.
4	**UNIT IV**		
	Machine learning techniques: Artificial Neural Networks in protein secondary structure prediction,	5	228-231(T1)
	Hidden Markov Models for gene finding	4	188-192(T1)
	Decision trees, Support Vector Machines		
	Introduction to Systems Biology and Synthetic Biology	2	68
	Microarray analysis	10	519-526(T2)
	DNA computing, Bioinformatics approaches for drug discovery	5	242
	Applications of informatics techniques in genomics and proteomics	3	68, 207-255
	Assembling the genome, STS content mapping for clone contigs,		
	Functional annotation	2	75
	Peptide mass fingerprinting		
5	**UNIT IV**		
	Basics of PERL programming for Bioinformatics:	1	18-23(T1)
	Datatypes: scalars and collections	1	18-23(T1)
	Operators,	1	18-23(T1)
	Program control flow constructs	1	18-23(T1)
	Library Functions	1	18-23(T1)
	String specific functions	3	58-60(T1)
	User defined functions	2	18-24(T1)
	File handling	1	18-24(T1)

UNIT I **9**

Introduction to Operating systems, Linux commands, File transfer protocols ftp and telnet, Introduction to Bioinformatics and Computational Biology, Biological sequences, Biological databases, Genome specific databases, Data file formats, Data life cycle, Database management system models, Basics of Structured Query Language (SQL).

1. Give two examples of popular dialects of SQL? (November/December 2016)
   The SQL dialect, derived from the Structured Query Language, uses human-readable expressions to define query statements.
   Example: SelectConnectByConditionStep connectBy(Condition condition);

2. Mention the types of data organized by KEGG. (November/December 2016)
   It is an ontology database containing hierarchical classifications of various entities including genes, proteins, organisms, diseases, drugs, and chemical compounds.

3. Define Operating system.
   An operating system is system software which may be viewed as an organized collection of software consisting of procedures for operating a computer and providing an environment for execution of programs. It acts as an interface between users and the hardware of a computer system.

4. What are the functions and components of an Operating system?
   An operating system is an essential component of a computer system. The primary objectives of an operating system are to make computer system convenient to use and utilize computer hardware in an efficient manner. An operating system is a large collection of software which manages resources of the computer system, such as;
   - Memory
   - Processor
   - File system
   - Input/output devices.

5. What are types of an Operating system?
   - Batch operating system
   - Multiprogramming operating system
   - Network operating system
   - Distributed operating system

6. What are Unix operating system and its features?
   Unix is a multi-programming operating system. Some high-level features of the UNIX system are
   - The file system,
   - The processing environment, and
   - The building block primitives

7. Define Unix kernel.
   The kernel is the essential center of a computer operating system, the core that provides basic services for all other parts of the operating system.

8. What is the role of a kernel?
   Kernel or operating systems provides the following services;
   - Controlling the execution of processes
   - Scheduling processes fairly for execution
   - Allocating main memory for an executing process
   - Allocating secondary memory for efficient storage and retrieval of user data
   - Allowing processes controlled access to peripheral devices such as terminals, tape drivers, disk drivers and network devices.
9. What is network and network hardware?
   A network is a set of nodes and links. Networking hardware includes all computers, peripherals, interface cards and other equipment needed to perform data-processing and communications within the network. The figure below depicts the components (hardware) required for a networking.

10. Define local area network (LAN)
    A local area network (LAN) is usually privately owned and links the devices in a single office, building, or campus. Depending on the needs of an organization and the type of the technology used, a LAN can be as simple as two PCs and printer in some home's office or it can be extended through out the company
    - LAN size is limited to a kilometre
    - LANs are designed to allow resources to be shared between personal computers or work station
    - LAN uses only one type of transmission medium
    - The most common LAN topologies are bus, ring and star

11. Define network topology and its types.
    Network topology is the study of the arrangement or mapping of the elements (links, nodes, etc.) of a network, especially the physical (real) and logical (virtual) interconnections between nodes. The most common of these basic types of topologies are:
    - Bus (Linear, Linear Bus)
    - Star
    - Ring
    - Mesh
    - Tree
    - Hybrid

12. What is Protocol and its types.

    A protocol is a set of rules that governs the communications between computers on a network. These rules include guidelines that regulate the following characteristics of a network: access method, allowed physical topologies, types of cabling, and speed of data transfer. The most common protocols are:

    - Ethernet
    - LocalTalk
    - Token Ring
    - FDDI
    - ATM

13. Define Transmission Control Protocol/Internet Protocol (TCP/IP)

    The transmission control protocol/Internet protocol is a set of protocols, or a protocol suite, that defines how all transmissions are exchanged across the internet.

14. Define File Transfer Protocol (FTP)

    File Transfer Protocol (FTP) is a standard mechanism provided by TCP/IP for copying a file from one host to another. Transferring files from one computer to another is one of the most common tasks expected from a networking or internetworking environment.

15. What are web browsers? Give a few examples and their suitability

    A browser is the software that is used to view web pages. There are two types of browsers

    - Text based browsers
    - Graphical browsers

16. What is HTML tag? How are the represented? Give two examples

    Hypertext Markup Language is a language for creating a web page.

17. What is DBMS? Mention the four main types of data organization.

    A database management system is software that defines a database, stores the data, supports a query language, produces reports, produces reports and creates data entry screens.

18. What are different types of Biological database?

    Primary database, secondary database and composite database.

19. Write any two methods available for alignment of pair of sequence.

    - Local alignment
    - Global alignment

20. What are Primary biological databases? Give example.

Primary biological database contains collection of crude rudimentary sequence submissions i.e., raw data. Some of the primary databases are GenBank, DDBJ and EMBL etc.

21. What are Secondary biological databases? Give example.

In addition to the numerous primary and composite resources, there are many secondary (or pattern) databases, so-called because they contain the fruits of analyses of the sequences in the primary sources. Some of the main secondary resources are; Prosite, Profiles, PRINTS, BLOCKS etc.

22. What are Structural biological databases? Give example.

Proteins share structural similarities, reflecting common evolutionary origins. The evolutionary process involves substitutions, insertions and deletions in amino acid sequences. For distantly related proteins, such changes can be extensive, yielding folds in which the numbers and orientations of secondary structures vary considerably. Example: SCOP, CATH etc.

23. What are the tools available for gene finding?

S.No	Software	Description
1.	GeneMark	Family of gene prediction programs
2.	Geneparser	Parse a DNA sequence into introns and exons
3.	GLIMMER	finding genes in microbial DNA
4.	ORF FINDER	a graphical analysis tool which finds all open reading frames

24. Give any two applications of decision tree in computational biology.

Decision trees have been applied to problems such as assigning protein function and predicting splice sites.

25. Give one major advantage of DNA computing.

The DNA computer has clear advantages over conventional computers when applied to problems that can be divided into separate, non-sequential tasks. The reason is that DNA strands can hold so much data in memory and conduct multiple operations at once, thus solving decomposable problems much faster. On the other hand, non-decomposable problems, those that require many sequential operations are much more efficient on a conventional computer due to the length of time required to conduct the biochemical operations.

26. Write a note on dot matrix method?

A dot matrix analysis is a method for comparing two sequences to look for possible alignment. One sequence (A) is listed across the top of the matrix and the other (B) is listed down the left side. Starting from the first character in B, one moves across the page keeping in the first row and placing a dot in many column where the character in A is the same. The process is continued until all possible comparisons between A and B are made. Any region of similarity is revealed by a diagonal row of dots. Isolated dots not on diagonal represent random matches.

27. What are Genome specific databases?

These databases collect genome sequences, annotate and analyze them, and provide public access. Some add curation of experimental literature to improve computed annotations. These databases may hold many species genomes, or a single model organism genome. Example, OMIM, Mouse genome etc.

28. Define file format.

File format is a format for encoding information in a file. Each different type of file has a different file format. The file format specifies first whether the file is a binary or ASCII file, and second, how the information is organized.

29. Write a note on Data life cycle?

Data lifecycle management is the process of managing business information throughout its lifecycle, from requirements through retirement.

30. What are different types of DBMS models?

- Hierarchical database model.
- Network model.
- Relational model.
- Entity–relationship model. Enhanced entity–relationship model.
- Object model.
- Document model

31. What is SQL?

SQL is a database computer language designed for the retrieval and management of data in relational database. SQL stands for Structured Query Language.

32. What are the sequence submission tools?

- BankIt, Sequin for GenBank
- Sakura for DDBJ
- Webin for EMBL

Part B

1. Describe the various database management models (November/December 2016).
2. Describe the various databases that deal with DNA and protein structure (November/December 2016).
3. Database heterogeneity is very common in bio-databases. How would you classify bio-databases based on the sources of data? Cp. 3, Pg.3-12, Nov-2013.
4. Explain the classification of biological databases. Give some information about applications of databases in molecular biology, Cp.3, Pg.3-12, Jan- 2014.
5. Explain in detail Data life cycle and database management system, Cp. 3, Pg.3-12,
6. What is SRS? Define composite database with an example, Cp.3, Pg.3-22, Nov-2013.

Part C

1. Define Operating system? Explain the architecture and organization of an operating system.
2. Explain genome specific databases in detail.
3. Explain Database management with reference to biological and clinical data.

UNIT II **9**

Sequence Analysis, Pairwise alignment, Dynamic programming algorithms for computing edit distance, string similarity, shotgun DNA sequencing, end space free alignment. Multiple sequence alignment, Algorithms for Multiple sequence alignment, Generating motifs and profiles, Local and Global alignment, Needleman and Wunsch algorithm, Smith Waterman algorithm, BLAST, PSIBLAST and PHIBLAST algorithms.

1. Mention the two important differences between global and local alignment (November/December 2016).

Global alignment	Local alignment
Aligns the entire sequence	Finds the local regions with highest level of similarity between the two sequences
Compares and contains all letters from the target and the query sequences	Aligns a substring of the query sequence to the substring of the target sequence
If two sequences are of same lengt h and similar in length, they are suitable for global alignment	Local alignment finds stretches of sequences with high level of matches without considering the alignment of rest of the sequence region.
Suitable for closely related sequences	Suitable for aligning more divergent sequences or distantly related sequences
Needleman-wunsch algorithm	Smith-watermann algorithm
EMBOSS Needle	BLAST

2. Write a short note on ExPaSy. (November/December 2016)
   ExPASy is the SIB Bioinformatics Resource Portal which provides access to scientific databases and software tools (i.e., resources) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc.

3. What is Pattern matching? Give some its application?
   Automated pattern matching is defined as the ability of a program to compare novel and known patterns and determine the degree of similarity which forms the basis for automated sequence analysis, modelling of protein structures, locating of homologous genes, data mining, search engines and dozens of other activities in bioinformatics. Some of the key bioinformatics applications of pattern recognition and matching (pattern matching) are

4. Define Sequence alignment.
   Sequence alignment is fundamental to inferring homology and function. For example, if two sequences are in alignment-part or the entire pattern of nucleotides match-then they are similar and may be homologous.

5. What are types of sequence alignment?
   There are three categories of sequence alignment
   - Pairwise sequence alignment
   - Global versus local alignment
   - Multiple sequence alignment

6. What are the methods of sequence alignment?
   There are various methods of sequence alignments. These methods differ in approach, computational complexity and accuracy of results.
   The various methods are;
   - Brute force alignment
   - Dot matrix alignment
   - Dynamic programming
   - Heuristics methods

7. What are Sequence comparison algorithms? Give example.
   Sequence comparison algorithms deal with two sequences and the similarities between them. Sequences are compared to assign function to a new sequence, predict and construct model protein structures, and design and analyse gene expression experiments. Example: Dotplot.

8. What are scoring matrices?
   A scoring matrix gives the score for aligning two amino acids (match or mismatch) in a pairwise alignment. A scoring matrix can be considered a measure of the evolutionary change. The most widely used matrices are PAMs and BLOSUMs. Both calculates substitution frequencies between amino acids, and both are derived from known protein alignments

9. Define Edit distance.
   The process of alignment can be measured in terms of the number of gaps introduced and the number of mismatches remaining in the alignment. A metric relating such parameters represents the distance between two sequences is referred to as edit distance. In other words, edit distance is referred to as the number of operations required to transform one of them into the other.

10. Define Levenstein distance.
    Levenshtein distance is a string metric which is one way to measure edit distance. The levenshtein distance between the two strings needed to transfer / transform one string into another, where an operation is an insertion, deletion or substitution of a single character.

11. What is FASTA format? Give an example of nucleotide sequence in FASTA format.
    FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which base pairs or amino acids are represented using single-letter codes. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data.
    >gi|317410865|gb|HQ108711.1|
    GTGAAGCCTTTCCAGACAGACGCTCTTGTTATAACACCAGGCCAGACCACCAATGTATTA
    TTCACGGCCA

12. Distinguish between bits score and e-value in BLAST.

Bits score	E-value
Raw scores have little meaning without detailed knowledge of the scoring system used, or more simply its statistical parameters K and lambda. $S' = \frac{\lambda S - \ln K}{\ln 2}$ Unless the scoring system is understood, citing a raw score alone is like citing a distance without specifying feet, meters, or light years.	The E-value corresponding to a given bit score is simply $E = mn\,2^{-S'}$
By normalizing a raw score using the formula one attains a "bit score" S', which has a standard set of units	Bit scores subsume the statistical essence of the scoring system employed, so that to calculate significance one needs to know in addition only the size of the search space

13. Write any two methods available for alignment of pair of sequence?
    Smith-watermann algorithm, Needlemann-wunsch algorithm, Dotplot etc.Define Hamming distance. A measure of the difference between two messages each consisting of a finite string of characters, expressed by the number of characters, expressed to obtain one from the other.

14. Define Dynamic programming and its types.
    Dynamic programming (DP) is an efficient recursive method to search through all possible alignments and finding the one with the optimal score. Dynamic programming is good example for pairwise sequence alignment. There are two types of dynamic programming such as
    - Global sequence alignment (Needleman-wunsch algorithm)
    - Local sequence alignment (Smith waterman algorithm)

15. What are the steps involved in dynamic programming?
    Dynamic programming usually consists of three components.
    - Recursive relation
    - Tabular computation
    - Traceback

16. What are the applications of Smith-watermann algorithm?
    Local alignment has many applications in the field of
    - Sequence comparison of different lengths

- Comparison of long sequences containing both coding and non-coding regions
- Proteins from different protein families are compared to find conserved domains.
- Sequence comparison using global alignment does not give the expected score.

17. What are Heuristic algorithms?

    Heuristics algorithms are faster algorithm that are based on assumptions and approximations. These algorithm do not make all possible pairwise comparison of all of the database sequences and thus they are not expensive. The process of knowing i.e., learning to solve a solution by trying rather than by following some pre-established formula is the approach of such algorithm. Thus based on trial and error method i.e., successive approximations, heuristics algorithms solve similarity search and alignment problems. These are the methods devised to search a small fraction of a dynamic programming matrix by looking at all the high scoring alignments. But heuristic algorithms compromise on sensitivity and selectivity.

18. What is FASTA? What are types of FASTA?

    FASTA is a heuristic sequence searching and local alignment tool found by Pearson and Lipmann in 1988. It has restrictions on word and window size. Various types of FASTA algorithm are

    - FASTA
    - TFASTA
    - LFASTA
    - FASTX/FASTY
    - FASTF/TFASTF
    - FASTS/TFASTS
    - TFASTX/FASTY

19. What is BLAST? What are types of BLAST?

    BLAST is a sequence alignment program similar to FASTA. It has speed faster than FASTA and very good sensitivity. It is the most popular sequence alignment algorithm. It finds the ungapped local alignments between a query sequence and a target database by either looking for any short stretch of identities or a very high scoring match. Both the query and the target database can be either nucleotide sequence or amino acid sequence, but it is very effective for amino acid sequences than DNA sequences. Various types of BLAST are

    - BLASTN - nucleotide query of the nucleotide database.
    - BLASTP - protein query of the protein database.
    - BLASTX - translate DNA to protein and query protein database.
    - TBLASTN - protein query of the translated nucleotide database.

- TBLASTX - translate DNA to protein and query the translated nucleotide database.

20. Define alignment score.

    Alignment score is defined as the sum of no. of alignments, no. of gaps and no. of inexact matches when comparing two biological sequences.

    Alignment score = No. of alignments + No. of gaps + No. of inexact matches

21. What is Gap penalty?

    Gap penalty is the total gap penalty including the opening gap penalty and extension gap penalty and the length of the sequence.

    Gap penalty or gap score (w) = g+rL

    Where

    g = opening gap penalty value

    r = extension gap penalty value

    L= length of the gap in characters

22. Define Multiple Sequence Alignment (MSA).

    Multiple sequence alignment of three or more biological sequence, generally protein or DNA. MSA is a tool to determine levels of homology and phylogeny, hence from the members of globally related sequence. Visual depictions illustrate point mutation, insertions or deletions that appear as gaps in one or more sequences in the sequence alignment. MSA is used to assess sequence conservation of protein domains, tertiary and secondary structures.

23. What is Sum of Pair method (SOP)?

    The SP score of a multiple sequence alignment is the sum of the scores of the pair-wise alignments implied by the multiple alignments. DP is similar to that of DP for two sequences. However, instead of aligning two sequences, one need to align three or more simultaneously. This method requires three dimensions making this method as a cumbersome and practically difficult methodology to work.

24. What is PHIBLAST?

    PHI-BLAST(Pattern Hit Initiated BLAST) searches protein sequences using a combination of pattern matching and local alignment to reduce the probability of false positives.

25. What is PSIBLAST?

    It is a position specific iterated BLAST, it incorporates elements of both pairwise and multiple sequences method. It begins with one at a time search and then derives information from a multiple sequence alignment.

26. Write note on shotgun sequencing?
    Shotgun sequencing, also known as shotgun cloning, is a method used for sequencing long DNA strands.

27. What are Next Generation Sequencing platforms?
    Illumina sequencing, pyrosequencing, roche 454, Ion torrent are few examples of Next Generation Sequencing techniques.

28. What are the implementation FASTA program available in the database?
    The FASTA package provides SSEARCH, an implementation of the optimal Smith-Waterman algorithm.

29. Define Gene Annotation.
    DNA annotation or genome annotation is the process of identifying the locations of genes and all of the coding regions in a genome and determining its functions.

30. What is Molecular Dynamics?
    Molecular dynamics (MD) is a computer simulation method for studying the physical movements of atoms and molecules, and is thus a type of N-body simulation. The atoms and molecules are allowed to interact for a fixed period of time, giving a view of the dynamical evolution of the system.

31. What are the algorithms used for Multiple Sequence Algorithm (MSA)?
    - Dynamic programming
    - Sum of pair method
    - Progressive alignment
    - Iterative methods
    - Hidden Markov Models
    - Genetic Algorithms
    - Automated tools (Macaw, Meme etc)

Part B

1. Describe the methods of sequence analysis (November/December 2016).
2. Write a detailed account of dynamic programming algorithms used for shotgun DNA sequencing (November/December 2016).
3. What are the BLAST programmes and modes available?, Cp.4, Pg.154-88(T1), Jan-2012;13;14, 15
4. Enumerate the various types of multiple sequence alignment. Give the programs under each category, and mention the drawbacks of progressive alignment, Cp.4, Pg. 154, Jan 2014, 2015.
5. Explain Needleman and Wunsch algorithm, Smith Waterman algorithm in detail, Cp.4, Pg.154, Nov-2016.

Part C

1. Explain PSIBLAST and PHIBLAST algorithms, Cp.4, Pg.154-88.
2. Explain shotgun sequencing, Cp.3, Pg. 53.
3. Describe Dynamic programming with any one algorithm, Cp.4, Pg. 154, Jan-12;13

UNIT III **8**

Introduction to phylogenetics, Distance based trees UPGMA trees, Molecular clock theory, Ultrametric trees, Parsimonious trees, Neighbour joining trees, trees based on morphological traits, Bootstrapping. Protein Secondary structure and tertiary structure prediction methods, Homology modeling, *abinitio* approaches, Threading, Critical Assessment of Structure Prediction, Structural genomics.

1. What is DALI? (November/December 2016)
   DALI stands for Distance Alignment Matrix Method. DALI is a common and popular method that breaks down the protein that is inputted into hexapeptide fragments and then calculates a distance matrix through the understanding of the contact pattern between successive fragments.

2. Write short note on ultrametric trees? (November/December 2016)
   Ultrametric trees, or approximations of them, can be used to deduce both the branching patterns of evolutionary history and some measure of the time that has passed along each branch.

3. Define mutation
   Any change in the DNA sequence of an organism is a mutation. Mutations are the source of the altered versions of genes that provide the raw material for evolution. Most mutations have no effect on the organism, especially among the eukaryotes, because a large portion of the DNA is not in genes and thus does not affect the organism's phenotype. Of the mutations that do affect the phenotype, the most common effect of mutations is lethality, because most genes are necessary for life. There are various types of mutations such as
   - Point mutation
   - Insertion mutation
   - Deletion mutation
   - Frameshift mutation

4. What is mutation rate?
   Mutation rate is defined as the probability of a particular type of mutation per unit time (or generation).

5. Define Mutation frequency
   Mutation frequency = number of times a particular mutation occurs in a population of cells or individuals.

6. What are the methods to detect mutations?
   - Ame's test
   - Southern blot
   - Sequencing denaturing HPLC

- Dideoxy fingerprinting
- Oligonucleotide arrays
- Protein truncation test
- Single stranded conformation polymorphism
- Denaturing gradient gel electrophoresis

7. What are cladogram?
   It is a branching diagram representing the most parsimonious distribution of derived characters within a set of taxa. The branching pattern of a cladogram is intended to show the relative relationship among taxa. It is not a true evolutionary tree of how those relationships came to be. Branch lengths are equal in a cladogram.

8. What are Phylogram?
   It is a phylogenetic tree that indicates the relationships among the taxa. It also shows evolution and the rate of evolution. Branch lengths are proportional to distance.

9. What are the methods of analyzing Phylogenetic trees?
   There are various methods of building and analyzing phylogenetic trees. Broadly these methods are of two types
   - Optimization methods
   - Algorithmic methods

10. What is distance based methods?
    Distance methods are also called phonetic methods. The trees are constructed by similarities of sequences and the resulting tree is called a dendogram. A dendogram does not necessarily reflect evolutionary relationships. Distance methods compress all of the individual differences between pairs of sequences into a single number.

11. What is character based methods?
    Character based approaches are also called cladistic methods. The trees are calculated by considering the various possible pathways of evolution and are based on parsimony or likelihood methods. The resulting tree is called cladogram. Cladistic methods use each alignment position as evolutionary information to build tree.

12. What are evolutionary distance methods?
    In evolutionary distance method or distance matrix method, all possible pairs of sequences are aligned to determine when pairs are the most similar or closely related. These alignments provide a measure of the genetic distance between the sequences. These distance measurements are then used to predict the evolutionary relationships.
    There are various distance matrix algorithms such as

- UPGMA
- WPGMA
- Neighbouring-joining
- Fitch-Morgalish

13. What is UPGMA?

    The clustering procedure called UPGMA stands for **U**nweighted **P**air **G**roup **M**ethod using **A**rithmetic averages. The method is simple and intuitively appealing. It works by clustering the sequences, at each stage amalgamating two clusters and at the same time creating a new node on a tree. The tree can be imagined as being assembled upwards, each node being added above the others, and the edge lengths being determined by the difference in heights of the nodes at the top and bottom of the edge.

14. What are Ultrametric trees? (November/December 2016)

    It is a rooted tree where each internal node is labelled with a number. Each internal node has atleast two offsprings and the labels decrease along the path from any root to leaf. The branch lengths of an ultrametric tree are proportional to the divergent time.

15. What are Parsimonous trees?

    It is an algorithm that works by finding the tree which can explain the observed sequences with minimal number of substitutions. There is various parsimony methods are

    - Maximum parsimony
    - Camin-Sokal parsimony
    - Dollo parsimony
    - Wagner parsimony
    - Polymorphism parsimony

16. What are Neighbour joining trees?

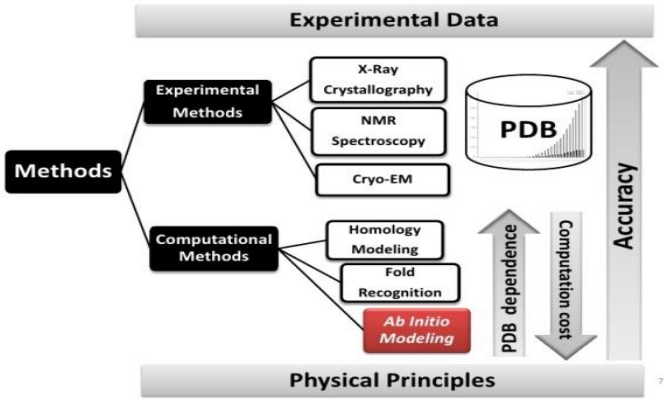
    NJ is a clustering method related to UPGMA that is able to solve problems similar to distance matrix algorithms such as UPGMA. These algorithms are computationally fast and don't make the assumption of additivity. It begins by choosing the two most closely related sequences and then adding the next most distant sequence as a third branch by the tree.

17. Write a note on Bootstrapping?

    Bootstrap resampling is sampling with replacement. In the case of MSA, sites are sampled at random until the data set is equal in length to the original alignment. In bootstrapped replicates, most sites are sampled once, some sites are sampled twice

and a small number of sites are sampled three times. Some For bootstrap resampling of a sequence alignment, it is best to create at least 100 bootstrapped data sets, and redo the phylogeny for each one. A consensus tree can then be built which indicates, for each branch in the tree, how often it occurs in the population of replicate samples. Certain positions are biased in each replicates, while some are underrepresented with enough replicates being weighted equally.

18. What are Protein Secondary structure and tertiary structure prediction methods?



19. Define parsimony.

Maximum parsimony is an optimality criterion under which the phylogenetic tree that minimizes the total number of character-state changes is to be preferred. Under the maximum-parsimony criterion, the optimal tree will minimize the amount of evolution.

20. What is an outgroup? How to select one?

An out-group is a group of organisms that serve as a reference group when determining the evolutionary relationship among three or more monophyletic groups of organisms. The out-group is used as a point of comparison for the ingroup—the set of organisms under study that specifically allows the phylogeny to be rooted.

21. What is Fitch-Morgalisch algorithm?

The Fitch-Margoliash algorithm, commonly called the FM-algorithm, is used to cluster taxa using evolutionarily related distances calculated using the Jukes-Cantor.

22. What is Homology modelling?

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. Ex. Swiss modeller and Modeller.

23. Brief out *abinitio* approaches of protein structure modelling.
    The ab initio prediction methods consist in modelling all the energetics involved in the process of folding, and then in finding the structure with lowest free energy. This approach is based on the 'thermodynamic hypothesis', which states that the native structure of a protein is the one for which the free energy achieves the global minimum.

24. Define Threading
    A protein fold recognition technique that involves incrementally replacing the sequence of a known protein structure with a query sequence of unknown structure. The new "model" structure is evaluated using a simple heuristic measure of protein fold quality. The process is repeated against all known 3D structures until an optimal fit is found.

25. How will you evaluate phylogenetic tools?
    A number of methods exist to test the amount of support for a phylogenetic tree, either by evaluating the support for each sub-tree in the phylogeny (nodal support) or evaluating whether the phylogeny is significantly different from other possible trees (alternative tree hypothesis tests). Methods include Bootstrapping, Jack-knife etc.

26. Define Molecular Clock hypothesis.
    A hypothesis that predicts a constant rate of molecular evolution among species. It is also a method of genetic analysis that can be used to estimate evolutionary rates and timescales using data from DNA or proteins.

27. Define a Root, Node and Clade.
    - Root - Rooted trees have a single lineage at the base representing a common ancestor that connects all organisms presented in a phylogenetic diagram
    - Node - nodes on the tree represent the common ancestors of those descendants
    - Clade - A clade is by definition monophyletic, meaning it contains one ancestor (which can be an organism, a population, or a species) and all its descendants.

28. Write a note on Critical Assessment of Structure Prediction (CASP)?
    The primary goals of CASP are to establish the capabilities and limitations of current methods of modeling protein structure from sequence.

29. How is the distance between species calculated for phylogenetic tree construction?
    The distance between species is calculated using distance based or character based algorithms. There are various algorithms such as UPGMA, Fitch-morgalisch, Neighbouring Joining, Parsimony methods etc.

30. What are the criteria to select the template in homology modeling?
    The criteria for selecting templates also depend on the purpose of a comparative model. For example, if a protein-ligand model is to be constructed, the choice of the template that contains a similar ligand is probably more important than the resolution of the template. On the other hand, if the model is to be used to analyze the geometry of the active site of an enzyme, it may be preferable to use a high-resolution template structure.

31. What are protein visualization tools? Give examples.
    Protein structure visualization softwares is a compilation of bioinformatics software used to view 3-dimensional protein structures. Example: PyMol, RasMol, UCSF Chimera.

32. What are synonymous and non-synonymous mutations?

Synonymous substitution	Nonsynonymous substitution
A synonymous substitution (often called a silent substitution though they are not always silent) is the evolutionary substitution of one base for another in an exon of a gene coding for a protein, such that the produced amino acid sequence is not modified.	A nonsynonymous substitution is a nucleotide mutation that alters the amino acid sequence of a protein. It is contrasted with synonymous substitution which do not alter amino acid sequences. As nonsynonymous substitutions result in a biological change in the organism, they are subject to natural selection.

Part B

1. List out the steps involved in phylogenetic tree construction and discuss with a distance based method?, Cp.4, Pg. 89
2. What is a neighbour joining tree? How it is different from UPGMA? Cp. 4, Pg. 89, Nov-2013, 2014.
3. Compare maximum parsimony and maximum likelihood methods for tree building for method, advantages and limitations. Cp.4, Pg. 89, Nov-2013.
4. How protein structure prediction methods are useful for research and also explain the process of Ab initio protein modeling? Cp.4, Pg.89, Jan 2014, Nov 2016.
5. Differentiate between Ab-initio and Heuristic methods of protein structure prediction, Cp.4, Pg.89, Nov 2016.
6. Explain in detail major steps in pattern recognition and discovery process, Nov 2016.

Part C

1. Explain CASP. Cp.5, Pg. 233.
2. Explain the applications of Structural genomics. Cp.5, Pg.233.
3. What is homology modeling? Cp.5, Pg.233, Jan 2012, 2013.

UNIT IV **11**

Machine learning techniques: Artificial Neural Networks in protein secondary structure prediction, Hidden Markov Models for gene finding, Decision trees, Support Vector Machines. Introduction to Systems Biology and Synthetic Biology, Microarray analysis, DNA computing, Bioinformatics approaches for drug discovery, Applications of informatics techniques in genomics and proteomics: Assembling the genome, STS content mapping for clone contigs, Functional annotation, Peptide mass fingerprinting.

1. Give applications of peptide mass fingerprinting. (November/December 2016)
   Peptide mapping is an important technique for investigating protein primary structures and determining surface-exposed sites or epitopes within proteins. It can be adapted to obtain internal protein sequences. Others include identification of microbes and their metabolic profile, disease biomarker discovery etc.

2. What is the importance of RMS deviation of atoms? (November/December 2016)
   In bioinformatics, the root-mean-square deviation of atomic positions (or simply root-mean-square deviation, RMSD) is the measure of the average distance between the atoms (usually the backbone atoms) of superimposed proteins.

3. Define support vector machines.
   Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

4. What are the applications of support vector machines?
   SVMs have been widely applied to many areas of bioinformatics including protein function prediction, protease functional site recognition, transcription initiation site prediction and gene expression data classification.

5. What is Sequence-Tagged Site (STS)?
   Sequence-Tagged Site (STS) is a relatively short, easily PCR-amplified sequence (200 to 500 bp) which can be specifically amplified by PCR and detected in the presence of all other genomic sequences and whose location in the genome is mapped. STS include such markers as microsatellites (SSRs, STMS or SSRPs), SCARs, CAPs, and ISSRs.

6. Define DNA computing and its applications.
   DNA computing is a branch of computing which uses DNA, biochemistry, and molecular biology hardware, instead of the traditional silicon-based computer technologies. Various applications are; DNA sequencing; DNA fingerprinting; DNA mutation detection or population screening.

7. Define Functional annotation.
   The "unit" of genome annotation is the description of an individual gene and its protein (or RNA) product, and the focal point of each such record is the function assigned to the gene product.

8. What is genome assembly?
   The genome assembly is simply the genome sequence produced after chromosomes have been fragmented, those fragments have been sequenced, and the resulting sequences have been put back together by a process of compiling.

9. Write a note on Machine learning techniques?
   Machine learning is an area of artificial intelligence concerned with the study of computer algorithms that improve automatically through experience. In practice, this involves creating programs that optimize a performance criterion through the analysis of data. The pattern matching and pattern discovery components of data mining are peformed by machine learning techniques.

10. What are Neural Networks?
    An artificial neural network (ANN) is a mathematical model or computational model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase.

11. What are the different architectures used in ANN?
    - Feed-forward networks
    - Feedback networks
    - Back propagation networks

12. What are Hidden Markov Models (HMMs)?
    Hidden Markov Models (HMMs) are probabilistic models for studying sequences of symbols. In particular, HMMs can model matches, mismatches, insertions and deletions of symbols in the given sequence.

13. What are the methodologies in gene identification/gene finding?
    Gene finding can be performed using Artificial Neural Networks, Hidden Markov models and ORF finder etc.

14. Write a note on Decision trees?
    Decision Trees are excellent tools for helping you to choose between several courses of action. They provide a highly effective structure within which you can lay out

options and investigate the possible outcomes of choosing those options. They also help you to form a balanced picture of the risks and rewards associated with each possible course of action.

15. Write a note on Peptide mass fingerprinting?
    Peptide mass fingerprinting, also known as protein fingerprinting is an analytical technique for protein identification. In this method, the unknown protein of interest is first cleaved into smaller peptides, whose absolute masses can be accurately measured with a mass spectrometer such as MALDI-TOF or ESI-TOF.

16. What are the steps involved in drug designing?
    Drug designing is another area of biotechnology research. Drug designing will involve at least three steps;
    - Target or site structure
    - Ligand will have to be designed will fit the binding site or it should be complementary to the site.
    - Once the ligand is designed to fit the binding site, ligand needs to be modified to have pharmacological and toxicological properties while maintaining its affinity for the binding site.

17. Write a note on Microarray?
    Micro arrays are miniature devices comprising a large number of DNA sequences immobilized on a substrate such as glass micro slide. The sequences known as features are arranged as a grid. Arrays are hybridized with a complex probe. The intensity of the hybridization signal for each feature corresponds to the amount of that particular molecule in the probe and this is directly proportional to the level of gene expression in the cell type or tissue from which the probe was prepared.

18. What are the steps involved in Microarray?
    - Building a chip
    - Sample preparation
    - Image generation
    - Scanning and capturing
    - Image analysis
    - Image processing

19. Define clustering in Microarray.
    Clustering is the most popular method currently used in the first step of gene expression matrix analysis. Clustering, much like PCA that is discussed above, reduces the dimensionality of the system and by this allows easier management of the data set. The goal of clustering is to group together objects (i.e. genes or experiments) with similar properties.

20. What are the types of clustering?
    - K-mean clustering
    - Hierarchical clustering
    - Self-organizing feature maps
    - Binning

21. Distinguish between cDNA and Oligonucleotide microarray.

cDNA Microarray	Oligonucleotide Microarray
• cDNAs, clones, or short and long oligonucleotides deposited onto glass slides • Each gene (or EST) represented by its purified PCR product • Simultaneous analysis of two samples (treated vs untreated cells) provides internal control.	• Short oligonucleotides synthesized in situ onto glass wafers • Each gene represented multiply - using 16-20 (preferably non-overlapping) oligomers. • Each oligonucleotide has single-base mismatch partner for internal control of hybridization specifity.
• Flexible and cheaper • Allows study of genes not yet sequenced (spotted ESTs can be used to discover new genes and their functions) • Variability in spot quality from slide to slide • Provide information only on relative gene expressions between cells or tissue samples • Relative gene expressions	• More expensive yet less flexible • Good for whole genome expression analysis where genome of that organism has been sequenced • High quality with little variability between slides • Gives a measure of absolute expression of genes • Absolute gene expressions

22. Define Genomics.

    Genomics is defined as the science dealing with mapping, sequencing and analyzing the gene.

23. What are the advanced techniques used in Genomics.
    a. Nucleic acids-
        i. DNA sequencing- Chain termination method
        ii. High throughput sequencing
    b. Expression studies
        i. SAGE- Serial Analysis of Gene Expression

ii. DNA Microarray – Gene expression study of cDNA or mRNA
iii. Protein chips
iv. Antibody arrays
v. ICAT
vi. Yeast two hybrid systems

c. High throughput methods
    i. DNA sequencing-Shot gun sequencing, Clone by clone method

d. Gel analysis
    i. SDS-PAGE-One Dimensional PAGE
    ii. 2DE-Two Dimensional Gel Electrophoresis

24. Define SAGE.

Serial analysis of gene expression, or SAGE, is an experimental technique designed to gain a direct and quantitative measure of gene expression. The SAGE method is based on the isolation of unique sequence tags (9-10 bp in length) from individual mRNAs and concatenation of tags serially into long DNA molecules for lump-sum sequencing. The SAGE method can be applied to the studies exploring virtually any kinds of biological phenomena in which the changes in cellular transcription are responsible.

25. Define Proteomics

Proteomics is the study of protein expression, regulation, modification, and function in living systems for understanding how living systems use proteins. Using a variety of techniques, proteomics can be used to study how proteins interact within a system, or how proteins change due to applied stresses.

26. Define Proteomics.

Proteomics is the study of protein expression, regulation, modification, and function in living systems for understanding how living systems use proteins. Using a variety of techniques, proteomics can be used to study how proteins interact within a system, or how proteins change due to applied stresses.

27. What are types of proteomics?

- Expression proteomics
- Cell map proteomics
- Functional proteomics
- Structural proteomics

28. Define a shotgun method.

- In short gun methods, two or three preparations of genomic DNA are made.
- One is sheared into short fragments of about 2kb, another is sheared into longer fragments of 15-20kb, and another preparation of 200-300kb fragments might be made.

- A genomic library is constructed from each preparation.
- Clones are selected at random from each library and sequenced.
- Software is used to assemble long stretches of sequences from overlapping short fragments from the libraries, using the sequences from the larger clones as a framework

29. Define Gene annotation.

    Gene annotation means obtaining useful information, that is, the structure and function of genes and other genetic elements, from raw sequence data.

30. Write a note on System biology and its types?

    Systems Engineering + Genomics = Systems Biology

    - Synthesis and integration of data types
    - Systems analysis

    There are two types of system biology

    - Integrative system biology
    - Predictive system biology

31. Write a note on Structural genomics.

    It is also called as the classical genomics. In this approach, genetic mapping, physical mapping and the complete sequencing is done.

32. What are High throughput methods for Proteomics

    - 2D Gel Electrophoresis
    - Immunoprecipitation
    - Yeast two hybrid system
    - Protein chips –Gene expression study
    - Peptide arrays - Gene expression study
    - Affinity purification
    - ICAT
    - MS
    - MS/MS (Tandem Mass spectroscopy)
    - MALDI-TOF
    - MALD-TOF-MS
    - Automated LC-MS/MS
    - Peptide mass finger printing
    - SELDI

Part B

1. Explain how neural network is exploited in gene and secondary structure prediction. Cp.5, Pg.228, Nov-2013, 2014, 2016.
2. What is HMM? How is it used in gene prediction? Cp.4, Pg. 188, Jan 2013.
3. Explain Microarray methodology and clustering types and techniques. Cp.5, Pg.242, Nov-2013, 2015.
4. Explain proteomics and genomics techniques and methodologies. Cp.5, Pg. 225
5. Explain STS mapping in detail. Cp.5, Pg. 225.
6. Describe the bioinformatics tools employed in drug discovery, Nov 2016.

Part C

1. Explain the basic machine learning process with neat diagram. Describe following the machine learning process in brief a) Neural Networks b) Decision trees, Nov 2016.
2. Explain Peptide mass fingerprinting. Cp.5, Pg.225.
3. Explain the advanced techniques used in Genomics.

UNIT V **8**

Basics of PERL programming for Bioinformatics: Datatypes: scalars and collections, operators, Program control flow constructs, Library Functions: String specific functions, User defined functions, File handling.

1. What is SCOP? (November/December 2016)
   The Structural Classification of Proteins (SCOP) database is a largely manual classification of protein structural domains based on similarities of their structures and amino acid sequences. ... The SCOP database is freely accessible on the internet.

2. Write a short note on string specific functions in PERL (November/December 2016)
   Perl provides a very useful set of string handling functions. For example, chop(STRING) OR chop(ARRAY)-- Removes the last character from a string or the last character from every element in an array. The last character chopped is returned.

4. What is Shell script? Give example.
   A *shell script* is just a text file containing a sequence of shell commands
   Example
   $ cat testscript
   #!/bin/sh
   echo Here is a long listing of the current directory
   ls -l

5. What is Scalar data? Give example.
   A single number or string, depending on context
   - References to scalars always begin with $
   - Variable names may contain characters, numbers and underscores
   - Assignment is done using the = operator
   - Examples:
     - $pi = 3.14159;
     - $color = 'red';
     - $old_color = "was $color before";
     - $host = `hostname`; # command substitution #

6. What are strings?
   Sequences of characters
   - No end of string character as in C
   - Single-quoted (note: ', not `)
   - Example
     - $x = "dog"; print 'bob $x'; # displays bob $x

7. What are arrays or ordered lists?
   Arrays or Ordered lists of scalar data items, indexed by an integer, variable starts with a @. There is a separate namespace for scalar and array variables. Arrays are subscripted using square brackets, where indexing begins at 0. The (scalar) variable $#arry is the highest assigned index of the array @arry. Arrays need not be declared; they come into existence when used.

8. Give some array functions?
   Example 1:
   push/pop
   – Add/remove an element to/from the end of an array
   – Either a scalar or a list can be added
   Code

```
push (@a, $b); # same as @a = (@a, $b);
@x = (1,2);
push (@a, @x); # same as @a = (@a, 1, 2);
$c = pop (@a); # returns and removes last element of @a
```

   Example2:
   shift/unshift
   – Add/remove element(s) at the beginning of a list
   – Either a scalar or a list can be added
   Code

```
unshift (@a, $b); # same as @a = ($b, @a);
$c = shift (@a); # returns and removes first element
# of @a
```

9. What are Associative arrays or Hashes?
   An array indexed by arbitrary scalars (not necessarily integers) –Index values are called keys
   - Associative array variable names begin with a %
   - Subscripted using curly braces {}
   - Elements have no particular order

10. Write a note on string operators?
    - Concatenation: .
      ”hello”. ”world”# the same as “helloworld””

    - Repetition: x
      fred”x 3 # same as ”fredfredfred”
      ”Bob”x (1+1) # same as ”BobBob”
      (3+2) x 4 # same as ”5”x 4 or 5555# (note auto-conversion
      # of 5 to”5”)

11. Like C, Perl provides shorthand for incrementing or decrementing variables
    - "Postfix"form: –In an expression, $j first used, and afterwards incremented
      $j++; # the same as $j = $j + 1$j--; # the same as $j = $j -1
    - "Prefix"form: –In an expression, $j first incremented, and then used in the expression
      ++$j; # the same as $j = $j + 1--$j; # the same as $j = $j -1

12. Write a syntax note on if/unless statement?

```
if (expression) {
    true_statement1;
    true_statement2;…
} else {
    false_statement1;
    false_statement2;
}
```

13. Write a syntax note on while/until statement?

```
while (expression) {
statement1;
statement2;
…
}
```

14. Write a syntax note on simple constructs?

    To loop on or branch around a singlestatement, you can use

```
print "Odd\n"if $n % 2 == 1;
print "Even\n"unless ($n % 2 == 1);
print $num--,"\n"while $num >
```

15. What are the data types used in PERL language?
    - Scalar –Scalars are simple variables. They are preceded by a dollar sign ($). A scalar is a number, a string, or a reference. A reference is actually an address of a variable, which we will see in the upcoming chapters.
    - Arrays –Arrays are ordered lists of scalars that you access with a numeric index which starts with 0. They are preceded by an "at" sign (@).
    - Hashes –Hashes are unordered sets of key/value pairs that you access using the keys as subscripts. They are preceded by a percent sign (%).

16. What is the difference between pipe and output redirectory symbol?

17. Pipe operator	18. Output redirectory symbol
19. The "|" (pipe) operator which sends the standard output of one command to another command as standard input.	20. Redirection is done using either the ">" (greater-than symbol)

21. Write a note on output (Print) file handlers?

print–

Takes a list of strings as argument, and sends each to STDOUTin turn

No additional characters are added

Returns a true/false value indicating whether the print succeeded

Examples:

print (”Hello”,”world”,”\n”); # prints “Hello world”with newline

22. What are metacharacters?

These characters have a special meaning inside a regular expression

Character	Meaning
.	Any single character except new line(\n)
*	zero or more of the preceding RE

23. Write a syntax note on string modifiers?

String modification is performed using the “substitute”operator.

$var=~ s/regexp/replacement-string/;

24. Write a syntax note on modifiers

Regular Expressions can have optional modifying suffixes. These include “g”(global; substitute as many times as possible), “i”(case insensitivity), “m”(treat string as multiple lines), and “s”(treat string as a single line)

```
$_ = ”foobaringup the foobarroad”;
$_ =~ s/bar/BEAR/g;      # $_ is now “fooBEARing
                         # up the fooBEARroad”
if ($str=~ /abc/i) { # same as /[Aa][Bb][Cc]/
…
}
$_ = ”fooBARingup the foobarroad”;
$_ =~ s/bar/bear/ig;     # $_ is now “foobearing
                         # up the foobearroad”
```

25. Write a syntax note on diamond operator (I/O function)?
    If the input operator is used without a filehandle, as<>, data are read from the files specified on the command line
    STDINinstead$ cat cat.pl
    #!/usr/local/bin/perl
    while (<>) { print $_; }
    $ cat.plfile1 file2 …filen

26. Write a note on local variables?
    By default, most variables are global in Perl
    @_is local to each function, however
    Can define other local variables using
    local ($var1, $var2, …)
    sub greater_than {
    local ($n, @values) = @_; # create some local variables
    local (@result); # to hold the return value
    foreach$val(@values) { # step through arglist
    if ($val> $n) { # is it eligible?
    push (@result, $val);
    # include it
    }
    }return (@result); # return final list
    }

27. Write a syntax coded for sort?
    By default, sort sorts the elements of a list according to their ASCII values
    The sorting function should assume two arguments $aand $b, and return
    –any negative number if $a is "less than"$b(i.e., if $ashould come before$bin the sorted list)
    –zero if $a"equals"$band
    –any positive number if $a is "greater than"$b (i.e., if$a should come after$b)

28. Write a note on regular expression?
    A function returns a value to the code that called it, which may be assigned or used in some other way. The return value of a function is the value of the last expression evaluated in the body of the function.
    sub double_a {
    $a *= 2;
    }
    $a = 3;
    $c = &double_a; # $c is now 6

29. Write a syntax code for arguments?

Arguments may be passed (in parentheses) to a function.
Any arguments passed to the function appear in the special array@_
@_ is local to the function
–If there is a global variable @_, it is saved and restored after the function exits
No formal (dummy) parameters

```
sub print_msg{
print “First argument: $_[0]\n”;
print “Second argument: $_[1]\n”;
}
&print_msg(“foo”, 42);
```

30. What are subroutines?
    Functions are also called subroutines, or sometimes just “subs”.
    General construct

```
sub my-subname{
statement_1;
statement_2;
...
}
```

31. Write a note on globbing?
    Globbing pattern is used to expand shell wildcards (“globbing”) by putting the globbing pattern inside <>. Multiple patterns are allowed inside the glob, for example
    !@foo_bar_files= <foo* bar*>;
    - Generally, anything you could send to the shell for expansion will work in a glob
    - Note: looks similar to regular expressions, but the meaning of the various metacharactersis very different!

32. How to concatenate two DNA strings?

```
# Store two DNA fragments into two variables called $DNA1 and $DNA2
$DNA1 = 'ACGGGAGGACGGGAAAATTACTACGGCATTAGC';
$DNA2 = 'ATAGTGCCGTGAGAGTGATGTAGTA';
# Print the DNA onto the screenprint "Here are the original two DNA fragments:\n\n";print $DNA1, "\n";
print $DNA2, "\n\n";# Concatenate the DNA fragments into a third variable and print them
# Using "string interpolation"$DNA3 = "$DNA1$DNA2";
print "Here is the concatenation of the first two fragments (version 1):\n\n";print "$DNA3\n\n";
```

33. Write a Perl code how to transcribe DNA to RNA?

```
#!/usr/bin/perl-w
# Transcribing DNA into RNA
# The DNA
$DNA =’ACGGGAGGACGGGAAAATTACTACGGCATTAGC’;
# Print the DNA onto the screen
print "Here is the starting DNA:\n\n";
print "$DNA\n\n";# Transcribe the DNA to RNA by substituting
# all T’s with U’s.
$RNA = $DNA;
$RNA =~ s/T/U/g;
# Print the RNA onto the screen
print "Here is the result of transcribing the
DNA to RNA:\n\n";
print "$RNA\n";
# Exit the program.
exit;
```

34. Write a code to reverse complement a DNA to RNA?

```
#!/usr/bin/perl-w
# Calculating the reverse complement of strand of DNA
# The DNA
$DNA =ACGGGAGGACGGGAAAATTACTACGGCATTAGC’;
# Print the DNA onto the screen
print "Here is the starting DNA:\n\n";
print "$DNA\n\n";
# Make a new copy of the DNA
$revcom= reverse $DNA;
# See the text for a discussion of tr///
$revcom=~ tr/ACGTacgt/TGCAtgca/;
# Print the reverse complement DNA onto the screen
print "Here is the reverse complement DNA:\n\n";
print "$revcom\n";
exit;
```

Part B

1. Explain the importance of arrays, lists and hashes in computational biology with example, Cp.1, Pg. 18 January 2012, 2015.
2. Discuss in detail the different operators used in with respect to manipulation and file handling, Cp.1, Pg.18, January 2012, 2015
3. Write a PERL program to perform nucleotide count in a DNA sequence, Cp.1, Pg.18, January 2012
4. Explain file handling operators in detail, Cp.1 Pg.18.
5. Explain program control flow constructs in detail Cp.3, Pg. 24.
6. Explain library functions, user and string specific functions in detail. Cp.1,Pg.18. Nov 2016.

Part C

1. Describe in detail the program control flow constructs using PERL, Nov 2016.
2. Discuss the importance of string specific functions in PERL, Nov 2016.
3. What are the data types in PERL? Discuss the methods of file handling in PERL, Nov 2016.

Reg. No. :

Question Paper Code : 50202

B.E./B.Tech. DEGREE EXAMINATION, NOVEMBER/DECEMBER 2017

Seventh Semester

Biotechnology

BT 6701 – BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

(Regulations 2013)

Time : Three Hours | Maximum : 100 Marks

Answer ALL questions

PART – A (10×2=20 Marks)

1. How will you erase all files in the current directory including all its sub-directories ?
2. Give the schematic diagram of SQL Architecture.
3. Write about Dot matrix plot.
4. What is BLAST ?
5. What are the uses of phylogenetic trees ?
6. How is the tertiary structure of protein predicted ?
7. What is the difference between synthetic biology and systems biology ?
8. What is Peptide Mass Fingerprinting ?
9. What are Perl features ?
10. What are the basics of handling files in Perl ?

PART – B (5×16=80 Marks)

11. a) Demonstrate any two file transfer mechanisms with examples.

(OR)

b) Describe the key steps in Data Life Cycle process.

12. a) Elucidate the main applications of multiple sequence alignment.

(OR)

b) Trace the optimal alignment for the given sequences using Smith-Waterman algorithm. Sequence-1 : AGGTTTC, Sequence-2 : ACGTTT, (Match score : 2, Mismatch score : 1, Gap score : -1).

13. a) Elaborate the procedure of bootstrap scheme.

(OR)

b) Illustrate structural genomics with its objectives, process and techniques.

14. a) Outline the applications of machine learning approach in biological science.

(OR)

b) Discuss on Adleman's solution of the Hamiltonian Directed Path Problem in detail.

15. a) Give the overview of Perl-syntax.

(OR)

b) Explain how to call a function using user defined functions with an example.

B.E/B.Tech. DEGREE EXAMINATION, NOVEMBER/DECEMBER 2017
Seventh Semester
BT 6701: BIOINFORMATICS AND COMPUTATIONAL BIOLOGY
(Regulation 2013)

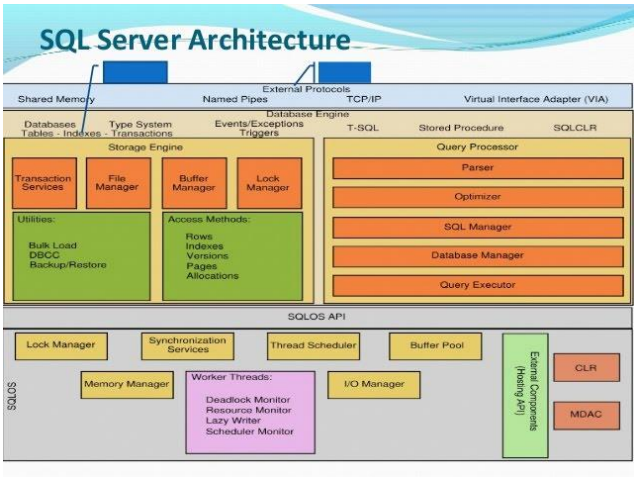
Time : Three Hours **Maximum : 100 Marks**

Part A

Answer all the questions (2 × 10 = 20)

1. How will you erase all files in the current directory including all its sub-directories?
   For just deletion of files on current directory: rm ./* and For deletion of files and folders inside in is rm -R ./*

2. Give the schematic diagram of SQL Architecture

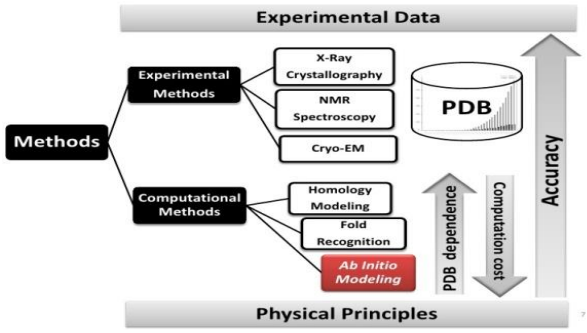


3. Write about Dot matrix plot.
   A dot plot is a graphical method that allows the comparison of two biological sequences and identify regions of close similarity between them. It is a type of recurrence plot.

4. What is BLAST?
   BLAST is a sequence alignment program similar to FASTA. It has speed faster than FASTA and very good sensitivity. It is the most popular sequence alignment algorithm. It finds the ungapped local alignments between a query sequence and a target database by either looking for any short stretch of identities or a very high scoring match. Both the query and the target database can be either nucleotide sequence or amino acid sequence, but it is very effective for amino acid sequences than DNA sequences.

5. What are the uses of phylogenetic trees?
   Phylogenetic trees are used to represent the evolutionary history of a set of n organisms which are often also called taxa within this context. A multiple alignment of a suitable, in a biological context, small region of their DNA or protein sequences can be used as input for the computation of phylogenetic trees.

6. How is the tertiary structure of protein predicted?



7. What is the difference between synthetic biology and systems biology?

Synthetic biology	System biology
The emerging field of synthetic biology combines knowledge from various disciplines including molecular biology, engineering, mathematics, and physics to design and implement new cellular behaviors.	It is a holistic approach to deciphering the complexity of biological systems that starts from the understanding that the networks that form the whole of living organisms are more than the sum of their parts.
The goal of synthetic biology is both to improve our quantitative understanding of natural phenomenon as well as to foster an engineering discipline for obtaining new complex cell behaviors in a predictable and reliable fashion.	It is collaborative, integrating many scientific disciplines – biology, computer science, engineering, bioinformatics, physics and others – to predict how these systems change over time and under varying conditions, and to develop solutions to the world's most pressing health and environmental issues.

8. What is Peptide Mass Fingerprinting?
   Peptide mass fingerprinting, also known as protein fingerprinting is an analytical technique for protein identification. In this method, the unknown protein of interest is first cleaved into smaller peptides, whose absolute masses can be accurately measured with a mass spectrometer such as MALDI-TOF or ESI-TOF.

9. What are Perl features?
   - Perls database integration interface DBI supports third-party databases including Oracle, Sybase, Postgres, MySQL and others.
   - Perl works with HTML, XML, and other mark-up languages.
   - Perl supports Unicode.

10. What are the basics of handling files in Perl?
    Three basic file handles are - STDIN, STDOUT, and STDERR, which represent standard input, standard output and standard error devices respectively.

Part B

Answer all questions (5×16 = 80)

1. Demonstrate any two file transfer mechanisms with example. (OR)
2. Describe the key steps in Data Life Cycle process. **Cp. 3, Pg.3-12, Nov-2017**
3. Elucidate the main applications of multiple sequence alignment. **Cp.4, Pg. 154, Jan 2014, 2015, Nov 2017.**
4. Trace the optimal alignment for the given sequences using Smith-Waterman algorithm. Sequence-1: AGGTTTC, Sequence-2: ACGTTT (Match score: 2, Mismatch score: 1, Gap score: -1) **Cp.4, Pg.154, Nov-2016, 2017**.
5. Elaborate the procedure of bootstrap scheme. **Cp.4, 198 Nov-2017**(OR)
6. Illustrate structural genomics with its objectives, process and techniques. **Cp.5, Pg.233.Nov 2017.**
7. Outline the applications of machine learning approach in biological science **Cp.5, Pg.228, Nov-2013, 2014, 2016, 2017** (OR)
8. Discuss on Adleman's solution of the Hamiltonian Directed Path Problem in detail.
9. Give the overview of Perl-syntax **Cp.1, Pg. 18 January 2012, 2015, Nov-2017**
10. Explain how to call a function using user defined functions with an example. **Cp.1,Pg.18. Nov 2016, 2017**