BIOINFORMATICS & BIG DATA ANALYSIS

Syllabus of the theory papers

BiSEP1: Bioinformatics

Total Hours : 52

Unit - I

Introduction to Bioinformatics **10 Hrs**

History of Bioinformatics, Role of Bioinformatics in biological sciences, scope of bioinformatics. Introduction to internet: WWW, network basics, LAN & WAN standards. Network topologies and protocols: ftp, http. Introduction to Database: Types of database. Biological Database: Need of biological database, Sequence and Structure database – (NCBI, EMBL, DDBJ, and PDB), other databases - KEGG, PubMed, OMIM, PubChem, NCI, ZINC, Drug Bank, Ligand. Format of Databases: GenBank and PDB flat file. Protein Structure Visualization: RasMol, PyMol, Jmol, CN3D, Swiss PDB viewer, Chimera and Discovery Studio visualizer. Protein Structure Comparison: Intra-molecular Method, Intermolecular method, combined method. Protein Structure Comparison: SCOP and CATH.

Unit - II **10 Hrs**

Sequence Alignment and Motif, Domain Prediction

Pairwise Alignment: Dot Matrix Method, Dynamic programming - (Local and Global Alignment) Gap Penalties, POA Alignment. Scoring Matrices: Amino acid scoring matrices, PAM, BLOSUM. Database Similarity Searching: BLAST. BLAST variants. BLAST output format. FASTA. Multiple Sequence Alignment: Scoring function, exhaustive algorithms, and Heuristic algorithms. PSSM, Markov Model and Hidden Markov Model. Protein Motif and Domain Prediction: Motif and Domain Databases PROSITE. Sequence Logos and Web-logo.

Unit - III **10 Hrs**

Gene and Promoter Prediction and Phylogenetic

Gene Perdition in Prokaryotes**:** Conventional determination of Open Reading Frames (ORF), Markov model and HMM. Gene Prediction in Eukaryotes: An Initio based program, Neural Networks. Promoter and Regulatory Element Prediction: Prokaryotes and Eukaryotes. Introduction to Phylogenetic: Phylogenetic Basics, Terminologies. Phylogenetic Tree construction Methods: Distant based method - (UPGMA, NJ) Character Based Method – (MP and ML), Phylogenetic Tree Evaluation: Bootstrapping.

Unit - IV **12 Hrs**

Protein Structure Prediction and Molecular Dynamics

Globular Proteins: Ab-Inition, Homology Based, Neural networks method. Transmembrane Proteins: Prediction of Helical membrane, β-barrel membrane proteins. RNA Structure Prediction: Ab Initio approach, dot matrices. Introduction to Homology modeling: Model refinement, model evaluation, homology model databases. Threading and fold recognition, CASP. Introduction of Molecular Modeling: Coordinate system, potential energy. Steps in

Molecular Modeling: introduction to Quantum Mechanics, introduction to Molecular mechanics. Force Filed: Types of force fields: Amber force field, CHARMM force field. Introduction about molecular dynamics (GROMACS).

Unit - V **10 Hrs**

Drug Discovery

Introduction: Drug Discovery Process, Molecular Modeling in Drug Discovery, Molecular Docking, Quantitative Structure-Activity Relationship (QSAR). Chemoinformatics: Introduction, stereoisomer, origin of stereospecificity in molecular recognition, importance of stereochemistry in drug design. Docking and Virtual Screening: Using different docking algorithms, Optimization of docking algorithms based on different target. Ligand - Receptor Interactions: Docking software's (AUTODOCK, LEAD IT), Post docking analysis. Pharmacokinetics: Absorption, Distribution, Metabolism, Excretion and Toxicity of drugs.

Text Books

1. David W Mount, "Bioinformatics sequence and Genome analysis", Second Edition, Cold Spring Harbor Laboratory Press, 2013.
2. Attwood T K, D J Parry-Smith, "Introduction to Bioinformatics", Pearson Education, 2005.

References

1. Neil C. Jones and Pavel A. Pevzner, "An Introduction to Bioinformatics Algorithms", MIT Press, 2005.
2. Steffen Schulze-Kremer, "Molecular Bioinformatics: Algorithms and Applications", Walter de Gruyter, 1996.

BiSEP2: Genome Informatics and Big Data Analytic

Total Hours : 52

Unit - I **10 Hrs**

Introduction to Genome Structure and Organization

Basics of Molecular Biology, Introduction to DNA & RNA, Structural and Functional aspects of DNA, Structural and Functional aspects of RNA, Types and variants of RNAs (Coding & Non-coding), Central Dogma, Genes and Proteins, Prokaryotic & Eukaryotic CELL structures, Genome size - sequence complexity -Introns and Exons, Genome sequences and database subscriptions. Genome organization: Mitochondrial and Chloroplast genome.Genomic regulatory elements and their role: Promoters, Enhancers & CpG Islands; Genomic Repetitive Elements & their role: Long Repeats, Short Repeats, SSRs, Transposons / Transposable elements, Miniature Inverted Repeat TE, etc.; DNA Modifications: Variations and Base modifications, Polymorphisms: types of polymorphism Single Nucleotide Polymorphisms (SNPs), mutations, other genetic variations, Introduction to Structural Variation and implications on genomes.

Unit - II **12 Hrs**

Introduction to Genome Informatics

Microarray analysis definition, types of microarray, microarray analysis life cycle (sample preparation and labeling, hybridization, washing and image acquisition), microarray data analysis, tools, databases and software for microarray data analysis. Past, present and feature of sequencing technology. Platform overview: Illumina, Pacific Biosciences. Comparison of NGS Systems: Recent scientific breakthroughs using NGS technology. Major biological databases and its classification, sequence database - NCBI, GenBank, EMBL, DDBJ. NGS Database: SRA, DRA, ENA. File/Data formats overview: FASTA, FASTQ, FNA, CSFASTA, GFF, SAM and BAM. Genome alignment and analysis tools- BWA (Burrows-Wheeler Aligner), SAMtools, GATK (The Genome Analysis Toolkit), IGV (Integrative Genomics Viewer), HISAT, StringTie, Cuffcompare, Velvet, Oases, Trinity. Advantage and disadvantage of NGS Technology.

Unit - III **10 Hrs**

Whole Genome / Exome / Targeted Resequencing Analytics

Introduction to genome Re-Sequencing, Indexing the reference genome, Sequence Alignment Tools and its Parameters, Alignment quality Assessment, Exome Enrichment Analysis, Target /Non-Target Enrichment Analysis, Statistical Analysis and genome Visualization, Introduction to Variation Analysis, Variation analysis to identify SNV / MNV / SV, dbSNP Annotation / Variation Effect Prediction, Variation Frequency Analysis, Exome Copy Number Variation Analysis, Data Visualization, Function & Structure based Comparative Genome Analysis.

Unit - IV **10 Hrs**

Transcriptome Resequencing and Chip Sequencing Analytics

Introduction to RNA-Seq Sequencing Alignment, Indexing the reference genome, Alignment Tools and its Parameters, Aligning Single End / Paired End reads to the indexed genome, Alignment quality Assessment, Statistical Analysis and genome Visualization, Qualitative &Quantitative Gene Expression Profiling, Gene Modifications & Alternative Splicing Analysis,

Gene Fusion identification, Differential Gene Expression Profiling, Gene Ontology and Pathway Analysis; Introduction to ChIP Sequencing Experimental Designs, Aligning ChIPSeq data to genome, Peak Calling Analysis, Replicate / IDR Peak Analysis, Peak Annotation, Peak Visualization tools, Motif Analysis, Statistical Analysis, Significant Biology Analysis for Annotated Genes.

Unit - V **10 Hrs**

Denovo Whole Genome and Transcriptome Assembly analytics

Introduction to Whole Genome De-novo Sequencing, Understanding Various Assembly Algorithms, Assembly Tools and its Parameters, Scaffolding and Constructing Draft Genome, Repeat Identification and Masking Analysis, SSR Marker Identification & Analysis, Introduction to Gene Prediction Algorithms (Coding & Non-coding), Gene Prediction Tools and its parameters, Sequence Homology Based Annotation & Gene Ontology / Pathway Mapping, Introduction to Transcriptome De-novo Sequencing, Assembly Tools & Parameters, Transcriptome Clustering and Assembly Evaluation, Qualitative & Quantitative Analysis of Assembled Transcriptome, SSR Marker Identification & Analysis, Differential Gene Expression Profiling (For Multiple Samples), Gene Ontology and Pathway Analysis.

Text Books

1. Ali Masoudi-Nejad, Zahra Narimani, Nazanin Hosseinkhan; “Next Generation Sequencing and Sequence Assembly”, Methodologies and Algorithms, Springer; 2013.
2. Sumitabha Das, “Unix Concepts and Applications”, McGraw - Hill; 4 edition. Units I and II – Chapters in book 1,2,3,4,7,10,14, (2006).

References

1. Mark I. Rees, “Challenges and Opportunities of Next-generation Sequencing for Biomedical Research”, Academic Press, 2012.
2. Wu, Wei, Choudhry, Hani (Eds.), “Next Generation Sequencing in Cancer Research: Volume 1: Decoding the Cancer Genome”, Springer, 2013.

BiSEP3: Computer - Programming

Total Hours : 52

Unit - I **8 Hrs**

Operating System Concepts and Linux Environment

Introduction to O.S., types of O.S., O.S services, system calls, system components, system structures, virtual machines. Linux: Introduction to Linux, basic commands (Navigation and Directory Control Commands. File Maintenance Commands, Display Commands, Print Commands etc) , working with the files, file attributes, pipes, wildcards, working with processes working with basic editors (vi, emacs). Basic regular expressions, string search applications using regular expressions.

Unit - II **12 Hrs**

Computer Environment

Introduction: What is a grid? -Infrastructure of hardware and software -Main Projects and Applications –The Open Grid Forum -International Grid Trust Federation. Grid Architecture - Overview of Resource Managers - Overview of Grid Systems - Application Management : Grid Application Description Languages –Application Partitioning -Meta-scheduling –Mapping – Monitoring - Web Services - Grid Portals - Clouds.Cluster computing at a glance – cluster classifications- cluster middleware – cluster applications – cluster setup and administration – multi path communication – distributed shared memory - representative cluster system: Biowulf – RWC PC cluster II- Parallel Processing on Linux Clusters. Java for HPC: java and different flavors of parallel programming models. HPC program optimization. API. Remote Desktop in Windows and linux operating systems.

Unit - III **12 Hrs**

Awk / Shell Scripting Fundamentals

Execution, Fields and Records, Scripts, Operations, Patterns, Actions, Associative Arrays, String Functions, String Functions, Mathematical Functions, User – Defined Functions, Using System commands in awk, Applications, awk and grep, sed and awk.Unix Session, Standard Streams, Redirection, Pipes, Tee Command, Command Execution, Command-Line Editing, Quotes, Command Substitution, Job Control, Aliases, Variables, Predefined Variables, Options, Shell/Environment Customization.

Unit - IV **10 Hrs**

Introduction to Perl And Python

An overview of Perl: Getting started, Statement blocks, ASCII, Unicode, Escape sequences, White spaces, Numerical data types, strings in Perl. Operators, Variables: special variables, regex (regular expression) variables, Input/output variables, Filehandle variables, error and system variables. Perl statements, Introduction to statements, Types - Input/Output statements, conditional statements, looping, andjumping statements.

Python : Simple values – Booleans, Integers, Floats and Strings, Expressions – Numerical operators, Logical Operations, String Operations, Names, Functions and Modules – Assigning

Names, Defining the functions – Function parameters, Comments and Documentation, Assertions, Default parameter values, Using Modules – Importing, Python Files.

Unit - V **10 Hrs**

Introduction to R and MATlab

Overview of the R language: Defining the R project, Obtaining R, Generating R codes, Scripts, Text editors for R, Graphical User Interfaces (GUIs) for R, Packages. R Objects and data structures: Variable classes, Vectors and matrices, Data frames and lists, Data sets included in R packages, Summarizing and exploring data, Reading data from external files, Storing data to external files, Creating and storing R workspaces. Manipulating objects in R: Mathematical operations, Basic matrix computation, Textual operations, Basic graphics. Introduction to MATLab and molecular forces; Bioinformatics ToolBox, Statistics ToolBox, Distributed computing server, Signal Processing ToolBox. The Matlab working environment.Variables, constants and reserved words. Arrays and matrices. Scripts. The debugger. Generating 2D and 3D Graphics. Simple statistical analysis. String manipulation. Boolean logic and if statements. Loops (while, for). Functions & Files. Program design. MATlab structures. Complexity.

TEXT BOOKS

1. RajkumarBuyya, “High Performance Cluster Computing: Programming and Applications”, Prentice Hall, 1999.
2. James D. Tisdall, “Beginning Perl for Bioinformatics”, O’Reilly, 2001.
3. Unix and shell Programming Behrouz A. Forouzan, Richard F. Gilberg.Thomson

REFERENCES

1. Michael J Quinn, “Parallel programming in C with MPI and OpenMP”, Tata McGraw-Hill, 2003
2. Ahmar Abbas, “Grid Computing: A Practical Guide to Technology and Applications”, Charles River Media, 2003.
3. Kevin Dowd, “High Performance Computing”, O’Reilly, 1993.

BiSEP4 : Syllabus of the Elective papers

(Choose from any one of the following)

BiSEP4a: Business Development

Total Hours : 52

Unit - I **10 Hrs**

Essentials of product development

Company protocols for research, privacy policies, institutional and professional code of ethics and standards of practice, IPR guidelines, Knowledge of basic laboratory procedures, GLP and GMP, relevant EOPs, SOPs, process flows in manufacturing, product life cycle and product properties, competitor products. Stability studies – generate stability data & prepare stability reports for innovation products

Unit – II **10 Hrs**

Reporting and documentation

Reporting – different standard reference materials used like drugs, products, side effects, adverse reactions, process details, statistical analysis of test data. Documentation – methods and procedures of writing and maintaining lab, research records, research performance reports, schemes and guidelines, power point presentations, tables, charts, word documents, development of research objectives and proposal writing for funding and contractual purposes, publications and technical writing, Regulatory compliance of the final documents.

Unit – III **08 Hrs**

Planning and communication

Research planning – resource, time, timeline & budget considerations, technical feasibility analysis on the NPD ideas by analyzing current development plans, plan day to day activities. Research communications - preparation of progress reports/ research outcomes for steering groups/ bodies, principal investigator, communication with upstream and downstream teams.

Unit - IV **08 Hrs**

Problem solving and decision making

Research initiatives – use new areas of research, techniques and methods, extend research/ product portfolio, creative analysis & interpretation of research data. Decision making – collaborative, appropriate, optimum & best possible solution, Trouble- shoot & Resolve problems to avoid delays.

Unit - V **08 Hrs**

Safety and Security at workplace

Different types of occupational health hazards, knowledge of chemical substances, characteristics & safety measures, use of safety gears, masks, gloves & accessories, evacuation procedures for workers & visitors. Health, safety & security issues – types (illness, fire accidents), company policies and procedures, When and how to report, summon medical assistance & emergency services

Unit - VI **08 Hrs**

Interpersonal Skills

Understand work output requirements, company rules, guidelines & policies related to the process flow, identifying and reporting issues requiring intervention, delivery of quality work on time & report any anticipated reasons for the delay, effective interpersonal communication, conflict-resolution techniques, importance of collaborative working, multi-tasking, training the team members, knowledge of project management.

BiSEP4b :Product Development

Total Hours : 52

Unit 1 – Essentials of product development **12 h**

Company protocols for research, privacy policies, institutional and professional code of ethics and standards of practice, IPR guidelines, Knowledge of basic laboratory procedures, GLP and GMP, relevant EOPs, SOPs, competitor products. Biosafety assessment procedures for transgenic food crops, case studies of relevance. Use of transgenics and their release in environment, legal implications. Stability studies – generate stability data & prepare stability reports for innovation products.

Unit 2 - Reporting and documentation **10 h**

Reporting – product development, adverse reactions, process details, statistical analysis of test data- Correlation and Regression. Chi-square test- Analysis of variance and Covariance. Documentation – methods and procedures of writing and maintaining lab, research records, research performance reports, schemes and guidelines, power point presentations, tables, charts, word documents, development of research objectives and proposal writing for funding and contractual purposes, publications and technical writing, Regulatory compliance of the final documents.

Unit 3 - Planning and communication **8 h**

Research planning – resource, time, timeline & budget considerations, technical feasibility analysis, plan day to day activities. Research communications - preparation of progress reports/ research outcomes for steering groups/ bodies, principal investigator, communication with laboratory and field trial teams.

Unit 4 - Problem solving and decision making **6 h**

Research initiatives – use new areas of research, techniques and methods, extend research/ product portfolio, creative analysis & interpretation of research data. Decision making – collaborative, appropriate, optimum & best possible solution, Trouble- shoot & Resolve problems to avoid delays.

Unit 5 – Safety and Security at workplace **8 h**

Different types of occupational health hazards, knowledge of chemical substances, characteristics & safety measures, use of safety gears, masks, gloves & accessories, evacuation procedures for workers & visitors. Health, safety & security issues – types (illness, fire accidents), company policies and procedures, When and how to report, summon medical assistance & emergency services

Unit 6 – Interpersonal Skills **8 h**

Understand work output requirements, company rules, guidelines & policies related to the process flow, identifying and reporting issues requiring intervention, delivery of quality work on time & report any anticipated reasons for the delay, effective interpersonal communication,

conflict-resolution techniques, importance of collaborative working, multi-tasking, training the team members, knowledge of project management

References:

- Kothari C.R, Research Methodology- Methods and Techniques, New Age International, New Delhi
- IPR in Agricultural Biotechnology by Erbisch F H and Maredia K M. Orient Longman Ltd.
- Safety Considerations for Biotechnology, Paris, OECD. Biosafety Management by P.L. Traynor, Virginia polytechnic Institute Publication.

Bryman, Alan & Bell, Emma (2011). Business Research Methods (Third Edition), Oxford University Press.

Tzosts, GT, Head, GP and Hull, R. 2010. Genetically Modified Plants: Assessing Safety and Managing Risk. Academic Press.

Syllabus of the practical papers

BiSEP5L: Bioinformatics Laboratory

LABORATORY EXPERIMENTS

1. Scientific article retrieval from bibliography database and working with referencing format
2. Physicochemical properties of Nucleic acid and Protein using online and commercial software's.
3. Sequence retrieval from Nucleic Acid and Protein databases and Pair wise sequences comparison.
4. Sequence (FASTA and BLAST) searches – Analysis of parameters affecting alignment.
5. Multiple sequences alignment and Protein and DNA motif searches.
6. Evolutionary studies / Phylogenetic analysis – Analysis of parameters affecting trees.
7. Gene Prediction for Prokaryotes and Eukaryotes genome.
8 .Bio-molecules structure visualization and analysis using free and commercial software's.
9. Predict secondary structures of Globular and Membrane proteins.
10. Predict tertiary structures and validation - using online and commercial software's.
11 .Ligand design using Marvin sketch and identification of biological activity using PASS Server.
12 .Molecular Docking using Autodock and LeadIT.
13 .Demo: Molecular Simulation using GROMACS.
14. Demo: Molecular Simulation using Discovery Studio 3.5.

BiSEP6: Genome Informatics and Big Data Analysis Laboratory

1. Working with NGS databases and NGS file formats.
2. Quality checking and trimming using free and commercial software.
3. Bacterial genome assembly using Velvet and Soap Denovo assembly.
4. References genome assembly using BWA and CLC Genomic Workbench.
5. Genome Annotation using Gene Ontology (GO).
6. Identification of SNPs using Cancer genome datasets (GATK Pipeline).
7. Genome browser- UCSC and Ensemble genome browser and comparative genomics.
8. Whole genome (WGS), Transcriptome (RNA, Exome) and Chip-Seq analysis using Cloud based server.
9. Unix/Linux Command Line mode, file and directory handling, Vi Editor.
10. Unix shell scripts – conditional operators, looping, string handling.
11. Working with Hadoop and Basic operations in Excel sheet and calculation of big datasets.
12. Basic R commands, Normalization and Gene expression studies on GEO datasets.
13. Overview of Python – Working with nucleic acid and protein sequences.
14. Analyzing the 3D structure using Python programming.