

**PONDICHERRY UNIVERSITY**  
**SCHOOL OF LIFE SCIENCES**

Centre for Bioinformatics

**Modular Courses**

(under MIT)

Centre for Bioinformatics

Eligibility: B.Sc., (All science Degrees), B.Sc.,(Agri.), B.Tech., M.B.B.S.,  
B.Pharm., B.D.S., B.V.Sc.

---

**CONTENTS**

Module I	Bioinformatics and IT	Credits:10
Paper I	Introduction to Bioinformatics	3
Paper II	Application of computing in Bioinformatics	3
Paper III	Lab – Bioinformatics	2
Paper IV	Lab – C programming, Perl in Bioinformatics	2

Module II	Bioalgorithms and sequence Analysis	Credits:10
Paper I	Algorithms in Bioinformatics	3
Paper II	Genome and Protein sequence Analysis	3
Paper III	Lab – Algorithms in Bioinformatics	2
Paper IV	Lab – Sequence Analysis and Phylogenetic Analysis	2

Module III	Structural Bioinformatics	Credits:10
Paper I	Biophysics and Structural Biology	3
Paper II	Molecular Modeling and Drug Designing	3
Paper III	Lab – Simulations in Macromolecules and Molecular Interactions	2
Paper IV	Lab – Docking and QSAR Analysis	2

Module IV	Applications of Bioinformatics / Experimental Bioinformatics	Credits:10
Paper I	Systems Biology and Profile Analysis	3
Paper II	Lab – Pathway Reconstruction and e-modeling	2
Paper III	Project – Viva Voce	5

## MODULE – I: BIOINFORMATICS AND IT

### PAPER - I – BIOINFORMATICS (3 CREDITS)

#### Unit-I

**Bioinformatics: an overview** - Introduction to Computational Biology and Bioinformatics; some of the biological problems that require computational methods for their solution; Role of internet and www in bioinformatics.

#### Unit-II

**Biological Data Acquisition** – The form of biological information; DNA sequencing methods – basic DNA sequencing, automated DNA sequencing, DNA sequencing by capillary array and electrophoresis; Types of DNA sequences – genomic DNA, cDNA, recombinant DNA, Expressed sequence tags (ESTs), Genomic survey sequences (GSSs); RNA sequencing methods; Protein structure determination methods; gene expression data.

#### Unit-III

**Databases : Format and Annotation** – Conventions for databases indexing and specification of search terms; Common sequencing file formats – NBRF/PIR, FASTA, GDE; Files for multiple sequence alignment – multiple sequence format (MSF), ALN format; Files for structural data – PDB format and NMR files; Annotated sequence databases – primary sequence databases (GenBank-NCBI, the nucleotide sequence database-EMBL, DNA sequence databank of Japan-DDBJ); Subsidiary data storage (ESTs, dbESTs, GSSs), unfinished genomic sequence data, organisms specific databases (EcoGene, SGD, MatDB, TAIR, FlyBase, OMIM, etc.); Protein sequence and structure databases (PDB, SWISS-PROT and TrEMBL); List of Gateways (NCBI, GOLD, MIPS, TIGR, UniGene)

#### Unit-IV

**Data : Access, Retrieval and Submission** – Data access – standard search engines, Data retrieval tools – Entrez, DBGET and SRS (sequence retrieval systems); Software for data building; Submission of new and revised data.

#### Unit-V

**Sequence Similarity Searches** – Sequence homology as product of molecular evolution; Sequence similarity searches; Significance of sequence alignment; Sequence alignment – global, local and free-space; Alignment scores and gap penalties; Measurement of sequence similarity; Similarity and homology.

**Recommended Texts:**

1. Mount, D. (2004) “Bioinformatics: Sequence and Genome Analysis”; Cold Spring Harbor Laboratory Press, New York. (ISBN 0-87969-712-1)
2. Baxevanis, A.D. and Francis Ouellette, B.F. (1998) “Bioinformatics – a practical guide to the analysis of Genes and Proteins”; John Wiley and Sons, New Jersey, USA.
3. Pevzner, P.A. (2004) “Computational Molecular Biology”; Prentice Hall of India Ltd, New Delhi.

## **MODULE – I : BIOINFORMATICS AND IT**

### **PAPER - II – APPLICATIONS OF COMPUTING IN BIOINFORMATICS (3 CREDITS)**

#### **Unit-I**

**Concepts in Computing** – Overview and functions of computer systems – Operating System Concepts – Linux Operating System – DOS Commands, , Batch Commands.

#### **Unit-II**

**C/C++** – Algorithms, flow-charts, programming languages, compilation, linking and loading, testing and debugging, documentation – C programming – variables and identifiers, data types, Conditional statements and loops, Structured Programming, Library Functions – C++ Programming – Introduction and Concept of OOP.

#### **Unit-III**

**Perl - Data Structures and modular programming** – Basic Perl Data Types, References, Matrices, Complex/Nested Data Structures, Scope (my, local, our), Function/Subroutines, System and User Function, File handle and File Tests, stat and lstat Functions, Formats, System Information, Perl Modules, CPAN Modules

#### **Unit-IV**

**Perl- Regular expressions and Pattern Matching** – Uses of Regular Expressions, Patterns, Single-Character Patterns, Grouping Patterns (Sequence, Multipliers, Parentheses as memory, Alternation) Anchoring Patterns, Precedence, Matching Operators, Ignoring Case, Different Delimiter, Variable Interpolation, Special Read-Only Variables, Substitutions, Split and Join Functions, Dynamic Programming, Approximate String Matching

#### **Unit-V**

**BioPerl** – Installing Bioperl, General Bioperl Classes, Sequences (Bio::Seq Class, Sequence Manipulation), Features and Location Classes (Extracting CDS), Alignments (AlignIO), Analysis (Blast, Genscan), Databases (Database Classes, Accessing a local database), Implementing REBASE

## Reference Books

1. Balaguruswamy, E. (1985) “Computer Fundamentals and Applications”, Second Edition, Tata McGraw-Hill Publishing Co. Ltd., India.
2. Fundamentals of Data Structures, E. Horowitz and S. Sahani, Galgotia Booksource Pvt. Ltd., (1999)
3. Ritchie, D.M. (1996) “The C programming language”, Second Edition, Prentice Hall Publishers, USA.
4. Lafore, R. (2002) “Object Oriented Programming using C++”, Fourth Edition, Sams Publishers.
5. Wall, W., Christiansen, T. and Orwant, J. (2000) “Programming Perl”, Third Edition, O’Reilly Publishers.
6. Tisdall, J. (2004) “Beginning Perl for Bioinformatics”, First Edition, O’Reilly Publishers.

## **MODULE – I : BIOINFORMATICS AND IT**

### **PAPER - III – LAB – BIOINFORMATICS (2 CREDITS)**

#### **Exercises:**

1. Entrez and Literature Searches.
2. Sequence Retrieval System of Biological Databases
3. File format conversion
4. Sequence Analysis
5. Phylogenetic analysis using PHYLIP, Phylodraw, PAUP, Treeview, JalView.
6. Usage of Softwares:
  - a. BioEdit
  - b. GeneDoc
  - c. ClustalW / X, MEGA, MEME
7. Usage of Visualization Tool
  - a. RasMol
  - b. Cn3D
  - c. MolMol

## MODULE – I : BIOINFORMATICS AND IT

### PAPER - IV – LAB – C Programming and PERL in Bioinformatics (2 CREDITS)

#### Exercises:

- 1. Operating System :** Overview of Linux Architecture  
File Management  
DOS Commands
  
- 2. C Programming :** Exercises in Basic Programming
  
- 3. C++  
Programming :** Exercises in OOP  
Class Handling – Examples
  
- 4. PERL :** String Matching  
Sequence Manipulation  
File Format Conversions  
Identification of SNPs / CNVs in Genome Sequences

## MODULE – II : BIO-ALGORITHMS AND SEQUENCE ANALYSIS

### PAPER - I - ALGORITHMS IN BIOINFORMATICS

(3 CREDITS)

#### Unit-I

**Computing Algorithms** – Algorithms in Computing, Analyzing algorithms, Designing algorithms, Asymptotic notation, Standard notations, Big ‘O’ notations, Time and space complexity of algorithms and common functions

Sets: Union and Intersections, Differences, Disjoint Sets, Counting Elements, Relations

Matrices: Adding and Multiplying, Extracting a sub-matrix, Combining, Inverting

#### Unit-II

**Sorting, Searching & Strings Matching** – Sorting: Bubble Sort, Insertion sort, Selection sort, Quick Sort, Radix sort, Exchange sort, Shellsort, Mergesort. External sort (K-way mergesort, balanced mergesort, polyphase mergesort) Sorting in Linear time, Heaps (Binary Heaps, Janus Heap, Heap sort, Binomial Heaps, Fibonacci Heaps)

Searching: Binary Search, Fibonacci Search, Hash Search, Lookup Searches, Generative Searches

String Matching: Naïve algorithm, Boyer-Moore algorithm, Knuth-Morris-Pratt algorithm

#### Unit-III

**Graphs** – Representation of Graphs, Breadth First Search, Depth First Search, Topological Sort, Connected Components, Minimum Spanning Tree, Single-Source Shortest Path (Dijkstra’s and Bellman Fort Algorithm), All-Pairs Shortest Paths (Floyd-Warshall algorithm), Coloring of Graphs (Kruskal’s Algorithm, Prim’s Algorithm),

#### Unit-IV

**Trees** – Forests, DAGs, Ancestors, and Descendants, Binary Search Trees, Querying a Binary search tree, Insertion and Deletion, Tree Traversals, Red-Black Trees, Properties of Red-Black Trees, AVL-Trees, Rotations, Insertion, Deletion, B+ Tree, B\* Trees.

#### Unit-V

**Algorithm Design and Analysis** – The substitution method, The iteration method, The master method, Divide and Conquer, Greedy Algorithms, Dynamic Programming (Traveling Sales Person Problem, Hamiltonian Path Problem), Backtracking Algorithms (8-queens Problem, Graph Coloring), Branch and Bound Algorithms.

**Recommended Texts:**

1. E. Horowitz and S. Sahani, "Fundamentals of Data structures", Galgotia Booksource Pvt. Ltd., (1999).
2. Ellis Horwitz, Sartaz Sahani and Sanguthevar Rajasekaran, (1999), "Computer Algorithms", Galgotia Publications.
3. T .H. Cormen, C. E. Leiserson, R .L. Rivest (2001) "Introduction to Algorithms", 3<sup>rd</sup> Ed PHI.

## MODULE – II : BIO-ALGORITHMS AND SEQUENCE ANALYSIS

### PAPER - II - GENOME AND PROTEIN SEQUENCE ANALYSIS (3 CREDITS)

#### Unit I

**Sequence Analysis** – Methods of sequence alignment: graphic similarity comparison; Dot plots; Hash tables; Scoring matrices – identify matrix, genetic code matrices (GCM); Substitution matrices, Mutation Data Matrices (MDM), Percentage accepted Mutation (PAM). Block Substitution Matrices (BLOSUM), mutation probability matrices; Sequence similarity searches and alignment tools – dynamic programming algorithms; Needleman-Wunch and Smith Waterman; alignment scores and gap penalties; measurement of sequence similarity; percentage of identically aligned residues; Optimal global alignment and optimal local alignment;

#### Unit-II

**Pairwise Sequence Alignment** – Concept; Programmes (Dot matrix, Dot plot, Dynamic programming); Similarity Searches; Sequence repeats and inversion; Database searching (BLAST and FASTA).

#### Unit-III

**Multiple Sequence alignment (MSA)** – significance; softwares (PIMA, Clustal, Pileup, ClustalW, Meme, MACAW); Considerations while choosing a MSA software for analysis; sensitivity and specificity of each software.

#### Unit-IV

**Comparative Genome Analysis** – Relevance of comparative genomics; orthologs and paralogs; Comparative genomics of prokaryotes; Minimal genome; Vertical and horizontal gene transfer. Comparative genomics of organelles; Comparative genomics of eukaryotes. Differences and similarities in genomes of organisms; Evolution of protein families; Applications of comparative genomics in reconstruction of metabolic pathways.

#### Unit-V

**Phylogenetic analysis** – Phylogenetics, cladistics and ontology; Phylogenetic representations – graphs, trees and cladograms; Classification and ontologies; Steps in phylogenetic analysis; Methods of phylogenetic analysis – similarity and distance tables, distance matrix method; Method of calculation of distance matrix (UPGMA, WPGMA); The Neighbour Joining Method; The Fitch/Margoliash method; Character-based Methods – maximum parsimony, maximum likelihood; Reliability of Phylogenetic trees; Steps in constructing alignments and phylogenies; Limitations of phylogenetic algorithms; Phylogenetic softwares – PAUP, PHYLIP, MacClade.

**Recommended Texts:**

1. Mount, D. (2004) "Bioinformatics: Sequence and Genome Analysis"; Cold Spring Harbor Laboratory Press, New York.
2. Baxevanis, A.D. and Francis Ouellette, B.F. (1998) "Bioinformatics – a practical guide to the analysis of Genes and Proteins"; John Wiley & Sons, UK.

**Reference Books**

1. Pevzner, P.A. (2004) "Computational Molecular Biology"; Prentice Hall of India Ltd, New Delhi.
2. Pevsner, J. (2003) "Bioinformatics and Functional Genomics"; John Wiley and Sons, New Jersey, USA.
3. Lesk, A.M. (2002) "Introduction to Bioinformatics", First edition, Oxford University Press, UK.
4. Sensen, C.W. (2002) "Essentials of Genomics and Bioinformatics"; Wiley-VCH Publishers, USA.

## **MODULE – II : BIO-ALGORITHMS AND SEQUENCE ANALYSIS**

### **PAPER - III - ALGORITHMS IN BIOINFORMATICS (2 CREDITS)**

#### **1. Exercises in Matrices Implementation**

- a. Addition
- b. Subtraction
- c. Multiplilcation

#### **2. Exercises in Sorting**

- a. Bubble Sort
- b. Insertion Sort
- c. Quick Sort

#### **3. Exercises in Searching**

- a. Binary Search

#### **3. Exercises in String Matching**

- a. Naïve Algorithm

#### **4. Exercises in Graph**

- a. Single Source Shortest path algorithm
- b. Dynammic Programming
- c. Backtracking Algorithms

## **MODULE – II : BIO-ALGORITHMS AND SEQUENCE ANALYSIS**

### **PAPER - IV – LAB – SEQUENCE AND PHYLOGENETIC ANALYSIS (2 CREDITS)**

#### **Exercises:**

1. Sequence Analysis Packages – EMBOSS, NCBI ToolKit
2. Dynamic programming.
3. Analysis of Biological Sequences.
4. FASTA
5. Multiple sequence alignment
6. MEME/MAST, eMotif, InterproScan, ProSite, ProDom, Pfam
7. Phylogenetic analysis – PAUP, PHYLIP, MacClade
8. Genome annotation – Artemis.
9. Hypothetical Protein analysis
10. Genome Comparison

## **MODULE – III : STRUCTURAL BIOINFORMATICS**

### **PAPER - I – BIOPHYSICS AND STRUCTURAL BIOLOGY (3 CREDITS)**

#### **Unit-I**

Structural features of biomolecules; techniques used to determine the structure of biomolecules; Methods for single crystal X-ray Diffraction of macromolecules: molecular replacement method and direct method – Fiber diffraction; analysis of structures and correctness of structures; submission of data to PDB: atomic coordinates and electron density maps.

#### **Unit-II**

Anatomy of proteins; Ramachandran Plot; secondary structures; motifs; domains; tertiary and quaternary structures. Anatomy of DNA; A, B, Z-DNA, DNA bending. Structure of RNA. Structure of Ribosome.

#### **Unit-III**

Methods for prediction of secondary and tertiary structures of proteins – knowledge-based structure prediction; fold recognition; *ab initio* methods for structure prediction, Comparative protein modeling.

#### **Unit-IV**

Methods for comparison of 3D structures of proteins; Methods to predict three dimensional structures of nucleic acids, rRNA; Electrostatic energy surface generation.

#### **Unit-V**

Molecular Mechanics and Molecular dynamics of Oligopeptides, Proteins, Nucleotides and small molecules – Mechanism and dynamics of bio-macromolecules, Simulation of molecular mechanics and dynamics, Simulations of Free energy changes; Force fields. Molecular interactions of protein-protein, protein-DNA, protein-carbohydrate and DNA-small molecules.

**Recommended Texts:**

1. Andrew R. Leach (2001) “Molecular Modeling – Principles and Applications”; Second Edition, Prentice Hall, USA.
2. Creighton, T.E. (1993) “Proteins: structure and molecular properties”; Second edition, W.H. Freeman and Company, New York, USA.

**Reference Books**

1. Mount, D. (2004) “Bioinformatics: Sequence and Genome Analysis”; Cold Spring Harbor Laboratory Press, New York.
2. Lesk, A.M. (2001) “Introduction to Protein Architecture”, Oxford University Press, UK.
3. Mcpherson, A. (2003) “Introduction of Molecular Crystallography”, John Wiley Publications, USA.

## MODULE – III : STRUCTURAL BIOINFORMATICS

### PAPER - II – MOLECULAR MODELING AND DRUG DESIGN (3 CREDITS)

#### Unit-I

**Concepts in Molecular Modeling** – Introduction; Coordinate System; potential energy surfaces molecular graphics; Computer hardware and software; Mathematical concepts – introduction of molecular mechanics & quantum mechanics.

#### Unit-II

**Molecular Mechanics** – Features of molecular mechanics, force fields; Bond structure and bending angles – electrostatic, van der Waals and non-bonded interactions, hydrogen bonding in molecular mechanics; Derivatives of molecular mechanics energy function; Calculating thermodynamic properties using force field; Transferability of force field parameters, treatment of delocalised  $\pi$  system; **Force field for metals and inorganic systems** – Application of energy minimization.

#### Unit-III

**Molecular Dynamics Simulation Methods** – Molecular Dynamics using simple models; Molecular Dynamics with continuous potentials and at constant temperature and pressure; Time-dependent properties; Solvent effects in Molecular Dynamics; Conformational changes from Molecular Dynamics simulation.

#### Unit-IV

**Molecular Modeling in Drug Discovery** – Deriving and using 3D pharmacophore; Molecular Docking; Structure-based methods to identify lead compounds; *de novo* ligand design; Applications of 3D Database Searching and Docking

#### Unit-V

**Structure Activity Relationship** - QSARs and QSPRs, QSAR Methodology, Various Descriptors used in QSARs: Electronic; Topology; Quantum Chemical based Descriptors. Use of Genetic Algorithms, Neural Networks and Principle Components Analysis in the QSAR equations.

### **Recommended Texts:**

1. Andrew R. Leach (2001) "Molecular Modeling – Principles and Applications"; Second Edition, Prentice Hall, USA.

### **Reference Books**

1. Fenniri, H. (2000) "Combinatorial Chemistry – A practical approach", Oxford University Press, UK.
2. Lednicer, D. (1998) "Strategies for Organic Drug Discovery Synthesis and Design"; Wiley International Publishers.
3. Gordon, E.M. and Kerwin, J.F. (1998) "Combinatorial chemistry and molecular diversity in drug discovery"; Wiley-Liss Publishers.
4. Swatz, M.E. (2000) "Analytical techniques in Combinatorial Chemistry"; Marcel Dekker Publishers.

## MODULE – III : STRUCTURAL BIOINFORMATICS

### PAPER - III – LAB – SIMULATIONS IN MACROMOLECULES AND MOLECULAR INTERACTIONS (2 CREDITS)

#### Exercises

1. Advanced Visualization Software and 3D representations.
2. Coordinate generations and inter-conversions.
3. Secondary Structure Prediction
4. Fold Recognition, *ab initio* (Rosetta Server)
5. Homology based comparative protein modeling.
6. Energy minimizations.
7. Validation of models.
  - a. WHATIF
  - b. PROSA
  - c. PROCHECK
  - d. VERIFY 3D
8. Protein Structure Alignment.
9. Modeller
10. Geno-3D
11. Discovery Studio Server.

**MODULE – III : STRUCTURAL BIOINFORMATICS**  
**PAPER - IV – LAB – DOCKING AND QSAR ANALYSIS**  
**(2 CREDITS)**

**Exercises**

1. Binding Site Identification.
2. Pharmacophore Identification
3. Rigid body Docking using AutoDock and ADT
4. Molecular dynamics simulations using Gromacs
5. Visual molecular Dynamics (VMD)
6. Docking with LigandFit (Discovery studio)
7. Receptor and Ligand Optimization.
8. Conformational Analysis
9. BABEL, MOPAC
10. QSAR and QSPR analysis

## MODULE – IV : APPLICATIONS OF BIOINFORMATICS

### PAPER - I – SYSTEMS BIOLOGY AND PROFILE ANALYSIS (3 CREDITS)

#### Unit-I

**Systems Biology** - Objectives of Systems Biology, Strategies relating to *In silico* Modeling of biological processes, Metabolic Networks, Signal Transduction Pathways, Gene Expression Patterns. E-cell and V-cell Simulations and Applications.

#### Unit-II

**Reconstruction of pathways and annotation** – Reconstructing metabolic pathways from sequence and function information in microbial species; statistical profiling and function annotation of genomes with a microbial genome as an example.

#### Unit-III

**Profile analysis** – Expression profile analysis of cells, Mining data from Yeast. Microarray and genome wide expression analysis: transcriptomes, proteome: Genomics in medicine, disease monitoring, profile for therapeutic molecular targeting.

#### Unit-IV

**Drug Designing Related Applications** – Finding new drug targets to treat diseases – Pharmacophore identification - Structure based drug design – Molecular Simulations.

#### Unit-V

**Commercial Bioinformatics** - Definition of Bioinformatics company. Genome Technology: high throughput sequencing and assembly. Diagnostic drug discovery and genomics. Pharmacogenomics and its application. SNPs and their applications. Proteomics in medicine, Toxicology.

## **Recommended Texts**

1. Hunt, S.P. and Livesey, F.J. (2000) "Functional Genomics – a practical approach", Oxford University Press, UK.
2. Wilkins, M.R., Williams, K.L., Appel, R.D. and Hochstrasser, D.F. (1997) "Proteome Research: New frontiers in Functional Genomics", Springer Verlag, New York, USA.
3. Witten, I.H. and Frank, E. (2005) "Data mining: Practical Machine Learning Tools and Techniques", Morgan Kauffman Publishers, USA.

## **MODULE – IV : APPLICATIONS OF BIOINFORMATICS**

### **PAPER - II – LAB– PATHWAY RECONSTRUCTION AND E-MODELING (2 CREDITS)**

#### **Exercises:**

1. Generating interaction networks for set of gene and proteins using on line tools
2. Usage of Cytoscape
3. Generating networks for the interaction of systems of host and pathogens
4. Constructing an integrated network for a sample human disease and its analysis

**MODULE – IV : APPLICATIONS OF BIOINFORMATICS**

**PAPER – III – PROJECT  
(5 CREDITS)**

