**M.Sc. in Big Data Biology (CBCS)**
**(Effective from the academic year 2023-2024)**
**SCHEME OF INSTRUCTION AND EXAMINATION**
**SEMESTER SCHEME**

| Course Code | Course Title | Nature of Course | Periods / Week | Duration of Exam (Hours) | IA | EA | Maximum Marks | Credits |
|---|---|---|---|---|---|---|---|---|
| **I Semester** | **Theory** | | | | | | | |
| BDBH-101 | Computer Programming and Operating Systems – Linux and Python | H Core | 4 | 3 | 30 | 70 | 100 | 4 |
| BDBH-102 | **Data Structures and Algorithms** | H Core | 4 | 3 | 30 | 70 | 100 | 4 |
| BDBH-103 | **Mathematics for Data Science** | H Core | 4 | 3 | 30 | 70 | 100 | 4 |
| BDBS-104 | **Cell and Molecular Biology & Human Genetics** | S Core | 2 | 2 | 15 | 35 | 50 | 2 |
| | **Laboratory** | | | | | | | |
| BDBP-105 | **Cell and Molecular Biology & Human Genetics – Laboratory** | Pract | 8 | 4 | 30 | 70 | 100 | 4 |
| BDBP-106 | Computer Programming and Operating System – Linux and Python- Laboratory | Pract | 8 | 4 | 30 | 70 | 100 | 4 |
| **Total Marks and Credits** | | | | | | | **600** | **22** |
| | | | | | | | | |
| **II Semester** | **Theory** | | | | | | | |
| BDBH-201 | **Biostatistics and Probability** | H Core | 4 | 3 | 30 | 70 | 100 | 4 |
| BDBH-202 | **Bioinformatics** | H Core | 4 | 3 | 30 | 70 | 100 | 4 |
| BDBH-203 | Algorithms for Computational Biology | H Core | 4 | 3 | 30 | 70 | 100 | 4 |
| BDBH-204 | **Machine Learning** | H Core | 4 | 3 | 30 | 70 | 100 | 4 |
| BDBS-205 | **Database Design and Cloud Computing** | S Core | 2 | 2 | 15 | 35 | 50 | 2 |
| BDBS-206 | Microbiology and Immunology | S Core | 2 | 2 | 15 | 35 | 50 | 2 |
| | **Laboratory** | | | | | | | |
| BDBP-207 | Machine Learning- Laboratory | Pract | 8 | 4 | 30 | 70 | 100 | 4 |
| BDBP-208 | **Database Design and Cloud Computing - Laboratory** | Pract | 4 | 2 | 15 | 35 | 50 | 2 |
| BDBP-209 | Biostatistics and 'R' Programming – Laboratory | Pract | 4 | 2 | 15 | 35 | 50 | 2 |
| **Total Marks and Credits** | | | | | | | **700** | **28** |
| | | | | | | | | |
| **III Semester** | **Theory** | | | | | | | |
| BDBH-301 | **Next Generation Sequencing and Genomic Data Analysis** | **H Core** | **4** | **3** | **30** | **70** | **100** | **4** |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **(bulk and single cell omics)** | | | | | | | |
| BDBH-302 | **Biomedical Imaging and Informatics** | H Core | 4 | 3 | 30 | 70 | 100 | 4 |
| BDBH-303 | **Deep Learning and Applications** | H Core | 4 | 3 | 30 | 70 | 100 | 4 |
| BDBE-304 | **(A)Computational Drug Design and AI applications/ (B)Systems Biology and Network Modelling** | Elective | 2 | 2 | 15 | 35 | 50 | 2 |
| BDBS-305 | **Clinical Trial Data Analysis** | S Core | 2 | 2 | 15 | 35 | 50 | 2 |
| BDBO-306 | Optional Elective: (A)Scientific Writing and Communication (B)Physical and Mental Health | OE | 2 | 2 | 15 | 35 | 50 | 2 |
| | **Laboratory** | | | | | | | |
| BDBP-307 | **Data Analysis in Genomics, Transcriptomics and Proteomics - Laboratory** | Pract | 8 | 4 | 30 | 70 | 100 | 4 |
| BDBP-308 | **Image Informatics Laboratory** | Pract | 8 | 4 | 30 | 70 | 100 | 4 |
| BDBP-309 | **Deep Learning Laboratory** | Pract | 4 | 2 | 30 | 70 | 100 | 4 |
| **Total Marks and Credits** | | | | | | | **750** | **30** |
| | | | | | | | | |
| **IV Semester** | **Theory** | | | | | | | |
| BDBP-401 | Capstone Project/Dissertation | Pract | 36 | | | | 200 | 18 |
| | Project Viva | | 4 | | | | 50 | 2 |
| **Total Marks and Credits** | | | | | | | **250** | **20** |
| **Grand Total Marks and Credits** | | | | | | | **2250** | **100** |

**SEMESTER I**

| | Course title | L-T-P-C | | |
|---|---|---|---|---|
| 1 | Computer Programming and Operating Systems – Linux and | 3-1-0-4 | H | 4 |
| 2 | Data Structures and Algorithms | 3-1-0-4 | H | 4 |
| 3 | Mathematics for Data Science | 3-1-0-4 | H | 4 |
| 4 | Cell and Molecular Biology & Human Genetics | 2-0-0-2 | S | 2 |
| 5 | Cell and Molecular Biology & Human Genetics – Laboratory | 0-0-4-4 | P | 4 |
| 6 | Computer Programming and Operating System – Linux and | 0-0-4-4 | P | 4 |
| | **TOTAL** | | | **22** |

## 1. Computer Programming and Operating Systems – Linux and Python H4

**Unit 1: Linux and Shell Programming (15 hours)**

**Linux:** Basic Linux Operating System - concepts - key commands useful for developers - Introduction to Operating Systems; History and features of UNIX and GNU/Linux. Unix file system, file and directory commands, file permissions. Basic commands, I/O redirection and piping, simple and advanced filters, vi as text editor - Archives and file compressions - Processes: background processes and scheduled processes - alias and environmental variables

**BasH Shell Programming:** Multiple commands as a shell script- simple shell script creation and execution. Variables: System variables and User defined variables, read values to variables, Mathematic and String handling. Decisions and loopings: if, for and while loops, case statement; awk programming, terminal formatting using echo and tput. Functions: calling functions; passing arguments; receiving parameters; local variables; returning values from functions; unsetting functions. Signals: handling signals; ignoring signals.

**Unit 2: Python Programming (30 hours)**

**Introduction to Python Programming:** Memory concepts-arithmetic, string formatting. Decision making: equality and relational operators-control structures-functions, basic sorting searching with data structures, Algorithm analysis tutorial

**Python exception hierarchy:** finally clause-exception objects and trace backs-programmer-defined exception classes

**String Processing and Regex:** Fundamentals of characters and strings-string presentation-searching strings-joining and splitting strings

Regular expressions-compiling regular expressions and manipulating regular expression objects, regular expression repetition and placement characters, classes and special sequences, regular expression-string manipulation functions-grouping – biological application assignments exploiting regular expressions

**File Processing:** File processing and serialization: Introduction, data hierarchy, files and streams, creating a sequential access file-reading data from a sequential access file, updating sequential access files

**Unit 3: Python Data Analysis and Panda (15 hours)**

Python data analysis - (statistical methods - Numpy and Scipy) key statistical methods like correlation regression, etc.to be covered along with plotting the results using matplotlib.

Python Panda library - spread sheet processing – using spreadsheets having multi parameter biological data, ingestion and statistical analysis will be demonstrated

**PythonDB- application development:** Python DB-API specification-database query example-querying the database-reading, inserting and updating a database – Bio Python

**Textbooks and Reference Books**


1. Richard Blum, Christine Bresnahan, Linux Command line and Shell Scripting Bible, Wiley, 2015, 3e
2. Tom Christiansen, Jon Orwant, Larry Wall, Brian Foy, Programming Perl, O'Reilly Media, 2012, 4e
3. Alex Martelli, Anna Ravenscroft, Steve Holden, Python in a Nutshell - A Desktop Quick Reference, O'Reilly Media, 2017, 3e
4. Bill Lubanovic, Introducing Python: Modern Computing in Simple Packages, O'Reily Media, 2014,1e
5. Martin Jones, Python for Biologists, Create Space Independent Edition, 2013,1e
6. William Mckinney, Python for Data Analysis: Data Wrangling Pandas, NumPy and IPython, O'Reilly Media, 2017, 2e
7. Robert Love, Linux System Programming, O-ReillyMedia,2013,2e


## 2. Data Structures and Algorithms H4

### Unit 1: Data Types and structures ( 8 hours)

Elementary Data types, Data structure and Abstract date types (ADT) – specification and implementation - Dynamic memory allocation- Asymptotic notations and common functions

### Unit 2: Stacks and Queues ( 10 hours)

Basic Concept of Stack - Stack as an ADT, Stack Operations, Stack Applications Conversion from infix to postfix/prefix expression, Evaluation of postfix/ prefix expressions -
Basic Concept of Queue, Queue as an ADT, Primitive Operations in Queue - Linear Queue, Circular Queue, Priority Queue, Queue Applications

### Unit 3: Linked List ( 8 hours)

List and ADT, Array Implementation of Lists, Linked List -Types of Linked List: Singly Linked List, Doubly Linked List, Circular Linked List -Basic operations in Linked List: Node Creation, Node Insertion and Deletion from Beginning, End and Specified Position -Stack and Queue as Linked List

### Unit 4 : Sorting, Searching and Hashing ( 12 hours)

Types of sorting: Internal and External sort - Comparison Sorting Algorithms: Bubble, Selection and Insertion Sort, Shell Sort -Divide and Conquer Sorting: Merge, Quick and Heap Sort - Efficiency of Sorting Algorithms -Search Algorithms – Sequential and binary search – efficiency of search algorithms – hashing- hash function and hash tables, collision resolution

### Unit 5 : Trees and Graphs ( 10 hours)

Basic Operations in Binary Tree, Tree Height, Level and Depth -Binary Search Tree, Insertion, Deletion, Traversals, Search in BST - Applications of Trees Definition and Representation of Graphs, Graph Traversal

### Unit 6 : Algorithms ( 12 hours)

Algorithms: Insertion sort - analysing algorithms, growth of functions – recurrences -heapsort, quicksort, medians and order statistics, dynamic programming, greedy algorithms, Graph algorithms, Caching

**Text Books and References**


Bradley N. Miller and David L. Ranum, Problem-Solving with Algorithms and Data structures in

Python", Franklin Beedle & Associates,2013,2e
Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Introduction to Algorithms, 2009, The MIT Press,3e.

## 3. Mathematics for Data Science H4

**Unit 1: Analytic Geometry (10 hours)**

Cartesian coordinate system in two, three and higher dimensions; Right handed and left handed coordinates; Polar and spherical polar coordinates; direction cosines; distance formula; Translation and rotations in two and three dimensional systems; Euler angles; Extension to higher dimensional systems

Equations of straight lines in two and three dimensions; direction cosines; Orthogonal and parallel lines; condition for orthogonality; Equation of plane in three dimensions; parallel and intersecting planes

**Unit 2: Vector Algebra (10 hours)**

Vectors in two, three and higher dimensions; magnitude and direction of vectors; direction cosines; resultant of vectors and parallelogram law; scalar product of vectors; vector product; projection vector; orthogonality of vectors; scalar triple product and determinant; vector triple product; translation and rotation of vectors in two and three dimensions; Vectors in higher dimension

**Unit 3: Simultaneous Equations (8 hours)**

Systems of simultaneous linear algebraic equations in two and three dimensions; Homogeneous and inhomogeneous equations; Geometrical interpretations of their solutions in terms of intersection of straight lines and planes; Conditions for the existence of solutions; Cramer's rule

**Unit 4: Matrices and Determinants (12 hours)**

Definition of matrix; representing matrix elements in index form; addition, subtraction and multiplication of matrices; outer and inner product of two matrices; Diagonalization of a matrix; matrix inverse; rank of a matrix; transpose of a matrix; writing a set of simultaneous equations in matrix form; representing vectors as column matrices; vector translation as matrix addition; vector rotation as matrix multiplication; eigenvectors and eigenvalues of matrices and their determination; properties of eigenvalues and eigenvectors; conditions for a matrix to have real eigenvalues and orthogonal eigenvectors; solutions of linear algebraic equations in matrix form; scalar triple product of vectors as determinants; properties of determinants; Singular Value Decomposition (SVD) of a matrix

**Unit 5: Calculus (20 hours)**

Continuity of a function at a point, Derivative of a function, differentiation, second and higher order derivatives, partial derivatives; The tangent and normal lines to a curve at a point; Finding extreme values of functions, local and global extrema; Linear approximations to functions at a point; Differentials, computing absolute, relative and percentage changes in formulas. Indefinite integrals, integration methods. Summation of series, Riemann sums and definite integral, evaluation of definite integrals. Area under the curve, average values of continuous functions. First order differential equations and their solutions, variable-separable method, general form and solution of first order linear differential equations, use of differential equations in biology with examples.

**Textbooks and Reference Books**

1. G. B. Thomas, J. R. Hass and M. D. Weir; Thomas' Calculus; Pearson 2010, 12th edition.
2. M. R. Spiegel, S. Lipschutz and D. Spellman; Schaum's Outline: Vector Analysis; McGraw Hill 2009, 2nd edition.
3. R. A. Horn and C. R. Johnson; Matrix Analysis; Cambridge 2016, 2nd edition.

4. G. B. Arfken and H. J. Weber; Mathematical Methods for Physicist; Elsevier 2011, 6th edition.

## 4. Cell and Molecular Biology & Human Genetics S2

**Unit 1: Basic structure and function of cells and their organelles (6 hours)**
Differences in the basic structure and composition of prokaryotic cells and eukaryotic animal and plant cells; structure and function of eukaryotic (plant and animal) cell organelles.

**Cell cycle and its relevance to cancer**
Introduction to cell cycle, cell division and types: mitosis and meiosis; typical phases of mitosis and meiosis; brief discussions on the molecular mechanisms, relationship with growth, differentiation and reproduction. Hall marks of Cancer, understanding of proto-oncogenes and tumour suppressors.

**Unit 2: The Central Dogma of biology, DNA replication and repair (6 hours)**
DNA as a universal genetic material in cells across species and the flow of genetic information from DNA to RNA to proteins - Types of replication. Characteristics and functions of prokaryotic and eukaryotic DNA polymerases
Mechanism of prokaryotic and eukaryotic DNA replications: replication bubbles, replication forks, the role of topoisomerases, SSBs and helicases. Mechanism and significance of telomerases. Types of DNA damage and repair mechanisms. Late and early replicating regions of the genome.

**RNA synthesis and processing**
Gene organization: promoters, exons and introns -Mechanism of transcription. Molecular control of transcription initiation: transcription factors, core promoters and RNA polymerases; Types of (rRNA, tRNA, mRNA coding) genes and their promoter features post-transcriptional processing; 5'-capping and poly-adenylation of transcripts; types of alternative splicing and significance of splicing in cell functions. An overview of splicing mechanism.

**Unit 3: Prokaryotic and Eukaryotic Gene expression regulation (5 hours)**
Prokaryotic operon models: inducible and repressible systems, lac and trp operon Eukaryotic core and distal promoters, cis elements, enhancers and insulators. Transcription factors and their binding sites. Relationship of gene-expression-control with cellular and physiological events; role of hormones and signal transduction systems

**Unit 4: Physical basis of heredity, Cytogenetics and Molecular Genetics (10 hours)**
Introduction, concepts and theories of Mendelian genetics, chromosome theory of inheritance, Non-Mendelian Inheritance, gene interaction, linkage and crossing over, two point and three point mapping, linkage map. Chromosomestructure, organization and classification, normal karyotype, chromosomal abnormalities, idiogram, Robertsonian translocation, imprinting, mutations and repair. DNA fingerprinting, DNA markers, types of repeats. Identification of the disease allele

**Unit 5: Clinical and Cancer Genetics (5 hours)**
Deciphering pathways for metabolic disorders, signatures in the genome used for generating haplotypes, single-gene inheritance, pedigree analysis, Human Genome Project, identification of the gene involved in rare monogenic disorders, mapping genomes. Familial cancers and sporadic cancer. Drivers of Cancer, Copy number variations in Cancer.

**Textbooks and Reference books:**
1. B. Alberts et. al.; Molecular biology of the cell; Taylor & Francis Publishers, 2008; 5th edition.
2. H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore and J. Darnell; Molecular Cell Biology; W. H. Freeman & Comp., 2007; 6th edition
3. G. M. Cooper and R. E. Hausman; The cell: A molecular approach; ASM Press, 2009; 5th edition.

4. B. Alberts et. al.; Molecular biology of the cell; Taylor & Francis Publishers, 2014.; 6th edition.
5. T. A. Brown; Genomes 3; Oxford: Wiley-Liss, 2007; 3rd edition
6. J. E. Krebs, E. S. Goldstein and S. T. Kilpatrick, Lewin's Genes XII; Jones & Bartlett Publishers, Inc., 2017; 12th edition.
7. D. L. Hartl and M. Ruvolo; Genetics: Analysis of Genes and Genomes, Jones & Bartlett Learning, 8th edition, 2011.
8. M. J. Simmons, D. P. Snustad and E. J. Gardner; Principles of Genetics, John Wiley and Sons Inc, 8th edition; 1991.
9. T. A. Brown; Introduction to Genetics: A Molecular Approach, Garland Science, 2011.
10. E. S. Tobias, M. Connor, M. Ferguson-Smith; Essential Medical Genetics, Wiley Blackwell, 6th edition, 2001
11. D. L Hartl and E. W. Jones; Genetics: Analysis of genes and genomes, Jones and Bartlett Publishers; 4th edition; 1998.
12. S Moody; Principles of Developmental Genetics; Academic Press, 2nd edition; 2014.
13. F Bunz; Principles of Cancer Genetics; Springer; 2008.

## 5. Cell and Molecular Biology & Human Genetics – Laboratory P4

1. Experiments in accuracy and precision (pipetting weighing balance, pH meter)
2. Cleanliness, sterilization, and safety (autoclaving, reagent, and media preparation)
3. Estimation of protein concentration using Lowry and Bradford methods
4. Isolation, estimation and analysis of nucleic acids by agarose gel electrophoresis, restriction digestion and visualization
5. Separation of proteins by SDS-PAGE, followed by coomassie and silver staining
6. Karyotyping and G and C banding of chromosomes using normal and cancerous cell lines of humans.
7. Mammalian cell culturing and viability testing by MTT, LDH and trypan blue exclusion assay. Transfection of mammalian cells. Splitting and cryopreservation of mammalian cells. Flow cytometry (Demo)
8. Gradient PCR, Realtime PCR, AFLP
9. RNA isolation and agarose and urea-PAGE
10. Whole Genome and RNA library preparation, estimation of DNA using Qubit and profiling by tape station

**Textbooks and Reference books:**

1. D. R. Randall; Molecular Biology Laboratory Manual; Available online (available free of cost at https://archive.org/details/MolecularBiologyLaboratoryManual_456).
2. S. Surzycki; Human Molecular Biology Laboratory Manual; Wiley-Blackwell, 2003.
3. K. V. Chaitanya; Cell and Molecular Biology: A Lab Manual; PHI Learning Press, 2013.
4. Practicing Safe Science video https://www.youtube.com/watch?v=sTzW6XENAXQ
5. Cell Fractionation and Organelle Isolation https://www.thermofisher.com/us/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/cell-fractionation-organelle-isolation.html
6. http://www.biology.arizona.edu/cell_bio/cell_bio.html
7. R. I. Freshney; Culture of Animal Cells: A Manual of Basic Technique; 4th Ed. Wiley Liss; New York, 2000.
8. J. M. Davis; Basic Cell Culture; Oxford University Press, New York, 93–134, 1996.

## 6. Computer Programming and Operating System – Linux and Python- Laboratory P4

**Python:**

1. Environment and environment setup (python setup with correct parameterization).
2. Basic syntax; variables; operators (basic python programming)
3. Decision making; loops (selection and iterations in python programming)
4. Numbers, lists, strings, tuples, dictionary (advanced data structures in python programming).
5. Functions in python; modules in python (function and module orientation in python programming)
6. Files and file operations in python (data storage and retrieval using files in python).
7. Classes and objects (object-oriented programming in python)
8. Regular expressions; CGI programming; multi threading; exception handling; XML processing; GUI programming (advanced python programming)

**Linux:**

9. Shell: Basic commands of a shell, internal vs external commands (working on multiple shell commands and understanding of how it interacts with Linux)
10. Shell programming (working on core shell programming in sync with Linux)
11. Linux file management: file and file system, file system related system calls (programming exercises on core Linux file system and file system related calls)
12. Linux process management: creation of a process-to-process system management (assignments on core Linux process management, life cycle, context and state, /proc system)
13. Make utility and make file (understanding the practical importance of make utility)

**Textbooks and Reference books:**


1. Martelli, A. Ravenscroft and S. Holden; Python in a Nutshell; O'Reilly Publication, 3rd edition, 2017.
2. D. Beazley and B. K. Jones; Python Cookbook – Recipes for Mastering Python; O'Reilly Publication, 3rd edition, 2013.
3. J. Jose and S P. Lal; Introduction to Computing and Problem Solving with Python; Khanna Publishers, 1st edition, 2015.
4. B. W. Kernighan and D. Ritchie; The C Programming Language; Pearson Education India, 2nd edition, 2015.
5. Balagurusamy; Programming in ANSI C; McGraw Hill Education India Private Ltd, 7th edition, 2017.
6. H. Schildt; C – The Complete Reference; McGraw Hill, 4th edition, 2017.
7. E. Siever; Linux in a Nutshell; O'Reilly Publication, 6th edition, 2009.
8. L. Robert; Linux System Programming; Shroff Publishers and Distributors Private Ltd, 2nd revised edition, 2014.
9. M. J. Bach; The Design of the UNIX Operating System; Pearson Education India, 1st edition, 2015.

**SEMESTER II**

| | Course title | L-T-P-C | | |
|---|---|---|---|---|
| 1. | **Biostatistics and Probability** | 3-1-0-4 | H | 4 |
| 2. | **Bioinformatics** | 3-1-0-4 | H | 4 |
| 3. | Algorithms for Computational Biology | 3-0-1-4 | H | 4 |
| 4. | **Machine Learning** | 3-0-1-4 | H | 4 |
| 5. | **Database Design and Cloud Computing** | 2-0-0-2 | S | 2 |
| 6. | Microbiology and Immunology | 2-0-0-2 | S | 2 |
| 7. | Machine Learning- Laboratory | 0-0-4-4 | P | 4 |
| 8. | **Database Design and Cloud Computing - Laboratory** | 0-0-2-2 | P | 2 |
| 9. | Biostatistics and 'R' Programming – Laboratory | 0-0-2-4 | P | 2 |
| | **TOTAL** | | | **28** |

## 1. Biostatistics and Probability H4

**Unit 1: Statistical Parameters (6 hours)**

Data types, Concept of population and random samples; Statistical parameters - parameters of centrality: Arithmetic, geometric and harmonic means, median; parameters of location – percentiles and quartiles; parameters of spread: variance and standard deviation; parameters of shape – skewness and kurtosis; Box and Whisker plots

**Unit 2: Probability Distributions (8 hours)**

Concept of discrete and continuous distributions; Binomial, Poisson, hypergeometric and negative binomial distributions and their applications; Gaussian (normal) probability density distribution and its properties; Unit Gaussian distribution and its application

**Unit 3: Sampling Distributions and Error Analysis (12 hours)**

Sampling distribution of mean and standard error; The central limit theorem; t-distribution; Confidence interval for a population mean; Independent and dependent samples; Distribution of difference between two sample means and the confidence interval; distribution of sample proportion and the confidence interval; distribution of difference between two sample proportions and the confidence interval; F-distribution; distribution of a sample variance and the confidence interval; distribution of ration of two sample variances and the confidence interval;

**Error Analysis**

Understanding the terms error, uncertainty, accuracy and precision; Derivations of error propagation formulae mathematical operations like addition, subtraction, multiplication, division and exponentiation of variables; representation of standard error and Confidence intervals as error bars in plots and their interpretation; bar plots with errors;

**Unit 4: Hypothesis Testing (18 hours)**

Introduction to hypothesis testing; Understanding null and alternate hypothesis; Level of statistical significance;

Parametric Tests : One sample Z-test and t-test on a population mean; Two sample Z-test on the difference between two population means; Two sample independent and dependent t-tests for the difference between two population means; Welsch t-test; hypothesis testing a sample proportion; hypothesis testing the difference between two sample proportions; hypothesis testing a population variance; hypothesis testing the difference between two population variances; Analysis of Variance

(ANOVA) – one factor and two factor ANOVA.
Nonparametric Tests: Wilcoxon signed rank test for one variable and two paired variables; Wilcoxon rank-sum test for unpaired variables; Mann-Whitne test; Kruskal Wallis test; Chi-square goodness of fit test; Statistical tests for contingency tables – Fisher's exact test, Chi-square test.

Statistical Inference: Type-I, Type-II errors and the power of statistical tests; p-value correction and the False Discovery Rate (FDR) for multiple comparisons; Sample size estimations.

**Unit 5: Regression Analysis and Data Reduction Methods (16 hours)**
Covariance and correlation; Pearson's correlation coefficient; Linear regression theory – Least square fit to a straight line; Non-linear regression; Multiple linear Regression; Logistic Regression.

**Data reduction methods**
Principal Component Analysis, Markov chains and Hidden Markov Models.Clustering algorithms. Demonstration of these methods with bioinformatics data.

**Text Books**

1. Wayne W. Daniel, Chad L. Cross, Biostatistics – Basic concepts and methodologies for Health Sciences, Wiley Students Edition, 2016,10e
2. Robert V.Hogg, Tanis, Zimmermann, Probability and Statistical Inference, Pearson Education Asia, 2015,9e

## 2. Bioinformatics H4

**Unit 1. Introduction to Bioinformatics (10 hours)**
Data access and analysis for biological research; Biological databases; Comparisons of formats, volumes and utilities of data across databases such as NCBI, GEO, Array Express, SRA, UCSC, Uniprot and HPRD. Bioinformatic tools for solving biological problems.

**Unit 2. Sequence Analysis (15 hours)**
DNA, RNA and protein sequences, genome and transcriptome, open reading frames, gene structure in prokaryotes and eukaryotes, coding and non-coding genes, extracting, collecting and storing sequences; File formats for bio-molecular sequences.
Biological sequence similarity, identity and homology. Dot plots. Definitions of homologues, orthologues, paralogues, concepts behind scoring matrices, database searching. Derivation of PAM and BLOSUM matrices, evolution model for nucleic acids. Global and local pairwise alignment methods – Smith-Waterman and Needleman-Wunsch algorithms. Concepts behind multiple sequence alignment. Using online and offline BLAST tool and its different versions to assign homology.

**Unit 3: Motifs and phylogeny (15 hours)**
Basic concept and definition of sequence patterns, motifs and profiles, various types of pattern representations viz. consensus, regular expression (prosite-type) and profiles. Use of Hidden Markov model (HMM) in assigning homology. Phylogeny: sequence evolution, distance matrices, phylogeny construction by UPGMA, neighbour joining and parsimony methods.

**Unit 4. Structural Bioinformatics (20 hours)**
Molecular structures – visualizing and graphical representations. Calculation of geometric parameters (bond distance, bond angle, dihedral angle). Identifying intramolecular and inter molecular interactions from crystal structures (using GUI).

Protein secondary structure calculation – DSSP, membrane topology prediction, ligand-receptor

interactions, composition of active sites in functional proteins, conformation change and activity, allostery, effects of point mutations on protein structure and function

**Textbooks:**

1. N. C Jones, Pavel and Pevzner; Introduction to Bioinformatics Algorithms; The MIT Press, 2004
2. P. G. Higgs and T. K Attwood; Bioinformatics and Molecular Evolution; Blackwell Publishing, 2005

**Reference books:**

1. A. M. Lesk; Introduction to Bioinformatics; Oxford University Press, 2002

2. D. W. Mount; Bioinformatics Sequence and Genome Analysis; Cold Spring Laboratory Press,2001

## 3. Algorithms for Computational Biology H4

**Unit 1: Strings and Graphs (12 hours)**

Strings - naive string search; suffix tree; KMP Algorithms; Boyer-Moore Algorithms; graphs- types of graphs; interval graphs; hybridization, partial digest and optical mapping, consecutive ones - property and associated algorithms.

**Unit 2: Sequence Analysis Algorithms (16 hours)**

Sequence alignment – Dynamic Programming - alignment algorithms for two sequences; Gapped BLAST and PSI BLAST; PHI-BLAST; Matrices-PAM and BLOSUM - Sequence alignment algorithms- Global and local alignment - Advanced alignment techniques: linear space, affine gaps, time warping

**Unit 3: Stochastic & Evolutionary Models (16 hours)**

Markov Chains, Hidden Markov models - Baum-Welch and Viterbi algorithms; applications of HMMs (HMMER, PFAM).

Evolutionary models - Jukes-Cantor and Kimura models; Phylogenetic trees – building a tree from pairwise distances, neighbour-joining, parsimony distance base method and character based methods - algorithms with problems

**Unit 4: Algorithms for understanding transcriptional regulation (16 hours)**

Transcriptional regulation, Transcription factor binding sites, position weight matrices, sequence logos, motif- finding via expectation maximization (MEME) and Gibbs sampling

**Textbooks and Reference books:**

1. Phillip Compeau and Pavel Pevzner, Bioinformatics Algorithms - An Active Learning Approach, Active Learning Publishers,2018,3e
2. Neil C.Jones and Pavel A. Pevzner, An Introduction to Bioinformatics Algorithms, MIT Press, 2004,1e
3. David M. Mount, Bioinformatics: Sequence and Genome Analysis, CSHL Press,2004,2e
4. Dan Gusfield, Algorithms on Strings, Trees and Sequences, Computer Science And Computational Biology' Series, Cambridge University Press, 2010

## 4. Machine Learning H4

**Unit 1: Introduction to convex optimization (6 hours)**

Unconstrained optimization, Gradient descent, Constrained optimization, Lagrange multipliers, KKT conditions

**Unit 2: Introduction to Machine Learning (8 hours)**

Overview of Machine Learning terminologies and applications -Basic definitions, types of learning, hypothesis space and inductive bias, evaluation, cross-validation - Linear regression, Decision trees, overfitting - Instance based learning, Feature reduction, Collaborative filtering-based recommendation.

**Unit 3: Bayesian Networks (8 hours)**

Probability and Bayes learning - Inference and learning of Bayesian network, BN and Other Probabilistic Models -Bayesian Methods in Computational Biology – expectation maximization – Supervised Learning – classification - Bayesian networks in medical diagnosis and healthcare

**Unit 3: Unsupervised Learning (10 hours)**

Unsupervised learning - Introduction- Methods - K-Means Algorithm - Optimization Objective- Random Initialization - Auto encoders, probabilistic graphical models

**Unit 4: Support Vector Machines and Neural Networks (16 hours)**

Logistic Regression, Support Vector Machine - Introduction, hyperplane separation (maximum and soft margin hyperplanes)- linear classifier, bioinformatics Kernel function - Kernel Methods and Support Vector Machines

Introduction to Perceptron -Topology of neural network architecture- single layer ANN - multilayer perceptron - multilayer network - back propagation- probabilistic neural networks

**Unit 5: Machine Learning for Big Data Biology (12 hours)**

Statistical structure of large-scale biological datasets using ML algorithms - scalable versions for implementation on a distributed computing framework- distributed ML algorithms for: matrix factorization - convex optimization- dimensionality reduction – clustering and classification of data

**Textbooks and Reference books:**


1. Bishop CM, Pattern Recognition and Machine Learning, Springer, 2010, 3e.
2. Jason Bell, Machine Learning for Big Data: Hands-On for Developers and Technical Professionals, John Wiley and Sons, 2014, 2e.


## 5. Database Design and Cloud Computing S2

**Unit 1: SQL (6 hours)**

Introduction to DBMS, RDBMS, SQL. Creating a database; altering a database; creating tables; altering tables; dropping tables. Entity relationship diagrams; normalization; primary keys and foreign keys. SQL Queries, datatypes and expressions: INSERT, UPDATE, DELETE, SELECT; conditions using WHERE, AND and OR, sorting using ORDER BY, grouping using GROUP BY, selecting unique records using DISTINCT. Integers, character strings, ppoints, date; Boolean expressions, date expressions, numerical expressions. SQL operators and functions: mathematical operators, logical operators, max(), min(), sum(), mid(), len(), round(), first(), count(), average(). Joins, views, and other miscellaneous SQL concepts.

**Unit 2: Database design principles and data architecture (4 hours)**

Introduction to normalization theory query paradigms, query languages (SQL, Psql and MySQL) similarity searches, approx. retrieval integrated data analysis and applications VIEWs across databases, pathway modeling, network queries. Developing data architecture using MySQL and oracle. Principles and developing modules on biological databases using relational system, developing genome annotation datasets, parsing, developing tables and relational systems.

**Unit 3: Overview of internet, HTML and Basic HTML Tags (4 hours)**

Overview of the internet: Internet; web server; web browser; HTTP and HTML; domains and IP addresses. HTML tags; pairing tags; attributes. Basic HTML Tags: html; head; body; title; headings; paragraphs; bold, italics and underline, horizontal rule. Colour models and representations; coloured text and background. Working with frames, images and tables: using frames; embedding images; creating tables; formatting tables. Creating hyperlinks to other pages; creating local hyperlinks; images as hyperlinks. Abbreviations, pre-formatted text; lists – ordered and unordered. Forms: need, form elements.

**Unit 4: XML and DTD (6 hours)**

Need for XML; differences between XML and HTML; application of XML; XML examples. XML basics: elements; attributes; validation. DTD: DTD declarations-internal and external; use of DTD; DTD-XML building blocks; DTD basics: elements, attributes, entity – internal and external; validation; examples.

**Schema**

Introduction; use of schema; examples; simple types: elements, attributes, restrictions, examples; complex types: elements – empty, elements only, text only, mixed, examples; indicators: order, occurrence, group, examples; data types: string, date, misc.

**Unit 5: CGI, PHP and HTML forms (6 hours)**

History of CGI; architecture and application of CGI and PHP. HTML forms revisited: HTML form elements and attributes; text fields; text areas; radio buttons; check boxes; submit button; reset button, drop down boxes; how to create an HTML form. CGI environment variables; simple responses. Sending data to the server; GET and POST method; use of CGI.pm module.

**Unit 6: Cloud Computing (6 hours)**

Cloud Architectures and Infrastructure Design -Virtual Clusters- Parallel Computing in the Cloud • SPMD and HPC-Style Parallelism - Data Collection, Mining and Analytics on the cloud - Scalable Parallel Computing - Hadoop programming with HDFS/MapReduce and YARN; Spark Core and Resilient Distributed Data Set-; Amazon Elastic MapReduce - Cloud-Based Machine Learning Frameworks (e.g., Azure HDInsight and Data Lake; Spark MLlib for Machine Learning and GraphX for Graph Processing; Azure Machine Learning Workspace; Amazon Machine Learning Platform)

**Textbooks and References:**


1. T. King, G. Reese, R. Yarger; MySQL and mSQL; O'Reilly Media Publishers, 1999.
2. M. Gruber; Mastering SQL; M. BPB Publications, 2000.
3. D. Zak; CGI/Perl; Thomson Learning, 1st edition, 2002.
4. A. Silberschatz; Database system concepts; McGraw Hill, 4th edition, 2002.
5. R. Elmasari and S. B. Navathe; Fundamentals of database systems; 5th edition, 2006.
6. J. D. Ullman and J. Widom; A first course in database systems; Pearson Education, 3rd edition, 2008.
7. T. A. Powell; Complete reference HTML; Tata McGraw Hill, 1999.
8. D. Hunter; Beginning XML; Shroff Publishers & Distributors, 2000.


## 6 Microbiology and Immunology S2

**Unit 1: Prokaryotes (6 hours)**

Introduction to the microbial world (bacteria, yeasts/fungi, and algae); molecular genetic methods in taxonomy (ribosomal RNA/DNA, PCR, and DNA fingerprinting, DNA base composition, rDNA sequencing etc); phylogeny of living organisms; archaebacteria: taxonomic position (extremophiles, relatedness to eukaryotes and prokaryotes, unique molecular and biochemical features); bacteriophages (lysogenic and lytic cycles).

**Unit 2: Viruses (6 hours)**
History and principles of virology; viruses as acellular entities; virus structure and morphology; virus taxonomy; hierarchical classification; introduction to replication strategies; Baltimore classification. Characteristics of major bacterial, fungal, and viral pathogens; concepts of pathogenicity and virulence; microbes of medical importance (E. coli, B. anthracis, Salmonella, Influenza virus, Ebola virus, Rabies virus, Candida albicans, Histoplasma capsulatum, etc.)

**Unit 3: Introduction to immunology (8 hours)**
Innate and acquired immunity, structure and functions of immune cells. Components and organs ofthe immune system (primary and secondary lymphoid organs). Primary and secondary immune responses; clonal selection theory; antigens and antibodies – structure determination; classes of immunoglobulins, CDRs, generation of polyclonal and monoclonal antibodies, haptens, adjuvants, CD and immunophenotyping, complement system – components, properties and functions; complement pathways and their biological significance, inflammation, TLRs, cross talk of innate and adaptive immune responses.

**Unit 4: Immunological diversity and Immune Systems in diseases (6 hours)**
Genetic control of immune response (VDJ recombination); antigen processing and presentation-major histocompatibility complex (structure and functions of MHC and HLA); B cell maturation and lymphocyte activation, tissue and organ transplantation (allograft and xenograft, host reaction, and rejection); immune suppressive therapy.

**Immune system in diseases**
Host-parasite interaction; allergy and types of hypersensitivity; lymphokines and cytokines (interleukins and interferons, their production, biological functions and assay methods); immunological tolerance. Autoimmunity (Hashimoto's disease, systemic lupus erythematosus, multiple sclerosis, myasthenia gravis) and its treatment; immunotherapy (immune suppression and immune stimulation). Immune profile in the diseased states

**Unit 6: Immunological methods/techniques, vaccines and other immunotherapies (6 hours)**
Immuno diagnosis; antigen-antibody based techniques (ELISA, dot blotting, RIA, western blotting, immunofluorescence). Conventional, peptide vaccines, DNA vaccines. Toxoids, antisera, edible vaccines, plantibodies, ISCOMs, recombinant antibodies, T-cell therapy, NK-cell therapy etc

**Textbooks and Reference books:**

1. M. J. Pelczar, E. C. S Chan, and N. R. Krieg; Microbiology; Tata McGraw Hill, 2001, 5th edition.
2. C. J. Alexopoulos, C. W. Mims and M. Blackwell; Introductory Mycology; Wiley, 1996, 4th edition.
3. S. J. Flint, L.W. Enquist, V. R. Racaniello, and A. M. Skalka; Principles of Virology: Molecular Biology, Pathogenesis, and Control of Animal Viruses; ASM Press, 2004, 2nd edition.
4. J. G. Black; Microbiology: Principles and Explorations, John Wiley and Sons, 2008, 7th edition.
5. M. T. Madigan, J. M. Martinko, P. V. Dunlap and D. P. Clark; Brock Biology of Microorganisms; Pearson Education, 2009, 12th edition.
6. J. Willey, L. Sherwood and C. Woolverton; Prescott, Harley, Klein's Microbiology; McGraw Hill Higher Education, 2008, 7th edition.
7. K. Murphy, P. Travers and M. Walport; Janeway's Immunobiology; Taylor & Francis Publishers, 2008, 7th edition.
8. K. Todar, Todar's Online Textbook of Bacteriology (available free of cost at http://textbookofbacteriology.net/).
9. R.A. Goldsby, T.J. Kindt and B.A. Osborne; Kuby Immunology; W. H. Freeman & Co, 2000, 6th edition.

## 7. Machine Learning – Laboratory P4

1. Taxonomy of Machine Learning Methods
2. Supervised Regression and Classification Methods (Linear Regression, Decision Trees for Machine Learning, Bayesian Classifier, Support Vector Machines)

3. Unsupervised Machine Learning Algorithms (Association Analysis, Clustering Methods Without Labels, Dimensionality Reduction)

4. Clustering and Dimensionality Reduction Methods (Cluster Analysis and K-Means Clustering, and Principal Component Analysis)

5. Model Development and Selection for Machine Learning (Model Over- and Under-fitting)

## 8. Database Design and Cloud Computing – Laboratory P2

1. Developing data architecture using principles and developing modules on biological databases using relational system
2. Developing genome annotation datasets, parsing, developing tables and relational systems
3. HTML form elements and attributes; text fields; text areas; radio buttons; check boxes; submit button; reset button
4. Creation of a HTML form - CGI environment variables; simple responses
5. Sending data to the server; GET and POST method; use of CGI.pm module
6. Installing the Cloud Monitoring Agent and allowing traffic to the Compute Engine Instances
7. Connecting to the Cloud SQL Database
8. Performing Operations on the Cloud SQL Database

## 9. Biostatistics and 'R' Programming – Laboratory P2

**Download and install R**

Downloading and installing specific version of R in linux and windows machines; Configuring the software; Learn to download and install external libraries

**Learning R language**

Declaring variables; Math library and mathematical operations; String operations; Vectors; Data frames; lists; Matrix operations; Writing R scripts; File I/O – reading text, excel and other format files into R; Manipulating columns and rows of data tables; Filtering of data; R graphics package – learn to make plots of points, lines, scatter plots, histograms, Pie-charts, bar charts and more specific plots; ggplot2 library for enhanced plots; Writing Functions and creating packages in R;

**Biostatistics in R**

Permutations and combinations of a set of objects, Venn Diagrams; Set theory operations; Creating and displaying Contingency tables; Creating ROC curves;

Computing the statistical parameter of the data in R; Box-Whiker plots; Confidence intervals; Plotting points and bar plots with error bars; One sample and two sample hypothesis testing with R library functions and user written scripts;

Non-parametric tests for one and two variables; Analysis of the variance (ANOVA); Computing covariance and correlation

Linear Regression, non-linear regression and multiple linear regression in R; Logistic regression ; Sample size estimations with R packages; p-value correction and False Discovery Rate (FDR) computation with R library functions;

Clustering Algorithms in R.

**References**

1. Wayne W. Daniel, Chad L. Cross, Biostatistics – Basic concepts and methodologies for Health Sciences, Wiley Students Edition, 2016,10e
2. Robert V.Hogg, Tanis, Zimmermann, Probability and Statistical Inference, Pearson Education Asia, 2015, 9e

**SEMESTER III**

| | Course title | L-T-P-C | | |
|---|---|---|---|---|
| 1 | **Next Generation Sequencing and Genomic Data Analysis (bulk and single cell omics)** | 3-1-0-4 | H | 4 |
| 2 | Biomedical Imaging and Informatics | 3-1-0-4 | H | 4 |
| 3 | Deep Learning and Applications | 3-1-0-4 | H | 4 |
| 4 | (A)Computational Drug Design and AI applications/ (B)Systems Biology and Network Modelling | 2-0-0-2 | E | 2 |
| 5 | Clinical Trial Data Analysis | 2-0-0-2 | S | 2 |
| 6 | Optional Elective: (A)Scientific Writing and Communication (B)Physical and Mental Health | 2-0-0-2 | O | 2 |
| 7 | Data Analysis in Genomics, Transcriptomics and Proteomics - Laboratory | 0-0-4-4 | P | 4 |
| 8 | Image Informatics Laboratory | 0-0-4-4 | P | 4 |
| 9 | Advanced Machine Learning Laboratory | 0-0-4-4 | P | 4 |
| | **TOTAL** | | | **30** |

## 1. Next Generation Sequencing and Genomic Data Analysis (bulk and single cell omics) H4

**Unit 1: Concepts of genomics and genome sequencing technologies (6 hours)**

Introduction to genome sequencing - human genome project - Genomic elements, SNPs and genome-wide association studies - Sanger's dideoxy method versus NGS – Next Generation Sequencing-platforms and technologies

**Unit 2: Genome Assembly (10 hours)**

Sequence assembly concepts and challenges in assembling short reads; Algorithms for assembling short reads using graph theory such as Hamiltonian cycle and de Brujin; Writing code for assembling reads- Gene prediction and annotation; gene ontology (GO); Mapping algorithms - Burrow-Wheeler algorithm

**Unit 3: Transcriptomics and Epigenomics (12 hours)**

Types of RNAs and the respective roles in cells – microRNAs and non coding RNAs -. Techniques used in transcriptomics; RNA-seq.; analysis of RNA sequencing data (lncRNA, microRNA)

Chromatin organisation- Chromatin Immunoprecipitation - ChIP-sequencing – read mapping, peak calling and visualization- TSS Analysis, differential binding analysis – functional and motif analysis- analysis of chromatin signatures from ChIP-sequencing data - DNA methylation - 3D organisation of the genome

**Unit 4: Whole Genome Sequencing and exome sequencing (12 hours)**

Whole genome sequencing and exome sequencing - Identification genetic variations from genome sequencing: SNPs, SNVs, translocation, copy number variation. Concepts behind genome-wide association studies

**Unit 5: Single Cell Sequencing (12 hours)**

Need for single cell sequencing – intra tumoural heterogeneity – single cell transcriptome sequencing and data analysis- computational challenges in analysing single cell transcriptomic data- single cell and multivariate approaches in CRISPR/Cas/ genetic perturbation screens

**Unit 6 : Agricultural Genomics (8 hours)**

Use of genomic data in the localization and description of gene functions and their interactions involving: mapping of genes/QTL with agronomic relevance- Tomato fruit shape gene QTL mapping and cloning- detection and development of molecular markers ideally perfect markers, cloning, deletion and overexpression of agronomically relevant genes in pathogens, plants and animals, generation and characterization of transgenic plants and animals, gene editing, identifying gene networks and the disclosure of gene functions in relation to relevant traits - Genetics and genomics of abiotic stresses- Applications to plant breeding and genetics- Nematode-host interaction genomics

**Text Books and References**


1. Xinkun Wang, Next-Generation Sequencing Data Analysis, CRC Press,2016,1e
2. Publications on NGS data analysis from literature


## 2. Biomedical Imaging and Informatics H4

**Unit 1: Basics of Biological Imaging (15 hours)**

Analytical and diagnostic imaging tools in research and health care- Introduction to biological imaging – Fundamentals of Microscopy- Principles for different types of microscopy - fluorescence, confocal and multiphoton microscopy – light sheet and super resolution microscopy-electron microscopy-Protein Interaction and dynamics– FRET Microscopy and FRAP- 3 Dimensional imaging of cell cultures, tissue and cancer cell spheroids – high throughput microscopy

**Unit 2: Biological Image Processing (15 hours)**

Basics of image processing and methods to extract quantitative image features- Image types- bits, pixels and voxels- Introduction to the spatial and frequency domains - Image Processing and enhancements - Contrast adjustments, histogram operations, look-up tables, filters for denoising, sharpening, finding edges - Intensity and spatial measures, manual and threshold-based measurements, RoIs, counting, binary images, improving accuracy with pre-processing - linking image data to genome-scale data and clinical data and phenotypes – analysis of images from high content imaging experiments – Digital Pathology

**Unit 3: Medical Imaging (15 hours)**

General properties of medical imaging- X-ray, Ultrasound imaging, Computed Tomography- Magnetic Resonance Imaging (MRI) and fMRI - Positron emission tomography (PET) scans

**Unit 4: Medical Image Processing (15 hours)**

Uses of Information Theory in Imaging- Image format: HIS, RIS, PACS. IHE and workflow integration - Image storage and retrieval - Imaging interpretation – Imaging process-sampling, Quantization, Image Enhancement - Filtering in frequency domain- Orthonormal Transforms, Image compression - Image De-noising, Segmentation, Feature extraction and Object Recognition – Tissue Contrast-Spatial Localisation-Phase coded and frequency coded domain – Image artefacts in MR Imaging.

**Textbooks and Reference books:**


1. Toshiyuki Furukawa, Biological Imaging and Sensing, Springer Verlag, 2004
2. Gonzales, RC and Woods RE, Digital Image Processing, Pearson,2017, 4e.
3. Wolfgange Birkfellner, Applied Medical Image Processing – A Basic Course, CRC Press, 2014, 2e.
4. James Pawley, Handbook of Biological Confocal Microscopy, Springer, 2006, 3e.


## 3. Deep Learning and Applications H4

**Unit 1: Deep Learning (12 hours)**

Fundamentals of Deep Learning – Differences between classical machine learning and deep learning - building and training deep neural networks -implementation of vectorized neural networks- architecture parameters –

**Unit 2 : Deep reinforcement learning ( 12 hours)**

Deep Reinforcement Learning – implementing a neural network using Tensor Flow and Keras- transfer and multi-task learning – Deep Generative Modelling

**Unit 3: Convolution Neural Networks (CNNs) (18 hours)**

Foundations of CNNs–Foundational layers of CNNs (pooling, convolutions) - implementation and stacking layers in a deep network to solve multi-class image classification problems- deep convolution models-building a CNN to image, video and 3D data

**Unit 4: Recurrent Neural Networks (18 hours)**

Basic concepts of recurrent neural networks- language models and sequence generation- vanishing gradients with recurrent neural networks-building and training Recurrent Neural Networks and its variants (GRUs, LSTMs) –bidirectional and deep RNNs - applications of RNNs to character-level language modelling - NLP and Word Embeddings

**Text Books**


1. Goodfellow, I., Bengio, Y., and Courville, A. (2016) 'Deep Learning'. http://www.deeplearningbook.org/ and https://github.com/HFTrader/ Deep Learning Book

2. Matthew D. Zeiler and Rob Fergus (2014) 'Visualizing and Understanding Convolutional Networks' by "Convolutional Neural Networks for Visual Recognition" (Stanford course given by Fei-Fei Li, Andrej Karpathy, and Justin Johnson, 2016): http: //cs231n.github.io/

3. Deep Natural Language Processing course offered at the University of Oxford: https://github.com/oxford-cs-deepnlp-2017/lectures

4. Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle1, "Deep learning for computational biology", Molecular Systems Biology (2016) 12: 878.


**4. A. Computational Drug Design and AI applications E2**

**Unit 1: Drug Target Interactions (6 hours)**

Drug targets classification, drug-target interaction and dose-response relationships. -drugs as agonist and antagonists -Pharmacokinetics and Pharmacodynamics: ADME -Bioavailability of drugs - Lipinski's rule

**Unit 2: Drug Discovery Process (6 hours)**

Overview of drug discovery: Properties of a drug like molecule, structure based and ligand-based drug design. Predicting potential binding sites on a protein structure, high-throughput virtual screening to identify lead molecules

**Unit 3: Drug Interactions (8 hours)**

Algorithms for molecular docking Protein-protein docking methods. Ligand based drug design – pharmacophore mapping and QSAR studies. Lead optimization and additional in silico validation- Virtual Screening - Quantitative Structure Activity Relationship (QSAR) – 3D QSAR and COMFA

**Unit 4: AI in drug discovery processes (12 hours)**

Machine learning algorithms for target identification in structure-based drug design – Predicting Protein-ligand Binding Affinities– deep learning algorithms for high throughput virtual screening and for predicting solubility of molecules- smiley notations – Generative Adversarial Networks (GANs) in drug discovery processes- Machine Learning, Natural Language Processing, and Deep Learning in Drug Safety and Pharmacovigilance

**Text Book**

Harry Yang, Data Science, AI, and Machine Learning in Drug Development (2022), Taylor and Francis 1e

## 4B Systems Biology and Network Modelling E2

**Unit 1: Numerical solutions to differential equations (4 hours)**

Numerical solutions of differential equations: Numerical methods of Ordinary Differential Equations (ODE); Computational methods of solving Partial differential equations; Solving a system of ODEs;

**Unit 2: Nonlinear systems and their biological applications (10 hours)**

1D ordinary differential equations; fixed points and their stability; bifurcation; 2D flows; 2D linear systems; 2D non-linear dynamics including 2D oscillations; Systems with more than 2 variables; Stability analysis; Parameter sensitivity analysis -Simple applications of a system of equations; Population models – insect breakout model, SIR model and its applications; Biological examples of simple gene regulation, glycolytic oscillations, heart beat.

**Unit 3: Reaction Kinetics (3 hours)**

Mathematical formulation of elementary biochemical reactions, Law of mass action, Co-operativity and Hill function, Enzyme reaction kinetics and Michaelis-Menten approximation

**Unit 4: Network analysis (12 hours)**

Mathematics of networks: Networks and their representation. Adjacency matrix, weighted networks, bipartite networks. Network Models (ER, Small-world and scale-free). Network measures: Clustering coefficient, Centrality, Clusters and Motifs. Network perturbations. Different types of biological networks. Methods to build protein interaction and gene regulatory networks. Biological network databases and visualization tools. Biological applications.

**Unit 5: Metabolic Networks and Flux Balance analysis (3 hours)**

Flux balance analysis of simple metabolic networks with specific examples.

**Text Book**

Uri Alon, An Introduction to Systems Biology: Design Principles of Biological Circuits, Chapman & Hall/CRC Mathematical and Computational Biology), 2006, 1e

## 5. Clinical Trial Data Analysis S2

**Unit 1: Introduction (6 hours)**

Introduction to clinical trials, different types and stages in clinical trials, clinical trial organizations.

**Unit 2: Basic Study designs (10 hours)**

Control design, parallel group design, Cross-over design, blinding and randomization, Randomization methods, cohort studies and case-control studies.

**Unit 3: Data processing and visualization (8 hours)**

Data types, Variables, precision, missing value imputation, visual representation of data in two and

three dimensions, data summary and presentation.

**Unit 4: Statistical analysis of data (4 hours)**
Measurements of association, relative risk, absolute risk, odds ratio, confidence intervals. Sample size and power calculations, Correlation, Hypothesis testing, Regression analysis of data, Logistic regression, Survival analysis and Factor analysis of clinical data.

**Unit 5: The clinical trial landscape (4 hours)**
The use of above-mentioned methodologies to interrogate clinical trial landscapes (such as in the US or in India).

**Textbooks and Reference books:**

1. Cleophas, Ton J, Zwinderman, Aeilko H, Understanding Clinical Data Analysis, Springer International Publishing, Switzerland,2017
2. Shein-Chung Chow, Jen-Pei Liu, Design and Analysis of Clinical Trials: Concepts and Methodologies, Wiley Press, 2013,3e


**6. Optional Elective: (A)Scientific Writing and Communication (B)Physical and Mental Health** **O2**

**Scientific Writing and Communication**

**Unit 1: Basics of good writing (8 hours)**
Brushing up grammar: nouns and pronouns; adjectives; verbs; adverbs; prepositions; conjunctions, articles; order of words; punctuation, vocabulary; idioms; figures of speech; active vs. passive voice, etc. Writing skills: paragraph construction; common mistakes while writing, better sentences, story writing; tools to edit your write-up. Comprehension, expansion of passages, precis, paraphrasing.

**Unit 3: Scientific writing (8 hours)**
Structure of a paper: abstract, title, introduction, methods, results and conclusions; general rules of formatting scientific documents; writing a paper in a particular journal's style; how to cite a publication; identifying and avoiding plagiarism.

**Unit 4: Professional writing (5 hours)**
Composing a variety of professional emails and letters; preparing an effective resume.

**Unit 5: Effective oral presentations (5 hours)**
Essential principles of a good talk; preparing to make presentations; preparing effective power point presentations; talks for different kinds of audiences.

**Unit 6: Other modes of communication (6 hours)**
Short accounts of scientific discoveries for the layman; social media and science communication; using social media to transmit a scientific discovery that has influenced you.

**Textbooks and Reference books:**

1. N. D. V. Prasada Rao (revised); Wren & Martin's High School English Grammar and Composition Book; S. Chand & Co. Ltd; Revised edition, 1995
2. T. L. H. Smith-Pearse; English Errors of Indian Students; Oxford, 5th revised edition; 1997.
3. J. Butcher and C Drake; Butcher's Copy-editing: The Cambridge Handbook for Editors, Copy-editors and Proofreaders; Cambridge University Press; 4th edition, 2006.
4. P. Sebranek, V. Meyer and D. Kemper; Writers Inc: A Student Handbook for Writing & Learning; Great Source Education Group Inc,1995.
5. R. A. Day; Scientific English: A Guide for Scientists and other Professionals. Greenwood Press; 3rd Revised edition, 2011.
6. M. Raman and S. Sharma; Technical Communication; Principles and Practice; Oxford University Press, New Delhi; 2nd Edition, 2012.

7. D. Carnegie; The Quick and Easy Way to Effective Speaking; Simon & Schuster; Reissue edition, 1990.
8. J. V. Vilanilam; More effective communication; A manual for professionals, Response Books (SAGE Publications); 1st edition, 2000.

**Optional Elective 2 : Physical and Mental Health**

**Unit 1: Illness, Health and Wellbeing (6 hours)**

Illness, Health and Wellbeing; Health continuum; models of health and illness: Medical, Bio psychosocial; Holistic Health; Health and Wellbeing.

**Unit 2: Stress and Coping with stress (8 hours)**

Nature and Sources of Stress; Personal and Social Mediators of Stress; Effects of Stress on Physical and Mental Health; Coping and Stress management

**Unit 3: Health Management (8 hours)**

Health Management: Health enhancing behaviours: Exercise, Nutrition, Meditation, Yoga; Health compromising behaviours (alcoholism, smoking, internet addiction); Health Protective behaviours, Illness Management.

**Module 4: Human Strengths and Life Enhancement (10 hours)**

Promoting Human Strengths and Life Enhancement: Strength: Meaning; Realising one's strength; Maximizing Unrealized Strength. Weakness – Identifying and Overcoming Weakness. Strategies to develop hope and optimism

## 7. Data Analysis in Genomics, Transcriptomics and Proteomics - Laboratory P4

**1: Genome Analysis**
Genome Assembly- identification of SNPs from genome sequencing data, identification of repetitive elements, gene prediction and discovery - cis-regulatory elements.

**2: Transcriptome Analysis**
Analysis of RNA-sequencing data- differential gene expression- heat maps- isoforms and novel transcripts – analysis of data from small RNA sequencing experiments

**3: Epigenetics**
Analyzing ChIP- sequencing data to understand genome wide chromatin signatures

**4: Whole genome sequencing and exome sequencing**
Analysis of WGS and exome sequencing data

5: Analysis of single cell RNA sequencing data

**6 Mapping of genes/QTL with agronomic relevance**

**7 Analysis of nematode host interaction genomics**

**Textbooks:**
1. Xinkun Wang, Next-Generation Sequencing Data Analysis, CRC Press, 2016,1e

**References:**
Publications on NGS data analysis from literature

## 8. Image Informatics Laboratory P4

1. Analysis of fluorescence and confocal microscopy images
2. Processing of 3D images from confocal microscopy
3. Digital image analysis and Quantitative Imaging
4. Imaging process, Sampling, Quantization, Image Enhancement
5. Filtering in frequency domain, Orthonormal Transforms, Image compression, Image De-noising
6. Segmentation, Feature extraction and Object Recognition
7. Medical imaging – CT, MRI and Positron Emission Tomography (PET) scans
8. Quantifying 3D images of cell culture and tissues
9. Quantifying confocal imaging of 3D Spheroids/organoids
10. Analysis of high content imaging modules

## 9. Deep Learning and Applications Laboratory P2

1. Building neural networks (NN) using PyTorch to predict patterns from real data
2. Building convolutional neural networks (CNNs) to classify images based on patterns
3. Implementing a recurrent neural network (RNN)
4. Implementing RNN variants LSTMs and GRUs) with PyTorch to build specific tools
5. Training a simple RNN in PyTorch to do time series prediction.
6. Implementing character level sequence RNN
7. Building generators and discriminators using convolutional, batch normalization and fully connected layers

### SEMESTER IV

| | Course title | L-T-P-C | | |
|---|---|---|---|---|
| 1 | Capstone Project/Dissertation | | | 18 |
| 2 | Project Viva | | | 2 |
| | **TOTAL** | | | **20** |

## 1. Capstone Project/Dissertation 18

## 2. Project Viva 2