School of Computational & Integrative Sciences, JNU

[P.G. Diploma Big Data Analytics in Biology](http://ccbb.jnu.ac.in/intranet/index.php/scis-course-syllabus/p-g-diploma)

Syllabus (All Courses)

**Semester-I**

**IT-601 N Programming in R and Python**

Syllabus no available

**IT 602 Data analytics and modeling**

Review of basic concepts in probability and mathematical statistics. Graph and network theory concepts.

Data Mining basic concepts and nature of the data mining problems.

Data summarization and visualization.

Data pre-processing: Data cleaning. Data integration, Data reduction. Data transformation and discretization.

Data classification techniques: Decision tree, Bayesian classification.

Performance measures and improvement of classification techniques. Ensemble learning.

Bagging boosting and random forests.

Advanced classification: Bayesian belief networks, Neural networks, Support vector machines

Regression techniques. Multiple linear regression, Regularization. Ridge regression and LASSO. Elastic Nets.

Mixture models.

Elementary ideas of Deep learning. Boltzman machine, CNN, LSTM and other variants.

Cluster analysis of large scale data sets: partitioning methods, hierarchical clustering.

Clustering graphs and network data

**Practicals:**

1. Re-implementing selected best machine learning tasks for selected applications in benchmark data sets such as in UCI and Kaggle.
2. Writing practical codes to solve selected biological problems using traditional machine learning algorithms such as SVM, MLR and neural networks.
3. Writing basic codes to implement CNN and LSTM in Tensorflow framework.
4. Assessing scalability and overhead in parallelization of simple tasks and one case study with advanced problem.

**IT 603 : Data warehousing and integration**

Introduction: Data warehousing: Definition, need and milestones of data warehousing.

Transactional systems and operational data storage.

Data warehouses and data marts. Architectures of a data warehouse. Generalized, federated and hub and spoke data warehousing. Architecture framework. Management and control modules.

Components of a data warehouse. Source data components. Data staging and storage components. Data delivery and meta data components. Management and control components.

Real time data warehousing. Query tools. Browsing tools. ERP, KM and CRM data warehousing. Active data warehousing.

Data fusion and integration. Data standards. Metadata and OLAP. Web-enabled data warehousing.

Workload management in data warehouse. Query classification, ETL and CDC workloads. Data loading techniques. Data transport.

Cloud computing. Infrastructure as a service. Platform as a service. Data warehousing on a cloud computer-technologies and issues.

Data virtualization. In-memory technologies.

Integration strategies in data warehouse. Data driven integration. Physical component integration and architecture. External data integration.

Hadoop and RDBMS. Semantic framework for data integration. Lexical processing. Semantic knowledge processing. Information Curation. Visualization.

Biological data integration using InterMine framework. InterMine powered data warehouses. FlyMine, MouseMine, YeastMine, INDIGOmine, TargetMine, HumanMine, PhytoMine for plant data.

Biological data warehousing with BioMart. ID conversion. Biological Sequence retrieval and enrichment analysis.

**Practicals:**

1. Setting up a batch queuing system such as Torque from scratch and configuring its multiple functionalities.
2. Using Apache Spark and Hadoop to perform tasks on a pre-configured Cloud.

**IT 604 Molecular modeling and simulation**

Simulation of biological processes. Equilibrium and steady state calculations. Hill plot. Van’t Hoff equation. Steady state enzyme kinetics. Lineweaver Burk fitting. Eadie-Hifstee fitting. Competitive inhibition.

Simulation of diffusion and transport. Fick’s law. Electrophoresis. Counter current diffusion.

Compartment models in physiology and pharmacokinetics. Periodic dose administration. Oscillations in calcium metabolism.

Regulation and control in metabolism. One-substrate, one-product reactions. Steady state flux calculation. Flux control coefficients.

Models of regulation. Feed-forward loops. Regulation of signalling. Bacterial chemotaxis. Morphogenesis.

Protein and nucleic acid structures and their interactions. Protein structural hierarchy. Turns and loops. Homology modeling.

Basics of molecular dynamics. Software tools for MD simulation in public domain. Gromacs, Amber, NAMD, VMD, Pymol.

Symplectic transformation. Harmonic oscillator example. Linear stability. Resonance condition and artifacts. Multiple-timestep methods. Langevin dynamics. Brownian dynamics.

Sampling methods. Coarse graining methods. Biasing approaches.

Similarity and diversity in chemical drugs. Molecular modeling in rational drug design. Problems in chemical libraries. Similarity and diversity sampling.

Advanced topics and current trends.

**IT 605 Genomics Data Analytics**

**Introduction to Genomics, sequencing and recombinant DNA technology**

**Next Generation Sequencing Technologies:** Methods and applications

**Whole Genome Sequencing and Analysis:** Concept, methods, assembly methods (de novo and reference-based) and algorithms, genome annotation (structural and functional), comparative genomics

**High-throughput Transcriptome Profiling:** Concept, methods and applications; transcriptome construction (de novo and reference-based), differential gene expression

**Non-coding RNAs:** Small RNAs, miRNAs, long non-coding RNAs; sequencing and prediction methods; biological relevance

**Single nucleotide polymorphisms:** Genome resequencing; data processing and SNP prediction; applications in agriculture/human health

**Practicals:**

1. Creating and benchmarking an NGS pipeline from scratch using tools such as Bowtie, Cufflinks, Samtools.
2. Writing advanced R and Bioconductor powered genomics applications for specific case studies.

**IT 606: Social Networks and Epidemiology**

Constructed vis-a-vis realized networks. Examples: Transportation, Communication, Biology, Health, Economics, Finance, Sociology; Basic Properties of networks: Degree, Clustering, Assortativity, Centrality, Modularity; Simple models of networks: Erdos-Renyi model (Random graphs), Watts-Strogatz model (Small-world networks), Barabasi-Albert model (Scale-free networks); Temporal networks and Multiplex networks: Examples: Economics, Finance, Sociology, Health, Disease spreading, Dynamical processes on networks: Diffusion on networks, Epidemiological models; Theoretical lectures to be followed by hands-on workshops on: (i) network simulations using R/Python, (ii) empirical data -- network construction, and (iii) network analyses using packages (Cytoscape, Gephi, etc.).

Classical epidemiological model (multi-compartmental models): SIR, SEIR and their extended models; Renormalization theory of infectious disease; Deterministic and stochastic analytical and computational methods of solving the epidemiological models: Equilibrium states, endemic properties, Oscillating properties, disease transmission; Epidemic waves: first, second etc and detection method/s; Immunity during infection; Reproduction number: method of calculation (from data and models); Epidemic forecasting: Malthus, Logistic, Gompertz, Winsor growth laws as means of forecasting; Epidemiological data: Data mining, Data analysis, Parameter estimation, Model fitting and forecasting; Epidemic model simulation: deterministic and stochastic simulation methods/software, Data modeling for intervention.

Recommended books:

1. *Data science and complex networks: real case studies with Python*. Guido Caldarelli and Alessandro Chessa (Oxford University Press, 2016)
2. *Network Science*. Albert-László Barabási (Cambridge University Press, 2016)
3. *A First Course in Network Science*. Filippo Menczer, Santo Fortunato, Clayton A. Davis

(Cambridge University Press, 2020)

1. M. J. Keeling and P. Rohani. Modeling infectious diseases in humans and animals (Princeton University Press, 2011).
2. Anderson, R. M. & May, R. M. Infectious Diseases of Humans (Oxford Univ. Press, 1991).

# **IT 607 Personalized medicine and clinical trials**

Genomics and personalized medicine. ENCODE project and mapping of the functional genome. Next generation sequencing and multi-gene approaches. Pharmacogenetics and pharmacogenomics.

Epigenetics. Diabetes and DNA methylation. Proteomics and metabolomics.

Integrative medicine. Nutrition and exercise.

Placebo effect. Psychological factors and neurobiology.

Personalized medicines for cancer. Oncogenes and molecular abnormalities. Tumor subtypes and personalized diagnostics. MicroRNA. Overview of Liquid biopsy and Immunotherapy. Personalized medicine approaches for other diseases such as Type II diabetes, Rheumatoid Arthritis, Multiple sclerosis, cardiovascular disease and AIDS.

Biomarkers for personalized medicine. Druggable genome. Genome-wide association studies (GWAS). International HapMap project. Clinical biomarker discovery and somatic mutations. Moral and ethical issues.

Clinical trial fundamentals. Phase I, Phase II, Phase III and Phase IV trials.

Ethical considerations in clinical trial. Safety, Privacy and confidentiality.

Trial with extensive data collection. Superiority vs non-inferiority trials. Intervention.

Response variables. Biomarkers and surrogate response.

Considerations for study population in clinical trials. Pharmacogenetics and generalizations. Recruitment.

Trial design. Randomized control traits. Concurrent studies. Historical data and limitations.

Cross-over design. Withdrawal studies. Factorial design. Hybrid designs.

Randomization process. Simple, blocked, stratified and adaptive randomization procedures.

Blinding. Single, double and triple blind trials.

Sample sizes and baseline assessments. Reporting harm. Health related quality of life (HRQL) measures. Methodological and design issues.

Multi center trials. Large, simple trials.

Regulatory issues. Pretrial requirements. Interventions. Post trial requirements. Comparative overview of pretrial, during trial and post trial regulatory requirements in Indian, USA and Canada.

**Practicals:**

1. Case studies on clinical data analysis, specially on large cohorts.

**Additional requirements for Project work:**

This will depend on individual labs. Large scale bid data bases project will be implemented using compute intensive tasks such as traditional and deep machine learning, high throughput docking, MD simulations of large libraries and long time scales, genomic data analysis protocols and other related works based in student/guide synergies.

**Semester-II**

IT-PR2 Project-2

-------

IT-717 Algorithms and Practical Implementation in Bioinformatics (BioAlgorithms)

Syllabus no available