YENEPOYA
(DEEMED TO BE UNIVERSITY)
Recognized under Sec 3(A) of the UGC Act 1956

Biological Data Analytics and Bioinformatics

Core course for Pre-PhD: 4 credits

Yenepoya Research Centre
Yenepoya (Deemed to be University)
University Road, Deralakatte
Mangalore – 575018

Course Name: Biological Data Analytics and Bioinformatics

1. Course Type: :Core
2. Level :Ph.D. (Pre-PhD course work)
3. Credit Value: :4 Credits
4. Total Hours : 60 (L:P:S: 10:25:25)
5. Total Marks: : 100 (IA= 40 + Final exam= 60)

6. Course Objectives

- Develop knowledge on advanced research topics pertaining to the principles behind genetic variation and enable them to comprehend the drivers of evolution over spatiotemporal realms.
- To expose the research scholars towards biological data analysis using bioinformatic tools.
- Develop skills to design and perform experiments that are relevant in emerging areas of data science and bioinformatics.Train the research scholars in programming in Python and R that are explicitly used in biological and biomedical research.

7. Learning Outcome

- This course will enable the students to analyze biological data using appropriate statistical, mathematical and computational measures.

8. Competencies

1. Describe the concepts of quantitative genetics, heritability and genetic diversity
2. Apply suitable statistical models for phylogeny reconstruction and population genetic analysis.
3. Identify the association between the phenotypic outcomes and the underpinning of the genetic make-up
4. Perform basic programming in Python and R computer programs
5. Perform biological data analysis using advanced statistical and computational tools
6. Interpret research data to offer valid interpretations to the given research problems.
7. Practice ethics and maintain privacy, confidentiality and integrity of the genetic data under investigation.
8. Practice biological/hazardous waste management protocols appropriately during sample collection, storage, analysis and disposal as per prescribed guidelines.

9. Content of the Course

Module 1: Quantitative Genetics (12 h)

1.1. Utilization of the data generated from human genome project in genomics research. Conceptual understanding and analysis of data related to polygenic and multifactorial traits and inheritance. Comprehending Heritability and estimating quantitative genetic diversity: board and narrow sense heritability. Construction and applications of additive, dominant and recessive models in respect to change in allele frequencies, over and under dominance. Application of statistical and mathematical models in genetics.

1.2. Concept of Inbreeding and Out breeding Depression and their applications in real-life scenarios. Ancestry based approaches to prevent inbreeding depression. Conceptual understanding of Neutral and Nearly Neutral Theory, performing Neutrality Tests on genomic data: Neutrality Tests – Tajima's D, HKA Test, MK Test, Ka/Ks Test, EHH, Determination of signatures of recent selective sweep: Extended Haplotype Homozygosity (EHH), XP-EHH. Applications of Linkage and Linkage Disequilibrium (LD) in genomic research.

Module 2: Population Genetics (12 h)

2.1. Statistical and mathematical measures of genetic variation in randomly and non-randomly mating populations. Applications of Hardy-Weinberg principle in population genetics. Conceptual and mathematical understanding of the sources responsible for changes in gene frequencies: mutation, selection, migration and isolation; random genetic drift. Insights and analysis of data pertaining human migration, natural selection and mutation-selection balance. Conceptual understanding and mathematical derivations of various types of selections (natural selection, sexual selection, kin selection).

2.2. Conceptual understanding and analysis of data related to Fixation indices, Heterozygosity, Effective Population Size and Coalescent Theory. Understanding importance of sexual reproduction in evolution. Applications of Haldane's Rule, Muller's Ratchet; Fisher-Muller Hypothesis and Red Queen Hypothesis in population genetic research.

2.3. Implication of population genetics on human diseases and dysfunctions; Understanding the association between genetic variants and drug responsiveness; personalized medicine; pharmacogenomics.

Module 3: Phylogenetics (12 h)

3.1. Use of sequence alignment tools – Clustal Omega, Sea View and PRANK alignments. Understanding gaps and gap penalties. Clustering and Phylogeny reconstruction.

Understanding methods for Phylogeny analysis: Distance and Character based methods. Computation of phyogenetic trees using distance matrix methods (Neighbor Joining, UPGMA), Minimum Evolution, Maximum Parsimony method, Maximum likelihood and Bayesian inference; Application of Information Criteria (AIC, AICc and BIC) in phylogeny reconstruction. Application of Molecular Systematics in evolutionary genetic research.

3.2. Use of multiple sequence editors such as Clustal Omega and PRANK. Learning the utilities of different types of multiple alignment files (fasta, nexus (interleaved/sequential), plylip (interleaved/sequential), and MEGA); converting one file type to another. Missing data/gaps in the sequences and its implication in multiple sequence alignment. Determination of genetic distances among sequences in multiple alignments. Use of Model Test to choose best nucleotide substitution model and Phylogeny reconstruction.

3.3. Conceptual understanding and applications of the Molecular Clock; Calibration of molecular clock; Limitation of molecular clock models. Deducing evolutionary histories through mitochondrial DNA and Y chromosome. Evolution of the genome: Genomic sequencing and mapping: Genome databases Whole Genome Projects.

Module 4: Bioinformatics (15 h)

4.1. Conceptual understanding and applications of PAM and BLOSSUM matrices. Database similarity search – BLAST and BLAT. Markov and Hidden Markov model.

4.2. Programming in Python and R. Use of R in Phylogenetics and Phylogeography; Use of R in Molecular Evolution; Use of R in sequence assembly and alignment; Application of molecular evolution software packages available in R.

4.3. Whole Genome Sequencing methods and analysis. Understanding the workflows for NGS experiments. Analysis of NGS data using PLINK, SAM tools, and VCF tool. File conversion and basic statistical analysis using PLINK, VCF tools and Eigensoft. Identification of disease genes using OMIM database. Construction of genomic maps.

4.4. Application of Literature Informatics in biological research. Biological Literature Information access, storage and retrieval systems- Primary and secondary databases of genomics and proteomics. Applications of genome browsers in bioinformatic research: UCSC Genome Browser, Ensemble browser. Variation Databases: dbSNP, ExAC, gnomAD, 1000 Genome, SNPedia; Phenotype–Genotype Associations: PheGenI, GWAS Catalogue; Cancer Genetics & Genomics: TCGA; Comparative Genomics: UCSC GB, MGI, Flybase, Wormbase, ZFIN; Clinical Relevance ClinVar, OMIM, GTR, Gene Tests;

Gene Regulation GTEx, RegulomeDB, HaploReg; Ontologies, Pathways & Networks Pathway Commons, Gene Ontology; Proteomics Miscellany UniProt.

4.5. Nucleic acid sequence analysis; Reading frames; Codon Usage analysis; Translational and transcriptional signals; Epigenetic signatures and prediction of regulatory regions in the genome. Splice site identification; Gene prediction methods; Use of GERP and Phylop scores.

Module 5: Multivariate Data Analytics (9 h)

5.1. Statistical genetic analysis: Correlation and Regression. Understanding association between correlation coefficient and LD.

5.2. Introduction to Bootstrap method – steps in Bootstrap method; Standard and percentile bootstrap confidence intervals; Use of transformations in bootstrap.

5.3. Non-parametric statistical tests: Wilcoxon signed rank test; Mann-Whitney U test; Kruskal-Wallis test; Spearman's Rho.

5.4. Multivariate statistical analysis: MANOVA, ANCOVA and their applications, Principal component analysis (PCA), hierarchical clustering, K-means clustering.

Teaching-learning methods

Modules	**Teaching-learning**		
	Lecture	Practical/Hands on	Self-study
Module 1: Quantitative Genetics	1.1		
	1.2	1.2 (Problem set)	
Module 2: Population Genetics	2.1	2.1 (Problem set)	
	2.2	2.2 (Problem set)	
			2.3
Module 3: Phylogenetics		3.1	3.1
		3.2	3.2 (Seminar)
	3.3	3.3 (Group Discussion)	3.3
Module 4: Bioinformatics		4.1	4.1 (Seminar)
		4.2 (Programming)	
		4.3	4.3 (Seminar)
		4.4	4.4 (Seminar)
	4.5	4.5 (Group Discussion)	
Module 5: Multivariate Data Analytics		5.1 (Problem set)	
		5.2 (Group discussion)	5.2
		5.3 (Problem set)	
		5.4	5.4

10. Assessments

Formative assessments: (40 Marks)

1	Internal Exams - 40 marks each (2)	20 M
2	Seminar (2)	8 M
3	Group discussion (2 Including ethical and regulatory issues)	6 M
4	Case studies (2)	6 M

Summative Assessment: (60 marks)

Sl. No.	Details	Q X M
1	Descriptive questions from module 1 and 2 Should also include assessing the knowledge/attitude regulatory/ethical issues	4X5M=20 M
2	Problem based questions (e.g. Solve problems related to population and quantitative genetics such as Hardy-Weinberg Equilibrium, Neutrality Tests, Heritability)	2X10M= 20 M
3	Question on Interpretation of biological data (E.g. interpretation of phylogenetic trees, graphs and multivariate data plots.)	
4	Question related to computational programs (e.g. Write short and relevant program(s) for a given problem; automation of NGS data analysis from multiple sources.)	1X20M=20 M

*A question bank will be maintained with multiple scenarios.

Learning Resources

Reference books:

1. Alberghina L and HV Westerhoff (2005). Systems Biology: Definitions and Perspectives. Springer.
2. Forsdyke D R (2010). Evolutionary Bioinformatics. Springer.
3. Ghosh Z and Mallick B (2011). Bioinformatics – Principles and Applications. Oxford University Press.
4. Hartl D L & Clark A G (2006). Principles of Population Genetics. Oxford University Press.
5. Higgins D and Taylor W (2000). Bioinformatics: Sequence, Structure and Databanks – A Practical Approach. Oxford University Press.
6. Hsiung-Li W (1997). Molecular Evolution. Sinauer Associates, Sunderland, MA, USA.

7. Mount D W (2011). Bioinformatics - Sequence and Genome Analysis. Cold Spring Harbour Laboratory Press.
8. Xiong J (2006). Essential Bioinformatics. Cambridge University Press.
9. National Ethical Guidelines for Biomedical and Health Research involving Human Participants, 2017, Indian Medical Research, 2017.

Other Resources:

- https://www.ncbi.nlm.nih.gov/genbank/
- https://genome.ucsc.edu
- https://m.ensembl.org/index.html
- https://www.cog-genomics.org/plink/
- http://vcftools.sourceforge.net/
- https://blast.ncbi.nlm.nih.gov/Blast.cgi
- https://www.genome.jp/kegg/pathway.html
- http://www.pantherdb.org/
- https://string-db.org/
- https://www.snpedia.com/
- https://www.ncbi.nlm.nih.gov/snp/
- https://regulomedb.org/regulome-search/
- https://www.ebi.ac.uk/gwas/
- https://www.omim.org/
- https://www.internationalgenome.org/
- https://genomeasia100k.org/
- https://reich.hms.harvard.edu/datasets
- http://www.cephb.fr/hgdp/
- https://xenabrowser.net/datapages/
- https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
- https://flybase.org/
- https://www.ncbi.nlm.nih.gov/clinvar/
- https://gnomad.broadinstitute.org/
- https://www.ebi.ac.uk/goldman-srv/webprank/
- https://www.ebi.ac.uk/Tools/msa/clustalo/
- https://spin.niddk.nih.gov/NMRPipe/doc1/

* * * * * * * * *