Savitribai Phule Pune University

(Formerly University of Pune)

Syllabus for M.Phil./Ph.D. (PET) Entrance Exam : Bioinformatics

Research Methodology

1. **Foundation of Research:** Meaning, Objectives, Motivation, Utility. Concept of theory, empiricism, deductive and inductive theory. Characteristics of scientific method - understanding the language of research - Concept, Construct, definition, Variable. Research Process –Steps of research, methods of research.
2. **Problem Identification & Formulation:** definition and formulating the research problem, Necessity of defining the problem, Importance of literature review in defining a problem, Literature survey: primary and secondary; web sources; critical literature review. Research Question - Investigation Question - Hypothesis Testing - Qualities of a good hypothesis - Null hypothesis & Alternative Hypothesis.
3. **Research Design:** Concept and Importance in Research - Features of a good research design - Exploratory Research Design - Concept, Types and uses, Descriptive Research Design - concept, types and uses. Experimental Design - Concept of Independent & Dependent variables. Biased and unbiased research design.
4. **Types ofBiological data and nature of research:** Qualitative data – Quantitative data Concept of measurement, causality, generalization, replication. Basic concepts in qualitative and quantitative data analysis.
5. **Data Collection:** Execution of the research - Observation and Collection of data - Methods of data collection Databases (GenBank, ENA, UniProtKB, PDB, GEO) searching databases, data retrieval, preparing data sets, data curation - Positive and negative data sets - hypothesis-testing - Generalization and Interpretation.
6. **Data Processing and archival:** Basic concepts in programming & scripting (Questions on any specific programming languages will not be asked) – Basic concepts in organization of data into databases
7. **Measurement:** Concept of measurement - what is measured? Problem in measurement in research - Validity and Reliability. Levels of measurement - Nominal, Ordinal, Interval, Ratio.
8. **Sampling:** Concept of Statistical population, Sample, Sampling Frame, Sampling Error, Sample size, Non Response. Characteristics of a good sample. Probability Sample - Simple Random Sample, Systematic Sample, Stratified Random Sample & Multi-stage sampling. Determining size of the sample - Practical considerations in sampling and sample size.
9. **Data Analysis and Interpretation:** Graphical representation of data, Univariate analysis (frequency tables, bar charts, pie charts, percentages), Bivariate analysis - Cross tabulations and Chi-square test including testing hypothesis of association. Correlation and regression analysis.

   Probability distributions: Discrete distribution (bionomial& Poisson) – Continuous distribution (Normal & Exponential)

   Bayesian Modeling Estimation of accuracy (sensitivity specificity, Mathew’s Correlation coefficient)

10. **Paper Writing:** Layout of a Research Paper, Journals, Choice of journals for publication based on various parameters such as impact factor, scope of journal etc. Ethical issues related to publishing, Plagiarism and Self-Plagiarism (basic understanding).
11. **Use of tools / techniques for referencing:** methods to search required information effectively, PubMed, effective literature search using Entrez, Google Scholar
12. **Scientific writing and communication:** Structure and components of scientific reports - Types of report - Technical reports and thesis - Significance - Different steps in the preparation - Layout, Structure and Language of typical reports - Illustrations and tables - Bibliography, referencing and footnotes - Oral presentation - Planning - Preparation - Practice - Making presentation - Use of visual aids - Importance of effective communication. Tools/Software for paper formatting like MSOffice, software for detection of Plagiarism.
13. **Application of results and ethics:** Environmental impacts - Ethical issues - ethical committees - Commercialization - Copy right - royalty - Intellectual property rights and patent law - Trade related aspects of intellectual property Rights - Reproduction of published material - Plagiarism - citation and acknowledgement - citation and acknowledgement - Reproducibility and accountability.
14. **Reasoning and Mental ability**: Logical reasoning, Aptitude, Analogy, Classification, Series, Coding-Decoding, Direction Sense, Representation Through Venn Diagrams, Mathematical Operations, Arithmetical Reasoning, Inserting the Missing Character, Number, Ranking and Time Sequence Test, Eligibility Test**,** Representation through Venn-diagrams, Number & symbols ordering.

Books Recommended

1) Research Methodology: An Introduction - Stuart Melville and Wayne
2) Practical Research Methods - Catherine Dawson
3) Research Methodology - C. R. Kothari Essential Bioinformatics – Jin Xiong (Cambridge University Press)

Additional References

1) Garg, B. L., Karadia, R., Agarwal, F. and Agarwal, U. K., 2002. An introduction to Research Methodology, RBSA Publishers.
2) Sinha, S. C. and Dhiman, A. K., 2002. Research Methodology, EssEss Publications. 2 columes.
3) Trochim, W. M. K., 2005. Research Methods: the concise knowledge base, Atomic Dog Publishing. 270p
4) Wadehra, B. L. 2000. Law relating to patents, trade marks, copyright designs and geographical indications. Universal Law Publishing.

Additional reading

1) Anthony, M., Graziano, A. M. and Raulin, M. L., 2009. Research Methods: A Process of Inquiry, Allyn and Bacon.
2) Carlos, C. M., 2000. Intellectual property rights, the WTO and developing countries: the TRIPS agreement and policy options. Zed Books, New York.
3) Coley, S. M. and Scheinberg, C. A., 1990, "Proposal Writing", Sage Publications.
4) Day, R. A., 1992. How to Write and Publish a Scientific Paper, Cambridge University Press.
5) Fink, A., 2009. Conducting Research Literature Reviews: From the Internet to Paper. Sage Publications
6) Leedy, P. D. and Ormrod, J. E., 2004 Practical Research: Planning and Design, Prentice Hall.
7) Satarkar, S. V., 2000. Intellectual property rights and Copy right. EssEss Publications.

Subject Concerned Syllabus
Bioinformatics

- **Major Bioinformatics Resources: NCBI, EBI, ExPASy, RCSB**
  The knowledge of various databases and bioinformatics tools available at these resources, organisation of databases: data contents and formats, purpose and utility in Life Sciences.
- **Open access bibliographic resources and literature databases:**
  Open access bibliographic resources related to Life Sciences viz., PubMed, BioMed Central, Public Library of Sciences (PloS), CiteXplore.
- **Sequence databases**: Formats, querying & retrieval
  - Nucleic acid sequence databases: GenBank, EMBL, DDBJ
  - Protein sequence databases: Uniprot-KB: SWISS-PROT, TrEMBL, UniParc
  - Repositories for high throughput genomic sequences: EST, STS, GSS, etc.
  - Genome Databases at NCBI, EBI, TIGR, SANGER
- Viral Genomes
- Archeal and Bacterial Genomes.
- Eukaryotic genomes with special reference to model organisms (Yeast, Drosophila, *C. elegans*, Rat, Mouse, Human, plants such as *Arabidopsis thaliana*, Rice, etc.)
- **Structure Databases:**
  PDB, NDB, PubChem, ChemBank
- **Derived Databases**
  Knowledge of the following databases with respect to: basic concept of derived databases, sources of primary data and basic principles of the method for deriving the secondary data, organization of data, contents and formats of database entries, identification of patterns in given sequences and interpretation of the same.
  - Sequence: InterPro, Prosite, Pfam, ProDom
  - Structure: FSSP, DSSP.

- **Extraction of knowledge from resources on**

  Immunology, Plant, animal & infectious diseases: databases & servers published in the NAR Database & Web server Issues and other Bioinformatics journals viz. BMC Bioinformatics etc.
- **Sequence Analysis**
  - Various file formats for bio-molecular sequences: GenBank, FASTA, GCG, MSF etc.
  - Basic concepts of sequence similarity, identity and homology, definitions of homologues, orthologues, paralogues and xenologues.
  - Scoring matrices: basic concept of a scoring matrix, Matrices for nucleic acid and proteins sequences, PAM and BLOSUM series, principles based on which these matrices are derived. Detailed method of derivation of the PAM and BLOSUM matrices.
- **Database Searches:**
  - Keyword-based Entrez and SRS
  - Sequence-based: BLAST & FASTA

- Use of these methods for sequence analysis including the on-line use of the tools and interpretation of results from various sequence and structural as well as bibliographic databases.
- **Pairwise sequence alignments:** basic concepts of sequence alignment, Needleman & Wunsch, Smith & Waterman algorithms (their implementations) for pairwise alignments, gap penalties, use of pairwise alignments for analysis of Nucleic acid and protein sequences and interpretation of results.
- **Multiple sequence alignments (MSA)**: the need for MSA, basic concepts of various approaches for MSA (e.g. progressive, hierarchical etc.). Algorithm of CLUSTALW, MUSCLE, DiAlign and PileUp and their application for sequence analysis (including interpretation of results), concept of dendrogram and its interpretation. Use of HMM-based Algorithm for MSA (e.g. SAM method).
- **Sequence patterns and profiles:** Basic concept and definition of sequence patterns, motifs and profiles, various types of pattern representations viz. consensus, regular expression (Prosite-type) and sequence profiles; profile-based database searches using PSI-BLAST, analysis and interpretation of profile-based searches.

  Algorithms for derivation of & searching sequence patterns: MeMe, PHI-BLAST, ScanProsite & PRATT.

  Algorithms for generation of sequence profiles: Profile Analysis method of Gribskov, HMMer, PSI-BLAST.
- **Taxonomy and phylogeny:** Basic concepts in systematics, taxonomy and phylogeny; molecular evolution; nature of data used in Taxonomy and Phylogeny, Definition and description of phylogenetic trees and various types of trees. Phylogenetic analysis algorithms such as Maximum Parsimony, UPGMA, Transformed Distance, Neighbors-Relation, Neighbor-Joining; Probabilistic models and associated algorithms such as Probabilistic models of evolution and Maximum likelihood algorithm, Bootstrapping methods, use of tools such as Phylip, Mega, PAUP. Whole genome phylogeny,

Alignment-free methods for clustering & phylogeny, Viral and bacterial typing methods

- **Protein and nucleic acid properties**: Computation of various parameters using proteomics tools at the ExPASy server, GCG utilities and EMBOSS.
- **Comparative genomics:** Basic concepts and applications, whole genome alignments: understanding significance. Artemis as an example.

Structural Biology

- **Proteins**: Principles of protein structure; anatomy of globular & membrane proteins – Hierarchical organization of protein structure – Primary, Secondary, Super secondary, Tertiary and Quaternary structure; Hydrophobicity of amino acids, Packing of protein structure, van der Waals and Solvent accessible surface, Internal coordinates of proteins; Derivation, significance and applications of Ramachandran Map, protein folding.
  Identification/assignment of secondary structural elements from the knowledge of 3-D structure of macromolecules using DSSP and STRIDE methods.
- **DNA and RNA**: types of base pairing – Watson-Crick and Hoogstein; types of double helices A, B, Z and their geometrical as well as structural features; structural and geometrical parameters of each form and their comparison; various types of interactions of DNA with proteins, small molecules.
  RNA secondary and tertiary structures, t-RNA tertiary structure.
- **Carbohydrates:**
  The various building blocks (monosaccharides), configurations and conformations of the building blocks; formations of polysaccharides and structural diversity due to the different types of linkages.
  Glyco-conjugates: various types of glycolipids and glycoproteins.
- **Structure analysis & validation:** Procheck, ProsaII, PDBsum
- **3-D structure visualization and simulation:**
  - Visualization of structures using Rasmol or SPDBV or CHIME or VMD.
  - Basic concepts in molecular modeling: different types of computer representations of molecules. External coordinates and Internal Coordinates
  - Concepts of force fields: representations of atoms and atomic interactions, potential energy representation.
- **Classification and comparison of protein 3D structures**:
  - Purpose of 3-D structure comparison and concepts, Algorithms such as FSSP, CE, VAST and DALI, Fold Classes.
  - Databases of structure-based classification: CATH and SCOP

- **Secondary structure prediction**: Algorithms viz. Chou Fasman, GOR methods; analysis of results and measuring the accuracy of predictions using Q3, Segment overlap, Mathew's correlation coefficient.
  Ph.D. and PSI-PRED methods.
- Structures of oligomeric proteins and study of interaction interfaces
- **Tertiary Structure prediction**: Fundamentals of the methods for 3D structure prediction (sequence similarity/identity of target proteins of known structure,

fundamental principles of protein folding etc.) Homology Modeling, fold recognition, threading approaches, and a*b-initio* structure prediction methods.
- Detailed protocols/algorithms for Homology modeling, fold recognition and *ab-initio* approaches

Genomics

- Genome mapping Physical and genetic mapping tools and markers
- Next generation genome sequencing platforms: instrumentation and chemistry
- Sequence data file formats, Quality checks, polishing the sequence files.
- Short read alignment algorithms, file formats and conversions.
- Genome assembly (*de novo* and reference guided) and annotation.
- Computational genomics approaches to address research questions: ChIP-Seq data analysis, Bisulfite Sequencing data analysis, Exome and RNA sequencing data analysis, differential gene expression analysis etc.
- NGS data archival: SRA@NCBI, ENA-EMBL-EBI, Encode
- Genome databases and browsers for Plants, animals and pathogens
- Metagenomics, computational resources available for metagenomics data analysis – Protocols for characterization of & diversity of metagenomes, methods for functional characterization of metagenomes.
- Gene networks: basic concepts, computational model such as Lambda receptor and *lac* operon.
- Genome Annotation process: Importance of annotations, multilevel nature of annotation process, genomic features, their significance and application of bioinformatics tools for annotations of the features.
- Homology-based and *ab initio* methods for prediction of genes (GLIMMER, GenScan, Augustus, GenScan); Methods for prediction of promoters, splice sites, regulatory regions: basic principles; applications of methods to prokaryotic and eukaryotic genomes and interpretation of results.
- Basic concepts on identification of disease genes, role of Bioinformatics- OMIM database, reference genome sequence, integrated genomic maps, gene expression profiling; identification of SNPs, SNP database (DbSNP). Role of SNP in Pharmacogenomics, SNP arrays.
- Basic concepts in identification of Drought stress response genes, insect resistant genes, nutrition enhancing genes
- Computational Epigenetics: Databases and methods to predict epigenetic information/features from genome sequence
- DNA microarray: databases and basic tools, Gene Expression Omnibus (GEO), ArrayExpress, SAGE databases.
- DNA microarray: understanding of microarray data, normalizing microarray data, detecting differential gene expression, correlation of gene expression data to biological processes and computational analysis tools (especially clustering approaches).
- **Comparative genomics:**
- Basic concepts and applications, BLAST2, MegaBlast algorithms, PipMaker, AVID, Vista, MUMmer, applications of Suffix tree in comparative genomics, synteny and gene order

comparisons, Population structure, recombination & selection pressure analysis using genomes.

⍰ **Functional genomics:**

o Application of sequence based and structure-based approaches to assignment of gene functions - e.g. sequence comparison, structure analysis (especially active sites, binding sites) and comparison, pattern identification, etc. Use of various derived databases in function assignment, use of SNPs for identification of genetic traits.

o Gene/Protein function prediction using Machine learning tools viz. Neural network, SVM etc

⍰ **Proteomics**

o Protein arrays: basic principles.

o Computational methods for identification of polypeptides from mass spectrometry

o Protein arrays: bioinformatics-based tools for analysis of proteomics data (Tools available at ExPASy Proteomics server); databases (such as InterPro) and analysis tools.

o Protein-protein interactions: databases such as DIP, PPI server and tools for analysis of protein-protein interactions

⍰ **Modeling biological systems**

Syllabus for Entrance Examination (PET) for admission to Ph.D. and M. Phil in Bioinformatics. (Revised in March, 2017)

o Systems biology – Use of computers in simulation of cellular subsystems

o Metabolic networks, or network of metabolites and enzymes

o Metabolic pathways: databases such as KEGG, EMP

o Study of plant pathways –MetaCyc, AraCyc

o Signal transduction networks

o Gene regulatory networks

⍰ **Bioinformatics Resources at the species level**

o ICTV Database, Viral genomes at NCBI, VBRC, VBCA, PBRC and Subviral RNA database, Species 2000, TreeBASE etc.

⍰ **Vaccine design:**

o Reverse vaccinology & immunoinformatics

o Databases in Immunology

o B-cell epitope prediction methods

o T-cell epitope prediction methods

o Resources to study antibodies, antigen-antibody interactions

⍰ **Cheminformatics and Drug design**

o 1D, 2D and 3D representation of chemical compounds

o Substructure, superstructure and similarity searching

o QSAR & descriptors used for the same

o Pharmacophore modelling

o Properties of drug like molecules

o High-throughput and fragment based screening

o Target identification and validation, Polypharmacology, Lead optimization and validation, drug repurposing o Docking and forcefields

o ADMET and computational predictions

o Pharmacogenomics

 Molecular modeling and simulations

o Force fields

o Energy Optimization algorithms

o Molecular dynamics simulations

o Electrostatics

 Bioinformatics Methods based on Mathematical concepts and algorithms:

o Functions and Graphs: Functions, Relations, notation and representation. Graphs. Review of basic functions. Functions of several variables.

o 2D coordinate geometry: Equation of a line, circle, ellipse, parabola, hyperbola

o 3D geometry: Equation of sphere, cone, direction cosines, equation of line.

o Basic trigonometric functions.

o Matrix algebra: Addition, subtraction, multiplication, transpose.

o Numerical integration. Interpolation and approximate methods.

o Vector - addition, subtraction, dot, cross, scalar triple product, divergence and curl.

o System of linear equations. Matrix inverse, eigen value, eigen vector, principal component analysis

o Mathematical modeling and simulation

Syllabus for Entrance Examination (PET) for admission to Ph.D. and M. Phil in Bioinformatics. (Revised in March, 2017)

 Bioinformatics Methods based on Statistics theories and algorithms:

o Principles of statistical sampling from a population, random sampling

o Frequency distributions and associated statistical measures, Probability distributions – normal and binomial.

o Methods of least squares, chi-square test, systematic and random sampling, accidental and systematic errors, correlation and regression analysis. Poisson and extreme value distributions.

o Multivariate analysis, Hypothesis testing, Markov process.

o Bayesian Statistics and applications in Bioinformatics