MASTER OF SCIENCE IN BIOINFORMATICS (MSc BIOINFORMATICS)

PROGRAMME STRUCTURE

SEMESTER 1: CORE COURSES

Module	Course Code	Course Title	T	P	Credit
1	BINF601	Principles of Bioinformatics	3	0	3
2	BINF603	Biocomputing - Python, Linux	1	2	3
3	BINF605	Statistical computing using R	1	2	3
4	BINF607	Structural Bioinformatics I	2	1	3
5	BINF609	OMICS I: NGS technologies and analysis tools	1	2	3
					15

SEMESTER 1: ELECTIVE

Module	Course Code	Course Title	T	P	Credit
7	BINF611	Molecular informatics	2	1	3
Sub-total for coursework					**15 - 18**

SEMESTER 2: CORE COURSES

Module	Course Code	Course Title	T	P	Credit
8	BINF602	OMICS II: Metagenomics	2	1	3
9	BINF604	Structural Bioinformatics II	2	1	3
10	BINF606	Biological databases	2	1	3

SEMESTER 2: ELECTIVES*

Module	Course Code	Course Title	T	P	Credit
12	BSTT601	Methods in Biostatistics	3	0	3
13	BINF612	Proteomics	2	1	3
14	BINF614	Machine Learning	2	1	3
Sub-total for coursework					**15 – 18**

* Students must take a minimum of 2 elective courses.

6	BINF610	Seminar I			3
11	BINF600	Project			6
TOTAL					**39 - 45**

Course Description

CORE COURSES

BINF 601: Principles of Bioinformatics

The course content covers the central dogma of molecular biology, interdisciplinary aspects of bioinformatics, and its application in understanding complex diseases. Students will learn about big biological data, sequence retrieval from databases, sequence alignment principles, and the Basic Local Alignment Search Tool (BLAST). Develop an understanding of phylogenetic analysis and interpretation.

Course Objectives

The overall objective of the course is to discover the multidisciplinary realm of bioinformatics in biomedical research.

- Introduce students to bioinformatics as a multidisciplinary field in biomedical research
- Initiate discussions on the applications of big data in biomedical science research, their potential benefits and challenges

Course Content

The major topics to be covered include the concept of big biological data; application of bioinformatics in biomedical research (infectious and non-infectious); sequence retrieval from biological databases and principles of sequence alignment and Basic Local Alignment Search Tool (BLAST); and phylogenetic analysis and interpretation.

Learning Outcome

It is expected that at the end of the course, students should be able to:

- *Define* bioinformatics as a multidisciplinary field in biomedical research and identify its key components.
- *D*iscuss and analyse the applications of big data in biomedical science research, demonstrating an understanding of their potential benefits and challenges.
- *Proficiently* conduct effective database *queries*, retrieving relevant sequences, and critically interpreting the results.
- Independently utilize BLAST and phylogenetic analysis tools, demonstrating their understanding of their purpose, functionality and proper application.

Mode of delivery

This will be delivered through a combination of team-taught lectures, which will introduce students to the theory and concepts, and hands on practicals which will reinforce the principles learnt in class. There will also be tutorials and journal clubs during the course to discuss current bioinformatics research papers.

Reading List

Byron, K., Herbert, K., & Wang, J. (2016). *Bioinformatics database systems* (1st ed.). CRC Press.

Higgs, P. & Attwood, T.K. (2005). *Bioinformatics & Molecular evolution.* Blackwell: Oxford, UK. ISBN: 1405106832

Kazutaka Katoh., (2020) Multiple Sequence Alignment: Methods and Protocols. Humana. [2231, 1 ed.].
Lesk, M. A. (2014). *Introduction to Bioinformatics* (4th ed). Oxford University Press: Oxford, UK.
Paul S. Ganney., (2023) Introduction to Bioinformatics and Clinical Scientific Computing. CRC Press.

BINF 603: Biocomputing -Python, Linux

This course provides students with a comprehensive foundation in biocomputing, focusing on the utilization of Python programming language and Linux environment for effective analysis of biological data. Students will learn the essentials of Python, including data types, functions, control flow, file processing, modules, and scientific computing libraries. The Linux component covers topics such as file systems, permissions, shell scripting, text manipulation, and software installation. By the end of the course, students will acquire the skills to load and analyze biological data, develop automated bioinformatics pipelines, simulate biological processes, navigate the Linux environment, manipulate large text files, and deploy bioinformatics tools for data analysis.

Course Objectives

Overall, this course is intended to provide students a foundation in computer programming for big data analysis in biological science applications. Specifically, the course will:

- Introduce students to Python programming language and Linux environment and their applications to biological science.
- Enhance proficiency of students in efficient handling and manipulation of big datasets in biomedical research using Python and Linux.
- Introduce students to automation and development of analyses pipelines for biomedical data.

Course Content

Major topics that will be covered on Python include basic data types, functions (structure, arguments, and parameters), conditional and Boolean expression, iteration, loops, control flow, file and text processing, exceptions, module, and their imports class, methods and class instances and regular expressions. Scientific Python (SciPy) stack together with NumPy, Matplotlib and pandas are some of the packages required for scientific computing. For Linux, topics include introduction to the Linux operating system and its history and distributions, Linux file system and tree structure; files, file permission and security; User environment, command-line operations (Basic operations, searching files and manipulating files), shell scripting (basic editors, shell variables, basic syntax, and features), text and string manipulation (Modifying files, grep command, file manipulation utilities).

Learning Outcome

At the end of this module, it is expected that students will be able to:

- Use Python to load and analyse text files containing biological data.
- Develop analyses pipelines in python to automate bioinformatics tasks.
- Simulate biological processes such as DNA replication, transcription and translation using python.
- Navigate through the Linux environment.
- Perform pattern searches, understand file permission, and manipulate large text files in Linux.

- Install and deploy bioinformatics tools and software on Unix platforms for biological data analysis.

Mode of delivery
The course will be delivered through team-taught lectures, to introduce students to concepts of programming. The lecture will be supported by tutorials and hands-on practicals to apply all the theory covered in class. Furthermore, students will be given programming assignments to test their understanding of the material covered. Demonstrations will be done mostly with biological data types.

Reading List

Barrett, J. D. (2018). *Linux Pocket Guide* (2nd ed). O'Reilly Media, Inc: California, USA. ISBN:1449316697

Chang J., Chapman, B., Friedberg, I., Hamelryck, T., de Hoon, M., Cock, P., Antao, T., Talevich, E., Wilczynski, B. (2014). *Biopython Tutorial and Cookbook*. http://www.biopython.org/

Crawley, M. J., (2013). *The R Book* (2nd ed). Wiley and Sons Publications: USA

Ken Youens-Clark., (2021) Mastering Python for Bioinformatics. O'Reilly Media. ISBN: 9781098100889

Martin J., (2013). *Python for Biologists*. CreateSpace Independent Publishing Platform: California, USA. ISBN: 1492346136.

Wickham, H., & Grolemund, G., (2016). *R for Data Science.* O'Reilly Media, Inc: California, USA

BINF 605: Statistical computing with R
This course will provide students with a comprehensive understanding of statistical computing techniques using the R language for effective analysis and visualization of biological data. Students will learn the fundamentals of biological statistical computing, including data loading and exploration, pattern recognition, quality control checks, and generating publication-worthy graphs. The course also introduces students to Bioconductor and its packages for statistical genomics. Topics covered include working with various data types, manipulating data objects, reading and writing files, loops and conditionals, creating and using functions, and graphical representations using ggplot2. Students will gain practical experience in statistical analysis using R and conducting genomics analysis using Bioconductor packages.
Course Objectives
The objectives of the course include:
- Explanation of the fundamentals of biological statistical computing using the R language and how to load and explore biological data
- Exploring pattern recognition and performance of quality control checks.
- Using R to generate publication worthy graphs and install and use Bioconductor packages for statistical genomics.

Course Content

The course covers statistical computing using the R statistical environment for biological data exploration and visualization, and introduces students to Bioconductor and packages for statistical genomics. Major topics to be covered include: Introduction to the R statistical environments; basic statistics and interactively using R; installing and working with packages; where and how to get help; data (objects) types, their mode and their manipulation; vectors, factors, working with array matrices, data frames, matrices and list; reading and writing to files; loops and conditionals; writing and using functions; graphical representations (ggplot2); Introduction and installation of Bioconductor; working with genomic datasets in Bioconductor and Bioconductor data objects and structure.

Learning Outcome

The learning outcome of this module will include:

- Load biological data into R and perform exploratory analysis to observe patterns.
- Performs statistical biological data analysis in the R statistical environment.
- Produce publication worthy graphs/figures in the R statistical environment.
- To install Bioconductor packages in R and perform simple genomics analysis.
- To manipulate, modify and extract features of interest in a Bioconductor object.

Mode of delivery

The course will mainly be delivered through didactic team-taught lectures supported by hands-on statistical analysis practice in R and tutorials to reinforce the understanding of the materials covered. These will include exercises done in class.

Reading List

Crawley, M. J., (2013). *The R Book* (2nd ed). Wiley and Sons Publications: USA

Curry, E. (2020). *Introduction to Bioinformatics with R: A Practical Guide for Biologists (Chapman & Hall/CRC Computational Biology Series)* (1st ed.). Chapman and Hall/CRC.
Mathur, S. K. (2009). *Statistical Bioinformatics with R* (1st ed.). Academic Press.

Wickham, H., & Grolemund, G., (2016). *R for Data Science.* O'Reilly Media, Inc: California, USA

BINF 607: Structural Bioinformatics I

In this course, students will explore the fascinating world of protein structures and their functional implications. The course emphasizes the interdisciplinary nature of structural bioinformatics, integrating principles, concepts, and computational techniques from structural biology and bioinformatics. Students will gain an understanding of protein structure organization, the relationship between structure and function, and the role of proteins in cellular mechanisms. The course covers protein structure determination techniques such as X-ray crystallography, nuclear magnetic resonance spectroscopy, and cryo-electron microscopy. Students will become proficient in querying protein sequences, motifs, domains, and structures using online databases. They will also learn visualization and manipulation of protein structures, as well as modeling and molecular dynamics simulations using bioinformatics tools.

Course Objectives

The following are the course objectives:

- Description of how protein structures relate to function.

- Exploring the principles of protein structure organization
- Familiarizing students with online databases for querying protein sequences, motifs, domains, and structures.
- Exploring protein structure elucidation techniques such as X-ray crystallography, nuclear magnetic resonance spectroscopy and cryo-electron microscopy,
- Visualization and manipulation of protein structures, and modelling of protein structures and molecular dynamic simulations.

Course Content

The course adopts an interdisciplinary approach by integrating various principles, concepts and computational techniques underlying structural biology and structural bioinformatics. It covers the elucidation of the three-dimensional structures of proteins and their relationships to sequences, functions, and roles in cellular mechanisms. Topics to be addressed include principles of protein structure, structural molecular biology, structural databases, and molecular visualization. Protein structure determination techniques involve X-ray crystallography, nuclear magnetic resonance spectroscopy, cryo-electron microscopy, and molecular modelling. Databases to be utilized include Protein Data Bank, SCOP and CATH, whilst some of the applications are MODELLER, PyMOL and visual molecular dynamics (VMD).

Learning Outcome

At the end of this module, students should:

- Be able to apply the principles, concepts and theories underlying structural biology to protein structures and functions.
- Use computational techniques for amino acid sequence searches, alignments, homology, sequence motif identification, and prediction of secondary and tertiary structures.
- Be able to evaluate and interpret the quality of experimentally determined protein structures.
- Be able to model protein structures using plethora of bioinformatics tools and undertake molecular dynamics simulations.
- Be able to apply structural bioinformatics techniques in rational drug design, vaccine development and diagnostic biomarker discovery.

Mode of delivery

The lectures will cover the theories, concepts and principles underlying structural bioinformatics. It will involve hands-on practical tutorials, paper discussions as journal club, and student presentations. The course delivery includes team-taught lectures, tutorials, practical exercises, and assignments.

Reading List

Engel, T. & Gasteiger, J. (2018). *Applied chemoinformatics: achievements and future opportunities*.Wiley-VCH: Weinheim, Germany.

Kihara, D. (2017). *Protein function prediction: methods and protocols*. Humana Press Springer: New York, N.Y

Nurit H, Filip J, Kevin Molloy., (2022) Algorithms and Methods in Structural Bioinformatics. Springer. ISBN 3031059131, 9783031059131

Senda, T. & Maenaka, K. (2016). *Advanced methods in structural biology*. Springer: Japan.

Wei, D., Qin., Zhao, T. & Dai, H. (2015). *Advances in structural bioinformatics*. Springer: Dordrecht.

BINF 609: OMICS I: NGS Technologies and analysis tools

This course provides a comprehensive overview of omics technologies and analysis tools, with a focus on Next Generation Sequencing (NGS) platforms. Students will gain an understanding of the theory and practical aspects of genomic library preparation, data quality control checks, and NGS analysis procedures. The course equips students with the skills to critically evaluate NGS analyses, from genome mapping to SNP calling, and perform differential expression analysis using NGS data. Students will learn to interpret and communicate NGS analysis results effectively for scientific publication.

Course Objective

The following are the course objectives:

- Introduction to different NGS platforms.
- Explanation of genomic library preparation theory and data quality control checks.
- Equipping students with the skills to critically evaluate existing NGS analyses procedures from genome mapping to SNP calling.
- Performing differential expression analysis using NGS and learn to critically evaluate results from NGS analysis for scientific publication.

Course Content

Omics technologies are increasingly being used in biomedical research and generated datasets are often large, complex, and difficult to analyse and interpret. This module will cover fundamentals of next generation technologies, generated data types and complexity, and their analysis from whole genome sequencing to transcriptomics. Major topics include Introduction to sequencing technology platforms (from Sanger sequencing to next generation sequencing, NGS); NGS data types, format, and management; quality control of sequencing reads and trimming; reference-based mapping and *de novo* assemblies and visualisation tools; variant calling and filtering; disease association analysis (candidate gene and genome-wide); transcriptomic analysis (whole RNA-Seq and scRNA-Seq) and pathway analysis and pathway biology.

Learning Outcome

At the end of this module, students will be able to:

- Critically evaluate and explain the steps of quality control for NGS experiments.
- Map NGS raw data to the canonical genomes, visualise and thoroughly describe the sequencing results.
- Perform SNP calling and other genomic analysis, filter variants, and critically explain and communicate results.
- Perform de novo genome assembly and quality control on genome assemblies.
- Perform NGS transcriptional analysis and critically evaluate genomic and transcriptomic approaches.
- Critically discuss results from bioinformatic analysis for scientific publication.

Mode of delivery

This course will be delivered through team-taught lectures to cover theory and concepts of next generation sequencing technologies and analysis methods. This will be supported with hands-on tutorials on NGS data analysis and journal club presentations of current papers in the field.

Reading List

Almagro C. (2019). *Nasopharyngeal Microbiota in Children with Invasive Pneumococcal Disease: Identification of Bacteria with Potential Disease-Promoting and Protective Effects.* Frontiers Microbiology; 10:11. doi: 10.3389/fmicb.2019.00011

Biesbroek G., Tsivtsivadze E., Sanders E.A., Montijn R., Veenhoven R.H., Keijser B. J. (2014). *Early respiratory microbiota composition determines bacterial succession patterns and respiratory health in children.* Am J Respir Crit Care Med.;190(11):1283–92. doi:10.1164/rccm.201407- 1240OC.

Camelo-Castillo A, Henares D, Brotons P, Galiana A, Rodriguez JC, Mira A and Muñoz-Poretsky R, Rodriguez-R LM, Luo C, Tsementzi D., & Konstantinidis K.T. (2014*). Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics.* PLoS ONE 9(4): e93827.

Sharpton, J. T. (2014). *Introduction to the analysis of shotgun metagenomic data.* Review article, Frontiers in Plant science. doi: 10.3389/fpls.2014.00209

Richardson, J. E., & Mick Watson, M. (2013). *The automatic annotation of bacterial genomes* *Briefings in Bioinformatics*, Volume 14, Issue 1, Pages 1–12, https://doi.org/10.1093/bib/bbs007

BINF 611: Molecular informatics

This course will provide an interdisciplinary exploration of molecular informatics, focusing on cheminformatics, computer-aided drug design, and pharmacological profiling. Students will gain insights into repurposing existing drugs, understanding polypharmacology, and navigating hits-to-leads generation. The course will delve into the diverse scaffolds found in synthetic and natural product libraries, employing biomolecular networks and chemogenomics. Through the utilization of bioinformatics tools and databases like DrugBank, ZINC, PubChem, and Chembl, students will learn to manipulate and analyze chemical structures, calculate molecular descriptors, and evaluate pharmacological properties. Engaging lectures, hands-on tutorials, journal club discussions, student presentations, and project work will empower students in the application of rational drug design techniques.

Course Objective

The objectives of this course include:

- Repurposing existing drugs and polypharmacology
- Describing the use of computer-aided drug design to explain the concepts of hits-to-leads.
- Exploring the diverse scaffolds of both synthetic and natural product libraries.
- Using different bioinformatics tools and databases for cheminformatics, and principles underlying pharmacological profiling of compounds.

Course Content

The course adopts an interdisciplinary approach by integrating various principles and concepts underlying cheminformatics, quantitative structure-activity relationships, combinatorial chemistry, and computer-aided molecular/drug design to explain hits-to-leads generation. The course will also cover the design and prediction of novel small molecules of therapeutic

interest, as well as incorporating polypharmacology and repurposing of existing drugs. Natural products are also considered since they are structurally and chemically diverse and serve as useful baseline scaffolds for drug discovery. Approaches involve biomolecular networks and systems, chemogenomics and pharmacological profiling. The databases and tools include DrugBank, ZINC, PubChem, Chembl, AfroDB, NANPDB, SANC, AnalytiCon Discovery and Open Babel.

Learning Outcome

At the end of this module, students should:

- Be familiar different approaches for prediction of protein biological activities and cellular mechanisms.
- Be able to manipulate, store and retrieve chemical structures and calculate various physicochemical properties.
- Be familiar with representation and manipulation of 1D, 2D and 3D molecular structures, as well as computation of molecular descriptors.
- Be able to compute pharmacological properties such as absorption, distribution, oral bioavailability and "drug-likeness".
- Be able to apply cheminformatics techniques in rational drug design.

Mode of delivery

The lectures will cover the theories, concepts and principles underlying molecular informatics. It will involve hands-on practical tutorials, paper discussions as journal club, and student presentations. The course delivery includes team-taught lectures.

Reading List

Bajorath, J. (2014). *Chemoinformatics for drug discovery*. Wiley & Sons Publication: Hoboken, New Jersey

Brown, J. (2018). *Computational chemogenomics*. Humana Press Springer: New York, N.Y.

Engel, T. & Gasteiger, J. (2018). *Applied chemoinformatics: achievements and future opportunities*. Wiley-VCH: Weinheim, Germany.

Engel, T. & Gasteiger, J. (2018). *Chemoinformatics: basic concepts and methods*. Wiley-VCH: Weinheim, Germany.

Guha, R. & Bender, A. (2012). *Computational approaches in cheminformatics and bioinformatics*. Wiley: Hoboken, N.J.

Jacoby, E. (2013). *Computational chemogenomics*. Pan Stanford Publishing: Boca Raton.

BINF 602: OMICS II

This course delves into the processing and analysis of data generated using high throughput omics technologies including next-generation sequencing (NGS) and mass spectrometry-based proteomics. The course covers data generating technologies including but not limited to short read Illumina NGS platforms (MiSeq, NextSeq 2000, NovaSeq etc) and long read Oxford Nanopore Technologies (GridION, MinION), types of data produced, processing steps used to prepare the raw data from these platforms for downstream analysis. Data processing steps including methods for QC of raw reads, mapping reads reference assemblies, for de novo

assembly of genomes and transcriptomes, for analysis of alternative splicing, genomic variants, and for expression level estimation. Furthermore, the course covers approaches for detection of differential expression at the level of individuals genes/proteins as well as of pathways/systems using various techniques, for example over-representation analysis. Data sources will include genetic material recovered directly from clinical and environmental (water, soil, food etc) sources. Methods for profiling the taxonomic structure of microbes and viruses and function of microbial communities from high throughput microbial community multi-omics data, alongside statistical methods for associating those profiles with community metadata (e.g host phenotypes). Student will focus on analyzing shotgun (or 16S RNA seq) metagenomes, metatranscriptomes using the bioBakery - a computing environment containing reproducible methods for microbial community analysis.

Course Objective

The overall objective of the course is to enhance the understanding of students in exploring 'omics' data from high throughput Next-generation sequencing. Specifically, on completion of the course student should knowledge, skills and general competences relating to:

- a deep understanding of key technologies used to generate omics data.
- understanding and able to implement methods for quality control, filtering and normalization of NGS and proteomics data.
- understanding and able to implement methods for sequence mapping, and for de novo assembly of genomes and transcriptomes, and detection and annotation of genome variants.
- understanding and able to implement methods for protein identification and analysis post-translational modifications.
- Understanding and able to implement methods for analysis of differential expression and over representation analysis.

Course Content

Next-generation sequencing (NGS) technologies have greatly impacted the field of 'omics', which is the crossover application of multiple high-throughput screening technologies represented by genomics, metagenomics transcriptomics, single-cell transcriptomics, proteomics, and so on. Omics provides multiple approaches to power discovery across multiple levels of host and pathogen biology. This course will focus on grounding the students in theory and practical bioinformatic skills, including project design and planning, sample quality control, specific omics sequencing workflows for data generation using ONT and Illumina platforms and downstream data processing and practical bioinformatics. This course will focus on omics data generation (library prep, sequencing), types of data formats, read processing and QC, read mapping and variant calling, annotation, and downstream analysis. Microbial metagenomics analysis be covered in this module include, shotgun sequence processing and quality control, methods of taxonomic assignment and clustering of targeted gene data; assembly, functional classification, and characterization of shotgun metagenomic data; tools for estimating microbial diversity; and microbial community comparison methodology and metrics. Overall, omics data generation from several sources including patient environmental sources poses a challenge for data integration, sharing with researcher worldwide. Therefore, there is need for omics data to be made more Findable, Accessible, Interoperable and Reusable (FAIR) for humans and machines. This course will highlight the FAIR principles and its relevance to omics data sharing within the broader

scientific community. Principles of data sharing will be covered more broadly in our introductory courses on research ethics.

Learning Outcome

At the end of this module, students should be able to:

- apply state of the art tools to analyze genome sequencing and proteomics data sets.
- conceptualize and design omics-based experiments to address specific biological questions.
- demonstrate understanding of multi-omics and its application in research.
- perform raw read processing and QC to address specific omics questions.
- Critically interpret results from omics studies for a scientific publication.

Mode of delivery

This course will be delivered through team-taught lectures to cover theory and concepts of omics. These will be supported with hands-on practical using freely available datasets and journal club presentations of current papers in the field.

Reading List

Camelo-Castillo A, Henares D, Brotons P, Galiana A, Rodriguez JC, Mira A and Muñoz-Almagro C. (2019). *Nasopharyngeal Microbiota in Children with Invasive Pneumococcal Disease: Identification of Bacteria with Potential Disease-Promoting and Protective Effects.* Frontiers Microbiology; 10:11. doi: 10.3389/fmicb.2019.00011

Hamady M, Knight R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res. 2009 Jul;19(7):1141-52. doi: 10.1101/gr.085464.108. Epub 2009 Apr 21. PMID: 19383763; PMCID: PMC3776646.

Calle ML. Statistical Analysis of Metagenomics Data. Genomics Inform. 2019 Mar;17(1):e6. doi: 10.5808/GI.2019.17.1.e6. Epub 2019 Mar 31. PMID: 30929407; PMCID: PMC6459172.

Santos A, Colaço AR, Nielsen AB, Niu L, Strauss M, Geyer PE, Coscia F, Albrechtsen NJW, Mundt F, Jensen LJ, Mann M. A knowledge graph to interpret clinical proteomics data. Nat Biotechnol. 2022 May;40(5):692-702. doi: 10.1038/s41587-021-01145-6. Epub 2022 Jan 31. PMID: 35102292; PMCID: PMC9110295.

Jackson AP, Gamble JA, Yeomans T, Moran GP, Saunders D, Harris D, Aslett M, Barrell JF, Butler G, Citiulo F, Coleman DC, de Groot PW, Goodwin TJ, Quail MA, McQuillan J, Munro CA, Pain A, Poulter RT, Rajandream MA, Renauld H, Spiering MJ, Tivey A, Gow NA, Barrell B, Sullivan DJ, Berriman M. Comparative genomics of the fungal pathogens Candida dubliniensis and Candida albicans. Genome Res. 2009 Dec;19(12):2231-44. doi: 10.1101/gr.097501.109. Epub 2009 Sep 10. PMID: 19745113; PMCID: PMC2792176.

BINF 604: Structural Bioinformatics II

This course delves into the intricacies of structural bioinformatics, equipping students with the knowledge and skills required to analyze protein structures using quantitative structure-activity

relationship concepts. Students will engage with a range of molecular graphics and modeling tools, gaining practical experience in their application. The course encompasses diverse molecular modeling techniques, including homology modeling, threading/fold recognition, and ab-initio methods. The concepts underlying molecular docking and virtual screening techniques will be explored, along with atomistic molecular dynamics simulations of protein complexes using GROMACS. Throughout the course, students will develop expertise in predicting protein functions, performing molecular modeling, evaluating docking tools, and applying structural bioinformatics in drug design, vaccine development, and biomarker discovery.

Course Objective

The specific objectives of this course are to:

- Explore the quantitative structure-activity relationship concepts in protein structures
- Equip students in the practice and use of a variety of molecular graphics and modelling tools.
- Introduce students to the application of different molecular modelling techniques including homology modelling, threading/fold recognition and ab-initio methods.
- Explain concepts underlying molecular docking and virtual screening techniques
- Describe the atomistic molecular dynamics simulations of protein complexes.

Course Content

This course integrates concepts, principles, and computational methods in structural bioinformatics, including quantitative structure-activity relationships, molecular modelling, molecular docking, and molecular dynamic simulations of proteins. Molecular modelling techniques involve homology modelling, threading/fold recognition and ab-initio methods. The molecular docking techniques include structure- and ligand-based virtual screening, as well as pharmacophore modelling using LigandScout and AutoDock. The validating approaches for the various docking protocols include area under the curve of receiver operating characteristic curve and enrichment factor. Atomistic molecular dynamics simulations of protein complexes involve conformational and mechanistic analysis using GROMACS.

Learning Outcome

At the end of this module, students should be:

- Familiar with different approaches for prediction of biological activities and functions of proteins.
- Able to use computational tools to perform different molecular modelling techniques such as homology modelling, threading/fold recognition and ab-initio methods
- Able to perform different molecular docking techniques including structure- and ligand-based methods, as well as pharmacophore modelling
- Able to evaluate molecular docking tools by using metrics such as enrichment factors (EFs), receiver operating characteristics (ROC) curves, the area under the ROC curve (ROC AUC) and the Boltzmann-enhanced discrimination of ROC (BEDROC).
- Familiar with the theoretical basis of molecular dynamics simulations and associated computational cost.
- Able to apply structural bioinformatics techniques in rational drug design, vaccine development and diagnostic biomarker discovery.

Mode of delivery

The lectures will cover the theories, concepts and principles underlying structural bioinformatics. It will involve hands-on practical tutorials, paper discussions as journal clubs, and student presentations. The course delivery includes team-taught ectures.

Reading List

Engel, T. & Gasteiger, J. (2018). *Applied chemoinformatics: achievements and future opportunities*. Wiley-VCH: Weinheim, Germany.

Senda, T. & Maenaka, K. (2016). *Advanced methods in structural biology*. Springer: Japan.
Wei, D., Qin., Zhao, T. & Dai, H. (2015). *Advances in structural bioinformatics*. Springer: Dordrecht.

Nurit H, Filip J, Kevin Molloy., (2022) Algorithms and Methods in Structural Bioinformatics. Springer. ISBN 3031059131, 9783031059131

Kihara, D. (2017). *Protein function prediction: methods and protocols*. Humana Press Springer: New York, N.Y

BINF 606: Biological databases

In this course, students will gain a comprehensive understanding of biological databases and their applications in bioinformatics. They will learn how to retrieve data from various databases, including pathogen-specific genomics databases and gene expression data repositories. Topics covered include an introduction to databases (curated versus noncurated), DNA and protein databases, pathogen-specific databases (e.g., EuPathDB), gene expression data repositories (ArrayExpress, Gene Expression Omnibus), pathway biology databases (InnateDB, STRING, Reactome, DAVID), and pathway interpretation. Students will also be introduced to the National Center for Biotechnology Information (NCBI) and learn procedures for literature search, gene retrieval, and genome retrieval. The course delivery includes didactic lectures, hands-on demonstrations, and practical exercises for retrieving biological information, performing gene expression analysis, and interpreting pathway and protein interaction networks.

Course Objectives

This course has the following objectives:

- Introduction to biological databases and how to retrieve data from them.
- Introduction to pathway and network analyses and how perform these analyses.
- Exploring pathogen specific databases relevant to biomedical research.
- Introduction to gene expression data repositories, dataset retrieval and using online tools for analysis.

Course Content

This will cover some of the major biological databases used in bioinformatics, ranging from pathogen-specific genomics databases to transcriptomics databases. Topics to be covered include Introduction to databases (curated versus noncurated databases); DNA databases; Protein databases; pathogen specific databases (E.g., EuPathDB); gene expression data repositories and databases (Array express, Gene expression Omnibus); gene expression analysis in data repositories and pathway biology databases (InnateDB, STRING, Reactome, DAVID) and pathway interpretation. Furthermore, students will be introduced to NCBI, including procedures for literature search, and gene and genome retrievals.

Learning Outcome
At the end of the module students are expected to be able to:

- Search and retrieve biological data from biological databases.
- Critically explain the use of biological databases in medical research.
- Perform gene expression analysis using available online databases.
- Perform and interpret results of pathway analyses of biological data.
- Perform, interpret, and communicate network of protein interaction.

Mode of delivery
The course will be delivered mainly through didactic team-taught lectures followed by hands on demonstrations of the use of individual databases to retrieve biological information, differential expression, network, and pathway analysis.

Reading List
The course will be based on primary literature and database-specific protocols.

Agarwala, R., Barrett, T., Beck, J., Benson, D., Bollin, C., & Bolton, E. et al. (2017). *Database resources of the National Center for Biotechnology Information*. Nucleic Acids Research, 46(D1), D8-D13. doi: 10.1093/nar/gkx1095

Benson, D. (2018). *GenBank. Nucleic Acids Research*, *46*(1), 41-47. doi: 10.1093/nar/gkg057
Byron, K., Herbert, K., & Wang, J. (2016). *Bioinformatics database systems* (1st ed.). CRC Press.

Chen, C., Huang, H., & Wu, C. H. (2017). *Protein Bioinformatics Databases and Resources.* Methods in molecular biology: Clifton, New Jersey, *1558*, 3–39. doi:10.1007/978-1- 4939- 6783-4_1

Dua, S., & Chowriappa, P. (2013). *Data mining for bioinformatics* (1st ed.). CRC Press, Taylor & Francis Group: Boca Raton, Fl.

Hamid Ismail., (2021) Bioinformatics: A Practical Guide to NCBI Databases and Sequence Alignments. Chapman & Hall/CRC Computational Biology Series. CRC Press. ISBN 1032123699, 9781032123691

Sharma, S., Ciufo, S., Starchenko, E., Darji, D., Chlumsky, L., Karsch-Mizrachi, I., & Schoch, C. (2018). *The NCBI BioCollections Database.* Database. doi: 10.1093/database/bay006

ELECTIVES

***BSTT 601: Methods in Biostatistics 1**

This course introduces the fundamentals of biostatistics for analyzing data in public health. Students will learn graphical and exploratory data analysis techniques using scatterplots, boxplots, and multivariate data displays. The course covers the distinctions between statistical methodologies and explores maximum likelihood and frequentist methods including hypothesis testing and confidence intervals. Topics include classical inference, population vs. sample distinctions, data types, analysis of continuous, binary, and count data, sampling distributions, estimation, and hypothesis tests. Students will gain proficiency in Bayesian techniques, P-value interpretation, estimation of single group summaries, two-group

comparisons, and basic ANOVA concepts. Lectures and hands-on practical sessions with freely available datasets support learning.

Course objective

The objectives of this course include:

- Introduction to the display and communication of statistical data.
- Graphical and exploratory data analysis using tools like scatterplots, boxplots and the display of multivariate data.
- Exploring distinctions between the fundamental paradigms underlying statistical methodology, the basics of maximum likelihood,
- Explanation of the basics of frequentist methods: hypothesis testing, confidence intervals.

Course content

This module introduces the basic statistical concepts and methods as applied to diverse problems in public health. Students should be familiar with data handling commands in Stata. Topics to be covered are an introduction to classical inference including the distinctions between population and sample, and between statistics and population values, and types of data. This component will also include analysis of continuous data (including linear regression), analysis of binary data, and analysis of count data within the concept of sampling distributions, estimation, confidence intervals, hypothesis tests, types I and II errors. Also included is the comparison of groups, association (contingency tables), stratification (Mantel-Haenzel methods) and interaction.

Learning outcomes

At the end of this module, students should be able to:

- Know the basic Bayesian techniques
- Create and interpret P values
- Estimate, test and interpret single group summaries such as means, medians, variances, correlations, and rates
- Estimate, test and interpret two group comparisons such as odds ratios, relative risks and risk differences, and the basic concepts of ANOVA

Mode of delivery

This course will be delivered through lectures to cover theory and concepts of biostatistics. These will be supported with hands-on practical using freely available datasets.

Reading List

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley and Sons: USA.

Marcello P, Kimberlee G, Heather M., (2022) Principles of Biostatistics. Chapman and Hall/CRC. ISBN 0367355809, 9780367355807

Richard J. Rossi., (2022) Applied biostatistics for the health sciences. WILEY. 2nd Edition. ISBN 9781119722694, 1119722691

Sullivan, L. M. (2012). *Essentials of biostatistics in public health* (2nd ed.). Jones and Bartlett Learning, LLC: Burlington, MA: USA.

Susan White., (2019) Basic & Clinical Biostatistics. McGraw Hill. ISBN 1260455378, 9781260455373

BINF 612: Proteomics

Proteomics is a comprehensive course that introduces students to the fundamental principles and applications of mass spectrometry in the fields of proteomics and metabolomics. The course explores the functionality of mass spectrometers, covering topics such as mass analyzers, parent and daughter ions, b and y ion series, chromatography, and data outputs. Students will gain proficiency in using bioinformatic tools to identify and quantify ions present in spectra. Additionally, the course delves into the analysis pipeline, from spectra to inferred proteins or molecules, emphasizing the importance of data quality control to monitor false discovery rates. Lectures, hands-on practical sessions with freely available datasets, and journal club presentations will facilitate learning in this field. Upon completion of the course, students will be able to comprehend mass spectrometry operations, interpret output data, analyze mass spectrometry data, and assess data quality.

Course Objectives

The objectives of this course include:

- Introduction into the basic principles and function of a mass spectrometer
- Exploring the output of the data for analyses and validation of mass spectrometry data

Course Content

Mass spectrometry as a tool has greatly advanced the fields of both proteomics and metabolomics. The ability to acquire high resolution spectra at increasing speeds has allowed for a whole host of bioinformatic tools to be developed that can leverage this information to identify and quantify ions present in spectra. The basics of mass spectrometry will be covered, including mass analysers, parent and daughter ions, b and y ion series, chromatography, data outputs, experimental variations (such as labelling), and targeted vs discovery approaches. In addition, a typical pipeline for analysis from spectra to inferred protein or molecule will be discussed, including the need for data quality controls to monitor false discovery rates.

Learning Outcome

At the end of this module, students should be able to:

- Describe how a mass spectrometer works when acquiring data.
- Interpret the output data obtained from a mass spectrometer.
- Analyse mass spectrometry data
- Determine the quality of the data generated.

Mode of delivery

This course will be delivered through team-taught lectures to cover theory and concepts of proteomics. These will be supported with hands-on practical using freely available datasets and journal club presentations of current papers in the field.

Reading List

Cox, J., & Mann, M. (2011). *Quantitative, High-Resolution Proteomics for Data-Driven Systems Biology*. Annual Review of Biochemistry, *80*(1), 273-299. doi: 10.1146/annurev-biochem- 061308-093216

Daniela Cecconi., (2021) Proteomics Data Analysis. Methods in Molecular Biology, 2361. Humana. ISBN 71616404, 9781071616406

Katrin M, Martin E, Barbara S., (2021) Quantitative Methods in Proteomics. Methods in Molecular Biology, 2228. Humana. ISBN 1071610236, 9781071610237

Larance, M., & Lamond, A. (2015). *Multidimensional proteomics for cell biology*. Nature Reviews Molecular Cell Biology, *16*(5), 269-280. doi: 10.1038/nrm3970

RC Sobti, Manishi M, Aastha S., (2022) Genomic Proteomics and Biotechnology. Translating Animal Science Research. CRC Press. ISBN 2022023880, 2022023881, 9781032116334, 9781032116341, 9781003220831

Rune Matthiesen., (2020) Mass Spectrometry Data Analysis in Proteomics. Methods in Molecular Biology 2051. Springer New York;Humana. ISBN 978-1-4939-9743-5, 978-1-4939-9744-2

Sanjeeva Srivastava., (2022) From Proteins to Proteomics: Basic Concepts, Techniques, and Applications. CRC Press. ISBN 0367566206, 9780367566203

BINF614: Machine learning

The course provides a comprehensive exploration of machine learning in bioinformatics. Students will gain an understanding of the principles, algorithms, and concepts underlying various machine learning techniques. Topics include supervised and unsupervised learning, application of machine learning in genomics, genetics, text mining, and drug discovery. Practical aspects cover programming languages such as R, and Python, for data analysis. Students will learn to implement machine learning algorithms, analyze biological data using supervised and unsupervised techniques, biological dataanalysis with Support Vector Machines (SVMs), and apply neural network algorithms in bioinformatics. The course employs lectures, hands-on tutorials, journal club discussions, student presentations, and project work to enhance learning.

Course Objective

The objectives of this course include:

- Introducing students to the principles, concepts, and algorithms underlying different types of machine learning techniques
- Identifying differences between supervised and unsupervised machine learning techniques.
- Application of machine learning in bioinformatics including genomics, genetics, text mining and drug discovery.
- Using different computer programming languages for machine learning-based data analysis
- Evaluating machine learning parameters and how to apply these tools to functionally relevant biological problems.

Course Content

The course covers the application of machine learning techniques in genomics, proteomics, microarrays, systems biology, evolution, drug discovery and text mining. It addresses algorithms, theories and concepts underpinning several techniques, including both supervised and unsupervised machine learning approaches. Some of the approaches to be considered are Hidden Markov Model, Artificial Neural Networks, Deep Learning, Random Forest, Support

Vector Machine and Naïve Bayes. Some of the topics to be covered include classification, feature subset selection, predictive models, clustering, and optimisation. Programming environment is flexible, and any language can be used including R, Python or Perl.

Learning Outcome
At the end of this module, students should:

- Be able to implement commonly used machine learning algorithms.
- Be able to analyse biological data using supervised and unsupervised machine learning techniques.
- Be able to use computer languages including R, Python or Perl for machine learning analysis.
- Be able to undertake predictive sequence analysis using Support Vector Machines (SVMs).
- Be able to apply neural networks algorithms in bioinformatics

Mode of delivery
The lectures will cover the theories, concepts and principles underlying different machine learning techniques. It will involve hands-on practical tutorials, paper discussions as journal club, and student presentations. The course delivery includes team-taught lectures.

Reading List

Kasim, A. (2016). *Applied bioclustering methods for big and high dimensional data using R* Taylor & Francis, CRC Pres: Boca Raton

Mitra, S. (2008). *Introduction to machine learning and bioinformatics.* Boca Raton: CRC Press.

Scutari, M. & Denis, J. B. (2014). *Bayesian networks: with examples in R*. CRC Press Taylor & Francis Group: Boca Raton.

Yang, Z. (2010). *Machine learning approaches to bioinformatics.* World Scientific: Singapore.

Zhang, Y. & Rajapakse, J. (2009). *Machine learning in bioinformatics.* Wiley: Hoboken, N.J.

PROJECT AND SEMINAR

BINF 600: Project
BINF600 is a research-based course designed to provide students with the opportunity to undertake an in-depth research project in the field of Bioinformatics. Students will develop their research skills and gain practical experience in conducting scientific investigations. The course will culminate in the production of a written report and an oral presentation at a departmental seminar, allowing students to demonstrate their research findings and defend their work.

Course Objectives:

- To develop students' research skills in the field of Bioinformatics.
- To provide practical experience in conducting scientific investigations.
- To foster critical thinking and problem-solving abilities within the realm of Bioinformatics research.

- To enhance students' communication and presentation skills through written and oral defense of their research findings.
- To promote independent learning and self-directed study.

Course Content:
This module will lead students to practicalize bioinformatics research methodologies and techniques. Students will practice the selection and formulation of research questions and problem statements; will practice the conduct of literature review and identification of relevant prior research; design and implement research experiments and computational analysis; collect and analyze data, with appropriate statistical analysis and data visualization tools; write a research report that strictly adheres to current scientific standards; prepare and deliver compelling oral presentations that best reflect the quality of their work.

Course Outcomes:

By the end of the course, students will be able to:

- Demonstrate an understanding of research methodologies and techniques employed in Bioinformatics.
- Formulate a clear research question or problem statement.
- Conduct a comprehensive literature review and critically evaluate prior research.
- Design and execute research experiments or computational analyses.
- Analyze and interpret research data using appropriate statistical methods.
- Communicate research findings effectively through a written research report.
- Present research findings confidently and professionally in an oral defense.
- Demonstrate ethical awareness and adhere to responsible conduct of science principles.

Mode of Delivery:
The course will be student-led, where students will independently identify research problems, and work to address them under the supervision of a faculty member. Preliminary lectures will introduce students to key concepts, methodologies, and techniques in Bioinformatics research, in addition to domain learnings students have acquired through other taught causes. Seminars will provide opportunities for discussion, peer feedback, and guidance on research progress. Students will be expected to engage in independent study to conduct their research, analyze data, and prepare their written report and oral presentation. Supervisors will continually provide feedback to students on the progress of their work.

BINF610: Seminar I

Course description and objective

This course will provide an opportunity for students to enhance their scientific communication skills and gain a comprehensive understanding of research in Bioinformatics. Through departmental seminars, students will present project proposals, progress reports, and articles, while also attending all seminars. The course objectives are focused on developing strong science communication abilities and fostering familiarity with various bioinformatic research methods. The content of the course will encompass researching, critically analyzing, and summarizing scholarly literature, presenting research methods and data interpretations to diverse audiences, communicating findings clearly and concisely, effectively responding to audience questions, constructively evaluating presentations, and asking relevant, thoughtful questions. By the end of the course, students will have acquired the skills to understand bioinformatic research methods, generate and analyze bioinformatic data, and prepare for scientific communication.

Course Objective

This course is intended to:

- Provide students with a broad overview of research in Bioinformatics and provide students with strong science communication skills.

Course Content

Researching, critically analysing and summarizing scholarly literature; Presenting research methods, analyses and research data and interpretations to a diverse audience; Communicating findings and views clearly and concisely; Learning how to respond to audience questions; Constructively evaluating their own presentation as well as presentation of others; Asking relevant, thoughtful questions.

Learning outcomes

Students will learn how to understand bioinformatic research methods, generate bioinformatic data, analyse and interpret bioinformatic data, and prepare for scientific communication.

Mode of Delivery:

The course will be in the form of student- delivered presentations on current research articles and trends in bioinformatics. There will be a faculty member in charge of organizing the weekly seminars. All faculty members will score the presentations for a final cumulative average seminar grade. Seminars will provide opportunities for discussion, peer feedback, and guidance on research progress.