Universitat Autònoma de Barcelona

2019/2020

Core Bioinformatics

Code: 42397
ECTS Credits: 12

Degree	Type	Year	Semester
4313473 Bioinformatics	OB	0	1

Contact

Name: Sònia Casillas Viladerrams
Email: Sonia.Casillas@uab.cat

Use of Languages

Principal working language: english (eng)

Teachers

Antoni Barbadilla Prados
Leonardo Pardo Carrasco
Pere Puig Casado
Alfredo Ruíz Panadero
Miquel Àngel Senar Rosell
Jean-Didier Pierre Marechal
Isaac Salazar Ciudad
Oscar Conchillo Solé
Raquel Egea Sánchez

External teachers

Cedric Notredame
Emanuel Raineri
Josep Abril
Sebastián Ramos

Prerequisites

Level B2 of English or equivalent is recommended.

Objectives and Contextualisation

This module focuses on the development of diverse bioinformatic tools and resources commonly used in Omics research. Our intention is that it covers several aspects of bioinformatics in a series of brief topics, in the form of "tastings". Therefore, it is not an accummulative module, but a transversal one, which should provide with a wide range of ideas and approaches that bioinformatics offers, through the hands of experts. The main

objective is to provide students with the necessary foundation to apply bioinformatics to different areas of scientific research. Over time, each student will be able to gain all the depth they propose on any of these topics, the one which finally represents their research framework.

Competences

- Analyse and interpret data deriving from omic technology using biocomputing methods .
- Design and apply scientific methodology in resolving problems.
- Possess and understand knowledge that provides a basis or opportunity for originality in the development and/or application of ideas, often in a research context.
- Propose biocomputing solutions for problems deriving from omic research.
- Propose innovative and creative solutions in the field of study
- Student should possess the learning skills that enable them to continue studying in a way that is largely student led or independent.
- Understand the molecular bases and most common standard experimental techniques in omic research (genomics, transcriptomics, proteomics, metabolomics, interactomics, etc.)
- Use and manage bibliographical information and computer resources in the area of study
- Use operating systems, programs and tools in common use in biocomputing and be able to manage high performance computing platforms, programming languages and biocomputing analysis.

Learning Outcomes

1. Create and promote algorithms, calculation and statistical techniques and theories to resolve formal and practical problems deriving from the handling and analysis of biological data.
2. Design and apply scientific methodology in resolving problems.
3. Identify and apply algorithms in which the programs are based bioinformatic analysis.
4. Identify and classify the principle types of biomolecular data obtained from omic technology.
5. Possess and understand knowledge that provides a basis or opportunity for originality in the development and/or application of ideas, often in a research context.
6. Propose innovative and creative solutions in the field of study
7. Search for specific bioinformatics tools and bioinformatics resources in the network.
8. Student should possess the learning skills that enable them to continue studying in a way that is largely student led or independent.
9. Synthesise and interpret in a logical and reasoned manner the information from the molecular data bases and analyse it using biocomputing tools.
10. Understand the theoretical, statistical and biological bases, in which the programs are based bioinformatic analysis: sequence alignment, similarity search and multiple alignment, structure prediction, genome annotation, phylogenetic and evolutionary analysis.
11. Use and manage bibliographical information and computer resources in the area of study
12. Use the main molecular databases, the main standard formats of molecular data and integrate data from different data sources

Content

BLOCK 1. STATISTICS

Statistical Inference
Professor Antonio Barbadilla

- Statistics: bridge between data and models
- Data Types
- Population and sample
- Experimental design
- Data Quality
- Exploration of Data
- Sample distribution and law of large numbers

- Statistical inference
- Central Limit Theorem
- Point estimation
- Estimation of confidence interval
- Hypothesis
- Elements of a test: H0, H1, statistical test, p value, significance level, type I and II errors, power
- Z test, t test, chi-square test, correlation test, regression, analysis of variance
- Interpretation of statistical significance
- Parametric versus nonparametric tests
- Selecting the appropriate statistical test (decision tree)
- Multivariate Testing
- Resampling

Statistics and Stochastic Processes for Sequence Analysis
Professor Pere Puig

a. Probability basics
Sets and events. Properties. Conditional probability. Independence. Alphabet and sequences. Probabilistic models.

b. The multinomial model
Simulating a multinomial sequence. Estimating probabilities.

c. The seqinr package

d. Markov chain models
Concept and examples. Classification of states. R code. Simulating a Markov chain sequence. Estimating the probabilities of transition. The probability of a sequence. Using Markov chain for discrimination.

e. Higher order Markov chain models
Concept and examples. Estimating the probabilities of transition. Comparison of higher order Markov chains.

f. Hidden Markov chain models
Concept and examples. Parameter estimation. Hidden states estimation.

g. An introduction to Generalized Linear Models
GLM basics. The Logistic model. The Poisson model.

Bayesian Inference
Professor Emmanuele Raineri

1. Curve fitting.
- Estimation of parameters of probability distributions: binomial, poisson and gaussian.
- Example: fitting a noisy dataset.
- Cross validation, overfitting and regularization.

2. Dimensional reduction.
- Principal component analysis, multidimensional scaling.
- Example: distinguishing cell types using methylation profiles.

3. Lasso regression.
- Variable selection in linear models.
- Penalized regression: Lasso and Elastic Net.
- Example: lasso regression in R.

BLOCK 2. BASIC UTILITIES

The Human Genome

Professor Alfredo Ruiz

a. Introduction to genomes
Sequenced genomes. Organization and size of eukaryotic genomes. Building a genome: NGS methods for genomics and transcriptomics.

b. The human genome: where are we now?
Current assembly of the human genome. The ENCODE project: functional elements in the human genome. Repetitive content of the human genome.

Databases and Sequence Formats

Professor Oscar Conchillo

a. Sequence formats
Nomenclature. Text editors. FASTA format and its variants. Raw/Plain format. Genbank sequence format. EMBL sequence format. GCG, NBRF/PIR, MSA, PHYLIP, NEXUS. Format conversion.

b. Databases
Concept. Boolean searches. Wildcards and regular expressions. Identifiers and accession numbers. Classification. NAR databases compilation. GenBank and other NCBI databases. EMBL. DDBJ. Integrated Meta-Databases. Main nucleotide, protein, structure, taxonomy, etc. databases.

Software Engineering

Professor Miquel Àngel Senar

a. Version control system with Git and GitHub
b. Parallelization strategies and HPC
c. Cloud computing with Amazon Web Services

Workflows with Galaxy

Professor Raquel Egea

a. Introduction to workflow managers
Concept, origin and design of workflow managers. Workflow patterns. Existing workflow managers. APIs and Web Servers.

b. Galaxy: basics, interface and practical uses.

BLOCK 3. STRUCTURAL BIOINFORMATICS

Protein structure

Professors Leonardo Pardo and Óscar Conchillo

a. Introduction
Amino acids, proteins, and peptide bonds. Four levels of protein structure. Protein folding and stability. Molecular interactions. Experimental methods for structure determination.

b. Motifs and domains

c. Analysis
UNIPROT, PDB, PFAM, CATH, and SCOP databases. Protein alignment, morphing, molecular surfaces, molecular electrostatic potential.

d. Cell membrane
Membrane proteins, transmembrane segments

Molecular modeling

Professors Leonardo Pardo *and Jean-Didier Maréchal*

a. Homology modeling

b. Molecular modeling
Atomic models. Potential energy. Quantum and molecular mechanics. Conformational exploration techniques

BLOCK 4. GENOMICS

Sequence Alignment

Professor Cedric Notredame

a. Evolution and comparison Models
Molecular clock. Protein structure and evolution. Substitution Matrices.

b. Dynamic Programming based sequence comparisons
Needlman and Wunsch algorithm. Smith and Waterman algorithm. Afine gap penalties computation. Linear space computation of pairwise algorithms.

c. Blast and Database searches
The Blast algorithm. E-values and estimates of statistical significance. Database search strategies. PSI-Blast and other evolutionary approaches.

d. Multiple Sequence Alignments: algorithms and strategies
Main applications of multiple sequence alignments. Most common algorithms. Multiple sequence alignment strategies.

Gene andControl Region Finding
Professor Josep Abril

a. Gene prediction
Annotation: concept, databases, problems. Gene finding: search by signal, search by content, approaches (ab-initio, homology search, comparative genomics, NGS), evaluation of software accuracy.

b. Finding regulatory motifs
DNA motifs: exact matching, regular expressions, position weight matrices, search trees, profiles, randomized algorithms, logos and pictograms, software for motif finding. Regulatory domains. Histones. CRMs architecture & networking. Regulatory networks. Meta-alignment. Conservation, phylogenetic footprinting and phylogenetic shadowing. NGS.

Population Genomics
Professor Alfredo Ruiz

a. Population genomics under neutrality in a finite population
Introduction. Genetic drift. Effective population size. Probability of fixation of neutral mutants.

b. Population genomics under selection
Natural selection. Probability of fixation of selected mutants. Fitness distribution of new mutants. Rate of evolution.

c. Adaptive evolution and population size

Phylogeny and Molecular Evolution
Professor Sebastián Ramos

a. Models of sequence evolution
DNA sequence. Jukes and Cantor model. More realistic models. Model selection.

b. Phylogeny
Concept. Species trees versus gene trees. Tree-reconstructionmethods: distance methods, maximum parsimony, maximum likelihood, Bayesian inference. Support. Phylogenomics. Building trees with R.

Systems Biology

Professor Isaac Salazar

a. Classical and Genomic age Systems Biology
The systems biology paradigm in light of technological developments over the last 100 years. Data integration bottlenecks.

b. Mathematical modeling of molecular circuits.
Conceptual models. From conceptual models to mathematical models. Mathematical formalisms. Data driven models.

c. Design and organization principles in molecular circuits.
Conceptof design principle. Mathematically controlled comparisons. Feasibility analysis. Design Spaces. Synthetic Biology.

Methodology

The methodology will combine master classes, solving practical problems and real cases, working in the computing lab, performing individual and team work, reading articles related to the thematic blocks, and independent self-study. The virtual platform will be used.

Activities

Title	Hours	ECTS	Learning Outcomes
Type: Directed			
Solving problems in class and work in the biocomputing lab	39	1.56	7, 10, 1, 2, 3, 4, 6, 8, 9, 5, 11, 12
Theoretical classes	39	1.56	7, 10, 1, 2, 3, 4, 6, 8, 9, 5, 11, 12
Type: Supervised			
Performing individual and team works	40	1.6	7, 10, 1, 2, 3, 4, 6, 8, 9, 5, 11, 12
Type: Autonomous			
Regular study	178	7.12	7, 10, 1, 2, 3, 4, 6, 8, 9, 5, 11, 12

Assessment

The evaluation system is organized in three main activities. There will be, in addition, a retake exam. The details of the activities are:

Main evaluation activities

- Student's portfolio (55%): work done and presented by the student all along the course. None of the individual assessment activities will account for more than 50% of the final mark.
- Individual theoretical and practical test (35%): a final exam will take place at the end of this module. It will consist of one or two multiple-choice or short questions by each professor teaching in this module.

- Soft skills (10%): assistence, arrival on time and active participation in class.

Retake exam

To be eligible for the retake process, the student should have been previously evaluated in a set of activities equaling at least two thirds of the final score of the module. The teacher will inform the procedure and deadlines for the retake process. Please note that soft skills cannot be recuperated.

Not valuable

The student will be graded as "Not Valuable" if the weight of the evaluation is less than 67% of the final score.

Assessment Activities

Title	Weighting	Hours	ECTS	Learning Outcomes
Individual theoretical and practical test	35%	4	0.16	7, 10, 1, 2, 3, 4, 6, 8, 9, 5, 11, 12
Soft skills	10%	0	0	2, 6, 8, 5
Student's portfolio	55%	0	0	7, 10, 1, 2, 3, 4, 6, 8, 9, 5, 11, 12

Bibliography

Updated bibliography will be recommended in each session of this module by the professor, and links will be made available on the Student's Area of the MSc Bioinformatics official website (http://mscbioinformatics.uab.cat/base/base3.asp?sitio=bioinformaticsintranet&anar=module_2&item=&subitem=)