

Degree programme	Type	Course
Bioinformatics	OB	1

Contact lecturer

Name : Sònia Casillas Viladerrams

Email : sonia.casillas@uab.cat

Teaching staff

Isabel Serra Mochales

Antoni Barbadilla Prados

Leonardo Pardo Carrasco

Miquel Àngel Senar Rosell

Jean-Didier Pierre Marechal

Jaime Martinez Urtaza

Isaac Salazar Ciudad

Oscar Conchillo Solé

Marta Puig Font

Angel Gonzalez Wong

Teaching staff (external to UAB)

Sebastián Ramos

Emanuel Raineri

Group languages

You can consult this information at the [end](#) of the document.

Prerequisites

Level B2 of English or equivalent is recommended.

Objectives

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 accumulative 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.

Learning outcomes

- CA03 (Manage the methodological process in the approach to complex bioinformatics problems in a rigorous manner.) Manage the methodological process in the approach to complex bioinformatics problems in a rigorous manner.
- CA04 (Develop bioinformatic approaches based on the integration of bibliographic information and specialised computer resources.) Develop bioinformatic approaches based on the integration of bibliographic information and specialised computer resources.
- CA05 (Manage resources effectively, assessing possible inequalities arising from social or sex/gender factors.) Manage resources effectively, assessing possible inequalities arising from social or sex/gender factors.
- KA04 (Explain the fundamentals and applications of the main methodologies, techniques and tools used in the sequencing, assembly and annotation of genomes, incorporating concepts of statistical inference, stochastic processes and Bayesian inference applied to sequence analysis.) Explain the fundamentals and applications of the main methodologies, techniques and tools used in the sequencing, assembly and annotation of genomes, incorporating concepts of statistical inference, stochastic processes and Bayesian inference applied to sequence analysis.
- KA05 (Explain the structure, content, and interrelatedness of the main sources of biological data that include information on the organisation, evolution, expression, variation, and three-dimensional structure of proteins, understanding their relevance in molecular modelling and functional annotation.) Explain the structure, content, and interrelatedness of the main sources of biological data that include information on the organisation, evolution, expression, variation, and three-dimensional structure of proteins, understanding their relevance in molecular modelling and functional annotation.
- KA06 (Describe the types of information present in omics databases and their contribution to the study of biological diversity and complexity in different ecological and social contexts.) Describe the types of information present in omics databases and their contribution to the study of biological diversity and complexity in different ecological and social contexts.
- SA04 (Use the most common experimental techniques used in omics research (genomics, transcriptomics, proteomics, metabolomics, interactomics, etc.) based on their molecular bases.) Use the most common experimental techniques used in omics research (genomics, transcriptomics, proteomics, metabolomics, interactomics, etc.) based on their molecular bases.
- SA05 (Handle different operating systems, programs and high-performance computing platforms, as well as different programming languages commonly used in bioinformatics.) Handle different operating systems, programs and high-performance computing platforms, as well as different programming languages commonly used in bioinformatics.
- SA06 (Analyse bioinformatic data derived from omics technologies and information from molecular databases.) Analyse bioinformatic data derived from omics technologies and information from molecular databases.
- SA07 (Apply bioinformatics solutions for the management and analysis of biological data derived from omics research.) Apply bioinformatics solutions for the management and analysis of biological data derived from omics research.

Contents

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: H_0 , H_1 , 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 Isabel Serra

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 Marta Puig

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

BLOCK 3. STRUCTURAL BIOINFORMATICS

Protein structure

Professors Leonardo Pardo, Óscar Conchillo and Ángel Gonzalez

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

Introduction: Genome and omic data

Professor Jaime Martínez Urtaza

a. Main milestones in the genome sequencing project: sequencing, assembly and annotation.

b. Sequencing. Classic sequencing by Sanger method. Next-generation sequencing (NGS) techniques. Second generation techniques: 454/Roche (pyrosequencing), Illumina (reversible termination), SOLiD (sequencing by ligation), Ion Torrent (proton detection). Third generation techniques. Challenges and differences with second generation techniques. Pacific Biosciences (PacBio; Single Molecule Real Time, SMART). Oxford Nanopore (Minion).

c. Assembly. De novo assembly versus mapping against reference. Reads and contigs. Measurement of the quality of an assembly: quality of a base or Phred score (Q), redundancy (coverage), N50 and L50. Paired-end reads and scaffolds.

Population Genomics

Professor Isaac Salazar

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-reconstruction methods: 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.

Concept of design principle. Mathematically controlled comparisons. Feasibility analysis. Design Spaces. Synthetic Biology.

Learning activities and methodology

Title	Hours	ECTS	Learning outcomes
Solving problems in class and work in the biocomputing lab	39	1.56	CA03, CA04, CA05, KA04, KA05, KA06, SA04, SA05, SA06, SA07
Regular study	178	7.12	CA03, CA04, CA05, KA04, KA05, KA06, SA04, SA05, SA06, SA07
Theoretical classes	39	1.56	CA03, CA04, CA05, KA04, KA05, KA06, SA04, SA05, SA06, SA07
Performing individual and team works	40	1.6	CA03, CA04, CA05, KA04, KA05, KA06, SA04, SA05, SA06, SA07

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.

Annotation: within the schedule set by the centre or degree programme, 15 minutes of one class will be reserved for students to evaluate their lecturers and their courses or modules through questionnaires.

Assessment

Continuous assessment activities

Title	Weight	Hours	ECTS	Learning outcomes
Individual theoretical and practical test	35%	4	0.16	CA03, CA04, CA05, KA04, KA05, KA06, SA04, SA05, SA06, SA07
Student's portfolio	55%	0	0	CA03, CA04, CA05, KA04, KA05, KA06, SA04, SA05, SA06, SA07
Soft skills	10%	0	0	CA03, CA04, CA05, KA04, KA05, KA06, SA04, SA05, SA06, SA07

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%): assistance, 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 "Non-gradable" if the weight of the evaluation is less than 67% of the final score.

Unique assessment

This subject/module does not provide for the single assessment system.

Any irregularity committed during an assessment activity (academic misconduct, plagiarism, or improper use of AI, unless such use is expressly authorized in the course guide) that may lead to a significant change in the grade will result in that assessment activity being graded with a 0. If the course guide establishes that obtaining a minimum grade in that assessment activity is an essential requirement to pass the course, or if several irregularities occur in the assessment activities of the same course, the final grade for that course will be 0. In addition, disciplinary proceedings may be initiated against any student who commits any of these irregularities.

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.

Software

Updated software 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.

Course groups and languages

The information provided is provisional until November 30. After this date, you will be able to consult the language of each group through this [link](#). To access the information, you will need to enter the course CODE

Type of teaching	Group	Language	Semester	Shift
(TEm) Theory (master)	1	English	first semester	morning-mixed
(PLABm) Practical laboratories (master)	1	English	first semester	morning-mixed