Core Bioinformatics 2014/2015

Code: 42397
ECTS Credits: 12

Degree	Type	Year	Semester
4313473 Bioinformàtica/Bioinformatics	OB	0	1

Contact

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

Use of languages

Principal working language: anglès (eng)
Some groups entirely in English: No
Some groups entirely in Catalan: Yes
Some groups entirely in Spanish: No

Teachers

Leonardo Pardo Carrasco
Pere Puig Casado
Alfredo Ruíz Panadero
Miquel Àngel Senar Rosell
Jean Didier Pie Marechal
Daniel Yero Corona
Marta Puig Font
Raquel Egea Sánchez
Xavier Daura Ribera

External teachers

Cedric Notredame
Josep Abril

Prerequisites

Level B2 of English or equivalent is recommended.

Objectives and Contextualisation

This module focuses on the development of bioinformatic tools and resources commonly used in Omics research. The main objective is to provide students with the necessary foundation to apply bioinformatics to different areas of scientific research.

Skills

- 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 classify the principle types of biomolecular data obtained from omic technology.
4. 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.
5. Propose innovative and creative solutions in the field of study
6. Student should possess the learning skills that enable them to continue studying in a way that is largely student led or independent.
7. Synthesise and interpret in a logical and reasoned manner the information from the molecular data bases and analyse it using biocomputing tools.
8. Use and manage bibliographical information and computer resources in the area of study
9. Use the main molecular databases, the main standard formats for molecular data and integrate data from different sources

Content

Session 1. Bioinformatics Formats and Databases
Professor Daniel Yero

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.

Sessions 2- 4. 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.

Sessions 5. Bayesian Inference
Professor TBD

Session 6. 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.

Session 7. Gene and Control 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.

Session 8. Introduction to Genomes
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.

Session 9. 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

Session 10. 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.

Session 11-14. Structural Bioinformatics
Professors Xavier Daura, Leonardo Pardo and Jean-Didier Maréchal

a. Structural biology and interactions
Biomolecules. Xenobiotics. Proteins, aminoacids and peptide bonds. Four levels of protein structure. Protein folding and stability. DNA structure. Experimental methods for structure determination.

b. Structure databases
PDB. PyMOL.

c. Molecular modeling
Models. Potential energy. Quantum and molecular mechanics techniques. Energy minimization. Molecular Dynamics. Homology modeling.

d. Docking
Concept and applications. Exploring the conformational space. Defining molecular interactions. Scoring functions. Workflows and software. Online resources.

Sessions 15. Systems Biology
Professor TBD

Sessions 16. Web Development
Professor Raquel Egea

a. Web development
Why developing webs as a bioinformatician? Web developing basics: OS, Web servers, databases, programming languages. Apache. HTML, CSS, JavaScript, PHP.

b. Self-learning: bibliography and useful online resources
Online bioinformatics communities.

Session 17-18. Software Engineering
Professors Miquel Àngel Senar and Raquel Egea

a. Parallelization strategies

b. Introduction to workflow managers
Concept, origin and design of workflow managers. Workflow patterns. Existing workflow managers. APIs and Web Servers. Galaxy: basics, interface and practical uses.

c. Version control system with Git and GitHub

d. Cloud computing with Amazon Web Services

Methodology

The methodology will combine master classes, solving practical problems and real cases, working in the computing lab, performing individual and team work, readings 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	29	1.16	1, 2, 3, 4, 5, 6, 7, 8, 9
Theoretical classes	34	1.36	1, 2, 3, 4, 5, 6, 7, 8, 9
Type: Supervised			
Performing individual and team works	40	1.6	1, 2, 3, 4, 5, 6, 7, 8, 9
Type: Autonomous			
Regular study	193	7.72	1, 2, 3, 4, 5, 6, 7, 8, 9

Evaluation

- Work done and presented by the student (student's portfolio) (50%).
- Individual theoretical and practical test (50%). A final exam will take place at the end of this module. It will consist of one or two short questions by each professor teaching in this module.

Evaluation activities

Title	Weighting	Hours	ECTS	Learning outcomes
Individual theoretical and practical test	50%	4	0.16	1, 2, 3, 4, 5, 6, 7, 8, 9
Student's portfolio	50%	0	0	1, 2, 3, 4, 5, 6, 7, 8, 9

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).