High Performance Computing and Big Data Analytics

Code: 43917
ECTS Credits: 12

2024/2025

Degree	Type	Year
4313473 Bioinformatics	OT	0

Contact

Name: Miquel Àngel Senar Rosell

Email: miquelangel.senar@uab.cat

Teachers

José Eduardo Cabrera Díaz

Oscar Lao Grueso

Emanuele Raineri

Teaching groups languages

You can view this information at the end of this document.

Prerequisites

To carry out this module is necessary to have passed previously both compulsory modules: Programming in Bioinformatics and Core Bioinformatics.

It is recommended you have a Level B2 of English or equivalent.

Objectives and Contextualisation

This module aims to provide students with the necessary knowledge and skills (1) to implement performance engineering approaches into modern computing platforms and (2) to perform statistical analyses of Big Data.

Competences

- Communicate research results clearly and effectively in English.
- 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
- 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. Apply advanced statistical methods (automatic learning, graph theory) to model and analyse bioinformatics problems involving massive biological data.
2. Communicate research results clearly and effectively in English.
3. Describe and apply clustering techniques and common classification algorithms.
4. Describe the operation, characteristics and limitations of the techniques, tools and methodologies to describe, analyze and interpret the amount of data produced by high-throughput technologies.
5. Design and apply scientific methodology in resolving problems.
6. Generate efficient parallel computing algorithms and applications for CID.
7. Know and handle open-source tools for parallel, distributed and scalable analysis through automatic learning.
8. Know the principles of massive data storage and management.
9. Know the principles of process parallelisation.
10. Learn new ways to model, store, recover and analyse abstract data types (graphs).
11. Learning to handle new platforms computing platforms, paradigms, and design applications that require massive computing and data handling.
12. 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.
13. Propose innovative and creative solutions in the field of study
14. Provide parallel solutions to specific bioinformatic problems.
15. Train, evaluate and validate predictive models.
16. Use and manage bibliographical information and computer resources in the area of study

Content

Modern Computer Architecture

- Cluster systems
- System Middleware and Programming Frameworks

Advanced Programming Models

- Advanced shell scripting
- Using system tools for bioinformatics analysis
- Principles of performance engineering (tools and methods)
- High Performance Computing with Python
- Performance engineering applied to common bioinformatics algorithms and tools (genome indexing, read alignment…).

Big Data Analytics

- Theory and tools of advanced statistics in Big Data analytics (dimensionality reduction, variable selection and Spark)
- Machine learning theory and algorithms. Applications in Bioinformatics
- Predictive modelling: data mining, model evaluation and validation
- Data classification: naïve Bayes and decision trees learning
- Association rule learning
- Clustering analysis: k-means algorithm
- Graph Theory for Big Data

Activities and Methodology

Title	Hours	ECTS	Learning Outcomes
Type: Directed			
Solving problems in class and work in the biocomputing lab	32	1.28	1, 3, 4, 6, 7, 8, 9, 10, 11, 12, 14, 15
Theoretical classes	38	1.52	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
Type: Autonomous			
Regular study	226	9.04	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 15, 16

By following a problem-oriented approach, students will get insight about efficient computational algorithms, methods and platforms and the statistical methods to be applied to challenging bioinformatics problems dealing with Big Data.

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

Continous Assessment Activities

Title	Weighting	Hours	ECTS	Learning Outcomes
Individual theoretical and practical tests	30%	4	0.16	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
Works done and presented by the student (student's portfolio)	70%	0	0	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16

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

Main evaluation activities

- Student's portfolio (60%): works 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 (40%): for each of the main modules of the subject, an individual evaluation mechanism will be established through an oral or written test.

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.

Not valuable

The student will be graded as "Not Valuable" 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.

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

Linux + SLURM and other tools from Linux enviroments

Python and other tools from its ecosystem

R and other tools from its ecosystem

Language list

Name	Group	Language	Semester	Turn
(PLABm) Practical laboratories (master)	1	English	first semester	morning-mixed
(TEm) Theory (master)	1	English	first semester	morning-mixed