

Course guide

230819 - DSBIO - Data Science in Bioinformatics and Computational Biology

Last modified: 01/06/2023

Unit in charge:	Barcelona School of Telecommunications Engineering	
Teaching unit:	707 - ESATII - Department of Automatic Control.	
Degree:	BACHELOR'S DEGREE IN TELECOMMUNICATIONS TECHNOLOGIES AND SERVICES ENGINEERING (Syllabus 2015). (Optional subject). BACHELOR'S DEGREE IN DATA SCIENCE AND ENGINEERING (Syllabus 2017). (Optional subject).	
Academic year: 2023	ECTS Credits: 6.0	Languages: Catalan

LECTURER

Coordinating lecturer:	Consultar aquí / See here: https://telecos.upc.edu/ca/estudis/curs-actual/professorat-responsables-coordinadors/responsables-assignatura
Others:	Consultar aquí / See here: https://telecos.upc.edu/ca/estudis/curs-actual/professorat-responsables-coordinadors/professorat-assignat-idioma

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Generical:

12 CPE N1. They will be able to identify, formulate and solve engineering problems in the ICC field and will know how to develop a method for analysing and solving problems that is systematic, critical and creative.

11 CDIO N3. They will be able to apply a comprehensive view of the entire life cycle (conception, design, implementation and operation) of a product, process or service in the ICC field, and identify users' needs and develop a set of requirements for the product, process or service and a set of initial specifications. They will be able to explore possible solutions and select the best one. They will be able to carry out a design process following a standardised methodology. They will know how to evaluate and propose improvements to the design. They will take into account economic and social aspects of the project or product.

Transversal:

04 COE N3. EFFICIENT ORAL AND WRITTEN COMMUNICATION - Level 3. Communicating clearly and efficiently in oral and written presentations. Adapting to audiences and communication aims by using suitable strategies and means.

05 TEQ N1. TEAMWORK - Level 1. Working in a team and making positive contributions once the aims and group and individual responsibilities have been defined. Reaching joint decisions on the strategy to be followed.

06 URI N3. EFFECTIVE USE OF INFORMATION RESOURCES - Level 3. Planning and using the information necessary for an academic assignment (a final thesis, for example) based on a critical appraisal of the information resources used.

07 AAT N2. SELF-DIRECTED LEARNING - Level 2: Completing set tasks based on the guidelines set by lecturers. Devoting the time needed to complete each task, including personal contributions and expanding on the recommended information sources.

07 AAT N3. SELF-DIRECTED LEARNING - Level 3. Applying the knowledge gained in completing a task according to its relevance and importance. Deciding how to carry out a task, the amount of time to be devoted to it and the most suitable information sources.

05 TEQ. TEAMWORK. Being able to work as a team player, either as a member or as a leader. Contributing to projects pragmatically and responsibly, by reaching commitments in accordance to the resources that are available.

07 AAT. SELF-DIRECTED LEARNING. Detecting gaps in one's knowledge and overcoming them through critical self-appraisal. Choosing the best path for broadening one's knowledge.

03 TLG. THIRD LANGUAGE. Learning a third language, preferably English, to a degree of oral and written fluency that fits in with the future needs of the graduates of each course.

TEACHING METHODOLOGY

The course is structured in three different stages. On first section (theory sessions):

- Kick off class (2 hours). All the class. Explanation of the dynamics of the class. Introduction of the first theme
- Subsequent classes will include introductory lectures and advanced lectures on the use of bioinformatics, online databases, analysis of datasets, Scientific R and interfacing to databases and visualization. Students will be asked to read and understand papers about a specific analysis in the area of bioinformatics. Classes will be mainly theoretical, intermingled with hands on interactive sessions aimed to R programming small tasks while incrementally approaching to the bioinformatic complete data analysis workflows.
- This set will mix synchronous sessions and online material in form of videos, learning scheme will generally adopt flipped classrooms.

The second section (laboratory sessions):

- Will consist of guided laboratory sessions, performed in groups. Laboratory sessions will cover practical examples of the different aspects of the course. These sessions will be performed online with Q&A slot sessions synchronously at lab times.

The third stage:

- Classes will contain lectures and the activity will be focused to the development of a guiding hackathon.
- This hackathon project involves the design of the programmatic approach on a biomedical through the synthesis and analysis of a biomedical dataset. In this team work project, students will learn, use the introduced data mining techniques with the integration of R, to extract relevant information from the database, including statistics and visualization.
- This third stage is conceived as a learn on demand system, where students will deeply comprehend the teaching material and complete knowledge as required for the project assigned to each team.
- The third stage end with a joint presentation with the results of the hackathon. This presentation will be held face-to-face or online depending on the sars-cov-2 risk index.

LEARNING OBJECTIVES OF THE SUBJECT

- Gain general working knowledge of the principles of biology, bioinformatics, and basic clinical science
- Understand the issues involved in bioinformatics databases and their structure
- To gain fluency in R software as biomedical data science programming language
- To gain skills for the statistical analysis of a disease, understanding the role of different factors
- Learn to implement best practices for data analysis and stewardship
- To gain skills to programmatic access to biological databases through R
- Demonstrate through a project the understanding and analysis of biological markers datasets
- Learn to explain detail complex information including technical data and results to a broad audience
- Learn ability to work well on teams and communicate effectively in both verbal and written venues
- Avoid behavioral beliefs and learn to act and understand reality based on evidence
- To strengthen professional and critical thinking
- Conduct research in an ethical manner that maintains data security and privacy

STUDY LOAD

Type	Hours	Percentage
Self study	98,0	65.33
Hours large group	26,0	17.33
Hours small group	26,0	17.33

Total learning time: 150 h

CONTENTS

Data Science in Bioinformatics and Computational Biology

Description:

1. Introduction.
2. Introduction to genetics
 1. The diversity of life and genetic material
 2. Genes
 3. DNA structure
 4. Transfer of information between DNA and proteins
 5. How is DNA analysed?
 6. Sources of genetic variability. Between individuals, between species
3. Data sources and algorithms
 1. Organization and levels of abstraction
 2. Genetics databases
 3. Gene Ontologies
 4. Programming languages with bio orientation.
 5. Sources of Algorithms
4. Sequence analysis
 1. Measurements of distance between sequences
 2. Models by Alignment.
 3. Needleman-Wunsch
 4. Smith-Waterman algorithms, BLAST
 5. Multiple alignment. Algorithms and interpretation
 5. Analysis of gene expression
 1. Technologies for gene expression measurement
 - a. Micro-Arrays
 - b. High Throughput Sequencing
 2. Data analysis workflow for gene expression data
 3. Introduction to Pattern recognition for gene expression
 4. Supervised and unsupervised learning
 5. Advanced Visualization
 6. Dimensionality reduction
 7. Methods of association
 6. Functional Analysis
 1. Introduction to Functional Analysis
 2. Overrepresentation analysis
 3. GSEA
 4. Network-based methods
 5. Applications into other omics (metabolomics)
 7. Analysis of protein structures
 1. Primary structure
 2. Secondary and tertiary or 3D structure
 3. Prediction of secondary structures
 4. Prediction of tertiary structures
 5. Analysis of protein structures
 8. Hackathon proposal

Related activities:

1. Introduction to R for bioinformatics
2. Ensembl! and NCBI. Access to biomedical databases with R and bioconductor
3. Homology
4. Introduction to R (III). Pattern Recognition, supervised and unsupervised analysis
5. Introduction to R (IV). Pattern Recognition, supervised and unsupervised analysis (if required)
6. Analysis of gene expression with R and Bioconductor
7. Functional enrichment

Full-or-part-time: 55h

Theory classes: 28h



Laboratory classes: 16h
Guided activities: 11h

GRADING SYSTEM

Grading will be built from continuous evaluation, including.:

- 60% for small individual assignments and laboratory report (approx. 2 assignments plus laboratories)
- 40% graded from an individual final assignment, dedicated to an analysis of a biomarkers dataset, presentation of results and presentation.

BIBLIOGRAPHY

Basic:

- Jones, Neil C.; Pevzner, Pavel A. An Introduction to bioinformatics algorithms [on line]. Cambridge, MA: MIT Press, cop. 2004 [Consultation: 30/06/2021]. Available on: <https://ebookcentral.proquest.com/lib/csuc-ebooks/detail.action?docID=3338847>. ISBN 0262101068.
- Gentleman, Robert. Bioinformatics and computational biology solutions using R and Bioconductor [on line]. New York: Springer Science+Business Media, cop. 2005 [Consultation: 30/06/2021]. Available on: <http://link.springer.com/openurl?genre=book&isbn=978-0-387-25146-2>. ISBN 0387251464.
- Wickham, Hadley; Golemund, Garrett. R for data science : import, tidy, transform, visualize, and model data. Beijing: O'Reilly, 2016. ISBN 9781491910399.
- Durbin, Richard. Biological sequence analysis : probabilistic models of proteins and nucleic acids [on line]. Cambridge [etc.]: Cambridge University Press, 1998 [Consultation: 30/06/2021]. Available on: <https://ebookcentral.proquest.com/lib/csuc-ebooks/detail.action?docID=320915>. ISBN 0521629713.