# UAB
## Universitat Autònoma de Barcelona

**2021/2022**

**Big Data Analysis in Bioinformatics**

Code: 104886
ECTS Credits: 6

| Degree | Type | Year | Semester |
|---|---|---|---|
| 2503852 Applied Statistics | OT | 4 | 0 |

The proposed teaching and assessment methodology that appear in the guide may be subject to changes as a result of the restrictions to face-to-face class attendance imposed by the health authorities.

## Contact

Name: Natalia Isabel Vilor Tejedor

Email: NataliaIsabel.Vilor@uab.cat

## Use of Languages

Principal working language: catalan (cat)

Some groups entirely in English: No

Some groups entirely in Catalan: Yes

Some groups entirely in Spanish: No

## Teachers

Natalia Isabel Vilor Tejedor

Angel González Wong

## Prerequisites

None. Recommended to have taken the Bioinformatics subject.

## Objectives and Contextualisation

The course aims to provide a view on the possibilities of big data analysis in bioinformatics. The course consists of two parts: 1) computational methodologies applied to drug discovery and 2) analysis of omics data. The course is part of the Mention in Statistics for Health Sciences.

## Competences

- Analyse data using statistical methods and techniques, working with data of different types.
- Correctly use a wide range of statistical software and programming languages, choosing the best one for each analysis, and adapting it to new necessities.
- Critically and rigorously assess one's own work as well as that of others.
- Formulate statistical hypotheses and develop strategies to confirm or refute them.
- Identify the usefulness of statistics in different areas of knowledge and apply it correctly in order to obtain relevant conclusions.
- Interpret results, draw conclusions and write up technical reports in the field of statistics.
- Make efficient use of the literature and digital resources to obtain information.
- Select statistical models or techniques for application in studies and real-world problems, and know the tools for validating them.
- Students must be capable of applying their knowledge to their work or vocation in a professional way and they should have building arguments and problem resolution skills within their area of study.

- Students must be capable of collecting and interpreting relevant data (usually within their area of study) in order to make statements that reflect social, scientific or ethical relevant issues.
- Students must be capable of communicating information, ideas, problems and solutions to both specialised and non-specialised audiences.
- Use quality criteria to critically assess the work done.
- Work cooperatively in a multidisciplinary context, respecting the roles of the different members of the team.

## Learning Outcomes

1. Apply statistical methods to the analysis of data on gene expression.
2. Critically assess the work done on the basis of quality criteria.
3. Design and conduct hypothesis tests in the different fields of application studied.
4. Draw conclusions that are consistent with the experimental context specific to the discipline, based on the results obtained.
5. Draw up technical reports that clearly express the results and conclusions of the study using vocabulary specific to the field of application.
6. Interpret statistical results in applied contexts.
7. Justify the choice of method for each particular application context.
8. Make effective use of references and electronic resources to obtain information.
9. Reappraise one's own ideas and those of others through rigorous, critical reflection.
10. Recognize the importance of the statistical methods studied within each particular application.
11. Recognize the statistical inference methods most commonly used in bioinformatics.
12. Students must be capable of applying their knowledge to their work or vocation in a professional way and they should have building arguments and problem resolution skills within their area of study.
13. Students must be capable of collecting and interpreting relevant data (usually within their area of study) in order to make statements that reflect social, scientific or ethical relevant issues.
14. Students must be capable of communicating information, ideas, problems and solutions to both specialised and non-specialised audiences.
15. Use different programmes, both open-source and commercial, associated with the different applied branches.
16. Work cooperatively in a multidisciplinary context, accepting and respecting the roles of the different team members.

## Content

PART 1. Big Data in Drug Design

1. Introduction to big data in drug design.
2. Structure of proteins and chemical space in small molecules (drug-like).
3. Protein-drug interactions.
4. Virtual screening.
5. Molecular dynamics.

PART 2. Big Data in Omics Analysis

1. Introduction to Bioconductor and bioinformatics tools for the analysis of omic data.
2. Genetic Association Studies and GWAS (Genome-wide association studies).
3. Multivariate Methods for Omics Data Analysis Integration and Big Data.

*Unless the requirements enforced by the health authorities demand a prioritization or reduction of these contents.

## Methodology

The course is organized in sessions of 3 hours. Each session consists of a theoretical part (theory classroom) that will introduce the new concepts followed by a practical part (computer room) where the students will work on the implementation of concepts explained in the theoretical part. In each session the teacher will indicate the students some tasks to do autonomously, such as reading articles or sending reports. The material used by the teachers will be available on the Virtual Campus of the course.

***The proposed teaching methodology may experience some modifications depending on the restrictions to face-to-face activities enforced by health authorities.***

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.

## Activities

| Title | Hours | ECTS | Learning Outcomes |
|---|---|---|---|
| Type: Directed | | | |
| Practical sessions | 21 | 0.84 | 1, 2, 3, 4, 6, 15, 8 |
| Presentation of Research Project | 3 | 0.12 | 14, 12 |
| Theory classes | 21 | 0.84 | 14, 13, 11, 10 |
| Type: Supervised | | | |
| Tutoring | 10 | 0.4 | 4 |
| Type: Autonomous | | | |
| Preparation of Research Project | 20 | 0.8 | 9, 3, 16 |
| Study | 70 | 2.8 | 12 |

## Assessment

PART 1. Big Data in Drug Design (50%):

- practical exercises (10%)

- theoretical-practical test (20%)

- bioinformatics oral presentation of a project (20%)

BLOCK 2. Big Data in Data Analysis (50%):

- practical exercises (30%)

- theoretical-practical test (20%)

The minimum global qualification required to pass the subject will be 5 points. The minimum mark of each of the evaluated activities must be equal to or greater than 4 points. Students who have any of the parts suspended will be able to do the recovery exam where they can be re-examined from the suspended part.

***Student's assessment may experience some modifications depending on the restrictions to face-to-face activities enforced by health authorities.***

## Assessment Activities

| Title | Weighting | Hours | ECTS | Learning Outcomes |
| --- | --- | --- | --- | --- |
| Presentation Research Project | 20 | 0.5 | 0.02 | 9, 5, 14, 16, 8 |
| Presentation of practicum reports | 40 | 0.5 | 0.02 | 5 |
| Theoretical-practical exams | 40 | 4 | 0.16 | 1, 2, 3, 4, 6, 7, 14, 12, 13, 11, 10, 15 |

## Bibliography

- Lesk A.M. *Introduction to Bioinformatics.* Oxford University Press 2005.
- Attwood, T.K., Parry-Smith, D.J*., Introducción a la Bioinformática.* Pearson Education, 2002.
- Foulkes A.S. *Applied Statistical Genetics with R. For Population-based Association Studies.*Springer Dordrecht Heidelberg London New York. ISBN 978-0-387-89553-6
- Gonzalez JR, Cáceres A. *Omic association studies with R and Bioconductor*. Chapman and Hall/CRC, ISBN 9781138340565, 2019.
- https://www.bioconductor.org/

## Software

R