Universidad Zaragoza

Academic Year/course: 2021/22

68461 - Big Data in Biology

Syllabus Information

Academic Year: 2021/22
Subject: 68461 - Big Data in Biology
Faculty / School: 100 - Facultad de Ciencias
Degree: 626 -
ECTS: 6.0
Year: 01
Semester: Second semester
Subject Type: Optional
Module:

1. General information

1.1. Aims of the course

The objectives of the course are the following:
Goal 1: Identify the main fields and applications in contemporary Bio-medical research where data volume, variety and generation velocity make data science approaches pivotal.
Goal 2: Understand the basics of data production in Next Generation Sequencing technologies (NGS).
Goal 3: Study the most important steps of a paradigmatic RNA-sequencing data analysis pipeline, as a key example of data science application in contemporary Biology; including QC, mapping, statistical modeling and biological interpretation.
Goal 4:Understand the basics of single-cell sequencing technologies: describe data generation techniques, data features that differ with respect to bulk sequencing data, and describe, and use, analytical approaches to deal with them.

Goal 5: Stablish a critical debate around ethical and epistemological aspects related to the role of the bio-informatician / data-scientist in contemporary computational genomics.

1.2. Context and importance of this course in the degree

This course is intended to offer a first approach to data science applications in current bio-medical research. To do that, we first provide a general overview of some of the fields where big data science approaches are used more frequently nowadays. After this theoretical contextualization, we will focus on one among such fields where data variety, volume and ever-increasing production velocity and affordability are most remarkable, and bound to deepest epistemological implications. This is the case of the study of -omics data compiled using NGS technologies. Focusing on transcriptomics as one amongst the most popular -omics data modalities, we will discuss how statistical models, applied on the analysis of these big-data-sets, both at bulk, and single-cell resolutions, can be used to shed light about fundamental aspects of cell Biology, evolutionary Biology, or immunology, among others.

From a conceptual point of view, the systems and Biological concepts studied in this course are complementary to the contents presented in the course on Systems and Synthetic Biology. The study of the central dogma of cell Biology, -and more specifically the regulation of gene expression, its variation across cell types and conditions- is central to both courses, even though the methodological toolboxes presented in each of them is totally different (Boolean and continuous models to describe time-series evolution in longitudinal designs in systems Biology; versus statistical modelling for the study of cross-sectional data cohorts in this course). From a methodological point of view, the methods presented in the course complement some of the tools presented in the course on Biostatistics and Bio-informatics.

1.3. Recommendations to take this course

The course, just like the entire master, is conceived for an interdisciplinary audience composed indistinctly by students coming from backgrounds in formal/quantitative sciences and bio-medical programs alike. In order to complete this course, it is highly recommended to prioritize developing strong R programming skills during the first semester *"Introduction to computational methods in Biology"*, as well as to choose this course along with the optional course in *"Bio-statistics and Bio-informatics"*, where many statistical concepts that are central here are also presented.

2. Learning goals

2.1. Competences

Basic and General
1. Order, analyze critically, and interpret information from different types of sources.
2. Develop the learning skills needed to continue studying autonomously new data, methods and applications.
3. Communicate results clear and unambiguously, using suitable presentation tools and with the limitations imposed by time or space.
4. Learn to manage the resources and time available for solving a problem or developing a project.
5. Use to quantitative data to discriminate complex hypotheses, and translate data analysis results into biologically articulated conclusions.
6. Develop critical judgement with respect to the results of one?s own data analyses.
7. Get acquainted with multidisciplinary research environments, and learn an efficient language suitable for communication within multidisciplinary collaborations in Bio-medicine.
Specific
1. Identify the main fields of application of data science in Biology and Medicine.
2. Acquire fluidity in basic computational management of big datasets of biological information
3. Implement complete and reproducible analysis pipelines of bulk and single cell RNA-seq data.
4. Identify the main features of an experimental design in -omics data, and translate them into optimal modeling strategies.
5. Acknowledge the ethical and societal implications of the decision-making process in biological data analysis.

2.2. Learning goals

At the end of the course, the student will know what are the main fields of application of data science in biomedical research nowadays. Furthermore, the student will know the basics concerning data production in NGS technologies, and will be able to design, and implement, a complete pipeline for the analysis of RNA-seq transcriptomic data: from QC, and mapping, to statistical modeling and critical and biological interpretation of the results. The student will recognize the main differences between bulk and single-cell sequencing data, and will be familiar with the ethical, societal, and epistemological implications of the data analysis tasks covered in the course.

2.3. Importance of learning goals

Fueled by technical developments, many fields in current bio-medical research have turned into data-intensive disciplines that require of substantial expertise regarding storage, handling, and, very importantly, analysis and interpretation of big datasets. A paramount example is that of NGS technologies, that nowadays allow retrieving genomic data of many different types for ever-increasing sample sizes at an affordable cost, which has supposed a revolution in the study of genomes, their regulation, variation within and across species and their implications in fields spanning from evolutionary Biology to clinical care. The explosive growth of the usage of NGS data in biomedical research has been followed by an intense demand, both in industry and academic environments of technicians and scientists with a simultaneous know-how in statistics, computation and data science techniques as well as a proper background in the biological concepts needed to be able not only to analyze big-datasets of biological information, but to interpret them properly from a biological point of view. The structure of this course is oriented towards the acquisition of such profile, which is under strong demand in contemporary biomedical research.

3. Assessment (1st and 2nd call)

3.1. Assessment tasks (description of tasks, marking system and assessment criteria)

1: (40% of the final grade). Continuous evaluation of the student's progress during the practical and theoretical sessions, through the correction of the practice reports, as well as through direct interaction in the classroom, rewarding active participation during the lectures and practices.
2: (10% of the final grade). Seminars on the topics proposed by the teacher, where coherence, and understanding of the subject, as well as clearness of presentation will be assessed and evaluated.
3: (50% of the final grade) Written exam, possibly including computer exercises, and/or resorting to the Moodle platform, on the topics discussed throughout the course.

4. Methodology, learning tasks, syllabus and resources

4.1. Methodological overview

The methodology followed in this course is oriented towards the achievement of the learning objectives through the implementation of a wide range of teaching and learning tasks, including lectures, practical sessions where lectures, or analogous material will be combined with the step-by-step execution of template code scripts by the teacher, and practice sessions in the computer laboratory room. The virtual platform Moodle will be used to distribute lecture notes, as well as to propose practices, and to broadcast relevant news. Students are expected to participate actively in the class throughout the semester. Course material: Notes written by the lecturers will be available on the course Moodle webpage.

4.2. Learning tasks

The course includes the following learning tasks:
- Theoretical lectures: using slides, R-markdown documents, or analogous materials, (and possibly also videoconferencing tools as required) deal with the explanation of theory and methods, organized according to the syllabus of the course.
- Practical lectures, where examples of computational implementations of the analysis described in the theory sessions will

be presented to the students, using a combination of slides, R markdown, and code scripts.
- Computer lab sessions, where students will be asked to solve specific problems, and implement analytical pipelines applying what was presented in theoretical and practical lectures.
- The presentation of a short seminars, distributed along the course, in small groups, on a topic proposed by the teacher.

4.3. Syllabus

The course is structured in seven blocks:
Block 1: Reviewing the main fields and applications of data science in Bio-medicine.
Block 2: NGS transcriptomic data as a paradigmatic example (1): Data generation, QC, and mapping.
Block 3: NGS transcriptomic data as a paradigmatic example (2): Data transformation and modeling.
Block 4: Complex designs, batch effects and empirical null models.
Block 5: Biological interpretation of modeling outcomes. Functional enrichment analyses.
Block 6: Single cell-omics technologies.
Block 7: Ethical and epistemological aspects of data analysis in Bio-medicine.

4.4. Course planning and calendar

The course is taught during 10 weeks in the second semester, indicatively from February to April.
Lectures will be held according to the schedule published on https://ciencias.unizar.es/calendario-y-horarios.
Typically, every week will include two sessions of three-hours, the first of which will be predominantly devoted to theoretical and practical lectures or seminars, whilst the second lecture will be devoted to practices. The precise dates and places will be reminded to the students via the virtual platform Moodle, so the students are advised to check their official (unizar) email account.

Evaluations of the practice sessions will take place throughout the course, structured by blocks. Seminars schedules will be agreed with the students throughout the semester. The exam sessions will be established on the dates and places reported in https://ciencias.unizar.es/consultar-examenes

4.5. Bibliography and recommended resources

http://psfunizar10.unizar.es/br13/egAsignaturas.php?codigo=68461