COURSE DATA

DATA SUBJECT

Code: 33142
Name: Bioinformatics
Cycle: Undergraduate Studies
ECTS Credits: 4.5
Academic year: 2025-26

STUDY (S)

Degree	Center	Acad. year	Period
1109 - Degree in Biochemistry and Biomedical Sciences	Facultat de Ciències Biològiques	4	Second quarter

SUBJECT-MATTER

Degree	Subject-matter	Character
1109 - Degree in Biochemistry and Biomedical Sciences	Materia de asignaturas optativas	ELECTIVES

COORDINATION

LUCAS LLEDO JOSE IGNACIO

SUMMARY

Bioinformatics is an interdisciplinary field of science that develops methods and software tools for the understanding of complex and abundant biological data, such as genomic sequences, or gene expression measurements. It has applications in biomedicine, conservation biology, systems biology, etc.

We designed the syllabus as an introduction to this wide discipline. When choosing and ordering the contents, we followed two main principles. First, that every bioinformatics analysis must be reproducible. This requirement entails the adption of certain good practices, which are also useful in other contexts. And second, we conceive bioinformatic analyses as a collective endeavour, aimed at collaborating, often in an interdisciplinary team. Thus, it is necessary to be able to communicate results in comprehensible ways, as well as being able to collaborate to produce them. This trait also conditions the fundamental habits at work with computers and the choice of tools.

The subject is oriented to practice, with almost twice as many hours in the lab than in the classroom.

PREVIOUS KNOWLEDGE

RELATIONSHIP TO OTHER SUBJECTS OF THE SAME DEGREE

There are no specified enrollment restrictions with other subjects of the curriculum.

OTHER REQUIREMENTS

There is no need for previous knowledge of Linux, nor of any programming language. The course departs from zero, without assuming any previous experience.

COMPETENCES / LEARNING OUTCOMES

-

Acceder a las principales bases de datos biológicos y recuperar y emplear la información contenida en ellas.

Aplicar correctamente los métodos de inferencia filogenética e interpretar los resultados.

Conocer los mecanismos evolutivos a escala molecular.

Conocer los métodos que permiten manejar grandes cantidades de datos derivados de las técnicas ómicas.

Know how to use the different bibliographic sources and biological databases and be able to use bioinformatic tools.

Saber utilizar los principales métodos bioinformáticos.

DESCRIPTION OF CONTENTS

1. Text files and alignments

Inconvenients of graphic user interfaces. The plain text and common formats: FASTA, FASTQ, SAM and VCF. Bash command line. Interpreted languages. Basics of R or python. Automatization of processes (scripts). Alignment of sequences.

2. Principles of data management

Inconvenients of spreadsheets. Metadata. Good practice in organizing folders. Permissions and properties of files in Linux. The XML and JSON formats as models of data bases. Notions of relational data bases. Operating with tables or data frames. Fisher's exact test.

3. Public data bases

The European Nucleotide Archive (ENA). Uniprot. Interpro. Other permanent repositories. Programatic queries through the aplication programming interfaces.

4. Visualization and exploration of data

Advanced graphics tools (e.g., ggplot2). Exploratory data analysis (PCA, biplots, etc.). Graphs.

5. Reproducibility

Pillars of reproducibility in bioinformatics. Version control with git. Tools for controlling the computation environment.

6. BLAST

The BLAST algorithm and its implementations as an online service and a command line tool. Building of a BLAST data base. Basic and advanced searches.

7. Phylogenetic inference

Markov chains and molecular evolution. Models of nucleotide substitution. Genetic distances. Neighbor-joining. Maximum likelihood. Bayesian methods.

8. Sequence assembly

Main methods and concepts related to genome assemby. De Bruijn graphs and k-mers. Quality control.

9. Genome anotation

Identification and masking of repeats. Gene prediction by homology and ab initio. Gene ontology. Formats GTF/GFF and BED. Genome browsers.

10. Gene expression analysis

RNA sequencing. Demultiplexing. Normalization. Count data. Differential expression.

11. Protein bioinformatic analyses

Prediction of secondary and tertiary structures. Protein-protein interactions. The PDB data base. Visualization of protein structures.

WORKLOAD

PRESENCIAL ACTIVITIES

Activity	Hours
Tutorials	3,00
Theory	16,00

Computer classroom practice	26,00
Total hours	**45,00**

NON PRESENCIAL ACTIVITIES

Activity	Hours
Attendance at other activities	0,00
Individual or group project	12,00
Independent study and work	31,50
Preparation of lessons	24,00
Preparation for assessment activities	0,00
Resolution of case studies	0,00
Total hours	**67,50**

TEACHING METHODOLOGY

There are two kinds of theoretical contents: the basic theory needed to practice bioinformatics (how the algorithms and the programs work, etc.) and the theoretical reasoning about the practice (what tools are more adequate, why reproducibility is important, etc.). Thus, the relationship between theory and practice will be two-ways. Theory classes, in master-class style, will inform the practice sessions. And these, in turn, will motivate theoretical reasoning. Practice sessions will combine live-coding and supervised autonomous work.

The first five units will develop the basic pragmatical skills: passive use of a programming language, version control, data manipulation and graphical representation. The automatization of processes and literate programming will be facilitated by the distribution of scripts previously prepared by instructors.

Once students become confortable with basic tools, we will address more complex bioinformatic challenges (BLAST searches, phylogenetic reconstruction, etc.). In this second stage, we will keep using the basic tools (literate programming, version control and so on), in order to reinforce the good practice in bioinformatics.

Mentoring sessions in smaller groups (tutoria) will serve to help students prepare their digital notebook, as well as to go over previously taught material or to broaden it, upon demand.

EVALUATION

There will be two evalutation tools. First, the digital practice notebook, where students will record the procedure and results of all the exercices from the practice sessions. The notebook will count 50% of the final grade.

The second instrument of evaluation will be an exam that will combine theory and practice. It will constitute the remaining 50% of the final note. In order to pass the whole subject, one must obtain a score of at least 4 out of 10 in the exam, and a global score of at least 5 out of 10.

In the second call, the evaluation criteria will be the same. Upon failure to pass in first call, any grade above 5 of either the notebook or the exam, will be respected, and only the failed part will have to be repeated.

REFERENCES

Basic References

References b1: https://software-carpentry.org/lessons/

References b2:

https://ucdavis-bioinformatics-training.github.io/2022_February_Introduction_to_R_for_Bioinformatics/

References b3: https://a-little-book-of-r-for-bioinformatics.readthedocs.io/en/latest/index.html

References b4: https://doi.org/10.18637/jss.v059.i10

References b5: https://www.ahl27.com/OtherTutorials/articles/BuildingTrees.html

References b6: https://doi.org/10.1371/journal.pcbi.1000424

References b7: https://www.ncbi.nlm.nih.gov/books/NBK279690/

References b8: Haddock, S.H.D & Dunn, C.W. 2011. Practical Computing for Biologists. Sinauer Associates, Inc. Sunderland (MA). 568 pp.

References b9: Allesina, S. & Wilmes, M. 2019. Computing skills for biologists; a toolbox. Princeton University Press. Princeton (NJ). 441 pp.

Complementary References

References c1: https://www.h3abionet.org/training/ibt

References c2: https://doi.org/10.1093/bib/bbad375

References c3: https://datasciencebox.org/

References c4: https://datacarpentry.github.io/R-ecology-lesson-alternative/index.html

References c5: https://www.ahl27.com/CompGenomicsBioc2022/

References c6: https://www2.decipher.codes/Tutorials.html

References c7: https://www.huber.embl.de/msmb/

References c8: https://doi.org/10.1080/00031305.2017.1375989