Course guide
270965 - BSG - Bioinformatics and Statistical Genetics

Last modified: 14/07/2025

Unit in charge: Barcelona School of Informatics
Teaching unit: 723 - CS - Department of Computer Science.
715 - EIO - Department of Statistics and Operations Research.

Degree: MASTER'S DEGREE IN INFORMATICS ENGINEERING (Syllabus 2012). (Optional subject).
MASTER'S DEGREE IN INNOVATION AND RESEARCH IN INFORMATICS (Syllabus 2012). (Optional subject).
MASTER'S DEGREE IN DATA SCIENCE (Syllabus 2021). (Optional subject).

Academic year: 2025 **ECTS Credits:** 6.0 **Languages:** English

LECTURER

Coordinating lecturer: MARTA JANIRA CASTELLANO PALOMINO

Others:

PRIOR SKILLS

Basic knowledge of algorithms and data structures.
Basic knowledge of statistics.
Basic knowledge of the Python programming language.
Basic knowledge of the R programming language.

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CE1. Develop efficient algorithms based on the knowledge and understanding of the computational complexity theory and considering the main data structures within the scope of data science
CE2. Apply the fundamentals of data management and processing to a data science problem
CE5. Model, design, and implement complex data systems, including data visualization
CE6. Design the Data Science process and apply scientific methodologies to obtain conclusions about populations and make decisions accordingly, from both structured and unstructured data and potentially stored in heterogeneous formats.
CE9. Apply appropriate methods for the analysis of non-traditional data formats, such as processes and graphs, within the scope of data science

Generical:
CG4. Design and implement data science projects in specific domains and in an innovative way

Transversal:
CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CT5. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.

Basic:
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.

TEACHING METHODOLOGY

All classes consist of a theoretical session (a lecture in which the professor introduces new concepts or techniques and detailed examples illustrating them) followed by a practical session (in which the students work on the examples and exercises proposed in the lecture). On the average, two hours a week are dedicated to theory and one hour a week to practice, and the professor allocates them according to the subject matter. Students are required to take an active part in class and to submit the exercises at the end of each class.

LEARNING OBJECTIVES OF THE SUBJECT

1.Introduce the student to the algorithmic, computational, and statistical problems that arise in the analysis of biological data.
2.Reinforce the knowledge of discrete structures, algorithmic techniques, and statistical techniques that the student may have from previous courses.

STUDY LOAD

Type	Hours	Percentage
Self study	96,0	64.00
Hours large group	54,0	36.00

Total learning time: 150 h

CONTENTS

Introduction to bioinformatics

Description:
Combinatorial introduction to molecular biology.

ILP and SAT in bioinformatics

Description:
Brief Introduction to ILP. Solving an integer linear program. AMPL. Brief introduction to SAT. Solving a SAT formulation. PySAT.

Longest common substring and subsequence

Description:
Longest common substring. ILP and SAT models. Longest common subsequence. RNA folding. ILP and SAT models.

Shortest common superstring and supersequence

Description:
Shortest common superstring. Genome assembly. ILP and SAT models. Shortest common supersequence. ILP and SAT models.

Sequence alignment and multiple sequence alignment

Description:
Sequence alignment. Edit distance. ILP and SAT models. Multiple sequence alignment. ILP and SAT models.

Other string selection problems

Description:
Closest string. ILP and SAT models. Closest substring. ILP and SAT models.

Introduction to statistical genetics

Description:
Basic genetic terminology. Population-based and family-based studies. Traits, markers and polymorphisms. Single nucleotide polymorphisms and microsatellites. R-package genetics.

Hardy-Weinberg equilibrium

Description:
Hardy-Weinberg law. Hardy-Weinberg assumptions. Multiple alleles. Statistical tests for Hardy-Weinberg equilibrium: chi-square, exact and likelihood-ratio tests. Graphical representations. Disequilibrium coefficients: the inbreeding coefficient, Weir's D. R-package HardyWeinberg.

Linkage disequilibrium

Description:
Definition of linkage disequilibrium (LD). Measures for LD. Estimation of LD by maximum likelihood. Haplotypes. The HapMap project. Graphics for LD. The LD heatmap.

Phase estimation

Description:
Phase ambiguity for double heterozygotes. Phase estimation with the EM algorithm. Estimation of haplotype frequencies. R-package haplo.stats.

Population substructure

Description:
Definition of population substructure. Population substructure and Hardy-Weinberg equilibrium. Population substructure and LD. Statistical methods for detecting substructure. Multidimensional scaling. Metric and non-metric multidimensional scaling. Euclidean distance matrices. Stress. Graphical representations.

Family relationships and allele sharing

Description:
Identity by state (IBS) and Identity by descent (IBD). Kinship coefficients. Allele sharing. Detection of family relationships. Graphical representations.

Genetic association analysis

Description:
Disease-marker association studies. Genetic models: dominant, co-dominant and recessive models. Testing models with chi-square tests. The alleles test and the Cochran-Armitage trend test. Genome-wide assocation tests.

ACTIVITIES

Development of syllabus topics

Specific objectives:
1, 2

Related competencies :
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.
CE5. Model, design, and implement complex data systems, including data visualization
CE2. Apply the fundamentals of data management and processing to a data science problem
CE1. Develop efficient algorithms based on the knowledge and understanding of the computational complexity theory and considering the main data structures within the scope of data science
CE6. Design the Data Science process and apply scientific methodologies to obtain conclusions about populations and make decisions accordingly, from both structured and unstructured data and potentially stored in heterogeneous formats.
CE9. Apply appropriate methods for the analysis of non-traditional data formats, such as processes and graphs, within the scope of data science
CG4. Design and implement data science projects in specific domains and in an innovative way
CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CT5. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.

Full-or-part-time: 114h
Theory classes: 15h
Laboratory classes: 24h
Self study: 75h

Final exam Bioinformatics

Specific objectives:
1, 2

Related competencies :
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.
CE5. Model, design, and implement complex data systems, including data visualization
CE2. Apply the fundamentals of data management and processing to a data science problem
CE1. Develop efficient algorithms based on the knowledge and understanding of the computational complexity theory and considering the main data structures within the scope of data science
CE6. Design the Data Science process and apply scientific methodologies to obtain conclusions about populations and make decisions accordingly, from both structured and unstructured data and potentially stored in heterogeneous formats.
CE9. Apply appropriate methods for the analysis of non-traditional data formats, such as processes and graphs, within the scope of data science
CG4. Design and implement data science projects in specific domains and in an innovative way
CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CT5. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.

Full-or-part-time: 18h
Guided activities: 3h
Self study: 15h

Final exam Statistical Genetics

Specific objectives:
1, 2

Related competencies :
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.
CE5. Model, design, and implement complex data systems, including data visualization
CE2. Apply the fundamentals of data management and processing to a data science problem
CE1. Develop efficient algorithms based on the knowledge and understanding of the computational complexity theory and considering the main data structures within the scope of data science
CE6. Design the Data Science process and apply scientific methodologies to obtain conclusions about populations and make decisions accordingly, from both structured and unstructured data and potentially stored in heterogeneous formats.
CE9. Apply appropriate methods for the analysis of non-traditional data formats, such as processes and graphs, within the scope of data science
CG4. Design and implement data science projects in specific domains and in an innovative way
CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CT5. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.

Full-or-part-time: 18h
Guided activities: 3h
Self study: 15h

GRADING SYSTEM

For the first half (Bioinformatics), students are evaluated in a mid-term exam, in which they model and solve new string problems in Bioinformatics using ILP and SAT. In the second half (Statistical Genetics), students are evaluated during class, and in a final exam. Every student is required to submit one exercise each week, graded from 0 to 10, and the final grade consists of 50% for the exercises and 50% for the final exam, also graded from 0 to 10.

BIBLIOGRAPHY

Basic:
- Gusfield, Dan. Integer linear programming in computational and systems biology : an entry-level text and course. Cambridge University Press, [2019]. ISBN 9781108421768.
- Foulkes, Andrea S. Applied Statistical Genetics with R: For Population-based Association Studies. New York, NY: Springer, 2009. ISBN 9780387895536.
- Laird, Nan M.; Lange, Christoph. The Fundamentals of modern statistical genetics. New York: Springer, 2011. ISBN 9781461427759.

Complementary:
- Pappalardo, Elisa; Pardalos, P. M; Stracquadanio, Giovanni. Optimization Approaches for Solving String Selection Problems [Recurs electrònic]. New York: Springer, 2013. ISBN 9781461490531.
- Ziegler, Andreas; König, Inke R.. Statistical Approach to Genetic Epidemiology. 2nd ed. Weinheim an der Bergstrasse, Germany: Wiley, 2011. ISBN 9783527633654.