Course guide
270656 - BSG - Bioinformatics and Statistical Genetics

Last modified: 12/07/2021

Unit in charge:	Barcelona School of Informatics
Teaching unit:	715 - EIO - Department of Statistics and Operations Research. 723 - CS - Department of Computer Science.
Degree:	MASTER'S DEGREE IN INNOVATION AND RESEARCH IN INFORMATICS (Syllabus 2012). (Optional subject).

Academic year: 2021 **ECTS Credits:** 6.0 **Languages:** English

LECTURER

Coordinating lecturer:	GABRIEL ALEJANDRO VALIENTE FERUGLIO
Others:	Primer quadrimestre: JAN GRAFFELMAN - 10 GABRIEL ALEJANDRO VALIENTE FERUGLIO - 10

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:
CEC1. Ability to apply scientific methodologies in the study and analysis of phenomena and systems in any field of Information Technology as well as in the conception, design and implementation of innovative and original computing solutions.
CEC2. Capacity for mathematical modelling, calculation and experimental design in engineering technology centres and business, particularly in research and innovation in all areas of Computer Science.
CEC3. Ability to apply innovative solutions and make progress in the knowledge that exploit the new paradigms of Informatics, particularly in distributed environments.

Generical:
CG3. Capacity for mathematical modeling, calculation and experimental designing in technology and companies engineering centers, particularly in research and innovation in all areas of Computer Science.

Transversal:
CTR6. REASONING: Capacity for critical, logical and mathematical reasoning. Capability to solve problems in their area of study. Capacity for abstraction: the capability to create and use models that reflect real situations. Capability to design and implement simple experiments, and analyze and interpret their results. Capacity for analysis, synthesis and evaluation.

Basic:
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.
CB9. Possession of the learning skills that enable the students to continue studying in a way that will be mainly self-directed or autonomous.

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

Total learning time: 150 h

CONTENTS

Introduction to bioinformatics

Description:
Computational biology and bioinformatics. Algorithms in bioinformatics. Strings, sequences, trees, and graphs. Algorithms on strings and sequences. Representation of trees and graphs. Algorithms on trees and graphs.

Phylogenetic reconstruction I

Description:
Character-based phylogenetic reconstruction. Compatibility. Perfect phylogenies. Distance-based phylogenetic reconstruction. Additive trees. Ultrametric trees.

Agreement of phylogenetic trees

Description:
Partition distance. Triplets distance. Quartets distance. Transposition distance. Edit distance and alignment of phylogenetic trees.

Phylogenetic reconstruction II

Description:
Phylogenetic networks. Galled trees. Tree-child networks. Tree-sibling networks. Time consistency of phylogenetic networks.

Agreement of phylogenetic networks

Description:
Path multiplicity distance. Tripartition distance. Nodal distance. Triplets distance. Edit distance and alignment of phylogenetic networks.

Phylogenetic reconstruction III

Description:
Mutation trees. Clonal trees. Clonal deconvolution.

Phylogenetic and taxonomic reconstruction

Description:
Phylogenies and taxonomies. Classification of metagenomic samples. Agreement of classifications.

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.

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.

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.

ACTIVITIES

Development of syllabus topics

Specific objectives:
1, 2

Related competencies :
CG3. Capacity for mathematical modeling, calculation and experimental designing in technology and companies engineering centers, particularly in research and innovation in all areas of Computer Science.
CEC3. Ability to apply innovative solutions and make progress in the knowledge that exploit the new paradigms of Informatics, particularly in distributed environments.
CEC2. Capacity for mathematical modelling, calculation and experimental design in engineering technology centres and business, particularly in research and innovation in all areas of Computer Science.
CEC1. Ability to apply scientific methodologies in the study and analysis of phenomena and systems in any field of Information Technology as well as in the conception, design and implementation of innovative and original computing solutions.
CTR6. REASONING: Capacity for critical, logical and mathematical reasoning. Capability to solve problems in their area of study. Capacity for abstraction: the capability to create and use models that reflect real situations. Capability to design and implement simple experiments, and analyze and interpret their results. Capacity for analysis, synthesis and evaluation.
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.
CB9. Possession of the learning skills that enable the students to continue studying in a way that will be mainly self-directed or autonomous.
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.

Full-or-part-time: 117h
Theory classes: 15h
Laboratory classes: 30h
Self study: 72h

Final exam

Full-or-part-time: 33h
Theory classes: 3h
Self study: 30h

GRADING SYSTEM

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:
- Valiente, Gabriel. Algorithms on trees and graphs. 2nd ed. Springer Nature, 2021. ISBN 9783030818845.
- Valiente, Gabriel. Combinatorial pattern matching algorithms in computational biology using Perl and R. Boca Raton: Chapman and Hall/CRC, 2009. ISBN 9781420069730.
- Foulkes, Andrea S. Applied statistical genetics with R : for population-based association studies. New York: Springer, 2009. ISBN 9780387895536.
- Laird, Nan M.; Lange, Christoph. The Fundamentals of Modern Statistical Genetics. Springer, 2011. ISBN 9781441973375.

Complementary:
- Gusfield, Dan. Algorithms on strings, trees, and sequences : computer science and computational biology. Cambridge [England]: Cambridge University Press, 1997. ISBN 0521585198.
- Paradis, Emmanuel. Analysis of phylogenetics and evolution with R [on line]. Second edition. New York: Springer, 2012 [Consultation: 16/07/2021]. Available on: https://doi.org/10.1007/978-1-4614-1743-9. ISBN 9781461417439.
- Weir, B.S. Genetic data analysis II: methods for discrete population genetic data. Sinauer Associates, 1996. ISBN 0878939024.
- Ziegler, Andreas; König, Inke R.. Statistical Approach to Genetic Epidemiology [on line]. 2nd ed. Weinheim: Wiley-VCH, 2011 [Consultation: 16/07/2021]. Available on: https://onlinelibrary-wiley-com/doi/book/10.1002/9783527633654. ISBN 9783527633654.

RESOURCES

Hyperlink:
- http://rosalind.info/
- http://www.r-project.org/