Guia docent
270965 - BSG - Bioinformàtica i Genètica Estadística

Última modificació: 23/11/2023

Unitat responsable: Facultat d'Informàtica de Barcelona
Unitat que imparteix: 723 - CS - Departament de Ciències de la Computació.
715 - EIO - Departament d'Estadística i Investigació Operativa.

Titulació: MÀSTER UNIVERSITARI EN CIÈNCIA DE DADES (Pla 2021). (Assignatura optativa).

Curs: 2023 **Crèdits ECTS:** 6.0 **Idiomes:** Anglès

PROFESSORAT

Professorat responsable: GABRIEL ALEJANDRO VALIENTE FERUGLIO

Altres: Primer quadrimestre:
MARTA JANIRA CASTELLANO PALOMINO - 10
GABRIEL ALEJANDRO VALIENTE FERUGLIO - 10

CAPACITATS PRÈVIES

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.

COMPETÈNCIES DE LA TITULACIÓ A LES QUALS CONTRIBUEIX L'ASSIGNATURA

Específiques:
CE1. Desenvolupar algoritmes eficients fonamentats en el coneixement i comprensió de la teoria de la complexitat computacional i les principals estructures de dades, dins de l'àmbit de ciència de dades
CE2. Aplicar els fonaments de la gestió i processament de dades en un problema de ciència de dades
CE5. Modelar, dissenyar i implementar sistemes complexos de dades, incloent-hi la visualització de dades
CE6. Dissenyar el procés de Ciència de Dades i aplicar metodologies científiques per a obtenir conclusions sobre poblacions i prendre decisions en conseqüència, a partir de dades estructurades o no estructurades i potencialment emmagatzemades en formats heterogenis.
CE9. Aplicar mètodes adequats per a l'anàlisi d'altres tipus de formats, com ara processos i grafs, dins l'àmbit de ciència de dades

Genèriques:
CG4. Dissenyar i posar en marxa projectes de ciència de dades en dominis específics de forma innovadora

Transversals:
CT4. ÚS SOLVENT DELS RECURSOS D'INFORMACIÓ: Gestionar l'adquisició, l'estructuració, l'anàlisi i la visualització de dades i informació de l'àmbit d'especialitat, i valorar de forma crítica els resultats d'aquesta gestió.
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

Bàsiques:
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.
CB6. Que els estudiants sàpiguen aplicar els coneixements adquirits y la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contexts més amplis (o multidisciplinaris) relacionats amb la seva àrea d'estudi.
CB7. Que els estudiants siguin capaços d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, essent incomplerta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.

METODOLOGIES DOCENTS

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.

OBJECTIUS D'APRENENTATGE DE L'ASSIGNATURA

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.

HORES TOTALS DE DEDICACIÓ DE L'ESTUDIANTAT

Tipus	Hores	Percentatge
Hores grup gran	54,0	36.00
Hores aprenentatge autònom	96,0	64.00

Dedicació total: 150 h

CONTINGUTS

Introduction to bioinformatics

Descripció:
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

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

Agreement of phylogenetic trees

Descripció:
Partition distance. Nodal distance. Triplets distance. Transposition distance. Edit distance. Alignment of phylogenetic trees.

Phylogenetic reconstruction II

Descripció:
Phylogenetic networks. Galled trees. Tree-child networks. Tree-sibling networks. Time consistency of phylogenetic networks. A hierarchy of phylogenetic networks.

Phylogenetic reconstruction III

Descripció:
Phylogenies and taxonomies. Classification of metagenomic samples. The taxonomic assignment problem. Accuracy and coverage. The LCA skeleton tree.

Agreement of phylogenetic networks

Descripció:
Path multiplicity distance. Tripartition distance. Nodal distance. Triplets distance. Edit distance. Alignment of phylogenetic networks.

Introduction to statistical genetics

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

Hardy-Weinberg equilibrium

Descripció:
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

Descripció:
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

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

Population substructure

Descripció:
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

Descripció:
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

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

ACTIVITATS

Development of syllabus topics

Objectius específics:
1, 2

Competències relacionades:
CG4. Dissenyar i posar en marxa projectes de ciència de dades en dominis específics de forma innovadora
CE2. Aplicar els fonaments de la gestió i processament de dades en un problema de ciència de dades
CE5. Modelar, dissenyar i implementar sistemes complexos de dades, incloent-hi la visualització de dades
CE1. Desenvolupar algoritmes eficients fonamentats en el coneixement i comprensió de la teoria de la complexitat computacional i les principals estructures de dades, dins de l'àmbit de ciència de dades
CE6. Dissenyar el procés de Ciència de Dades i aplicar metodologies científiques per a obtenir conclusions sobre poblacions i prendre decisions en conseqüència, a partir de dades estructurades o no estructurades i potencialment emmagatzemades en formats heterogenis.
CE9. Aplicar mètodes adequats per a l'anàlisi d'altres tipus de formats, com ara processos i grafs, dins l'àmbit de ciència de dades
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

CT4. ÚS SOLVENT DELS RECURSOS D'INFORMACIÓ: Gestionar l'adquisició, l'estructuració, l'anàlisi i la visualització de dades i informació de l'àmbit d'especialitat, i valorar de forma crítica els resultats d'aquesta gestió.
CB7. Que els estudiants siguin capaços d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, essent incomplerta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.
CB6. Que els estudiants sàpiguen aplicar els coneixements adquirits y la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contexts més amplis (o multidisciplinaris) relacionats amb la seva àrea d'estudi.
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.

Dedicació: 114h
Grup gran/Teoria: 15h
Grup petit/Laboratori: 24h
Aprenentatge autònom: 75h

Final exam Bioinformatics

Objectius específics:
1, 2

Competències relacionades:
CG4. Dissenyar i posar en marxa projectes de ciència de dades en dominis específics de forma innovadora
CE2. Aplicar els fonaments de la gestió i processament de dades en un problema de ciència de dades
CE5. Modelar, dissenyar i implementar sistemes complexos de dades, incloent-hi la visualització de dades
CE1. Desenvolupar algoritmes eficients fonamentats en el coneixement i comprensió de la teoria de la complexitat computacional i les principals estructures de dades, dins de l'àmbit de ciència de dades
CE6. Dissenyar el procés de Ciència de Dades i aplicar metodologies científiques per a obtenir conclusions sobre poblacions i prendre decisions en conseqüència, a partir de dades estructurades o no estructurades i potencialment emmagatzemades en formats heterogenis.
CE9. Aplicar mètodes adequats per a l'anàlisi d'altres tipus de formats, com ara processos i grafs, dins l'àmbit de ciència de dades
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

CT4. ÚS SOLVENT DELS RECURSOS D'INFORMACIÓ: Gestionar l'adquisició, l'estructuració, l'anàlisi i la visualització de dades i informació de l'àmbit d'especialitat, i valorar de forma crítica els resultats d'aquesta gestió.
CB7. Que els estudiants siguin capaços d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, essent incomplerta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.
CB6. Que els estudiants sàpiguen aplicar els coneixements adquirits y la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contexts més amplis (o multidisciplinaris) relacionats amb la seva àrea d'estudi.
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.

Dedicació: 18h
Activitats dirigides: 3h
Aprenentatge autònom: 15h

Final exam Statistical Genetics

Objectius específics:
1, 2

Competències relacionades:
CG4. Dissenyar i posar en marxa projectes de ciència de dades en dominis específics de forma innovadora
CE2. Aplicar els fonaments de la gestió i processament de dades en un problema de ciència de dades
CE5. Modelar, dissenyar i implementar sistemes complexos de dades, incloent-hi la visualització de dades
CE1. Desenvolupar algoritmes eficients fonamentats en el coneixement i comprensió de la teoria de la complexitat computacional i les principals estructures de dades, dins de l'àmbit de ciència de dades
CE6. Dissenyar el procés de Ciència de Dades i aplicar metodologies científiques per a obtenir conclusions sobre poblacions i prendre decisions en conseqüència, a partir de dades estructurades o no estructurades i potencialment emmagatzemades en formats heterogenis.
CE9. Aplicar mètodes adequats per a l'anàlisi d'altres tipus de formats, com ara processos i grafs, dins l'àmbit de ciència de dades
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

CT4. ÚS SOLVENT DELS RECURSOS D'INFORMACIÓ: Gestionar l'adquisició, l'estructuració, l'anàlisi i la visualització de dades i informació de l'àmbit d'especialitat, i valorar de forma crítica els resultats d'aquesta gestió.
CB7. Que els estudiants siguin capaços d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, essent incomplerta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.
CB6. Que els estudiants sàpiguen aplicar els coneixements adquirits y la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contexts més amplis (o multidisciplinaris) relacionats amb la seva àrea d'estudi.
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.

Dedicació: 18h
Activitats dirigides: 3h
Aprenentatge autònom: 15h

SISTEMA DE QUALIFICACIÓ

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.

BIBLIOGRAFIA

Bàsica:
- Valiente, Gabriel. Algorithms on Trees and Graphs. 2nd ed. Cham: 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, NY: Springer, 2009. ISBN 9780387895536.
- Laird, Nan M.; Lange, Christoph. The Fundamentals of modern statistical genetics. New York: Springer, 2011. ISBN 9781461427759.

Complementària:
- Gusfield, Dan. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambbridge [England]; New York: Cambridge University Press, 1997. ISBN 9780521585194.
- Paradis, Emmanuel. Analysis of phylogenetics and evolution with R [en línia]. 2nd ed. Springer, 2012 [Consulta: 10/01/2024]. Disponible a: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=884307. ISBN 9781461417439.
- Ziegler, Andreas; König, Inke R.. Statistical Approach to Genetic Epidemiology. 2nd ed. Weinheim an der Bergstrasse, Germany:

Wiley, 2011. ISBN 9783527633654.

RECURSOS

Enllaç web:
- http://rosalind.info/. Rosalind
- http://www.r-project.org/. The R Project for Statistical Computing