Guía docente
270965 - BSG - Bioinformática y Genética Estadística

Última modificación: 23/11/2023

Unidad responsable: Facultad de Informática de Barcelona
Unidad que imparte: 723 - CS - Departamento de Ciencias de la Computación.
715 - EIO - Departamento de Estadística e Investigación Operativa.

Titulación: MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS (Plan 2021). (Asignatura optativa).

Curso: 2023 **Créditos ECTS:** 6.0 **Idiomas:** Inglés

PROFESORADO

Profesorado responsable: GABRIEL ALEJANDRO VALIENTE FERUGLIO

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

CAPACIDADES PREVIAS

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.

COMPETENCIAS DE LA TITULACIÓN A LAS QUE CONTRIBUYE LA ASIGNATURA

Específicas:
CE1. Desarrollar algoritmos eficientes basados en el conocimiento y comprensión de la teoría de la complejidad computacional y las principales estructuras de datos dentro del ámbito de ciencia de datos
CE2. Aplicar los fundamentos de la gestión y procesamiento de datos en un problema de ciencia de datos
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos
CE6. Diseñar el proceso de Ciencia de Datos y aplicar metodologías científicas para obtener conclusiones sobre poblaciones y tomar decisiones en consecuencia, a partir de datos estructurados o no estructurados y potencialmente almacenados en formatos heterogéneos.
CE9. Aplicar métodos adecuados para el análisis de otro tipo de formatos, tales como procesos y grafos, dentro del ámbito de ciencia de datos

Genéricas:
CG4. Diseñar y poner en marcha proyectos de ciencia de datos en dominios específicos de forma innovadora

Transversales:
CT4. USO SOLVENTE DE LOS RECURSOS DE INFORMACIÓN: Gestionar la adquisicion, la estructuracion, el analisis y la visualizacion de datos e informacion en el ambito de la especialidad y valorar de forma critica los resultados de esta gestion.
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

Básicas:
CB10. Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.
CB6. Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.
CB7. Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios

METODOLOGÍAS DOCENTES

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.

OBJETIVOS DE APRENDIZAJE DE LA ASIGNATURA

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.

HORAS TOTALES DE DEDICACIÓN DEL ESTUDIANTADO

Tipo	Horas	Porcentaje
Horas aprendizaje autónomo	96,0	64.00
Horas grupo grande	54,0	36.00

Dedicación total: 150 h

CONTENIDOS

Introduction to bioinformatics

Descripción:
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

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

Agreement of phylogenetic trees

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

Phylogenetic reconstruction II

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

Phylogenetic reconstruction III

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

Agreement of phylogenetic networks

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

Introduction to statistical genetics

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

Hardy-Weinberg equilibrium

Descripción:
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

Descripción:
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

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

Population substructure

Descripción:
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

Descripción:
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

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

ACTIVIDADES

Development of syllabus topics

Objetivos específicos:
1, 2

Competencias relacionadas:
CG4. Diseñar y poner en marcha proyectos de ciencia de datos en dominios específicos de forma innovadora
CE2. Aplicar los fundamentos de la gestión y procesamiento de datos en un problema de ciencia de datos
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos
CE1. Desarrollar algoritmos eficientes basados en el conocimiento y comprensión de la teoría de la complejidad computacional y las principales estructuras de datos dentro del ámbito de ciencia de datos
CE6. Diseñar el proceso de Ciencia de Datos y aplicar metodologías científicas para obtener conclusiones sobre poblaciones y tomar decisiones en consecuencia, a partir de datos estructurados o no estructurados y potencialmente almacenados en formatos heterogéneos.
CE9. Aplicar métodos adecuados para el análisis de otro tipo de formatos, tales como procesos y grafos, dentro del ámbito de ciencia de datos
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

CT4. USO SOLVENTE DE LOS RECURSOS DE INFORMACIÓN: Gestionar la adquisicion, la estructuracion, el analisis y la visualizacion de datos e informacion en el ambito de la especialidad y valorar de forma critica los resultados de esta gestion.
CB7. Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios
CB6. Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.
CB10. Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.

Dedicación: 114h
Grupo grande/Teoría: 15h
Grupo pequeño/Laboratorio: 24h
Aprendizaje autónomo: 75h

Final exam Bioinformatics

Objetivos específicos:
1, 2

Competencias relacionadas:
CG4. Diseñar y poner en marcha proyectos de ciencia de datos en dominios específicos de forma innovadora
CE2. Aplicar los fundamentos de la gestión y procesamiento de datos en un problema de ciencia de datos
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos
CE1. Desarrollar algoritmos eficientes basados en el conocimiento y comprensión de la teoría de la complejidad computacional y las principales estructuras de datos dentro del ámbito de ciencia de datos
CE6. Diseñar el proceso de Ciencia de Datos y aplicar metodologías científicas para obtener conclusiones sobre poblaciones y tomar decisiones en consecuencia, a partir de datos estructurados o no estructurados y potencialmente almacenados en formatos heterogéneos.
CE9. Aplicar métodos adecuados para el análisis de otro tipo de formatos, tales como procesos y grafos, dentro del ámbito de ciencia de datos
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

CT4. USO SOLVENTE DE LOS RECURSOS DE INFORMACIÓN: Gestionar la adquisicion, la estructuracion, el analisis y la visualizacion de datos e informacion en el ambito de la especialidad y valorar de forma critica los resultados de esta gestion.
CB7. Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios
CB6. Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.
CB10. Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.

Dedicación: 18h
Actividades dirigidas: 3h
Aprendizaje autónomo: 15h

Final exam Statistical Genetics

Objetivos específicos:
1, 2

Competencias relacionadas:
CG4. Diseñar y poner en marcha proyectos de ciencia de datos en dominios específicos de forma innovadora
CE2. Aplicar los fundamentos de la gestión y procesamiento de datos en un problema de ciencia de datos
CE5. Modelar, diseñar e implementar sistemas complejos de datos, incluyendo la visualización de datos
CE1. Desarrollar algoritmos eficientes basados en el conocimiento y comprensión de la teoría de la complejidad computacional y las principales estructuras de datos dentro del ámbito de ciencia de datos
CE6. Diseñar el proceso de Ciencia de Datos y aplicar metodologías científicas para obtener conclusiones sobre poblaciones y tomar decisiones en consecuencia, a partir de datos estructurados o no estructurados y potencialmente almacenados en formatos heterogéneos.
CE9. Aplicar métodos adecuados para el análisis de otro tipo de formatos, tales como procesos y grafos, dentro del ámbito de ciencia de datos
CT5. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

CT4. USO SOLVENTE DE LOS RECURSOS DE INFORMACIÓN: Gestionar la adquisicion, la estructuracion, el analisis y la visualizacion de datos e informacion en el ambito de la especialidad y valorar de forma critica los resultados de esta gestion.
CB7. Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios
CB6. Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.
CB10. Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.

Dedicación: 18h
Actividades dirigidas: 3h
Aprendizaje autónomo: 15h

SISTEMA DE CALIFICACIÓN

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.

BIBLIOGRAFÍA

Básica:
- 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.

Complementaria:
- 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 línea]. 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=8843

07. ISBN 9781461417439.
- Ziegler, Andreas; König, Inke R.. Statistical Approach to Genetic Epidemiology. 2nd ed. Weinheim an der Bergstrasse, Germany: Wiley, 2011. ISBN 9783527633654.

RECURSOS

Enlace web:
- http://rosalind.info/- http://www.r-project.org/