Universitat Autònoma de Barcelona

2020/2021

Computación de Altas Prestaciones y Análisis de Big Data

Código: 43917
Créditos ECTS: 12

Titulación	Tipo	Curso	Semestre
4313473 Bioinformática / Bioinformatics	OT	0	1

La metodología docente y la evaluación propuestas en la guía pueden experimentar alguna modificación en función de las restricciones a la presencialidad que impongan las autoridades sanitarias.

Contacto

Nombre: Miquel Àngel Senar Rosell

Correo electrónico: MiquelAngel.Senar@uab.cat

Uso de idiomas

Lengua vehicular mayoritaria: inglés (eng)

Equipo docente

Juan Carlos Moure Lopez

Santiago Marco Sola

Equipo docente externo a la UAB

Emanuele Raineri

Oscar Lao

Prerequisitos

Para cursar esta asignatura deben haberse superado previamente los dos módulos obligatorios: Programming in Bioinformatics y Core Bioinformatics.

Se recomienda disponer del nivel B2 (o equivalente) de inglés.

Objetivos y contextualización

Este módulo tiene como objetivo proporcionar a los estudiantes los conocimientos y habilidades necesarios (1) para implementar aproximaciones de ingeniería de rendimiento en plataformas informáticas modernas y (2) para realizar análisis estadísticos de Big Data.

Competencias

- Comunicar en lengua inglesa de manera clara y efectiva los resultados de sus investigaciones.
- Diseñar y aplicar la metodología científica en la resolución de problemas.
- 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.
- Proponer soluciones bioinformáticas a problemas derivados de las investigaciones ómicas.
- Proponer soluciones innovadoras y emprendedoras en su campo de estudio.

- Utilizar sistemas operativos, programas y herramientas de uso común en bioinformática, así como, manejar plataformas de cómputo de altas prestaciones, lenguajes de programación y análisis bioinformáticos.
- Utilizar y gestionar información bibliográfica y recursos informáticos en el ámbito de estudio.

Resultados de aprendizaje

1. Aplicar métodos estadísticos avanzados (aprendizaje automático, teoría de grafos) para modelar y analizar problemas bioinformáticos que manejan datos biológicos masivos.
2. Aprender a entrenar, evaluar y validar modelos predictivos.
3. Aprender a manejar las nuevas plataformas de cómputo paralelo, paradigmas, y el diseño de aplicaciones que requieren un manejo masivo de cómputo y datos.
4. Aprender nuevas formas de modelar, almacenar, recuperar y analizar tipos de datos abstractos (grafos).
5. Comunicar en lengua inglesa de manera clara y efectiva los resultados de sus investigaciones.
6. Conocer los principios de la paralelización de procesos.
7. Conocer los principios del almacenamiento y la gestión de datos masivos.
8. Conocer y aprender a manejar herramientas de código abierto para el análisis paralelo, distribuido y escalable mediante aprendizaje automático.
9. Describir el funcionamiento, características y limitaciones de las técnicas, las herramientas y las metodologías que permiten describir, analizar e interpretar la enorme cantidad de datos producidos por las tecnologías de alto rendimiento.
10. Describir y aplicar técnicas de agrupamiento (clustering) y algoritmos de clasificación comunes.
11. Diseñar y aplicar la metodología científica en la resolución de problemas.
12. Generar algoritmos de computación paralela eficientes y aplicaciones para la CID.
13. 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.
14. Proponer soluciones innovadoras y emprendedoras en su campo de estudio.
15. Proporcionar soluciones paralelas a problemas bioinformáticos concretos.
16. Utilizar y gestionar información bibliográfica y recursos informáticos en el ámbito de estudio.

Contenido

Arquitectura Moderna de Ordenadores

- Arquitectura de procesadores de uso general y especializado
- Jerarquía de memoria
- Sistemas de clúster
- Infraestructuras en la nube y virtualización de sistemas
- Sistema *Middleware* y marcos de programación

Modelos de Programación Avanzada

- Memoria compartida y programación paralela distribuida
- Usando herramientas del sistemas para análisis bioinformáticos
- Shell scripting avanzado
- Principios de ingeniería de rendimiento (herramientas y métodos)
- Computación de Altas Prestaciones con Python
- Ingeniería de rendimiento aplicada a algoritmos y herramientas comunes de bioinformática (indexación del genoma, alineamiento de *reads*, ...)

Análisis de ***Big Data***

- Teoría y herramientas de estadística avanzada en análisis de *Big Data* (reducción de dimensionalidad, selección de variables y Spark)
- Teoría y algoritmos de *machine learning*. Aplicaciones en bioinformática
- Modelado predictivo: minería de datos, evaluación y validación de modelos

- Clasificación de datos: aprendizaje de Bayes ingenuo y árboles de decisión
- Aprendizaje de reglas de associación
- Análisis de *clusterización*: algoritmo *k-means*
- Teoría de grafos para *Big Data*

**A menos que las restricciones impuestas por las autoridades sanitarias obliguen a una priorización o reducción de estos contenidos.*

Metodología

Siguiendo una aproximación basada en problemas, el alumnado aprenderá sobre algoritmos, métodos y plataformas computacionales eficientes y los métodos estadísticos que se aplicarán a los desafiantes problemas de bioinformática que tratan con Big Data.

**La metodología docente propuesta puede experimentar alguna modificación en función de las restricciones a la presencialidad que impongan las autoridades sanitarias.*

Actividades

Título	Horas	ECTS	Resultados de aprendizaje
Tipo: Dirigidas			
Clases teóricas	38	1,52	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 15, 14, 13, 16
Resolución de problemas en clase y tareas en el laboratorio biocomputacional	32	1,28	1, 2, 3, 4, 6, 7, 8, 9, 10, 12, 15, 13
Tipo: Autónomas			
Estudio autónomo individual	226	9,04	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 14, 16

Evaluación

El sistema de evaluación está organizado en dos actividades principales. Habrá, además, un examen de recuperación. Los detalles de las actividades son:

Actividades de evaluación principales

- Portafolio del estudiante (60%): trabajos hechos y presentados por el alumno a lo largo del curso. Ninguna de las actividades de evaluación individuales representará más del 50% de la nota final.
- Prueba teórica y práctica individual (40%): habrá un examen al final de este módulo.

Examen de recuperación

Para poder participar en el proceso de recuperación, el alumno deberá previamente haber participado en como mínimo el equivalente a dos tercios de la nota final del módulo en actividades de evaluación. El profesorado informará de los procedimientos y plazos para el proceso de recuperación.

No evaluable

El alumno será calificado como "No evaluable" cuando el peso de la evaluación en la que ha participado sea inferior al equivalente al 67% de la nota final del módulo.

**La evaluación propuesta puede experimentar alguna modificación en función de las restricciones a la presencialidad que impongan las autoridades sanitarias.*

Actividades de evaluación

Título	Peso	Horas	ECTS	Resultados de aprendizaje
Prueba teórica y práctica individual	40%	4	0,16	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 14, 13
Trabajos hechos y presentados por el alumnado (portafolio del estudiante)	60%	0	0	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 14, 13, 16

Bibliografía

El profesor recomendará la bibliografía actualizada en cada sesión de este módulo, y los enlaces se pondrán a disposición en el Área del Estudiante del sitio web oficial de MSc Bioinformatics.