UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Elementi di Bioinformatica

2627-3-E3101Q116

Aims

The student will know some fundamental problems and algorithms in bioinformatics. The student will be able to write small Python programs to solve some problems in bioinformatics, using also data originating from publicly available databases.

Knowledge and understanding

This course provides basic knowledge and understanding on:

- Algorithms on biological sequences.
- Data structures to index biological sequencing.
- Algorithms for phylogeny reconstruction.
- Python programming in bioinformatics.
- File Formats used in bioinformatics.

Applied Knowledge and Understanding

At the end of the course the students will be able to:

- decide which algorithms and data structures can be used to solve some problems in bioinformatics.
- write Python programs for bioinformatics problems.
- write small pipelines that process bioinformatics data.

Independent Judgment

At the end of the course, the students will be able to interpret the results of a bioinformatics analysis.

Communication skills

At the end of the course, the students will be able to communicate the results of a bioinformatics analysis.

Learning skills

At the end of the course, the students will have developed the ability to:

- learn new libraries and software tools in bioinformatics.
- tackle new computational problems in genomics.

Contents

Fundamental problems and algorithms in bioinformatics. Pattern matching.Sequence Alignment. DNA sequencing. Evolutionary histories. Managing biological data (standard formats in Bioinformatics) and genomic databases.

Detailed program

1. Introduction to genomic data: primary sequences of DNA, RNA, and protein sequences (standard FASTA format).
2. Gene expression and alternative splicing (standard BED and GTF formats).
3. Genome sequencing and sequencing data quality (standard FASTQ format).
4. Global and local alignment of two sequences: Needleman–Wunsch, Smith–Waterman, and Gotoh algorithms.
5. Semi-global alignment of two sequences.
6. Banded alignment of two sequences.
7. Multiple sequence alignment.
8. Exact pattern matching: Rabin–Karp and Dömölki algorithms (shift-and paradigm).
9. Text indexing: suffix trees, suffix arrays, and their applications to string problems (including exact pattern matching and longest common substring search).
10. Genome assembly: overlap graphs and de Bruijn graphs.
11. Evolutionary trees (phylogenetic trees). Character-based models: Gusfield's algorithm for perfect phylogeny. Distance-based models: ultrametric and additive distances. UPGMA and Neighbor Joining algorithms. Introduction to maximum likelihood reconstruction.
12. Genotypes and haplotypes. Distinction between a single individual and a pedigree.
13. Introduction to Python.
14. Pandas (data frame manipulation).
15. Biopython.

Prerequisites

1. Time and space complexity.
2. Basic data structures: lists, arrays, search trees, dictionaries.
3. Sorting algorithms: radix sort, merge sort.
4. Memory hierarchy.
5. Wiriting a short program, in any programming language.

Teaching form

Lectures and Laboratory. The individual study can use the e-learning platform to enrich the standard activities and to self assess the level of competence acquired during the course.

All activities are in-person and will be recorded but not streamed. The teaching language of this course is Italian. The activities will be:

- 36 hours of lectures, delivered in 2–3-hour sessions in unidirectional mode
- 24 lab activities, 2 hours each, with an initial part in unidirectional mode and a second part in interactive mode

Textbook and teaching resource

The adopted textbook is “Algorithms on Strings, Trees and Sequences”, by Daniel Gusfield, Cambridge Univ. Press. The library has some copies, also as ebook.

The book "An Introduction to Bioinformatics Algorithms" by N. Jones, P. Pevzner is used only for some parts on phylogeny reconstruction and on genome sequencing.

The books "Theoretical Evolutionary Genetics" by J. Felsenstein and Population and Quantitative Genetics by Graham Coop are used for some topics on phylogeny reconstruction and on haplotypes.

The book Think Python by A. B. Downey is used for introducing the Python language.

The Pandas library is covered in the book Python Data Science Handbook by VanderPlas.

Semester

Second semester

Assessment method

The assessment has a written exam and a project work.

The written exam is taken individually, on the algorithmic topics presented during the lectures. This part consists of open-ended questions. The written exam is 1 hour long and contains 4 questions. Of those questions, you have to answer to 3 of them.
The competencies evaluated are: the ability to abstract a model, knowledge of the algorithms covered in the course, and the ability to analyze computational complexity.
The evaluation of the written exam is is graded out of thirty (30) based on the correctness and completeness of the answers, and on the ability to identify the essential elements of a topic.

The project work consists of devloping a Python program by a single student. Students can choose the topic of the project among alternatives proposed by the teacher at the end of the course.

The project will be discussed after the student has passed the written exam and within a year since the date of the written exam.

The final grade is obtained by weighting 50% of the degree (on a 30-point scale) of the written exam and 50% the project work, but you have to pass both parts. There are no in-progress written exams.
Beware that you must be registered via "segreterie online" to take the exam. If you are not registered, you will not allowed to take the exam. No exceptions will be made.

There will be no midterm assessments.

Office hours

Office hours with Professor Raffaella Rizzi will be held in person. Yoi can send an email to raffaella.rizzi@unimib.it.
Office hours with Professor Della Vedova will be held online. You can book a meeting at https://www.unimib.it/gianluca-della-vedova

Sustainable Development Goals