UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Introduction To Bioinformatics

2122-3-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

- Unix shell

- Python programming in bioinformatics
- File Formats used in bioinformatics

Ability to apply 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 use Python and the Unix shell

Contents

Fundamental problems and algorithms in bioinformatics. Pattern matching.Sequence Alignment. DNA sequencing. Evolutionary histories. Managing biological data and databases.

Detailed program

1. Pattern matching: Algorithms and Data Structures. Karp-Rabin algorithm, Dömölki algorithm.
2. Suffix trees and suffix arrays: management, pattern matching and applications to the longest substring problem.

3. Sequence Alignment of two strings. Global alignment, local alignment, band alignment. Linear and generic gap cost. Multiple sequence alignment.

4. DNA Sequencing. Overlap graphs and de Bruijn graphs.

5. Evolutionary trees. Character-based models. Gusfield's algorithm for the perfect phylogeny. Distance-based models: ultrametrics and additive distance. UPGMA and Neighbor Joining algorithms. Max likelihood.
6. Genotypes and haplotypes. Single individual and pedigree cases.

7. Biological Data: file formats
8. Bionformatics open source software development methodologies
9. Linux shell
10. Python
11. Pandas
12. Biopython

Prerequisites

Algorithms and data structures; Programming Languages

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.

This course is taught in Italian.

We will use a discussion

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

First 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 project work consists of writing a Python program and it can be done by a single individual or a small group (max 3 students). The students decide the group composition. The specifications of the project are decided by the lecturers and are the same for all students in a given session.

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 of the written exam and 50% the project work, but you have to pass both parts.

There are no in-progress written exams.

Office hours

Please reserve a meeting via email.

?