UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Data Analysis

2223-1-F0901D043

Aims

Basic knowledge of the most important statistical-methodological tools of the descriptive and inferential statistics for: design of experiments, data collection and analysis, interpretation of scientific literature. Introduction to the main problems related to the computational analysis of biological sequences (DNA, RNA, proteins).

The student will be able to: understand the main concepts of study design, implement statistical analysis, read the scientific literature presenting descriptive and inferential statistic results, acquire the basic knowledge and concepts related to computational methods and techniques for collecting, managing and analyzing data in molecular biology and will master the main computational tools necessary to extract information of interest for biomedical research from the main sequencing databases.

Contents

The goal of the course is to contribute to the education of the medical biotechnologist in order to be able to:

- understand the principles of the experimental design in medicine and biology
- understand the most important statistical techniques for data analysis
- use a software for data analysis
- understand the literature presenting results from statistical analysis
- understand the motivations, problems and methodologies.
- be introduced to NGS technologies
- be able to access, query and entry data in the main databases;
- understand the main data analysis techniques: genome reconstruction and annotation; sequence comparison: global, local and multiple alignment algorithms; reconstruction of phylogenies; transcriptome analysis.

Detailed program

The module of Biostatistics is organized in two parts: descriptive statistics, inferential statistics, and interpretation of scientific literature. The first and the second part share the following characteristics:

- inclusion of methodological aspects of study design and programming of experiments
- are thought using motivating examples from the applied literature
- involves the STATA package

Part one – Basic descriptive statistics, graphical representation of quantitative and qualitative variables, indicators of position and variability, Gaussian distribution, concepts of probability.

Part two– Basics on inferential statistics, Hypothesis testing on continuous variables, T test for paired and unparired data, test on association between categorical variables, Chi square test, McNemar test, analysis of variance, sample size and power.

The module of Bioinformatics is organized in 8 chapters:

- Data management in life sciences
- Basics of informatics: Algorithms and programs, Alphabets, word, graphs, Databases
- The NGS technology: Second generation NGS platforms, Third generation NGS platforms, Genomic data formats, Genome reconstruction and annotation
- Basi di dati di sequenze molecolari: Genomic databases (EMBL – GenBank), Protein databases (SwissProt, PDB), Database query systems
- Sequence Analysis in molecular biology: Exact String matching algorithms, Sequence alignments, Motivations, Dot matrices, Substitution matrices (PAM, BLOSUM), Global alignment: Needleman-Wunsch Algorithm, Local alignment: Smith-Waterman Algorithm, Euristic Algorithms: BLAST, Fasta, BWA, Multiple alignment algorithms; CLUSTALW
- Functional motifs finding in sequences: Suffix trees, Pattern discovery algorithms
- Transcriptome Analysis: Gene Annotation and d alternative transcripts, RNA-seq data analysis
- Molecular evolution: philogenetic trees reconstruction: Clustering algorithms, k-means, Neighbor joining, UPGMA, Maximum parsimony methods, Maximum likelihood methods

Prerequisites

The student is expected to have a basic knowledge on the use of personal computer, informatics and molecular biology.

Teaching form

Standard classes, on-line quiz, video clip.

Textbook and teaching resource

https://www.pearson.it/opera/pearson/0-7623-fondamenti_di_statistica

- M. Helmer Citterich, F. Ferrè, G. Pavesi, C. Romualdi, G. Pesole, Fondamenti di bioinformatica (Zanichelli editore)
- Notes written by the teachers
- Students are recommended to subscribe to the 3 e-learning pages (the one of the course, those of the modules)

Semester

First semester.

Assessment method

Written exam (Biostatistics) and Oral exam (Bioinformatics). The grade will be calculated by averaging the grades of the two modules.

Communications relating to organizational aspects of the appeals will be given through the forum on the "Data Analysis" page.

Office hours

To be defined with the student by email contact.

Sustainable Development Goals