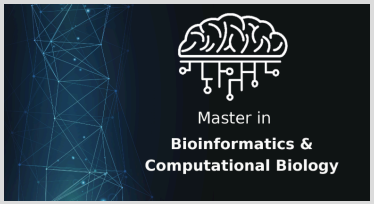
Algorithms for Bioinformatics

2018/2019

Bioinformatics: an introduction
Course organization

Pedro G. Ferreira
pgferreira@fc.up.pt

Master in Bioinformatics and Computational Biology
[dCC] FCUP



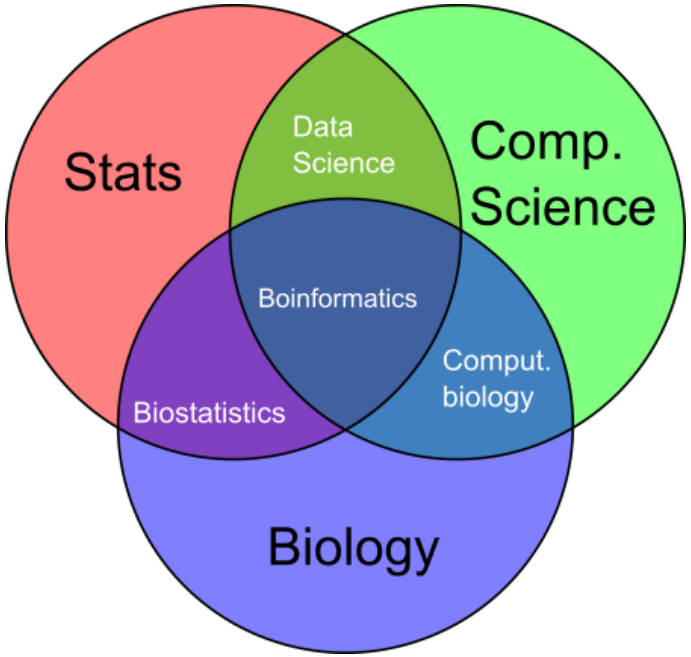
Outline

- What is Bioinformatics?
- Course Organization

What is Bioinformatics

- Requires a large set of skills: programming/scripting, cloud computing, statistics, visualization, molecular biology, …

- Interdisciplinary field with many applications:



Adapted from Genome Jigsaw

From Wikipedia :

"Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data."

From NIH:

"Bioinformatics is research, development, or application of computational approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize and analyze such data.

Goals of Bioinformatics

- Develop new algorithms and statistics to assess the relationships in large datasets;
- Analyze and interpret different types of data (e.g. nucleotide and amino acid sequences, protein structures, gene expression, ...)
- Develop and implement tools and databases to handle and access different types of data and information.
- Bioinformatics provides methods and databases for the efficient:
  - storage of data
  - annotation
  - search and retrieval
  - data integration
  - data mining and analysis

Applications

- Application areas from the *Bioinformatics* journal:

OXFORD ACADEMIC
Bioinformatics

Genome analysis
Comparative genomics, genome assembly, genome and chromosome annotation, identification of genomic features such as genes, splice sites and promoters.

Phylogenetics
Multiple sequence alignment, sequence searches and clustering; prediction of function and localisation; novel domains and motifs; prediction of protein, RNA and DNA functional sites and other sequence features.

Structural Bioinformatics
New methods and tools for structure prediction, analysis and comparison; new methods and tools for model validation and assessment; new methods and tools for docking; models of proteins of biomedical interest; protein design; structure based function prediction.

Systems Biology
Whole cell approaches to molecular biology. Any combination of experimentally collected whole cell systems, pathways or signaling cascades on RNA, proteins, genomes or metabolites that advances the understanding of molecular biology or molecular medicine.

Genetics and Population Analysis

Gene Expression
Approaches to data analysis to be considered include statistical analysis of differential gene expression; expression-based classifiers; methods to determine or describe regulatory networks; pathway analysis; integration of expression data; expression-based annotation (e.g., Gene Ontology) of genes and gene sets, and other approaches to meta-analysis.

Data and Text Mining
New methods and tools for extracting biological information from text, databases and other sources of information.

Bioimage Informatics
methods for the acquisition, analysis and modeling of images produced by modern microscopy, with an emphasis on the application of innovative computational methods to solve challenging and significant biological problems at the molecular, sub-cellular, cellular, and tissue levels.

Databases and Ontologies
Curated biological databases, data warehouses, eScience, web services, database integration, biologically-relevant ontologies.

Importance of Bioinformatics

- Important technological advances in the last decades including the HGP in the 1990's/early 2000's have revolutionized the biological and biomedical fields;
- High-throughput technologies for measuring gene expression, protein interaction or sequencing genomes are generating massive amounts of data. This opens new and interesting possibilities in biomedical research and related areas, but also many challenges.
- Life sciences are becoming more and more very computationally oriented.
- Bioinformatics plays an essential role in handling these large volumes of biological data by using computers to organize and unravel new knowledge from raw data.

Evaluation

Grading:

- Homework: 45%
  - Four homework projects (individual work; if done in group (max 2 students) should clearly indicate in the header of the document which were the elements of the group);
  - one homework is the revision of a scientific paper (individual work).
  - Both programming and written exercises

- Participation in class: 5%

- Final Exam/Test: 50%
  - Both programming and theory exercises

Plagiarism and **duplication** of work **are not allowed** at any means. Students are encouraged to discuss their work and solutions with their colleagues **but must submit original versions** of their homework. Tools such as Turnitin will be used to verify the originality of the work.

Tentative Program

Date	Topic
11/13 February	Introduction to Python & Molecular Biology concepts
18/20 February	Basic Processing of Biological Sequences I
25/27 February	Basic Processing of Biological Sequences II and Python Object Oriented
4/6 March	Finding Patterns in Sequences I
11/13 March	Finding Patterns in Sequences II
18/20 March	Pairwise Sequence Alignment I
25/27 March	Pairwise Sequence Alignment II
1/3 April	Searching Similar Sequences in Databases
8/10 April	Multiple Sequence Alignment
29 April and 13/15 May	Clustering and Trees
13/15 May	Motif Finding
27/29 May	Graphs and Biological Networks

Literature

Recommended Literature

- Bioinformatics Algorithms(1st Edition): Design and Implementation in Python. Miguel Rocha and Pedro G. Ferreira
- An Introduction to Bioinformatics Algorithms (Computational Molecular Biology) 1st Edition. Neil C. Jones and Pavel A. Pevzner

Complementary Books:

- B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Molecular Biology of the Cell, 4th edition, Garland Science, New York, USA, 2002.
- Sebastian Bassi, Python for Bioinformatics, CRC Press, 2016.
- R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins

and Nucleic Acids, Cambridge University Press, 1998.

- Dan Gusfield, Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology, 1st edition, Cambridge University Press, May 1997.

Python Resources

- The python language website, http://www.python.org/.
- Python tutor, visualization of code execution, http://pythontutor.com/.
- The python tutorial, https://docs.python.org/3/tutorial/.