Università degli Studi di Milano
Computer Science Department

# Bioinformatics

Instructors: *Matteo Re and Giorgio Valentini*

The main goal of the course consists in providing methodological tools to analyze and infer biological knowledge from complex biomolecular data through machine learning methods.
The course is intrinsically interdisciplinary and open to students in Computer Science, Physics, Mathematics, Biology, Biotechnology and other scientific disciplines.

*Introduction*.
Basic molecular biology. Computational biology problems and types of biomolecular data. Genomic and proteomic databases.

*I. Pattern matching methods and probabilistic models.*
Sequence alignment through dynamic programming algorithms.
Heuristic methods for genome-wide alignments. Markov chains, Hidden Markov Models and their applications in Computational Biology.

*II. Machine learning methods*
*A. Fundamentals*
1. Main computational biology applications of Machine Learning methods
2. An introductory example: phenotype prediction based on expression profiles. Look-up table and nearest neighbours. Probabilistic approaches and the Bayes theorem. Curse of dimensionality and the Naive-Bayes approach. From the probability density estimation to the direct assessment of the discriminant function.
3. Learning, generalization, and generalization assessment:
(a) Supervised, Unsupervised and Semi-supervised Learning
(b) Learning, over and underfitting, generalization
(c) Experimental methods for the assessment of the generalization error
4. Supervised learning
- Neural networks:
(a) Linear perceptron
(b) MLP and backpropation algorithm
(c) Multi-class ensembles of perceptrons for the biomolecular diagnosis of patients
- SVMs and their Computational Biology applications
5. Unsupervised Learning
- Clustering algorithms for the analysis of omics data: k-means and fuzzy k-means algorithms, hierarchical algorithms and self-organizing maps. Ensemble clustering methods.
- Analysis of the reliability of clusters: methods based on the structural characteristics of clusters and stability-based approaches. Applications to the discovery of pathological and clinically relevant subclasses of patients.

*B. Some relevant problems in Bioinformatics*
1. The Automated Function Prediction problem(AFP)
(a) AFP as a multiclass, multilabel and hierarchical classification problem
(b) The Princeton approach based on ensembles and Bayesian networks
(c) The True Path Rule approach based on hierarchical ensemble methods

2. Biomolecular network analysis
   (a) Modelling biomolecular networks as graphs
   (b) Node label ranking problems in computational biology: AFP, disease gene prioritization and drug repositioning.
   (c) Random walk and random walk with restart algorithms
   (d) Kernel-based algorithms and  kernelized score functions
   (e) Cost-sensitive algorithms based on Hopfield networks

Web page of the course: http://homes.di.unimi.it/valentini/CorsoBioinformatica1314.html

Web page of *AnacletoLab*, the Computational Biology Laboratory of the Computer Science Department: http://anacletolab.di.unimi.it