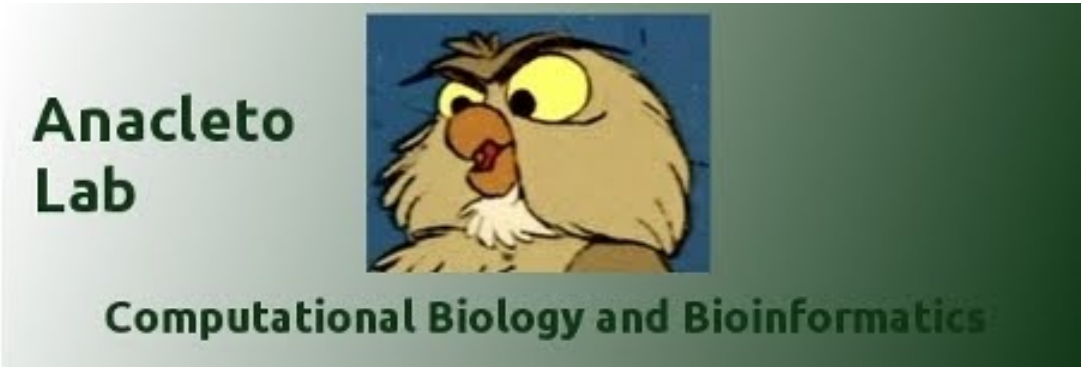
Bioinformatics theses available at AnacletoLab

Computer Science Department

UNIVERSITÀ DEGLI STUDI DI MILANO



For informal discussions about possible Bioinformatics theses, please fix an appointment by e-mail: valentini@di.unimi.it

Prediction of “deleterious” germline and somatic variants with Machine Learning methods

Thesis topics:

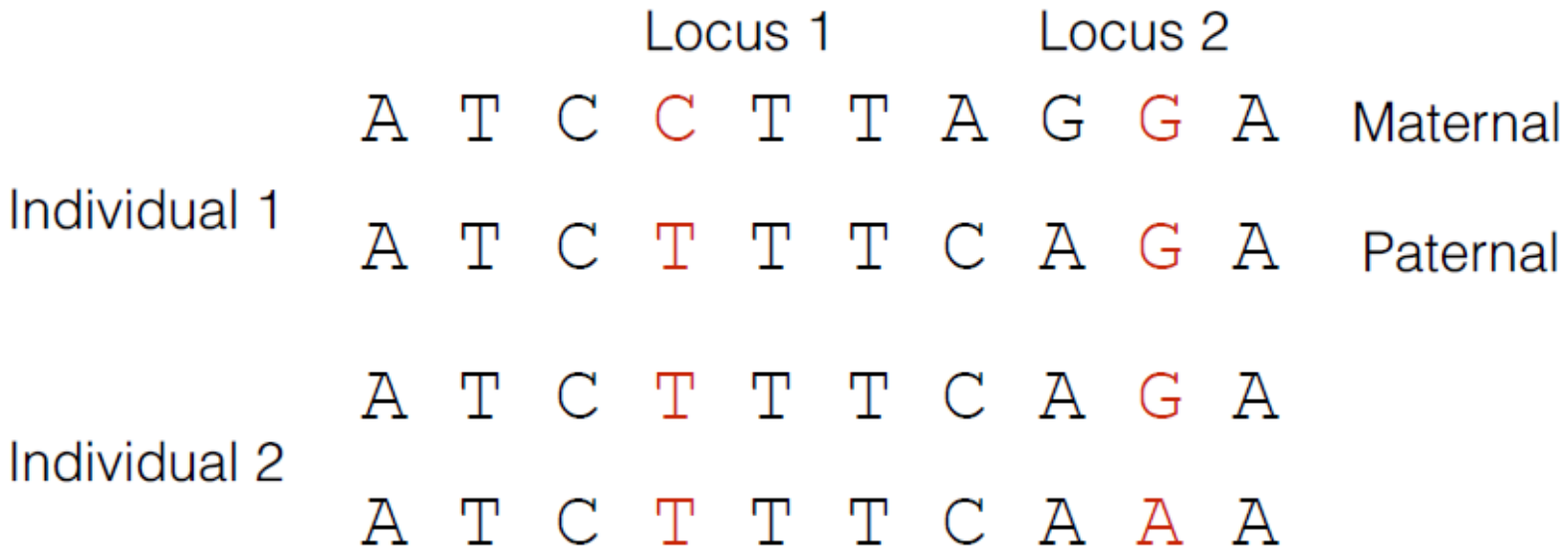
- Design and implementation of ML methods for ranking and classification of genetic variants.
- Application to real genomic data:
  - A) Genetic diseases
  - B) Cancer
- Parallel implemention of ML methods on HPC clusters (in collaboration with CINECA)

Theses in collaboration with:

- ✔ Jackson Lab, CT USA
- ✔ Charite', von Humboldt Universitat Berlin,
- ✔ CINECA Supercomputing Applications, Italy

Genetic variation at level of single nucleotides

Single Nucleotide Polymorphism (SNP)



Form the basis of most genetic analyses

Easy to study in high-throughput (million at a time)

Common (80 million SNPs discovered in 2500 individuals)

Two human chromosomes have a SNP every ~1000 bases

Genetic variation at level of single nucleotides

Single Nucleotide Polymorphism (SNP)

How to distinguish between deleterious and neutral variants?

State-of-the-art ML approaches fail on this task …

but imbalance-aware ML can work:

- M. Schubach, M. Re, P.N. Robinson and G. Valentini. *Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants* Scientific Reports, Nature Publishing, 2017
- Cost-Sensitive (CS) ML methods could be explored too: CS version of CADD (Kircher et al, Nature Genetics, 2014)

Form

E

C

T

Big biological network analysis using secondary memory based or distributed computation

"Local" implementation

+

"disk-based" computation

=

analysis of big biological graphs **on single PCs**

Theses topics:

- Development of semi-supervised graph-based ML methods using a vertex-centric paradigm
- Scalable implementation of vertex-centric algorithms using secondary-memory-based technologies or distributed computation
- Applications to the analysis of big biomolecular networks

M. Mesiti, M. Re, G. Valentini Think globally and solve locally: secondary memory-based network learning for automated multi-species function prediction, GigaScience, 3:5, 2014

Development and application of hierarchical ensemble methods (HEM) to the structured prediction of biological ontologies

Thesis topics:

- Design and implementation of novel top-down hierarchical learning strategies
- Design and implementation of novel bottom-up hierarchical learning strategies
- Application of HEM to the Gene Ontology
- Integration of HEM with network-based semi-supervised ML methods for the prediction on non coding RNA interactions

Theses in collaboration with:

- Berlin Institute of Health, Germany
- Department of Computer Science and Artificial Intelligence, University of Granada, Spain

Other theses available at AnacletoLab

- ML methods for the predictions of response of patients to drugs (in collaboration with the research group in neuro-degenerative diseases, Ospedale S.Raffaele, Milano).
- Transport protein classification through structured prediction and Multiple Kernel Learning (in collaboration with Aalto University, Helsinki).
- Methods for the visualization of big biomolecular networks
- Parallelization of network-based algorithms with GPU techniques
- Design of network ensembling methods with applications in molecular biology or medicine.