M2 Master Internship offer

Internship supervisor and Host laboratory:

Lab: Laboratoire Biométrie et Biologie Évolutive

Supervisor for the internship: Laurent GUÉGUEN

Co-supervisors: Bastien BOUSSAU, Laurent DURET

Contact e-mail: laurent.gueguen@univ-lyon1.fr

Address of the internship: UMR CNRS 5558 - LBBE

Adresse: UCB Lyon 1 - Bât. Grégor Mendel
43 bd du 11 novembre 1918
69622 VILLEURBANNE cedex

Research project title: Phylogenetic inference in recombinant sequences using Hidden Markov Models. Application to the study of gene conversion dynamics in paramecium.

Keywords : Phylogeny, Recombination, Hidden Markov Model, Paramecium

Project description:

Usual methods of phylogenetic and evolutionary inference consider that all the sites in an alignment have followed the same evolutionary history. However, there are many biological processes that contradict this homogeneity. For example, homologous recombination is ubiquitous in all living organisms and viruses. In combination with incomplete lineage sorting or gene conversion between duplicate genes (paralogs), it can have the effect that different parts of a single gene have different genealogies. This latter phenomenon of gene conversion is frequent within multi-genic families and can affect the entire genome following polyploidization events.

To correctly model the evolution of such sequences, they must be cut into segments in which the sites are considered to have the same evolutionary history. It has been shown that in the absence of considering such an heterogeneity, inference programs mostly fail. We have developed in 2009 an efficient method, "PhyML-Multi", which proposes to infer a hidden Markov model between

phylogenetic trees on alignments, to dynamically model homogeneous segments and their evolutionary history. This method has not been maintained and is no longer operational. However, we are involved in the development of the Bio++ library (https://github.com/BioPP/bpp-documentation/wiki), whose latest version makes it possible to integrate such an approach into a richer family of models (networks, complex models of sequence evolution, etc.) .

The purpose of this internship is to develop in Python a phylogenetic method that will use Bio++ programs to search for recombination points in alignments using hidden Markov Modeling. The student will test the accuracy of this method on simulated sequences, and then apply it to the characterization of conversion events in paramecia genomes. Three whole genome duplications have occurred at different times in the evolutionary history of this clade, and this study will allow measuring the dynamics of the conversion process as a function of the divergence between paralogs.

The student will have the opportunity to study probabilistic sequence analysis in a phylogenetic background, as well as to study the complex history of a large clade.

References:

Boussau B, Guéguen L, Gouy M (2009). "A Mixture Model and a Hidden Markov Model to Simultaneously Detect Recombination Breakpoints and Reconstruct Phylogenies", Evolutionary Bioinformatics, Jun 25;5:67-79.-

Guéguen L et *al* , Bio++: Efficient Extensible Libraries and Tools for Computational Molecular Evolution, *Molecular Biology and Evolution*, Volume 30, Issue 8, August 2013, Pages 1745–1750,https://doi.org/10.1093/molbev/mst097

Aury J-M et *al* (2006) Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature 444: 171–178. https://doi.org/10.1038/nature05230

Gout J-F, et *al* (2023) Dynamics of gene loss following ancient whole-genome duplication in the cryptic Paramecium complex. Molecular Biology and Evolution: msad107. https://doi.org/10.1093/molbev/msad107