Universitat Pompeu Fabra Barcelona | Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Salvador Capella Gutierrez
Email	salvador.capella@bsc.es
Institution	Barcelona Supercomputing Center (BSC)
Website	
Group	INB Coordination Node

Project

Web development & bioinformatic tools

Project Title:

Uplifting trimAl for handling thousands of sequences

Keywords:

Comparative Genomics; Multiple Sequence Alignment; Benchmarking; Machine learning

Summary:

Many evolutionary studies involving coding and non-coding genomic regions rely on the correct identification of positional homology for sequences being considered. Thus, alignments of two or more sequences are often involved in those studies. Indeed, Multiple Sequence Alignments (MSAs) are a powerful tool to establish the positional homology among sequences being compared. Generated MSAs are then used in many downstream analyses including but not limited to phylogenetic trees reconstruction. MSAs were generally used to align up to a few hundred (200 ~ 300) sequences until recently. It was observed that the MSAs quality dropped dramatically after that threshold. Various biological aspects, e.g. accelerate rates of evolution, the closeness of the subjacent species, etc; can account partially for it. Importantly, MSA algorithms belong to the non-deterministic polynomial-time family, known popularly as NP-Hard. In practice, it represents that algorithms make extensive use of heuristics trying to maximize (or minimize) a global score. Heuristic decisions might maximize a score by introducing or increasing the existing signal-to-noise ratio. Depending on the complexity of the region that is aligned, the signal-to-noise ratio can vary significantly even within the same alignment. In 2009 we presented trimAl [1], an algorithm capable of automatically identifying high signal-to-noise ratio positions within MSAs and then remove them to improve downstream analysis. trimAl was designed and tested to work with a few hundred sequences. Since its publication and extensive use across many evolutionary and non-evolutionary studies, new algorithmic solutions have been proposed to reliably align thousand of sequences [2]. This proposal aims to review the state-of-the-art methods to improve MSAs prior to any downstream analysis in the context of large-scale alignments. Building on this revision, the objective is to better characterize the behaviour of existing automated methods already available as well as explore the incorporation of new ones. The plan to incorporate new methods is to use machine learning techniques to understand what are the most relevant aspects when treating MSAs. To this end, extensive data sets can be used [3]. Finally, the resulting method will be incorporated in trimAl 2.0, which we aim to publish in the frame of this project.

References:

[1] Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. "trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses." Bioinformatics 25.15 (2009): 1972-1973. [2] Garriga, Edgar, et al. "Large multiple sequence alignments with a root-to-leaf regressive method." Nature Biotechnology 37.12 (2019):1466-1470. [3] Tan, Ge, et al. "Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference." Systematic Biology 64.5 (2015):778-791.

Expected skills::

C++ programming experience. Knowledge of ML.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed