Phyloformer: Fast, accurate and versatile phylogenetic reconstruction with deep neural networks

Luca Nesterenko[1], Luc Blassel[2], Philippe Veber[3], Laurent Jacob[4], and Bastien Boussau*[5]

[1]Département écologie évolutive [LBBE] – Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – France

[2]Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – CNRS, Université de Lyon, Université Lyon 1 – France

[3]Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 (LBBE) – Université Claude Bernard Lyon 1, Institut National de Recherche en Informatique et en Automatique, VetAgro Sup - Institut national d'enseignement supérieur et de recherche en alimentation, santé animale, sciences agronomiques et de l'environnement, Centre National de la Recherche Scientifique – France

[4]Biologie Computationnelle et Quantitative = Laboratory of Computational and Quantitative Biology – Sorbonne Universite, Centre National de la Recherche Scientifique, Institut de Biologie Paris Seine – France

[5]Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – Université Claude Bernard Lyon 1, Institut National de Recherche en Informatique et en Automatique, VetAgro Sup - Institut national d'enseignement supérieur et de recherche en alimentation, santé animale, sciences agronomiques et de l'environnement, Centre National de la Recherche Scientifique, Centre National de la Recherche Scientifique : UMR5558 – France

Résumé

Phylogenetic inference aims at reconstructing the binary tree describing the evolution of a set of sequences descending from a common ancestor.
The high computational cost of state-of-the-art Maximum likelihood and Bayesian inference methods limits their usability under realistic evolutionary models.
Harnessing recent advances in likelihood-free inference and geometric deep learning, we introduce Phyloformer, a fast and accurate method for evolutionary distance estimation and phylogenetic reconstruction.
Sampling many trees and sequences under an evolutionary model, we train the network to learn a function that enables predicting the latter from the former.
Under a commonly used model of protein sequence evolution and with GPU acceleration, it outpaces fast distance methods while matching maximum likelihood accuracy, on simulated and empirical data.
Under more complex models, some of which include dependencies between sites, it outperforms other methods.
Our results pave the way for the adoption of sophisticated realistic models for phylogenetic inference.

*Intervenant

Mots-Clés: phylogenetic reconstruction, deep learning, molecular phylogeny