Transformers for EpiDemiological DYnamics: from genomic data to epidemiological parameters

Vincent Garot*[1,2], Luca Nesterenko[3], Luc Blassel[1,4], Anna Zhukova[5], Samuel Alizon[6], and Laurent Jacob[1]

[1]Biologie Computationnelle et Quantitative = Laboratory of Computational and Quantitative Biology – CNRS, Sorbonne Universités, UPMC, CNRS – Biologie Computationnelle et Quantitative UMR 7238 CNRS-Sorbonne-Université, Site des Cordeliers Bât. A - 4ème étage, 15, Rue de l'Ecole de Médecine 75006 Paris, France, France

[2]Centre interdisciplinaire de recherche en biologie – CNRS, Institut National de la Santé et de la Recherche Médicale - INSERM, Université de recherche Paris Sciences Lettres (PSL), Collège de France – 11 place Marcellin Berthelot 75005 Paris, France

[3]Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – CNRS, Université de Lyon, Université Lyon 1 – 43 Bld du 11 Novembre 1918 69622 VILLEURBANNE CEDEX, France

[4]Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – CNRS, Université Claude Bernard - Lyon I – 43 Bld du 11 Novembre 1918 69622 VILLEURBANNE CEDEX, France

[5]Institut Pasteur [Paris] – Université de Paris – 25-28, rue du docteur Roux, 75724 Paris cedex 15, France

[6]Centre interdisciplinaire de recherche en biologie – CNRS, Institut National de la Santé et de la Recherche Médicale - INSERM, Université de recherche Paris Sciences Lettres (PSL), Collège de France, Collège de France – 11 place Marcellin Berthelot 75005 Paris, France

Résumé

The way a virus spreads leaves footprints in its genome. Phylodynamics leverages these footprints to estimate epidemiological parameters from collected virus genetic data. The estimation is typically done in a likelihood-based framework. The epidemiological process is modeled on a virus transmission tree. This tree is approximated by time-scaled phylogenetic trees reconstructed from virus sequences. However, as the epidemiological models become more realistic, their complexity increases, and the likelihood might become intractable, impeding the use of standard likelihood-based inference methods.

We introduce Teddy, a likelihood-free inference method where likelihood computations are replaced by data sampling from the epidemiological model. More precisely, we use this data to learn a function that takes observed data (dated virus sequences) and returns a posterior distribution of the epidemiological parameters given the data. Our function is parameterized by a neural network, with self-attention layers to handle permutation invariances among sequences and positional embeddings to incorporate the dates. The output contains an estimation of the epidemiological parameters and a measure of uncertainty in the form of credible intervals.

Under the common and tractable birth-death model on simulated data and early COVID

*Intervenant

data, the inference obtained by Teddy matches the one obtained by BEAST2, a state-of-the-art Bayesian inference method relying on MCMC. Unlike BEAST2, however, Teddy does not require tree reconstruction or likelihood evaluation. We also show that model mispecifications have the same effect on Teddy and BEAST2. These results are a proof of concept and suggest that Teddy may allow inference under models where likelihoods are intractable and BEAST2 could not be used.