M2 Internship Project

Assessment of Bayesian phylogeographic inference techniques with an application to the analysis of the *Rice yellow mottle virus* in Africa.

KIM Data & Life Sciences

LIRMM - IRD - IMAG - Montpellier

Project Description

The analysis of georeferenced genetic sequences using methods from the field of phylogeography makes it possible to characterize the spatial dynamics of an evolving species or population. During the course of an epidemic, these methods can be applied to reconstruct both the mutation history of the sequences, represented by a phylogenetic tree, and the geographical spread of the pathogen (Lemey et al., 2010). These methods have for instance proven very useful to describe the spread of viruses that infect animal or plant species (Dellicour et al., 2017). Phylogeography, as an inference tool, is thus of interest both in terms of public health and agronomy. More generally, it advances our knowledge of evolutionary and ecological processes and their interactions (Baele et al., 2017).

Phylogeographic analysis relies on probabilistic models describing both the evolution of sequences along a phylogenetic tree, i.e., the accumulation of mutations that have given rise to the differences observed between genes and genomes of contemporary organisms, as well as the fluctuation of the spatial coordinates of the different lineages during evolution (Lemey et al., 2010). Fitting these complex models to the data is a considerable methodological challenge that involves sophisticated software and significant calculation times (Baele et al., 2018). It is therefore essential to develop robust techniques to test the accuracy of these methods. Using simulated data for which we can control both the true phylogeographic scenarios and the level of difficulty, our objective in this project is to verify that the available tools are able to correctly recover the general dynamic of the process, as well as the values of the parameters of the phylogeographic models.

The simulations carried out in this context will be all the more relevant that the generated data is close to real data, for which we would like to have guarantees on the good behavior of the inference tools used. The internship subject that we are proposing here will focus on the *Rice yellow mottle virus* (RYMV) in Africa. Here we have heterochronous sequences of isolates sampled from 1966 to 2018 for which the rate of evolution of the virus was estimated (Fargette et al., 2008). The integration of the spatial and temporal components of diversity was carried out according to the continuous model developed by Lemey et al. (2010). It results in the reconstruction of the phylogeography of RYMV in the different regions of Africa (see Issaka et al., 2021, and references therein). Our simulations will thus be carried out from the *a posteriori predictive distribution*, using parameter values inferred from observed RYMV data to simulate new synthetic data conditionally on these values.

Particular attention will be given to the inference of dispersal statistics, characterizing the speed at which organisms colonize their habitat (Pybus et al., 2012). Defining summary statistics for phylodynamic models is a challenging task, that requires both relevant modeling assumptions and rigorous mathematical developments, as different definitions of the same statistic can sometimes give very different results. Preliminary analyzes that we conducted recently show that currently used dispersal statistics could be defined in an ambiguous way, sometimes in contradiction with the assumptions of the underlying model. Thus, depending on the affinities of the successful candidate, it will be possible to focus on questions related to the mathematical modeling and the inference of spatial diffusion phenomena, rather than on simulations. The subject is designed in such a way that the results obtained from the simulations will feed the developments on dispersal statistics, and vice versa.

Working Environment

The internship will be directed by **Stéphane Guindon**, **Denis Fargette** and **Paul Bastide**.

Stéphane Guindon is a CNRS reseacher at the *Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier* (LIRMM, UMR UM-CNRS 5506)). His researches focus on the deveolpement and test of algorithms and probabilistic models to understand the forces that drive evolution and ecology. [www.lirmm.fr/users/utilisateurs-lirmm/stephane-guindon]

Denis Fargette is an IRD Senior Researcher at the Plant Heath Institute of Montpellier (UMR UM-IRD-INRA-CIRAD-SupAgro). He is a specialist on the evolutionary and epidemiologic dynamic of African crop viruses. [umr-ipme.ird.fr/annuaire/denis-fargette]

Paul Bastide is a CNRS researcher at the *Institut Montpelliérain Alexander Grothendieck* (IMAG, UMR UM-CNRS 5149). He is working on the mathematical and statistical modeling of evolution. [pbastide.github.io]

The main internship hosting lab will be the LIRMM, in the team *Méthodes et Algorithmes pour la Bioinformatique*. The intern will also have the opportunity to work at the IRD, in the *Virus Cereals In Tropical Agro-ecosystems*, and at the IMAG, in the *Probability and Statistics* team.

The internship is financed by the *Data and Life Sciences* Key Initiative, and will take place in a highly interdisciplinary environement, offering many opportunities to collaborate.

Candidacy

Applications should include a short CV and cover letter and sent to:
stephane.guindon@lirmm.fr
denis.fargette@ird.fr
paul.bastide@umontpellier.fr

References

Baele, Dellicour, Suchard, et al. 2018. *Current Opinion in Virology*. 31:24–32.
Baele, Suchard, Rambaut, et al. 2017. *Systematic Biology*. 66:47–65.
Dellicour, Rose, Faria, et al. 2017. *Molecular Biology and Evolution*. 34:2563–2571.
Fargette, Pinel-Galzi, Sérémé, et al. 2008. *PLOS Pathogens*. 4:e1000125.
Issaka, Traoré, Longué, et al. 2021. *Virus Evolution*. .
Lemey, Rambaut, Welch, et al. 2010. *Molecular Biology and Evolution*. 27:1877–1885.
Pybus, Suchard, Lemey, et al. 2012. *Proceedings of the National Academy of Sciences*. 109:15066–15071.