ABS24

The Applied Bayesian Statistics summer school

BAYESIAN PHYLOGENETICS AND INFECTIOUS DISEASES

Lecturer: Marc Suchard
Assistant lecturer: Filippo Monti

Villa del Grumello, Como, Italy, 26-30 August 2024

Abstracts

Participants' presentations

CONTENTS

Flexible phylodynamic modelling using general branching processes

Frederik Mølkjær Andersen[(1)], *Samir Bhatt*[(1)], *Carsten Wiuf* [(2)].

(1) Department of Epidemiology, University of Copenhagen, Denmark,
(2) Department of Mathematical Sciences, University of Copenhagen, Denmark.

fman@sund.ku.dk ; samir.bhatt@sund.ku.dk ; wiuf@math.ku.dk

Abstract: Stochastic models for phylogenetic trees are still not well understood. Whilst an array of stochastic processes (e.g. birth-death processes and the coalescent) are frequently used, a large class of them fit empirical phylogenetic trees poorly. Age-dependent speciation/extinction models have been shown to explain data better but have long been considered intractable. Here, we show that, at least mathematically, this is not so. Inspired by its application in phylogenetics, we define and study a reduced time-varying Bellman-Harris process and the related reduced branching tree, representing respectively the counting process of the number of lineages through time and the reconstructed phylogenetic tree with observed speciation times. A full distributional characterization of this process and its reduced tree are given through a set of integral equations, whose solutions are simple to approximate numerically. The joint distribution of the tree shape and the observed speciation times characterized, allow for estimation of evolutionary dynamics given a reconstructed phylogenetic tree.

Keywords: Phylodynamics; General branching processes; Tree prior.

Bayesian Birth-Death Skyline Model - A Case Study on Heterochronous Maltese SARS-CoV-2 Genomic Data

Gianluca Ursino[(1)], Monique Borg Inguanez[(1)], David Suda[(1)], Joseph Borg[(2)], Graziella Zahra[(3)].

(1) Department of Statistics and Operations Research University of Malt, Malta,
(2) Department of Applied Biomedical Science, Faculty of Health Sciences, University of Malta, Malta,
(3) Molecular Diagnostics – Infectious Diseases Mater Dei Hospital, Malta.

monique.inguanez@um.edu.mt; david.suda@um.edu.mt; joseph.j.borg@um.edu.mt; graziella.zahra@gov.mt

Abstract: When studying viral genome sequence data the Bayesian framework has the advantage that it can simultaneously construct phylogenetic trees, which allow us to analyse the relationships between different genomes (phylogenetic analysis), and at the same time infer viral dynamic across time (phylodynamic analysis). This requires the specification of three models: (i) the transmission model (ii) the substitution model and (iii) the molecular clock model used to infer the root of a phylogenetic tree. In this study as transmission model we consider the Bayesian birth-death skyline (BDSKY) model which can be applied for both homochronous and heterochronous genome datasets under specific formulations of it. When it comes to defining a substitution model we resort to the bModelTest method. As a case study we consider 681 heterochronous genome sequences of COVID-19 sampled by the Molecular Diagnostics-Infectious Diseases at the Mater Dei Hospital in Malta between 19/8/2020 and 5/1/2022. For this purpose, we consider both serial BDSKY and the multi-rho BDSKY models, with the former being more suitable for genome sequences individually sampled at several time points, and the latter being more suitable for genome sequences sampled in batches at multiple time points. For each transmission model we consider four different molecular clock models which are a combination of the strict and relaxed molecular clock setups, and two settings for the number of intervals over which the reproductive number is considered constant (m=15 and m=30). The relaxed molecular clock option with m=15 intervals resulted to be the preferred molecular clock model of the data being studied. In general the serial BDSKY and the multi-rho BDSKY transmission models gave considerably similar results yet some discrepancies were observed and these will be discussed.

A full-length paper related to this study has been accepted for publication in the Springer book Quantitative Methods and Data Analysis in Applied Demography - Volume 1.

Keywords: Bayesian Birth-Death Skyline Models, Phylogenetic tree, Genomics, SARS-CoV-2

Time depth and the limits of phylogenetic inference in linguistics

Emma Kopp[(1)], *Robin J Ryder*[(1)], *Thomas Pellard*[(2)], *Guillaume Jacques*[(2)].

(1) Paris-Dauphine University, France,
(2) INALCO, France.

kopp@ceremade.dauphine.fr; ryder@ceremade.dauphine.fr; thomas.pellard@cnrs.fr; rgyalrongskad@gmail.com

Abstract: Computational methods have been used to reconstruct the history of languages over several millennia, based on data from modern languages. Using stochastic models of evolution along a phylogenetic tree, these methods infer language relationships (the topology of the tree) along with the ages of ancestral languages, usually in the Bayesian setting. Language phylogenies in the literature rarely reconstruct ages beyond 8 to 10 thousand years; additionally, all the more ancient proposed language groupings are subject to debate within the scientific community.
We investigate the threshold beyond which phylolinguistics trees reconstruction is unreliable. We apply theoretical results from the mathematics of phylogenies literature, which give upper bounds on the probability of correct reconstruction of the tree topology and the values at the root. In particular, we show that for languages evolving at the rates typically reported in the literature, it is impossible to reconstruct the topology of a tree whose root age is older than 12,000 years. For trees older than this threshold, the inferred topology will not be more reliable than a random guess.
To arrive at this result, we reproduce three previous analyses on cognatized lexical data from 50 Sino-Tibetan languages (Sagart et al. [2019]), 422 Bantu languages (Grollemund et al. [2015]) and from 161 Indo-European (Heggarty et al. [2023]). We use Markov Chain Monte Carlo to produce samples from the posterior distribution of model parameters. We then apply results from percolation theory and information theory to bound the probability of correct reconstruction. In both cases, we find that the bound decreases rapidly from 1 to 0, with a threshold between 9 and 12 thousand years. To our knowledge, this is the first theoretical quantitative bound on phylolinguistics methods. It demonstrates that reconstructing the deep topology of more ancient language families based on cognatized lexical data is a hopeless enterprise.

Keywords: phylogeny; linguistics.

Bayesian inference of mixed Gaussian phylogenetic models

Bayu Brahmantio[(1)], *Krzysztof Bartoszek*[(1)], *Etka Yapar*[(2)].

(1) Department of Computer and Information Science, Linköping University, Sweden,
(2) Department of Biology, Lund University, Sweden.

bayu.brahmantio@liu.se; krzysztof.bartoszek@liu.se; etka.yapar@biol.lu.se

Abstract: Continuous trait evolution is commonly modelled using stochastic differential equations (SDEs) to represent deterministic change of trait through time, while incorporating noises that represent different unobservable evolutionary pressures. Two of the most popular classes of SDEs are Brownian motion (BM) and Ornstein-Uhlenbeck (OU) process, which fall under a larger family of models called GLInv that has a Gaussian transition probability with expectation that is linear with respect to ancestral value and variance that is invariant with respect to it. This framework enables multiple different GLInv models under a single phylogenetic tree to capture diversity of traits from different species on the tips.
In this work, we consider a Bayesian method to infer the posterior distribution of parameters for heterogeneous branching Gaussian processes that involve GLInv family of models. We implement a particle-based method for posterior simulation and an evaluation metric based on the posterior predictive accuracy. Our work is inspired by evolutionary questions, and we illustrate it by a biologically inspired simulation study.

Keywords: Phylogenetic comparative methods; Bayesian inference; Monte Carlo.

Unravelling spatiotemporal heterogeneities of wild and vaccine-derived poliovirus spread: past and present

Darlan da Silva Candido[(1)*], *Simon Dellicour*[(2),(3),(4)], *Laura V Cooper*[(1)], *Carlos A Prete Jr*[(1),(5)], *David Jorgensen*[(1)], *Christopher B Uzzell*[(1)], *Arend Voorman*[(6)], *Hil Lyons*[(7)], *Dimitra Klapsa*[(8)], *Manasi Majumdar*[(8)], *Kafayat Arowolo*[(8)], *Corey M Peak*[(6)], *Ananda S Bandyopadhyay*[(6)], *Javier Martin*[(8)], *Nicholas C Grassly*[(1)], *Isobel M Blake*[(1)]

(1) MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London; London, UK. (2) Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles; Bruxelles, Belgium. (3) Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory for Clinical and Epidemiological Virology, Katholieke Universiteit Leuven; Leuven, Belgium. (4) Interuniversity Institute of Bioinformatics in Brussels, Université Libre de Bruxelles, Vrije Universiteit Brussel; Brussels, Belgium. (5) Departamento de Engenharia de Sistemas Eletrônicos, Escola Politécnica, Universidade de São Paulo ; São Paulo, Brazil. (6) Bill and Melinda Gates Foundation; Seattle, USA. (7) Institute for Disease Modeling, Global Health Division, Bill and Melinda Gates Foundation; Seattle, USA. (8) Division of Vaccines, National Institute for Biological Standards and Control, Medicines and Healthcare products Regulatory Agency; Potters Bar, UK.

ddasilva@ic.ac.uk; simon.dellicour@ulb.be; l.cooper@imperial.ac.uk; carlos.prete@usp.br; david.jorgensen13@imperial.ac.uk; c.uzzell@imperial.ac.uk; Arend.Voorman@gatesfoundation.org; Hil.Lyons@gatesfoundation.org; dimitra.klapsa@mhra.gov.uk; manasi.majumdar@mhra.gov.uk; kafayat.arowolo.@mhra.gov.uk; corey.peak@gatesfoundation.org; ananda.bandyopadhyay@gatesfoundation.org; javier.martin@mhra.gov.uk; n.grassly@imperial.ac.uk; isobel.blake@imperial.ac.uk

Abstract: Outbreaks of vaccine-derived poliovirus type 2 (cVDPV2) have become a major threat to polio eradication. We use cVDPV2 cases and Wild-type poliovirus 1 (WPV1) sequences to uncover the spatiotemporal patterns and drivers of poliovirus spread. Between May 2016 and September 2023, 3120 cVDPV2 cases were reported across 76 outbreaks and 39 countries globally. Outbreaks have mostly been small (median = 5 cases, range 1-578), have spread to a median maximum distance of 231 km (0-4442) and for median duration of 202 days (0-1905). Wavefront velocity analysis of large outbreaks reveals a median velocity of spread of 2.3 km/day (1.0-4.4). International borders are associated with a slower velocity of spread ($p < 0.001$), when in the presence of high immunity. Finally, phylogeographic analysis of 1572 global sequences, including 38 newly generated, reveals that historic WPV1 spread resembles recent cVDPV2 patterns and that international spread is largely sustained by unidirectional movement between neighbouring countries. Our findings offer insights for enhancing the geographical scope of poliovirus surveillance and response, crucial steps in the final phases of poliovirus eradication.

Keywords: Poliovirus; Phylogeography; Spatiotemporal analysis.

Quantifying the cross-species transmission events of SARS-CoV-2 between farmed mink and humans in Denmark between June and November 2020

Amanda Gammelby Qvesel [1,2,3], *Marlies Jilles Francine Goedknegt* [1,2], *Esben Rahbek Thuesen* [1,2,3], *Thomas Bruun Rasmussen* [3], *Anders Gorm Pedersen* [1,2].

(1) Department of Health Technology, Section for Bioinformatics, Technical University of Denmark, Kgs. Lyngby, Denmark,
(2) PandemiX Center, Department of Science and Environment, Roskilde University, Roskilde, Denmark,
(3) Department of Virus and Microbiological Special Diagnostics, Statens Serum Institut, Copenhagen, Denmark.

amaqve@dtu.dk; marliesgoedknegt@hotmail.com; esth@ssi.dk; tbru@ssi.dk; agpe@dtu.dk

Abstract: During 2020, SARS-CoV-2 was infecting not only humans but also farmed minks in many countries, including Denmark, a country that then comprised 40 % of the global mink production with more than 1100 farms. Of these, infected mink were identified on 290 farms, before all mink were culled by the end of the year to halt the spread after the rise of so-called mink mutations. Among these are the Y453F mutation in the spike protein that has been found to increase the binding of the protein to the ACE2 receptor in mink without changing the binding to human ACE2, and the H69/V70del deletion.
In this work, we use viral sequences from mink and humans to estimate the number of cross-species transmission events between mink and humans in 2020. We do so by four different methods; A) Counting direct transmission events identified by Transphylo between two observed sequences, B) As A but also including indirect transmission events between observed sequences, C) Counting jumps on the Transphylo transmission tree with unobserved node states (i.e. host species state) inferred by parsimony, D) using a structured coalescent as implemented in BEAST2 via MASCOT.
We find that the mean number of transmission events from mink to human ranges from 8.6-59.0, and that the mean number of transmission events from human to mink ranges from 13.2-265.1, depending on the method. We discuss the drawbacks of working with highly similar sequences and how the different methods employed have different strengths with respect to capturing the underlying structure of the data but also come with different shortcomings.

The work presented is published as part of the article "*Emergence and spread of SARS-CoV-2 variants from farmed mink to humans and back during the epidemic in Denmark, June-November 2020*" by Rasmussen et al., PLOS Pathogens, 2024.

Keywords: SARS-CoV-2; cross-species transmission; phylogenetic tree, transmission tree.