Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

A comprehensive study of the phylodynamics of SARS-CoV-2 in Europe

Master Thesis

Cecilia Valenzuela Agüí
ceciliav@student.ethz.ch

Department of Biosystems Science and Engineering
Computational Evolution
ETH Zürich

Supervisors:
Prof. Dr. Tanja Stadler, Dr. Timothy Vaughan, Sarah Nadeau

February 22, 2021

Acknowledgements

I thank Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Abstract

The abstract should be short, stating what you did and what the most important result is. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Contents

Acknowledgements i

Abstract ii

1 Introduction 1

2 Material and Methods 2

2.1 The multytipe birth death model 2

2.2 Infering epidemic trajectories 3

2.3 Modeling the early dynamics of SARS-CoV-2 in Europe 3

2.4 Incorporating travel data 4

2.5 SARS-CoV 2 genome data set 4

2.6 GLM predictos data 4

2.7 Bayesian inference of phylodynamic parameters 4

2.8 Processing of epidemic trajectories 5

2.9 Data availability 5

2.10 Code availability and analyses reproducibility 6

3 Discussion 7

Bibliography 8

A Data A-1

B Priors B-1

C Supplementary Figures C-1

Chapter 1

Introduction

We have many sources of information, but all of them are imperfect in some aspect. Bayesian phylodynamics incorporates different beliefs and data to get a detailed picture of the epidemic. Genetic sequences as an objective source of information about virus evolution and transmision together with viral phylogenetics, transpotation data and geographic data to understand the dynamics and case counts to guide the magnitutde of the numbers.

We lack absolute numbers of the pandemic time resoluted without reporting issues dependent. We lack understading of the dynamics, from where to where, how often, how long.

We can use travel data as a proxy, or we can incorporate it in a model. Lemey et al. Difference we get absolute numbers, not only the rates. And we use BD model instead of coalescent, better to describe beginning of a epidemic. While due to the complexite and time consuming model we dow scaled to europe and only thhe initial phase with the first countries to experiment relevant sars cov 2 epidemics.

Chapter 2

Material and Methods

2.1 The multytipe birth death model

Introduction?

In a phylodynamic analysis, we aim to integrate genetic and epidemiological information to understand pathogen evolution and transmission dynamics. This is possible for RNA viruses, and in specific for SARS-CoV-2, because the population genetic process and the spread of the epidemic are in the same time scale [?]. Thus, from virus sequences we can gain information about the interactions between hosts and the epidemic development [?]. More precisely, the field of phylodynamics studies how phylogenetic trees are being generated and infers the population parameters behind that process.

To study the early dynamics of SARS-CoV-2 in the European countries we have used the multitype birth-death model described in [?]. Birth death models are classical compartmental population models with high flexibility. In the multitype version, we can describe a structured population in types or subpopulations. The process starts with one individual in one of the subpopulations, e.g in subpopulation i, who can give rise to another individual at rate β_i, die at death δ_i, migrate to another deme j at rate $m_i j$ or be sampled at rate ϕ_i to become part of the phylogenectic tree.

The parameters that we will consider are the R0, the becoming uninfectious rate, the migration rates and the sampling proportion (proportion of no longer infectious individual that arer sampled and included in the phylogenetic tree).

Thus, the probability that we want to know is:

$P(T, P|D)$

This process describes the generation of the tree as part of the dynamics of the whole population, it is very rich and detailed in information. Therefore it is also possible to obtain information about the epidemic trajectory of the whole population. We can obtain the full sequence of events that describes the epidemic in each of the demes.

From the phylogenetic tree and the phylodynamics parameters we use Stochas-

tic mapping to infer the ancestral traits and then to infer the epidemic trajectories.

$P(Tc|T, P)$

$P(E|Tc, P)$

2.2 Infering epidemic trajectories

Stochastic mapping of ancestral traits

Read paper and tim presentation

Stochastic mapping of epidemic trajectories

From the phylogenie with ancestral locations and the set of phylodynamic parameters we can simulate epidemic trajectories using an algorithm similar to Gillespie (with tau leaping and particle filtering? the subsampling). And most improtant, we can compute the probability of that specific trajectory given those parrameters and tree.

$P(E|Tc, P)$

Any trajectories without the events that are respresent in the tree has probability 0, so to avoid the simualtion of this trajectories, no time effcieint, we will enforce the events that arer represented in the tree (and will include a weigth). Also to avoid low probabilities trajectories, we will importance subsample one trajectory after a certain time of simulation according to the weigths detemrined by its probabilities.

2.3 Modeling the early dynamics of SARS-CoV-2 in Europe

In our case, we have 6 demes, one for each of the countries/geographic location. A birth represents a transmission event of covid-19, a death event is an individual becoming uninfectious by overcoming the disease, being isolated or dying. From the birth rate and death rate we can estuimate the R0, key vallue in the epidemic understanding and control. A migration event happens when an individual travels from one country to another, we know that is an important way of covid-19 spreading and we are interested on the dynamics.

We focus our analysis in the early spread in Europe of SARS-CoV-2, therefore we constrain the analysis to the period from the origin of the epidemic to March 8, when the first European lockdown in the Lombardy region started. During this time, for the European countries we expect an unimpeded spread of the virus that could be described by a constant R0 particular to each deme. However,

the migration rates could have changed during these months due to the increase public awareness, so we will consider a different migration rate for origin to 23 jan, 23 jan to end of february and 1 to 8 of March. We assume a constant sampling proportion different for each of the demes, to account for difference in the sequencing efforts. The becoming uninfectious rate we will fixed in 36.5, assuming that one individual is infectious for a exponential time distirbuted with mean 10.

2.4 Incorporating travel data

To incorporatet travel data in the model we have use a GLM model described in Lemey et al 2009. We use a generalized linear model to describe the migration matrix KxK, where the rpedictors are the number of dayly flight from one country to another from EUROSTATS, the average distance between the centrroids of the country and the population sizes from origin and destinations. The predictors are log transformed, we add a pseudocount to make them all positive, and standardized as described in ..

$mij = cexp(\sum_{i=1}^{p} \delta_i \beta_i xij)$

The coefficients beta describe the effect size of the predictors in the migrations rate and the delta coefficients act as a model selection variables, taking 0-1 values including orr not that specific predictor in the model.

2.5 SARS-CoV 2 genome data set

GISAID. Sequences from France, Germany, Italy, Spain and other European countries. Subsampling a sequence with probability equal to the probability of having a case that day in the country, inverseley weigthed by the probability of being sequenced that day.

2.6 GLM predictos data

EUROSTAT, countries considers. Distance and population datasets.

2.7 Bayesian inference of phylodynamic parameters

We use bayesian inference and Metropolis Hasting Monte Carlo algorith to infer our phylodynamics parameters and phylogeneit tree. We use the following priors

Clock rate and substitution model. Fixed to 8x10-4, HKY 4 gamma categories with priors... BDMM-Prime epi parameterization: R0 Become uninfectious rate Sampling rates Migration rates

R0

R0 constant except for China (and Italy? maybe not in the main analysis).

Prior Log Normal (0.8, 0.5) Median 2.2, 95IQR 0.8 to 5.9. (Lai A. 2020 prior log normal (0,1) median 1.0 0.1-7.10)

- Main analysis: We fix China R0 to 2.7 (23 jan) 1.3 (10 feb) 0.8 - Use a decreasing paramater as Sarah with migration rate? But then we would need to add sequences. - Or instead of fixing it we can put a strong informative prior. - Allow Italy R0 to change March 1.

Meta-analyses R0:

R0 in China (3.14,95%CI,2.40–4.09).[?]

R0value 2.90 (2.32, 3.63 2.9 (95% CI: 2.1–4.5)[?]

Periods: until 23 jan R0 2.9 23 jan - 10 feb 1.1 dashboard cevo 10 feb - 8 mar 0.43 dashboard cevo

Become uninfectious rate same for all demes, fixed to 36.5. Migrations among demes, 3 epochs. Sampling proportion 0 before first sample. Upperbounded by number of cases.

Implementation BEAST 2.6.3 and BDMM-Prime package. We run 5 chains of 10e7 iterations witth different seeds.

2.8 Processing of epidemic trajectories

In our case, for each set of parameters and tree we will simulate 10000 trajectories with an epsilon of 0.3 and select one of them with importance sampling. We have one trtajectory for each step in the MCMC. We subsampled x.

2.9 Data availability

GISAID data table.

2.10 Code availability and analyses reproducibility

All the code is availables at ... We have implement the full workflow with Snakemake so its fully reproducible.

Chapter 3

Discussion

Bibliography

Appendix A

Data

Appendix B

Priors

Appendix C

Supplementary Figures