Published: June 2018

Regression ABC phylodynamics to infer epidemiological parameters

Samuel Alizon - evolutionary biologist and Research Director at the CNRS based in Montpellier, France.

From phylogenies to phylodynamics

Using DNA to reconstruct the history of species (i.e. phylogeny) is now commonplace in evolutionary biology. This is done by using two principles: 1. parsimony, i.e. more similar sequences cluster closer together in the phylogenetic tree; and 2. molecular clock, i.e. the number of differences between two sequences indicates the time since divergence. In the end, we can display the relatedness between species in so called phylogenetic trees (diagrams).

With the advent of mass sequencing, these phylogenetic methods have been applied to infections instead of individuals. By obtaining the genetic sequence of a virus in a set of hosts, it is possible to infer a phylogenetic tree that is somehow related to the transmission chain. The underlying assumption here is that the virus evolves rapidly enough for mutations to take place on the time scale of the epidemics such that the virus transmitted from one person differs from the virus that caused the infection (1). The evolutionary rate of the virus determines the time scale on which inferences can be made: for human immunodeficiency virus (HIV) or hepatitis C virus (HCV), the resolution can be of days, whereas for more slowly evolving viruses such as human papillomaviruses (HPVs) we need hundreds of years to get enough resolution (2).

In 2004, Grenfell et al. (3) coined the term "phylodynamics" to capture the idea that the way viruses spread leaves footprints in their phylogenies and that by analysing these via phylogenetic methods, we can make inferences regarding key epidemiological parameters (like the basic reproduction number, R_0). For several years, phylodynamics used generic population dynamics models, but the last decade has seen the development of new methods to use classical epidemiological models, such as the SIR model, with Susceptible, Infected and Removed hosts (4,5).

Successes and limits of phylodynamics

The Ebola epidemic in West Africa marked a qualitative change in terms of the amount of viral infection sequences, but also showed how rapid and shared this sequencing was (6). The development of *ad hoc* models allowed estimation of R_0, which is the number of secondary infections caused by an infected host in a completely susceptible population. Classically, the greater the R_0, the more difficult the epidemic is to control. Another strength of phylodynamics is that it helped to understand the geographical spread of the epidemics (7).

Several limits of phylodynamics also appeared. Firstly, implementing a detailed epidemiological model is difficult for methods that require the true likelihood of observing the phylogeny for a given set of parameters. Furthermore, computing time typically increases faster than linear as the number of sequences goes up, hence larger phylogenies are more difficult to handle with current approaches. Finally, existing phylogenetic methods tend to only use dated sequences and cannot include other types of data such as time series of incident cases, which were particularly detailed during the recent Ebola epidemics.

ABC phylodynamics

In a recent study (8), we introduced a new approach that relies on Approximate Bayesian Computation (ABC). Contrarily to existing methods, it does not require the derivation of the exact likelihood of observing the phylogeny. In a nutshell, the idea of ABC is to use numerical simulations to generate thousands of phylogenies of infection using a given epidemiological model. Then, we identify the simulated phylogenies that are the most similar to the target phylogeny obtained from the empirical data. Since we know how we simulated the phylogenies, we can estimate the model parameters that are most likely to generate the target phylogeny.

An important difficulty in this approach resides in the comparison between two phylogenies. This is almost like finding a metric that captures the difference between two actual trees: the only way to do this is to summarise the tree into quantitative variables (height, diameter, number of branches and number of leaves) that can be compared between trees. We do the same by breaking our phylogenies into dozens of summary statistics such as branch lengths, ratio of internal to external branches, general balance of the phylogeny, timing of events etc. We then use all of our 83 summary statistics to calculate a Euclidean distance (length of a true straight line) between each of the simulated phylogenies and the target one.

Being able to tell how different phylogenies are allows us to perform the first step of the ABC, which is called the "rejection step". This consists of keeping only the phylogenies that are sufficiently close to the target one. Classical ABC algorithms then re-iterate this rejection by re-simulating data using a Markov Chain Monte-Carlo (MCMC) algorithm. However, here we used a more recent method called regression ABC. It consists of using the phylogenies kept after the rejection step to inform a regression model. If successful, we can then predict model parameters for any given set of summary

statistics. As we will see later, regression ABC has several advantages compared to the canonical ABC approach. Since this method is quite new, we compared two algorithms to perform the regression step, both of which rely on machine learning: Least Absolute Shrinkage Selector (LASSO) regression and neural networks – an information processing paradigm that is inspired by the way the brain processes information.

ABC regression is powerful

In general, our cross-validation analysis shows that regression ABC is powerful. By definition, we know ABC cannot outperform a method that is based on the exact likelihood. However, for a simple SIR model, we reach an inference precision for R_0 and infection duration close to the "gold-standard" obtained by maximising the true likelihood function (via the BDSIR model in the software BEAST). Regression ABC even outperforms BDSIR when it comes to estimating the host population size, which is due to the fact that the algorithm implemented in BEAST has to make numerical approximations to compute the likelihood function (Figure 1).

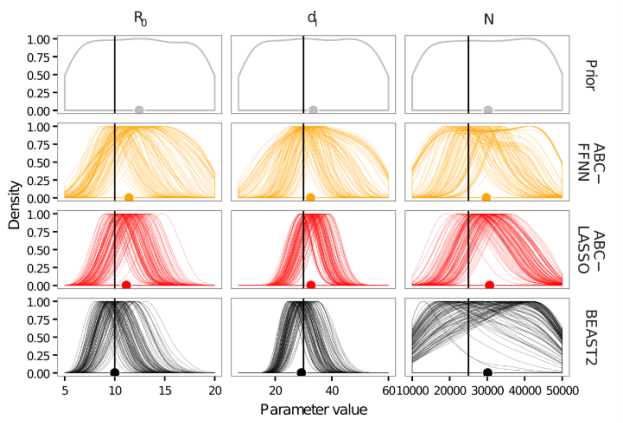

Figure 1: Inference power of the regression ABC for three parameters of the SIR model. The first row shows the prior that was explored (in gray), the second row is the regression-ABC using neural networks (in yellow), the third row is the regression-ABC using LASSO regression (in red) and the bottom row is the results using the maximum likelihood approach implemented in BEAST (in black). The inference power is close to that of BEAST for R_0 and infection duration and is somehow better for the population size (right column). Vertical lines show the target value. This estimation was performed for 100 target phylogenies and each line shows the inferred distribution for the respective target. The large dot shows the median value for all 100 targets. This figure is reproduced from (8).

Regarding the more technical details, we first showed that it is not possible to capture the information contained by the phylogeny with few summary statistics as this information seemed to be spread out in our summary statistics. However, one of the differences between LASSO and neural networks was that the former was more efficient at selecting between summary statistics, making it more reliable as a generic regression tool.

To further investigate the power of our model, we focused on the 2014-2016 Ebola epidemics and used the first outbreak phylogeny published by Gire et al. (6) to assess our ability to infer parameters with a more elaborate epidemiological model. To this end, we simulated phylogenies assuming an SEIR model (where E stands for an Exposed state, that is infected but not yet infectious) and then ran our regression ABC to infer R_0, duration of latency, and duration of infectiousness. Compared to the approach encoded in BEAST (9), we have a lower ability to infer R_0, but a greater ability to infer other parameters such as the duration of latency. This is interesting because duration parameters are typically difficult to infer from incidence data and require contact tracing data. However, the small size of the phylogeny (72 leaves) and its low resolution seem to be close to the limit of our method, which is much more powerful for larger phylogenies.

Perspectives on regression ABC

We showed that a "naive" approach that only requires the ability to simulate phylogenies of infections can reach a precision close to that obtained by state-of-the art methods that rely on the true likelihood of observing a phylogeny given some parameter values. We also showed that the accuracy of this method, in terms of minimising the error made when inferring parameters, increases with the size of the phylogeny. From a more technical point of view, regression techniques now allow us to efficiently deal with a large number of summary statistics, which has been a long-standing limitation of ABC approaches (few summary statistics could be used and choosing them was near impossible), by either giving different weights to summary statistics or by sorting out the redundant ones. This advance may explain the difference of estimation accuracy in our study compared to earlier ones (10).

We expect regression ABC to be most appropriate in situations where we have a large phylogeny and where there is enough biological knowledge to simulate a rather detailed individual-based model. Indeed, the number of simulated phylogenies required to perform regression ABC does not increase with phylogeny size (although the time to simulate each phylogeny does increase), whereas maximum likelihood functions tend to take much more time to converge as the phylogeny size and the model complexity increase. Another advantage of the ABC approach, which we are currently investigating, is that it can readily combine different sources of data. In the case of the Ebola epidemics for instance, there was an extremely detailed follow-up of incident cases. By

extracting summary statistics of these time series, we can further improve the power of the ABC and hopefully combine the precision of classical epidemiological methods to infer parameters such as R_0 with the ability of phylodynamics to access more hidden parameters such as, for Ebola, the fraction of transmission through dead hosts.

Samuel Alizon *- evolutionary biologist and Research Director at the CNRS based in Montpellier, France. Research interests: how human infectious diseases spread and evolve within and between their hosts.* *samuel.alizon@cnrs.fr*

References:

1. Campbell F, Strang C, Ferguson N, Cori A, Jombart T. When are pathogen genome sequences informative of transmission events? PLoS Pathog. 2018;14:e1006885.
2. Pimenoff VN, de Oliveira CM, Bravo IG. Transmission between Archaic and Modern Human Ancestors during the Evolution of the Oncogenic Human Papillomavirus 16. Mol Biol Evol. 2016;34:4-19.
3. Grenfell BT, Pybus OG, Gog JR et al. Unifying the epidemiological and evolutionary dynamics of pathogens. Science. 2004;303:327-32.
4. Stadler T, Kouyos R, von Wyl V et al. Estimating the Basic Reproductive Number from Viral Sequence Data. Mol Biol Evol. 2012;29:347-357.
5. Volz EM. Complex population dynamics and the coalescent under neutrality. Genetics. 2012;190:187-201.
6. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014; 345:1369-72.
7. Dudas G, Carvalho LM, Bedford T et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic Nature. 2017;544:309-315.
8. Saulnier E, Gascuel O, Alizon S. Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. PLoS Comput Biol. 2017;13:e1005416.
9. Stadler T, Kühnert D, Rasmussen DA, du Plessis L. Insights into the Early Epidemic Spread of Ebola in Sierra Leone Provided by Viral Sequence Data. PLoS Curr. 2014.
10. Ratmann O, Donker G, Meijer A, Fraser C, Koelle K. Phylodynamic inference and model assessment with approximate bayesian computation: influenza as a case study. PLoS Comput Biol. 2012;8:e1002835.