Inferring phylogenies from pandemic-scale genome datasets

Reconstructing phylogenetic trees from large collections of genome sequences is a computationally challenging task. We developed MAPLE, a method for performing phylogenetic inference on large numbers of closely related genomes, which might be useful when studying the evolution and spread of SARS-CoV-2 and of infectious pathogens in future pandemics.

This is a summary of:
De Maio, N. et al. Maximum likelihood pandemic-scale phylogenetics. *Nat. Genet.* https://doi.org/10.1038/s41588-023-01368-0 (2023).

Published online: 12 April 2023

The problem

Genome sequence data provide important insights into pathogen transmission and evolution, and phylogenetic methods are fundamental in the analysis of this data. For example, these methods enable the identification and tracking of pathogen variants[1], the tracking of infectious disease spread across and within countries[2], and the identification of mutations that are key to the pathogen's spread[3]. As sequencing technologies advance and are more widely adopted, genome data from infectious pathogens are becoming more important and ubiquitous in the analyses that guide the public health response to disease outbreaks. The COVID-19 pandemic provides a good example, as several million SARS-CoV-2 genomes from around the world are now available for analysis. However, algorithmic limitations of existing state-of-the-art phylogenetic inference methods mean that, at most, only a few thousand genome sequences can be analysed at a time, severely limiting the applicability of current methods to very large genomic datasets[4].

The solution

Our aim was to develop algorithms tailored for the inference of phylogenetic trees of many closely related genomes. More specifically, we wanted to improve the efficiency of probabilistic phylogenetic methods in this scenario, which is relevant to the COVID-19 pandemic and will likely also be relevant to large-scale infectious disease outbreaks in the future. We developed mathematical approximations that are both computationally convenient and accurate, assuming that the analysed genomes are closely related. This assumption also allowed us to concisely represent the genomes of the samples and their unsampled ancestors, greatly reducing the computational demand of probabilistic phylogenetics.

We have now developed a maximum likelihood phylogenetic inference software based on these principles and algorithmic ideas, called maximum parsimonious likelihood estimation (MAPLE). Using real and simulated SARS-CoV-2 genome data, we show that our software can infer phylogenetic trees more rapidly and from much larger collections of genomes than other, pre-existing maximum likelihood methods (Fig. 1). MAPLE can also perform more extensive phylogenetic tree searches owing to its reduced computational demand, resulting in more accurate inferred phylogenetic trees. Therefore, MAPLE enables the application of accurate probabilistic phylogenetic methods to genomic epidemiology datasets that are at least 1–2 orders of magnitude larger than was previously possible.

Future directions

Probabilistic phylogenetic frameworks allow seamless integration of different forms of data — such as geographic and temporal information — within phylogenetic analyses, and the use of sequence evolution models to realistically describe and reconstruct the complex interplay of different features of genome evolution — such as selection and mutational forces. We are currently working on extending the mathematical models in MAPLE to include biological complexities such as variation in mutation rates and selective pressure along the genome. We are also working on modelling sequence errors, such as assembly errors or contamination, which, if not accounted for, can adversely affect phylogenetic and downstream analyses. Our methods will now allow us to scale up popular phylogenetic and phylogenetic-based genome analysis tools to larger collections of genomes and therefore to perform more informative analyses. However, it is important to consider that our approach assumes that the analysed genomes are closely related, and that higher genomic divergence negatively affects both the computational demand and accuracy of MAPLE. Therefore, while our methods are useful in the context of genomic epidemiology where dense sampling over a short timescale is possible, they are less useful when comparing, for example, the genomes of different species.

The methods we have developed not only enable large-scale maximum likelihood phylogenetic inference of many closely related genomes but also can be used in the context of Bayesian phylogenetic inference[5], potentially leading to similar reductions in computational demand. In this respect, we plan to use our methods within existing software such as BEAST[5], which is commonly used for genome data analyses that are based on probabilistic phylogenetics, such as phylogeography[2] and phylodynamics[3].

Nicola De Maio & Nick Goldman
European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK.

EXPERT OPINION

"This is a technically sound and well-executed 'bespoke' solution to very large phylogeny inference problems. Overall, I think this is an innovative and important manuscript that will be of broad interest to the scientific community." **Sergei Kosakovsky Pond, Temple University, Philadelphia, PA, USA.**

FIGURE

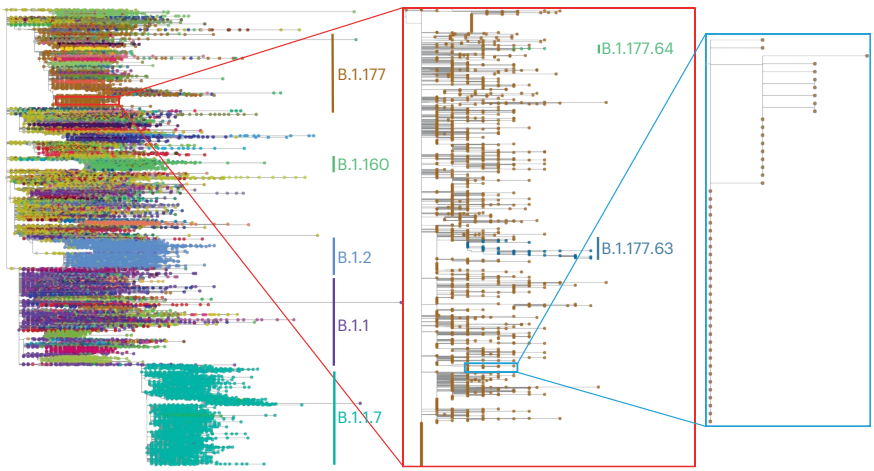

Fig. 1 | Use of MAPLE to infer an example phylogeny from a large SARS-CoV-2 genome dataset. Different colors distinguish different SARS-CoV-2 lineages and some clades have been labeled to illustrate the level of detail shown. Left: MAPLE was applied to SARS-CoV-2 genome sequence data containing 500,000 samples to estimate their phylogeny. Centre: a subtree of 3,600 B.1.177 lineage samples (red inset, ~100x magnification of portion of left panel). Right: a subtree of 49 B.1.177 lineage samples (blue inset, ~100x magnification of portion of red inset).

BEHIND THE PAPER

The development and application of phylogenetic methodologies, both for genomic epidemiology and more generally, has been our research focus for many years, but during the first few months of the COVID-19 pandemic in 2020, it became clear that most of our computational tools for genome data analysis were not adequate to handle large datasets of the size that became available for sequenced SARS-CoV-2 genomes. For this reason, we have since focused on the development of computational tools tailored for the analysis of many closely related genomes, which is the typical scenario in genomic epidemiology, with the aim, for example, of studying the emergence and proliferation of infectious pathogen variants. **N.D.M.**

REFERENCES

1. O'Toole, A. et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* **7.2**, veab064 (2021).
A research article presenting the popular phylogenetic tool 'pangolin' for lineage assignment.
2. Lemey, P. et al. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* **5**, e1000520 (2009).
A research article presenting one of the most widely used phylogeographic methods.
3. Volz, E. M., Koelle, K. & Bedford, T. Viral phylodynamics. *PLoS Comput. Biol.* **9**, e1002947 (2013).
A review article describing the field of phylodynamics.
4. Hodcroft, E. B. et al. Want to track pandemic variants faster? Fix the bioinformatics bottleneck. *Nature* **591**, 30–33 (2021).
Commentary on the need for advances in computational methods for the analysis of pandemic-scale genome data.
5. Bouckaert, R. et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).
A research article describing one of the latest versions of BEAST, the most widely used Bayesian phylogenetic software in genomic epidemiology.

FROM THE EDITOR

"Sequencing of SARS-CoV-2 genomes during the pandemic has produced a flood of data, but extant tools are incapable of handling the millions of sequences now available. MAPLE presents a way to analyse this amount of data, and with the need for variant monitoring, will contribute to ongoing efforts to characterise this and other infectious disease, as well as other applications of phylogenetics." **Editorial Team, *Nature Genetics*.**