Supplementary Material S1: A review of phylogenetic concepts

Phylogenetic models

*Phylogenetic models* have become a common tool in the study of infectious disease transmission [1–5] and are used to detect transmission chains. Those models use pathogen genetic sequences, collected from infected individuals, to infer the history of the epidemic, which is represented by a tree structure known as a *phylogeny*, e.g. Fig 1. The phylogeny can be fully represented with two components,

1. The *topology*: Represented by a list indicating the order in which the tips of the tree meet to form internal nodes,

2. The *branch lengths*: Expressed in expected nucleotide substitutions per base pair (nt/bp), they indicate genetic distance.

The first component summarizes *clades* in the tree. In Fig 1, that list would be (with \(x\) designating the viral DNA sequence from patient \(x\)): \([\{1\}, ..., \{7\}, \{3, 4\}, \{6, 7\}, \{2, 3, 4\}, \{5, 6, 7\}, \{1, ..., 4\}, \{1, ..., 7\}, \{\text{Outgroup}, 1, ..., 7\}\]. Any set of genetic sequences forms a clade if and only if it contains all sequences descended from an arbitrary node. Trivially, all sets of size 1 and the set comprising the entire sample form clades. A set including only viral sequences from patients 1 and 2 would not constitute a clade however, since their most recent common ancestor has four descendants, namely, the sequences from patients 1 to 4.

Due to the availability of sizeable viral DNA sequence databases collected in the context of antiretroviral drug resistance testing [6–8], phylogenetics has been used extensively to study *Human Immunodeficiency Virus (HIV)* epidemics [4,10]. The Quebec *HIV* genotyping program database for example, as of 2017, contains 27,487 *HIV* sequences from 9,687 *HIV*-positive individuals, living mostly in Montreal, Quebec, Canada [11].
Fig 1. Phylogeny for a sample of seven HIV sequences obtained from different HIV-positive patients, with an outgroup used to place the root of the tree. The sampled viral sequences are placed at the tips of the tree, and the root corresponds to the sample’s most recent common ancestor. Branches link the different nodes in the tree. Their lengths are expressed in expected nucleotide substitutions per base pair (nt/bp), a measure of genetic distance, indicating dissimilarity between sequences.

Transmission clusters and transmission chains

Disagreements persist in the literature regarding the difference between transmission chains and transmission clusters [12,13]. When studying viral transmission among members of a population, we often consider only one viral DNA sequence per infected individual, and therefore, a given clade contains sequences from all sampled individuals who became infected following a specific outbreak or introduction of the pathogen into the population. As a result, those cases must belong to the same transmission chain. All conventional phylogenetic clustering methods require transmission clusters to be non-nested clades, and so, individuals that co-cluster must belong to the same
transmission chain. Clustering algorithms therefore identify, first and foremost, distinct transmission chains, that are then termed “transmission clusters” if they satisfy certain criteria, usually a small enough genetic distance between their members and a minimum confidence level for their estimation. In other words, we can use standard phylogenetic clustering algorithms to find transmission chains: the only difference is in the requirements for the clades. In this paper, the terms partitioning and clustering are synonymous, and we use the term “transmission chain”.

**Confidence thresholds for transmission chain inference**

Phylogenetic studies for the inference of transmission chains in HIV-1 epidemics have relied mostly on methods deemed *nonparametric*, as they tend to depend on a number of *ad hoc* rules applied a posteriori to phylogenetic estimates [14]. In particular, availability of software like MEGA and PAUP* [15,16] has led to widespread adoption of maximum likelihood phylogenetic reconstruction, coupled with the bootstrap to evaluate confidence in the inferred clades. The scheme involves repeated resampling of site indices, and construction of simulated sequences based on the indices obtained. A phylogeny is then fitted to each simulated sample, and clades are listed. The proportion of times each clade appears in the obtained phylogenies is computed, which serves as the previously-mentioned measure of confidence. In that context, chain estimation depends on an arbitrary cutoff applied to bootstrap support estimates, usually between 70% and 95% [11,17,18].

Alternatively, software like BEAST and MrBayes [19,20] have popularised Bayesian phylogenetic estimation. Both are based on versions of the *Markov Chain Monte Carlo* (MCMC) algorithm, that numerically approximate posterior distributions for a variety of evolutionary and phylogenetic parameters. They also provide posterior probability support for clades, a Bayesian alternative to bootstrap support. Most of the times, studies require posterior probability support of 1 to conclude in a clade forming a genuine transmission chain [21,22].

A popular alternative to both bootstrap-based and Bayesian estimates of clade support is the approximate likelihood ratio test (aLRT) for branches [23], more specifically the aLRT-SH non-parametric branch support estimate. It is available in
PhyML \cite{24} and IQtree \cite{25}. It consists in a test statistic that indicates to what extent a given branch contributes to a gain in the phylogenetic likelihood, in comparison to the case where its length is reduced to zero, thus eliminating it altogether.

**Distance requirements for transmission chain inference**

In addition to clade confidence requirements, studies often impose a within-chain genetic distance requirement, usually between 0.01 nt/bp and 0.05 nt/bp \cite{26}. Distance requirements may be applicable to mean \cite{8}, median \cite{27}, or maximum *patristic* distances \cite{11}, also called tree or *cophenetic* distances, that is, distances calculated by summing branch lengths along the shortest path between any two tips in the phylogeny. The *ClusterPicker* algorithm instead formulates that requirement in terms of maximum within-chain *p*-distances, e.g. the Hamming distance \cite{26}. Both *p*-distances and patristic distances are measures of genetic distance, with the former being computed separately for each pair of sequences, and the latter being based on information obtained from the whole sample \cite{22}. As noted previously, sets of sequences that meet both the confidence and distance requirements are usually termed “transmission clusters”.

**Summarising samples of trees**

Both the bootstrap and the MCMC algorithms produce samples of trees that must be summarised before the application of confidence or genetic distance criteria. One strategy involves using the maximum likelihood (ML) or maximum posterior probability (MAP) tree, and applying criteria solely to the clades they contain. However, it is common for phylogenetic estimation procedure to produce several trees with likelihood or posterior probability very close to the maximum. The data may therefore support a wide variety of phylogenies, and this is not properly reflected by the ML or MAP trees. To address the issue, a majority-rule consensus tree can be constructed instead: in it, bifurcations support clades found in more than 50% of sampled trees; otherwise, multifurcations are used \cite{28}. The majority-rule consensus tree can be shown to always exist, but it lacks branch lengths \cite{22}. In other words, it provides only a nesting order for clades, which precludes the application of patristic
distance requirements.

**Alternative partitioning approaches**

Cutoffs are however hard to justify rigorously and so, methods grounded in more explicit definitions of chains have been published. For example, proposed the so-called *Gap Procedure*, a fast pure distance-based approach that requires minimal tuning. Indeed, it involves a single tuning parameter, whose purpose is to eliminate the effect of outliers on estimation and whose value rarely need to be changed. In a similar vein, formulated DM-PhyClus, a Bayesian algorithm that aims to limit reliance on hard thresholds and to offer a straightforward measure of uncertainty for estimates of chain membership. Other options are also available.

**Computational challenges**

The heavy computational burden of conventional phylogenetic inference is problematic in light of the fast increase in the size of sequence databases, and can therefore restrict its applicability. Thankfully, software designed to handle larger datasets is now available. RAxML and FastTree, for example, make use of heuristics in phylogenetic optimisation to improve scalability of the maximum likelihood phylogenetic methods. Partitioning of large datasets in a purely Bayesian paradigm is a computational challenge that has not yet been fully overcome, although vast progress has been made thanks in part to GPU computing.

**References**


