Review

CellPress

Clinical applications of pathogen phylogenies

Matthew Hartfield, Carmen Lía Murall, and Samuel Alizon

Laboratoire MIVEGEC (UMR CNRS 5290, IRD 224, UM1, UM2), 911 avenue Agropolis, B.P. 64501, 34394 Montpellier Cedex 5, France

Innovative sequencing techniques now allow the routine access of whole genomes of pathogens, generating vast amounts of data. Phylogenetic trees are a common method for synthesizing this information. Unfortunately, these trees are often seen only as a visual support to guide researchers, thus neglecting the value of employing phylogenetic techniques to perform hypothesis testing on clinical questions. These include investigating how a pathogen spreads within a patient, or whether the infection severity (often measured by virus load) is controlled by viral genetics. Advances in methodology mean the time is ripe for combining phylogenies with clinical data to better understand and fight infectious diseases.

Phylogenies and their current use in clinical research

Several human RNA viruses, including HIV, hepatitis C virus (HCV), and influenza, are characterized by high mutation rates, so that samples taken from different patients will differ vastly from one another [1,2]. This fact even remains true for isolates taken from different locations within a single patient [3]. Phylogenetic analyses (see Glossary) are now a common tool in clinical studies that aim to illustrate, and then determine, the impact of this inherent diversity.

Phylogenetic analysis is a method that aims to ascertain how individual samples are genetically related, assuming that differences arise through mutation only (and not through other evolutionary processes, such as recombination). In a clinical context, it may be used to infer whether the common genetic background of pathogens means that they arise from a similar source. The output takes the shape of a 'phylogenetic tree'. Figure 1 outlines an example of a phylogeny created from a transmission chain, and what information is portrayed with this type of analysis.

One of the first applications of phylogenetic methods in clinical studies was to use 16s rRNA sequences obtained from Whipple's-disease-associated bacterium in order to classify its bacterial genus and determine related species [4,5]. Another seminal study applied phylogenetic methods to HIV samples taken from a single host over several years to determine how the infection progresses with time [6]. Phylogenies were also employed to reveal single transmission events; notable examples include detecting HIV infection from a dentist to patients [7], and nosocomial transmission of HCV either during cardiac surgery [8] or via contaminated saline vials [9].

Since these initial studies, phylogenetic analysis has become a common tool for investigating disease spread, aided by increased computing power and low sequencing

Glossary

Ancestral state: the trait status of the founding individual, either for the entire phylogeny or for a subclade, from which all other samples are descended (see Figure 3 in main text).
Burn-in (for MCMC): the first few results produced by an MCMC algorithm (generally set at 10–20%). These outputs are usually discarded because they produce unrealistic estimates.
Clades, subclades: a collection of tips on a phylogeny, representing a branch and its related individuals. The clade size is simply its size, in terms of number of tips it encompasses, or the length of time it has existed for.
Cluster: a collection of tips that share a common state value, which could indicate a pathogenic effect over a trait.
Convergence (for MCMC): when the MCMC algorithm starts to sample around the most likely outcomes, rather than produce unrealistic outputs.
Demographic history: how the size of a population changes over time. That is, does it stay approximately constant, or change over time (increase, decrease, or oscillate)? For example, most pathogenic agents are reduced in population size following transmission to a new host.
Genetic divergence: a measure of to what extent two different genomic samples differ with regards to their genome.
Markov Chain Monte Carlo (MCMC): a computational technique for statistical analysis, where the parameters are continually shifted so that the most likely outcomes are explored, instead of producing a single output.
Maximum likelihood (ML): a statistical technique for data analysis, where a single model is fitted (a phylogeny, for example) that matches up best to the samples. However, this single output can change depending on the input parameters used.
Molecular clock: the rate of change of nucleotides (substitution rate) over time. If nucleotides are changed from one type to another at the same rate for each clade, the clock is fixed or strict: otherwise if there are widespread changes in the fixation rate in different clades, it is relaxed.
Node: a point in the phylogenetic tree where one ancestor breaks into two descendants.
Phylogenetic analysis: a systematic method for visually inferring the relationship between different samples, and determining the most likely chain of descent among them. Historically applied to determining how species were related in evolutionary studies, but now used to analyse transmission chains of infectious diseases.
Phylogenetic signal: a measure of to what extent traits (such as infection outcome, location, or others) can be explained by how related they are in the phylogeny. More intuitively, a measure of to what extent proximity in the phylogeny is associated with proximity in trait value.
Phylogeography: a biological research area that aims to use phylogenies to determine the geographic context of isolates. It is contemporarily applied to ascertaining the global spread of epidemics.
Substitution models: a mathematical model of how likely that a specific part of a genome changes over time; that is, the probability that each nucleotide will mutate into a different state (see Box 1 in main text). These are used to determine how related different samples are, which is important for building the right phylogeny.
Tips: the points at the end of a phylogeny, representing the isolated sampled and analysed. They can bear a trait (or marker) value.
Virulence: the degree of harm that a pathogen causes to its host.

Corresponding author: Hartfield, M. (matt.hartfield@gmail.com).
Keywords: phylogenetic analysis; transmission network; immune escape; HIV; HCV.

1471-4914/
© 2014 Elsevier Ltd. All rights reserved. http://dx.doi.org/10.1016/j.molmed.2014.04.002

CrossMark

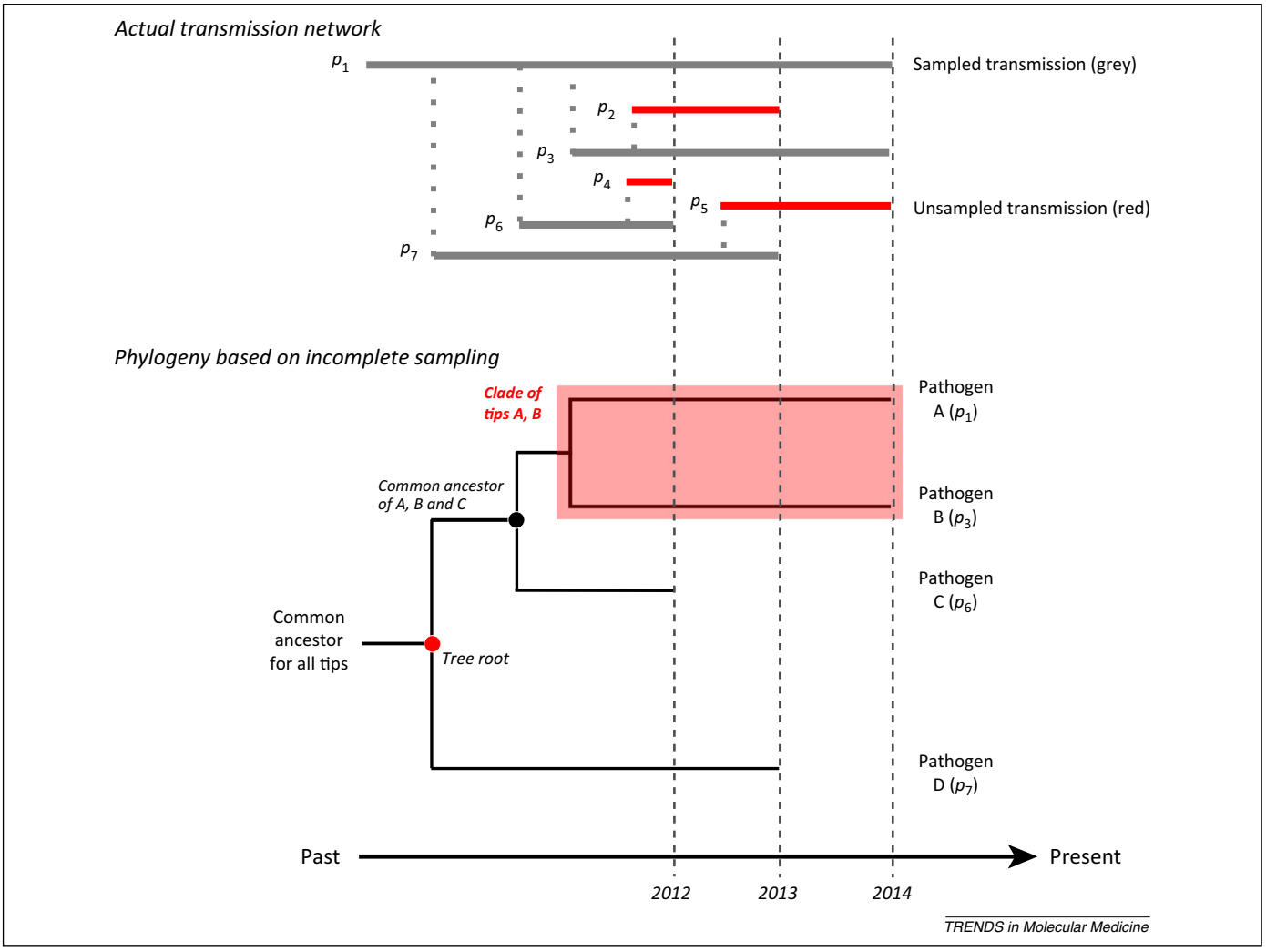

Figure 1. How a pathogen phylogeny reflects a transmission chain. Top: a schematic describing a transmission network, with horizontal lines representing cases, and vertical broken lines showing infection. Bottom: how sampling from the transmission network determines the layout of a clinical phylogenetic tree. Patients are denoted p_1 to p_7; in this example, patients 1, 3, 6, and 7 are sampled to produce the phylogeny, represented as pathogens A to D on the tips. Patients 2, 4, and 5 (red lines in the transmission network) are not sampled. The nodes of each branch represent the common ancestor to them. A collection of tips and their nodes is called a clade; the clade for tips A and B is highlighted in red. The root represents the common ancestor for all samples. The horizontal axis is scaled by the degree of genetic change (for example, nucleotide substitutions per site) over time. The vertical axis has no scale; spacing is included purely for clarity. Note that all tips do not line up exactly with the present: this is because isolates can be collected at different times, and this information determines the phylogenetic shape.

costs. However, these methods are not often being used to their full potential. Conceptual breakthroughs have recently taken place in genetics research that can link specific pathogen genomes, and thus a pathogen phylogeny, to clinical information.

The power of modern phylogenetic methods lies with the fact that they approximate the full transmission chain of an outbreak, as opposed to single transmission events. These complete chains are otherwise very hard to re-create [10,11]; Figure 1 illustrates how a phylogeny reflects an incompletely-sampled epidemic. If the marker values (or 'traits' in evolutionary biology) on phylogeny tips were measured from individual patients, each node represents at least one transmission event leading to the creation of a new sample. However, more than one transmission could have taken place, with intermediate infections not being sampled (as is the case in Figure 1). Furthermore, if the transmission chain was large, the same pathogen strains can circulate rapidly between already-infected individuals (co-infections). In addition, fast-evolving viruses diversify extensively in patients, reducing their relatedness over time. These are some reasons why phylogenies cannot be used to prove with certainty that one patient directly infected another [12].

Although the phylogeny only approximates the transmission network, combining it with clinical data (e.g., infection outcome) can be useful in, for instance, shedding light on how pathogen genotypes control infection morbidity and mortality. Advanced phylogenetic methods also utilize techniques from the biology field of phylogeography, which aims to characterize the geographical spread of species (especially infectious diseases). These methods have provided great insights with delineating the local and worldwide transmission routes of viruses, such as H5N1 avian influenza [13], as well as the potential sources of new outbreaks [2,14].

These conceptual breakthroughs have been applied to improve treatment and medical-care delivery. For instance, phylogenetic methods have identified HIV risk groups in the UK [15] and Switzerland [16], along with

Download English Version:

https://daneshyari.com/en/article/2838543

Download Persian Version:

https://daneshyari.com/article/2838543

Daneshyari.com