Infection, Genetics and Evolution 11 (2011) 1825–1841

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Review

Phylogenetic and epidemic modeling of rapidly evolving infectious diseases

Denise Kühnert [1], Chieh-Hsi Wu [1], Alexei J. Drummond *

Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Auckland, New Zealand
Department of Computer Science, University of Auckland, Auckland, New Zealand

ARTICLE INFO

Article history:
Received 1 April 2011
Received in revised form 9 August 2011
Accepted 9 August 2011
Available online 31 August 2011

Keywords:
Coalescent
Phylodynamics
Statistical phylogeography
Phylogenetic epidemiology
Rapidly evolving viruses
Stochastic SIR

ABSTRACT

Epidemic modeling of infectious diseases has a long history in both theoretical and empirical research. However the recent explosion of genetic data has revealed the rapid rate of evolution that many populations of infectious agents undergo and has underscored the need to consider both evolutionary and ecological processes on the same time scale. Mathematical epidemiology has applied dynamical models to study infectious epidemics, but these models have tended not to exploit – or take into account – evolutionary changes and their effect on the ecological processes and population dynamics of the infectious agent. On the other hand, statistical phylogenetics has increasingly been applied to the study of infectious agents. This approach is based on phylogenetics, molecular clocks, genealogy-based population genetics and phylogeography. Bayesian Markov chain Monte Carlo and related computational tools have been the primary source of advances in these statistical phylogenetic approaches. Recently the first tentative steps have been taken to reconcile these two theoretical approaches. We survey the Bayesian phylogenetic approach to epidemic modeling of infection diseases and describe the contrasts it provides to mathematical epidemiology as well as emphasize the significance of the future unification of these two fields.

Contents

1. Introduction

Molecular phylogenetics has had a profound impact on the study of infectious diseases, particularly rapidly evolving infectious

* Corresponding author.
E-mail address: alexei@cs.auckland.ac.nz (A.J. Drummond).
[1] These authors contributed equally to this work.

1567-1348/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2011.08.005

agents such as RNA viruses. It has given insight into the origins, evolutionary history, transmission routes and source populations of epidemic outbreaks and seasonal diseases. One of the key observations about rapidly evolving viruses is that the evolutionary and ecological processes occur on the same time scale (Pybus and Rambaut, 2009). This is important for two reasons. First, it means that neutral genetic variation can track ecological processes and population dynamics, providing a record of past evolutionary events (e.g., genealogical relationships) and past ecological/population events (geographical spread and changes in population size and structure) that were not directly observed. Second, the concomitance of evolutionary and ecological processes leads to their interaction that, when non-trivial, necessitates joint analysis.

Arguably the most studied infectious disease agent to date has been human immunodeficiency virus (HIV) and it has been the subject of thousands of phylogenetic studies. These have shed light on many aspects of HIV evolutionary biology, epidemiology, origins, phylogeography, transmission dynamics and drug resistance. In fact, the vast body of literature on HIV makes it clear that almost every aspect of the biology of a rapidly evolving pathogen can be better understood in the context of the evolution of the virus. Whether it is retracing the zoonotic origins of the HIV pandemic or describing the interplay between the virus population and its host's immune system, a phylogenetic analysis frequently sheds light.

Although probabilistic modeling approaches to phylogenetics predate Sanger sequencing (Edwards and Cavalli-Sforza, 1965), it was not until the last decade that probabilistic modeling became the dominant approach to phylogeny reconstruction. Part of that dominance has been due to the rise of Bayesian inference (Huelsenbeck et al., 2001), with its great flexibility in describing prior knowledge, its ability to be applied via the Metropolis-Hastings algorithm to complex highly parametric models, and the ease with which multiple sources of data can be integrated into a single analysis. The history of probabilistic models of molecular evolution and phylogenetics is a history of gradual refinement; a process of selection of those modeling variations that have the greatest utility in characterizing the ever-growing empirical data. The utility of a new model has been evaluated either by how well it fits the data (formal model comparison or goodness-of-fit tests) or by the new questions that it allows a researcher to ask of the data. In this review we will describe the modern phylogenetic approach to the field of infectious diseases, and particularly with reference to Bayesian inference of the phylogenetic epidemiology of rapidly evolving viral pathogens such as Hepatitis C virus (HCV), HIV and Influenza A virus. The review is separated into two main sections. In Section 2 we discuss phylogenetic methods for reconstructing the history of infectious epidemics, including identification of origins, dating of common ancestors, relaxed phylogenetics and coalescent-based population dynamics. In Section 3 we review epidemiological models and finish by outlining progress in the development of phylodynamical models that marry statistical phylogenetics with dynamical modeling.

2. Reconstructing the history of infectious epidemics

The introduction of an efficient means of calculating the probability of a sequence alignment given a phylogenetic tree (known as the phylogenetic likelihood; Felsenstein, 1981) heralded the beginning of practical phylogenetic tree reconstruction in a statistical framework. At around the same time the coalescent was introduced: a theory relating the shape of the genealogy of a random sample of individuals to the size of the population from which they came (Kingman, 1982; see Section 2.3 for details). Both of these advances have been subsequently developed to the point that, together they enable the estimation of viral evolutionary histories and past population dynamics.

Bayesian inference brings together the *likelihood*, $\Pr(D|\theta)$ (the probability of the data given the model parameters) and the *prior*, $P(\theta)$ (the probability of the model parameters prior to seeing the data), so that the *posterior* probability of the model parameters (θ) given the data is:

$$P(\theta|D) = \frac{\Pr(D|\theta)P(\theta)}{\int \Pr(D|\theta)P(\theta)d\theta} \quad (1)$$

In a standard phylogenetic setting, the probabilistic model parameters include the phylogenetic tree, coalescent times and substitution parameters, and a prior probability distribution over these parameters must be specified. By using Kingman's coalescent as a prior density on trees, Bayesian inference can be used to simultaneously estimate the phylogeny of the viral sequences and the demographic history of the virus population (Drummond et al., 2002, 2005, 2006, see Box 1). Extension of phylogenetic inference methods to accommodate time-stamped sequence data (Rambaut, 2000; Drummond et al., 2002) and relaxation of the assumption of a strict molecular clock (Thorne et al., 1998; Kishino et al., 2001; Sanderson, 2002; Drummond et al., 2006; Rannala and Yang, 2007) provided sophisticated methods for ancestral divergence time estimation. For virus species that occupy more than one host species (e.g Influenza A), models that aim to detect cross-species transmission may provide clues to the origin of a virus strain in a host population (Reis et al., 2009).

2.1. Reconstructing the origins of an infectious disease

When a new epidemic emerges, one of the first goals is to trace it back to its genetic and geographic origin. The reconstruction of phylogenetic trees to infer the evolutionary relationships has been a key tool to uncover the origin of regional epidemics such as those resulting from HIV (Gao et al., 1999; Santiago et al., 2002), HCV (Pybus et al., 2009; Markov et al., 2009) and SARS coronavirus (SARS-CoV) (Li et al., 2005). Some studies have also attempted to use phylogenetic trees to draw conclusions about transmission history and geographic spread of viral epidemics (Motomura et al., 2003; Santiago et al., 2005; Gilbert et al., 2007). However, great care should be taken when coming to conclusions about aspects of the epidemic process that are not explicitly modeled in the reconstruction of the phylogenetic tree and even if they are, the user needs to consider the appropriateness of the underlying model assumptions.

One common and straightforward method used to identify the origin of an epidemic involves determining the non-epidemic genotype or lineage most closely related to the epidemic, i.e., the molecular sequences clustered most closely with the epidemic strain on a phylogenetic tree. While the method is intuitive, its success heavily depends on the collected data.

The closest simian immunodeficiency virus (SIV) relative of HIV-1 is SIVcpz (Gao et al., 1999; Santiago et al., 2002), which is harbored in chimpanzee sub-species *Pan troglodytes troglodytes* and *P.t. schweinfurthii* in the form of the respective sub-species specific SIV lineages SIVcpz*Ptt* and SIVcpz*Pts*. Although SIVcpz became the prime candidate for the zoonotic source of HIV-1 as soon as it was identified, alternative sources could not be ruled out due to the paucity of identified chimpanzee infections (Vanden Haesevelde et al., 1996). The source of HIV-1 was confirmed much later after the collection of SIVcpz from fecal samples of wild *P. t. troglodytes* apes in the Cameroon forest (Keele et al., 2006). HIV-1 groups M and N are much more closely related to sequences from the fecal samples than previously identified SIVcpz strains. This finding uncovered the distinct origins of HIV-1 group M (pandemic) and group N (non-pandemic) traced to chimpanzee communities of southeastern and central Cameroon respectively. The