J Med Microbiol Infec Dis, 2016, 4 (1-2): 8-10

Mini Review

A Note on Evolutionary Rate Estimation in Bayesian Evolutionary Analysis: Focus on Pathogens

Kayhan Azadmanesh[1], Sana Eybpoosh[2, 3*]

[1]Department of Virology, Pasteur Institute of Iran, Tehran, Iran;
[2]Department of Epidemiology and Biostatistics, Research Centre for Emerging and Reemerging Infectious Diseases, Pasteur Institute of Iran, Tehran, Iran;
[3]HIV/STI Surveillance Research Center, and WHO Collaborating Center for HIV Surveillance, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran.

Received May 30, 2017; Accepted Jun 29, 2017

Bayesian evolutionary analysis provide a statistically sound and flexible framework for estimation of evolutionary parameters. In this method, *posterior* estimates of evolutionary rate (μ) are derived by combining evolutionary information in the data with researcher's *prior* knowledge about the true value of μ. Nucleotide sequence samples of fast evolving pathogens that are taken at different points in time carry evolutionary information that allow for estimation of evolutionary rates and divergence dates. If the amount of genetic change in the data is proportional to the time elapsed since divergence from the common ancestor, then one can directly estimate the μ from the data. Otherwise, external sources should be used to select the μ value, and use it as a fixed *prior* in Bayesian evolutionary analysis. This note provides a brief overview on how to assess the adequacy of the evolutionary information in the data and provides some recommendations for obtaining proper evolutionary rate *priors* from external sources. The recommendations generally highlight the need for the candidate μ prior to be a good representative of the evolutionary rate in the data at hand. This will be achieved by ensuring that the samples that are the source of the candidate μ value have been under relatively similar evolutionary forces as the data at hand. As the evolutionary forces acting on a particular set of samples varies across different study settings and species type, selection of prior for μ should be founded on a thorough understanding of the species under study at biological and social levels. *J Med Microbiol Infec Dis, 2016, 4 (1-2): 8-10.*

Keywords: Evolution, Evolutionary rate, Bayesian Evolutionary Analysis, Phylogeny.

Recently, there has been growing interest in the analysis of serial nucleotide sequence samples that are taken at different points in time, for fast evolving pathogens [1]. The various sampling dates allow for calibration of evolutionary changes and estimation of absolute evolutionary rates (μ) and divergence times [2]. Bayesian molecular clock dating methods provide a *posterior* estimate of rate and time. The *posterior* is derived by combining evolutionary information in serially sampled sequence data with researcher's *prior* knowledge about the value of these evolutionary parameters [3].

Generally speaking, for accurate estimation of evolutionary rates, sampled sequences need to cover a wide time span, so that substantial evolution can occur in the data. Also, given a fixed sampling time span, sequences with higher mutation rates, lower heterogeneity, and longer lengths are more suitable for evolutionary rate estimation as opposed to those with lower mutation rate, higher heterogeneity, and shorter sequence length. Moreover, for proper estimation of evolutionary rates, it is desired for the amount of divergence in the dataset to increase linearly with time. Such "clock-like" behavior of the data can be statistically tested using root-to-tip linear regression or maximum likelihood methods. These methods test if the magnitude of genetic divergence of the sequences significantly increases with time (*i.e.,* if μ and its confidence interval excludes 'zero.' For a detailed explanation of these methods see [2]). Root-to-tip regression of genetic distances against sampling time can be done using tools such as TempEst (formerly known as Path-O-Gen) [4]. In cases of data with no clock-like behavior, direct estimation of evolutionary rates from data is not recommended. In such situations, Bayesian estimation of evolutionary rate should be based on evolutionary rate *priors* obtained from external sources. Two major external sources can be used for this purpose, including 1) external sequence datasets and 2) available literature reporting estimates for μ [5, 6].

Several studies have estimated evolutionary rates for different pathogen genes. Online genetic databases, such as

***Correspondence:** Sana Eybpoosh
Department of Epidemiology and Biostatistics, Pasteur Institute of Iran, No. 69, Pasteur Ave, Tehran, Iran, 1316943551.
Email: sana.eybpoosh@gmail.com
Tel: +98 (21) 64112121 **Fax:** +98 (21) 66465132

National Center for Biotechnology Information (NCBI), are also rich sources for selecting clock-like datasets for a given gene/pathogen. However, the question is which of the many rate estimates or data subsets should be used for a Bayesian evolutionary analysis? Here, we will provide a brief guideline on how to make this decision, using the HIV as an example.

First, it is recommended that the study sample - referred here as "study dataset"- and the dataset used for external estimation of evolutionary rate - referred here as "external dataset"- resemble each other regarding factors affecting the evolutionary rate value. It is highly recommended to choose sequences that have been sampled from the same pathogen and the same strain (*e.g.,* extracting external dataset from the same HIV-1 groups and preferably the same HIV-1 subtypes as study dataset).

Moreover, it is necessary for the nucleotide sequences included in the external dataset to be under the same evolutionary forces as those sequences in the study dataset. In practice, ensuring such evolutionary similarities between the two datasets may not be so trivial; however, the following recommendations can be helpful in approaching this goal.

Generally speaking, it is recommended to select sequences from similar host populations, genes, and genomic regions. The rationale here is that within a population, individuals are usually under similar selective pressures, while different communities can experience different evolutionary forces. For example, people within a country or particular geographic region have relatively similar culture, diet, healthcare services, genetic composition, and inherent immunity, which impose similar selective pressures on the pathogens infecting them [7].

The epidemic growth rate also affects the evolutionary rate of the pathogen and is relatively homogeneous within a population [8]. However, it is notable that some sub-populations with different epidemic growth rates may exist within a population. For example, individuals within a specific HIV risk groups, such as homosexuals, usually have similar epidemic growth rates, but might show different epidemic growth rates than other risk groups within the same major population, such as people who inject drugs. If this is the case, it is recommended to select the external dataset and the study dataset from the same subpopulation (here, similar HIV risk groups).

Moreover, factors such as care and treatment interventions and disease stage of the host impose considerable selective pressures on the infecting pathogen. In the case of HIV infection, for example, Anti-Retroviral Treatment (ART) can result in specific selective pressures on virus genome [9]. Therefore, if the study samples are ART-free individuals, or if the genomic region(s) under study are free of drug-resistance mutations, the μ estimated from an external dataset with similar characteristics would better represent the actual evolutionary rate of the study dataset.

Also, hosts with rigorous immune systems impose higher pressure on the pathogen, causing immune escape mutations in the pathogen genome [10, 11]. Therefore, selecting an external dataset that includes patients with relatively similar levels of immunity as the ones in the study dataset would also be beneficial. In the circumstances such as HIV infection, the host's immune system aggravates as the disease progresses. In such cases, selecting an external dataset that comprises patients with relatively similar disease stage as the ones in the study dataset would be preferred.

Finally, different genes and genomic regions are believed to be subjected to various selective pressures and evolutionary constraints (*e.g.,* higher pressure on coding vs. non-coding regions, and on 1st and 2nd codons vs. 3rd codon) [8]. Therefore, it is strongly recommended to pick similar genes or genomic regions from external datasets for evolutionary rate estimation.

This note provides a brief overview of the factors affecting the evolutionary rate of pathogens that need to be considered when obtaining evolutionary rate priors from external sources. The recommendations given here, focus on selecting an external dataset for obtaining evolutionary rate priors. However, these recommendations are also applicable when obtaining the evolutionary rate prior from the existing literature. It is noteworthy that these recommendations may vary depending on the context of the research study, and species under investigation. Selection of external datasets or evolutionary rate estimates should be founded on a thorough understanding of factors affecting the evolutionary rate of species under study at biological and social levels.

ACKNOWLEDGEMENT

The study was not supported by any grants or funds.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest associated with this manuscript.

REFERENCES

1. Rodrigo AG, Goode M, Forsberg R, Ross HA, Drummond A. Inferring evolutionary rates using serially sampled sequences from several populations. Mol Biol Evol. 2003; 20 (12): 2010-8.

2. Drummond A, Pybus OG, Rambaut A. Inference of Viral Evolutionary Rates from Molecular Sequences. Adv Parasitol. 2003; 54: 331-58.

3. Stadler T, Yang Z. Dating phylogenies with sequentially sampled tips. Syst Biol. 2013; 62 (5): 674-88.

4. Rambaut A, Lam TT, Carvalho LM, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2016; 2 (1): vew007.

5. Drummond AJ, Rambaut A, BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007; 7: 214.

6. Drummond AJ, Rambaut A. Bayesian evolutionary analysis by sampling trees. In: Lemey P, Salemi M, Vandamme AM. The Phylogenetic Handbook. A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. 2nd ed. Cambridge: Cambridge University Press; 2008; 564-91.

7. Laland KN, Odling-Smee J, Myles S. How culture shaped the human genome: bringing genetics and the human sciences together. Nat Rev Genet. 2010; 11 (2): 137-48.

8. Vandamme AM. Basic concepts of molecular evolution. In: Lemey P, Salemi M, Vandamme AM. The Phylogenetic Handbook. A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. 2nd ed. Cambridge: Cambridge University Press; 2008; 3-28.

9. Carvajal-Rodriguez A, Crandall KA, Posada D. Recombination favors the evolution of drug resistance in HIV-1 during antiretroviral therapy. Infect Genet Evol. 2007; 7 (4): 476-83.

10. Moore CB, John M, James IR, Christiansen FT, Witt CS, Mallal SA. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. Science. 2002; 296 (5572): 1439-43.

11. Goulder PJR, Brander C, Tang Y, Tremblay C, Colbert RA, Addo MM, Rosenberg ES, Nguyen T, Allen R, Trocha A, Altfeld M, He S, et al. Evolution and transmission of stable CTL escape mutations in HIV infection. Nature. 2001; 412 (6844): 334-8.