PhyloGeographeR: phylogeography simulation incorporating geographic information

Luiz Max F. de Carvalho
Program for Scientific Computing (PROCC) - Fiocruz
Brazil

October 16, 2013

Abstract

This PDF is just a very basic proposal of some strategies for structuring a continuous time Markov chain (CTMC) emission matrix ($\mathbf{Q}$) incorporating geographic information. I introduce some definitions and notation and then explain the idea behind PhyloGeographeR and some aspects of the silly piece of code I have written to implement it.[1]

1 Background and motivation

Since Lemey et al (2009) [6] proposed a Markov model for discrete phylogeography, great attention has been paid to applying the discrete phylogeography methods implemented in the BEAST package to the study of the spatial spread of many organisms, mainly human pathogens, such as influenza [7, 1] and HIV [3, 4]. In Faria et al (2011) [5] the authors argue that a quantitative understanding of the spatio-temporal spread of viruses could provide not only means for formal hypothesis testing, but also an integrated prediction framework.

This representation of geography is, however, quite limited, and relies on strong simplifications of the physical scenario. This may, however, be ameliorated by incorporating neighborhood structure into the CMTC rate matrix formulation, as well as exploring other geographic variables and models, such as source-sink, population gravity and distance-weighted flow.

In this context, a simulation tool that is able to incorporate geographic information into a CTMC-based discrete trait simulator seems to be a good option for generating synthetic data under different phylogeographic scenarios. Such simulations could prove useful for the task of testing the available inference frameworks as well as gaining insight into the spatial dynamics itself.

This document and the R code described herein concern the problem of incorporating geographic information in CTMC models of discrete phylogeography.

[1] DISCLAIMER: This is **not** a vignette. This text is intended solely to help explaining some aspects of the simulation tools included in "phylogeographeR.R" and its companion, "phylogeographeR_aux.R"

2 Some useful definitions and notation

In what follows, I introduce some notation, outline the basics on Continuous-time Markov chains and present some ideas on how to incorporate geographic information on CTMC-based discrete trait generation along a phylogeny.

2.1 Continuous-time Markov chain (CTMC)

Continuous-time Markov chains (CTMCs) are very useful modeling tools, widely used in models of chemical reactions, queueing systems and molecular evolution. A CTMC is a discrete stochastic process $X(t), t \geq 0$ taking values in {0,1,2,. . . , K} for which the Markov property holds, i.e, for $s, t \geq 0$ and all states i, j, $x(u)$:

$$\begin{aligned} P(X(t+s)) = j|X(s) = i, X(u) = x(u))) &= P(X(t+s) = j|X(s) = i) \\ &= \pi_{ij}(t) \quad \forall 0 \leq u < s \end{aligned} \tag{1}$$

Where π_{ij} denotes the stationary transition probability.

A desirable property is that the process is completely governed by a transition matrix $\mathbf{Q}$. We formulate $\mathbf{Q}$ so that upon exponentiation it yields a stochastic matrix:

$$\pi_{ij}(t) = e^{t\mathbf{Q}} \tag{2}$$

$$\mathbf{Q} = \mu\mathbf{S}\Pi \tag{3}$$

Where μ is a rate scalar, $\mathbf{S}$ is symmetric or nearly symmetric K-order matrix and $\Pi = \pi_1, \pi_2, \ldots, \pi_k$ is a diagonal matrix.

For such, the following condition is needed:

$$\mathbf{Q_{ii}} = -\sum_{i \neq j}^{K} \mathbf{Q_{ij}} \quad \forall i \in \{1, 2, 3, \ldots, K\} \tag{4}$$

Lemey et al (2009) [6] proposed using a CTMC to model the diffusion of a given organism through time and space. In this framework, the locations of sampling for each taxon in a phylogeny are treated as CTMC observed states, and branch lengths are seen as the elapsed times between transitions. This allows for the inference of transitions through time, therefore providing insight into the transmission network of pathogens. Moreover, it is possible to infer the probability distribution for each state (location) in every node of the phylogeny. This is specially important for the root of the given phylogeny, for it offers insight into the spatial origins of epidemics.

Although this approach provides an efficient statistical framework to the study of phylogeography, it is also restricted to a simple representation of geographic space.

Several studies have adressed this problem by incorporating environmental information into prior distributions for the rates of the transition matrix [6, 3, 2, 7], for example, geographic distances, differences in population sizes and vaccination status between locations. These “informed” priors have played an important role in testing evolutionary hypothesis about spatial diffusion. In this document, I study how to incorporate such environmental data into phylogeography simulation. Importantly, such development should take CTMC-associated constraints into account, and further analysis is in order to avoid ill-posed decomposition problems on matrix exponentiation.

2.2 Neighborhood structure

Consider a region $\mathbf{R}$ consisting of K areas A_i, $i \in \{1, 2, 3, \ldots, K\}$, to which belongs a vector of attributes $\mathbf{v_i}$ each.

3 General Structure

To attain to the restriction in 4, a wrapper named `reg.matrix()` is called to regularize the matrix by adjusting its principal diagnonal so that row-wise summation equals zero.

3.1 Models

3.1.1 Homogeneous

We begin by specifying the simplest possible model for $\mathbf{Q}$. If there is no preference of flux between locations, we make each non-diagonal entry of $\mathbf{Q}$ equal to $1/K$ and subsequentially parameterize diag($\mathbf{Q}$) to meet restriction (4), which yields $\mathbf{Q_{ii}} = -(K-1)/K$. In the limit of large K, $\mathbf{Q_{ii}} \to -1$, placing almost all mass outside the diagonal.

This model is intended to serve as a base-line, for instance to investigate the sole influence of tree topology, i.e., population dynamics when there is no detectable geographical signal.

3.1.2 Contiguity-constrained Homogeneous

Sometimes it can be of interest to assume that flow can only occur between neighbor areas, whilst not assuming any directionality. This "model" can be achieved by formulating $\mathbf{Q}$ as the element-wise multiplication of the $\mathbf{Q}$ described above and the contiguity matrix. Let k_i be the degree – i.e., number of neighbors – of area i the and $\hat{k}$ be the average degree of the contiguity graph. It is easy to see that:

$$\mathbf{Q_{ii}} = -\frac{k_i}{K} \quad \forall i \in \{1, 2, \ldots, K\} \tag{5}$$

and $E[diag(\mathbf{Q})] = -\hat{k}/K$.

3.1.3 Distance-informed

In the original paper Lemey et al (2009) [6] proposed a distance-informed prior for **Q**, in which infinitesimal rates are proportional to:

$$\mathbf{Q_{ij}} = C \frac{d_{ij}^{-1}}{\sum_{i<j} d_{ij}^{-1}} \tag{6}$$

where d_{ij} is the great-circle distance between locations i and j.

3.1.4 "Patchy"

3.1.5 Gravity

Gravity priors are built so that they attain to the form:

$$d(i,j) = f(\mathbf{v_i}), \sum_{i \neq j}^{K} d(i,j) = 1 \tag{7}$$

3.1.6 Source-sink

The function `build.ss.matrix()` builds **Q** according to a "source-sink" model, where some locations act mainly as emiters (sources) and others as receptors (sinks). A parameter, ρ, controls the magnitude of this flow. Let $\delta_{ss}(i,j)$ be an indicator function that assigns 2 to source-to-sink entries in **Q**, 1 to sink-to-source ones and 0 otherwise .Then the model is specified according to:

$$\mathbf{Q_{ij}} = \delta_{ss}\{i,j\} \tag{8}$$

4 Implementation

4.1 Dependencies

4.2 Main Features

At the moment, a the code has a few features:

- Ability to simulate coalescent processes with varying growth rates;
- Benefit from the R (**seqgen**) adaptation of the original C code by Prof. Andrew Rambaut of **Seq-Gen** to generate nucleotide/aminoacid sequences under quite complex models;

4.3 Options

4.4 TODO list

- Further theoretical study on the geographical models to structure **Q**, to ensure real eigen decomposition is possible.
- More population growth models. Currently only exponential (and constant, of course) growth is implemented.
- Maybe make a function of it, which would read a file of parameters.

References

[1] J. Bahl, M. I. Nelson, K. H. Chan, R. Chen, D. Vijaykrishna, R. A. Halpin, T. B. Stockwell, X. Lin, D. E. Wentworth, E. Ghedin, Y. Guan, J. S. Peiris, S. Riley, A. Rambaut, E. C. Holmes, and G. J. Smith. Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. *Proc. Natl. Acad. Sci. U.S.A.*, 108(48):19359–19364, Nov 2011.

[2] L. M. de Carvalho, L. B. Santos, N. R. Faria, and W. de Castro Silveira. Phylogeography of foot-and-mouth disease virus serotype O in Ecuador. *Infect. Genet. Evol.*, 13:76–88, Jan 2013.

[3] N. R. Faria, I. Hodges-Mameletzis, J. C. Silva, B. Rodes, S. Erasmus, S. Paolucci, J. Ruelle, D. Pieniazek, N. Taveira, A. Trevino, M. F. Goncalves, S. Jallow, L. Xu, R. J. Camacho, V. Soriano, P. Goubau, J. D. de Sousa, A. M. Vandamme, M. A. Suchard, and P. Lemey. Phylogeographical footprint of colonial history in the global dispersal of human immunodeficiency virus type 2 group A. *J. Gen. Virol.*, 93(Pt 4):889–899, Apr 2012.

[4] N. R. Faria, M. A. Suchard, A. Abecasis, J. D. Sousa, N. Ndembi, I. Bonfim, R. J. Camacho, A. M. Vandamme, and P. Lemey. Phylodynamics of the HIV-1 CRF02_AG clade in Cameroon. *Infect. Genet. Evol.*, 12(2):453–460, Mar 2012.

[5] N. R. Faria, M. A. Suchard, A. Rambaut, and P. Lemey. Toward a quantitative understanding of viral phylogeography. *Curr Opin Virol*, 1(5):423–429, Nov 2011.

[6] P. Lemey, A. Rambaut, A. J. Drummond, and M. A. Suchard. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.*, 5(9):e1000520, Sep 2009.

[7] M. I. Nelson, P. Lemey, Y. Tan, A. Vincent, T. T. Lam, S. Detmer, C. Viboud, M. A. Suchard, A. Rambaut, E. C. Holmes, and M. Gramer. Spatial dynamics of human-origin H1 influenza A virus in North American swine. *PLoS Pathog.*, 7(6):e1002077, Jun 2011.