Journal of Theoretical Biology 420 (2017) 144–151

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Full reconstruction of non-stationary strand-symmetric models on rooted phylogenies

Benjamin D. Kaehler

Research School of Biology, Australian National University, Canberra, Australian Capital Territory, Australia

ARTICLE INFO

Keywords:
Identifiability
Phylogenetic inference
Markov process
Maximum likelihood

ABSTRACT

Understanding the evolutionary relationship among species is of fundamental importance to the biological sciences. The location of the root in any phylogenetic tree is critical as it gives an order to evolutionary events. None of the popular models of nucleotide evolution currently used in likelihood or Bayesian methods are able to infer the location of the root without exogenous information. It is known that the most general Markov models of nucleotide substitution also cannot identify the location of the root or be fitted to multiple sequence alignments with fewer than three sequences. We prove that the location of the root and the full model can be identified and statistically consistently estimated for a non-stationary, strand-symmetric substitution model given a multiple sequence alignment with two or more sequences. We also generalise earlier work to provide a practical means of overcoming the computationally intractable problem of labelling hidden states in a phylogenetic model.

1. Introduction

The location of the root in a molecular phylogeny has contributed to criminal convictions (González-Candelas et al., 2013), been used to understand the source and epidemiology of human viruses (Podsiadlo and Polz-Dacewicz, 2013), determined how biodiversity conservation resources were distributed (Faith and Baker, 2006), been used to develop potential HIV vaccines (Nickle et al., 2003), and played an important role in our understanding of the tree of life (Murphy et al., 2007). While an unrooted phylogenetic tree can be used to infer relatedness between species, without the location of the root the order of evolutionary events is open to conjecture. It might then be surprising that none of the commonly used Markov models of nucleotide or codon substitution can be used to identify the location of the root without incorporating information that is exogenous to the model. The class of Markov models to which we refer will be made precise in the next section.

ModelTest is one of the most popular pieces of software for selecting phylogenetic models of character substitution (Posada, 2008). It allows users to determine which of the 88 time-reversible (hereafter *reversible*) substitution models best fits their data. By definition, for a reversible model the location of the root in a phylogeny cannot change the probability distribution of the observed data, the column frequencies. Some software (Knight et al., 2007) allows users to fit non-stationary models of character substitution such as that of Barry and Hartigan (1987a). Unfortunately, the theoretical results that exist around fully general models (Chang, 1996) explicitly state that for such models the location of the root is not statistically identifiable, that one cannot use such models to ask where on an edge a root resides, and that it is possible to reformulate the model so that any node is the root.

Before continuing it is worth clarifying the relationship between non-stationarity and reversibility of Markov processes. A process is stationary if the distribution of states does not change through time and non-stationary otherwise. A stationary process is reversible if the joint distribution of the states of the process taken at any two time points does not depend on the order of the points in time. It is straightforward to show that a non-stationary process cannot be reversible. A review of these properties that is relevant to the current context can be found in Jermiin et al. (2008).

In practice, the location of the root is usually determined by declaring that a specific taxon in a phylogeny is an *outgroup* or by making a *molecular clock* assumption (Felsenstein, 2004). The first method assumes that the location of the root is already known, that it is on the edge connected to the outgroup. The second method comes in varying degrees of complexity. In its most simple form it assumes that the tree is *ultrametric*, that the genetic distance from the root node to each tip is identical. In more sophisticated Bayesian approaches the location of the root enters the calculation as part of the prior distribution of tree topologies and branch lengths, so that the tree is not necessarily ultrametric but that in some sense the evolutionary time from the root of the tree to the tips is the same along every lineage (Drummond and Rambaut, 2007).

E-mail address: benjamin.kaehler@anu.edu.au.

http://dx.doi.org/10.1016/j.jtbi.2017.03.007
Received 14 November 2016; Received in revised form 6 March 2017; Accepted 8 March 2017
Available online 09 March 2017

A third method is to use a substitution model that is able to identify the location of the root. Such a model must not be reversible, but using a model that is not reversible does not automatically ensure that the model is able to identify the root. This statement is easily justified using the findings in Chang (1996), where a model that is non-stationary, so also non-reversible, is not able to recover the root. That a non-reversible model might not be even theoretically able to discover the root seems to have been missed by some authors.

Yang and Roberts (1995) fitted a non-stationary model to rooted topologies of real data using maximum likelihood and found that the location of the root of the tree had a significant effect on likelihood estimates. This is useful empirical evidence but the authors made no attempt to prove that their model is identifiable.

Huelsenbeck et al. (2002) fitted a non-reversible but stationary model to real and simulated data and found that while the outgroup and molecular clock methods were able to recover the location of the root in many cases, their model was not. Again, they made no effort to show that their model is theoretically capable of recovering the location of the root, so the poor performance of their model is not necessarily a reflection on the ability of all substitution models to recover the location of the root.

Yap and Speed (2005) systematically reproduced the results in Yang and Roberts (1995) and Huelsenbeck et al. (2002) and, with some small discrepancies with the earlier studies, found again that a non-stationary model was able to make statistically significant inferences about the location of the root, but that a non-reversible model did little better than a reversible model. This also was empirical research that left unanswered questions about whether any of the models were theoretically able to identify the location of the root.

The contribution of the present work is to constructively prove that there is a non-stationary substitution process that identifies the location of the root that can be statistically consistently estimated from data. Indeed, the model is shown to be consistently estimable for two taxa. This is not possible for general non-stationary processes (Chang, 1996; Bonhomme et al., 2014), so we make the additional assumption that the process is *strand-symmetric*; that the process of evolution is identical on the sense and antisense strands of DNA. That is, the rate of substitution between two nucleotides is the same as the rate of substitution between their Watson–Crick base pair complements. For instance, the rates for A → G and T → C must be equal as they are strand complementary substitutions. The conditions of the proof imply that the process is non-stationary, and we show that a non-reversible, stationary model is not identifiable so cannot be consistently estimated. This observation sheds some light on the success of non-stationary processes and the failure of non-reversible, stationary processes at detecting the root in the literature.

Much has been written about the biological mechanisms that result in nucleotide substitution processes being strand-asymmetric and there is now substantial empirical evidence to support strand asymmetry's existence in nature. Touchon and Rocha (2008) provide a good review of the subject. Strand asymmetry seems to be a localised phenomenon, existing on the scale of genes rather than genomes, and appears to be common in prokaryote and organelle genomes but not in eukaryotes. Nucleotide compositional asymmetry is the most common measure used for statistical inference. Under very loose assumptions (Lobry, 1995; Lobry and Lobry, 1999) a strand-symmetric process should result in the proportions of As and Ts being equal and the proportions of Gs and Cs being equal on a single strand. Strand asymmetry can also be inferred by directly comparing estimated rates of nucleotide substitution, although most of the evidence seems to come from counting substitutions in ancestral state reconstructions based on maximum parsimony (e.g. Wu and Maeda, 1987; Bulmer, 1991; Francino and Ochman, 2000).

Strand-symmetric models have been used in a maximum likelihood context, although rarely for the purpose of establishing whether strand symmetry is a reasonable assumption. Yap and Pachter (2004) comment that the reversible models that they fitted to real data seemed to exhibit strand symmetry. Squartini and Arndt (2008) fitted a continuous-time, non-stationary, strand-symmetric model on a known, rooted, four-taxon topology to two genome-scale data sets. They comment that their model is not identifiable on the edges incident to the root, but that it is identifiable on the other two edges. This is not strictly correct. As stated in Chang (1996, Remark 2), the labelling of states at the internal nodes is not automatically identifiable even for a continuous-time model. Also, as the mapping from the discrete-time process considered in Chang (1996) to the continuous-time process fitted in Squartini and Arndt (2008) is not always unique (Higham, 2008, Section 2.3), continuous-time models are not identifiable under the results in Chang (1996) without further constraints.

Strand-symmetric substitution models have also been approached from a theoretical perspective (Casanellas and Sullivant, 2005; Jarvis and Sumner, 2016). Jarvis and Sumner (2016) make the observation that strand-symmetric models enjoy the property of *closure*; that a model which is strand-symmetric on two adjacent phylogenetic branches is strand-symmetric across the two branches as well.

The proofs in this work mirror those in Chang (1996), but apart from adding the assumption of strand symmetry and removing the assumption of an unrooted tree, we also relax an important assumption that Chang (1996) makes about the structure of the model. As noted in Zou et al. (2011) and addressed in Mossel and Roch (2006) the model of Barry and Hartigan (1987a) is only identifiable up to an arbitrary relabelling of states at internal nodes of the phylogeny. Chang (1996) addresses this problem by assuming that the transition probability matrix for every edge in the topology is *reconstructible from rows*. As we shall demonstrate, this assumption can be restrictive in practice, so, motivated by Remark 4 in Chang (1996), we partially relax it.

In Section 2 we briefly introduce the necessary notation and theoretical context. Section 3 contains the main results of the paper, where we state that the full topology and parameters of a discrete-time, non-stationary, strand-symmetric Markov model can be recovered from the joint probability distributions of states between pairs of extant taxa. In this section we also extend the result to continuous-time models which are used more commonly in practice. Section 4 reveals that the results in Section 3 provide the necessary basis for consistent statistical estimation of the models in question for multiple sequence alignments of increasing length. Section 5 gives some concluding remarks. The proofs are relegated to the Appendix.

2. Markov models on trees: definitions and notation

We consider a finite set of extant taxa T whose phylogenetic history we wish to infer. The history is modelled as a tree. The tree consists of a set of nodes S that correspond to extant and ancestral species and edges E that indicate direct genetic descent between species. The edges are a set of unordered pairs of nodes, so if $\{r, s\} \in E$, an edge exists between nodes r and s. The nodes consist of the *terminal nodes*, which are just T, and the *internal nodes* N, so that $N = S \backslash T$. The internal nodes represent ancestral taxa at branching points.

The *degree* of a node is the number of edges incident on that node. Terminal nodes have degree one. We say that a tree is *unrooted* if it contains no internal nodes with degree two. A tree is *rooted* if it contains a single node with degree two, which we call the *root*. We do not consider trees with more than one internal node of degree two.

The model of character substitution is properly considered a probabilistic graphical model defined on the tree. That is, we associate a random variable $X_s \in C$ with each node $s \in S$ so that $\{X_s\}_{s \in S}$ represents the history of a single column in a multiple sequence alignment. Each X_s is independent of all other X_r, conditional on the states at the nodes neighbouring s. As we will focus on the assumption of strand symmetry, we will assume that $C = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. We also assume that each column in the alignment is an independent observation of the multivariate random variable $\{X_s\}_{s \in T}$.