Journal of Theoretical Biology 407 (2016) 362–370

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Reversible polymorphism-aware phylogenetic models and their application to tree inference

Dominik Schrempf [a,b], Bui Quang Minh [c], Nicola De Maio [d,e], Arndt von Haeseler [c], Carolin Kosiol [a,*]

[a] Institut für Populationsgenetik, Vetmeduni Vienna, Wien, Austria
[b] Vienna Graduate School of Population Genetics, Wien, Austria
[c] Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, Austria
[d] Nuffield Department of Medicine, University of Oxford, UK
[e] Oxford Martin School, University of Oxford, UK

HIGHLIGHTS

- Species tree inference from genome-wide population data.
- Takes incomplete lineage sorting into account.
- Analytical solution of stationary distribution and formal proof of reversibility.
- Reversibility ensures swiftness and stability.
- Increase of sample size per species improves estimations without raising runtime.
- Comparison to the Wright-Fisher diffusion.

ARTICLE INFO

Article history:
Received 11 April 2016
Received in revised form
25 July 2016
Accepted 27 July 2016
Available online 29 July 2016

Keywords:
Species tree
Phylogenetics
Incomplete lineage sorting
Substitution model
Reversible polymorphism-aware phylogenetic model

ABSTRACT

We present a reversible Polymorphism-Aware Phylogenetic Model (revPoMo) for species tree estimation from genome-wide data. revPoMo enables the reconstruction of large scale species trees for many within-species samples. It expands the alphabet of DNA substitution models to include polymorphic states, thereby, naturally accounting for incomplete lineage sorting. We implemented revPoMo in the maximum likelihood software IQ-TREE. A simulation study and an application to great apes data show that the runtimes of our approach and standard substitution models are comparable but that revPoMo has much better accuracy in estimating trees, divergence times and mutation rates. The advantage of revPoMo is that an increase of sample size per species improves estimations but does not increase runtime. Therefore, revPoMo is a valuable tool with several applications, from speciation dating to species tree reconstruction.

1. Introduction

Molecular phylogenetics seeks to understand evolutionary phenomena such as speciation dynamics and biodiversity by estimating evolutionary parameters at the species level. The reconstruction of the species history gives insights into the basic mechanisms of biology. However, the topology of the species tree is not always clear, especially when phylogenies from different genomic regions (i.e., gene trees or genealogies) differ from each other (Degnan and Rosenberg, 2006).

Statistical approaches to tree reconstruction such as maximum likelihood and Bayesian methods rely on substitution models (Tavaré, 1986). These models describe and quantify the probabilities of how sequences may evolve along a phylogeny. They are defined by an instantaneous rate matrix $\boldsymbol{Q}$ that contains the substitution rates between the different character states. For computational convenience, most substitution models are *reversible*. That is, the process describing the evolution of the sequence is independent of the direction in time. Reversibility is important in phylogenetics for tree inference from large data sets with many species because it simplifies the likelihood function (Yang, 2006, p. 34) and reduces the number of trees by a factor of $2l - 3$, where l is the number of tips of the tree (Hein et al., 2004, p.

* Corresponding author.
E-mail addresses: dominik.schrempf@vetmeduni.ac.at (D. Schrempf), carolin.kosiol@vetmeduni.ac.at (C. Kosiol).

http://dx.doi.org/10.1016/j.jtbi.2016.07.042

70). Finally, rate matrices of reversible substitution models have real eigenvalues (Kelly, 1979) which enable a fast and stable eigendecomposition during matrix exponentiation (Golub and Loan, 1996). Many software packages use reversible substitution models (e.g., HyPhy, Pond et al., 2005; PhyML, Guindon et al., 2010 and MrBayes, Ronquist et al., 2012). RAxML (Stamatakis, 2014) and IQ-TREE (Nguyen et al., 2015) additionally offer efficient tree search algorithms for very large phylogenies.

Substitution models, when naively applied to species trees (concatenation methods, e.g., Gadagkar et al., 2005), assume the species or population to be fixed for a specific character state and do not account for effects on the population genetics level such as Incomplete Lineage Sorting (ILS; Maddison, 1997; Knowles, 2009). Incompletely sorted lineages coalesce deep in the tree and their coalescent events do not match the speciation events. The probability of ILS is large and consequently tree reconstruction is difficult if the time between speciation events is short or if the effective population size is large (Pamilo and Nei, 1988). The multispecies coalescent model can be used to quantify the phylogenetic distortion due to ILS. It simulates a coalescent process (Kingman, 1982) on each branch of the species tree and combines these separate processes when branches join together. This model predicts that for specific evolutionary histories the gene trees with highest abundance conflict the species tree topology (anomaly zone; Degnan and Rosenberg, 2009; Degnan, 2013). These are extreme cases where common tree inference methods not accounting for ILS such as concatenation (Gadagkar et al., 2005) or democratic vote (Pamilo and Nei, 1988) fail because they are statistically inconsistent (e.g., Ewing et al., 2008). However, ILS considerably deteriorates estimates already when species trees are not in the anomaly zone (Pollard et al., 2006).

We have recently developed an approach called **Po**lymorphism-Aware Phylogenetic **Mo**del (PoMo, De Maio et al., 2013). PoMo builds on top of substitution models but makes use of within-species data and considers present and ancestral polymorphisms thereby accounting for ILS. Similar to multispecies coalescent models it uses multiple sequence alignments of up to several hundred species while allowing for many within-species sequences to infer base composition and mutational parameters. Recently, we applied PoMo to infer species trees (De Maio et al., 2015). We showed in a large scale simulation study with various demographic scenarios and evaluation against other state-of-the-art methods like BEST (Liu, 2008),*BEAST (Heled and Drummond, 2010), SNAPP (Bryant and et al., 2012) and STEM (Kubatko et al., 2009) that PoMo is approximately as fast as standard DNA substitution models while being more accurate in terms of the branch score distance (Section 3.1). Furthermore, application to great apes data leads to phylogenies consistent with previous literature and also with the geographic distribution of the populations.

Here, we prove the reversibility of PoMo when an associated reversible mutation model (Section 2.3) is used and derive the corresponding stationary distribution. This will open the PoMo approach to a new area of applications because a reversible model can take advantage of existing algorithms that efficiently reconcile the species tree. We will discuss the reversible solution of PoMo, provide connections to the diffusion equation and introduce an implementation in IQ-TREE (Nguyen et al., 2015).

Finally, we present a simulation study and an application to real data to demonstrate the performance of the reversible PoMo (revPoMo) and to confirm its relevancy in medium-to-large-scale tree search.

2. Materials and methods

2.1. DNA substitution models

DNA substitution models assume that a DNA sequence evolves as a series of independent substitution events which replace a nucleotide by another one. Substitutions are modeled as a time-continuous, time-homogeneous Markov process (Yang, 1994). Additionally, the different sites of a sequence are assumed to evolve independently. The four nucleotides A, C, G and T form the alphabet $\mathcal{A}$. The rates of change q_{xy} from nucleotide x to nucleotide y are summarized in an instantaneous rate matrix $\mathbf{Q} = (q_{xy})_{x,y\in\mathcal{A}}$ which completely describes the time-continuous Markov process. The assumption of time-homogeneity implies that the entries of $\mathbf{Q}$ are constant in time. One also assumes stationarity, i.e., the existence of a stationary distribution $\boldsymbol{\pi} = (\pi_x)_{x\in\mathcal{A}}$ which is the solution to $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$. If the Markov process is reversible, then detailed balance $\pi_x q_{xy} = \pi_y q_{yx}$ is fulfilled. Thus, for the General Time Reversible (GTR, Tavaré, 1984) model the rate matrix has the following structure:

$$\mathbf{Q} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} * & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ r_{CA}\pi_A & * & r_{CG}\pi_G & r_{CT}\pi_T \\ r_{GA}\pi_A & r_{GC}\pi_C & * & r_{GT}\pi_T \\ r_{TA}\pi_A & r_{TC}\pi_C & r_{TG}\pi_G & * \end{pmatrix} \end{array}, \tag{1}$$

with $q_{xy} = \pi_y r_{xy}$ and exchangeabilities $r_{xy} = r_{yx} > 0$. The diagonal entries are chosen such that the row sums are zero. The expected number of events on a branch of length d is $\mathbb{E}(d) = -d\sum_x \pi_x q_{xx}$. Usually, $\mathbf{Q}$ is normalized such that $\sum_x \sum_{x\neq y} \pi_x q_{xy} = 1$ or $\mathbb{E}(1) = 1$.

2.2. The alphabet of revPoMo

Standard DNA substitution models are limited in the sense that they assume that species are always fixed for a specific nucleotide (i.e., the changes are substitutions). For revPoMo, we use standard DNA models such as HKY (Hasegawa et al., 1985) or GTR (Tavaré, 1984) as mutation models introducing variation into populations that are no longer assumed to be fixed for one nucleotide. We expand the alphabet to include characters that represent polymorphisms so that populations can have polymorphic states. Thereby, revPoMo introduces a virtual haploid population of constant size N and distinguishes between fixed (*boundary*) $\{Nx\} = \{Nx, 0y\} = \{0y, Nx\}$ and polymorphic characters $\{ix, (N-i)y\}$ $(1 \leq i \leq N-1;\ x, y \in \{A, C, G, T\};\ x \neq y)$, where x and y are the nucleotides of the associated mutation model (Fig. 1). For convenience, we call the set of boundary characters the *boundary*. To keep the alphabet of revPoMo $\mathcal{A}_{PoMo}$ manageable, we assume that at most two different nucleotides per site are present simultaneously. This is only a mild restriction and many real data sets meet this assumption. For example, no sites with three or four nucleotides have been found in the great apes data set described in Section 2.11. This restriction also agrees with the chosen mutation model (Section 2.3). The alphabet-size of revPoMo is

$$|\mathcal{A}_{PoMo}| = 4 + \binom{4}{2}(N-1). \tag{2}$$

To differentiate between revPoMo and the associated mutation model, we refer to the characters of the mutation model as nucleotides and to the characters of revPoMo as *states*. The instantaneous rate matrix of revPoMo $\mathbf{Q}_{revPoMo}$ is composed of the rates of mutations and genetic drift

$$\mathbf{Q}_{revPoMo} = \mathbf{Q}_{Mut} + \mathbf{Q}_{Drift}, \tag{3}$$