Letter to Editor

Consistency and identifiability of the polymorphism-aware phylogenetic models

1. Introduction

Phylogenies help to answer a multitude of questions regarding the origin of species, the tempo of evolution, the origin of particular traits and the processes (either neutral or selective) of molecular evolution (e.g. see Pagel, 1999). Phylogenies are therefore central to the study of patterns and processes by which evolution happens. However, phylogenies can only serve these purposes if they are correctly estimated. Fortunately, mathematical phylogenetics provides criteria that help us to assess whether a given phylogeny estimator is statistically sound. Statistical consistency and identifiability are two examples of such criteria.

In phylogenetic theory, a phylogenetic reconstruction method $\hat{\tau}$ is statistically consistent under a model if it converges in probability to the true tree τ as the number of sites S of the sequence alignment increases indefinitely (Wald, 1949; Felsenstein, 1973), i.e.

$$\lim_{S \to \infty} p(|\hat{\tau}_S \to \tau|) = 1 \quad . \tag{1}$$

Identifiability is met if the model of evolution is uniquely characterized by the probability distribution it defines (Chang and Hartigan, 1991). An identifiable model is a necessary condition for consistency. More formal conditions for identifiability and consistency are described in Steel (1994), Chang (1996) and Steel (2013); these are revisited later in this article. Lack of statistical consistency has long been an aspect that phylogeneticists cared for. For example, some maximum parsimony methods were criticized for lacking statistical consistency early on (Felsenstein, 1978).

Simple phylogeny estimation methods using standard substitution model (e.g. Bayesian or maximum likelihood inference under the Jukes and Cantor, 1969; Kimura, 1980; Tajima and Nei, 1984 substitution models) can be shown to enjoy statistical consistency (Chang, 1996). However, the same principle cannot be directly extended to more complex and general methods of tree inference that include rate variation across sites (Γ) and invariant sites (I). For such models of evolution, the model distribution is not fully understood, which complicates proving identifiability. Identifiability has been proven for the pure general time-reversible model (GTR, Tavaré, 1986). The GTR with rate variation (i.e. GTR+Γ) (Wu and Susko, 2010) also enjoys identifiability; however, a rigorous proof for the commonly utilized GTR+Γ+I is still lacking (Rogers, 2001; Allman et al., 2008; Chai and Housworth, 2011).

During the last years, alternative approaches to the classic nucleotide substitution models have been proposed. The polymorphism-aware phylogenetic models (PoMo) are such an example of an alternative approach for tree estimation that also accounts for incomplete lineage sorting (De Maio et al., 2013). While PoMo can be more broadly classified as a nucleotide substitution model; it adds a new layer of complexity by accounting for the population-level evolutionary processes (such as mutations, genetic drift, and selection) to describe the evolutionary process (De Maio et al., 2015; Schrempf et al., 2016; Borges et al., 2019). To do so, PoMo builds on a GTR-like mutation scheme and expands the {A, C, G, T} state-space to include polymorphic states, thereby accounting for current and ancestral intra-population variation. The latter aspect sets PoMo apart from the classic models of evolution, which traditionally only use a single representative DNA sequence per species.

PoMo has received substantial attention from the evolutionary community (see Mirarab et al., 2014; Szöllsi et al., 2015; Leaché and Oaks, 2017). Several publications have employed it to solve a wide range of evolutionary questions, e.g., disentangling phylogenetic relationships among baboon species (Rogers et al., 2019), describing the phylogeographic history of great apes (Schrempf et al., 2016), estimating patterns of GC-bias and mutational biases (De Maio et al., 2013; Borges et al., 2019) and inferring the site-frequency spectrum (Schrempf and Hobolth, 2017; Borges et al., 2019) from population data.

All this raised the question of whether PoMo is a statistically consistent phylogeny estimator for phylogenetic data sets. Building upon the formal results provided by Steel (2013) and identifiability of the PoMo rate matrix and stationary distribution, we present a proof that the MAP topology (i.e. the tree topology that has the greatest posterior probability) is a statistically consistent estimate of the true tree. This result shows that PoMo is a statistically sound method of phylogenetic inference, and it provides validity for further investigations and uses of PoMo methods on real data sets.

2. Polymorphism-aware phylogenetic models in a nutshell

PoMo assumes a Moran model (Moran, 1958) with N haploid individuals and defines the allele trajectory of a single locus with four possible alleles a_i, where $i \in \mathcal{A} = \{A, C, G, T\}$ (Table 1 includes a glossary of the symbols used in the present proof). The evolution of this population in the course of time is described by a continuous time Markov chain with a discrete state-space defined by N and the four alleles. States are monomorphic (or boundary) if all the N individuals have the allele i {Na_i}; or polymorphic, if two alleles are present in the population {$na_i, (N-n)a_j$} (Fig. 1).

A rate matrix $\mathbf{Q}$ describing the dynamic between the boundary and polymorphic states can be defined by considering several population-level processes. So far, PoMo includes mutation, genetic drift and allelic selection (Fig. 1) (De Maio et al., 2015; Schrempf et al., 2016; Borges et al., 2019).

Table 1
Glossary. The order of the symbols follows their first occurrence in the text.

| Symbol	Description
τ	tree topology
$\hat{\tau}$	tree topology estimator
S	number of sites
a_i	allele i
$\mathcal{A}$	nucleotide base {A, C, G, T}
N	population size
n	absolute frequency of an allele; relative frequency would be n/N
$\mathbf{Q}$	PoMo instantaneous rate matrix with elements q
π	component of the mutation rate; stationary frequencies in the GTR model of evolution
ρ	component of the mutation rate; base exchangeabilities in the GTR model of evolution
σ	allelic selection coefficients
k	normalization constant of the stationary vector
α	PoMo state
K	number of alleles; $K = 4$ for PoMo
L	number of taxa/sequences
$\mathbf{P}$	transition matrix
χ	site pattern or, better, PoMo frequency patterns
η	root vertex
$\epsilon = (v_1, v_2)$	directed edge
ω	branch lengths
θ	set of PoMo parameters and branch lengths
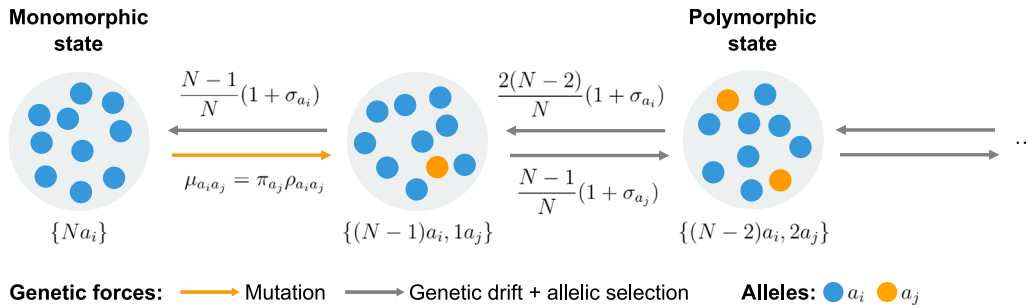
Fig. 1. PoMo state-space and transition rates. The two alleles represent any of the four nucleotide bases A, C, G, and T. Orange and grey distinguish the role of mutation and genetic drift plus selection, respectively. The PoMo state-space includes monomorphic (or boundary states) $\{Na_i\}$ and polymorphic states $\{na_i, (N-n)a_j\}$. Monomorphic states interact with polymorphic states via mutation and polymorphic states reach monomorphic states via drift or selection. Between polymorphic states only drift and selection occur. PoMo is thus a particular case of the multivariate Moran model with boundary mutations and selection when four alleles (i.e. the four nucleotide bases) are considered.

- Mutations are assumed to occur only in the boundary states $\{Na_i\}$ with mutation rates $\mu_{a_i a_j} = \rho_{a_i a_j} \pi_{a_j} = \rho_{a_j a_i} \pi_{a_j}$. Mutation rates are thus modeled according to a GTR model of evolution (Tavaré, 1986), where π represents the stationary frequencies and ρ the six exchangeability terms. Similar interpretations of these parameters are still valid for the neutral PoMo, as π immediately informs the stationary frequencies of the monomorphic states. However, the stationary frequencies of the monomorphic states are no longer only defined by π in the more general PoMo with allelic selection (Borges et al., 2019) (Eq. (3)). All in all, the GTR-like mutation scheme is a convenient strategy to obtain quantities of interest for PoMo.
- Genetic drift rules the allele frequency changes in a population. Genetic drift is modeled according to the Moran model, in which one individual is chosen to reproduce (i.e. to copy itself) and one to die in each time step (Moran, 1958). Therefore, the rates by which an allele with a starting frequency n/N is born or dies (i.e. the allele increases or decreases by one) are the same and equal to $\frac{n(N-n)}{N}$. The allele frequency changes are thus neutral.
- Allelic selection may favor one allele over the other by differentiated reproductive success. The Moran machinery described previously can be adapted to model relative fitnesses by permitting that a given allele a_i has a fitness advantage/disadvantage $(1 + \sigma_{a_i})$ over the others (Durrett, 2008;

Borges et al., 2019). σ refers to the vector of relative selection coefficients: i.e. a reference allele is chosen to have fitness 1, or selection coefficient 0.

Taking into account the population processes described so far, the PoMo instantaneous rate matrix $\mathbf{Q}$ is

$$q(\{n_1 a_i, (N - n_1)a_j\}, \{n_2 a_i, (N - n_2)a_j\})$$
$$= \begin{cases} \mu_{a_i a_j} = \rho_{a_i a_j} \pi_{a_j} & n_1 = N, \ n_2 = N - 1 \\ \mu_{a_j a_i} = \rho_{a_i a_j} \pi_{a_i} & n_1 = 0, \ n_2 = 1 \\ \frac{n_1}{N}(N - n_1)(1 + \sigma_{a_i}) & n_2 = n_1 + 1, \ 0 < n_1 < N \\ \frac{n_1}{N}(N - n_1)(1 + \sigma_{a_j}) & n_2 = n_1 - 1, \ 0 < n_1 < N \\ 0 & |n_1 - n_2| > 1 \end{cases}, \quad (2)$$

where n_1 and n_2 represent a shift in the allele frequencies. Frequency shifts larger than 1 are disallowed (last condition in Eq. (2)) making PoMo rate matrices typically sparse. The diagonal elements are defined such that the respective row sum is 0. The stationary distribution of PoMo is obtained by satisfying the condition $\psi \mathbf{Q} = \mathbf{0}$. ψ is the normalized stationary vector and has the solution

$$\psi(\alpha) = \begin{cases} \pi_{a_i}(1 + \sigma_{a_i})^{N-1} k^{-1} & \text{if } \alpha = \{Na_i\} \\ \pi_{a_i} \pi_{a_j} \rho_{a_i a_j}(1 + \sigma_{a_i})^{n-1} & \text{if } \alpha = \{na_i, (N-n)a_j\} \\ (1 + \sigma_{a_j})^{N-n-1} \frac{N}{n(N-n)} k^{-1} \end{cases}, \quad (3)$$

where k is the normalization constant and α a PoMo state (Borges et al., 2019).

When using PoMo to infer phylogenies, we make use of two important assumptions that are convenient for the proof of consistency presented here. Our first assumption is that sites evolve independently. Thus the probability that sequence A evolves to sequence B equals the product of the probability of evolutionary paths across all sites S. As a result, the site patterns created by the L species are independently and identically distributed (i.i.d.). The second assumption is that the evolutionary process is stationary and reversible with equilibrium measure $\boldsymbol{\psi}$. Proof of reversibility and stationarity have been provided by Schrempf et al. (2016) for the neutral PoMo model and by Borges et al. (2019) for the model with allelic selection.

3. Identifiability of PoMo

Identifiability is the inverse problem of finding the tree τ and the transition matrix $\boldsymbol{P}$ given just the probability of the various site patterns $\boldsymbol{\chi}$ (or frequency patterns, in the case of PoMo) (Steel et al., 1998). Identifiability comes from the restrictions that must be placed on $\boldsymbol{P}$ for τ to be uniquely described by the probability of generating a pattern $\boldsymbol{\chi}$. These restrictions have been extensively studied (Chang and Hartigan, 1991; Steel, 1994; Chang, 1996; Steel et al., 1998).

Let us assume a tree τ with η representing the most recent common ancestor of L species (i.e. the root of the tree). The edges $\epsilon = (\nu_1, \nu_2)$ of τ are directed away from the root η in such a way that ν_1 lies between η and ν_2. If we attribute states to each vertex in the tree τ, beginning from the root η to all the descending vertices, we can represent the probability of generating pattern $\boldsymbol{\chi}$ as

$$f_{\boldsymbol{\chi}} = \sum_{\boldsymbol{\chi}'} p(\chi_\eta) \prod_{\epsilon=(\nu_1,\nu_2)} P(\chi_{\nu_1}, \chi_{\nu_2}) \quad, \tag{4}$$

where $\boldsymbol{\chi}'$ extends $\boldsymbol{\chi}$ and represents the total assignment of states, and $p(\chi_\eta)$ is the probability of state χ_η at the root. The alphabet of PoMo has $4 + 6(N-1)$ states and thus there are $[4 + 6(N-1)]^L$ possible frequency patterns $\boldsymbol{\chi}$ in a tree with L leaves.

As PoMo makes some assumptions regarding the evolutionary process (Schrempf et al., 2016), we can further simplify Eq. (4): (i) frequency changes on edges are described by a continuous-time Markov process; (ii) the PoMo rate matrix $\boldsymbol{Q}$ is the same for all edges of the tree; and (iii) the distribution of frequency states at the root $p(\chi_\eta)$ is simply the equilibrium distribution $\boldsymbol{\psi}$.

$$f_{\boldsymbol{\chi}} = \sum_{\boldsymbol{\chi}'} \psi(\chi_\eta) \prod_{\epsilon=(\nu_1,\nu_2)} \exp\{t_\epsilon \boldsymbol{Q}\}(\chi_{\nu_1}, \chi_{\nu_2}) \quad, \tag{5}$$

where t_ϵ is the length of edge ϵ. Conditions represented by (4) and (5) are also known as the Markov property, which is a necessary though not sufficient condition for identifiability.

Following (Steel, 1994; Steel et al., 1998), another condition for identifiability additional to (5) needs to hold: $\det(\boldsymbol{P}) \neq 0, 1, -1$ for all edges ϵ in the tree and $\psi(\alpha) \neq 0$ for each PoMo state α. Then the tree τ can be uniquely recovered from the frequency patterns $\boldsymbol{\chi}$ (Steel, 1994; Steel et al., 1998). By Jacobi's formula (theorem 2.12 in Hall, 2015), we can write that $\det(\boldsymbol{P}) = \exp\{\text{tr } \boldsymbol{Q}\}$ and therefore

$$\det(\boldsymbol{P}) = \exp\{- \sum_{\substack{a_i a_j \in \mathcal{A} \\ a_i \neq a_j}} \rho_{a_i a_j}(\pi_{a_i} + \pi_{a_j}) - \frac{1}{2}(N-1)(N+1)$$

$$(4 + \sum_{a_i \in \mathcal{A}} \sigma_{a_i})\} \tag{6}$$

The selective pressures and the mutation rates, as modeled in PoMo, can only be real positive numbers. These rates cannot be 0 because they would represent very unlikely and biologically unreasonable scenarios: If $1 + \sigma_{a_i} = 0$ the individual carrying allele a_i would immediately die (i.e. an extremely deleterious allele); if $\pi_{a_i} = 0$ or $\rho_{a_i a_j} = 0$ (both imply $\mu_{a_i} = 0$) the allele a_i does not arise by mutation. In both situations, such allele should not be observed at all in the population, and one could simply use PoMo with $K - 1$ alleles, where K is the number of alleles in the population. Therefore, we can easily conclude that $0 < \det(\boldsymbol{P}_\epsilon) < 1$ for all edges ϵ of τ.

Because we assume the equilibrium distribution of allele frequencies in the root, we need to check whether any elements of $\boldsymbol{\psi}$ can be equal to 0. The PoMo stationary distribution is defined for two different types of states α, the monomorphic (or fixed) and the polymorphic ones. As shown in Eq. (3), these states can only have 0 probability if any of the population parameters are 0. Therefore, we conclude that $\psi(\alpha) > 0$ for all PoMo states α. Summarizing, PoMo models respect the conditions for identifiability and we conclude that PoMo trees are uniquely identifiable by the frequency patterns they induce. Identifiability can be extended to the multivariate case by defining an alphabet $\mathcal{A}$ over K alleles and a state-space of $K + K\frac{K-1}{2}(N-1)$ states. For simplicity we have considered the four-variate case, but the proofs shown here remain valid for the multivariate Moran model with allelic selection (Borges et al., 2019).

4. Consistency of the Bayesian PoMo

To evaluate the consistency property under PoMo, we have to prove that a given tree estimator converges in probability to the true tree as the number of sites S increases indefinitely. It has already been formally shown that the MAP tree (i.e. the tree topology that has the greatest posterior probability), estimated in a Bayesian framework, provides a statistically consistent estimator of the true tree (Steel, 2013). Consistency was proven under a wide variety of conditions (Steel, 2013), including tree inference from aligned sequences across the entire parameter range, and with the usage of general priors in models where the identifiability condition holds. Here, we show that these conditions are also met under PoMo.

Suppose we are given i.i.d. site patterns $\boldsymbol{\chi} = (\chi_1, \ldots, \chi_S)$ generated by some unknown parameters $(\tau, \boldsymbol{\theta})$ and we wish to identify the topology τ from $\boldsymbol{\chi}$ given prior densities on the set of fully resolved trees T and the continuous parameters $\boldsymbol{\theta}$. These parameters include the branch lengths $\boldsymbol{\omega}$, the mutations rates (defined by $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$) and the selection coefficients $\boldsymbol{\sigma}$. Suppose we have a discrete prior probability distribution $p(\tau)$ on T, and, for each $\tau \in T$, a continuous prior probability distribution on $\Theta(\tau)$ with a probability density function $p_\tau(\boldsymbol{\theta})$.

In particular, if the following four conditions hold:

- C1: $p(\tau) > 0$;
- C2: the density $p_\tau(\boldsymbol{\theta})$ is continuous, bounded and nonzero on $\Theta(\tau)$;
- C3: the function $\boldsymbol{\theta} \to p_{(\tau,\boldsymbol{\theta})}(\boldsymbol{\chi})$ is continuous and nonzero on $\Theta(\tau)$;
- C4: identifiability, guaranteeing that τ is uniquely identifiable by $\boldsymbol{\chi}$.

Steel (2013) has shown that

$$\lim_{S\to\infty} p(\tau^*, \boldsymbol{\theta}, S|\boldsymbol{\chi}) = 1 \quad, \tag{7}$$

where $p(\tau^*, \boldsymbol{\theta}, S|\boldsymbol{\chi})$ is probability that the MAP correctly selects τ^* for a i.i.d. sequence (χ_1, χ_S) generated by $(\tau^*, \boldsymbol{\theta})$. In other words, the MAP maximizes the posterior probability $p(\tau)E_{\boldsymbol{\theta}}[p(\boldsymbol{\chi}|\tau,$
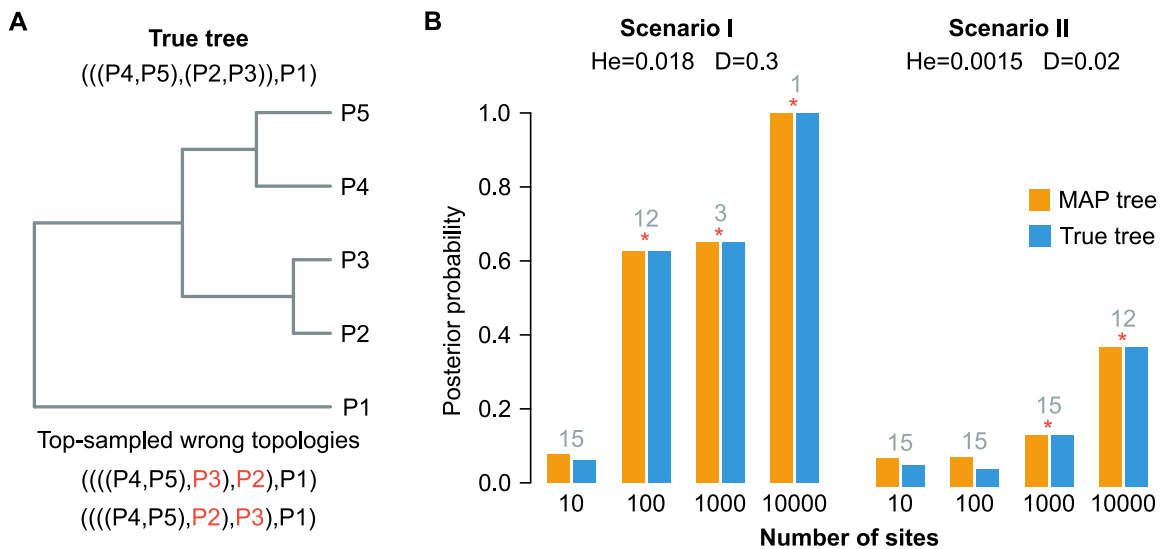
Fig. 2. Consistency of the Bayesian PoMo. **(A)** Topology (and relative branch lengths) used for simulations and the two top-visited wrong topologies sampled from the posterior. These alternative topologies follow the expected tree discordance due to incomplete lineage sorting between populations 2 and 3. **(B)** Posterior probabilities of the MAP and true topologies for the two simulated scenarios I and II. The expected heterozygosity H_e and divergence D used to simulate each scenario were taken from natural populations of *D. simulans* (scenario I (Begun et al., 2007)) and great apes (scenario II (Prado-Martinez et al., 2013)). The asterisk indicates that the MAP tree corresponds to the true tree. Numbers in grey indicate the number of sampled posterior topologies.

θ)] where E_θ is the expectation of the likelihood with respect to the prior probability of $\Theta(\tau)$.

Consistency can then be inherently gained by Bayesian inference. PoMo can be easily placed in a Bayesian framework. Indeed, we can easily meet conditions 1 and 2 in standard Bayesian phylogenetic inference software (e.g. BEAST (Bouckaert et al., 2019) and RevBayes (Höhna et al., 2016)). C1 requires that a nonzero prior is set on the tree topology, which can be easily implemented by setting a uniform prior. If θ takes the usual exponential/gamma priors on the branch lengths and rate parameters (ρ, σ and ω) or the Dirichlet distribution on π (which can actually be generated from a set of K-independent gamma random variables), condition C2 is met. Conditions C3 and C4 hold for all Markov processes on any tree with pendant edges of positive lengths (i.e. on any binary metric-tree) for which identifiability was proven. Therefore, as shown in the previous section, C3 and C4 hold for PoMo. Consequently, the MAP tree under PoMo is a consistent estimator of the species tree.

5. A simulation-based example of consistency with PoMo

Consistency guarantees the identification of the correct parameter values with infinite sequence lengths. In real data situations, the sequence length is finite as is the running time. We have nevertheless tested the consistency of the tree topology for the Bayesian PoMo estimator using simulated population data sets.

We simulated alignments of 10 000 sites under PoMo using a phylogeny of five populations as shown in Fig. 2A with relative branch lengths defined in such a way that we have two closely and two distantly (i.e. twice the expected divergence) related populations. We simulated two scenarios I and II by mimicking fast and slowly evolving populations (expected divergence D equal to 0.3 and 0.02 substitutions per site, respectively) with expected heterozygosity H_e as observed in *Drosophila simulans* populations and among great apes (0.0015 and 0.018, respectively) (Begun et al., 2007; Prado-Martinez et al., 2013).

To test for consistency, we created four alignments including the first 10, 100, 1000, 10 000 sites. We fed these alignments to

RevBayes (Höhna et al., 2016) and performed standard Bayesian phylogeny estimation on them. We ran PoMo for two chains and 20 000 generations, keeping every 10th iteration. A burning period of 10% was defined by checking mixing, convergence, and auto-correlation of the MCMC chains.

We observed that MAP already recovers the true tree for the average gene length of 1000 sites (Fig. 2B). These simulated scenarios complement our proofs that the MAP is a consistent estimator of the true tree. We observed further that populations with more heterozygous alignments (i.e. scenario I) converge to the true tree faster. The effect of incomplete lineage sorting is evident in these examples, as the top-sampled wrong topologies (Fig. 2A) cluster the closely related populations 3 and 2 with the clade including population 4 and 5. These topologies are, for a fewer number of sites, the MAP trees of scenarios I and II.

6. Conclusions

Here we prove that PoMo is identifiable and further that the MAP is a statistically consistent estimator of the species tree when PoMo is placed in a Bayesian framework. This is the first time identifiability and consistency were shown for the polymorphism-aware phylogenetic models.

Identifiability of PoMo can be easily extended to important, more general models. Identifiability should be kept for generalizations that work on the standard PoMo instantaneous rate matrix: This is, for example, the case with balancing selection acting along with genetic drift. More complicated extensions of the PoMo models would include the joint inference of gene trees with the species trees; currently, PoMo directly estimates the species tree. Steel (2013) suggested that identifiability and consistency should still apply, as the gene trees can be viewed as nuisance parameters. A formal proof for this statement, especially for the case of gene trees undergoing gene gain, loss, and transfer, is missing.

The consistency property, as already mentioned, says nothing about the performance of the method in real contexts, where data is finite. However, consistency is a desirable property, especially for

the type of data PoMo is applied to: population-scale and genome-wide. This data is costly and technically difficult to obtain and indeed the data sets available in the literature are few (e.g. humans (Altshuler et al., 2010), great apes (Prado-Martinez et al., 2013), fruit flies (Lack et al., 2015) and *Arabidopsis* (Long et al., 2013)). It is only natural to expect that higher samples sizes are repaid with less erroneous estimates of model parameters, including the species tree.

Our simulated examples, though simple, show that the MAP recovers the true tree even when the expected heterozygosity is very low and with incomplete lineage sorting is present. ILS is a very well-known cause of discordance between gene and species trees (Maddison and Knowles, 2006), by affecting the probability of visited wrong topologies. Statistical consistency is thus a desirable property for phylogeny estimation on closely related populations. As we have seen, the MAP tree corresponds already to the true topology for 1000 sites, even for less diverse species.

As future work, we would want to explore the sequence length requirement under PoMo. This is essentially the sequence length that a phylogeny reconstruction method needs to recover the true tree with a considerably small error (Atteson, 1999). A low sequence length requirement is a condition for a computationally efficient method. In particular, it would be important to determine whether PoMo is a fast converging method (i.e. a method that requires a sequence length of only $\mathcal{O}(poly(n))$).

Declaration of Competing Interest

The authors have no conflicts of interest.

References

Allman, E.S., Ané, C., Rhodes, J.A., 2008. Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. Adv. Appl. Probab. 40 (01), 229–249. doi:10.1239/aap/1208358894.

Altshuler, D.L., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Collins, F.S., De La Vega, F.M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S.B., Gibbs, R.A., Knoppers, B.M., Lander, E.S., Lehrach, H., Mardis, E.R., McVean, G.A., et al., 2010. A map of human genome variation from population-scale sequencing. Nature 467 (7319), 1061–1073. doi:10.1038/nature09534.

Atteson, K., 1999. The performance of neighbor-joining methods of phylogenetic reconstruction. Algorithmica 25 (2–3), 251–278. doi:10.1007/PL00008277.

Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.-P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N., Pachter, L., Myers, E., Langley, C.H., 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol. 5 (11), e310. doi:10.1371/journal.pbio.0050310.

Borges, R., Szöllősi, G.J., Kosiol, C., 2019. Quantifying GC-biased gene conversion in great ape genomes using polymorphism-aware models. Genetics 212 (4), 1321–1336. doi:10.1534/genetics.119.302074.

Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F.K., Müller, N.F., Ogilvie, H.A., du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., et al., 2019. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 15 (4), e1006650. doi:10.1371/journal.pcbi.1006650.

Chai, J., Housworth, E.A., 2011. On Rogers' proof of identifiability for the GTR + Γ + i model. Syst. Biol. 60 (5), 713–718. doi:10.1093/sysbio/syr023.

Chang, J., Hartigan, J., 1991. Reconstruction of evolutionary trees from pairwise distributions on current species. In: Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface, pp. 254–257.

Chang, J.T., 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math. Biosci. 137 (1), 51–73. doi:10.1016/S0025-5564(96)00075-2.

De Maio, N., Schlötterer, C., Kosiol, C., 2013. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. Mol. Biol. Evol. 30 (10), 2249–2262. doi:10.1093/molbev/mst131.

De Maio, N., Schrempf, D., Kosiol, C., 2015. PoMo: an allele frequency-based approach for species tree estimation. Syst. Biol. 64 (6), 1018–1031. doi:10.1093/sysbio/syv048.

Durrett, R., 2008. Probability Models for DNA Sequence Evolution. Probability and its Applications. Springer New York, New York, NY doi:10.1007/978-0-387-78168-6.

Felsenstein, J., 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Biol. 22 (3), 240–249. doi:10.1093/sysbio/22.3.240.

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Biol. 27 (4), 401–410. doi:10.1093/sysbio/27.4.401.

Hall, B., 2015. Lie Groups, Lie Algebras, and Representations: An Elementary Introduction, second ed. Springer International Publishing, Switzerland doi:10.1017/s0025557200177174.

Höhna, S., Landis, M.J., Heath, T.A., Boussau, B., Lartillot, N., Moore, B.R., Huelsenbeck, J.P., Ronquist, F., 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst. Biol. 65 (4), 726–736. doi:10.1093/sysbio/syw021.

Jukes, T., Cantor, C., 1969. Evolution of protein molecules. In: Mammalian Protein Metabolism. Elsevier, pp. 21–132. doi:10.1016/B978-1-4832-3211-9.50009-7.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16 (2), 111–120. doi:10.1007/BF01731581.

Lack, J.B., Cardeno, C.M., Crepeau, M.W., Taylor, W., Corbett-Detig, R.B., Stevens, K.A., Langley, C.H., Pool, J.E., 2015. The *Drosophila* genome nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. Genetics 199 (4), 1229–1241. doi:10.1534/genetics.115.174664.

Leaché, A.D., Oaks, J.R., 2017. The utility of single nucleotide polymorphism (SNP) data in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 48 (1), 69–84. doi:10.1146/annurev-ecolsys-110316-022645.

Long, Q., Rabanal, F.A., Meng, D., Huber, C.D., Farlow, A., Platzer, A., Zhang, Q., Vilhjálmsson, B.J., Korte, A., Nizhynska, V., Voronin, V., Korte, P., Sedman, L., Mandáková, T., Lysak, M.A., Seren, Ü., Hellmann, I., Nordborg, M., 2013. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. Nat. Genet. 45 (8), 884–890. doi:10.1038/ng.2678.

Maddison, W.P., Knowles, L.L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55 (1), 21–30. doi:10.1080/10635150500354928.

Mirarab, S., Bayzid, M.S., Boussau, B., Warnow, T., 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346 (6215), 1250463. doi:10.1126/science.1250463.

Moran, P., 1958. Random processes in genetics. Math. Proc. Cambridge Philos. Soc. 54 (01), 60. doi:10.1017/S0305004100033193.

Pagel, M., 1999. Inferring the historical patterns of biological evolution. Nature 401 (6756), 877–884. doi:10.1038/44766.

Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., Kelley, J.L., Lorente-Galdos, B., Veeramah, K.R., Woerner, A.E., O'Connor, T.D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A.E., Malig, M., et al., 2013. Great ape genetic diversity and population history. Nature 499 (7459), 471–475. doi:10.1038/nature12228.

Rogers, J., Raveendran, M., Harris, R.A., Mailund, T., Leppälä, K., Athanasiadis, G., Schierup, M.H., Cheng, J., Munch, K., Walker, J.A., Konkel, M.K., Jordan, V., Steely, C.J., Beckstrom, T.O., Bergey, C., Burrell, A., Schrempf, D., Noll, A., Kothe, M., et al., 2019. The comparative genomics and complex population history of papio baboons. Sci. Adv. 5 (1), eaau6947. doi:10.1126/sciadv.aau6947.

Rogers, J.S., 2001. Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. Syst. Biol. 50 (5), 713–722. doi:10.1080/106351501753328839.

Schrempf, D., Hobolth, A., 2017. An alternative derivation of the stationary distribution of the multivariate neutral wright fisher model for low mutation rates with a view to mutation rate estimation from site frequency data. Theor. Popul. Biol. 114, 88–94. doi:10.1016/j.tpb.2016.12.001.

Schrempf, D., Minh, B.Q., De Maio, N., von Haeseler, A., Kosiol, C., 2016. Reversible polymorphism-aware phylogenetic models and their application to tree inference. J. Theor. Biol. 407, 362–370. doi:10.1016/j.jtbi.2016.07.042.

Steel, M., 1994. Recovering a tree from the leaf colourations it generates under a Markov model. Appl. Math. Lett. 7 (2), 19–23. doi:10.1016/0893-9659(94)90024-8.

Steel, M., 2013. Consistency of Bayesian inference of resolved phylogenetic trees. J. Theor. Biol. 336, 246–249. doi:10.1016/j.jtbi.2013.08.012.

Steel, M., Hendy, M.D., Penny, D., 1998. Reconstructing phylogenies from nucleotide pattern probabilities: a survey and some new results. Discrete Appl. Math. 88 (1–3), 367–396. doi:10.1016/S0166-218X(98)00080-8.

Szöllsi, G.J., Tannier, E., Daubin, V., Boussau, B., 2015. The inference of gene trees with species trees. Syst. Biol. 64 (1), e42–e62. doi:10.1093/sysbio/syu048.

Tajima, F., Nei, M., 1984. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. 1 (3), 269–285. doi:10.1093/oxfordjournals.molbev.a040317.

Tavaré, S., 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. 17 (2), 57–86.

Wald, A., 1949. Note on the consistency of the maximum likelihood estimate. Ann. Math. Stat. 20 (4), 595–601. doi:10.1214/aoms/1177729952.

Wu, J., Susko, E., 2010. Rate-variation need not defeat phylogenetic inference through pairwise sequence comparisons. J. Theor. Biol. 263 (4), 587–589. doi:10.1016/j.jtbi.2009.12.022.

Rui Borges

Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, Wien 1210, Austria

Carolin Kosiol*

Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, Wien 1210, Austria

Centre for Biological Diversity, University of St Andrews, St Andrews, Fife KY16 9TH, UK

*Corresponding author.

E-mail address: ck202@st-andrews.ac.uk (C. Kosiol)