Evolutionary inference via the Poisson Indel Process

Alexandre Bouchard-Côté[a] and Michael I. Jordan[b,1]

[a]Department of Statistics, University of British Columbia, Vancouver, BC, Canada V6T 1Z4; and [b]Departments of Statistics and Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2010.

Contributed by Michael I. Jordan, November 26, 2012 (sent for review July 10, 2012)

We address the problem of the joint statistical inference of phylogenetic trees and multiple sequence alignments from unaligned molecular sequences. This problem is generally formulated in terms of string-valued evolutionary processes along the branches of a phylogenetic tree. The classic evolutionary process, the TKF91 model [Thorne JL, Kishino H, Felsenstein J (1991) J Mol Evol 33(2):114–124] is a continuous-time Markov chain model composed of insertion, deletion, and substitution events. Unfortunately, this model gives rise to an intractable computational problem: The computation of the marginal likelihood under the TKF91 model is exponential in the number of taxa. In this work, we present a stochastic process, the Poisson Indel Process (PIP), in which the complexity of this computation is reduced to linear. The Poisson Indel Process is closely related to the TKF91 model, differing only in its treatment of insertions, but it has a global characterization as a Poisson process on the phylogeny. Standard results for Poisson processes allow key computations to be decoupled, which yields the favorable computational profile of inference under the PIP model. We present illustrative experiments in which Bayesian inference under the PIP model is compared with separate inference of phylogenies and alignments.

phylogenetics | systematics | sequence homology | point process

The field of phylogenetic inference is being transformed by the rapid growth in availability of molecular sequence data. There is an urgent need for inferential procedures that can cope with data from large numbers of taxa and that can provide inferences for ancestral states and evolutionary parameters over increasingly large time spans. Existing procedures are often not scalable along these dimensions and can be a bottleneck in analyses of modern molecular datasets.

A key issue that renders phylogenetic inference difficult is that sequence data are generally not aligned a priori, having undergone evolutionary processes that involve insertions and deletions. Consider Fig. 1, which depicts an evolutionary tree in which each node is associated with a string of nucleotides, where the string evolves via insertion, deletion, and substitution processes along each branch of the tree. Even if we consider evolutionary models that are stochastically independent along the branches of the tree (conditioning on ancestral states), the inferential problem of inferring evolutionary paths (conditioning on observed data at the leaves of the tree) does not generally decouple into independent computations along the branches of the tree. Rather, alignment decisions made throughout the tree can influence the posterior distribution on alignments along any branch.

This issue has come to the fore in a line of research beginning in 1991 with a seminal paper by Thorne et al. (1). In the "TKF91 model," a simple continuous-time Markov chain (CTMC) provides a string-valued stochastic process along each branch of an evolutionary tree. This makes it possible to define joint probabilities on trees and alignments, and thereby obtain likelihoods and posterior distributions for statistical inference. A further important development has been the realization that the TKF91 model can be represented as a hidden Markov model, and that generalizations to a broader class of string-valued stochastic processes with finite-dimensional marginals are therefore possible (2–9). This has the appeal that statistical inference under these processes (known as transducers) can be based on dynamic programming (10–13). Unfortunately, however, despite some analytical simplification that is feasible in restricted cases (14), the memory needed to represent the state space in these models is generally exponential in the number of leaves in the tree (15). Moreover, even in the simple TKF91 model, there does not appear to be additional structure in the state space that allows for simplification of the dynamic program. Indeed, the running time of the most sophisticated algorithm for computing marginals (16) depends on the number of homology linearizations, which is exponential in sparse alignments (17).

As a consequence of this unfavorable computational complexity, there has been extensive work on approximations, specifically on approximations to the joint marginal probability of a tree and an alignment, obtained by integrating over the derivation (8, 18). A difficulty, however, is that these marginal probabilities play a role in tree inference procedures as the numerators and denominators of acceptance probabilities for Markov chain Monte Carlo (MCMC) algorithms. Loss of accuracy in these values can have large, uncontrolled effects on the overall inference. A second approach is to consider joint models that are not obtained by marginalization of a joint continuous-time, string-valued process. A range of combinatorial (19–24) and probabilistic (25–29) models fall into this category. Although often inspired by continuous-time processes, obtaining a coherent and calibrated estimate of uncertainty in these models is difficult.

A third possible response to the computational complexity of joint inference of trees and alignments is to retreat to methods that treat these problems separately. In particular, as is often done in practice, one can obtain a multiple sequence alignment (MSA) via any method (often based on a heuristically chosen "guide tree") and then infer a tree based on the fixed alignment. This latter inferential process is generally based on the assumption that the columns of the alignment are independent; in such case, the problem decouples into a simple recursion on the tree [the "Felsenstein" or "sum-product" recursion (30)]. Such an approach can introduce numerous artifacts, however, both in the inferred phylogeny (28, 29, 31), and in the inferred alignment (32, 33).

It is also possible to iterate the solution of the MSA problem and the tree inference problem (34, 35), which can be viewed as a heuristic methodology for attempting to perform joint inference. The drawbacks of these systems include a lack of theoretical understanding, the difficulty of getting calibrated confidence intervals, and overalignment problems (17, 36).

Finally, other methods have focused on analyzing only pairs of sequences at a time (17, 37–39). Although this approach can considerably simplify computation (40, 41), it has the disadvantage that it is not based on an underlying joint posterior probability distribution.

Author contributions: A.B.-C. and M.I.J. designed research; A.B.-C. and M.I.J. performed research; A.B.-C. analyzed data; and A.B.-C. and M.I.J. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

See Profile on page 1141.

[1]To whom correspondence should be addressed. E-mail: jordan@eecs.berkeley.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1220450110/-/DCSupplemental.

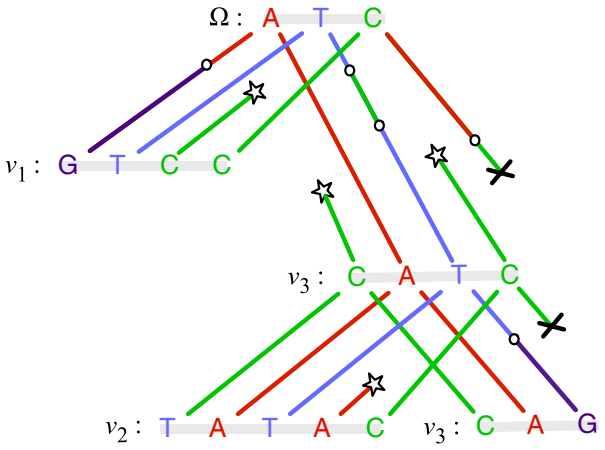

Fig. 1. Depiction of the evolution of a set of strings of nucleotides along the branches of a tree with leaves $\mathscr{S} = \{v_1, v_2, v_3\}$ and root Ω, where each string is subject to insertion, deletion, and substitution processes. Stars denote nucleotide insertion events, crosses denote deletion events, and circles denote substitution events.

In the current paper, we present a model-based approach to the joint probabilistic inference of trees and alignments. Our approach is based on a model that is closely related to TKF91, altering only the insertion process but leaving the deletion and substitution processes intact. Surprisingly, this relatively small change has a major mathematical consequence: Under our model, evolutionary paths have an equivalent global description as a Poisson process on the phylogenetic tree. We are then able to exploit standard results for Poisson processes (notably, Poisson thinning) to obtain significant computational leverage on the problem of computing the joint probability of a tree and an alignment. Indeed, under our model, this computation decouples in such a way that this joint probability can be obtained in linear time (linear in the number of taxa and the sequence length) rather than in exponential time, as in TKF91.

Our model has two descriptions: the first as a local continuous-time Markov process that is closely related to the TKF91 model and the second as a global Poisson process. We treat the latter as the fundamental description and refer to the newly developed process as the Poisson Indel Process (PIP). The global description not only sheds light on computational issues but opens up new ways to extend evolutionary models, allowing, for example, models that incorporate structural constraints and slipped-strand mispairing phenomena.

Under the Poisson process representation, another interesting perspective on our process is to view it as a string-valued counterpart to stochastic Dollo models (42, 43), which are defined on finite state spaces. In particular, the general idea of the two-step generation process used in Section 2 has antecedents in the literature on probabilistic modeling of morphological or lexical characters, but the literature did not address the string-valued processes that are our focus here.

The remainder of the paper is organized as follows. First, Section 1 provides some basic background on the TKF91 model. Next, in Section 2, we present the PIP model, in both its local and global formulations. Section 3 delves into the computational aspects of inference under the PIP model, describing the linear-time algorithm for computing the exact marginal density of an MSA and a tree. In Section 4, we present an empirical evaluation of the inference algorithm, and, finally, we present our conclusions in Section 5.

1. Background

We begin by giving a brief overview of the TKF91 model. Instead of following the standard treatment based on differential equations, we present a Doob–Gillespie view of the model (44, 45) that will be useful in our subsequent development.

Let us assume that at some point in time t, a sequence has length n. In the TKF91 process, the sequence stays unchanged for a random interval of time Δt, and after this interval, a single random mutation (substitution, insertion, or deletion) alters the sequence. This is achieved by defining a total of $3n + 1$ independent exponential random variables, n of which correspond to deletion of a single character, n of which correspond to the mutation of a single character, and $n + 1$ of which yield insertions after one of the n characters (including one "immortal" position at the leftmost position in the string). These $3n + 1$ exponential random variables are simulated in parallel, and the value of the smallest of these random variables determines Δt. The index of the winner determines the nature of the event at time Δt (whether it is a substitution, deletion, or insertion).

The random variables corresponding to a deletion have exponential rate μ_{TKF}, whereas those corresponding to an insertion have exponential rate λ_{TKF}. If the event is a mutation, a multinomial random variable with parameters obtained from the substitution rate matrix θ is drawn to determine the new value of the character. Finally, if an insertion occurs, a multinomial random variable is drawn to determine the value of the new character, with parameters generally taken from the stationary distribution of θ.

This describes the evolution of a string of characters along a single edge of a phylogenetic tree. The extension to the entire phylogeny is straightforward; we simply visit the tree in preorder and apply the single-edge process to each edge. The distribution of the sequences at the root is generally assumed to be the stationary distribution of the single-edge process (conceptually, the distribution obtained along an infinitely long edge).

Although the TKF91 model is reversible (and the PIP model as well, as we prove in Section 2.3), making the location of the root unidentifiable, it is useful to assume for simplicity that an arbitrary root has been picked, and we will make that assumption throughout. The likelihood is not affected by this arbitrary choice.

2. PIP

In this section, we introduce the PIP. This process has two descriptions: a local description that is closely related to the TKF91 model and a global description as a Poisson process.

We require some additional notation (Fig. 2). A phylogeny τ will be denoted by $(\mathscr{V}, \mathscr{E})$, where $\mathscr{V} \subset \tau$ is equal to the finite subset containing the branching points, the leaves $\mathscr{S} \subset \mathscr{V}$ and the root Ω, and where $\mathscr{E}$ is the set of edges. Parent nodes will be denoted by $\mathrm{pa}(v)$, for $v \in \mathscr{V}$, and the branch lengths will be denoted by $b(v)$, which is the length of the edge from $\mathrm{pa}(v)$ to v. For any $x \in \tau$ (whether x is a branch point in or an intermediate point on an edge), we write τ_x for the rooted phylogenetic subtree of τ rooted

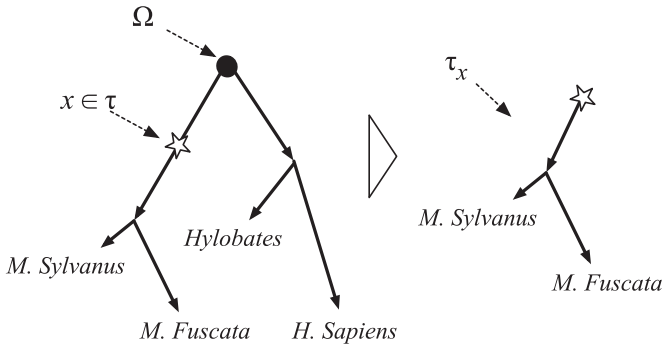

Fig. 2. Notation used for describing the PIP. Given a phylogenetic tree τ and a point $x \in \tau$ on that tree, τ_x is defined as the subtree rooted at x. *H. Sapiens*, *Homo sapiens*; *M. Fuscata*, *Macaca fuscata*; *M. Sylvanus*, *Macaca sylvanus*.

at x (dropping all points in the original tree that are not descendants of x). Finally, the set of characters (nucleotides or amino acids) will be denoted as Σ.

2.1. Local Description. The stochastic process we propose has a local description that is very similar to the TKF91 process, with the only change being that the insertion rate no longer depends on the sequence length. Therefore, instead of using $3n + 1$ competing exponential random variables to determine the next event as in the TKF91 model (n for substitutions, $n + 1$ for insertions, and n for deletions), we now have $2n + 1$ variables (n for substitutions, 1 for insertion, with rate λ, and n for deletion, each of rate μ). When an insertion occurs, its position is selected uniformly at random.* We assume that the process is initialized by sampling a Poisson-distributed number of characters, with parameter λ/μ. Each character is sampled independently and identically according to the stationary distribution of θ.

Note that if $\lambda/\lambda_{\mathrm{TKF}}$ is an integer and the sequence has the length $(\lambda/\lambda_{\mathrm{TKF}}) - 1$ at some point in time, the distribution over the time and type of the next mutation is the same as in TKF91, using the fact that the minimum of exponential variables with λ_i is exponential, with a rate equal to the sum of the λ_i. However, in general, the distributions are different. We discuss some of the biological aspects of these differences in Section 5; for now, we focus on the computational and statistical aspects of the PIP model.

2.2. Poisson Process Representation. We turn to a seemingly very different process for associating character strings with a phylogeny. This process consists of two steps, with the first involving insertions and the second involving deletions and substitutions.

In the first step, depicted in Fig. 3*A*, a multiset of insertion points is sampled from a Poisson process defined on the phylogeny τ (46). The rate measure for this Poisson process has an atomic mass at the root of the tree; hence, the need for multisets rather than simple point sets. Except for the root, no other points on the tree have an atomic mass (in particular, and in contrast to population genetics models, the probability that evolutionary events occur at branching points is 0). We denote this multiset of insertion points by $\mathbf{X}$.

In the second step, we visit the insertion points one at a time. The order of the visits of the insertions is sampled uniformly at random, $(X_1, X_2, \ldots, X_I) \sim \mathrm{Perm}(\mathbf{X})$. An insertion visit consists of two substeps. First, we extract the directed subtree rooted at the insertion location X_i. Examples of these subtrees are shown in Fig. 3*B* (*Left*). Second, we simulate the fate of the inserted character along τ_{X_i}. This is done via a substitution-deletion CTMC whose state space $\Sigma_\varepsilon = \Sigma \cup \{\varepsilon\}$ consists of the basic alphabet Σ augmented with an empty string symbol ε. As shown in Fig. 3*B* (*Right*), the substitution-deletion CTMC yields paths along subtrees in which a single character either mutates or is deleted. The latter event, represented by ε, is an absorbing state.

We define a homology path H_i as the single-character history generated by a substitution-deletion CTMC along a phylogeny. If a point $x \in \tau$ is a descendant of the insertion X_i, $H_i(x)$ is set to the state of the substitution-deletion CTMC at x. If $x \in \tau$ is not a descendant of X_i, we set $H_i(x)$ to the absorbing symbol ε. Thus, formally, a homology path H_i is a random map from any point on τ to Σ_ε.

Given a set of homology paths for each inserted character index i, the sequence at any point on the tree, $x \in \tau$, is obtained as follows (Fig. 3*C*, *Right*). First, we construct a list of all the values

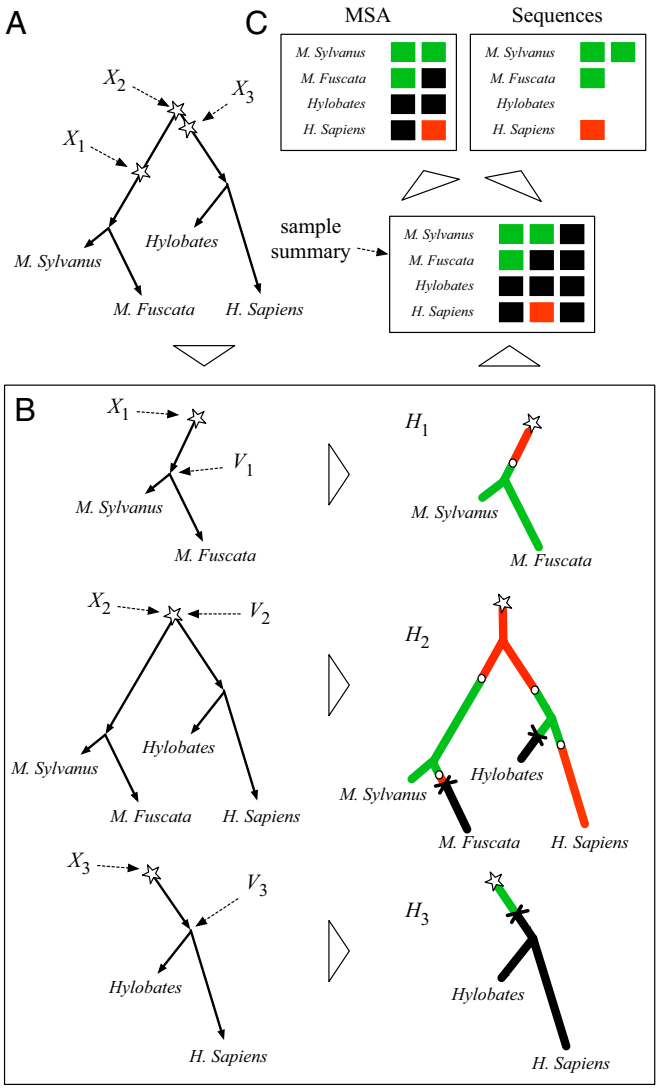

Fig. 3. Example of a PIP sample. Here, Σ has two symbols, represented by red and green squares, and the absorbing deletion symbol ε is represented in black. (*A*) Sample from a Poisson process on τ. (*B*) Each sampled point corresponds to a rooted tree on which a CTMC path is sampled. (*C*) Alignments and sequences are obtained as a deterministic function of the first two steps. *H. Sapiens*, *Homo sapiens*; *M. Fuscata*, *Macaca fuscata*; *M. Sylvanus*, *Macaca sylvanus*.

taken by $H_i(x)$ at the given point: $(H_1(x), H_2(x), \ldots, H_I(x))$. Second, we remove from the list any characters that are equal to the absorbing symbol ε. The string obtained thereby is denoted by $Y(x)$. The set of observed data comprises the values of Y at the leaves of the tree: $\mathcal{Y} = \{(v, Y(v)) : v \in \mathcal{L}\}$.

We can also construct an MSA M from a set of homology paths (Fig. 3*C*, *Left*). From each homology path H_i, we extract the characters at the leaves, arranging these characters in a column. Delete any column in which all the characters are the character ε. Arrange these columns in the order of the visits to the insertion points. The resulting matrix, whose entries range over the augmented alphabet Σ_ε, is the MSA M.

For a given rooted phylogenetic tree τ, we will denote by $p_\tau(m)$ the marginal probability that this process generates an MSA m, integrating over all homology paths, $p_\tau(m) = \mathbb{P}(M = m)$. For joint inference, we make the phylogenetic tree T random, with a distribution specified by a prior with density $p(\tau)$.

*More precisely, assume there is a real number r_i in the interval [0, 1] assigned to each character in the string in increasing order: $0 < r_1 < r_2 < \cdots < r_n < 1$. When an insertion occurs, sample a new real number r' uniformly in the interval [0, 1] and insert the new character at the unique position (with a probability one of 1), such that an increasing sequence of real numbers $0 < \cdots < r' < \cdots < 1$ is maintained.

2.3. Characterization. In this section, we show that the local and the global descriptions of the PIP given in the previous two subsections are, in fact, alternative descriptions of the same string-valued stochastic process. In stating our theorem, we let ν denote the rate measure characterizing the insertion process in the global description and Q and π denote the transition matrix and the initial distribution for the substitution-deletion CTMC.

2.3.1. Theorem 1. Let τ be a phylogenetic tree with an arbitrary rooting, and let us denote the Lebesgue measure on τ by the same symbol. For any insertion rate $\lambda > 0$, deletion rate $\mu > 0$, and reversible substitution rate matrix θ, the local and global processes described in Sections 2.1 and 2.2 coincide if we set for all $\sigma, \sigma' \in \Sigma_\varepsilon$:

$$\nu(\mathrm{d}x) = \lambda\left(\tau(\mathrm{d}x) + \frac{1}{\mu}\delta_\Omega(\mathrm{d}x)\right),$$

$$Q_{\sigma,\sigma'} = \begin{cases} -\sum_{\sigma'' \neq \sigma'} Q_{\sigma,\sigma''} & \text{if } \sigma = \sigma' \\ 0 & \text{if } \sigma = \varepsilon \\ \mu & \text{if } \sigma' = \varepsilon \\ \theta_{\sigma,\sigma'} & \text{o.w.} \end{cases},$$

and set π to be the quasi-stationary distribution of Q (47).

The proof is given in *SI Appendix*, Section 1. Note that in the case of interest here, where the rate of deletion does not depend on the character being deleted, π_σ is equal to the entry of the stationary distribution of θ corresponding to σ when $\sigma \neq \varepsilon$, and 0 otherwise. The following result establishes some basic properties of the PIP model. Its proof can be found in *SI Appendix*, Section 1.

2.3.2. Proposition 2. For all $\mu, \lambda > 0$ and reversible rate matrix θ, the PIP model is reversible, with a stationary length distribution given by a Poisson distribution with mean λ/μ.

The Poisson stationary length distribution represents a modeling advantage of PIP over TKF91, which has a geometrically distributed stationary distribution. Based on a study of protein-length distributions for the three domains of life (48), the Poisson distribution has been suggested (49) as a more adequate length distribution.

From proposition 2, we can also obtain an alternative reparameterization of the PIP model, in terms of asymptotic expected length $\eta = \lambda/\mu$ and insertion-deletion (indel) intensity $\zeta = \lambda \cdot \mu$.

3. Computational Aspects

We turn to a consideration of the computational consequences of the Poisson representation of the PIP model. We first consider how the Poisson process characterization allows us to compute the marginal likelihood, $p_\tau(m)$, in linear time, which is a significant improvement over methods based on the TKF91 model. In *SI Appendix*, Section 4, we provide a brief discussion of the role that the marginal likelihood plays in inference.

To compute the marginal likelihood, $p_\tau(m)$, we first condition on the number of homology paths, $|\mathbf{X}|$. Although the number of homology paths is random and unknown, we know that it can be no less than the number of columns $|m|$ in the postulated alignment m. We need to consider an unknown and unbounded number of birth events with no observed offspring in the MSA, but because they are exchangeable, they can be marginalized analytically. This is done as follows:

$$\begin{aligned} p_\tau(m) &= \mathbb{E}[\mathbb{P}(M = m||\mathbf{X}|)] \\ &= \sum_{n=|m|}^{\infty} \mathbb{P}(|\mathbf{X}| = n) \cdot \binom{n}{|m|} \cdot (p(c_\emptyset))^{n-|m|} \prod_{c \in m} p(c)\,, \end{aligned}$$

where the first factor captures the probability of sampling n homology paths, the second captures the number of ways to pick the $|m|$ observed homology paths (the columns, which contain at least one descendent character at the leaves) out of the n paths, the factor $p(c) = \mathbb{P}(C = c)$ is the likelihood of a single MSA column c, and $c_\emptyset$ is a column with an absorbing deletion symbol at every leaf $v \in \mathscr{L}$: $c_\emptyset \equiv \varepsilon$ (in this section, we drop subscripts for column-specific random variables, such as C, H, and X, because they are exchangeable). Note that such simplification is not possible in the TKF91 model, because the rate of insertion depends on the length of the internal sequences, and hence of the deletion events.

This expression can be simplified by introducing the function j defined as follows for all $z \in (0,1), k \in \{1,2,\ldots\}$:

$$\varphi(z,k) = \frac{1}{k!}\|\nu\|^k \exp\{(z-1)\|\nu\|\},$$

$$\|\nu\| = \lambda\left(\|\tau\| + \frac{1}{\mu}\right),$$

where $\|\tau\|$ is the normalization of the measure τ (i.e., the sum of all the branch lengths in the topology). We show in *SI Appendix*, Section 2 that this yields the simple formula:

$$p_\tau(m) = \varphi(p(c_\emptyset), |m|) \prod_{c \in m} p(c).$$

The next step is to compute the likelihood $p(c)$ of each individual alignment column c. We do this by partitioning the computation into subcases depending on the location of the tree at which the insertion point X is located for column c. More precisely, we look at the most recent common ancestor $V = v \in \mathscr{V}$ of the characters in c that are not equal to ε (Fig. 3*B*). If $v \neq \Omega$, this corresponds to the most recent end point of the edge $e \in \mathscr{E}$ where the insertion occurred.

Computing the prior probability of the insertion location is greatly simplified by the fact that $X||\mathbf{X}| \sim \bar{\nu}$ (ref. 50, chap. 2.4), where $\bar{\nu} = \nu/\|\nu\|$ denotes the probability obtained by normalizing the measure ν. We can therefore write:

$$\begin{aligned} \mathbb{P}(V = v) &= \begin{cases} \bar{\nu}(e \setminus \{\Omega\}) & \text{if } v \neq \Omega \\ \bar{\nu}(\{\Omega\}) & \text{o.w.} \end{cases} \\ &= \frac{1}{\|\tau\| + 1/\mu} \times \begin{cases} b(v) & \text{if } v \neq \Omega \\ 1/\mu & \text{o.w.} \end{cases} \end{aligned}$$

Finally, the column probabilities are computed as follows:

$$\begin{aligned} \mathbb{P}(C = c) &= \sum_{v \in \mathscr{V}} \mathbb{P}(V = v)\mathbb{P}(C = c|V = v) \\ &= \sum_{v \in \mathscr{V}} \mathbb{P}(V = v) f_v, \end{aligned}$$

where f_v is the output of a slight modification of Felsenstein's peeling recursion (30) applied on the subtree rooted at v (the derivation for f_v can be found in *SI Appendix*, Section 2). Because computing the peeling recursion for one column takes time $O(|\mathscr{L}|)$, we get a total running time of $O(|\mathscr{L}| \cdot |m|)$, where $|\mathscr{L}|$ is the number of observed taxa and $|m|$ is the number of columns in the alignment.

4. Experiments

We implemented a system based on our model that performs joint Bayesian inference of phylogenies and alignments. We used this system to quantify the relative benefits of joint inference relative to separate inference under the PIP and TKF91 models (i.e., the benefits of inferring trees on accuracy of the inferred

MSA and the benefits of inferring MSAs on the accuracy of the inferred tree).

We used synthetic data to assess the quality of the tree reconstructions produced by PIP, compared with the reconstructions of phylogenetic estimation using maximum likelihood (PhyML) 2.4.4, a widely used platform for phylogenetic tree inference (51). We also compared the inferred MSAs with those produced by Clustal 2.0.12 (52), a popular MSA inference system.

Although our implementation evaluated in this section is based on the Bayesian framework, we evaluate it using a frequentist methodology. More precisely, we use Bayes estimators (described in *SI Appendix*, Section 4) to obtain two point estimates from the posterior: one for the MSA and one for the phylogeny. Each point estimate is compared with the true alignment and tree. It is therefore possible to compare the method with the well-known frequentist methods mentioned above.

In this study, we explored four types of potential improvements: (*i*) resampling trees and MSAs increasing the quality of inferred MSAs, compared with resampling only MSAs; (*ii*) resampling trees and MSAs increasing the quality of inferred trees, compared with resampling only trees; (*iii*) resampling trees increasing the quality of inferred trees, compared with trees inferred by PhyML, and fixing the MSA to the one produced by Clustal; and (*iv*) resampling MSAs increasing the quality of inferred MSAs, compared with MSAs inferred by Clustal, and fixing the tree to the one produced by PhyML. The results are shown in Table 1. These experiments were based on 100 replicas, with each having seven taxa at the leaves, a topology sampled from the uniform distribution, branch lengths sampled from rate 2 exponential distributions, indels generated from the PIP with parameters $\eta = 100$ and $\zeta = 1$, and nucleotides sampled from the Kimura two-parameter model (53).

We measured the quality of MSA reconstructions using the F1 score, defined as the harmonic mean of the reconstructed alignment edge recall [called the sum-of-pairs score or developer's score in the MSA literature (54)] and alignment edge precision [modeler's score (55)]. We measured the quality of tree reconstructions using the partition (symmetrical clade difference) metric (56) and the weighted Robinson–Foulds metric (57). Relative improvements were obtained by computing the absolute value of the quality difference (in terms of the F1 for alignments and Robinson–Foulds distance for trees), divided by the initial value of the measure. We report relative improvements averaged over the 100 replicas.

We observed improvements of all four types. Comparing edge F1 relative improvements with Robinson–Foulds relative improvements, the relative additional improvement of type 2 is larger (13%) than that of type 1 (3%). Overall (i.e., comparing the baselines with the joint system), the full improvements of both trees and MSAs are substantial: 43% edge F1 improvement and 27% Robinson–Foulds improvement. A summary of the relative improvements is provided in Fig. 4.

Table 1. PIP results on simulated data

	Reconstruction accuracy			
Tree resampled?	No	Yes	No	Yes
MSA resampled?	No	No	Yes	Yes
Edge recall (SP)	**0.25**	—	0.22	0.24
Edge Precision	0.22	—	0.56	**0.58**
Edge F1	0.23	—	0.31	**0.32**
Partition Metric	0.24	0.22	—	**0.19**
Robinson-Foulds	0.45	0.38	—	**0.33**

Reconstruction accuracy using five different metrics. The bold font highlights the best-performing combination of resampling for each row.

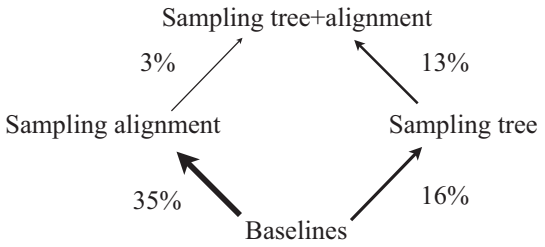

Fig. 4. Relative improvements for enabling each component of the sampler. Arrows on the left are relative alignment improvements, and arrows on the right are relative tree improvements.

We also tested our system on data generated from the TKF91 model instead of the PIP model. We used the same tree distribution and number of replicas as in the previous experiments and the same generating TKF91 parameters as Holmes and Bruno (2). We again observed improvements over the baseline, both in terms of MSA and tree quality. For MSAs, the relative improvement over the baseline was actually larger on the TKF91-generated data than on the PIP-generated data (47% vs. 43%, as measured by edge F1 improvement over Clustal), and it was lower but still substantial for phylogenetic trees (13% vs. 27%, as measured by Robinson–Foulds improvement over PhyML).

It should be noted that the MCMC kernels used in these experiments (described in the *SI Appendix*, Section 3) are based on simple Metropolis–Hastings proposals, and can therefore suffer from high rejection rates in large datasets. Fortunately, previous work in the statistical alignment literature has developed sophisticated MCMC kernels, some of which could be applied to inference in our model (e.g., ref. 28). Another potential direction would be to replace the MCMC by a sequential Monte Carlo posterior approximation (58).

It should also be emphasized that point indels are certainly not the exclusive driving force behind sequence evolution. In particular, "long indels" (atomic insertions and deletions of long segments, with a probability higher than the product of their point indels) are also prominent. As a consequence, any system purely based on point indels will have significant biases on biological data. In practice, these biases will introduce three undesirable artifacts: overestimation of the branch lengths; "gappy alignments," where the reconstructed MSA has many scattered gaps instead of a few long ones; and the related "ragged end" problem, where the prefix and suffix of sequences are poorly aligned because observed sequences are often truncated in practice. In the next section, we propose ways to address these limitations.

5. Discussion

We have presented a string-valued evolutionary model that can be used for joint inference of phylogenies and MSAs. As with its predecessor, the TKF91 model, our model can be used to capture the homology of characters evolving on a phylogenetic tree under insertion, deletion, and substitution events. Its advantage over TKF91 is that it permits a representation as a Poisson process on the tree. This representation has the consequence that the marginal likelihood of a tree and an alignment (marginalizing over ancestral states) can be computed in time linear in the number of taxa rather than exponentially, as in the case of TKF91. Poisson representations have played an important role in pure substitution processes (42, 43, 59), but in this work, we use Poisson representations for indel inference.

Although the insertion process in TKF91 might be argued to be more realistic biologically than that of the PIP model in that it allows the insertion rate to vary as the sequence length varies, in the common setting, in which all the sequences being aligned are of roughly similar lengths, this extra degree of freedom may be of

limited value for inference. Indeed, in our experiments, we saw that the PIP model can perform well even when data are generated from the TKF91 model. We might also note that there are biological processes in which insertions originate from a source that is extrinsic to the sequence (e.g., viruses, other genomic regions); in such case, the constant-rate assumption of PIP may actually be preferred.

It is also important to acknowledge, however, that neither TKF91 nor PIP is an accurate representation of biology. Their use in phylogenetic modeling reflects the hope that the statistical inferences they permit, most notably taking into account the effect of indels on the tree topology, will nonetheless be useful as data accrue. This hope is more likely to be realized in larger datasets, motivating our goal of obtaining a method that scales to larger sets of species. However, both models should also be viewed as jumping-off points for further modeling that is more faithful to the biology while retaining the inferential power of the basic models. For example, there has been significant work on extending TKF91 to models that capture the long indels that arise biologically but are not captured by the basic model (7, 60, 61).

In this regard, we wish to note that the Poisson representation of the PIP model provides avenues for extension that are not available within the TKF91 framework. In particular, the superposition property of Poisson processes makes it possible to combine the PIP model with other models that follow a Poisson law. For example, if the location X' of long indels, slipped-strand mispairing (62), or other nonlocal changes follows a Poisson point process, the union $U = X \cup X'$ of the nonlocal changes with the point indels X provided by a PIP will also be distributed according to a Poisson process. Moreover, the thinning property of Poisson processes provides a principled approach to inference for such superpositions. Indeed, an MCMC sampler for the superposition model can be constructed as follows. First, we can exploit the decomposition to analytically marginalize X (using the algorithm presented in this paper). Second, the other terms of the superposition and the sequences at these points in time can be represented explicitly as auxiliary variables. Because we have an efficient algorithm for computing the marginal likelihood, the auxiliary variables can be resampled easily. Note that designing an irreducible sampler without marginalizing X would be difficult: Integrating out X creates a bridge of positive probability between any pair of patterns of nonlocal changes.

Under the parameterization of the process used in this paper, the model assumes both an equal deletion rate for all characters and a uniform probability over inserted characters. It is worth noting that our inference algorithm can be modified to handle models relaxing both assumptions by replacing the calculation of $\beta(v)$ in *SI Appendix*, Section 2 by a quasi-stationary distribution calculation (47). It would be interesting to use this idea to investigate what nonuniformities are present in biological indel data.

Finally, another avenue to improve PIP models is to make the insertion rate mean measure more realistic: Instead of being uniform across the tree, it could be modeled using a prior distribution, hence forming a Cox process (63). This would be most useful when the sequences under study have large length or indel intensity variations across sites and branches (64).

ACKNOWLEDGMENTS. We thank Bastien Boussau, Ian Holmes, Michael Newton, and Marc Suchard for their comments and suggestions. This work was partially supported by a grant from the Office of Naval Research under Contract N00014-11-1-0688, by Grant K22 HG00056 from the National Institutes of Health, and by Grant SciDAC BER KP110201 from the Department of Energy.

1. Thorne JL, Kishino H, Felsenstein J (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *J Mol Evol* 33(2):114–124.
2. Holmes I, Bruno WJ (2001) Evolutionary HMMs: A Bayesian approach to multiple alignment. *Bioinformatics* 17(9):803–820.
3. Hein J (2001) An algorithm for statistical alignment of sequences related by a binary tree. *Pac Symp Biocomput* 6:179–190.
4. Steel M, Hein J (2001) Applying the Thorne-Kishino-Felsenstein model to sequence evolution on a star-shaped tree. *Appl Math Lett* 14(6):679–684.
5. Metzler D, Fleissner R, Wakolbinger A, von Haeseler A (2001) Assessing variability by joint sampling of alignments and mutation rates. *J Mol Evol* 53(6):660–669.
6. Hein J, Jensen JL, Pedersen CN (2003) Recursions for statistical multiple alignment. *Proc Natl Acad Sci USA* 100(25):14960–14965.
7. Miklós I, Lunter GA, Holmes I (2004) A "Long Indel" model for evolutionary sequence alignment. *Mol Biol Evol* 21(3):529–540.
8. Lunter G, Miklós I, Drummond A, Jensen JL, Hein J (2005) Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics* 6:83.
9. Novák A, Miklós I, Lyngsø R, Hein J (2008) StatAlign: An extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics* 24(20):2403–2404.
10. Allison L, Wallace CS, Yee CN (1992) Finite-state models in the alignment of macromolecules. *J Mol Evol* 35(1):77–89.
11. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* 235(5):1501–1531.
12. Searls DB, Murphy KP (1995) Automata-theoretic models of mutation and alignment. *Proc Inf Conf Intell Syst Mol Biol* 3:341–349.
13. Mohri M (2009) *Handbook of Weighted Automata, Monographs in Theoretical Computer Science*, eds Droste M, Kuich W, Vogler H (Springer, Berlin), pp 213–254.
14. Song YS (2006) A sufficient condition for reducing recursions in hidden Markov models. *Bull Math Biol* 68(2):361–384.
15. Dreyer M, Smith JR, Eisner J (2008) Latent-variable modeling of string transductions with finite-state methods. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* 13:1080–1089.
16. Miklós I, Drummond A, Lunter G, Hein J (2003) *Algorithms in Bioinformatics* (Springer, Berlin).
17. Schwartz AS, Pachter L (2007) Multiple alignment by sequence annealing. *Bioinformatics* 23(2):e24–e29.
18. Westesson O, Lunter G, Paten B, Holmes I (2012) Accurate reconstruction of insertion-deletion histories by statistical phylogenetics. *PLoS ONE* 7(4):e34572.
19. Sankoff D (1975) Minimal mutation trees of sequences. *SIAM J Appl Math* 28(1):35–42.
20. Wheeler WC, Gladstein DS (1994) MALIGN: A multiple sequence alignment program. *J Hered* 85(5):417–418.
21. Lancia G, Ravi R (1999) GESTALT: Genomic Steiner alignments. *Notes in Computer Science* 1645:101–114.
22. Varon A, Vinh LS, Wheeler WC (2010) POY version 4: Phylogenetic analysis using dynamic homologies. *Cladistics* 26(1):72–85.
23. Snir S, Pachter L (2011) Tracing the most parsimonious indel history. *J Comput Biol* 18(8):967–986.
24. Löytynoja A, Vilella AJ, Goldman N (2012) Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics* 28(13):1684–1691.
25. Hein J (1990) Unified approach to phylogenies and alignments. *Methods Enzymol* 183:626–645.
26. Knudsen B, Miyamoto MM (2003) Sequence alignments and pair hidden Markov models using evolutionary history. *J Mol Biol* 333(2):453–460.
27. Rivas E (2005) Evolutionary models for insertions and deletions in a probabilistic modeling framework. *BMC Bioinformatics* 6:63.
28. Redelings BD, Suchard MA (2005) Joint Bayesian estimation of alignment and phylogeny. *Syst Biol* 54(3):401–418.
29. Redelings BD, Suchard MA (2007) Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. *BMC Evol Biol* 7:40.
30. Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol* 17(6):368–376.
31. Wong KM, Suchard MA, Huelsenbeck JP (2008) Alignment uncertainty and genomic analysis. *Science* 319(5862):473–476.
32. Roshan U, Livesay DR, Chikkagoudar S (2006) Improving progressive alignment for phylogeny reconstruction using parsimonious guide-trees. *Proc IEEE Int Symp Bioinformatics Bioeng* 6:159–164.
33. Nelesen S, Liu K, Zhao D, Linder CR, Warnow T (2008) The effect of the guide tree on multiple sequence alignments and subsequent phylogenetic analyses. *Pac Symp Biocomput* 13:25–36.
34. Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* 324(5934):1561–1564.
35. Liu K, et al. (2012) SATe-II: Very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Syst Biol* 61(1):90–106.
36. Lunter G, Drummond A, Miklós I, Hein J (2004) *Statistical Methods in Molecular Evolution, Series in Statistics in Health and Medicine*, ed Nielsen R (Springer, New York).
37. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4(4):406–425.
38. Roch S (2010) Toward extracting all phylogenetic information from matrices of evolutionary distances. *Science* 327(5971):1376–1379.
39. Bradley RK, et al. (2009) Fast statistical alignment. *PLOS Comput Biol* 5(5):e1000392.
40. Holmes I (2004) A probabilistic model for the evolution of RNA structure. *BMC Bioinformatics* 5:166.

41. Crawford FW, Suchard MA (2012) Transition probabilities for general birth-death processes with applications in ecology, genetics, and evolution. *J Math Biol* 65(3): 553–580.
42. Alekseyenko AV, Lee CJ, Suchard MA (2008) Wagner and Dollo: A stochastic duet by composing two parsimonious solos. *Syst Biol* 57(5):772–784.
43. Nicholls G, Gray R (2006) *Phylogenetic Methods and the Prehistory Languages* (McDonald Institute for Archaeological Research, Cambridge, UK), pp 161–172.
44. Doob JL (1945) Markoff chains: Denumerable case. *Trans Am Math Soc* 58(3):455–473.
45. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81(25):2340–2361.
46. Huelsenbeck JP, Nielsen R (1999) Effect of nonindependent substitution on phylogenetic accuracy. *Syst Biol* 48(2):317–328.
47. Buiculescu M (1972) Quasi-stationary distributions for continuous-time Markov processes with a denumerable set of states. *Revue Roumaine De Mathématiques Pures et Appliqués* XVII(1):1013–1023.
48. Zhang J (2000) Protein-length distributions for the three domains of life. *Trends Genet* 16(3):107–109.
49. Miklós I (2003) Algorithm for statistical alignment of sequences derived from a Poisson sequence length distribution. *Discrete Appl Math* 127(1):79–84.
50. Kingman JFC (1993) Poisson Processes *(*Oxford Studies in Probabilities, Oxford, UK*)*.
51. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52(5):696–704.
52. Higgins DG, Sharp PM (1988) CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. *Gene* 73(1):237–244.
53. 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.
54. Sauder JM, Arthur JW, Dunbrack RL, Jr. (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins* 40(1):6–22.
55. Zachariah MA, Crooks GE, Holbrook SR, Brenner SE (2005) A generalized affine gap model significantly improves protein sequence alignment accuracy. *Proteins* 58(2): 329–338.
56. Bourque M (1978) Arbres de Steiner et réseaux dont varie l'emplagement de certains somments. Ph.D. thesis (Université de Montréal, Montreal, QC, Canada).
57. Robinson D, Foulds L (1979) Comparison of weighted labelled trees. *Lecture Notes in Mathematics* 748:119–126.
58. Bouchard-Côté A, Sankararaman S, Jordan MI (2012) Phylogenetic inference via sequential Monte Carlo. *Syst Biol* 61(4):579–593.
59. Huelsenbeck JP, Larget B, Swofford DL (2000) A compound Poisson process for relaxing the molecular clock. *Genetics* 154(4):1879–1892.
60. Thorne JL, Kishino H, Felsenstein J (1992) Inching toward reality: An improved likelihood model of sequence evolution. *J Mol Evol* 34(1):3–16.
61. Miklós I, Toroczkai Z (2001) An improved model for statistical alignment. *Lect Notes Comput Sci* 2149:1–10.
62. Kelchner SA (2000) The evolution of non-coding chloroplast DNA and its application in plant systematics. *Ann Mo Bot Gard* 87(4):482–498.
63. Cox DR (1955) Some statistical methods connected with series of events. *J R Stat Soc Series B Stat Methodol* 17(2):129–164.
64. Sniret S, Pachter L (2006) Phylogenetic profiling of insertions and deletions in vertebrate genomes. *Lect Notes Comput Sci* 3909:265–280.

Supporting Information: Evolutionary Inference via the Poisson Indel Process

Alexandre Bouchard-Côté Michael I. Jordan

1 Proofs for the Main PIP Properties

In this section, we prove Theorem 1 and Proposition 2. We begin by stating and proving two lemmas.

Lemma 1 *Let $U \sim \mathrm{Unif}(0,t)$ and $W \sim \mathrm{Exp}(\mu)$ be independent for fixed $t, \mu > 0$. Then*

$$\mathbb{P}(W + U > t) = \frac{1 - \exp(-t\mu)}{t\mu}.$$

Proof: By conditioning:

$$\begin{aligned}\mathbb{P}(W + U > t) &= \mathbb{E}\big[\mathbb{P}(W + U > t|U)\big] \\ &= \int_0^t \frac{\exp(-x\mu)}{t}\,\mathrm{d}x \\ &= \frac{1 - \exp(-t\mu)}{t\mu}.\end{aligned}$$

■

Lemma 2 *Let τ_0 denote a degenerate topology consisting of a root Ω connected to a single leaf v_0 by an edge of length t. Let H_i be a homology path as defined in the main paper, with $\tau = \tau_0$. For all $x \in \tau$, define $I(x) = \{i : H_i(x) \neq \varepsilon, 1 \leq i \leq I\}$ and:*

$$\begin{aligned}N &= |I(\Omega)| \\ N' &= |I(v_0)|.\end{aligned}$$

Then $N \sim \mathrm{Poi}(\lambda/\mu)$ implies $N' \sim \mathrm{Poi}(\lambda/\mu)$.

Proof: To prove the result, we decompose N and N' as follows (see Figure S.1):

$$\begin{aligned} N_1 &= |I(\Omega)\backslash I(v_0)| \\ N_2 &= |I(\Omega) \cap I(v_0)| \\ N_3 &= |I(v_0)\backslash I(\Omega)| \\ N_4 &= |I\backslash I(\Omega)\backslash I(v_0)| \\ N &= N_1 + N_2 \\ N' &= N_2 + N_3. \end{aligned}$$

By the Coloring Theorem [1],

$$N_2 \sim \text{Poi}\left(\nu(\{\Omega\})\mathbb{P}(W > t)\right),$$

where W is a rate μ exponential random variable, and ν is as in the condition of Theorem 1. Therefore $N_2 \sim \text{Poi}(\lambda \exp(-t\mu)/\mu)$. Similarly,

$$N_3 \sim \text{Poi}\left(\nu(\tau\backslash\{\Omega\})\mathbb{P}(W + U > t)\right),$$

where $U \sim \text{Unif}(0, t)$, and therefore from Lemma 1, $N_3 \sim \text{Poi}(\lambda(1-\exp(-t\mu))/\mu)$. It follows that:

$$\begin{aligned} N' &= N_2 + N_3 \\ &\sim \text{Poi}\left(\frac{\lambda}{\mu}e^{-\mu} + \frac{\lambda}{\mu}\left(1 - e^{-\mu}\right)\right) \\ &= \text{Poi}\left(\frac{\lambda}{\mu}\right), \end{aligned}$$

which concludes the proof of the lemma. ■

We can now prove Theorem 1:

Proof: In order to establish the equivalence, it is enough to show that for all edges $e = (v \to v')$ in the tree, the following two properties hold:

1. The distribution of the string length at the ancestral endpoint, $|Y(v)|$, is identical in the local and global descriptions: a Poisson distribution with rate λ/μ.

2. The distribution of the number and locations of mutations that fall on $e\backslash\{v, v'\}$ are also identical in the local and global descriptions.

We will enumerate the edges in the tree in preorder, using induction to establish these two hypotheses on this list of edges.

In the base case, hypothesis 1 is satisfied by construction: the local description is initialized with a $\text{Poi}(\lambda/\mu)$-distributed number of characters, and in the global description, the intensity measure ν of the Poisson process $\mathbf{X}$ assigns a point mass λ/μ to $v = \Omega$.

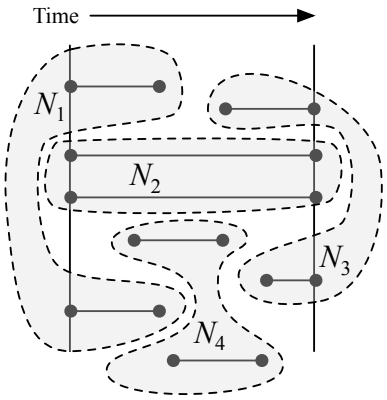

Figure S.1: Notation used in the appendix. The horizontal lines denote the times where each character is present in the sequence. The vertical line on the left denotes the sequence at Ω, and the vertical line on the right, the sequence at v_0. The sites are decomposed depending on whether they are present at each of two points Ω, v_0 in τ_0.

To establish hypothesis 1 in the inductive case, let $e' = (v'' \to v)$ denote the parent edge. By hypothesis 1 on e', $|Y(v'')| \sim \text{Poi}(\lambda/\mu)$, therefore by Lemma 2 and hypothesis 2 on e', hypothesis 1 is satisfied on e as well.

To establish hypothesis 2, it is enough to show that for all $x \in e \backslash \{v, v'\}$ the waiting time for each type of mutation given $Y(x)$ is exponential, with rates:

(a) λ for insertion,

(b) $\mu \cdot |Y(x)|$ for deletion, and

(c) $\sum_{\sigma \neq \varepsilon} \theta_{\sigma,\sigma'} |Y(x)|_\sigma$ for substitutions to $\sigma' \neq \varepsilon$, where $|s|_\sigma$ denotes the number of characters of type $\sigma \in \Sigma$ in the string $s \in \Sigma^*$.

Item (a) follows from the Poisson Interval Theorem [1]. Items (b) and (c) follow from the standard Doob-Gillespie characterization of CTMCs: if X_t is a CTMC with rate matrix $Q = (q_{i,j})$ and $Z_{i,j}$ are independent exponential random variables with rate $q_{i,j}$, then

$$(\Delta, J) | (X_0 = i) \stackrel{d}{=} (\min_{j \neq i} Z_{i,j}, \operatorname*{argmin}_{j \neq i} Z_{i,j}),$$

where $\Delta = \inf\{t : X_t \neq i\}$, $J = X_\Delta$. ■

We now turn to Proposition 2 and establish reversibility.

Proof: Let $h(n_1, n_2, n_3, n_4) = \mathbb{P}(N_i = n_i, i \in \{1, 2, 3, 4\})$. Using reversibility of θ, it is enough to show that h is invariant under the permutation (1 3); i.e., $h(n_1, n_2, n_3, n_4) = h(n_3, n_2, n_1, n_4)$.

We have that $h(n_1, n_2, n_3, n_4)$ is equal to:

$$\begin{aligned}
&\mathbb{P}\Big(N_i = n_i, \sum_i N_i = \sum_i n_i, N_1 + N_2 = n_1 + n_2, N_3 + N_4 = n_3 + n_4\Big) \\
&\quad = \mathbb{P}\Big(\sum_i N_i = \sum_i n_i\Big) \times \\
&\qquad \mathbb{P}\Big(N_1 + N_2 = n_1 + n_2, N_3 + N_4 = n_3 + n_4 \Big| \sum_i N_i = \sum_i n_i\Big) \times \\
&\qquad \mathbb{P}(N_1 = n_1, N_2 = n_2 | N_1 + N_2 = n_1 + n_2) \times \\
&\qquad \mathbb{P}(N_3 = n_3, N_4 = n_4 | N_3 + N_4 = n_3 + n_4) \\
&\quad = f_1(n_1 + n_2 + n_3 + n_4) \times \\
&\qquad \left(\frac{1/\mu}{1/\mu + t}\right)^{n_1+n_2} \left(\frac{t}{1/\mu + t}\right)^{n_3+n_4} \times \\
&\qquad \left(1 - e^{-\mu t}\right)^{n_1} f_2(n_2) \times \\
&\qquad \left(\frac{1 - e^{-\mu t}}{t\mu}\right)^{n_3} f_3(n_4),
\end{aligned}$$

where only the dependencies of the functions f_1, f_2 and f_3 is important in this argument, not their exact form. By inspection, it is clear that h is invariant under the permutation (1 3). ■

2 Proofs for the Likelihood Computation

First, we show how the function φ, defined in the main paper, simplifies the computation of $p_\tau(m)$:

$$\begin{aligned}
p_\tau(m) &= \mathbb{E}\big[\mathbb{P}(M = m||\mathbf{X}|)\big] \\
&= \sum_{n=|m|}^{\infty} \mathbb{P}(|\mathbf{X}| = n) \cdot \binom{n}{|m|} \cdot (p(c_\emptyset))^{n-|m|} \prod_{c \in m} p(c) \\
&= \frac{e^{\|\nu\|} \prod_{c \in m} p(c)}{|m|!(p(c_\emptyset))^{|m|}} \sum_{n=|m|}^{\infty} \frac{(\|\nu\| p(c_\emptyset))^n}{(n - |m|)!} \\
&= \frac{e^{\|\nu\|} (\|\nu\| p(c_\emptyset))^{|m|} \prod_{c \in m} p(c)}{|m|!(p(c_\emptyset))^{|m|}} \sum_{k=0}^{\infty} \frac{(\|\nu\| p(c_\emptyset))^k}{k!} \\
&= \frac{e^{\|\nu\|} (\|\nu\| p(c_\emptyset))^{|m|} \prod_{c \in m} p(c)}{|m|!(p(c_\emptyset))^{|m|}} \exp\left(\|\nu\| p(c_\emptyset)\right) \\
&= \varphi(p(c_\emptyset), |m|) \prod_{c \in m} p(c).
\end{aligned}$$

Next, we show how to compute $f_v = \mathbb{P}(C = c | V = v)$ for all $v \in \mathscr{V}$. The recursions for f_v are similar to those found in stochastic Dollo models [2]. Note first that f_v can be zero for some vertices. To see where and why, consider the subset of leaves S that that have an extant nucleotide in the current column c, $S = \{v \in \mathscr{L} : H(v) \neq \varepsilon\}$. Then f_v will be non-zero only for the vertices ancestral to all the leaves in S. Let us call this set of vertices A (see Figure S.2).

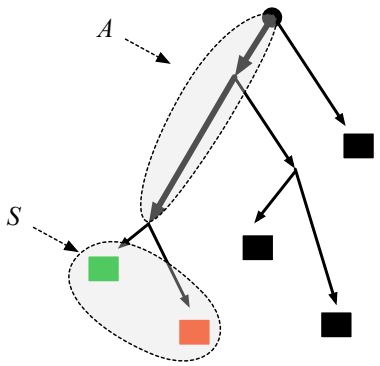

Figure S.2: Given a set S of leaves v with $H(v) \neq \varepsilon$, we define the set A of vertices with nonzero modified Felsenstein peeling weight to be those ancestral to the leaves in S. In this example, A contains three vertices.

To compute f_v on the remaining vertices, we introduce an intermediate variable, $\tilde{f}_v = \mathbb{P}(C = c|V = v, H(v) \neq \varepsilon)$. This variable can be computed using the standard Felsenstein peeling recursion (dynamic programming) as follows:

$$\tilde{f}_v(\sigma) = \begin{cases} \mathbf{1}(c(v) = \sigma) & \text{if } v \in \mathscr{L} \\ \sum_{\sigma' \in \Sigma_\varepsilon} \exp(b(v)Q)_{\sigma,\sigma'} \prod_{w \in \text{child}(v)} \tilde{f}_w(\sigma') & \text{o.w.} \end{cases} \tag{1}$$

$$\tilde{f}_v = \sum_{\sigma \in \Sigma} \pi_\sigma \tilde{f}_v(\sigma). \tag{2}$$

From Lemma 1, we have an expression for the survival probability at v given an insertion on the edge $(\text{pa}(v) \to v)$:

$$\begin{aligned} \beta(v) &= \mathbb{P}(H(v) \neq \varepsilon | V = v) \\ &= \frac{1}{b(v)} \frac{1}{\mu} \left(1 - e^{-\mu b(v)}\right). \end{aligned} \tag{3}$$

Finally, for $c \neq c_\emptyset$, we have:

$$\begin{aligned} f_v &= \mathbb{P}(C = c | V = v) \\ &= \mathbb{E}[\mathbb{P}(C = c | V = v, H(v))] \\ &= \begin{cases} \tilde{f}_v & \text{if } v = \Omega \\ \mathbf{1}[v \in A] \beta(v) \tilde{f}_v & \text{o.w.,} \end{cases} \end{aligned} \tag{4}$$

and for $c = c_\emptyset$:

$$f_v = \begin{cases} \tilde{f}_v & \text{if } v = \Omega \\ 1 + \beta(v)(\tilde{f}_v - 1) & \text{o.w.} \end{cases} \tag{5}$$

3 Proposal distributions

To perform full joint inference over trees and alignments using Markov chain Monte Carlo, several objects need to be resampled: the tree topology, the branch lengths, the MSA, and the parameters.

For trees and branch lengths, we use standard proposal mechanisms [3]. Our MSA proposal is inspired by the proposal of [4], avoiding the mixing problems of auxiliary variables [5, 6, 7]. Our proposal distribution consists of two steps. First, we partition the leaves into two sets A, B. Given a current MSA m_0, the support of the proposal is the set S of MSAs m satisfying the following constraints:

1. If e has both endpoints in A (or both in B), then $e \in m \iff e \in m_0$.
2. If e, e' have both endpoints in A (or both in B), then $e \prec_m e' \iff e \prec_{m_0} e'$.

The notation $\prec_m$ is based on the concept of posets over the columns (and edges) of an MSA [8].

We propose an element $m^* \in S$ with probability proportional to $\prod_{c \in m^*} p(c)$. The set S has exponential size, but can be sampled efficiently using standard pairwise alignment dynamic programming. A Metropolis-Hastings ratio is then computed to correct for φ. Note that the proposal induces an irreducible chain: one possible outcome of the move is to remove all links between two groups of sequences. The chain can therefore move to the empty MSA and then construct any MSA incrementally.

For the parameters, we used multiplicative proposals in the (λ, μ) parameterization [3].

4 Computational Aspects

In this section, we provide a brief discussion of the role that the marginal likelihood plays in both frequentist and Bayesian inference methods.

4.1 Maximum likelihood

In the case of maximum likelihood, the overall inference problem involves optimizing over the marginal likelihood:

$$\sup_{\tau \in \mathcal{T}(\mathscr{L}), m \in \mathcal{M}(y)} \log p_\tau(m),$$

where τ ranges over phylogenies on the leaves $\mathscr{L}$, and m ranges over the alignments consistent with the observed sequences y. This optimization problem can be approached using simulated annealing, where a candidate phylogeny and MSA pair (τ', m') is proposed at each step i, and is accepted (meaning that it replaces the previous candidate (τ, m)) according to a sequence of

acceptance functions $f^{(i)}(p, p')$ depending only on the marginal probabilities $p = p_\tau(m), p' = p_{\tau'}(m')$. Provided $\lim_{i\to\infty} f^{(i)}(p, p') = \mathbf{1}[p' > p]$ sufficiently slowly, this algorithm converges to the maximum likelihood phylogeny and MSA [9].

4.2 Bayes estimators

In order to define a Bayes estimator, one typically specifies a decision space D (for example the space of MSAs, or the space of multifurcating tree topologies, or both), a projection into this space, $(\tau, m) \mapsto \rho(\tau, m) \in D$, and a loss function $l : D \to [0, \infty)$ on D (for example, for tree topologies, the symmetric clade difference, or partition metric [10]; and for alignments, $1-$ the edge recall or Sum-of-Pairs (SP) score [11]).

Given these objects, the optimal decision in the Bayesian framework (also known as the consensus tree or alignment), is obtained by minimizing over $d \in D$ the risk $\mathbb{E}[l(d, \rho(T, M))|\mathcal{Y}]$. This expectation is intractable, so it is usually approximated with the empirical distribution of the output $(\tau^{(i)}, m^{(i)})$ of an Markov chain Monte Carlo (MCMC) algorithm. Producing MCMC samples boils down to computing acceptance ratios of the form:

$$\frac{p(\tau')p_{\tau'}(m')}{p(\tau)p_\tau(m)} \cdot \frac{q_{(\tau',m')}(\tau, m)}{q_{(\tau,m)}(\tau', m')},$$

for some proposal having density q with respect to a shared reference measure on $\mathcal{T}(\mathscr{L})\times\mathcal{M}(y)$. We thus see that for both maximum likelihood and joint Bayesian inference of the MSA and phylogeny the key problem is that of computing the marginal likelihood $p_\tau(m)$.

5 Pseudocode and Example

In this section, we summarize the likelihood computation. We also give a concrete numerical example to illustrate the calculation.

1. Inputs:
   (a) PIP parameter values (λ, μ), substitution matrix θ over Σ.
   Example: $(\lambda, \mu) = (2.0, 1.0), \Sigma = \{a\}$
   (b) Rooted phylogenetic tree τ
   Example: $\tau = ((v_2 : 1.0, v_3 : 1.0)v_0 : 1.0, v_4 : 2.0)v_1$;
   (c) Multiple sequence alignment m
   Example: $m =$

   ```
   v_2|-a
   v_3|aa
   v_4|a-
   ```

2. Computing modified Felsenstein recursion:

(a) For each site, compute $\tilde{f}_v(\sigma)$ in post-order using Equation (1), and from each $\tilde{f}_v(\sigma)$, compute $\tilde{f}_v$ using Equation (2)
Example:
for site 1, $(\tilde{f}_{v_2}, \tilde{f}_{v_3}, \tilde{f}_{v_0}, \tilde{f}_{v_4}, \tilde{f}_{v_1}) = (0.0, 1.0, 0.23, 1.0, 0.012)$*;*
for site 2, $(\tilde{f}_{v_2}, \tilde{f}_{v_3}, \tilde{f}_{v_0}, \tilde{f}_{v_4}, \tilde{f}_{v_1}) = (1.0, 1.0, 0.14, 0.0, 0.043)$*;*

(b) Do the same for an artificial site or column $c_\emptyset$ where all leaves have a gap
Example:
for site 3, $(\tilde{f}_{v_2}, \tilde{f}_{v_3}, \tilde{f}_{v_0}, \tilde{f}_{v_4}, \tilde{f}_{v_1}) = (0.0, 0.0, 0.40, 0.0, 0.67)$*;*

3. For each node v in the tree, compute the survival probability $\beta(v)$ using Equation (3) (setting it to 1 at the root for convenience)
Example:
$(\beta(v_2), \beta(v_3), \beta(v_0), \beta(v_4), \beta(v_1)) = (0.63, 0.63, 0.63, 0.43, 1.0)$

4. For each site, compute the set of nodes A ancestral to all extant characters, as described in the caption of Figure S.2
Example:
for site 1, $A = \{v_1\}$
for site 2, $A = \{v_0, v_1\}$

5. Computing f_v:

(a) For each site, compute f_v using Equation (4)
Example:
for site 1, $(f_{v_2}, f_{v_3}, f_{v_0}, f_{v_4}, f_{v_1}) = (0.0, 0.0, 0.0, 0.0, 0.012)$*;*
for site 2, $(f_{v_2}, f_{v_3}, f_{v_0}, f_{v_4}, f_{v_1}) = (0.0, 0.0, 0.086, 0.0, 0.043)$*;*

(b) For $c_\emptyset$, use Equation (5)
Example:
for site 3, $(f_{v_2}, f_{v_3}, f_{v_0}, f_{v_4}, f_{v_1}) = (0.37, 0.37, 0.62, 0.57, 0.67)$*;*

6. For each node v in the tree, compute $\iota_v = \mathbb{P}(V = v)$ as shown in Section 3 of the main paper
Example:
$(\iota(v_2), \iota(v_3), \iota(v_0), \iota(v_4), \iota(v_1)) = (0.17, 0.17, 0.17, 0.33, 0.17)$

7. Compute $p_\tau(m)$ from the ι_v's, f_v's as shown in Section 3 of the main paper
Example: $\log p_\tau(m) = -11$

References

[1] Kingman JFC (1993) *Poisson Processes* (Oxford Studies in Probabilities).

[2] Alekseyenko A, Lee C, Suchard MA (2008) Wagner and Dollo: a stochastic duet by composing two parsimonious solos. *Systematic Biology* 57 (5):772–784.

[3] Lakner C, van der Mark P, Huelsenbeck JP, Larget B, Ronquist F (2008) Efficiency of Markov Chain Monte Carlo Tree Proposals in Bayesian Phylogenetics. *Systematic Biology* 57:86–103.

[4] Lunter G, Miklós I, Drummond A, Jensen J, Hein J (2005) Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics* 6(83).

[5] Holmes I, Bruno WJ (2001) Evolutionary HMM: A Bayesian approach to multiple alignment. *Bioinformatics* 17:803–820.

[6] Jensen J, Hein J (2002) Gibbs sampler for statistical multiple alignment., (Dept of Theor Stat, U Aarhus), Technical report.

[7] Bouchard-Côté A, Jordan MI, Klein D (2009) Efficient inference in phylogenetic InDel trees. *In Proceedings of Advances in Neural Information Processing Systems* 21:177–184.

[8] Schwartz A, Pachter L (2006) Multiple alignment by sequence annealing. *Bioinformatics* 23:e24–e29.

[9] Delyon B (1988) Convergence of the simulated annealing algorithm., (Massachusetts Institute of Technology), Technical report.

[10] Bourque M (1978) Ph.D. thesis (Université de Montréal).

[11] Robert CP (2001) *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation* (Springer).