Version dated: January 17, 2017

Geometric Poisson Indel Process

A Poissonian model of indel rate variation for phylogenetic tree inference

Yongliang Zhai, Alexandre Bouchard-Côté

Department of Statistics, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada

Corresponding author: Alexandre Bouchard-Côté, Department of Statistics, University of British Columbia, 3182 Earth Sciences Building, 2207 Main Mall, Vancouver, British Columbia, V6T 1Z4, Canada; E-mail: bouchard@stat.ubc.ca.

Abstract.— While indel rate variation has been observed and analyzed in detail, it is not taken into account by current indel-aware phylogenetic reconstruction methods. In this work, we introduce a continuous time stochastic process, the geometric Poisson indel process, that generalizes the Poisson indel process by allowing insertion and deletion rates to vary across sites. We design an efficient algorithm for computing the probability of a given multiple sequence alignment based on our new indel model. We describe a method to construct phylogeny estimates from a fixed alignment using neighbor joining. Using simulation studies, we show that ignoring indel rate variation may have a detrimental effect on the accuracy of the inferred phylogenies, and that our proposed method can sidestep this issue by inferring latent indel rate categories. We also show that our phylogenetic inference method may be more stable to taxa subsampling than methods that either ignore indels or indel rate variation.

(Keywords: indel rate variation, Poisson indel process, evolutionary stochastic process, TKF91)

It is well known that different regions of nucleotide sequences evolve at different rates, both in terms of substitutions (Fitch and Margoliash 1967; Li *et al.* 1985; Nachman and Crowell 2000), and in terms of insertions-deletions (indels) (Mouchiroud *et al.* 1991; Wong *et al.* 2004; Lunter *et al.* 2006; Mills *et al.* 2006; Chen *et al.* 2009; Kvikstad and Duret 2014). In phylogenetic analyses based on substitutions, rate variation is viewed as an important phenomenon to include when building evolutionary models; consequently, virtually all modern phylogenetic methods explicitly model substitution rate variation across sites (Yang 1997; Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003; Suchard and Redelings 2006; Yang 2007; Guindon *et al.* 2010; Stamatakis 2014).

There is substantial previous work *analyzing* patterns of indel rate variation, but these analyses are typically done from trees and alignments inferred using standard models which ignore rate variation. This body of previous work has not only demonstrated that indel rate variation is widespread (Chen *et al.* 2009; Kvikstad and Duret 2014), but also identified correlates (and in some cases, mechanisms) behind indel rate heterogeneity, including sequence context (Tanay and Siggia 2008), substitution rate (Ananda *et al.* 2011; Jovelin and Cutter 2013), selection (Carvalho and Clark 1999; Kvikstad and Duret 2014), recombination (Nam and Ellegren 2012; Leushkin and Bazykin 2013) and short tandem repeats (Ellegren 2004).

There are now several approaches to phylogenetic tree inference that take indels into account (Thorne *et al.* 1991, 1992; Westesson *et al.* 2012), and some of them include substitution rate heterogeneity (Klosterman *et al.* 2006; Suchard and Redelings 2006; Redelings and Suchard 2007). However, these approaches generally do not incorporate

indel rate heterogeneity as part of the model specification. Although in the multiple sequence alignment literature, some methods do consider indel rate variation, those methods typically assume a fixed guide tree and are not based on a continuous-time stochastic process (Löytynoja and Goldman 2008), or are limited to fixed trees with a small number of leaves (Satija *et al.* 2009).

In this work, we present a simple indel rate heterogeneity model suitable for phylogenetic tree inference. As with substitution rate heterogeneity models, we approximate the distribution over rates using a discrete mixture. Given a discrete indel rate mixture, our model is obtained as the finite-dimensional marginal distributions Kallenberg (2002) of a reversible stochastic process defined on a phylogenetic tree. This continuous-time Markov process is called the geometric Poisson indel process (GeoPIP), which we introduce in this paper.

As its name suggests, the main building block of the GeoPIP model is the Poisson indel process (PIP) (Bouchard-Côté and Jordan 2013), and the GeoPIP model inherits the attractive computational properties of the PIP model. This means in particular that given a tree, computing the probability of an alignment (i.e., marginalizing over internal sequences) can be done in time polynomial in both the number of the sequences and the lengths of the sequences. This property forms the basis of an efficient algorithm which determines in an unsupervised fashion the indel rates, while inferring the tree and partitioning the sequences into segments taking on different indel rates.

Utilizing our efficient likelihood calculation algorithm to infer segmentations, we propose an algorithm to estimate phylogeny from a fixed multiple sequence alignment using the neighbor joining (NJ) algorithm (Saitou and Nei 1987; Studier *et al.* 1988; Gascuel 1997) as an illustration. It is also worth mentioning that a full likelihood approach, as well as joint inference of phylogeny and multiple sequence alignments, can also be implemented based on the GeoPIP model, using existing phylogenetic inference framework (Huelsenbeck

and Ronquist 2001; Suchard and Redelings 2006; Guindon *et al.* 2010; Bouchard-Côté *et al.* 2012; Hajiaghayi *et al.* 2014). Our inference method iteratively estimates a segmentation of the multiple sequence alignment, indel rates, phylogenetic tree and other relevant parameters, until convergence occurs or the full likelihood stops increasing. The exact marginalization still plays a key role because of the need to infer a segmentation and indel parameters. The segmentation of the multiple sequences alignment and indel rates are estimated using the GeoPIP model, based on our efficient algorithm to calculate the probability of multiple sequence alignment. The phylogenetic tree is constructed using neighbor joining based on pairwise distances which are calculated using GeoPIP model on pairwise sequence alignments that inherit the segmentation and indel rates estimated from the multiple sequence alignment. Our inference method is initialized using random starts, without requiring a guide tree.

Using our method, we investigate the effect of indel rate heterogeneity on phylogenetic inference. We provide some evidence that modelling indels enhances accuracy of phylogenetic inference, and that modelling indel rate heterogeneity can further improve the accuracy of phylogenetic inference. We demonstrate the accuracy of our method in both well-specified and misspecified synthetic experiments, including data generated using the software INDELible (Fletcher and Yang 2009) and aligned using the software MUSCLE (Edgar 2004a,b).

In this paper, we focus on modelling indel rate variation and consider only indels of size one. An important area of related work is the development of long indel models (Thorne *et al.* 1992; Miklós *et al.* 2004; Lunter *et al.* 2005b; Redelings and Suchard 2007). Modelling long indels is important in the context of phylogenetics because explaining the insertion or deletion of a segment with many single-character indels can lead to inaccurate tree estimation. Liu *et al.* (2009a) showed that using the affine gap penalty which models long indels directly can improve alignment and tree estimation accuracy. At the same time,

the indel rate is comparable with the substitution rate when the indel rate and the average indel length are separately estimated. This leads to more interpretable results which provide helpful insights into the ratio of indel event frequency and substitution event frequency. Unfortunately, the problem of reconciling long indels with a model that can be obtained as a tractable, exact marginalization of a continuous time stochastic process is still open and appears elusive. The state of affairs consists in complex approximations (Knudsen and Miyamoto 2003; Miklós *et al.* 2004), models that support insertions but not deletion (or vice versa) (Miklos and Toroczkai 2001), and methods limited to sequence pairs (Thorne *et al.* 1992).

For tractability reasons, we do not attempt to include long indels into our GeoPIP model. Instead, our strategy to avoid the branch overestimation is to have the GeoPIP model explain them with segment of very high indel rate. Our method shares a limitation of previous segment-based long indel methods (Thorne *et al.* 1992), namely that certain overlapping patterns of indels are not explained in the most parsimonious way (see Thorne *et al.* (1992) for examples). On the other hand, our method has better scaling properties as the number of taxa increases, compared to the TKF92 model which does not allow exact marginalization of internal nodes in polynomial time. To demonstrate that our strategy is sensible, we include synthetic experiments where the data are generated from models that include long indels. There is one potential caveat of modelling regions undergoing long indels using high indel intensity segments: indel rates in the GeoPIP model are not easily interpretable. This is because the rate categories conflate actual indel rate variation with higher indel intensity to explain long indels.

The statistical and computational properties of the GeoPIP model differentiate it from the model used in the alignment method of Lunter (2007). This previous work introduced a sequence aligner based on a string transducer. This transducer is equipped with groups of latent states encoding different indel rates. While Lunter's model is effective

for pairwise alignment, there are two important challenges in applying this model to phylogenetic tree inference. First, since Lunter's model is not defined as the finite-dimensional marginal distributions of a stochastic process on a phylogenetic tree, there is no straightforward approach to using this model for tree reconstruction. Second, summing over the sequences on the internal nodes of a tree using Lunter's transducer model leads to a worst-case running time exponential in the number of taxa (this can be derived using the results in Hirschberg (1975)). Consequently, Lunter's model has not been used for phylogenetic tree inference. Incidentally, we show that even if one only cares about identifying the rate segmentation (with a fixed guide tree), using more sequences jointly improves inference accuracy. Again, one would have to resort to approximations to do so with a transducer-based approach (Holmes and Bruno 2001; Holmes 2003; Miklós *et al.* 2004; Jensen and Hein 2005; Bouchard-Côté *et al.* 2008), while we can do this exactly in time linear in the number of sequences with the GeoPIP model.

Background and notation

Before describing the GeoPIP model, we introduce our notation, and review the PIP model, which is the foundation of our method. In the following, we assume that sequences from different species take the form of a multiple sequence alignment (MSA) of characters from a finite alphabet Σ (for example, $\Sigma = \{\text{A, C, G, T}\}$ for DNA data). MSAs are sets of homologous characters which can be visualized using an alignment matrix, where each row represents one aligned sequence and each column represents one set of homologous characters at a certain locus. When there are no homologous characters observed at a locus in one sequence, a gap symbol "–" is padded at the locus of that sequence so that two characters are in the same column of the alignment matrix if and only if they are homologous. Let $\Sigma_+ = \Sigma \cup \{-\}$ denote the expanded set of symbols including the gap symbol "–".

Let $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)'$ denote a fixed MSA of sequences from N different species with n columns, ($\mathbf{x}_i \in \Sigma_+^n$, $i = 1, 2, \ldots, N$). We will also use $\mathbf{x}$ as $\mathbf{x} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n)$ for a fixed MSA with columns ($\mathbf{c}_j \in \Sigma_+^N$, $j = 1, 2, \ldots, n$).

We let $\mathbf{Q}$ denote a reversible substitution rate matrix over a state space $\mathcal{X}$. Here, $\mathcal{X}$ could be taken to be the finite alphabet Σ, or $\mathcal{X}$ could be the set of pairs containing a character in Σ together with a substitution rate category annotation from a discrete set of substitution category indices. To simplify the notation, we take $\mathcal{X} = \Sigma$ in the following, but we note that substitution heterogeneity can be handled in our framework with no change on the algorithms or properties of the method. Let π denote the stationary distribution of the rate matrix $\mathbf{Q}$. Finally, we let τ denote an unobserved phylogenetic tree with leaves labelled with the same taxa as those indexing the rows of the MSA $\mathbf{x}$.

The Poisson indel process

Bouchard-Côté and Jordan (2013) proposed the PIP to model insertion, deletion and substitution of characters in string-valued continuous time processes. The description of the PIP model on a string of k characters consists of two steps: first, the type of the next change (insertion, deletion or substitution) is determined by a realization of $2k + 1$ exponential random variables; second, the exact change is determined based on the type of change and realization of some type-specific random variables.

The first step is generated as follows. For a sequence of length k, the PIP model assumes that the smallest of $2k + 1$ exponential random variables determines the nature of the next evolutionary event and the waiting time. The waiting time for a potential insertion event is exponentially distributed with rate $\lambda > 0$ (this random variable does not determine the location of the insertion since all $k + 1$ possible insertion sites share the same random variable for insertion). The waiting times for k potential deletion events are independently and identically exponentially distributed with rate $\mu > 0$ (these random

variables determine the location of the deletions since there is one random variable for deletion of each site). The waiting times for k potential substitution events are independently exponentially distributed with rates based on the substitution rate matrix $\mathbf{Q}$. We let $\theta = (\lambda, \mu)$ denote the two indel parameters of the PIP model.

The second step is generated as follows. If the next event is an insertion, the location of the insertion is uniformly selected from $k+1$ possible insertion positions, and a new character is randomly generated based on a multinomial distribution with parameter π, which is the stationary distribution of rate matrix $\mathbf{Q}$. If the next event is deletion, the character associated with the smallest realization of the k deletion random variables is deleted from the sequence. If the next event is substitution, a new character is randomly generated from a multinomial distribution based on respective rows of the rate matrix $\mathbf{Q}$ determined by the character to be substituted.

Bouchard-Côté and Jordan (2013) showed that under the PIP model, the marginal probability mass function of observing an alignment $\mathbf{x}$ at the leaves of a given tree τ is

$$\mathrm{PIP}(\mathbf{x}|\theta,\tau) = \psi(\Pr(\mathbf{c}_\emptyset|\theta,\tau), n, \theta, \tau) \prod_{i=1}^{n} \Pr(\mathbf{c}_i|\theta,\tau), \tag{1}$$

where $\mathbf{c}_\emptyset$ is a single MSA column with empty characters "–" at each leaf, θ is the indel rate, and n is the number of alignment columns. The function ψ in (1) is given by

$$\psi(z,k,\theta,\tau) = \frac{1}{k!}\|\nu_{\theta,\tau}\|^k \exp\{(z-1)\|\nu_{\theta,\tau}\|\}, \tag{2}$$

where $\|\nu_{\theta,\tau}\| = \lambda(\|\tau\| + 1/\mu)$ and $\|\tau\|$ is the sum of all branch lengths in τ. The stationary sequence length distribution is given by a Poisson distribution with mean λ/μ (Bouchard-Côté and Jordan 2013), which is a more adequate length distribution than the geometric sequence length distribution induced by the TKF model (Miklós 2003).

Bouchard-Côté and Jordan (2013) proposed a dynamic programming algorithm, which adds one row and one column representing deletion to the rate matrix, to calculate $\Pr(\mathbf{c}_i|\theta,\tau)$ efficiently based on a variation of Felsenstein's peeling recursion algorithm (Felsenstein 1981), as well as a Bayesian framework for phylogenetic inference based on the PIP model.

The geometric Poisson indel process

The GeoPIP model is based on the concept of *MSA segment*, which we define as a group of contiguous MSA columns in which indels are assumed to accumulate at a similar rate. We define a *segmentation* β of a fixed MSA $\mathbf{x}$ as a partition of the MSA columns $\mathbf{x}_1,\ldots,\mathbf{x}_N$ into MSA segments, i.e., $\beta=(\mathbf{s}_1,\mathbf{s}_2,\ldots,\mathbf{s}_Z)$ where $\mathbf{s}_k$ is the k-th segment and $Z=|\beta|$ is the number of segments ($k=1,2,\ldots,|\beta|$). To be specific, $\mathbf{s}_k=(\mathbf{c}_{d_{k-1}+1},\ldots,\mathbf{c}_{d_k})$ where $d_k=\sum_{j=1}^{k-1}|\mathbf{s}_j|$ ($k=1,2,\ldots,Z$) and $d_0=0$.

It is common in substitution rate variation models to assume a discrete set of possible rate categories (Yang 1996). Here we proceed similarly, and define a finite list of indel rate categories $\theta_1=(\lambda_1,\mu_1),\ldots,\theta_m=(\lambda_m,\mu_m)$, where each item in the list is just a distinct PIP indel parameter setting. However, in contrast to discrete substitution rate models, where each rate is often obtained using a discretized gamma distribution, we do not assume a specific parametric form for $\theta_1,\ldots,\theta_m$.

We assume that the number of segments $Z\geq 1$ follows a geometric distribution with parameter ρ, $(0<\rho\leq 1)$. The choice of a geometric distribution is motivated by computational considerations: the memoryless property allows a speedup of a factor n (the number of alignment columns). Given Z, we assume that the indel rate of each segment is independently and identically sampled from one of the m distinct indel rates $\theta_1,\ldots,\theta_m$. We denote the prior probabilities of each of the possible m categories as $\omega=(\omega_1,\cdots,\omega_m)$, $\sum_{j=1}^{m}\omega_j=1$. For each segment $i\in\{1,2,\ldots,Z\}$, we introduce a random variable R_i

indicating the rate category sampled for segment i:

$$\Pr(R_i = j) = \omega_j, \quad i = 1, 2, \ldots, Z \text{ and } j = 1, 2, \ldots, m.$$

Now that the sampling process for the segment-specific rate categories has been described, we can complete the description of the GeoPIP model by defining how the data are generated in each segment. This is done by using the PIP model to sample the data in each segment i independently using the indel parameter θ_{R_i} corresponding to the rate category associated with segment i. We assume a shared substitution rate matrix $\mathbf{Q}$ for substitution, with stationary distribution π in this paper.

To summarize, we obtain the following generative description of the GeoPIP model:

$$\begin{aligned}
Z &\sim \text{Geo}(\cdot|\rho) \\
R_i &\sim \text{Cat}(\cdot|\omega) \qquad i = 1, 2, \ldots, Z \\
\mathbf{s}_i|R_i &\sim \text{PIP}(\cdot|\theta_{R_i}, \tau) \quad i = 1, 2, \ldots, Z \\
\beta &= (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_Z), \\
\mathbf{x} &= \mathbf{x}(\beta) := \mathbf{s}_1 \circ \mathbf{s}_2 \circ \cdots \circ \mathbf{s}_Z,
\end{aligned}$$

where Geo and Cat are the geometric and categorical distributions, and "$\circ$" denotes concatenation of multiple sequence alignments. This gives us the following probability mass function of the GeoPIP model:

$$\text{GeoPIP}(\beta, \mathbf{r}|\gamma) = \text{GeoPIP}(\beta, \mathbf{r}|\theta, \tau, \rho, \omega) = (1-\rho)^{|\beta|-1}\rho \prod_{i=1}^{|\beta|} \omega_{r_i} \, \text{PIP}(\mathbf{s}_i|\theta_{r_i}, \tau), \tag{3}$$

where $\gamma = (\theta, \tau, \rho, \omega)$ denotes all the parameters involved, $\mathbf{R} = (R_1, R_2, \cdots, R_Z)$ are random variables that indicate the rate category for each segment, $\mathbf{r} = (r_1, r_2, \cdots, r_Z)$ is a

realization of $\mathbf{R}$, and $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$ are the m distinct indel rates.

The motivation behind this construction is that the GeoPIP model inherits the desirable properties of the PIP model. We start with a simple result to illustrate this:

Proposition 1 *For all $\mu > 0, \lambda > 0$, the GeoPIP model is explosion free (i.e., the expected sequence length is finite). Moreover, when the substitution rate matrix is reversible, the GeoPIP model is reversible. Its stationary length distribution has mean $(1/\rho)\sum_{j=1}^{m} \omega_j \lambda_j / \mu_j$ and a probability generating function given by*

$$\left(\left[\sum_{j=1}^{m} \omega_j \exp\{(s-1)\lambda_j/\mu_j\} \right]^{-1} - (1-\rho) \right)^{-1} \rho.$$

In particular, Proposition 1 means that the GeoPIP model can capture richer sequence length distributions than previous indel models. For example, the PIP model has a Poisson stationary length distribution, and therefore an equal mean and variance. In contrast, the GeoPIP model can capture the overdispersion found in real data because the distribution of the sequence length based on the GeoPIP model is a mixture of Poisson distributed random variables and thus has an unequal mean and variance. The TKF91 model has a stationary length distribution that is even more problematic, predicting a geometrically distributed stationary sequence length, which is undesirable because that probability mass function has its mode on the empty sequence (Zhang 2000; Miklós 2003). We emphasize that the GeoPIP model does *not* have this deficiency. The geometric reference in its name refers to the PIP mixing distribution, not the stationary length distribution. The most important property of the GeoPIP model, however, is its amenability to efficient phylogenetic inference, which we describe in detail in the next section.

Efficient phylogenetic inference with the GeoPIP

MODEL

Computational complexity is a key issue in phylogenetic inference. Approximation algorithms are proposed in order to explore the space of trees in practice, either using local search (Li *et al.* 2000; Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003; Barker 2004; Stamatakis 2005), or incrementally (Saitou and Nei 1987; Studier *et al.* 1988; Gascuel 1997; Bouchard-Côté *et al.* 2012). Given the large literature on phylogenetic inference, our goal is to show that our model can be incorporated into most existing phylogenetic inference frameworks with minimal changes. In the following, we view θ, ρ, ω as fixed for simplicity but discuss how they are jointly estimated in Appendix 1.

At the core of most modern phylogenetic inference methods is a likelihood function taking a phylogeny as an input, $\ell(\tau)$. Maximum likelihood methods optimize $\ell(\tau)$; Bayesian methods combine $\ell(\tau)$ with a prior and approximate the posterior via Markov chain Monte Carlo (MCMC) methods; and neighbor-joining (NJ) methods break the likelihood $\ell(\tau)$ optimization into many small problems, one for each pair of leaves $\{k_1, k_2\}$—these smaller problems can be viewed as optimization of a likelihood function over a two-leaf tree, $\ell(\tau_{\{k_1,k_2\}})$. In all these cases, the tree inference method usually views the evolutionary model as a black box function $\ell(\tau)$. Since this black box is evaluated at several putative trees, it is important to have efficient evaluation algorithms for calculating $\ell(\tau)$.

If the segmentation β^* and indel rate categories $\mathbf{r}^*$ were known, we could simply pick

$$\ell(\tau; \beta^*, \mathbf{r}^*) = \text{GeoPIP}(\beta^*, \mathbf{r}^* | \gamma).$$

Efficient evaluation in this case is a direct corollary of Section 3 from Bouchard-Côté and Jordan (2013):

Proposition 2 *Computing* $\text{GeoPIP}(\beta^*, \mathbf{r}^* | \gamma)$ *can be done in time* $O(Nn)$*, where* N *is the number of taxa, and* n *is the number of alignment columns.*

Importantly, this running time is of the same order as that of computing the likelihood of a substitution-only model.

Naturally, we need to take into account the fact that a true segmentation is not known in practice (and the notion of a "true" segmentation is only imperfectly applicable in real datasets). The most natural approach to address this issue is to marginalize over the space of segmentations compatible with the data $\mathbf{x}$:

$$\ell^{\Sigma}(\tau) = \sum_{\beta:\mathbf{x}(\beta)=\mathbf{x}} \sum_{r_1=1}^{m} \cdots \sum_{r_{|\beta|}=1}^{m} \mathrm{GeoPIP}(\beta, \mathbf{r}|\gamma).$$

However, in the following we use a different but closely related objective, given by:

$$\ell(\tau) = \max_{\beta:\mathbf{x}(\beta)=\mathbf{x}} \max_{r_1} \ldots \max_{r_{|\beta|}} \mathrm{GeoPIP}(\beta, \mathbf{r}|\gamma).$$

This second objective is motivated by a penalized likelihood approach. In this view, since the segmentation parameter is a combinatorial structure, standard regularization such as L_2 is not appropriate. Instead, our regularization is based on the probability model in Equation (3), where after taking the logarithm, the terms

$$(|\beta| - 1)\log(1-\rho) + \log\rho + \sum_{i=1}^{|\beta|} \log\omega_{r_i}$$

act as a penalty on segmentations that use a large number of blocks or rare indel categories.

The summation problem, $\ell^{\Sigma}(\tau)$, and the maximization problem, $\ell(\tau)$, can both be computed efficiently using dynamic programming. However, the algorithm is markedly simpler in the maximization case. In the summation case, the additional complexity stems from the fact that the set over which we sum, $\{\beta : \mathbf{x}(\beta) = \mathbf{x}\}$, is countably infinite, as segmentations with empty blocks need to be considered in the sum. To reduce the problem to a finite sum problem, an approach analogous to the one described in Supplementary

Information Section 2 of Bouchard-Côté and Jordan (2013) could be used, after which the two dynamic programming algorithms are similar, but we leave this to future work and describe the maximization algorithm in the following. In the maximization case, segmentation with empty blocks can trivially be ignored since the geometric probability mass function is strictly decreasing in $|\beta|$, so adding an empty segment can only reduce the probability of the data under the GeoPIP model.

Proposition 3 *Computing $\ell(\tau)$ can be done in time $O(mn^2 + Nn)$, where N is the number of taxa, n is the number of alignment columns, and m is the number of indel rate categories.*

We now describe an algorithm achieving this running time. First, as a preprocessing step, we calculate:

$$p_{i,j} = \Pr(\mathbf{c}_i|\theta_j, \tau), \quad i = 1, 2, \ldots, n, j = 1, 2, \ldots, m, \tag{4}$$

which is the probability of observing a single MSA column $\mathbf{c}_i$ with indel rate $\theta_j = (\lambda_j, \mu_j)$ on a tree τ. Second, we calculate

$$m_{k,j} = \psi(z_j, k, \theta_j, \tau), \quad k = 1, 2, \ldots, n; \;\; j = 1, 2, \ldots, m.$$
$$z_j = \Pr(\mathbf{c}_\emptyset|\theta_j, \tau), \quad j = 1, 2, \ldots, m.$$

which is used to calculate the factor in the PIP density determined by the length of the MSA segment. Here ψ is defined in Equation (2).

To calculate $m_{k,j}$ efficiently, we use the following recursion:

$$\log m_{k+1,j} = \log m_{k,j} - \log(k+1) + \log(\|\nu_j\|) \text{ for } k = 1, 2, \ldots, n-1,$$

where $\|\nu_j\| = \|\nu_{\theta_j,\tau}\| = \lambda_j(\|\tau\| + 1/\mu_j)$. The recursion is initialized with:

$$\log m_{1,j} = \log \|\nu_j\| + (\Pr(\mathbf{c}_\emptyset|\theta_j, \tau) - 1)\|\nu_j\|,$$

for all $j = 1, 2, \ldots, m$. Using this recursive formula for $m_{k,j}$ and the recursions described in Bouchard-Côté and Jordan (2013) for $p_{i,j}$, the computational cost for calculating all $p_{i,j}$ and $m_{k,j}$ is $O(nm)$.

Let l_i denote the maximum likelihood over all possible segmentations for the first i MSA columns $\mathbf{c}_{1:i} = (\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_i)$ $(1 \leq i \leq n)$. We set $l_0 = 1$ and start with $\mathbf{c}_{1:1}$. There are m possible choices for the rate assigned to this single column, yielding

$$l_1 = \max\left\{p_{1,j}\ m_{1,j}\ \omega_j\rho : j \in \{1, 2, \ldots, m\}\right\}.$$

The computational cost of this step is $\mathrm{O}(m)$. We calculate an intermediate quantity l_t based on $l_0, l_1, \ldots, l_{t-1}$ recursively. To do so, we define a $t \times m$ matrix $\mathbf{L}^{(t)}$ with entry (i, j) given by:

$$l_{i,j}^{(t)} = l_{i-1}\ p_{i,j}\ p_{i+1,j}\ \cdots\ p_{t,j}\ m_{t-i+1,j}\ \omega_j(1-\rho), \quad i \in \{1, \cdots, t\}, j \in \{1, \cdots, m\},$$

where $l_{i,j}^{(t)}$ represents the largest likelihood if the t-th column forms a segment with the last $t - i$ columns with the j-th indel rate, conditioning on knowing the first t columns only (i.e., no information on the columns $\{t+1, \ldots, n\}$). Therefore, the matrix $\mathbf{L}^{(t)}$ considers all possible segmentation choices for the i-th column, and utilizes previously calculated maximum likelihood for the segmentation choices of the first $t - 1$ columns to calculate the largest likelihood for all $t \times m$ possible segmentation choices when the t-th column is added

to the first $t-1$ columns. Then we compute

$$l_t = \max\left\{ l_{i,j}^{(t)} : i \in \{1, 2, \cdots, t\}, j \in \{1, 2, \cdots, m\} \right\}, \tag{5}$$

The largest value of $\mathbf{L}^{(t)}$ gives the maximum likelihood l_t of all possible segmentations and indel rate assignments of the first t columns.

The computational cost of naively calculating l_{t+1} is $\mathrm{O}(t^2 m)$. However, we notice that part of the product $p_{i,j} p_{i+1,j} \cdots p_{t,j}$ in $l_{i,j}^{(t)}$ can be stored and used to calculate part of product $p_{i-1,j} p_{i,j} \cdots p_{t,j}$ in $l_{i-1,j}^{(t)}$, so the computational cost can be reduced to $\mathrm{O}(tm)$. As a result, the computational cost of calculating all of $\{l_0, l_1, \ldots, l_n\}$ is $\mathrm{O}(\sum_{t=1}^{n} tm) = \mathrm{O}(n^2 m)$.

Hierarchical Poisson indel process

We also developed a more elaborate generalization of the PIP model that incorporates long indels. We use this more elaborate process, called the Hierarchical Poisson indel process (hPIP), as an additional mechanism to generate synthetic data that we then analyze using the simpler GeoPIP model. While it is easy to generate data using the hPIP model, it is not computationally tractable to perform tree inference. See Appendix 2 for more details on the hPIP model. As with the TKF92 model, the hPIP model allows long indels but in a manner that does not cover all types of long indels expected in a biologically realistic process (in both cases, there cannot be an overlapping long insertion and long deletion, for example).

Simulation studies

This section is organized as follows. First, we perform a simulation study to investigate the accuracy of our segmentation inference method, given the correct

alignment. Second, we perform simulation studies to assess the accuracy of the complete inference algorithm for the GeoPIP model in finding the true tree when the evolutionary model is correctly specified (i.e., data are simulated using the GeoPIP model, and the true alignment is given) and misspecified (e.g., data are simulated using the software INDELible (Fletcher and Yang 2009) or the hPIP model, and an estimated alignment is used). We compare inference results with a set of widely used phylogenetic inference methods.

Segmentation

We consider three sets of indel rates in the simulations. In the first scenario, we consider two indel rate categories, deletion rates $\mu_1 = 0.02$ and $\mu_2 = 2.0$, insertion rates $\lambda_j = 20 \cdot \mu_j$ $(j = 1, 2)$ and multinomial parameter for the stationary distribution of segments $\omega = (1/2, 1/2)$. In the second scenario, we set $m = 3$, $\mu_1 = 0.02$, $\mu_2 = 0.2$ and $\mu_3 = 2.0$, $\lambda_j = 20 \cdot \mu_j$ $(j = 1, 2, 3)$, and $\omega = (1/3, 1/3, 1/3)$. In the third scenario, we set $m = 4$, $\mu_1 = 0.01$, $\mu_2 = 0.1$, $\mu_3 = 1.0$ and $\mu_4 = 5.0$, $\lambda_j = 20 \cdot \mu_j$ $(j = 1, 2, 3, 4)$, and $\omega = (1/4, 1/4, 1/4, 1/4)$. The geometric parameter for the number of segments is $\rho = 0.05$ in all scenarios. A perfect binary tree with 32 leaves is used in this simulation study. All edge lengths are set to be 0.1.

In each simulation run, we generate the MSAs randomly using the GeoPIP model proposed in this paper. To focus on the accuracy of the segmentation inference method, we fix the tree τ, rate matrix $\mathbf{Q}$, indel rates θ, and the GeoPIP model parameters ρ and ω as true values. Instead of generating a geometric-distributed number of segments, we generate 20 segments at the root of the tree in all runs so that the lengths of MSA columns are less variable across simulation runs.

To measure the accuracy of the segmentation algorithm, we calculate the proportion of alignment columns being identified with incorrect rates. Since each alignment column belongs to exactly one segment and thus is associated with exactly one indel rate, we define

segmentation error as the percentage of alignment columns in the estimated segmentation which have a different indel rate than that of the true segmentation.

We vary the number of sequences used for segmentation inference (using 2, 4, 8, 16 or 32 sequences), and evaluate the segmentation error on MSA columns that are non-empty for the smallest set of sequences (i.e., 2 sequences), to make the absolute magnitude of the errors comparable when varying the number of sequences.

We observe a dramatic decrease in error rate when the number of sequences used for segmentation inference increases (Table 1). This decrease in error motivates the need for marginalization of internal sequences: the fact that the GeoPIP model allows such marginalization in a simple and exact fashion allows us to efficiently search over segmentations, even when the number of sequences increases.

Table 1: Simulation results on segmentation error and running time.

	segmentation error			running time (in seconds)		
Sequences	$m = 2$	$m = 3$	$m = 4$	$m = 2$	$m = 3$	$m = 4$
1-2	0.0276	0.2064	0.2219	3.0593	3.7787	2.9613
1-4	0.0064	0.1226	0.1180	5.6851	6.8666	6.3316
1-8	0.0035	0.0804	0.0899	14.9626	18.6748	19.0748
1-16	0.0011	0.0307	0.0437	44.4989	55.6419	61.4356
1-32	0.0011	0.0391	0.0397	142.829	169.2812	199.2880

Data are simulated based on the geometric Poisson indel process (GeoPIP) model with 2, 3, or 4 indel rates (m), on a perfect binary tree with 32 leaves. Average percentages of alignment columns with incorrectly inferred indel rates from 100 simulations are listed.

Well-specified synthetic examples

In this section, we perform simulation studies to assess tree reconstruction accuracy when the data are simulated according to the GeoPIP model. In this case, the the GeoPIP models and substitution-only models are both well-specified Truszkowski and Goldman

(2016). Our focus is on the effect of the additional information brought by the indels on tree reconstruction accuracy. To make the reconstruction accuracies more interpretable, we also include the accuracy of reconstructions from PhyML (Guindon *et al.* 2010), and from a standard PIP model.

Simulation setup.—

We set the number of indel categories $m = 2$ and the indel rate $(\lambda_1, \mu_1) = (0.4, 0.02)$ for the first segment. For the second segment, we consider three sets of indel rates, $(\lambda_2, \mu_2) = (10, 0.5), (40, 2.0)$, or $(80, 4.0)$. Note that when $(\lambda_2, \mu_2) = (80, 4.0)$, the data simulated using the GeoPIP model have fast-evolving regions making the synthetic alignments visually most similar to real datasets. We consider two phylogenetic trees in the simulation: a phylogenetic tree with 8 leaves and varying branch lengths and a perfect binary phylogenetic tree with 16 leaves and constant branch lengths (see Figure 1).

We focus on indel rate variation and ignore substitution rate variation for simplicity, but we note that substitution rate variation can be incorporated into our methods without technical difficulty. For the first set of simulations on the tree with 8 leaves, the estimated rate matrix $\widehat{\mathbf{Q}}$ from PhyML is used as starting values for CTMC+NJ, PIP+NJ and GeoPIP+NJ estimation algorithms and then updated iteratively together with other parameters. For the second set of simulations on the tree with 16 leaves, we fix the rate matrix $\widehat{\mathbf{Q}}$ in the CTMC+NJ, GeoPIP+NJ and PIP+NJ methods as the estimated rate matrix obtained from PhyML, so that the rate matrix is the same across all methods considered.

For the PIP results, we randomly generate a deletion rate $\mu \sim U(0, 1)$ and set $\lambda = \mu\eta$ as a starting value, where η is set as the total number of observed alignment columns. We use the true value $m = 2$ in the results based on the GeoPIP model. Since our iterative optimization algorithm requires a set of starting values for the indel rates θ,

the multivariate parameter ω, and the segmentation, we show two sets of results, one using the true values as initialization, and one using random values. For the random starting values, we randomly generate two deletion rates $\mu_1 \sim U(0,1)$ and $\mu_2 \sim U(1,2)$, then set $\lambda_i = \mu_i \eta$ $(i = 1, 2)$. We set $\eta = 20$ in all simulations. The choice of η is related to the minimum number of alignment columns in one segment. Similar results are observed when $\eta = 10$ is used instead of $\eta = 20$. We set starting values $p_s = 0.1$, and $\omega_s = (1/m, 1/m) = (0.5, 0.5)$. Again, we found that different choices of starting values p_s and ω_s did not markedly affect the inference results in our simulations studies. We simply set the initial segmentation as one segment containing all MSA columns.

Simulation results.—

We calculate the Robinson-Foulds (RF) and the weighted Robinson-Foulds (wRF) distance (Robinson and Foulds 1979; Felsenstein 2004) between each estimated unrooted tree and the true unrooted tree from 100 simulation runs. The RF and wRF distances are calculated using the Python package `dendropy` (Sukumaran and Holder 2010).

Table 2: Results on synthetic data simulated from the GeoPIP model on a phylogenetic tree of 8 leaves with varying branch lengths (see Figure 1a).

		wRF (unscaled trees)		wRF (scaled trees)		RF
Parameter	Method	mean (s.e.)	median	mean (s.e.)	median	mean
	PhyML	0.200 (0.006)	0.190	0.187 (0.006)	0.179	0.10 (0.07)
	CTMC+NJ	0.213 (0.005)	0.203	0.200 (0.005)	0.192	0.12 (0.06)
μ_2=0.5	PIP+NJ	0.150 (0.003)	0.144	0.137 (0.003)	0.136	0
	GeoPIP+NJ (true init.)	0.153 (0.003)	0.151	0.139 (0.003)	0.139	0
	GeoPIP+NJ (random init.)	0.153 (0.003)	0.151	0.139 (0.003)	0.138	0
	PhyML	0.222 (0.006)	0.208	0.208 (0.006)	0.199	0.24 (0.08)
	CTMC+NJ	0.240 (0.006)	0.227	0.223 (0.005)	0.218	0.30 (0.07)
μ_2=2.0	PIP+NJ	0.144 (0.003)	0.144	0.130 (0.003)	0.128	0
	GeoPIP+NJ (true init.)	0.134 (0.004)	0.130	0.116 (0.003)	0.115	0
	GeoPIP+NJ (random init.)	0.134 (0.004)	0.130	0.116 (0.003)	0.115	0
	PhyML	0.216 (0.007)	0.203	0.207 (0.006)	0.196	0.20 (0.06)
	CTMC+NJ	0.231 (0.007)	0.226	0.219 (0.006)	0.212	0.28 (0.07)
μ_2=4.0	PIP+NJ	0.203 (0.003)	0.203	0.201 (0.003)	0.203	0
	GeoPIP+NJ (true init.)	0.124 (0.003)	0.116	0.107 (0.002)	0.103	0
	GeoPIP+NJ (random init.)	0.124 (0.003)	0.116	0.107 (0.002)	0.105	0

All models are well-specified, except for the standard Poisson indel process (PIP). The weighted Robinson-Foulds (wRF) distances and the Robinson-Foulds (RF) distance of 100 simulation runs are summarized. For the "scaled tree" columns, we scale the total branch length of all estimated trees and the true tree to be equal to one.

The main comparison of interest is between the GeoPIP+NJ method and the CTMC+NJ method. Both models are well-specified here, but only the former uses indels. Our results show that the GeoPIP+NJ method reduces reconstruction error by a factor of up to two (Table 2 and Table 3) in terms of the wRF distance, and the GeoPIP+NJ method always outperforms CTMC+NJ in terms of the RF distance as well. Reconstructions based on the standard PIP model also outperform reconstructions solely based on substitutions, but by a much smaller margin.

Table 3: Simulation results on synthetic data generated from the GeoPIP model.

		wRF (unscaled trees)		wRF (scaled trees)		RF
Parameter	Method	mean (s.e.)	median	mean (s.e.)	median	mean (s.e.)
	PhyML	0.584 (0.013)	0.567	0.375 (0.008)	0.367	1.18 (0.17)
$\mu_2 = 4.0$	CTMC+NJ	0.660 (0.016)	0.651	0.424 (0.009)	0.414	1.82 (0.21)
$b = 0.05$	PIP+NJ	0.315 (0.004)	0.309	0.210 (0.003)	0.208	0
	GeoPIP+NJ	0.317 (0.007)	0.308	0.194 (0.004)	0.192	0
	PhyML	1.161 (0.038)	1.073	0.372 (0.011)	0.345	1.54 (0.20)
$\mu_2 = 4.0$	CTMC+NJ	30.19 (28.76)	1.236	0.422 (0.019)	0.387	2.20 (0.33)
$b = 0.1$	PIP+NJ	0.854 (0.011)	0.854	0.319 (0.004)	0.319	0
	GeoPIP+NJ	0.686 (0.016)	0.675	0.211 (0.004)	0.208	0.12 (0.05)
	PhyML	2.772 (0.094)	2.604	0.464 (0.019)	0.421	3.82 (0.43)
$\mu_2 = 4.0$	CTMC+NJ	31.80 (14.52)	3.203	0.658 (0.040)	0.505	5.44 (0.57)
$b = 0.2$	PIP+NJ	2.837 (0.035)	2.805	0.529 (0.005)	0.535	0.04 (0.03)
	GeoPIP+NJ	2.043 (0.054)	2.003	0.314 (0.007)	0.302	0.86 (0.13)
	PhyML	0.511 (0.010)	0.497	0.333 (0.006)	0.326	0.72 (0.12)
$\mu_2 = 0.5$	CTMC+NJ	0.569 (0.013)	0.547	0.371 (0.008)	0.361	1.28 (0.18)
$b = 0.05$	PIP+NJ	0.345 (0.006)	0.341	0.227 (0.004)	0.219	0
	GeoPIP+NJ	0.340 (0.008)	0.338	0.217 (0.005)	0.214	0.02 (0.02)
	PhyML	1.053 (0.037)	0.920	0.344 (0.014)	0.297	1.30 (0.30)
$\mu_2 = 0.5$	CTMC+NJ	15.42 (14.23)	1.068	0.378 (0.020)	0.338	1.78 (0.36)
$b = 0.1$	PIP+NJ	0.740 (0.023)	0.740	0.258 (0.007)	0.253	0
	GeoPIP+NJ	0.669 (0.022)	0.624	0.205 (0.004)	0.196	0.06 (0.03)
	PhyML	2.800 (0.437)	2.236	0.406 (0.019)	0.367	2.74 (0.34)
$\mu_2 = 0.5$	CTMC+NJ	37.14 (16.22)	2.794	0.643 (0.045)	0.461	5.18 (0.56)
$b = 0.2$	PIP+NJ	1.954 (0.092)	2.229	0.353 (0.018)	0.311	0
	GeoPIP+NJ	1.536 (0.063)	1.367	0.238 (0.008)	0.218	0.40 (0.08)

The true tree is a perfect binary tree of 16 leaves with the same branch length b for all branches (see Figure 1b). Different indel rates (i.e., μ_2) and different phylogenetic tree branch lengths (i.e., b) are considered. The weighted Robinson-Foulds (wRF) distances and the Robinson-Foulds (RF) distance of 100 simulation runs are summarized.

As a reference, we also include results obtained using PhyML, which uses a statistically superior tree estimation method (compared to NJ) (Roch 2010), and a well-specified model, but no indel information. Comparing PhyML and CTMC+NJ illustrates the discrepancy introduced by the slightly suboptimal NJ estimator. The accuracy gains obtained by modelling indel rate heterogeneity are larger than those obtained by using a more sophisticated tree estimation method under the simulation setups

we considered.

Table 2 also shows that the difference between initializing the GeoPIP model parameters with true values versus random values is negligible, supporting the robustness of our estimation procedure. In Table 3 and following tables, we show only the GeoPIP+NJ results with random initial values. The average running times of 100 simulations runs on the phylogenetic tree with 8 leaves are: 4.03 seconds for PhyML, 12.76 seconds for CTMC+NJ, 124.88 seconds for the PIP+NJ method, 182.46 seconds for the GeoPIP+NJ method (true initialization), and 234.51 seconds for the GeoPIP+NJ method (random initialization). The GeoPIP+NJ method is currently implemented in Python and it is not optimized for computation speed. The running times are provided as a general reference on methods implemented in the same languages (i.e., GeoPIP+NJ and PIP+NJ) and are not meaning for benchmarking the performance of methods implemented in different languages (for example, PhyML).

Misspecified synthetic examples from the hPIP model

In real applications, the substitution and indel processes are unknown. The gaps in MSAs may also be caused by long indels which are not directly captured by the GeoPIP model. The hPIP model can be viewed as a more realistic model since it explicitly incorporates long indel events. This motivates the experiments presented in this section, where we simulate data from the hPIP model, and show that tree reconstructions based on the GeoPIP model are still superior.

Simulation setup.—

We use the same evolutionary parameters as in the previous section for the phylogenetic tree with 8 leaves and set $(\lambda_2, \mu_2) = (80, 4.0)$. For the hPIP model, we set the segment insertion rate to $\lambda_{seg} = 2$ and the segment deletion rate to $\mu_{seg} = 0.1$ (see Appendix 2: Hierarchical Poisson Indel Process).

We estimate the phylogenetic tree using several tree inference methods and models. For the GeoPIP models, we use $m = 3$ and $m = 5$ as the numbers of indel rate categories. These two variants of the GeoPIP model are denoted by GeoPIP3 and GeoPIP5. Even though two indel rates are used in the hPIP simulation model, there is no "true" value in this setup for m in the GeoPIP model, since additional rate categories can be recruited as surrogates to long indels. Therefore, both the PIP model and the GeoPIP model are misspecified in this simulation study. The CTMC+NJ and PhyML are still correctly specified since they utilize only substitutions Truszkowski and Goldman (2016). Starting values for the PIP and GeoPIP estimators are randomly generated in the same way as in the previous section.

Simulation results.—

Both the GeoPIP+NJ and the PIP+NJ methods are based on misspecified models in this case, as neither capture long indels directly. However, Table 4 shows that the GeoPIP+NJ method provides a better approximation of the long indels introduced by the hPIP model, by assigning regions with possible long indels a larger indel rate. The GeoPIP+NJ method also compares favorably against models that use substitution only, which are still well-specified, but use only a subset of the data. At the same time, the region with long indel (dyed as dark gray in Figure 2) is perfectly identified by our inference method based on the GeoPIP model.

Table 4: Simulation results when the true model is the hierarchical Poisson indel process (hPIP).

	wRF (unscaled trees)		wRF (scaled trees)		RF
Method	mean (s.e.)	median	mean (s.e.)	median	mean
PhyML	0.232 (0.008)	0.224	0.215 (0.007)	0.209	0.20 (0.08)
CTMC+NJ	0.249 (0.007)	0.242	0.237 (0.007)	0.236	0.44 (0.09)
PIP+NJ	0.219 (0.004)	0.216	0.210 (0.005)	0.204	0
GeoPIP3+NJ	0.172 (0.006)	0.156	0.151 (0.005)	0.147	0.02 (0.02)
GeoPIP5+NJ	0.172 (0.006)	0.157	0.153 (0.006)	0.147	0.02 (0.02)

The true tree has 8 leaves with varying branch lengths (see Figure 1a). The PIP and GeoPIP models are misspecified, while the other, substitution-only methods are well-specified. Both wRF and RF are reported.

The average running times of 100 simulation runs are: 4.08 seconds for PhyML, 12.81 seconds for CTMC+NJ, 145.00 seconds for PIP+NJ, 304.15 seconds for the GeoPIP3+NJ method, and 270.71 seconds for the GeoPIP5+NJ method.

Misspecified synthetic examples using software INDELible and MUSCLE

We consider generating data using other popular indel models. We use the software INDELible to generate data in this section. INDELible provides several options for both the indel model and the substitution model, and it also allows data to be generated in blocks with different indel models and substitution models.

When data were generated using INDELible, the GeoPIP+NJ method utilizes both indels and substitutions to reconstruct the phylogenetic tree, but the indel model is misspecified, while the CTMC+NJ method utilizes only substitutions which are correctly specified. Therefore, the comparison of results from GeoPIP+NJ and results from CTMC+NJ illustrates the potential gain or loss of modelling indels using a misspecified indel model in real applications.

In a real application, the multiple sequence alignment is usually unknown. We use MUSCLE to obtain an alignment, then use this alignment for inference. We compare

results obtained using the MUSCLE estimated alignment with the results obtained using the true alignment generated by INDELible. MUSCLE does not require an input tree to estimate the alignment, so it can be used to obtain an estimated alignment before running our inference method when the alignment is unknown.

Simulation setup.—

We simulate data on a perfect binary tree with 16 leaves and branch length $b = 0.05$ for all branches using INDELible. The total branch length for this tree is 1.5. We consider two simulation scenarios. First, we simulate two blocks with the same indel length distribution but different indel rates: indel length distribution is set as a negative binomial with parameter $r = 1$ and $p = 0.1$ and the indel rate is set as 0.05 and 0.25 (same insertion and deletion rate within each block). The initial length is set to be 50 for both blocks. Second, we simulate three blocks with different indel length distributions and different indel rates: a negative binomial indel length distribution with parameter $r = 1$ and $p = 0.1$, no indels for the second block and a power law indel length distribution (Fletcher and Yang 2009) with parameter 1.7 and maximum length 30. The indel rate is 0.2 for the first block and 0.05 for the third block. The initial length is set to be 30 for all three blocks.

Simulation results.—

Table 5 shows that for the first simulation scenario, GeoPIP5+NJ and PIP+NJ outperform CTMC+NJ and PhyML in terms of the RF and the wRF of the scaled trees, on both the true alignment and the MUSCLE alignment. The GeoPIP5+NJ and PIP+NJ methods also outperform CTMC+NJ and PhyML in terms of the wRF of the unscaled trees on the true alignment, but not on the MUSCLE alignment. For the second simulation scenario, GeoPIP5+NJ and PIP+NJ outperform CTMC+NJ (but not PhyML) in terms of RF, but not in terms of wRF.

Table 5: Simulation results on synthetic data generated from the software INDELible and aligned by the software MUSCLE.

		wRF (unscaled trees)		wRF (scaled trees)		RF
Parameter	Method	mean (s.e.)	median	mean (s.e.)	median	mean
true alignment NB+NB	PhyML	0.612 (0.009)	0.614	0.400 (0.006)	0.397	1.06 (0.14)
	CTMC+NJ	0.653 (0.010)	0.664	0.427 (0.007)	0.423	1.40 (0.16)
	PIP+NJ	0.544 (0.008)	0.547	0.356 (0.006)	0.360	0.38 (0.10)
	GeoPIP5+NJ	0.548 (0.009)	0.550	0.358 (0.006)	0.364	0.40 (0.10)
MUSCLE alignment NB+NB	PhyML	1.301 (0.017)	1.306	0.433 (0.008)	0.419	1.68 (0.20)
	CTMC+NJ	1.349 (0.016)	1.355	0.442 (0.007)	0.442	1.86 (0.18)
	PIP+NJ	1.384 (0.014)	1.390	0.403 (0.007)	0.403	1.26 (0.15)
	GeoPIP5+NJ	1.349 (0.014)	1.357	0.408 (0.007)	0.405	1.32 (0.17)
true alignment NB+SUB+POW	PhyML	0.653 (0.011)	0.641	0.426 (0.007)	0.422	1.24 (0.16)
	CTMC+NJ	0.681 (0.011)	0.670	0.443 (0.007)	0.441	1.68 (0.17)
	PIP+NJ	0.724 (0.013)	0.719	0.472 (0.008)	0.472	1.02 (0.15)
	GeoPIP5+NJ	0.724 (0.014)	0.712	0.468 (0.008)	0.462	1.08 (0.15)
MUSCLE alignment NB+SUB+POW	PhyML	1.393 (0.015)	1.393	0.449 (0.008)	0.436	1.74 (0.18)
	CTMC+NJ	1.432 (0.015)	1.426	0.459 (0.007)	0.445	2.24 (0.21)
	PIP+NJ	1.589 (0.020)	1.585	0.471 (0.009)	0.458	1.88 (0.22)
	GeoPIP5+NJ	1.549 (0.020)	1.523	0.479 (0.010)	0.466	2.18 (0.23)

The true tree is a perfect binary tree of 16 leaves with the same branch length $b = 0.05$ for all branches (see Figure 1b). The true alignment generated using INDELible and the estimated alignment using the software MUSCLE are both considered. In this table, NB+NB indicates that the data are generated using two blocks with the same indel length model (negative binomial with parameter 1 and 0.1) but different indel rates (0.05 and 0.25 respectively), NB+SUB+POW indicates that the data are generated using three blocks with different indel length models (a negative binomial distribution with parameter 1 and 0.1, a substitution model with no indels, and a power law distribution with parameter 1.7 and maximum 30), and different indel rates (0.2 for the negative binomial block and 0.1 for the power law block).

The results show that even when the indel model is misspecified, the GeoPIP5+NJ method may still achieve a more accurate phylogenetic tree estimate, compared to the correctly-specified model CTMC+NJ that relies on the substitution only. The improvement in accuracy may depend on the true indel models. When the true alignment is not available, using the MUSCLE alignment provides an alternative to apply the GeoPIP5+NJ method which requires a fixed alignment.

On the other hand, PhyML always outperforms CTMC+NJ in all scenarios, which

shows the benefits of the likelihood approach versus the NJ approach in general, and the magnitude of potential improvement if the GeoPIP model is incorporated into a full likelihood inference approach in future work. At the same time, the comparison between the results using the true alignment and the MUSCLE alignment shows the potential gain in accuracy if the GeoPIP model is incorporated into a joint inference of phylogenetic tree and alignment for future work. Because exact boundaries of segments may not be easy to identify, our inference method based on the GeoPIP model does not always separate segments generated by different rules (NB, SUB and POW). However, the region with long indel (dyed as dark gray in Figure 3) is still perfectly identified by our method.

Data analysis

In this section, we apply our methods to a real data set. We compare results obtained using our methods and other tree reconstruction methods, and show some examples of inferred segmentations in real alignments.

Molluscs are a diverse group of well studied animals, but many phylogenetic relationships among molluscan species are still unresolved (Smith *et al.* 2011). Because of the vast diversity within this large group of species, insertions and deletions of nucleotides is prevalent in molluscan ribosomal RNA (rRNA) alignments. Lydeard *et al.* (2000) conducted a comparative analysis of complete mitochondrial large subunit (LSU) rRNA sequences of 10 molluscan species and two outgroups (*L. terrestris* and *D. melanogaster*), and obtained the MSAs of these sequences based on their secondary structure. Smith *et al.* (2011) obtained a different tree for some sub-groups of molluscs, in particular grouping Gastropoda with Bivalvia, instead of Gastropoda with Cephalopoda. A few other hypotheses on sub-grouping of molluscs can also be found in Kocot *et al.* (2011); Smith *et al.* (2011).

We re-analyze the dataset of Lydeard *et al.* (2000) using the following methods: CTMC+NJ, PIP+NJ, GeoPIP+NJ with four indel rates (denoted as GeoPIP4), PhyML with four substitution rates (denoted PhyML4), and BAli-Phy (Suchard and Redelings 2006; Redelings and Suchard 2007), a state-of-the-art Bayesian approach that takes long indels into account to simultaneously estimate both alignment and phylogeny. In the BAli-Phy experiments, we used RS07+GTR (Redelings and Suchard 2007) as the evolutionary model, and 10 000 MCMC iterations (10% burn-in). This data set can be downloaded from `http://www.rna.icmb.utexas.edu/SIM/4D/Mollusk/alignment.gb`. We used reference clades based on the fossil record (Lydeard *et al.* 2000; Smith *et al.* 2011) to assess the quality of the inferred trees. We describe these clades in Table 6.

Table 6: Description of the reference clades used for validation in terms of the species available in the dataset of Lydeard *et al.* (2000).

Reference clade	Constituents
Clausiliidae	*A. turrita* and *A. coerulea*
Helicoidea	*E. herktotsi* and *C. nemoralis*
Herterobranchia	Clausiliidae and Helicoidea
Bivalvia	*P. maximus* and *M. edulis*
Cerithioidea	*P. paludiformis* and *Cac. lacertina*

We ran each method on the full dataset, as well as on the subset excluding the two outgroups. Table 7 reports whether the reference clades were correctly reconstructed for all algorithm and data configurations. Among the three indel methods, both GeoPIP and BAli-Phy reconstruct all the reference clades, while the PIP reconstruction (from data excluding outgroups) fails to reconstruct one of the clades (Bivalvia). This supports that using constant rate, point-indel models can confound phylogenetic tree inference.

Table 7: Comparison of the clades identified by different methods, when the two outgroups are added, and in parentheses, when the outgroups are excluded.

	Substitution-based			Indel-aware		
Reference clade	MP	MP(t.o.)	PhyML4	BAli	PIP	GeoPIP
Clausiliidae	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)
Helicoidea	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)
Herterobranchia	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)
Bivalvia	0 (0)	1 (1)	1 (1)	1 (1)	1 (0)	1 (1)
Cerithioidea	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)

In this table, "1" indicates that the clade has been identified by the corresponding tree inference method (column), and "0" indicates that the clade has not been identified. Maximum-parsimony (MP) trees are taken from Lydeard *et al.* (2000), "MP(t.o.)" stands for MP analysis from transversions only, and "BAli," for BAli-Phy.

Prompted by the observation of Smith *et al.* (2011) that molluscan phylogenetic trees are influenced by the choice of outgroups, we assessed the robustness of each method by measuring the wRF distance and the RF distance between the tree inferred without outgroup and the subtree obtained after exclusion of the two outgroups from the tree inferred from the full dataset. Figure 4 shows that the wRF distance between the two GeoPIP trees is 0.253, which compares favorably to the wRF distance between results from other indel-aware methods. The RF distances tell a different story where the GeoPIP model has the largest value of 4 due to the change of the placement of *K.tunicata*. However, the total branch length *K.tunicata* travels is very small (0.042), which explains why the ΔwRF is small even though ΔRF is 4.

Moreover, one of the two outgroups, *D. melanogaster*, is severely misplaced in the CTMC tree, the PhyML tree, and the BAli-Phy tree. This can be explained by the fact that substitution-only models and some indel models cannot overcome the erroneous attraction due to the similar base compositions of *D. melanogaster* and *L.bleekeri*. To restore correct placement, a pruning and regraft operation would require moving the stem

of that outgroup by a total branch length of 0.861 (four branches) in the PhyML tree and 0.199 (four branches) in the BAli-Phy tree. In contrast, the placement of *D. melanogaster* is greatly improved in both the GeoPIP and PIP trees, requiring moving the stem by a total branch length of 0.012 (one branch) for both the PIP tree and the GeoPIP tree.

Figure 5 shows a subset of an inferred segmentation of the molluscan data. The four estimated deletion rates are $\widehat{\mu}_1 = 0.01$, $\widehat{\mu}_2 = 0.15$, $\widehat{\mu}_3 = 0.42$ and $\widehat{\mu}_4 = 1.41$. Similar results are obtained when 6 indel rates are used instead of 4 indel rates or when $\beta = \lambda_i/\mu_i$ is set to 10 as initial value instead of 20, which shows that the choice of category numbers for indel rates is not critical as long as it is large enough to allow sufficient indel rate variations. The choice of initial segment lengths does not markedly affect the results as long as this choice falls into a reasonable range. The running times are: 33.8 seconds for PhyML, 5.2 minutes for PIP+NJ, 48.9 minutes for GeoPIP+NJ, and 1 day and 3 hours for BAli-Phy (10 000 iterations).

Discussion

With the exception of hand-coded indel characters, mainstream methods for phylogenetic tree reconstruction have been refractory to the incorporation of the indel information present in the sequence data. Our experiments suggest that one potential factor behind this is that single rate point indel models tend to lack robustness when doing phylogenetic tree inference.

We show that a simple model of indel rate variation can restore robustness while improving the quality of the reconstructed phylogenies. The model is simple, both in the sense that its running time is the same as existing pure-substitution reconstruction algorithms, and also that its implementation involves components already present in standard phylogenetic software toolboxes. In particular, a promising direction is to

combine other tree inference methods with the GeoPIP model, for example Bayesian tree reconstruction methods (Li 1996; Mau 1996; Huelsenbeck and Ronquist 2001; Drummond *et al.* 2012). Calculating confidence intervals for indel parameters is not a simple task in our current GeoPIP+NJ framework. For example, the popular bootstrap approach is not directly applicable because resampling alignment columns breaks dependence of neighboring alignment columns, which is key in the GeoPIP model. The Bayesian approach would provide the additional advantage of outputting credible intervals for not only segmentations, but also indel parameters.

Alignment uncertainty is an important related issue. Using a point estimate for the alignment can cause underestimation of tree uncertainty downstream, and alignment errors can confound tree reconstruction (Suchard and Redelings 2006; Redelings and Suchard 2007; Wong *et al.* 2008). To address these issues while still taking indel rate heterogeneity into account, our model could be integrated into a Bayesian or maximum likelihood co-estimation method (Lunter *et al.* 2005a; Suchard and Redelings 2006; Redelings and Suchard 2007; Liu *et al.* 2009b, 2012). Note also that the GeoPIP model could potentially be modified to reduce the confounding effect of incorrect alignment regions, by correlating the indel rate with the substitution rate. The uncertain substitution information coming from high indel intensity regions could be discounted and therefore have a lesser effect on tree inference.

The GeoPIP model assumes a fixed segmentation for the entire phylogenetic tree. However, indel rate heterotachy, which has been measured in certain datasets, for example promoter regions (Taylor *et al.* 2006), can violate this assumption in real datasets. The model could be modified to take indel heterotachy into account, for example by splitting and merging segments at random points of the tree, but at the cost of making inference significantly more complicated. A similar trade-off is found in substitution rate variation modelling, where rate variation assumptions that ignore heterotachy are often preferred as

they are simple and generally effective.

On the other hand, there are ways in which the GeoPIP model can be improved without sacrificing its computational efficiency. For example, it would be simple to make the rate category of one segment depend on the previous rate segment category. This defines a model related to the phylogenetic HMM model used for substitution rate variation (Yang 1995; Felsenstein and Churchill 1996). Correlation of indel and substitution rates (Ananda *et al.* 2011; Jovelin and Cutter 2013) is another interesting future direction to explore. One simple method to model such correlations would be to estimate substitution rate matrices separately for different indel rate regions. The computation cost of rate matrix estimation would only increase by a factor of m (the number of indel rate categories). Source code and scripts of simulation studies can be obtained from `https://github.com/yzhai220/geopip`.

*

References

Ananda, G., Chiaromonte, F., and Makova, K. D. 2011. A genome-wide view of mutation rate co-variation using multivariate analyses. *Genome Biology*, 12(3): R27.

Barker, D. 2004. Lvb: parsimony and simulated annealing in the search for phylogenetic trees. *Bioinformatics*, 20(2): 274–275.

Bouchard-Côté, A. and Jordan, M. I. 2013. Evolutionary inference via the Poisson indel process. *Proceedings of the National Academy of Sciences*, 110(4): 1160–1166.

Bouchard-Côté, A., Jordan, M. I., and Klein, D. 2008. Efficient Inference in Phylogenetic InDel Trees. In *Advances in Neural Information Processing Systems 21 (NIPS)*, volume 21, pages 177–184.

641 Bouchard-Côté, A., Sankararaman, S., and Jordan, M. I. 2012. Phylogenetic Inference via
642 Sequential Monte Carlo. *Systematic Biology*, 61: 579–593.

643 Carvalho, A. B. and Clark, A. G. 1999. Genetic recombination: intron size and natural
644 selection. *Nature*, 401(6751): 344–344.

645 Chen, J.-Q., Wu, Y., Yang, H., Bergelson, J., Kreitman, M., and Tian, D. 2009. Variation
646 in the ratio of nucleotide substitution and indel rates across genomes in mammals and
647 bacteria. *Molecular Biology and Evolution*, 26(7): 1523–1531.

648 Drummond, A., Suchard, M., Xie, D., and Rambaut, A. 2012. Bayesian phylogenetics with
649 BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29: 1969–1973.

650 Edgar, R. C. 2004a. Muscle: a multiple sequence alignment method with reduced time and
651 space complexity. *BMC bioinformatics*, 5(1): 113.

652 Edgar, R. C. 2004b. Muscle: multiple sequence alignment with high accuracy and high
653 throughput. *Nucleic acids research*, 32(5): 1792–1797.

654 Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. *Nature*
655 *Reviews Genetics*, 5(6): 435–445.

656 Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood
657 approach. *Journal of Molecular Evolution*, 17(6): 368–376.

658 Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Incorporated.

659 Felsenstein, J. and Churchill, G. A. 1996. A hidden Markov model approach to variation
660 among sites in rate of evolution. *Molecular Biology and Evolution*, 13: 93–104.

661 Fitch, W. M. and Margoliash, E. 1967. A method for estimating the number of invariant

amino acid coding positions in a gene using cytochrome c as a model case. *Biochemical Genetics*, 1(1): 65–71.

Fletcher, W. and Yang, Z. 2009. Indelible: a flexible simulator of biological sequence evolution. *Molecular biology and evolution*, 26(8): 1879–1888.

Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7): 685–695.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3): 307–321.

Hajiaghayi, M., Kirkpatrick, B., Wang, L., and Bouchard-Côté, A. 2014. Efficient Continuous-Time Markov Chain Estimation. In *International Conference on Machine Learning (ICML)*, volume 31, pages 638–646.

Hirschberg, D. S. 1975. A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, 18(6): 341–343.

Hobolth, A. and Yoshida, R. 2005. Maximum likelihood estimation of phylogenetic tree and substitution rates via generalized neighbor-joining and the EM algorithm. *Algebraic Biology*.

Holmes, I. 2003. Using guide trees to construct multiple-sequence evolutionary HMMs. *Bioinformatics*, 19(suppl 1): i147–i157.

Holmes, I. and Bruno, W. J. 2001. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, 17(9): 803–820.

Huelsenbeck, J. P. and Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8): 754–755.

685 Jensen, J. L. and Hein, J. 2005. Gibbs sampler for statistical multiple alignment. *Statistica*
686 *Sinica*, 15(4): 889.

687 Jovelin, R. and Cutter, A. D. 2013. Fine-scale signatures of molecular evolution reconcile
688 models of indel-associated mutation. *Genome Biology and Evolution*, 5(5): 978–986.

689 Kallenberg, O. 2002. *Foundations of Modern Probability*. Springer, New York, 2nd ed. 2002
690 edition edition.

691 Klosterman, P. S., Uzilov, A. V., Bendaña, Y. R., Bradley, R. K., Chao, S., Kosiol, C.,
692 Goldman, N., and Holmes, I. 2006. XRate: a fast prototyping, training and annotation
693 tool for phylo-grammars. *BMC Bioinformatics*, 7(1): 428.

694 Knudsen, B. and Miyamoto, M. M. 2003. Sequence alignments and pair hidden Markov
695 models using evolutionary history. *Journal of Molecular Biology*, 333(2): 453–460.

696 Kocot, K. M., Cannon, J. T., Todt, C., Citarella, M. R., Kohn, A. B., Meyer, A., Santos,
697 S. R., Schander, C., Moroz, L. L., Lieb, B., *et al.* 2011. Phylogenomics reveals deep
698 molluscan relationships. *Nature*, 477(7365): 452–456.

699 Kvikstad, E. M. and Duret, L. 2014. Strong heterogeneity in mutation rate causes
700 misleading hallmarks of natural selection on indel mutations in the human genome.
701 *Molecular Biology and Evolution*, 31(1): 23–36.

702 Leushkin, E. V. and Bazykin, G. A. 2013. Short indels are subject to insertion-biased gene
703 conversion. *Evolution*, 67(9): 2604–2613.

704 Li, S. 1996. *Phylogenetic tree construction using Markov chain Monte carlo*. Ph.D. thesis,
705 Ohio State University.

706 Li, S., Pearl, D. K., and Doss, H. 2000. Phylogenetic tree construction using Markov chain
707 Monte Carlo. *Journal of the American Statistical Association*, 95(450): 493–508.

Li, W., Luo, C., and Wu, C. 1985. Evolution of DNA sequences. *Molecular Evolutionary Genetics*, pages 1–94.

Liu, K., Nelesen, S., Raghavan, S., Linder, C. R., and Warnow, T. 2009a. Barking up the wrong treelength: the impact of gap penalty on alignment and tree accuracy. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 6(1): 7–21.

Liu, K., Nelesen, S., Raghavan, S., Linder, C. R., and Warnow, T. 2009b. Barking up the wrong treelength: the impact of gap penalty on alignment and tree accuracy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6: 7–21.

Liu, K., Warnow, T. J., Holder, M. T., Nelesen, S. M., Yu, J., Stamatakis, A. P., and Linder, C. R. 2012. SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology*, 61(1): 90–106.

Löytynoja, A. and Goldman, N. 2008. A model of evolution and structure for multiple sequence alignment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1512): 3913–3919.

Lunter, G. 2007. Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. *Bioinformatics*, 23(13): i289–i296.

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

Lunter, G., Drummond, A. J., Miklós, I., and Hein, J. 2005b. Statistical alignment: recent progress, new applications, and challenges. In *Statistical Methods in Molecular Evolution*, pages 375–405. Springer.

Lunter, G., Ponting, C. P., and Hein, J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Computational Biology*, 2(1): e5.

731 Lydeard, C., Holznagel, W. E., Schnare, M. N., and Gutell, R. R. 2000. Phylogenetic
732 analysis of molluscan mitochondrial LSU rDNA sequences and secondary structures.
733 *Molecular Phylogenetics and Evolution*, 15(1): 83–102.

734 Mau, B. 1996. *Bayesian phylogenetic inference via Markov chain Monte carlo methods*.
735 Ph.D. thesis, University of Wisconsin, Madison.

736 Miklós, I. 2003. Algorithm for statistical alignment of two sequences derived from a
737 Poisson sequence length distribution. *Discrete Applied Mathematics*, 127(1): 79–84.

738 Miklos, I. and Toroczkai, Z. 2001. An improved model for statistical alignment. In *First*
739 *Workshop on Algorithms in Bioinformatics*, Berlin, Heidelberg. Springer-Verlag.

740 Miklós, I., Lunter, G., and Holmes, I. 2004. A long indel model for evolutionary sequence
741 alignment. *Molecular Biology and Evolution*, 21(3): 529–540.

742 Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., and
743 Devine, S. E. 2006. An initial map of insertion and deletion (indel) variation in the
744 human genome. *Genome Research*, 16(9): 1182–1190.

745 Mouchiroud, D., D'Onofrio, G., Aïssani, B., Macaya, G., Gautier, C., and Bernardi, G.
746 1991. The distribution of genes in the human genome. *Gene*, 100: 181–187.

747 Nachman, M. W. and Crowell, S. L. 2000. Estimate of the mutation rate per nucleotide in
748 humans. *Genetics*, 156(1): 297–304.

749 Nam, K. and Ellegren, H. 2012. Recombination drives vertebrate genome contraction.
750 *PLoS Genetics*, 8(5): e1002680.

751 Redelings, B. D. and Suchard, M. A. 2007. Incorporating indel information into phylogeny
752 estimation for rapidly emerging pathogens. *BMC Evolutionary Biology*, 7(1): 40.

753 Robinson, D. and Foulds, L. 1979. Comparison of weighted labelled trees. In
754 *Combinatorial Mathematics VI*, pages 119–126. Springer.

755 Roch, S. 2010. Toward extracting all phylogenetic information from matrices of
756 evolutionary distances. *Science*, 327(5971): 1376–1379.

757 Ronquist, F. and Huelsenbeck, J. P. 2003. MrBayes 3: Bayesian phylogenetic inference
758 under mixed models. *Bioinformatics*, 19(12): 1572–1574.

759 Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for
760 reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4): 406–425.

761 Satija, R., Novák, Á., Miklós, I., Lyngsø, R., and Hein, J. 2009. Bigfoot: Bayesian
762 alignment and phylogenetic footprinting with mcmc. *BMC evolutionary biology*, 9(1):
763 217.

764 Smith, S. A., Wilson, N. G., Goetz, F. E., Feehery, C., Andrade, S. C., Rouse, G. W.,
765 Giribet, G., and Dunn, C. W. 2011. Resolving the evolutionary relationships of molluscs
766 with phylogenomic tools. *Nature*, 480(7377): 364–367.

767 Stamatakis, A. 2005. An efficient program for phylogenetic inference using simulated
768 annealing. In *Parallel and Distributed Processing Symposium, 2005. Proceedings. 19th*
769 *IEEE International*, pages 8–pp. IEEE.

770 Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis
771 of large phylogenies. *Bioinformatics*, 30(9): 1312–1313.

772 Studier, J. A., Keppler, K. J., *et al.* 1988. A note on the neighbor-joining algorithm of
773 Saitou and Nei. *Molecular Biology and Evolution*, 5(6): 729–731.

774 Suchard, M. A. and Redelings, B. D. 2006. BAli-Phy: simultaneous Bayesian inference of
775 alignment and phylogeny. *Bioinformatics*, 22(16): 2047–2048.

776 Sukumaran, J. and Holder, M. T. 2010. DendroPy: a Python library for phylogenetic
777 computing. *Bioinformatics*, 26(12): 1569–1571.

778 Tanay, A. and Siggia, E. D. 2008. Sequence context affects the rate of short insertions and
779 deletions in flies and primates. *Genome Biology*, 9(2): R37.

780 Taylor, M. S., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Semple, C. A. 2006.
781 Heterotachy in mammalian promoter evolution. *PLoS Genetics*, 2(4): e30.

782 Thorne, J. L., Kishino, H., and Felsenstein, J. 1991. An evolutionary model for maximum
783 likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, 33(2): 114–124.

784 Thorne, J. L., Kishino, H., and Felsenstein, J. 1992. Inching toward reality: an improved
785 likelihood model of sequence evolution. *Journal of Molecular Evolution*, 34(1): 3–16.

786 Truszkowski, J. and Goldman, N. 2016. Maximum likelihood phylogenetic inference is
787 consistent on multiple sequence alignments, with or without gaps. *Systematic biology*,
788 65(2): 328–333.

789 Varin, C. and Vidoni, P. 2005. A note on composite likelihood inference and model
790 selection. *Biometrika*, 92(3): 519–528.

791 Westesson, O., Lunter, G., Paten, B., and Holmes, I. 2012. Accurate reconstruction of
792 insertion-deletion histories by statistical phylogenetics. *PLoS One*, 7(4): e34572.

793 Wong, G. K.-S., Liu, B., Wang, J., Zhang, Y., Yang, X., Zhang, Z., Meng, Q., Zhou, J., Li,
794 D., Zhang, J., *et al.* 2004. A genetic variation map for chicken with 2.8 million
795 single-nucleotide polymorphisms. *Nature*, 432(7018): 717–722.

796 Wong, K., Suchard, M., and Huelsenbeck, J. 2008. Alignment uncertainty and genomic
797 analysis. *Science*, 319: 473–476.

Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics*, 139: 993–1005.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9): 367–372.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences: CABIOS*, 13(5): 555–556.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8): 1586–1591.

Zhang, J. 2000. Protein-length distributions for the three domains of life. *Trends in Genetics*, 16(3): 107–109.

Appendix 1

Details of the phylogenetic inference method

In this section, we show how to optimize the parameters of the GeoPIP model via a coordinate ascent algorithm. The full algorithm is summarized in Algorithm 1. Note that Algorithm 1 can also be used for the PIP model, since the PIP model is a special case of the GeoPIP model.

One particularity of the approach is that we maximize rather than marginalize over the segmentations. The approach we took is inspired by a penalized likelihood approach on the segmentation. Our estimation procedure can thus be seen as an hard EM procedure. This choice simplifies the implementation of the algorithm.

Number of indel rate categories m.—

In this paper, we assume that m is fixed for simplicity. This is a reasonable assumption when the number of distinct indel rates can be roughly inferred. In cases that a rough estimate of distinct indel rates is not easy to obtain, choosing m to be a large number works in application as our algorithm will naturally choose a subset of indel rates from m available indel rates, but at a price of higher computational cost.

Algorithm 1 Iterative optimization algorithm for estimation of GeoPIP model parameters

Initialize parameters $\mathbf{Q}$, θ, β, $\mathbf{r}$, ρ, ω.
Calculate $\mathbf{B}$ given θ, $\mathbf{Q}$, β and $\mathbf{r}$.
Infer τ based on $\mathbf{B}$ using NJ and mid-point rooting.
Set tolerance level tol. Set $d = tol$. Set $\ell_{old} =$ 1.e-10. Set $\Delta\ell = 1$.
while $d \geq tol$ and $\Delta\ell > 0$ **do**
    Update β^* and $\mathbf{r}^*$ given θ, $\mathbf{Q}$ and τ using dynamic programming.
    Update ρ^* given $|\beta^*|$.
    Update ω^* given $\mathbf{r}^*$.
    Update θ^* given τ, $\mathbf{Q}$, β^* and $\mathbf{r}^*$.
    Update $\mathbf{Q}^*$ given τ.
    Update $\mathbf{B}^*$ given θ^*, $\mathbf{Q}^*$, β^* and $\mathbf{r}^*$.
    Update τ^* based on $\mathbf{B}^*$ using NJ and mid-point rooting.
    Set $d \leftarrow \max\{\|\mathbf{B}^* - \mathbf{B}\|, \|\theta^* - \theta\|, \|\mathbf{Q}^* - \mathbf{Q}\|\}$.
    Set $\mathbf{B} \leftarrow \mathbf{B}^*, \tau \leftarrow \tau^*, \theta \leftarrow \theta^*, \mathbf{Q} \leftarrow \mathbf{Q}^*, \beta \leftarrow \beta^*, \mathbf{r} \leftarrow \mathbf{r}^*, \rho \leftarrow \rho^*, \omega \leftarrow \omega^*$.
    Calculate full likelihood ℓ_{new}.
    Calculate change of likelihood $\Delta\ell = \ell_{new} - \ell_{old}$.
    Set $\ell_{old} = \ell_{new}$.
end while

Optimizing β and $\mathbf{r}$.—

See description in the Efficient Phylogenetic Inference with the GeoPIP Model section. Here we add the description of the backtracking algorithm. Note that in (5), the maximum is taken over a matrix $\mathbf{L}^{(t)} = (l_{i,j}^{(t)})$ of $t \times m$ elements. Let $(\eta_{t,1}, \eta_{t,2})$ denote the index of the largest element in $\mathbf{L}^{(t)}$. To find the optimal segmentation β for a fixed alignment with maximum likelihood l_n using the path of dynamic programming, we record a backward function $f : \{1, 2, \cdots, n\} \to \{1, 2, \cdots, n\}$ where $f(t)$ is the row index of the

maximum entry in $\mathbf{L}^{(t)}$, i.e.,

$$f(t) = \eta_{t,1}, \quad t = 1, 2, \cdots, n.$$

To find the indel rates $\mathbf{r}$ in each segment of the optimal segmentation β using the path of dynamic programming, we record another backward function $g : \{1, 2, \cdots, n\} \to \{1, 2, \cdots, m\}$ where $g(t)$ is the column index of the maximum entry in $\mathbf{L}^{(t)}$, i.e.,

$$g(t) = \eta_{t,2}, \quad t = 1, 2, \cdots, n.$$

We trace the optimal segmentation β with maximum likelihood and respective indel rates $\mathbf{r}$ by Algorithm 2. The lengths of all segments are given in the ordered array A and the indel rates of all segments are given in the ordered array C of Algorithm 2.

Algorithm 2 Backtracking for best segmentation

```
Set i = n. Set A = ∅. Set C = ∅.
while i > 0 do
    j ← f(i)
    Add element {i − j + 1} to A as the first element.
    Add element g(i) to C as the first element.
    i ← j − 1.
end while
```

It is easy to see that recording these two backward functions f and g does not change the order of the time complexity of the dynamic programming, and finding the best segmentation and rate category in each segment based on f and g does not increase the order of the total time complexity either.

Updating ρ and ω.—

We calculate $\widehat{\rho} = 1/|\beta|$ since $\mathrm{E}(|\beta|) = 1/\rho$. We estimate ω based on $\mathbf{R}$ only, by counting how many inferred states $\widehat{r}_i$ equal j for $i = 1, 2, \ldots, |\beta|$, and $j = 1, 2, \ldots, m$. We

use Laplace smoothing to ensure that all elements of ω are non-zero.

Updating τ.—

We focus on bifurcating tree topologies in this paper. We reconstruct τ using NJ (Saitou and Nei 1987; Gascuel 1997), based on updated pairwise distance matrix $\mathbf{B}$ and root the unrooted tree by midpoint rooting. Since the GeoPIP model is reversible, the root location will not affect the inference of evolutionary parameters.

When all other parameters are fixed, a composite log-likelihood (Varin and Vidoni 2005) ℓ_c of $\mathbf{B}$ can be written as

$$\ell_c(\mathbf{B}) \quad = \quad \sum_{1 \leq i < j \leq N} \log \text{GeoPIP}(\beta(\mathbf{x}_i, \mathbf{x}_j), \mathbf{r} | \theta, b_{ij}, \rho, \omega), \tag{6}$$

where $\beta(\mathbf{x}_i, \mathbf{x}_j)$ denotes the segmentation β on two sequences $\mathbf{x}_i$ and $\mathbf{x}_j$ only, and b_{ij} is the total branch length from sequence i to sequence j.

The parameter b_{ij} only appears in one composite log-likelihood component

$$\log \text{GeoPIP}(\beta(\mathbf{x}_i, \mathbf{x}_j), \mathbf{r} | \theta, b_{ij}, \rho, \omega), \tag{7}$$

thus the maximum composite likelihood estimate (MCLE) $\widehat{b}_{ij}$ can be obtained by maximizing (7) instead of (6). Given β, b_{ij} is conditional independent of ρ, and given θ, b_{ij} is conditional independent of ω. Therefore, the composite likelihood of b_{ij} depends only on β, θ, $\mathbf{Q}$.

Updating θ.—

We estimate indel rate θ by pooling all segments with same rates together.

$$\begin{aligned}
&\log \mathrm{GeoPIP}(\beta, \mathbf{r}|\gamma) \\
= \;& (|\beta|-1)\log(1-\rho) + \log\rho + \sum_{i=1}^{|\beta|} \log \omega_{r_i} + \sum_{i=1}^{|\beta|} \log \mathrm{PIP}(\mathbf{s}_i|\theta_{r_i}, \tau) \\
= \;& (|\beta|-1)\log(1-\rho) + \log\rho + \sum_{i=1}^{|\beta|} \log \omega_{r_i} + \sum_{l=1}^{m} \left\{ \sum_{k:r_k=l} \log \mathrm{PIP}(\mathbf{s}_k|\theta_l, \tau) \right\} \qquad (8)
\end{aligned}$$

where the inner summation is over all $k = 1, 2, \ldots, |\beta|$ satisfying that $r_k = l$, i.e., segments with the l-th indel rates ($l = 1, 2, \ldots, m$). The parameter θ_l appears only in the component

$$\sum_{k:r_k=l} \log \mathrm{PIP}(\mathbf{s}_k|\theta_l, \tau), \qquad (9)$$

therefore, the MLE of θ_l ($l = 1, 2, \cdots, m$) can be obtained by maximizing (9) given rate matrix $\mathbf{Q}$ and tree τ, instead of (8).

Updating $\mathbf{Q}$.—

The conditional substitution rate matrix is the same at all loci regardless of the indel rate of the segment. Based on this observation, we pool all data involving transitions only to estimate the rate matrix $\mathbf{Q}$. We explain this step only briefly as estimating rate matrix $\mathbf{Q}$ is not the focus of this paper, and refer readers to Hobolth and Yoshida (2005) for more details.

We use an EM algorithm to estimate $\mathbf{Q}$ based on substitutions of characters only. At E-step, we calculate expectations of stationary distribution of characters, transitions among all characters and the waiting times at each character type given a rate matrix $\widehat{\mathbf{Q}}$ and data. At M-step, we maximize a penalized likelihood function of $\mathbf{Q}$ based on the GTR model to find $\widehat{\mathbf{Q}}$ given all expectations from the E-step. We repeat the E-step and M-step iteratively until the change in penalized likelihood is smaller than a given tolerance.

The GeoPIP+NJ algorithm can simply incorporate the correlation of indel rates and substitution rates by estimating substitution rate matrices separately for different indel rate regions. The computation cost of updating $\mathbf{Q}$s will increase by a factor of m, which is the number of indel rate categories.

Convergence of the optimization algorithm.—

In our algorithm, the iterative updating procedure is terminated when the change of parameters is smaller than the tolerance level or the full likelihood decreases after one full iteration, as shown in Algorithm 1.

We calculate the full likelihood of the new set of all parameters updated at the end of each iteration and monitor the change of the full likelihood. This procedure is important. Because some updating steps for individual parameters, for example $\mathbf{B}$, are not based on optimizing the full likelihood, even though at each step for individual parameters, we obtain a new estimate which maximize the respective (composite) likelihood, it is possible that the full likelihood may decrease after one full iteration. The estimates obtained using our algorithm are not guaranteed to represent a global optimum in general.

Appendix 2

Hierarchical Poisson Indel Process

In this section, we describe the Hierarchical Poisson Indel Process (hPIP), the model we use in some of the synthetic data experiments to generate dataset containing long indels. The parameters of the hPIP model consist in θ, ω defined as in the GeoPIP model, in addition to an "upper level" PIP insertion and deletion parameters $\lambda_{\text{top}}, \mu_{\text{top}} > 0$.

The generative process of the hPIP model is as follows. First, at the root of the tree, sample a number of segments $Z \sim \text{Poisson}(\lambda_{\text{top}}/\mu_{\text{top}})$, and for each segment i, sample

an indel rate category θ_{R_i} as in the GeoPIP model. For each segment, also sample a sequence distributed according to the stationary distribution of the PIP model with parameters θ_{R_i} given by the previous step.

Next, assume recursively that a segmented sequence is given for some point on the tree. The sequence in the segments undergo independent but not identically distributed "lower level" PIP evolutionary models. They are not identically distributed because different segments have different indel rate categories. In addition to that, a new segment can be added, and a whole segment can be deleted. Insertion and deletion of segments obey the "top level" PIP distribution: deletion of a segment occurs at a rate μ_{top} per segment, and insertion of a segment, at a rate λ_{top} (independent of the number of segment). When a segment is inserted, its location is chosen uniformly at random.

Figure 1: The reference phylogenetic trees used in simulation studies. a). a phylogenetic tree with 8 leaves and varying branch lengths. b). a perfect binary phylogenetic tree with 16 leaves and same branch length b for all branches.

Figure 2: Inferred indel rate categories for alignment columns 150-300 of one set of simulated data: segments with low estimated deletion rate (0.006) are in white; segments with intermediate deletion rate (0.848) are in light gray; segments with high deletion rate (4.150) are in dark gray.

Figure 3: Inferred indel rate categories for alignment columns of one set of simulated data using NB+SUB+POW and MUSCLE alignment: segments with low estimated deletion rate (0.060) are in white; segments with intermediate deletion rate (0.500) are in light gray; segments with high deletion rate (1.211) are in dark gray.

Figure 4: Trees reconstructed by the three indel-aware methods (columns) for the data with and without outgroups (rows). The five numbers measure the wRF distance and the RF distance (in brackets) between each of the bottom tree and the corresponding top subtree obtained after excluding the two outgroups.

Figure 5: Inferred indel rate categories for alignment columns 701-850 of molluscan data: segments with lowest deletion rate (0.01) are in white; segments with low deletion rate (0.11) are in light gray; segments with high deletion rate (0.41) are in medium gray; segments with low deletion rate (1.27) are in dark gray.