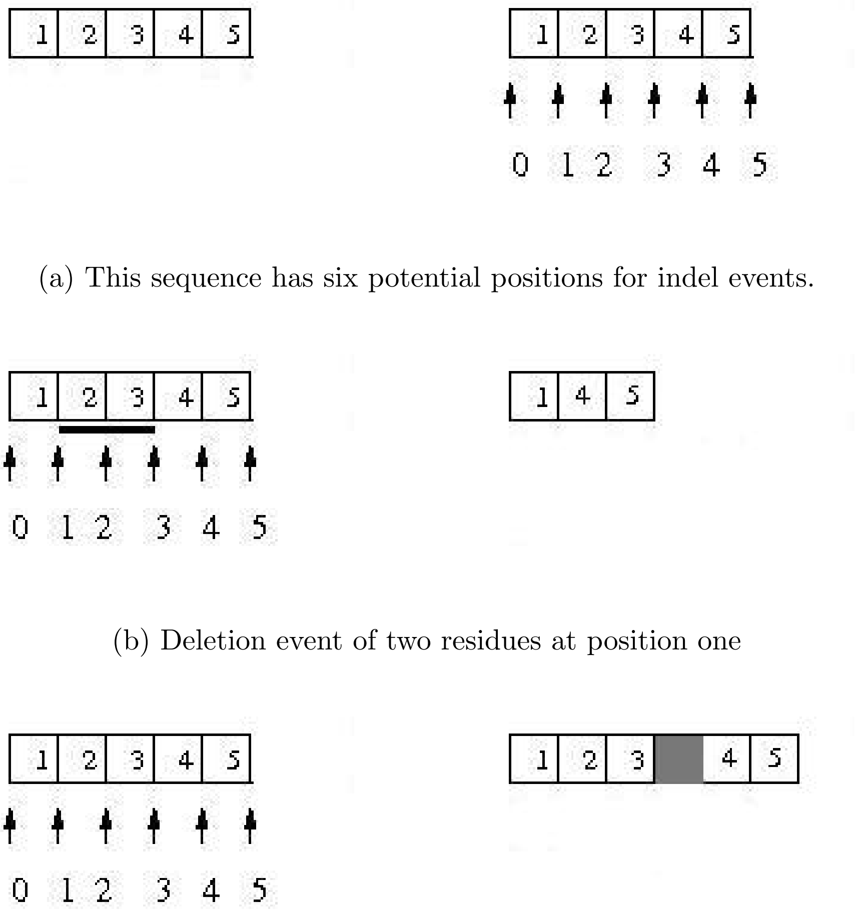
Web-based Supplementary Materials for “BayesCAT : Bayesian Co-estimation of Alignment and Tree” by Heejung Shim and Bret Larget

1 Web Figure 1. Indel events for a given sequence



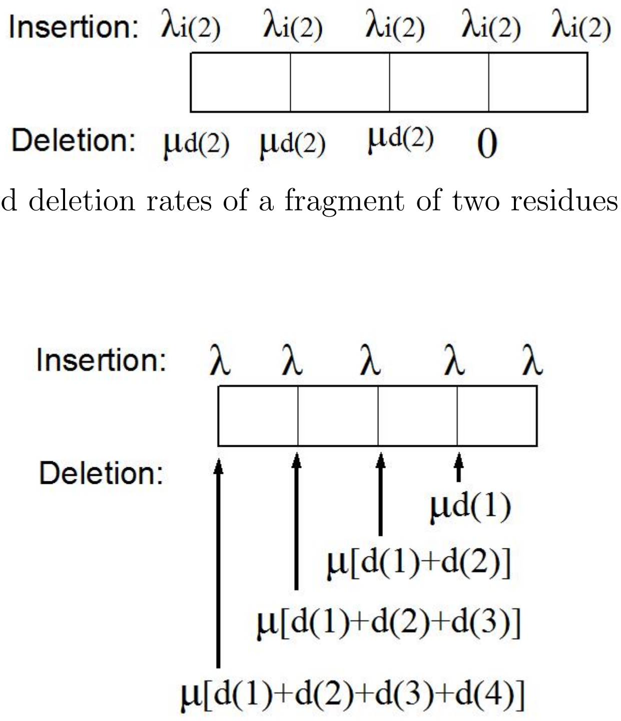
(a) This sequence has six potential positions for indel events.

(b) Deletion event of two residues at position one

Insertion event of one residue at position three

Figure 1: **Example : Indel events for a given sequence.** (a) For the given sequence of five residues, we assign numbers one to five to each residue. This sequence has six potential positions for indel events including one start position, one end position, and four positions between two residues. The positions are numbered from zero to five. (b) A deletion event of two residues occurs at position one. The two residues to the right of position one are deleted from the sequence. (c) An insertion event of one residue occurs at position three.

2 Web Figure 2. Insertion and deletion rates on a sequence



(a) Insertion and deletion rates of a fragment of two residues at each position.

(b) Total insertion and deletion rates per site.

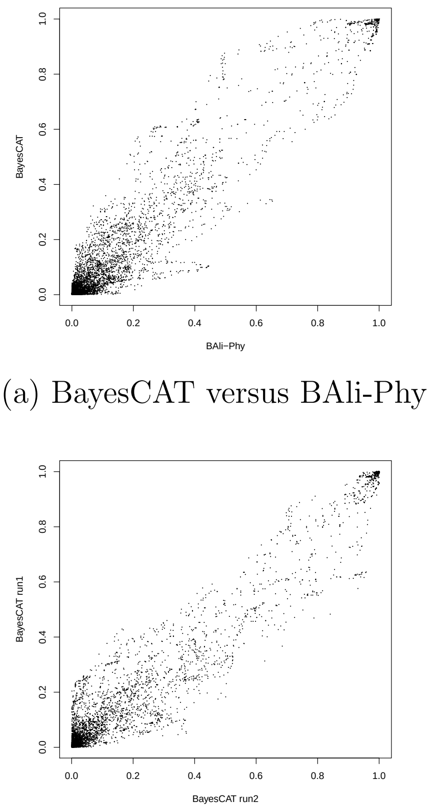
Figure 2: **Example : Insertion and deletion rates on a sequence.** A sequence of four residues has five and four possible positions for insertion and deletion events, respectively. (a) For a fragment of two residues, the insertion rate is $\lambda i(2)$ at all possible positions, but the deletion rate is $\mu d(2)$ at the first three positions, but zero at the fourth position as a deletion with a fragment of two residues cannot occur at the fourth position. (b) The total insertion rate per site is $\lambda \sum_{k=1}^{\infty} i(k) = \lambda$ at all positions, but the total deletion rate for a position j is $\mu \sum_{k=1}^{4-j} d(k)$ for $j \in \{0, 1, 2, 3\}$, which depends on the position on the sequence. The deletion fragment size distribution at a position $j \in \{0, 1, 2, 3\}$ is $\frac{d(x)}{\sum_{k=1}^{4-j} d(k)}$ for all $x \in \{1, \ldots, 4-j\}$, which is a truncated distribution of $d(\cdot)$

3 Web Table 1. 5S rRNA : The posterior estimates of parameters.

Parameter	Prior Distribution		Posterior Distribution	
	mean/median	95% *C.R.*	mean/median	95% *C.R.*
γ	$\infty/2$	(0.051, 78)	3.605/3.4	(1.540, 6.607)
κ	$\infty/2$	(0.051, 78)	1.812/1.77	(1.023, 2.849)
π_A	0.19	(0.107, 0.289)	0.173	(0.164, 0.185)
π_C	0.31	(0.208, 0.422)	0.325	(0.307, 0.349)
π_G	0.33	(0.226, 0.444)	0.336	(0.312, 0.347)
π_T	0.17	(0.092, 0.265)	0.166	(0.151, 0.186)
λ	0.005	(0.00013, 0.0185)	0.0216	(0.011, 0.036)
r	0.008	(0.006, 0.01)	0.0081	(0.0068, 0.0097)
r_d	0.166	(0.038, 0.364)	0.290	(0.148, 0.501)

Table 1: **5S rRNA : The posterior estimates of parameters.** For γ and κ, we list prior and posterior mean/median together. The third and fifth columns show 95% credible regions for the parameters.

4 Web Figure 3. 5S rRNA : Comparison of two alignment distributions (BayesCAT and BAli-Phy) using pairwise homology posterior probabilities.

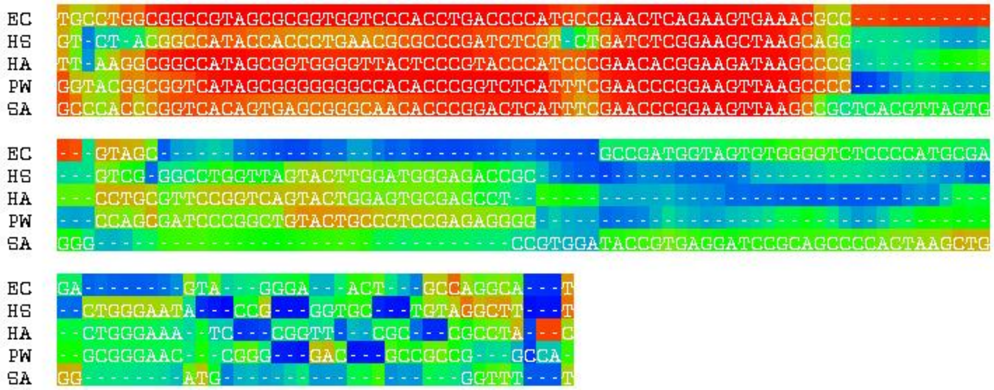


(a) BayesCAT versus BAli-Phy

(b) Two different MCMC chains of BayesCAT

Figure 3: **5S rRNA : Comparison of two alignment distributions (BayesCAT and BAli-Phy) using pairwise homology posterior probabilities.** Each point represents a homology for a pair of bases from two different sequences or a homology for one base and a gap. The scatter plots show two estimated posterior probabilities of the homologies. (a) Posterior probabilities from BayesCAT and BAli-Phy. (b) Posterior probabilities from different MCMC chains of BayesCAT.

5 Web Figure 4. 5s rRNA : Multiple sequence alignment estimated using FSA.



```
EC  TGCCTGGCGGCCGTAGCGCGGTGGTCCCACCTGACCCCATGCCGAACTCAGAAGTGAAACGCC-----------
HS  GT-CT-ACGGCCATACCACCCTGAACGCGCCCGATCTCGT-CTGATCTCGGAAGCTAAGCAGG-----------
HA  TT-AAGGCGGCCATAGCGGTGGGGTTACTCCCGTACCCATCCCGAACACGGAAGATAAGCCCG-----------
PW  GGTACGGCGGTCATAGCGGGGGGGCCACACCCGGTCTCATTTCGAACCCGGAAGTTAAGCCCC-----------
SA  GCCCACCCGGTCACAGTGAGCGGGCAACACCCGGACTCATTTCGAACCCGGAAGTTAAGCCGCTCACGTTAGTG

EC  ---GTAGC-----------------------------------GCCGATGGTAGTGTGGGGTCTCCCCATGCGA
HS  ---GTCG-GGCCTGGTTAGTACTTGGATGGGAGACCGC------------------------------------
HA  ---CCTGCGTTCCGGTCAGTACTGGAGTGCGAGCCT--------------------------------------
PW  ---CCAGCGATCCCGGCTGTACTGCCCTCCGAGAGGGG------------------------------------
SA  GGG---------------------------------CCGTGGATACCGTGAGGATCCGCAGCCCCACTAAGCTG

EC  GA--------GTA---GGGA---ACT---GCCAGGCA---T
HS  --CTGGGAATA---CCG---GGTGC---TGTAGGCTT---T
HA  --CTGGGAAA--TC---CGGTT---CGC---CGCCTA---C
PW  --GCGGGAAC---CGGG---GAC---GCCGCCG---GCCA-
SA  GG--------ATG-------------------GGTTT---T
```

Figure 4: **5s rRNA : Multiple sequence alignment estimated using FSA.**

6 Web Figure 5. 5s rRNA : Posterior estimate of realized indel fragment size distribution (BayesCAT).

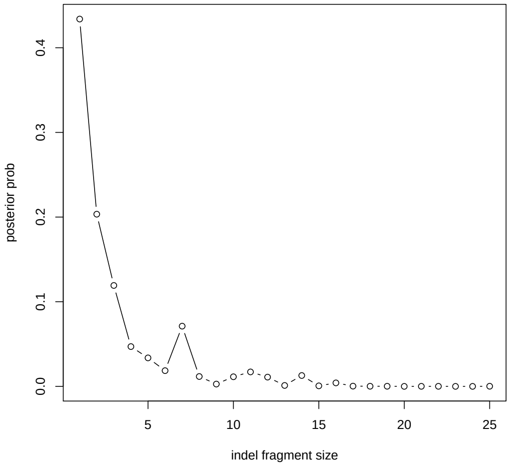


Figure 5: **5S rRNA : Posterior estimate of realized indel fragment size distribution (BayesCAT).**

7 Web Appendix A: A comparison among results from the traditional sequential approach with three fixed alignments

To motivate further approaches to joint estimation of the alignment and tree, we investigate the problems of the traditional sequential approach using a simulated data set with five taxa, which is generated from our joint model. We introduce our joint model in the main text and describe the procedure to generate the data set in Section 16. The true tree and simulated sequence alignment are displayed in Figure 6 and Figure 7 (a), respectively.

We apply MrBayes (Huelsenbeck and Ronquist, 2001), software for Bayesian phylogeny estimation, to three different fixed alignments. One is the true sequence alignment and the other two are estimated using ClustalW (Thompson *et al.*, 1994) and Muscle (Edgar, 2004).

The two alignments created with ClustalW and Muscle are shown in Figure 7 with the true alignment. The columns marked with $*$ in the two alignments also appear in

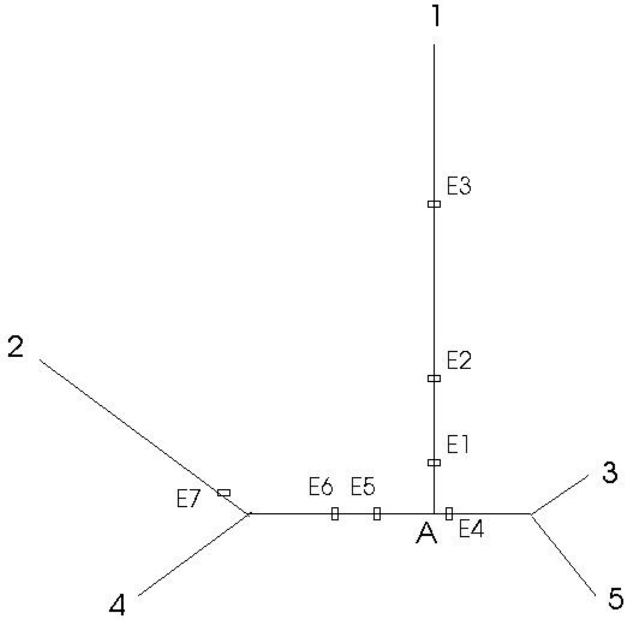


Figure 6: **Simulated true tree and history of indel events.** An internal node A has 67 bases. We specify seven indel events relative to rooting the tree at A: insertion of one base at position 25 ($E1$), deletion of five bases at position 43 ($E2$), deletion of two bases at position five ($E3$), deletion of four bases at position 40 ($E4$), insertion of one base at position 41 ($E5$), deletion of one base at position eight ($E6$), and deletion of six bases at position ten ($E7$). Section 3.3.1 (Description of an indel history on a tree) in the main text introduces the terms used here when specifying indel events on the tree.

```
TGGTG--TTCCACCTCTTTGCACAAGACGGCTAGCCCCATCT-T-----TCCGTTGAACATATTTTCCC
TGATCGTG-CC------TTGGCTCG-TTGTTCGACGCCATCGTATTACGCTTTTCATTCAGAACTTCAA
AGATGGGCTCCCACGTTCCGCACTA-TCGGCCGGCGCCATC-----TCACTTGTTATATACAACTTCAT
AGATGGGG-CCATCGTTTTACACAG-TTGAATGCCGCCATCGTATTACACCCCTTATTCTCATTTTCAC
AGGTAGGCTCCCACGTTCCGCACTA-TCGGCTGGCGCCATC-----CCATTTGTTATACACACTTTCAA
```

(a) True alignment

```
                                               ********************
-TGGTGTTCCACCTCTTTGCAC-AAGA-CGGCTAGCCCCATCTT---TCCGTTGAACATATTTTCCC
-----TGATCGTGCCTTGGCTC-GTTGTTCGACGCCATCGTATTACGCTTTTCATTCAGAACTTCAA
-AGATGGGCTCCCACGTTCCGC-ACTATCGGCCGGCGCCATCTCA--CTTGTTATATACAACTTCAT
AGATGGGGCCATCGTTTTACACAGTTGAATGCCGCCATCGTATTACACCCCTTATTCTCATTTTCAC
-AGGTAGGCTCCCACGTTCCGC-ACTATCGGCTGGCGCCATCCCA--TTTGTTATACACACTTTCAA
```

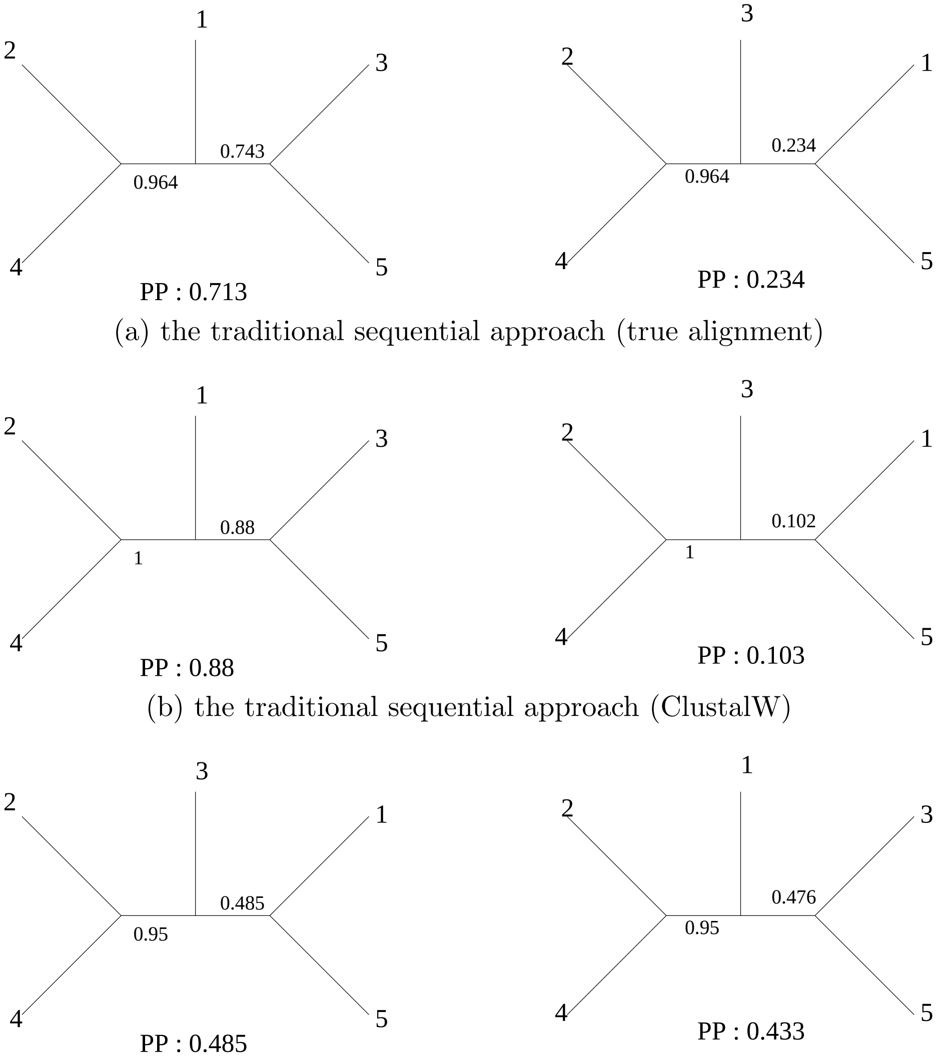
(b) ClustalW

```
*****                    *************** * ***  ********************
TGGTGTTCCACCTCTTTGCACAA-GACGGCTAGCCCCATC------TTTCCGTTGAACATATTTTCCC
TGATCGT-GCCTTGGCTC------GTTGTTCGACGCCATCGTATTACGCTTTTCATTCAGAACTTCAA
AGATGGGCTCCCACGTTCCGCACTATCGGCCGGCGCCATC-----TCACTTGTTATATACAACTTCAT
AGATGGG-GCCATCGTTTTACACAGTTGAATGCCGCCATCGTATTACACCCCTTATTCTCATTTTCAC
AGGTAGGCTCCCACGTTCCGCACTATCGGCTGGCGCCATC-----CCATTTGTTATACACACTTTCAA
```

(c) Muscle

Figure 7: **True alignment and alignments estimated with ClustalW and Muscle for a simulated data set.**

the true alignment. The alignment estimated with Muscle looks quite similar to the true alignment. The discrepancy between the Muscle and true alignments originates from the positioning of a small number of gaps. The alignment estimated using ClustalW agrees with the others only at the right part of the alignment. For this data set, ClustalW places most of the gaps at the beginning of the alignment.



(a) the traditional sequential approach (true alignment)

(b) the traditional sequential approach (ClustalW)

(c) the traditional sequential approach (Muscle)

Figure 8: **Inferred topologies with a posterior probability (PP) $>$ 0.1.**

Figure 8 shows all topologies with posterior probability greater than 0.1 from analyses using MrBayes with the three fixed alignments. The traditional sequential method using the true alignment highly supports the true topology (0.713). When the alignment is fixed to that estimated by ClustalW, the support for the true topology (0.88) is stronger than when the true alignment is used. A possible explanation is that the topology of the guide tree used by ClustalW in its alignment procedure is the true topology. Unlike the previous two cases, the most probable topology is not the true topology in the traditional sequential approach with the alignment determined by Muscle, although the posterior probability of the true topology (0.433) is comparable to that of the most probable topology (0.485). Based on the close similarity between the true alignment and the alignment from Muscle, this result demonstrates how much phylogeny estimation with a fixed alignment using the traditional sequential approach can be sensitive to small changes in the alignment.

This example shows the possibility of observing a great disparity between estimates of the posterior probability of the true topology when analyzing identical sequences under identical models, but with different fixed alignments, which provides motivation to develop a model that accounts for alignment uncertainty.

8 Web Appendix B: HKY model as a substitution model

We use the HKY model (Hasegawa *et al.*, 1985) as a substitution model in our analysis, so Θ_{sub} consists of κ, the ratio of the transition to transversion rates among nucleotides, and nucleotide frequencies in the equilibrium distribution of the rate matrix, denoted as $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$. Transitions are substitutions between purines ($A \leftrightarrow G$) or pyrimidines ($C \leftrightarrow T$) while transversions are substitutions from purine to pyrimidine or from pyrimidine to purine. The rate matrix under the HKY model is

$$Q_{\text{HKY}} = \phi \begin{pmatrix} -(\kappa\pi_G + \pi_Y) & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & -(\kappa\pi_T + \pi_R) & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & -(\kappa\pi_A + \pi_Y) & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & -(\kappa\pi_C + \pi_R) \end{pmatrix},$$

where $\phi = \frac{1}{2(\kappa(\pi_A\pi_G + \pi_C\pi_T) + \pi_R\pi_Y)}$ is a scaling parameter, $\pi_R = \pi_A + \pi_G$, and $\pi_Y = \pi_C + \pi_T$. The matrix is scaled so that there is one expected substitution per unit time at equilibrium. Then, under the HKY model, the transition probability that a base at each column changes from state i to state j during a given time t, $\mathsf{P}_{(i,j)}(t \mid \Theta_{\text{sub}})$, is

$$\begin{cases} \pi_j + \pi_j(\frac{1}{\psi_j} - 1)\exp(-\phi t) + (\frac{\psi_j - \pi_j}{\psi_j})\exp(-\phi t(1 + \psi_j(\kappa - 1))) & \text{if } i = j \\ \pi_j + \pi_j(\frac{1}{\psi_j} - 1)\exp(-\phi t) + (\frac{\pi_j}{\psi_j})\exp(-\phi t(1 + \psi_j(\kappa - 1))) & \text{if } i \neq j \text{ and transition} \\ \pi_j(1 - \exp(-\phi t)) & \text{if } i \neq j \text{ and transversion,} \end{cases}$$

where $\psi_j = \pi_A + \pi_G$ if $j \in \{A, G\}$ and $\psi_j = \pi_C + \pi_T$ if $j \in \{G, T\}$.

9 Web Appendix C: The proof for the proposition 1 in the main text

Proposition 1 Under the assumptions listed in Section "General indel model" of the main text., the equilibrium length distribution $q(\cdot)$ is

$$q(x) \quad = \quad r(1-r)^x \text{ for all } x \in \{0, 1, \ldots\},$$

where $1 - r = \frac{\lambda i(1)}{\mu d(1)}$ and $0 < r < 1$. The base deletion fragment size distribution $d(\cdot)$ can be any distribution with support on the positive integers and $d(1) > 0$, and the ratio of the insertion rate to the deletion rate is

$$\frac{\lambda}{\mu} \quad = \quad \sum_{k=1}^{\infty} (1-r)^k d(k) < 1.$$

The base insertion fragment size distribution $i(\cdot)$ is determined as

$$i(k) \quad = \quad \frac{\mu}{\lambda}(1-r)^k d(k) \text{ for all } k \in \{1, 2, \ldots\}.$$

Proof Time-reversibility leads to

$$\lambda i(k) q(x) = \mu d(k) q(x+k) \text{ for all } x \in \{0, 1, \ldots\} \text{ and } k \in \{1, 2, \ldots\}. \tag{1}$$

Taking $k = 1$, we see

$$\lambda i(1) q(x) \quad = \quad \mu d(1) q(x+1) \text{ for all } x \in \{0, 1, \ldots\}.$$

As $d(1) > 0$ and $\mu > 0$, we have

$$q(x+1) = \frac{\lambda i(1)}{\mu d(1)} q(x) \text{ for all } x \in \{0, 1, \ldots\},$$

which means

$$q(x) = r(1-r)^x \text{ for all } x \in \{0, 1, \ldots\}, \tag{2}$$

where $1 - r = \frac{\lambda i(1)}{\mu d(1)}$ and $0 < r < 1$. Substituting (2) into (1), we obtain

$$i(k) \quad = \quad \frac{\mu}{\lambda}(1-r)^k d(k) \text{ for all } k \in \{1, 2, \ldots\}.$$

For any probability distribution $d(\cdot)$ with support on positive integers and real number r with $0 < r < 1$, $\sum_{k=1}^{\infty}(1-r)^k d(k) < 1 < \infty$ as $\sum_{k=1}^{\infty} d(k) = 1$ and $(1-r)^k d(k) < d(k)$ for all $k \in \{1, 2, \ldots\}$. Therefore, we can satisfy the assumption of a probability distribution for $i(\cdot)$ by restricting $\frac{\lambda}{\mu} = \sum_{k=1}^{\infty} (1-r)^k d(k)$. □

10 Web Appendix D: Examples of general indel model

This section illustrates two examples of the general indel model, determined by the selection of a particular distribution for the deletion fragment size.

Case 1 : Negative binomial distribution

The negative binomial distribution is a two-parameter generalization of the geometric distribution that allows the mean and variance to be specified independently. Setting the deletion fragment size distribution to the negative binomial (b_d, p_d) distribution shifted to the positive integers yields a similar distribution for insertion fragment size. Selection of r, b_d and p_d, where $0 < r$, $p_d < 1$ and $0 < b_d$, leads to

$$\begin{aligned} q(x) &= r(1-r)^x \text{ for all } x \in \{0,1,\ldots\}, \\ d(x) &= \frac{\Gamma(b_d+x-1)}{(x-1)!\Gamma(b_d)} p_d^{b_d}(1-p_d)^{x-1} \text{ for all } x \in \{1,2,\ldots\}, \\ i(x) &= \frac{\Gamma(b_i+x-1)}{(x-1)!\Gamma(b_i)} p_i^{b_i}(1-p_i)^{x-1} \text{ for all } x \in \{1,2,\ldots\}, \\ \frac{\lambda}{\mu} &= \left(\frac{p_d}{p_i}\right)^{b_i} \frac{1-p_i}{1-p_d}, \end{aligned}$$

where $p_i = 1-(1-p_d)(1-r)$, $b_i = b_d$, $0 < r$, p_d, $p_i < 1$, b_i, $b_d > 0$, $p_i > p_d$, $p_i > r$, and $\mu > \lambda > 0$. Constraints leaves four free parameters, $\Theta_{\text{ID}} = (r, p_d, b, \lambda)$, where $b = b_i = b_d$. We note that the geometric distribution is a special case of the negative binomial distribution with $b = 1$ and $p_d = r_d$ ($b = 1$ and $p_i = r_i$). We also note that $(b(1-p_i)/p_i)+1$ and $(b(1-p_d)/p_d)+1$ are mean insertion and mean deletion fragment sizes under the negative binomial model, respectively. As $p_i > p_d$, the mean deletion fragment size is greater than the mean insertion fragment size.

For prior on p_d and b, we reparameterize them to $\omega_1 = b(1-p_d)/p_d$ and $\omega_2 = \frac{1}{p_d}-1$, and then assume gamma distributions with parameters (α_1, β_1) and (α_2, β_2) for ω_1 and ω_2, respectively. We note that deletion fragments must contain at least one residue and ω_1 equals the expected length of the remaining residues.

Case 2 : Power law distribution

We can consider the power law distribution recommended by Cartwright (2009) as the deletion fragment size distribution. Selection of $0 < r < 1$ and $\alpha > 1$ yields

$$\begin{aligned} q(x) &= r(1-r)^x \text{ for all } x \in \{0,1,\ldots\}, \\ d(x) &= C_\alpha x^{-\alpha} \text{ for all } x \in \{1,2,\ldots\}, \\ i(x) &= \frac{\lambda}{\mu}(1-r)^x C_\alpha x^{-\alpha} \text{ for all } x \in \{1,2,\ldots\}, \\ \frac{\lambda}{\mu} &= \sum_{k=1}^{\infty}(1-r)^k C_\alpha x^{-\alpha}, \end{aligned}$$

where $0 < r < 1$, $\alpha > 1$, $\mu > \lambda > 0$, and C_α is Riemann's Zeta function. Constraints leaves three free parameters, $\Theta_{\mathrm{ID}} = (r, \alpha, \lambda)$.

11 Web Appendix E: Basic ideas of four categories of MCMC proposals

Table 2 lists four categories of our MCMC proposal methods and the components of the state space updated by the methods in each category. This section provides an overview of the proposal methods in these four categories. Shim (2010) describes all the updates in detail.

Category of MCMC proposal methods	**Updated parameters**
Update a branch length	V, H
Update an indel history on a single edge	$H(A)$
Update an indel history on three edges adjacent to one internal node	V, $H(A)$, $\lvert S_{\text{internal node}}\rvert$
Subtree pruning and regrafting (SPR)	τ, V, $H(A)$, $\lvert S_{\text{internal node}}\rvert$

Table 2: **Four categories of MCMC proposal methods.** Our MCMC proposal methods are grouped into four categories. $H(A)$ indicates changes in an indel history (H) leading to variation of an alignment (A) while changes in an indel history without any modification to the alignment is denoted by H. V, τ, and $|S_{\text{internal node}}|$ indicate branch length, tree topology, and sequence length at internal nodes, respectively.

11.1 Update a branch length

This proposal method updates the branch length of a randomly selected edge. Although the times of the indel events on the edge change in proportion to the change of the edge length, this update method does not vary alignments. We note that this is the only update method which leads to changes in the total sum of branch lengths.

11.2 Update an indel history on a single edge

The proposal methods in this category select an edge of the tree at random and propose a new indel history on the edge, conditional on the fixed sequence lengths of the two nodes connected by the edge.

The selected edge separates the tree into two subtrees whose roots are the two nodes connected by the edge. As the indel history on each subtree is not modified, the alignment of sequences corresponding to leaves in each subtree remains the same. However, the proposed history can modify the complete alignment by changing homologies between subtrees. See Figure 9 for an example.

11.3 Update an indel history on three edges adjacent to one internal node

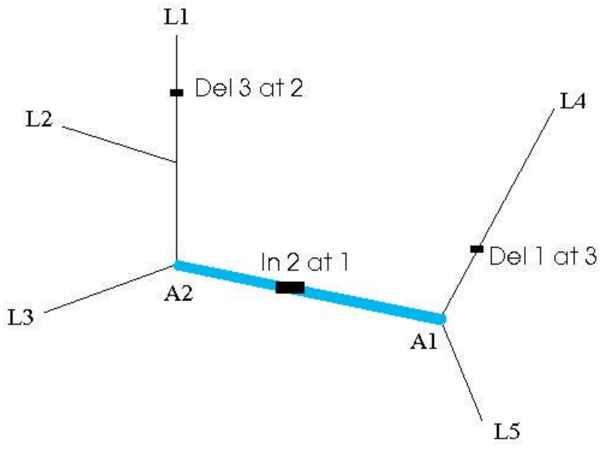
The idea is to first pick an internal node Z, which has three adjacent edges as depicted in Figure 10. We select two of the three edges, join the two edges into one edge, and propose a new indel history on this edge. Next, the method selects a new position for node Z (Z^*). Finally, propose a new indel history on the remaining edge beginning at node Z^*. The two proposals of a new indel history on an edge are conditional on sequence lengths of the two nodes connected by that edge.

The three edges connect node Z to three nodes whose sequence lengths are unchanged by this update, and separate the tree into three subtrees whose roots are the three nodes. As the method does not modify the indel history on each subtree, the alignment of sequences corresponding to leaves in each subtree remains the same. This method can yield a bigger alteration in a multiple alignment than the method to update an indel history on a single edge in the previous section because it keeps the alignment fixed within the three (not two) groups of sequences and changes the alignment between these three groups. Unlike the method in the previous section, this method updates a sequence length at an internal node (Z). In addition, two branch lengths are modified by sampling a new position of node Z, although the total sum of branch lengths remains unchanged.

11.4 Subtree pruning and regrafting (SPR)

As described in Figure 11, the basic idea is to first select an internal node Z, which has three adjacent edges. We pick an edge e_C from these three edges. The edge e_C connects node Z to node C, and separates the tree into two subtrees whose roots are the two nodes. We detach the subtree, which has node C as a root, together with edge e_C from the tree. Next, we pick an edge from the remainder of the tree, pick a location on this edge, and reconnect the detached subtree and edge to the remaining tree by placing node Z at the selected point (Z^*). Finally, we propose an indel history on e_C from Z^* to C.

SPR is the only update method which can change the tree topology (τ). In addition, it updates the sequence length at an internal node (Z), the indel history (H), and the collection of branch lengths (V). Although the method updates branch lengths, the total sum of branch lengths remains constant. SPR updates an indel history only on the selected edge e_C, which does not modify the alignment of sequences corresponding to leaves in each of the two disjoint subtrees (Figure 11). Alterations in the multiple alignment are limited to changes in the alignment between the two groups of sequences.



(a)

```
L1 : + + − − − + +
L2 : + + + + + + +
L3 : + + + + + + +
L4 : + − − + + − +
L5 : + − − + + + +
```

(b)

```
L1 : + + − − − + +
L2 : + + + + + + +
L3 : + + + + + + +
L4 : + + − − + − +
L5 : + + − − + + +
```

(c)

Figure 9: **Update an indel history on a single edge.** (a) Let an internal node $A1$ have five bases. We specify three indel events relative to rooting the tree at $A1$: an insertion of two bases at position one on the edge leading to node $A2$, a deletion of three bases at position two on the edge leading to node $L1$, and a deletion of one base at position three on the edge leading to node $L4$. The selected branch, which connects node $A1$ to $A2$, partitions the leaves into two groups, $\{L1,\ L2,\ L3\}$ and $\{L4,\ L5\}$. (b) The multiple alignment determined by the indel history mapped onto the tree in (a). (c) The modified alignment after the position of the event on the selected edge is changed from one to two. The changed homologies are indicated with rectangles. We observe that the alignment between the two leaf groups is changed while alignments within each group remain unchanged.

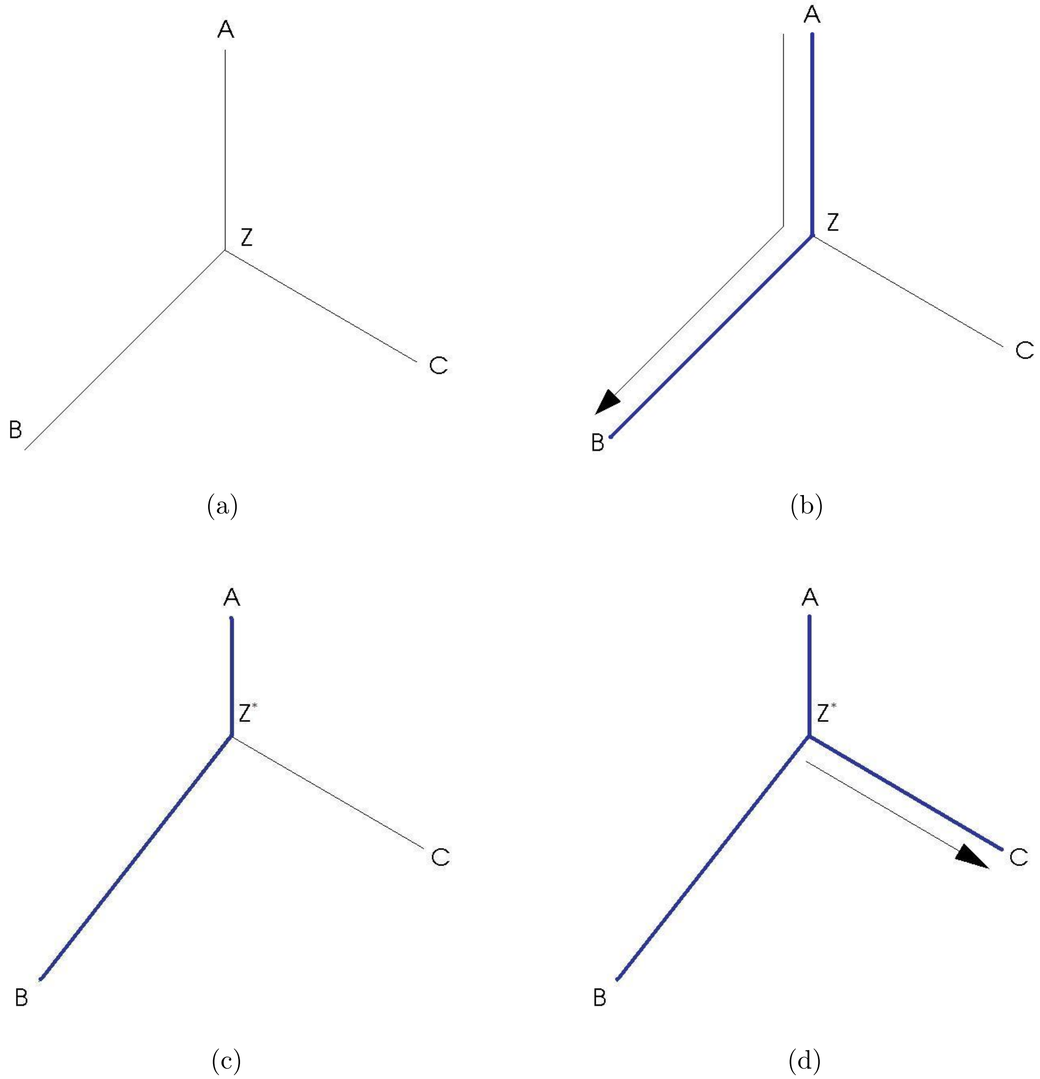
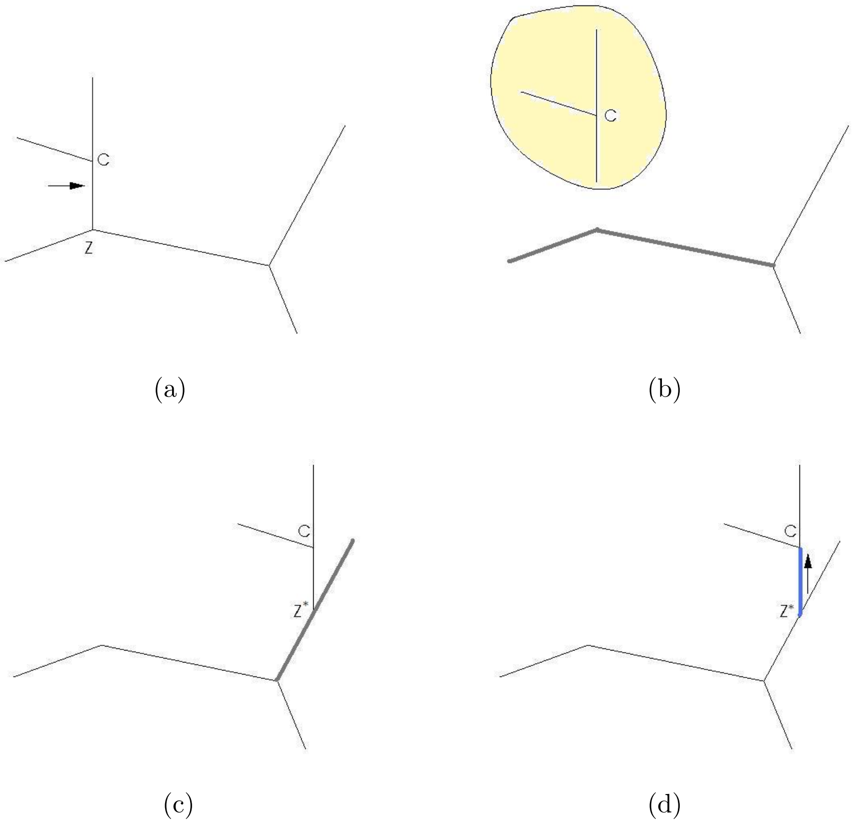


Figure 10: **Update an indel history on three edges adjacent to one internal node.** (a) The sampled internal node Z is connected to three nodes A, B, and C. (b) We pick one pair of nodes (A and B), and sample one direction (from A to B). Then, we propose a new indel history on the path from A to B. (c) A new position of Z (Z^*) is sampled on the path. (d) We propose new indel history on the remaining edge from Z^* to C.



(a) (b)

(c) (d)

Figure 11: **Subtree pruning and regrafting (SPR).** (a) We select one internal node Z, which has three adjacent edges. We pick one edge e_C (indicated by the arrow). This edge connects node Z to node C and separates a tree into two subtrees. (b) We detach the subtree, which has node C as a root, together with edge e_C from the tree. The detached part is circled. Joining two remaining edges into a single edge (thick line) leaves the remaining part as an unrooted tree. (c) We select one edge (thick line) from the remaining tree, and then reconnect the detached part to the remaining tree by placing node Z at the randomly selected point on the selected edge. (d) As the current indel history on edge e_C is not consistent with the sequence length at the new position of Z (Z^*), we update the indel history on e_C from Z^* to C.

12 Web Appendix F: Validation of the MCMC implementation

step 1: Generate parameters Θ from their prior distributions.
step 2: Pick an unrooted tree topology τ from a uniform distribution over unrooted tree topologies with n taxa.
step 3: Generate each branch length of $V = (v_1, \ldots, v_{2n-3})$ from independent exponential distributions with common mean $1/\gamma$.
step 4: Pick one node as a root, and then determine the parent node of each edge relative to the root position.
step 5: Sample the root sequence length from the equilibrium length distribution $q(x) = r(1-r)^x$, and then generate each base of the sequence from the stationary distribution $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$.
step 6: Starting from edges adjacent to the root, generate an indel history (h_i) on each single edge, conditional on a sequence length of the parent node (n_i) and the edge length (v_i), from the distribution under our indel model,
i.e., $h_i \sim \mathsf{P}(h \mid v = v_i, n = n_i, \Theta_{\text{ID}})$. The generated indel history determines a sequence length of the child node. Sample each base of the child sequence from the stationary distribution $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ if the base is inserted or from the transition probability distribution, conditional on the base at the parent node and the edge length, otherwise.
step 7: Repeat step 6 until all sequences at leaves are sampled.

Table 3: **Procedure to simulate sequences from the prior distribution.**

As some of our updates use the sequence data to propose new states, we cannot run MCMC without data to check that the sample is consistent with the prior distribution. Instead, we generate many data sets from the prior distribution, run MCMC on each one, calculate summary statistics of interest from each sample, average these across samples, and then compare these results to expected values from the prior distribution. Table 3 describes a procedure to generate one data set with n taxa from the prior distribution. We generate 4000 data sets with four taxa from the prior distribution and sample 1000 states using MCMC for each data set. The geometric distribution is used as the deletion fragment size distribution. Table 4 shows estimates of the means of external edge lengths and counts of insertion and deletion events on each external edge using samples from the prior distribution and averages over MCMC samples. We also investigate estimates of parameters (γ, κ, π, λ, r and r_d) in the model (not shown). The two results are congruent. For indel fragment sizes and sequence lengths at internal nodes, we compare two distributions, one using samples from the prior distribution and the other using MCMC samples over all data sets. The two distributions are quite similar in which the largest difference in relative frequencies is less than 0.001 for indel fragment sizes and 0.0001 for sequence lengths at internal nodes. In cases with five, six and seven taxa, we also get consistent results. We also get similar agreements when varying prior distributions (not shown).

	Edge number	prior distribution	averaged posterior distributions
Edge length	1	0.238 ± 0.0048	0.244 ± 0.0033
	2	0.246 ± 0.0054	0.244 ± 0.0033
	3	0.249 ± 0.0051	0.248 ± 0.0036
	4	0.248 ± 0.0048	0.245 ± 0.0033
Number of insertions	1	2.017 ± 0.1077	2.11 ± 0.1008
	2	2.085 ± 0.1156	2.152 ± 0.1148
	3	2.049 ± 0.1143	2.115 ± 0.1097
	4	2.074 ± 0.1196	2.088 ± 0.1031
Number of deletions	1	2.034 ± 0.1066	2.130 ± 0.1010
	2	2.079 ± 0.1135	2.141 ± 0.1128
	3	2.043 ± 0.1089	2.086 ± 0.1026
	4	2.066 ± 0.1161	2.086 ± 0.1015

Table 4: **Comparison of prior distribution and averaged posterior distributions using MCMC.** Edge number is determined by connected taxon number. Numerical values are means ± standard errors. Each mean is calculated over 4000 independent and random data sets.

13 Web Appendix G: Proposal algorithm to update a new indel history on a single edge

Detailed algorithm

```
1. Set $i = 1$ and $t_0 = 0$.
2. Repeat
    (a) Sample $t_{\mathrm{toNext}}$ from an exponential distribution with rate $\eta_i = (n_{i-1} + 1)\lambda + f(n_{i-1})\mu$, where $f(x) = \sum_{k=1}^{x}(x - k + 1)d(k)$.
    (b) If $t_{i-1} + t_{\mathrm{toNext}} > t$, go to step 3 (get out of step 2).
    (c) Set $t_i = t_{i-1} + t_{\mathrm{toNext}}$.
    (d) Set $id_i =$ in with probability $\frac{(n_{i-1}+1)\lambda}{\eta_i}$. Otherwise $id_i =$ del.
    (e) Check whether $id_i$ equals in.
        i. If yes, propose an insertion event.
            A. Sample $l_i$ from $q_{\mathrm{in}}$, a probability of proposing an insertion fragment size.
```

```
                B. Set $n_i = n_{i-1} + l_i$.
                C. Sample $p_i$ from a Unif$\{0, 1, \ldots, n_{i-1}\}$, which denotes a discrete uniform
                   distribution on a set $\{0, 1, \ldots, n_{i-1}\}$.
            ii. Otherwise, propose a deletion event.
                A. Sample $l_i$ from $q_{\text{del}}$, a probability of proposing a deletion fragment size.
                B. Set $n_i = n_{i-1} - l_i$.
                C. Sample $p_i$ from a Unif$\{0, 1, \ldots, n_{i-1} - l_i\}$.
        (f) Set $i = i + 1$.
3. Compare $n_{i-1}$ to $n_v$.
        (a) If $n_{i-1} < n_v$, propose one additional insertion event.
            i. Sample $t_i$ from a Unif($t_{i-1}$, $v$), which indicates a continuous uniform dis-
               tribution over an open interval ($t_{i-1}$, $v$).
            ii. Set $id_i$ = in, $l_i = n_v - n_{i-1}$, $n_i = n_v$.
            iii. Sample $p_i$ from a Unif$\{0, 1, \ldots, n_{i-1}\}$.
            iv. Set $K = i$.
        (b) If $n_{i-1} > n_v$, propose one additional deletion event.
            i. Sample $t_i$ from a Unif($t_{i-1}$, $v$).
            ii. Set $id_i$ = del, $l_i = n_{i-1} - n_v$, $n_i = n_v$.
            iii. Sample $p_i$ from a Unif$\{0, 1, \ldots, n_{i-1} - l_i\}$.
            iv. Set $K = i$.
        (c) If $n_{i-1}$ equals $n_v$, set $K = i - 1$.
4. Set $n_{K+1} = n_v$ and $t_{K+1} = v$.
```

14 Web Appendix H: Proposal algorithm to update a new indel history on a single edge considering the sequence length at the child node

All of these modifications are at step 2 of section 13. Modified step 2 is described as follows.

Detailed algorithm

```
Repeat
    1. If $n_{i-1}$ (current sequence length) equals $n_v$, go to the end (get out of this loop)
       with probability $\exp(-\eta_1(\frac{v-t_0}{c_1}))$ where $\exp(-\eta_1(v - t_0))$ is a probability of having
       no events on the remaining edge and $c_1$ is a tuning parameter.
    2. Sample $t_{\text{toNext}}$ from an exponential distribution with rate $\eta_i = (n_{i-1}+1)\lambda + f(n_{i-1})\mu$,
       where $f(x) = \sum_{k=1}^{x}(x - k + 1)d(k)$.
    3. If $t_{i-1} + t_{\text{toNext}} > v$, go to the end (get out of this loop).
    4. Set $t_i = t_{i-1} + t_{\text{toNext}}$.
    5. Set $p_{\text{I}} = \frac{(n_{i-1}+1)\lambda}{\eta_i}$.
```

6. Compare n_{i-1} to n_v.
    (a) If $n_{i-1} > n_v$, $p_{\text{I}} = p_{\text{I}} \times c_2$ where c_2 is a tuning parameter.
    (b) If $n_{i-1} < n_v$, $p_{\text{I}} = 1 - (1 - p_{\text{I}}) \times c_2$.
7. Set $id_i =$ in with probability p_{I}. Otherwise $id_i =$ del.
8. Check whether id_i equals in.
    (a) If yes, propose an insertion event.
        i. Check whether $0 < n_v - n_{i-1} < c_3$, where c_3 is a tuning parameter.
            A. If yes, set $l_i = n_v - n_{i-1}$ and then go to step 8.1.2 with probability c_4, which is tuning parameter.
            B. Sample l_i from q_{in}, a probability of proposing an insertion fragment size.
        ii. Set $n_i = n_{i-1} + l_i$.
        iii. Sample p_i from a Unif$\{0, 1, \dots, n_{i-1}\}$.
    (b) Otherwise, propose a deletion event.
        i. Check whether $0 < n_{i-1} - n_v < c_3$.
            A. If yes, set $l_i = n_{i-1} - n_v$ and then go to step 8.2.2 with probability c_4.
            B. Sample l_i from q_{del}, a probability of proposing a deletion fragment size.
        ii. Set $n_i = n_{i-1} - l_i$.
        iii. Sample p_i from a Unif$\{0, 1, \dots, n_{i-1} - l_i\}$.
9. Set $i = i + 1$.

Proposal probability

For a given edge of length v with n_0 and n_v bases at the parent node and the child node, the probability of proposing a history h of K indel events under this procedure, $Q_2(h \mid v, n_0, n_v)$ is

$$\begin{cases} \exp(-\eta_1(\frac{v-t_0}{c_1})) + (1 - \exp(-\eta_1(\frac{v-t_0}{c_1}))) \exp(-\eta_1(v - t_0)) & \text{if } K = 0 \\ \prod_{i=1}^{K-1} \mathsf{P}(e_i \mid t_{i-1}, n_{i-1}, v, n_v) \mathsf{P}(e_K \mid t_{K-1}, n_{K-1}, v, n_v) & \text{if } K > 0. \end{cases}$$

where $\mathsf{P}(e_i \mid t_{i-1}, n_{i-1}, v, n_v) = \exp(-\eta_i(t_i - t_{i-1}))A_i$. A_i and $\mathsf{P}(e_K \mid t_{K-1}, n_{K-1}, v, n_v)$ are calculated as follows.

$\langle$ *Calculation of A_i* $\rangle$

If $id_i =$ in,

$$A_i = \begin{cases} c_2 \lambda q_{\text{in}}(l_i) & \text{if } n_{\text{diff}} < 0 \\ (1 - \exp(-\eta_1(\frac{v - t_{i-1}}{c_1}))) \lambda q_{\text{in}}(l_i) & \text{if } n_{\text{diff}} = 0 \\ \frac{B_i(c_4 I_{\{n_{\text{diff}} = l_i\}} + (1-c_4) q_{\text{in}}(l_i))}{n_{i-1}+1} & \text{if } 0 < n_{\text{diff}} < c_3 \\ \frac{B_i q_{\text{in}}(l_i))}{n_{i-1}+1} & \text{if } n_{\text{diff}} \geq c_3, \end{cases}$$

where $n_{\text{diff}} = n_v - n_{i-1}$ and $B_i = \eta_i - f(n_{i-1})\mu c_2$.
If $id_i = \text{del}$,

$$A_i = \begin{cases} \frac{c_2 \mu f(n_{i-1}) q_{\text{del}}(l_i)}{n_{i-1} - l_i + 1} & \text{if } n_{\text{diff}} < 0 \\ \frac{(1 - \exp(-\eta_1(\frac{v - t_{i-1}}{c_1})))\mu f(n_{i-1}) q_{\text{del}}(l_i)}{n_{i-1} - l_i + 1} & \text{if } n_{\text{diff}} = 0 \\ \frac{D_i(c_4 I_{\{n_{\text{diff}} = l_i\}} + (1 - c_4) q_{\text{del}}(l_i))}{n_{i-1} - l_i + 1} & \text{if } 0 < n_{\text{diff}} < c_3 \\ \frac{D_i q_{\text{del}}(l_i))}{n_{i-1} - l_i + 1} & \text{if } n_{\text{diff}} \geq c_3, \end{cases}$$

where $n_{\text{diff}} = n_{i-1} - n_v$ and $D_i = \eta_i - (n_{i-1} + 1)\lambda c_2$.
$\langle$ *Calculation of* $\mathsf{P}(e_K \mid t_{K-1}, n_{K-1}, v, n_v) \ \rangle$
$\mathsf{P}(e_K \mid t_{K-1}, n_{K-1}, v, n_v)$ is

$$\frac{\exp(-\eta_K(v - t_{K-1}))}{(v - t_{K-1})(N_{K-1} - l_K I_{\{id_K = \text{del}\}} + 1)}$$
$$+[R + (1 - R)\exp(-\eta_K(v - t_K))]\frac{\exp(-\eta_K(t_K - t_{K-1}))U}{N_{K-1} - l_K I_{\{id_K = \text{del}\}} + 1},$$

where $R = \exp(-\eta_K(\frac{v - t_K}{c_1}))$ and U is

$$\begin{cases} B_K(c_4 I_{\{n_v - n_{K-1} < c_3\}} + (1 - c_4 I_{\{n_v - n_{K-1} < c_3\}}) q_{\text{in}}(l_K)) & \text{if } id_K = \text{in} \\ D_K(c_4 I_{\{n_{K-1} - n_v < c_3\}} + (1 - c_4 I_{\{n_v - n_{K-1} < c_3\}}) q_{\text{del}}(l_K)) & \text{if } id_K = \text{del}. \end{cases}$$

15 Web Appendix I: Multiple alignment with maximal expected accuracy under our model

One interesting point is that the objective function adopted in FSA is quite suitable in our setting, in which the goal is to summarize alignment samples from the joint model posterior $\mathsf{P}(H, \tau, V, \Theta \mid S)$. Bradley *et al.* (2009) introduce two restrictions to use only pairwise inference of alignment probabilities to approximate complete models of sequences evolving on trees. We review the authors' theoretical justification of their approach, and then point out that one of the two restrictions can be removed in our setting. We remark that this optimal alignment ideally minimizes the expected distance to a random alignment selected from the posterior distribution for some distance metric on alignments. The method implemented in FSA approximates this by seeking an alignment that minimizes a function of pairwise homology probabilities.

15.1 Theoretical justification of distance-based alignment (FSA)

Bradley *et al.* (2009) provide a theoretical justification that their distance-based approach to the multiple alignment problem, which uses only pairwise alignment probabilities, can be viewed as an approximation to more complex models of multiple alignment evolving on trees. First, we review their justification as follows. The goal is to find the optimal multiple alignment of sequences $X_1, \ldots, X_N$ related by a phylogenetic tree T. FSA seeks to find the alignment with maximal expected accuracy, which is defined as an alignment with minimum expected distance to the truth. Thus, the optimal alignment defined by FSA is

$$A_{optimal} = \text{argmin}_{A^*} E[d(A^*, A \mid T)]P(A \mid X_1, \ldots, X_N, T),$$

where $d(A^*, A \mid T)$ denotes a distance between two multiple alignments of sequences related by a tree T and $P(A \mid X_1, \ldots, X_N, T)$ indicates a posterior probability of a multiple alignment A given a tree T. To define the distance $d(A^*, A \mid T)$ and the probability $P(A \mid X_1, \ldots, X_N, T)$ using pairwise comparisons, the authors made two restrictions. First, they define the distance $d(A^*, A \mid T)$ as a weighted sum of pairwise distances, i.e.,

$$d(A^*, A \mid T) \ = \ \sum_{i,j} w_{ij}(T) d(A^*_{ij}, A_{ij}),$$

where A_{ij} denotes a pairwise alignment of sequences X_i and X_j and a distance $d(A^*_{ij}, A_{ij})$ between two alignments is defined as the number of characters for which they make different homology statements, taking into account both matches and gaps. They also use a pairwise approximation to the full probabilistic model, i.e.,

$$\sum_{A|A_{ij}} P(A \mid X_1, \ldots, X_N, T) \ = \ P(A_{ij} \mid X_i, X_j),$$

where $P(A_{ij} \mid X_i, X_j)$ is a posterior probability of a pairwise alignment A_{ij} and $A \mid A_{ij}$ refers to a multiple alignment A constrained to contain the pairwise alignment A_{ij}. Therefore, their objective function is

$$\begin{aligned}
&E[d(A^*, A \mid T)]P(A \mid X_1, \ldots, X_N, T) \\
&\quad = \ \sum_{A} P(A \mid X_1, \ldots, X_N, T) d(A^*, A \mid T) \\
&\quad = \ \frac{1}{\binom{N}{2}} \sum_{i,j} \sum_{A_{ij}} \sum_{A|A_{ij}} P(A \mid X_1, \ldots, X_N, T) d(A^*, A \mid T) \\
&\quad = \ \sum_{i,j} w_{ij}(T) \sum_{A_{ij}} d(A^*_{ij}, A_{ij}) P(A_{ij} \mid X_i, X_j).
\end{aligned}$$

In our setting, in which the goal is to summarize alignment samples from a joint model, the optimal alignment could be defined as

$$A_{optimal} = \text{argmin}_{A^*} E[d(A^*, A)]P(A \mid X_1, \ldots, X_N),$$

where a distance between two multiple alignments and a posterior probability of a multiple alignment are defined not for a given tree but by summing over all possible trees. The two restrictions could be modified as follows. The distance $d(A^*, A)$ is defined as a sum of pairwise distances, i.e.,

$$d(A^*, A) \;=\; \sum_{i,j} d(A^*_{ij}, A_{ij}),$$

and the pairwise approximation to the full probabilistic model is expressed as

$$\sum_{A|A_{ij}} P(A \mid X_1, \ldots, X_N) = P(A_{ij} \mid X_i, X_j). \tag{3}$$

Then, we still have the same objective function, which is

$$\begin{aligned} E[d(A^*, A)]P(A \mid X_1, \ldots, X_N) \;&=\; \sum_A P(A \mid X_1, \ldots, X_N) d(A^*, A) \\ &=\; \frac{1}{\binom{N}{2}} \sum_{i,j} \sum_{A_{ij}} \sum_{A|A_{ij}} P(A \mid X_1, \ldots, X_N) d(A^*, A) \\ &=\; \sum_{i,j} \sum_{A_{ij}} d(A^*_{ij}, A_{ij}) P(A_{ij} \mid X_i, X_j). \end{aligned}$$

We can see that (3) is not restricted in our setting.

16 Web Appendix J: A comparison among methods using simulated data

To evaluate the importance of alignment uncertainty in the estimation of phylogeny, we compare the posterior topology distribution from BayesCAT and the traditional sequential approach using the simulated data set which was introduced in Section 7.

This data set is generated by the procedure described in section 12, except that the parameters Θ are fixed to values in Table 5 instead of simulating them in step 1 of the procedure. We use the geometric distribution for the deletion fragment size. The true tree and sequence alignment for the simulated data are displayed in Figure 6 and 7 (a), respectively.

In what follows, Section 16.1 compares the posterior distributions of the tree topology from the various methods. Then, Section 16.2 presents summaries of alignment samples from the two joint estimation methods and compares the alignment distribution of the two methods. Section 16.3 summarizes information about the indel process. Finally, we conclude with a discussion.

Parameter	Value
r	0.02
r_d	0.25
r_i	0.265
λ	0.05
μ	0.054
γ	6.5
κ	2
$(\pi_A, \pi_C, \pi_G, \pi_T)$	(0.18, 0.32, 0.17, 0.33)

Table 5: **Parameter settings for the simulated data set.**

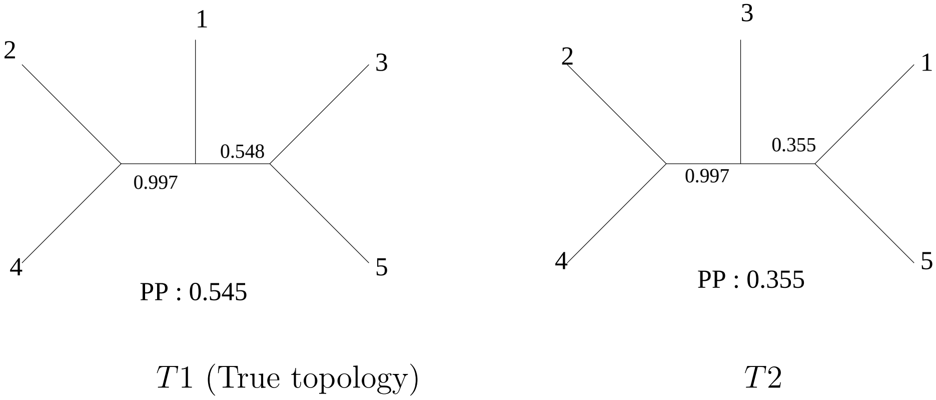


Figure 12: **Inferred topologies with a posterior probability (PP) > 0.1 from BayesCAT.**

16.1 Phylogeny estimation

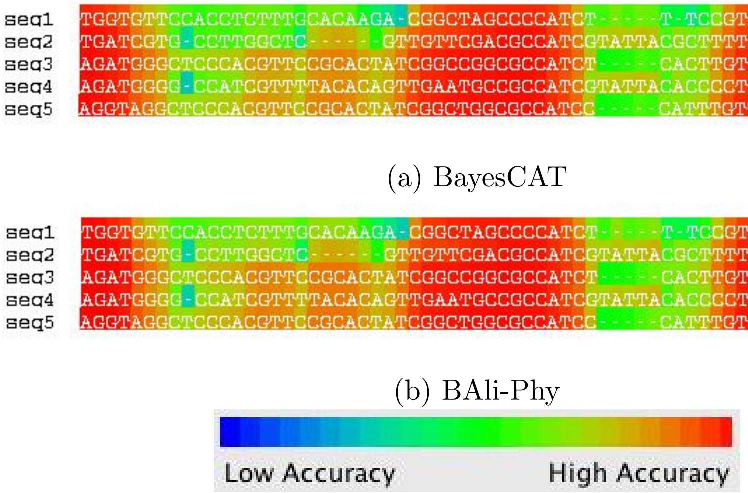
Figure 12 shows topologies with posterior probability greater than 0.1 using BayesCAT. The first topology $T1$ is the true topology. These two topologies are also the two with highest posterior probability from BAli-Phy (not shown) and the traditional sequential approach with three fixed alignments as shown in Figure 8. Table 6 lists the posterior probabilities from those five methods. The traditional sequential approach strongly supports the true topology when the true alignment (0.713) or the alignment estimated using ClustalW (0.88) is used. In the two joint models, as we expect, this high support is moderated (BayesCAT: 0.545 and BAli-Phy: 0.361) since alignment uncertainty is taken into account. The traditional sequential approach using different fixed alignments results in great disparity in estimates of the posterior probability of the true topology (the true alignment: 0.713, ClustalW: 0.88, and Muscle: 0.433). The joint models provide improved

Method	$T1$	$T2$	others
BayesCAT	0.545	0.355	0.100
BAli-Phy	0.361	0.347	0.292
MrBayes+true alignment	0.713	0.234	0.053
MrBayes+ClustalW	0.880	0.103	0.017
MrBayes+Muscle	0.433	0.485	0.082

Table 6: **Simulated data : Summary of posterior distributions of the topology.** $T1$ is the true topology.

estimates by considering uncertainty in the alignment, although different assumptions in the two joint models lead to dissimilar supports for the true topology.

16.2 Summary of alignment samples



(a) BayesCAT

(b) BAli-Phy

Figure 13: **Simulated data : Summary of alignment samples.** Alignment samples from BayesCAT (a) and BAli-Phy (b) are summarized using the procedure described in the main text.

Alignment samples from BayesCAT and BAli-Phy are summarized in Figure 13 using the procedure described in the main text. Although BAli-Phy provides its own summarization method, we use the same summarization procedure for both programs to focus the comparison on the alignment distributions and not the summarization methods. The two point estimates under the different joint models are exactly identical and have almost

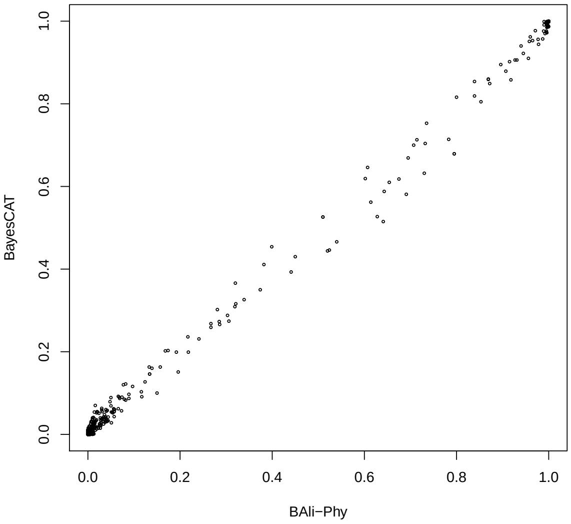


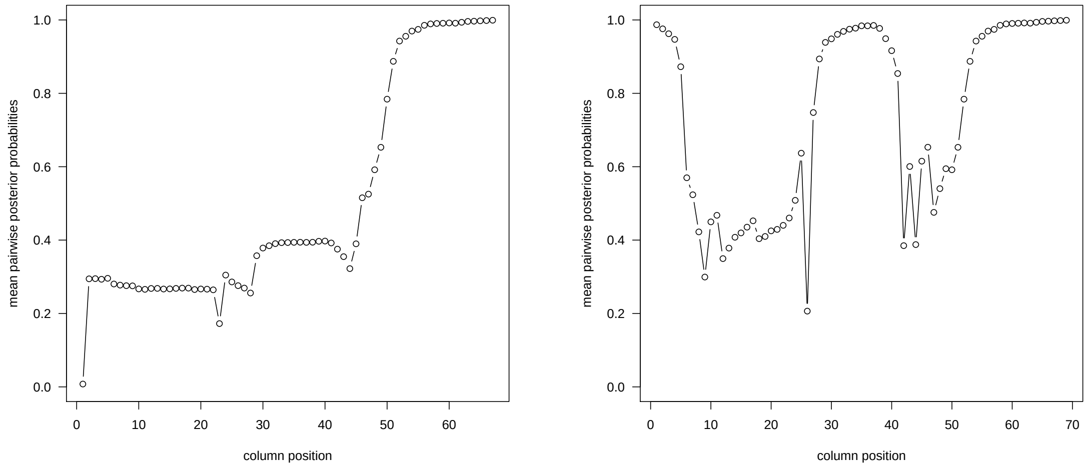
Figure 14: **Simulated data : Comparison of two alignment distributions (BayesCAT and BAli-Phy) using pairwise homology posterior probabilities.** Each point represents the posterior probabilities of homology for a pair of bases, one from one sequence and one from another, or between a base from one sequence and a gap. Many points are plotted at (0, 0) and (1, 1).

---

the same expected accuracy. We plot the pairwise homology posterior probabilities from each method in Figure 14. All points are around a diagonal, which is evidence that the alignment distributions from the two methods are very similar.

The alignment estimate contains a group of columns with high accuracy (red) and a second group with lower accuracy (other colors). All red columns appear in the true alignment as well as the alignment estimated using Muscle (Figure 7). Bases with dark orange color also show a similar homology relationship in both alignments. The red/dark orange regions show where the alignment is more certain.

In contrast, the alignment determined by ClustalW (Figure 7) includes only some red columns at the right part of the alignment in the summary of our distribution. To investigate the probability of the alignment estimated using ClustalW under our model, we plot the average pairwise homology posterior probability for each pair of bases (including pairs of gap and base) within each column of the alignment in Figure 15 (a). Only the last seventeen columns have high mean pairwise posterior probabilities (> 0.8) and are identical to the red columns at the right part of the alignment in our summary. The ClustalW alignment includes many columns which include pairs of bases that our analysis suggests have low probability of homology.



(a) The ClustalW alignment.

(b) The true alignment.

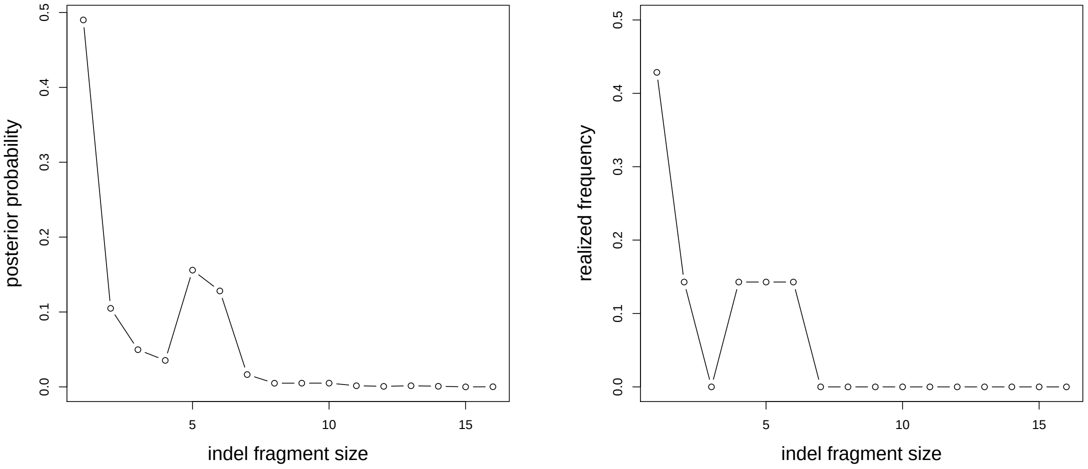
Figure 15: **Simulated data : Mean pairwise homology posterior probabilities of the ClustalW alignment and the true alignment.** We plot the average pairwise homology posterior probability for each pair of bases (including pairs of gap and base) within each column of the alignment estimated using ClustalW (a) and the true alignment (b).

Figure 15 (b) shows the similar plot for the true alignment. The columns with high mean pairwise posterior probabilities agree with all of the red columns in the alignment of our summary. Only six columns have mean pairwise posterior probabilities less than 0.4. Although our method shows uncertainty in the alignment of some regions, it still supports the true homologies in those regions with moderate amounts of accuracy.

16.3 Information on the indel process

Since we model indel events directly, some information about the indel process can be inferred by using our approach and not by BAli-Phy. Figure 16 (a) shows the posterior estimate of realized indel fragment size distribution, which is obtained by first collecting empirical indel fragment size distributions from each sample, and then averaging over all samples. This distribution has modes at sizes one and five. The realized indel fragment size distribution obtained from the true indel history is shown in Figure 16 (b). There are strong similarities between the two distributions in Figure 16, which indicates that the true indel history is similar to the sample of indel histories from our method.

Another quantity we can estimate is the number of indel events on each split as shown in Table 7. The edges leading to leaf one, leaf two, and a clade containing leaves two and



(a) Posterior estimate of realized indel fragment size distribution.

(b) The true realized indel fragment size distribution.

Figure 16: **Simulated data : Posterior estimate of realized indel fragment size distribution (BayesCAT).**

four include more than one indel event and the remaining splits contain fewer than one indel event on average. In this analysis, rounding the mean number of indels for each split in the true tree to the nearest integer matches the true number of indels on the corresponding edge. This is additional evidence that the true indel history is typical of those sampled by our MCMC method.

To investigate whether the expectation of the number of indel events vary with branch length, we also list the posterior mean edge length given occurrence of each split in the fourth column of Table 7. Edges with more than one indel event are longer than the remaining edges, but the estimated number of indel events on each edge is not always proportional to the branch length. The edge leading to a clade including leaves three and five (0.501) contains more indel events than the edge leading to leaf four (0.163), but the former (0.128) is shorter than the latter (0.164).

16.4 Conclusions

The traditional sequential approach using different fixed alignments of the simulated sequences results in great disparity in estimates of the posterior probability of the true topology. The joint models provide a solution by accounting for uncertainty in the alignment in phylogenetic inferences. In this example, two different joint estimation approaches provide moderately different posterior probabilities of the true topology, but their poste-

Split	PP of split	# of indels	edge length
1 | 2,3,4,5	1	2.74 (3)	0.536 (0.483)
2 | 1,3,4,5	1	1.42 (1)	0.350 (0.272)
3 | 1,2,4,5	1	0.001 (0)	0.085 (0.069)
4 | 1,2,3,5	1	0.163 (0)	0.164 (0.146)
5 | 1,2,3,4	1	0.001 (0)	0.071 (0.106)
1,2 | 3,4,5	0	-	-
1,3 | 2,4,5	0.11	0	0.035
1,4 | 2,3,5	0.002	0	0.178
1,5 | 2,3,4	0.347	0.006	0.067
2,3 | 1,4,5	0	-	-
2,4 | 1,3,5	0.998	2.14 (2)	0.323 (0.194)
2,5 | 1,3,4	0	-	-
3,4 | 1,2,5	0	-	-
3,5 | 1,2,4	0.543	0.501 (1)	0.128 (0.097)
4,5 | 1,2,3	0	-	-

Table 7: **Simulated data : Posterior mean number of indel events on each split (BayesCAT).** The second column lists posterior probabilities for each split. The posterior means of the number of indel events and the edge length given occurrence of each split are shown in the third and fourth columns, respectively. The true number of indel events and the true edge length are shown in parentheses for splits in the true tree.

rior alignment distributions are quite similar. In addition, the true realized indel history is typical among those sampled using BayesCAT both in fragment size distribution and location on the tree.

References

Bradley, R. K., Roberts, A., Smoot, M., Juvekar, S., Do, J., Dewey, C., Holmes, I. and Pachter, L. (2009) Fast Statistical Alignment. *PLoS Computational Biology*, **5**, e1000392.

Cartwright, R. A. (2009) Problems and Solutions for Estimating Indel Rates and Length Distributions. *Molecular Biology and Evolution*, **26**, 473–480.

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

Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160–174.

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

Shim, H. (2010) Bayescat : Bayesian co-estimation of alignment and tree. *PhD Thesis, Department of Statistics, University of Wisconsin at Madision.*

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, **22**, 4673–4680.