PNAS RESEARCH ARTICLE | EVOLUTION

Correlations between alignment gaps and nucleotide substitution or amino acid replacement

Tae-Kun Seo[a,1], Benjamin D. Redelings[b,c,d], and Jeffrey L. Thorne[e,f,1]

Edited by David Hillis, The University of Texas at Austin, Austin, TX; received March 12, 2022; accepted July 11, 2022

To assess the conventional treatment in evolutionary inference of alignment gaps as missing data, we propose a simple nonparametric test of the null hypothesis that the locations of alignment gaps are independent of the nucleotide substitution or amino acid replacement process. When we apply the test to 1,390 protein alignments that are informed by protein tertiary structure and use a 5% significance level, the null hypothesis of independence between amino acid replacement and gap location is rejected for ∼65% of datasets. Via simulations that include substitution and insertion–deletion, we show that the test performs well with true alignments. When we simulate according to the null hypothesis and then apply the test to optimal alignments that are inferred by each of four widely used software packages, the null hypothesis is rejected too frequently. Via further simulations and analyses, we show that the overly frequent rejections of the null hypothesis are not solely due to weaknesses of widely used software for finding optimal alignments. Instead, our evidence suggests that optimal alignments are unrepresentative of true alignments and that biased evolutionary inferences may result from relying upon individual optimal alignments.

gaps | insertion | deletion | substitution

Significance

We introduce a test of the null hypothesis that nucleotide substitution or amino acid replacement processes are independent of gap locations within sequence alignments. When applying this test to alignments that are informed by protein structure, the null is rejected about 2/3 of the time. This indicates that modifications are needed to the usual approach of ignoring gap locations when making evolutionary inferences. Additionally, we demonstrate that optimal alignments introduce spurious correlations between gap locations and nucleotide substitution patterns. Because these spurious correlations will not be eliminated by employing genomic-scale datasets, we emphasize the need for modifying the conventional approach of basing evolutionary inferences upon single optimal alignments.

DNA and protein alignments are evolutionary hypotheses about the positional correspondence between sequences. Gaps in alignments can arise from multiple sources, including historical insertion or deletion events. Sometimes, alignment gaps exist simply because data corresponding to certain sequences have not been acquired. When the sequences being compared are collected on the basis of their being transcribed or being both transcribed and translated, other sources of alignment gaps are possible (e.g., frameshift mutations and alternative splicing). The possibility of a gap being due to alignment error should also be considered.

In one of his pioneering contributions to molecular phylogenetics, Fitch (1) wrote "either the character was not examined or it does not exist" while explaining that the treatment of a gapped alignment position should be affected by whether the gap corresponds to uncollected data or whether it stems from insertion or deletion events. Fitch wrote that passage at a time when available sequence data were scarce and when there was therefore a substantial disincentive to discarding phylogenetic information that might be associated with insertions or deletions. Because sequence data are now far less scarce and because it is difficult to extract the evolutionary information that stems from insertion and deletion events in a statistically rigorous fashion, an attractive alternative might now be to ignore the evolutionary information that is associated with insertions and deletions and to instead rely solely on nucleotide substitutions or amino acid replacements.

When making evolutionary inferences from aligned sequences, one option is therefore to consider gaps as missing data. The treatment of gap locations as being independent of nucleotide substitution or amino acid replacement is widespread when analyzing aligned data with likelihood-based techniques, and the associated computations are detailed in Felsenstein (2) and Yang (3). This likelihood-based treatment of gaps as missing data was inspired by a corresponding handling of gaps in parsimony-based phylogenetics.

The statistical consequences of this conventional handling of alignment gaps warrant careful attention. When gaps represent uncollected data, the gap locations are not necessarily independent of the missing sequence information. Specifically, uncollected data can be substantially more diverged than observed data. For example, divergence at a primer location could cause certain loci in certain taxa to fail to amplify via PCR (e.g., ref. 4). The resulting ascertainment bias will be magnified if there is a strong correlation between primer divergence and the lengths of the branches that connect the uncollected loci to the phylogenies relating the loci that are sequenced. A similar potential ascertainment bias needs to be considered for divergence at restriction sites that leads to uncollected RAD-seq data (see ref. 5).

Author contributions: T.-K.S. and J.L.T. designed research; T.-K.S. and B.D.R. analyzed data; and T.-K.S., B.D.R., and J.L.T. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: seo.taekun@gmail.com or thorne@ncsu.edu.

Published August 16, 2022.

When gaps are caused by historical insertion and deletion events, the conventional treatment can also be problematic because of the possibility that the insertion and deletion processes are correlated with the nucleotide substitution or amino acid replacement processes. Lack of independence would mean that the presence of a gap provides information about substitution or amino acid replacement that should be reflected in the evolutionary inferences. Lack of independence between these processes is biologically plausible because natural selection will affect persistence and fixation of both point mutations and insertion or deletion mutations. The degree of selective constraint in a genomic region is likely to be correlated for these different kinds of mutations. For example, both amino acid replacement processes (e.g., refs. 6, 7) and insertion–deletion processes (e.g., ref. 8) are correlated with protein structure, and this correlation is presumably largely attributable to selective constraint. Zheng et al. (9) find a strong positive correlation between deletion and amino acid replacement in mammalian protein sequences. Beyond the correlation due to natural selection, additional correlation may stem from the mutation process. For example, Tian et al. (10) provide evidence that heterozygosity for insertion–deletion polymorphisms may be mutagenic (see also ref. 11).

Here we introduce a simple nonparametric test of the null hypothesis that the presence of gaps is independent of the amount of nucleotide substitution or amino acid replacement. If alignment uncertainty can be neglected and if gaps are exclusively attributable to historical insertion and deletion events, our test is an examination of whether the insertion–deletion processes are independent of the nucleotide substitution or amino acid replacement processes.

When we simulate according to the null hypothesis of independence between nucleotide substitution and insertion–deletion and then analyze the data using the simulated alignments, our simple test performs well in that the cumulative probability of the test statistic is approximately uniform. When we simulate according to the null hypothesis of independence and then infer the optimal alignments with widely used software packages, all of the aligners tend to generate optimal alignments with misleading signals of dependence between nucleotide substitution and insertion–deletion. Next, we apply our test to datasets where protein tertiary structure is employed to guide sequence alignment. At a significance level of 0.05, the test rejects the null hypothesis for 908 of 1,390 datasets. These investigations emphasize the need to improve evolutionary inference tools so that dependence between insertion–deletion and nucleotide substitution or amino acid replacement can be accommodated, and they reinforce the conclusion that the conventional practice of making evolutionary inferences from single optimal alignments can be problematic (e.g., ref. 12).

Theory and Methods

Nonparametric Test of Independence between Substitution and Alignment Gaps. Our test can be applied to alignments of either protein or DNA sequences, but we describe it with regard to DNA and the nucleotide substitution process. Consider an alignment with n columns that will each be grouped into exactly one of three categories according to the proportion of gap characters in the column. The number of columns in these groups will be n_1, n_2, and n_3 where $n = n_1 + n_2 + n_3$. The n_1 columns that have a low proportion of gaps will be termed "fewer-gaps" columns. The n_2 columns with higher proportions of gaps will be termed the "more-gaps" columns. The n_3 columns where only one taxon has a nucleotide and the other taxa all have gaps will be referred to as "uninformative" because these columns can do little to illuminate patterns of nucleotide substitution. We determine the sizes n_1 and n_2 so that size imbalance (e.g., $|n_1 - n_2|^2$) between n_1 and n_2 is minimized. Table 1 shows an example of five taxa in which $n = 17$ columns are separated into $n_1 = 9$, $n_2 = 6$, and $n_3 = 2$ columns.

Table 1. Example alignment between five sequences

	Site index	
	1	
Taxon name	123456789	012345 67
T1	AAAAAAAAA	AAAAAA --
T2	AATATATAA	-TAAT- --
T3	ATTATATTT	ATA-TA --
T4	TTAAATTAT	T----- --
T5	TAAA-----	---T-A AA

Columns 1-9 with gap proportions of 0 and 0.2 are grouped into the 'fewer-gaps' category. Columns 10-15 with gap proportion 0.4 are grouped into the 'more-gaps' category. Columns 16-17 are in the 'uninformative' category and are ignored in the subsequent analysis.

Denote s_{ik} as the entry for the ith taxon in the kth alignment column of the fewer-gaps category. This means that s_{ik} will either represent one of the four nucleotide types or a gap. We define p_{ijk} as

$$p_{ijk} := \begin{cases} 1 & \text{if } s_{ik} \neq s_{jk} \text{and neither is gap,} \\ 0 & \text{if} s_{ik} = s_{jk} \text{and neither is gap,} \\ - \text{ (undefined)} & \text{otherwise.} \end{cases}$$

For each pair i and j, $p_{ij\cdot}$ is defined as

$$p_{ij\cdot} := \begin{cases} \frac{\sum_{k=1}^{n_1} p_{ijk} I(p_{ijk} \neq -)}{\sum_{k=1}^{n_1} I(p_{ijk} \neq -)} & \text{if } \sum_{k=1}^{n_1} I(p_{ijk} \neq -) \neq 0, \\ - \text{ (undefined)} & \text{otherwise,} \end{cases}$$

where $I(\cdot)$ is 1 if the condition within the parentheses is satisfied and is 0 otherwise. In other words, $p_{ij\cdot}$ considers only alignment columns in the fewer-gaps category where both taxa have nucleotides and is the proportion of these columns where the residues differ. We note that the $p_{ij\cdot}$ terms are not independent among different combinations of (i, j) due to shared common ancestry. However, our nonparametric test accounts for phylogenetic correlations without relying upon an explicit model of evolutionary change (see the bootstrap approach below).

We define q_{ijk} for the columns in the more-gaps category in a way that parallels the p_{ijk} definition for the fewer-gaps category. Likewise, we will use the q_{ijk} to calculate $q_{ij\cdot}$ in the same way as the p_{ijk} determine $p_{ij\cdot}$. Table 2 shows the $p_{ij\cdot}$ and $q_{ij\cdot}$ values that are derived from the example depicted in Table 1.

Our test statistic is

$$\begin{aligned} T &= \frac{\sum_{i<j} \{q_{ij\cdot} - p_{ij\cdot}\} I(q_{ij\cdot} \neq - \text{ and } p_{ij\cdot} \neq -)}{\sum_{i<j} I(q_{ij\cdot} \neq - \text{ and } p_{ij\cdot} \neq -)} \\ &= \sum_{i<j} w_{ij}\, q_{ij\cdot} - \sum_{i<j} w_{ij}\, p_{ij\cdot} =: \overline{q_{ij\cdot}} - \overline{p_{ij\cdot}}, \end{aligned} \qquad [1]$$

where

$$w_{ij} := \frac{I(q_{ij\cdot} \neq - \text{ and } p_{ij\cdot} \neq -)}{\sum_{i<j} I(q_{ij\cdot} \neq - \text{ and } p_{ij\cdot} \neq -)},$$

and T compares how likely nucleotide types are to differ in the more-gaps category relative to how likely they are to differ in the fewer-gaps category. Thus, $\overline{p_{ij\cdot}}$ and $\overline{q_{ij\cdot}}$ are averages over sequence

Table 2. A pairwise comparison matrix to illustrate calculations based on the example alignment in Table 1

		p_{ijk} and q_{ijk}			
			1		
Row number	Taxa pair	123456789	012345	$p_{ij\cdot}$	$q_{ij\cdot}$
1	T1–T2	001010100	-1001-	3/9	2/4
2	T1–T3	011010111	010-10	6/9	2/5
3	T1–T4	110001101	1-----	5/9	1
4	T1–T5	1000-----	---1-0	1/4	1/2
5	T2–T3	010000011	-00-0-	3/9	0
6	T2–T4	111011001	------	6/9	—
7	T2–T5	1010-----	---1--	2/4	1
8	T3–T4	101011010	1-----	5/9	1
9	T3–T5	1110-----	-----0	3/4	0
10	T4–T5	0100-----	------	1/4	—

The average of $\{q_{ij\cdot} - p_{ij\cdot}\}$ over the eight rows where both $p_{ij\cdot}$ and $q_{ij\cdot}$ are defined (i.e., excluding rows 6 and 10) yields a test statistic value of $T \simeq 0.0569$.

pairs of the difference proportions in the fewer-gaps and more-gaps categories. The numerator of w_{ij} ensures that sequence pairs only contribute to these averages if the difference proportion is defined for both the fewer-gaps and more-gaps categories. The denominator of w_{ij} counts the number of sequence pairs where the difference proportion is defined in both categories.

For the null hypothesis that the proportion of gaps in a column is independent of the substitution process, T should be close to 0. If gaps are exclusively due to insertion or deletion events, then the null hypothesis is that the substitution process is independent of the insertion–deletion process, and significant positive (negative) values of T would imply a positive (negative) correlation between substitution and insertion–deletion rates.

To assess the significance of the deviation of T from 0, we use a bootstrap approach to approximate the null distribution of T. From the original alignment of n columns, we resample n columns with replacement to create each resampled dataset. By doing so, we make use of the common assumption that alignment columns are independently and identically distributed. Because the units being resampled are alignment columns, the resampled datasets reflect the phylogenetic correlations among sequences possessed by the original data. Thus, the bootstrap allows the null distribution of T to be approximated without an explicit evolutionary model and even though the calculation of the test statistic averages values from nonindependent pairwise comparisons.

For each resampled dataset, we apply our classification rules to assign each column into the fewer-gaps, more-gaps, or uninformative category. The number of columns in the three categories are denoted n_1^*, n_2^* and n_3^* with $n_1^* + n_2^* + n_3^* = n$ and where an asterisk indicates a quantity from the resampled data. We then calculate the test statistic T^*. If the null hypothesis is true, the expected value of T (i.e., $E[T]$) is 0, and the distribution of T minus 0 can be well approximated by the distribution of T^* minus its expected value $E[T^*]$. Following the guideline of bootstrap centering (13), we therefore approximate the null distribution of $\{T - 0\}$ with that of $\{T^* - E[T^*]\}$. Then, denoting the test statistic value from the rth resampled dataset as $T^{*(r)}$, we approximate the cumulative probability of the test statistic T under the null hypothesis as

$$F(T) = \text{Prob}(\{T^* - E[T^*]\} < \{T - 0\}) \approx \frac{1}{B}\sum_{r=1}^{B} I(\{T^{*(r)} - \overline{T^*}\} < \{T - 0\}), \quad [2]$$

where B is the number of bootstrap resampled datasets and where the sample mean of the resampled test statistic values is

$$\overline{T^*} = \frac{1}{B}\sum_{r=1}^{B} T^{*(r)}.$$

When the test is two-sided, $F(T) < 0.025$ and $F(T) > 0.975$ represent the critical regions for rejecting the null hypothesis at a 5% significance level.

Simulation Design: Null and Alternative Hypothesis Experiments. To investigate our nonparametric test, we designed simulation studies in which 12 taxa were related by the phylogeny shown in Fig. 1. For simplicity, all internal and terminal branches of the Fig. 1 phylogeny share an identical length. We used the INDELible program (14) to simulate nucleotide substitutions as well as insertions and deletions. Because INDELible has the insertion–deletion process be independent of the substitution process, our simulated datasets were generated by having INDELible simulate two partitions of sequences according to the phylogeny of Fig. 1 and then concatenating the sequences of the two partitions. At the root node, the length of each partition was set to 500 nucleotides. For the substitution process, our simulations used the Jukes–Cantor model (15). For all simulations, the length distributions of both insertion and deletion events were geometric, and both distributions had a mean length of 5/3 nucleotides in both partitions.

To satisfy the null hypothesis of independence between nucleotide substitution and insertion–deletion, we set the lengths of each branch in each partition to 0.1 nucleotide substitutions per site. For the first partition, the rates of insertion and deletion events relative to substitution were 0.08 and 0.12, respectively. For the second partition, the rates of insertion and deletion events relative to substitution were 0.12 and 0.18, respectively. Therefore, our datasets that were simulated according to the null hypothesis exhibited regional heterogeneity of insertion–deletion rates, but this heterogeneity was not linked to any variation in substitution rates.

To generate datasets that violate the null hypothesis of our test, all settings were identical to those used when the null hypothesis was satisfied with the exception that the first partitions were

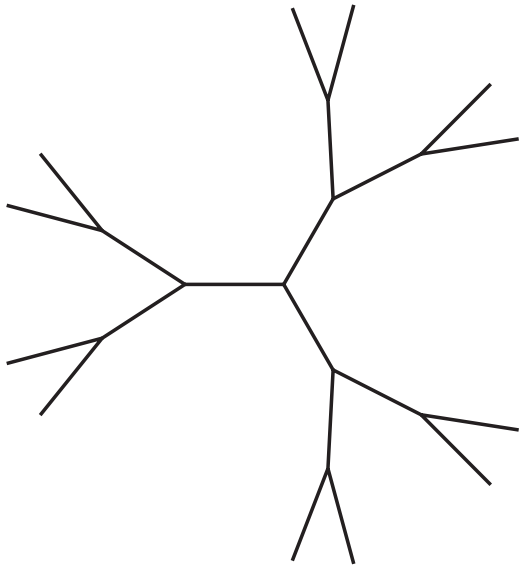

Fig. 1. Phylogeny of 12 taxa that was used in simulations. The central internal node was treated as the root.

simulated with all branches having 0.05 rather than 0.1 expected substitutions per site. Due to these shorter substitution lengths for the first partition as well as to the lower relative rates of insertion and deletion for the first partition, these simulated datasets had a positive correlation between the rates of substitution and insertion–deletion.

Because INDELible reports both the true simulated alignment and the nonaligned simulated sequences at the tips of the phylogeny, we were able to investigate the performance of our test for the true alignment and for the optimal alignments that were inferred by widely used software packages. We examined the test behavior using optimal alignments from Clustal Omega (version 1.2.2) (16), MAFFT (version 7.475) (17), Muscle (version 3.8.31) (18), and Prank (version 170427) (19). Our experiments with these sequence aligners were not intended to assess the relative merits of the four software packages. Their performance will be influenced both by the simulation conditions and the choice of analysis options. Instead, our motivation was to determine whether the optimal alignments from these packages might yield substantially different outcomes with our nonparametric test than is observed for the true alignments. All alignments from these software packages were produced using default settings.

Simulation Design: Probabilistic Alignment Experiment. Substantially different outcomes were observed when applying our test to the true simulated alignments and when applying our test to the optimal alignments from the four software packages. Our test performed well with the true alignments but rejected the null hypothesis too often when using inferred alignments (*Results*). Therefore, we sought to understand whether the differences can be completely explained by a disconnect between the evolutionary process used to simulate data and the default settings of the four aligners. With this motivation, we designed a simple simulation experiment that aimed to have settings of alignment software closely match the settings of the INDELible software that generated the data. This allows us to compare the performance of our test with true simulated alignments to its performance with optimal alignments that were inferred when closely matching the simulation and alignment parameter values. To do this, we used version 4.0-alpha4 of the model-based BAli-Phy software (20–22) to analyze the simulated sequence data.

For this experiment, INDELible simulated datasets of three sequences that were equally diverged from a common ancestral root sequence of 1,000 nucleotides. Each of the three branches emanating from the root had substitutions occur according to the Jukes–Cantor model (15) and had lengths of both insertion and deletion events be geometrically distributed with a mean of 5/3 nucleotides. We simulated and analyzed 1,000 different datasets under each of two simulation conditions. In the low-substitution scenario, branches had an expected 0.1 substitutions per site, and the rates of insertions and deletions relative to substitutions were each 0.5. In the high-substitution scenario, branches had an expected 0.5 substitutions per site, and the relative rates of insertions and deletions were both 0.1. We intentionally designed the low-substitution and high-substitution scenarios so that they would not differ in the expected amounts of insertion and deletion but would differ in the expected amounts of substitution.

BAli-Phy analyses specified the Jukes–Cantor model and specified (rather than estimated) the actual expected numbers of substitutions per site per branch. BAli-Phy analyses also specified the true length distribution for insertion and deletion events as well as the true rates of insertion and deletion relative to substitution. With these settings, BAli-Phy only had to infer the alignments between the simulated sequences. For each simulated dataset, two 1,000-generation Markov chain Monte Carlo (MCMC) runs were performed.

To assess MCMC convergence on each dataset, we computed the effective sample size (ESS) for each logged scalar variable. This includes the alignment length (|A|), the number of indels on the tree (#indels), the total length of indels (|indels|), and the maximum parsimony score (#substs), in addition to the log-prior, log-likelihood, and log-posterior. For each variable in each dataset, we computed a combined ESS by concatenating the results of the two MCMC runs, skipping the first 100 generations of each run as burn-in. (If the two runs yield differing posterior distributions, the combined ESS can be much lower than the ESS of an individual run.) We then recorded the variable in each dataset with the lowest ESS. The mean (and SD) of this minimum ESS across the 1,000 datasets was 706.2 (110.5) for the low-substitution scenario and 970.6 (145.9) for the high-substitution scenario. This indicates that the MCMC chains converged and were mixing well.

We also used BAli-Phy to construct a single optimal alignment for each simulated dataset. This alignment was constructed from samples of the posterior distribution of alignments. From the two MCMC runs per analyzed dataset, we treated the first 200 generations of each run as a burn-in of the Markov chain and then sampled an alignment every generation during the remainder of each run. Because there were two MCMC runs per dataset, we considered a total of $1{,}600 = 800 + 800$ sampled alignments for each dataset.

We constructed the optimal alignment from the 1,600 posterior samples for each dataset via posterior decoding. Posterior decoding means that the optimal alignment is chosen by maximizing a combination of the posterior probabilities of individual alignment columns, instead of maximizing the joint probability of those columns (which is sometimes called Viterbi decoding). Different posterior decoding options of BAli-Phy represent different ways of combining these (estimated) posterior probabilities of individual alignment columns into a single optimal sequence alignment. We selected the BAli-Phy option that finds the alignment that maximizes the product of the (estimated) posterior probabilities of the columns in the alignment (see also ref. 23).

In addition to the single posterior decoding alignment that was obtained from each analyzed dataset, we also considered a single alignment that was randomly sampled from the posterior distribution. For this, we used the alignment at the final generation of the first set of the two 1,000-generation MCMC runs that were performed with each simulated dataset.

Database Analyses. We applied our nonparametric test to alignments of protein sequences that are in the Balibase (version 4, updated 12 December 2016) (24), Homstrad (version 2/1/2021) (25), and Mattbench (version 1.0) (26) databases. We selected these three databases because they employ protein tertiary structure information to derive alignments. Although alignments based on tertiary structure do not necessarily reflect positional correspondence between sequences that is due to common ancestry (e.g., see ref. 19) and although alignment uncertainty cannot be eliminated by incorporating information from tertiary structure, tertiary structures tend to evolve very slowly relative to protein sequences (e.g., ref. 27). Therefore, tertiary structural information can substantially reduce alignment uncertainty, and this can facilitate interpretation of the results from our hypothesis test.

These three databases contained a total of 2,034 alignments. With our test, pairwise alignment columns with a gap are categorized as uninformative, and pairwise alignment columns with a match or mismatch would be assigned to the fewer-gaps category. Therefore, pairwise alignments cannot have columns in the

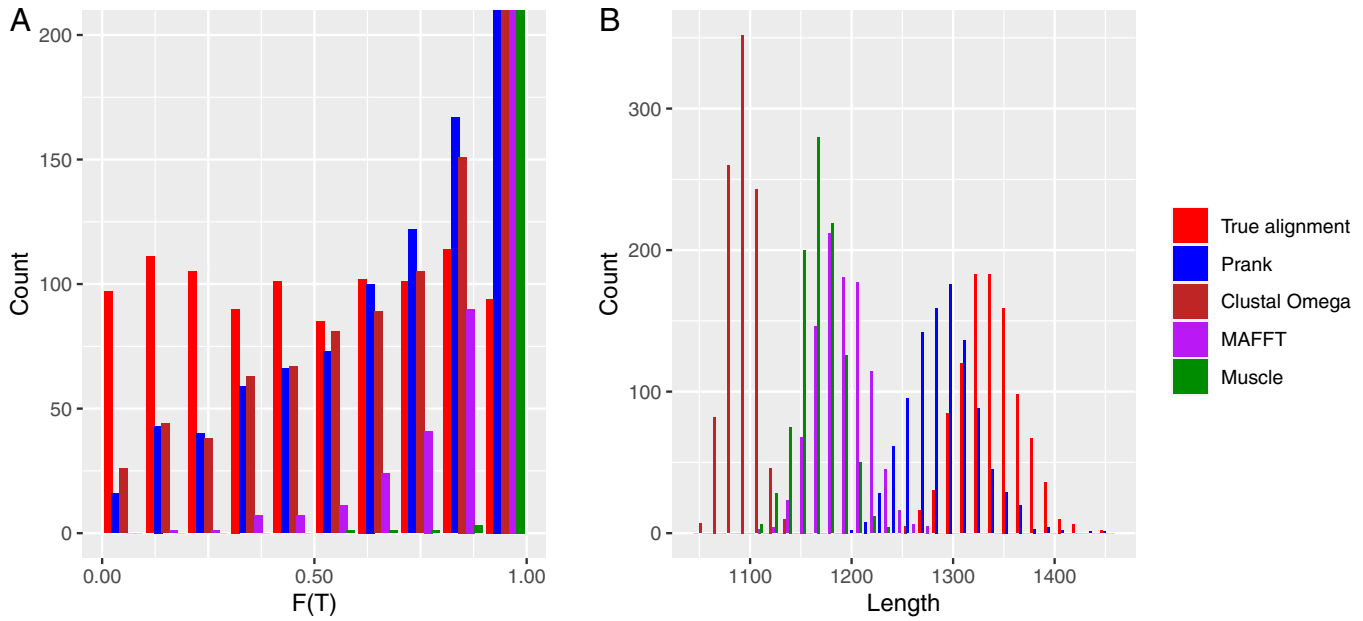

Fig. 2. Distribution of (*A*) F(T) and (*B*) the aligned sequence length of simulated data. Colors indicate the source of analyzed alignments. Truncated counts of the range $F(T) > 0.9$ for Prank, Clustal Omega, MAFFT, and Muscle were 314, 336, 818, and 994, respectively.

more-gaps category (i.e., $n_2 = 0$ for pairwise alignments). This means that our test cannot be applied to pairwise alignments. Out of the 2,034 alignments from the three databases, 629 were pairwise alignments and could not be analyzed with our test. An additional 15 alignments could not be analyzed because no columns could be assigned to the more-gaps category (i.e., $n_2 = 0$ for these datasets). We applied the nonparametric test to the remaining 1,390 alignments from the three databases.

Results

Results from Null Hypothesis Experiment. Using 1,000 simulated datasets and their corresponding true alignments, we investigated whether $F(T)$ of Eq. 2 has the expected uniform distribution between 0 and 1 when the null hypothesis is true. With the true alignment, the distribution of $F(T)$ closely approximates a Uniform(0,1) distribution (Fig. 2*A*). The mean and SD of the 1,000 $F(T)$ values were 0.501 and 0.292, respectively. These are close to the mean of 0.5 and SD of 0.289 for a Uniform(0,1) distribution. The number of $F(T)$ values belonging to the 5% rejection region, which corresponds to the cases of $F(T) < 0.025$ or $F(T) > 0.975$, was 56 out of 1,000, and this is close to the expected 50 out of 1,000 for a 5% significance level.

In addition to the behavior of our test when the null hypothesis of independence was true and there was no alignment error, we also investigated the test when the null hypothesis was true but the true alignment was not used and instead optimal alignments from four different software packages were employed. Although the null hypothesis was true, the distributions of $F(T)$ that result from the optimal alignments of all four aligners deviate from the expected uniform distribution and instead are shifted toward 1 (Fig. 2*A*). This implies a tendency for the T statistic of Eq. **1** to be positive, and it suggests that the optimal alignments incorporate positive dependency between the substitution and insertion–deletion processes. The number of cases rejecting the null hypothesis with a 5% significance level for Prank, Clustal Omega, MAFFT, and Muscle were 137, 167, 628, and 970, respectively.

Fig. 2*B* shows the distribution of the number of columns in the true alignments as well as the distributions in the optimal alignments from the default settings of the four aligners. The mean (SD) of the number of columns for the true alignments is 1,340.0 (30.2). In contrast, all four aligners produced optimal alignments that were too short. This tendency to overalign (i.e., put unrelated nucleotides in the same column) was strongest for Clustal Omega even though it produced a less skewed distribution of $F(T)$ values than MAFFT or Muscle (Fig. 2*A*). The mean (SD) of the number of columns in the optimal alignments for Prank, Clustal Omega, MAFFT, and Muscle were 1,294.2 (33.4), 1,091.4 (14.4), 1,187.0 (25.9), and 1,164.7 (20.6), respectively.

Results from Alternative Hypothesis Experiment. For the datasets that were simulated according to the aforementioned scenario that satisfies the alternative hypothesis of a positive correlation between nucleotide substitution and gap presence, we do not expect the distribution of $F(T)$ to be uniform. For the true alignments, the distribution of $F(T)$ is shifted to the right so that the mean of $F(T)$ exceeds the value of 0.5 that is expected under the null hypothesis. The mean and SD of the 1,000 $F(T)$ values from true alignments are 0.975 and 0.0652, respectively. With the true alignments, the null hypothesis of independence between substitution process and gap presence was rejected at a 5% significance level for 779 of the 1,000 datasets that were simulated according to the alternative hypothesis. For the optimal alignments from the four alignment software packages, the distributions of $F(T)$ values were shifted even further to the right than for the true alignments. The number of cases rejecting the null hypothesis with a 5% significance level for alignments from Prank, Clustal Omega, MAFFT, and Muscle were 937, 866, 993, and 1,000, respectively.

For the datasets simulated according to the alternative hypothesis, the mean and SD of the number of columns for the true alignments were 1,272.5 and 26.3, respectively. The four aligners show a tendency to overalign for these simulated datasets, although the tendency was not extreme for PRANK. The mean (SD) of the number of columns in the optimal alignments for

Table 3. Pattern counts from the BAli-Phy analyses of simulated data

Substitution	Alignment	XXX	XXY	XYZ	XXG	XYG	XGG	Total
Low	True	579.8 (18.0)	185.9 (12.7)	12.5 (3.5)	167.5 (14.8)	35.9 (6.3)	258.7 (24.4)	1,240.3 (24.2)
	Randomly sampled	575.0 (18.1)	185.2 (12.9)	12.5 (3.5)	171.8 (16.5)	36.6 (6.3)	265.2 (25.2)	1,246.3 (23.6)
	Posterior decoding	621.1 (16.0)	195.1 (12.9)	12.3 (3.4)	131.1 (13.8)	25.3 (5.2)	201.9 (20.9)	1,186.7 (20.7)
High	True	203.9 (14.1)	434.2 (16.7)	139.9 (11.2)	91.0 (10.8)	112.8 (12.1)	257.7 (25.3)	1,239.5 (24.9)
	Randomly sampled	203.3 (13.0)	435.6 (17.4)	140.3 (11.1)	90.1 (11.2)	112.3 (12.7)	256.8 (23.6)	1,238.4 (22.2)
	Posterior decoding	306.7 (10.6)	464.3 (17.3)	123.3 (10.1)	45.3 (7.9)	37.0 (6.9)	151.9 (19.9)	1,128.4 (18.5)

The table shows the mean (and SD) among the 1,000 simulated datasets of the pattern counts for the low-substitution and high-substitution simulation scenarios with three alignment types. The "Total" column shows the means (and SDs) of the number of columns per alignment.

Prank, Clustal Omega, MAFFT, and Muscle were 1,260.5 (34.0), 1,076.8 (12.8), 1,157.9 (22.0), and 1,142.6 (18.4), respectively.

Results from Probabilistic Alignment Experiment. We summarize the BAli-Phy analyses by classifying each column of three-sequence alignments into one of six categories: XXX columns have three identical nucleotide types; XXY columns have two nucleotides of one type and another with a different type, regardless of which of the three sequences has a different type from the others; XYZ columns have three different types; XXG columns have two identical types and one gap symbol; XYG columns have two different types and one gap symbol; and XGG columns have one nucleotide and two gap symbols. For the specific case simulated here where substitution is independent of insertion–deletion and where the three branches are equally long and follow the Jukes–Cantor model, the counts of these pattern categories are sufficient statistics for inferring the shared branch length from an alignment. Therefore, reconstructed alignments with pattern counts that are close to those of the true alignment are likely to provide a good basis for evolutionary inference.

Table 3 shows that the mean pattern counts are quite similar between the true and randomly sampled alignments. This strong similarity can be attributed to the fact that the BAli-Phy and INDELible parameters were intentionally set to closely correspond. Because the INDELible and BAli-Phy insertion–deletion treatments are close but not identical, there are some small but presumably real differences in pattern counts between the true and randomly sampled alignments. Whereas the pattern counts from the true and randomly sampled alignments are quite similar, they substantially differ from those of the posterior decoding alignments. This suggests that the randomly sampled alignments are preferable to the posterior decoding alignments for the purposes of evolutionary inference.

Because the three-sequence datasets were simulated according to the null hypothesis that substitution and insertion–deletion processes are independent, the $\overline{q_{ij\cdot}}$ and $\overline{p_{ij\cdot}}$ values of Eq. **1** should be similar. Table 4 shows that the means of these two statistics are very similar for the true alignments. This is also the case for randomly sampled alignments. Furthermore, randomly sampled alignments behave almost identically to true alignments with regard to SDs of $\overline{q_{ij\cdot}}$ and $\overline{p_{ij\cdot}}$. However, the behaviors of these statistics for the posterior decoding alignments are quite different from their distributions from true and from randomly sampled alignments (Table 4).

The average $F(T)$ when applying our independence test is close to the expected 0.5 for both the true and randomly sampled alignments (Table 4). In contrast, the posterior decoding alignments generate values of $\overline{q_{ij\cdot}}$ and $\overline{p_{ij\cdot}}$ that are too small on average relative to the true alignments. The posterior decoding alignments yield mean $F(T)$ values that are markedly less than the expected 0.5. Whereas the performance of the independence test is satisfactory for true and randomly sampled alignments, the 81 incorrect rejections of the null hypothesis for the low-substitution scenario with posterior decoding alignments (Table 4) are significantly different from the expected 5% rate of incorrect rejections (two-tailed exact binomial test, $P < 0.0001$). The same conclusion applies to the 89 incorrect rejections with posterior decoding alignments for the high-substitution scenario.

Results from Database Analyses. Fig. 3*A* is a histogram of the $F(T)$ values from the 1,390 structurally informed alignments that were analyzed with our test. The histogram deviates from the uniform distribution that would be expected if the null hypothesis were true for all datasets. The concentration of $F(T)$ values near 1 is consistent with a positive correlation between amino acid replacement rates and rates of insertion and/or deletion. The null hypothesis was rejected at a 5% significance level for 908 of the 1,390 datasets ($\simeq 65.3\%$). Fig. 3*B* shows that the tendency for $F(T)$ to be near 1 increases with alignment length.

Discussion and Conclusion

Our simple hypothesis test relies upon aligned DNA or protein sequences to examine the null hypothesis that positions of gaps within alignments are independent of the nucleotide substitution or amino acid replacement processes. When this null hypothesis

Table 4. Statistics from the BAli-Phy analyses of simulated data

Substitution	Alignment	$\overline{q_{ij\cdot}}$	$\overline{p_{ij\cdot}}$	$F(T)$	No. of rejections
Low	True	0.176 (0.0268)	0.175 (0.0107)	0.502 (0.287)	46
	Randomly sampled	0.176 (0.0264)	0.176 (0.0111)	0.489 (0.288)	61
	Posterior decoding	0.162 (0.0308)	0.172 (0.0101)	0.402 (0.292)	81
High	True	0.554 (0.0357)	0.552 (0.0128)	0.515 (0.293)	50
	Randomly sampled	0.555 (0.0351)	0.553 (0.0124)	0.518 (0.290)	46
	Posterior decoding	0.450 (0.0560)	0.484 (0.0091)	0.339 (0.265)	89

The table shows statistics from the BAli-Phy analyses of the low-substitution and high-substitution simulation scenarios with three alignment types. As defined in Eq. **1**, $\overline{q_{ij\cdot}}$ and $\overline{p_{ij\cdot}}$ represent proportions of differences among nongap characters in the more-gaps and fewer-gaps categories, respectively. The entries in the $\overline{q_{ij\cdot}}$ and $\overline{p_{ij\cdot}}$ columns are averages from the 1,000 simulated datasets. The $F(T)$ column shows the mean $F(T)$ of Eq. **2** from applying the test to the 1,000 cases. The number of rejections column lists the number of times out of 1,000 that the null hypothesis was incorrectly rejected at a significance level of 0.05. Parenthesized numbers are SDs of the 1,000 values.

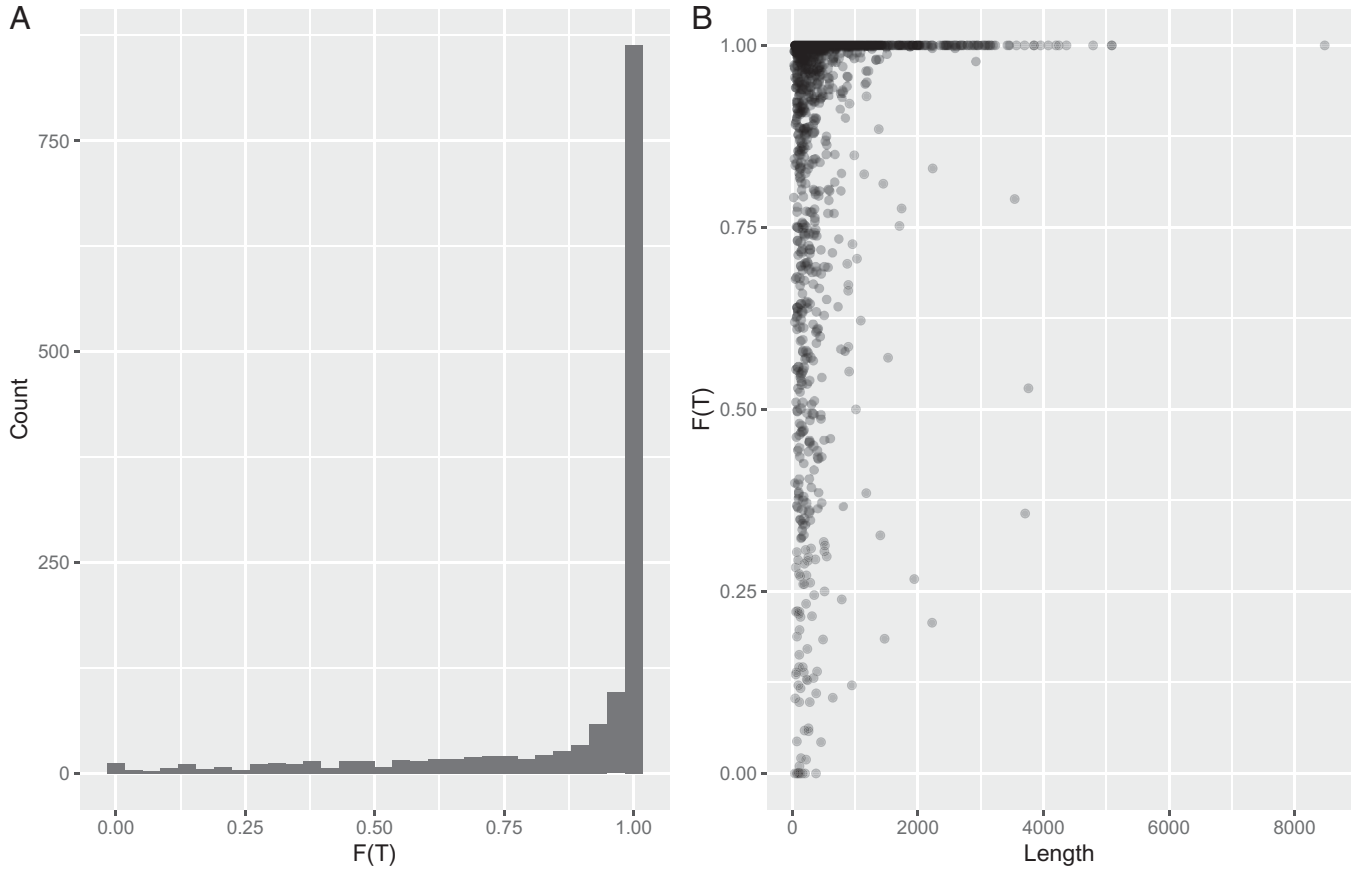

Fig. 3. Distributions of $F(T)$ from protein alignments that are informed by protein tertiary structure. (*A*) A histogram of the $F(T)$ values obtained by applying the hypothesis test to 1,390 aligned datasets. (*B*) A plot of alignment length (*x* axis) versus $F(T)$ (*y* axis) for the 1,390 datasets.

is correct and when alignment gaps can be attributed to insertions or deletions, the conventional treatment of gaps as missing data is problematic in the sense that it ignores evolutionary information from insertion and deletion events. However, when the null hypothesis is correct, the conventional treatment will not bias evolutionary inferences that are based solely on nucleotide substitution or amino acid replacement. One motivation for testing the null hypothesis is therefore to assess whether evolutionary inferences might be biased due to the conventional treatment of gap locations as being independent of the substitution or amino acid replacement processes.

We intentionally designed our hypothesis test to be nonparametric. A parametric test would be likely to be more powerful and could have the attractive feature of incorporating phylogenetic structure. However, a cost of the additional power is likely to be sensitivity to violations of model assumptions. Seo et al. (28) introduced a model adequacy test for analysis of aligned sequence data. That model adequacy test could be modified to examine the null hypothesis that our nonparametric test evaluates. We chose to instead explore the nonparametric test behavior specifically because of its simple and nonparametric nature.

When we simulated data according to the null hypothesis, our nonparametric test performed well. Similarly, our nonparametric test had the desired behavior of frequently rejecting the null hypothesis when the null hypothesis was violated for simulated data. When our test was applied to datasets consisting of proteins that were aligned on the basis of tertiary structure, the null hypothesis was often rejected. Two possible causes of these frequent rejections warrant particular attention. Although the two causes are not mutually exclusive, we separately consider them here. The first possibility is that amino acid replacement and insertion–deletion are not independent. The second possibility is that alignment errors have artifactually introduced a signal of dependence between these processes.

Because we analyzed alignments that are informed by protein structure, we suspect that alignment errors are not a major reason for the high proportion of rejected null hypotheses from the databases of protein alignments. A positive correlation across protein regions is biologically plausible because natural selection is likely to impose correlated constraints across regions on amino acid replacement and insertion–deletion. Probabilistic models of sequence change that incorporate this positive correlation should more often be applied to evolutionary inference. However, evolutionary inference that includes insertion and deletion can be computationally challenging, and correlations between these processes and amino acid replacement will exacerbate already daunting inference challenges. A less ambitious alternative might be a framework where the conventional handling of gaps as missing sequence data is modified so that the abundance of gaps in an alignment column (or in an alignment region) is used as prior information concerning the relative rate of nucleotide substitution or amino acid replacement. This modification would be inspired by the pioneering framework for modeling substitution rate heterogeneity among sites that Yang introduced (29). Whereas the simplest version of Yang's framework has a discretized probability distribution of relative rates across sites with all rate categories having equal prior probability, a modification could have the local pattern of gap presence/absence influence the prior probabilities of the different relative rate categories that might affect a particular alignment column.

Rather than modifying evolutionary models, another option for dealing with a correlation between substitution and insertion–deletion processes would be to determine when inferences are robust to such correlations. The degree of robustness will depend on details of the evolutionary process as well as on the evolutionary history. In addition, robustness will vary among evolutionary inference tasks. For example, it may be that tree topology estimation is more robust than divergence time estimation. Characterizing robustness is clearly preferable to ignoring violated assumptions.

We showed with simulated data that the null hypothesis of independence between substitution and insertion–deletion is prone to being incorrectly rejected due to alignment errors (Fig. 2). Different alignment software packages exhibited different tendencies to incorrectly reject the null hypothesis, but we do not view our results as being helpful concerning the relative merits of these aligners. We did not explore a wide variety of simulation scenarios and did not attempt to adjust the default settings of the software packages. Our purpose was solely to demonstrate that inferred alignments can include errors that incorrectly signal evidence for a correlation between substitution and insertion-deletion.

Adjustments to program settings and/or adoption of other software might reduce the tendency for incorrect rejection of the null hypothesis, but the BAli-Phy analyses indicate that the tendency will be difficult to eliminate because optimality criteria are prone to favoring alignments that are unrepresentative of true alignments. With conventional alignment optimality criteria, the distribution of patterns in incorrectly aligned columns will differ from the distribution of patterns among true alignment columns. True alignments can have regions that are relatively improbable while still being possible. Conventional optimality criteria are likely to resolve the relatively improbable region of the true alignment with an incorrect albeit more probable scenario for that region. Conventional optimality criteria will make this decision about resolving alignment regions in a deterministic way such that the evolutionary signal in the resulting alignment may be unrepresentative of the signal in the true alignment.

While an alignment is an evolutionary hypothesis about positional correspondence between sequences, alignment inference is often an intermediate step when studying evolutionary process or history. A common practice is to extract the values of sufficient statistics from an inferred alignment to make likelihood-based inferences that are conditional upon the inferred alignment being correct. In such a situation, it may be worthwhile to design simulation studies that assess an alignment procedure not on the basis of how close is the inferred positional correspondence to the true alignment but instead on how close are the sufficient statistics in an inferred alignment likely to be to their true values.

Probabilistic treatments for analyzing unaligned sequences are rapidly improving (e.g., refs. 20, 30, 31). By marginalizing over alignments, these procedures are able to better reflect variance in parameter estimates due to alignment uncertainty. The genomic era has gifted evolutionary biology with large datasets that can lead to very small variances in parameter estimates. Although genome-scale datasets can greatly reduce variance, they do not eliminate bias. We showed here that optimal alignments can lead to biased sufficient statistics (Table 3). It may be that the biggest advantage to probabilistic treatments of alignments is their ability to avoid the biased inferences that emerge when relying upon alignments that satisfy conventional optimality criteria.

While it is certainly most desirable to base evolutionary inferences on all alignments or on a large sample from the posterior distribution of alignments, relying on a single random sample from this posterior distribution may yield better evolutionary inferences than relying upon a single optimal alignment. In some cases, inferring evolutionary parameters based on a single alignment sampled from the posterior can be nearly as accurate as inference based on the joint posterior of the alignment and evolutionary parameters (32). Representing posterior distributions by a single summary alignment can be problematic when uncertainty is resolved in the same direction at many different locations. For example, when two sequences contain different residues in the same location, it may be ambiguous whether these two residues should be aligned (creating a mismatch) or unaligned (creating two gaps). When both alternatives have the same posterior probability, posterior decoding tends to resolve them all in the same direction: either always voting in favor of substitutions or always in favor of gaps, depending on which version of posterior decoding is used. In such cases, relying on a single sample from the posterior distribution may yield better evolutionary inferences than an optimal alignment because ambiguities are not all resolved in the same direction. The relative benefit of using a single posterior sample or an optimal alignment may depend on many factors, including the evolutionary distance between the sequences, the evolutionary parameter to estimate, the number of sequences in the alignment, and the optimality criterion.

Alignment algorithms have traditionally been judged based on various measures of accuracy. For example Mirarab and Warnow (33) describe the sum-of-pairs (SP) score, which is the fraction of true pairwise homologies found in the estimated alignment. BAli-Phy implements a version of posterior decoding that maximizes the expected SP score by maximizing the expected number of true pairwise homologies. Such posterior decoding alignments should be more accurate than a randomly sampled posterior alignment according to their respective scores. However, evolutionary inferences based on posterior decoding alignments will still suffer from the bias that results from resolving equiprobable events in a consistent direction. Thus, there may be a difference between alignments that are most accurate and alignments that yield the most accurate estimates of evolutionary parameters.

Data, Materials, and Software Availability. Analyzed data and software have been deposited in GitHub (https://github.com/diploid2n/IND_TEST) (34).

ACKNOWLEDGMENTS. We thank David Hillis, Joe Felsenstein, and an anonymous reviewer for their comments on an earlier version of this manuscript. We also thank Joe Felsenstein for illuminating the history of treating gaps as missing data. T.-K.S. was supported by Korea Polar Research Institute (KOPRI) grants funded by the Ministry of Oceans and Fisheries (KOPRI PE22060, PE22140). B.D.R. was supported by NSF (grant DBI-1759838) and NIH (grant R01TW010870). J.L.T. was supported by NSF (grant DEB-1754142).

Author affiliations: [a]Division of Life Sciences, Korea Polar Research Institute, Yeonsu-gu, Incheon 21990, Republic of Korea; [b]Biology Department, Duke University, Durham, NC 27708; [c]Ronin Institute, Durham, NC 27705; [d]Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS 66045; [e]Department of Biological Sciences, North Carolina State University, Raleigh, NC 27695; and [f]Department of Statistics, North Carolina State University, Raleigh, NC 27695

1. W. Fitch, Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.* **20**, 406–416 (1971).
2. J. Felsenstein, *Inferring Phylogenies* (Sinauer Associates, Sunderland, MA, 2004), pp. 255–256.
3. Z. Yang, *Computational Molecular Evolution* (Oxford University Press, 2006), pp. 107–108.
4. Y. Guo, A. Pais, A. Weakley, Q. Xiang, Molecular phylogenetic analysis suggests paraphyly and early diversification of *philadelphus* (hydrangeaceae) in western North America: New insights into affinity with *carpenteria*. *J. Syst. Evol.* **51**, 545–563 (2013).
5. D. A. R. Eaton, E. L. Spriggs, B. Park, M. J. Donoghue, Misconceptions on missing data in rad-seq phylogenetics with a deep-scale example from flowering plants. *Syst. Biol.* **66**, 399–412 (2017).

6. J. M. Koshi, R. A. Goldstein, Context-dependent optimal substitution matrices. *Protein Eng.* **8**, 641–645 (1995).
7. J. L. Thorne, N. Goldman, D. T. Jones, Combining protein evolution and secondary structure. *Mol. Biol. Evol.* **13**, 666–673 (1996).
8. M. S. Taylor, C. P. Ponting, R. R. Copley, Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. *Genome Res.* **14**, 555–566 (2004).
9. Y. Zheng, D. Graur, R. B. R. Azevedo, Correlated selection on amino acid deletion and replacement in mammalian protein sequences. *J. Mol. Evol.* **86**, 365–378 (2018).
10. D. Tian *et al.*, Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* **455**, 105–108 (2008).
11. P. Sjödin, T. Bataillon, M. H. Schierup, Insertion and deletion processes in recent human history. *PLoS One* **5**, e8650 (2010).
12. J. L. Thorne, H. Kishino, J. Felsenstein, An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33**, 114–124 (1991).
13. P. Hall, S. Wilson, Two guidelines for bootstrap hypothesis testing. *Biometrics* **47**, 757–762 (1991).
14. W. Fletcher, Z. Yang, INDELible: A flexible simulator of biological sequence evolution. *Mol. Biol. Evol.* **26**, 1879–1888 (2009).
15. T. Jukes, C. Cantor, "Evolution of protein molecules" in *Mammalian Protein Metabolism*, H. N. Munro, Ed. (Academic Press, New York, 1969), pp. 21–123.
16. F. Sievers *et al.*, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
17. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
18. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
19. A. Löytynoja, N. Goldman, An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 10557–10562 (2005).
20. B. D. Redelings, M. A. Suchard, Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.* **54**, 401–418 (2005).
21. M. A. Suchard, B. D. Redelings, BAli-Phy: Simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics* **22**, 2047–2048 (2006).
22. B. D. Redelings, M. A. Suchard, Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. *BMC Evol. Biol.* **7**, 40 (2007).
23. G. Lunter, I. Miklós, A. Drummond, J. L. Jensen, J. Hein, Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics* **6**, 83 (2005).
24. J. D. Thompson, P. Koehl, R. Ripp, O. Poch, BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark. *Proteins* **61**, 127–136 (2005).
25. K. Mizuguchi, C. M. Deane, T. L. Blundell, J. P. Overington, HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Sci.* **7**, 2469–2471 (1998).
26. N. M. Daniels, A. Kumar, L. J. Cowen, M. Menke, Touring protein space with Matt. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* **9**, 286–293 (2012).
27. C. Chothia, A. M. Lesk, The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826 (1986).
28. T. K. Seo, O. Gascuel, J. L. Thorne, Measuring phylogenetic information of incomplete sequence data. *Syst. Biol.* **71**, 630–648 (2022).
29. Z. Yang, Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994).
30. I. Holmes, A model of indel evolution by finite-state, continuous-time machines. *Genetics* **216**, 1187–1204 (2020).
31. N. De Maio, The cumulative indel model: Fast and accurate statistical evolutionary alignment. *Syst. Biol.* **70**, 236–257 (2021).
32. B. Redelings, Erasing errors due to alignment ambiguity when estimating positive selection. *Mol. Biol. Evol.* **31**, 1979–1993 (2014).
33. S. Mirarab, T. Warnow, FastSP: Linear time calculation of alignment accuracy. *Bioinformatics* **27**, 3250–3258 (2011).
34. T.-K. Seo, B. D. Redelings, J. L. Thorne, diploid2n/IND_TEST. GitHub. https://github.com/diploid2n/IND_TEST. Deposited 2 August 2022.