WILEY-BLACKWELL

Cladistics 28 (2012) 514–528

Cladistics

10.1111/j.1096-0031.2012.00402.x

Indel information eliminates trivial sequence alignment in maximum likelihood phylogenetic analysis

John S.S. Denton[a,b,*] and Ward C. Wheeler[b,c]

[a]*Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA;* [b]*Richard Gilder Graduate School, American Museum of Natural History, New York, NY 10024, USA;* [c]*Division of Invertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA*

Accepted 12 March 2012

Abstract

Although there has been a recent proliferation in maximum-likelihood (ML)-based tree estimation methods based on a fixed sequence alignment (MSA), little research has been done on incorporating indel information in this traditional framework. We show, using a simple model on a single character example, that a trivial alignment of a different form than that previously identified for parsimony is optimal in ML under standard assumptions treating indels as "missing" data, but that it is not optimal when indels are incorporated into the character alphabet. We show that the optimality of the trivial alignment is not an artefact of simplified theory assumptions by demonstrating that trivial alignment likelihoods of five different multiple sequence alignment datasets exhibit this phenomenon. These results demonstrate the need for use of indel information in likelihood analysis on fixed MSAs, and suggest that caution must be exercised when drawing conclusions from software implementations claiming improvements in likelihood scores under an indels-as-missing assumption.

© The Willi Hennig Society 2012.

Maximum likelihood (ML) has become a popular optimality criterion for inferring evolutionary trees following its introduction by Fisher (1912), its initial application to nucleotide data (Neyman, 1971; Felsenstein, 1981), and its popularization through early software releases (Felsenstein, 1989; Swofford, 2002). ML is a parametric method that requires estimation of the entries of a stochastic character transition matrix as well as branch lengths on a phylogenetic tree. Due to the large number of values requiring independent optimization in an ML framework, computation times for likelihood tree searches are much longer than those for parsimony, and ML analyses have been made more time-consuming by increased dataset sizes, greater numbers of partitions, and by increased model complexity. Yet despite these additional computational burdens, likelihood analyses have become increasingly tractable in the past decade with improvements in both computing speed and public access to multiprocessor clusters, as well as in the increasing sophistication of search heuristics in ML software (e.g. Jobb et al., 2004; Guindon et al., 2005, 2010; Stamatakis, 2006; Zwickl, 2006).

However, despite this increase in sophistication, the fundamental procedure remains unchanged. A gap cost, whether non-affine (Waterman et al., 1976), affine (Gotoh, 1982; Edgar, 2004; Katoh et al., 2005; Larkin et al., 2007), logarithmic (Mott, 1999), or log-affine (Cartwright, 2007) is employed during multiple sequence alignment via progressive algorithms, but most often not during topology search. The pattern of the inserted indels ("missing" data, in a phylogenetic context) may influence the topology found as optimal by either increasing the number of most parsimonious trees (Wheeler, 1994; Wiens, 1998, 2003a,b; Kearney, 2002) or by flattening the likelihood surface. When indels are treated as missing data in the character alphabet on a fixed alignment, adding more indels may improve the optimality score to the point where a "trivial" alignment, in which sequences are aligned completely out of phase, is mathematically optimal, i.e. having a parsimony cost of zero (Wheeler and Gladstein, 1994;

*Corresponding author:
E-mail address: jdenton@amnh.org

© The Willi Hennig Society 2012

Wheeler et al., 1995; Giribet and Wheeler, 1999). Trivial alignments are characterized by two undesirable properties: (a) they have zero cost (in a parsimony framework) despite contributing no biologically justified homology statements, and (b) they are invariant with respect to tree topology. Although the assumption of treating indels as "missing" data in the character alphabet remains standard practice in phylogenetic analyses, little work has been done to investigate the conceptual implications of this assumption under the likelihood criterion.

Several methods for incorporating indel information into phylogenetic analyses exist, among them likelihood-based sequence alignment (Bishop and Thompson, 1986; Thorne et al., 1991, 1992); simultaneous alignment and topology inference by Bayesian criteria (Redelings and Suchard, 2005, 2007; Suchard and Redelings, 2006), by likelihood criteria (Wheeler, 2006), and by parsimony (Wheeler, 1996; Varón et al., 2010) criteria; and by incorporation of indels into the character alphabet for topology search on a fixed alignment (McGuire et al., 2001a; Young and Healy, 2003; Rivas and Eddy, 2008). Utilization of indel information on a static MSA eliminates trivial alignments because it defines indels on a direct transformation path, thereby maintaining metricity of the substitution cost matrix (Wheeler, 1993). Additionally, incorporating indel information has been observed to contribute to phylogenetic resolution and node support (Whiting et al., 1997; Egan and Crandall, 2008; Simmons et al., 2008; Dwivedi and Gadagkar, 2009; Dessimoz and Gil, 2010; Pas′ko et al., 2011). However, the use of indel information is limited in application outside parsimony by computational constraints, and so the limited number of existing model-based methods for use on nucleotide data apply the assumption of atomistic indel events to make analyses tractable. Although the utility of atomistic indels has been questioned in several studies under parsimony (Simmons and Ochoterena, 2000, for example), to date no model-based software implementations exist that provide non-atomistic alternatives with heuristics that operate in polynomial time. It is therefore imperative that such implementations be developed and made accessible so that the behaviours of the assumptions can be studied.

Context of the "indel problem"

Parameterizing insertions and deletions is a more complex problem than determining the best-fit model of nucleotide substitution. In likelihood analysis of nucleotide data, we wish to estimate the likelihood of sequence data (D) given a tree topology (T) and a stochastic model of evolution (Θ) of any alphabet size:

$$L(\Theta, T) \propto Pr(D|\Theta, T) \tag{1}$$

There are two ends of the spectrum of general approaches to this problem. In the broad formulation, a given alignment of the data (a_i) in the space of alignments (A) can be treated as a random variable and marginalized, in which case

$$L(\Theta, T) = \sum_{a_i \in A} Pr(D, a_i|\Theta, T) \tag{2}$$

and therefore

$$L(\Theta, T) = \sum_{a_i \in A} Pr(D|a_i, \Theta, T) \times Pr(a_i|\Theta, T) \tag{3}$$

by the chain rule for probability distributions. The first term in the right hand side of the above equation treats only substitutions among alphabet characters, whereas the second term accounts for differences in alignment probability due to differing numbers of insertion and deletion events. Because insertion and deletion events are accounted for by the second term in this formulation, they do not have meaning as an entry in the character alphabet of the continuous-time stochastic process.

By contrast, the more widely-applied and traditional formulation of the problem is that of utilizing eqn (1), with several additional assumptions. First, in this approach, a heuristic MSA is taken as a fixed parameter value for the alignment A, from which the tree topology is inferred by search algorithms. Hence, under this "two-step" formulation, the second term in the right-hand side of eqn (3) is omitted entirely. Second, under this traditional approach, the conditional state transition probabilities $p(j|i, t)$ for an alphabet size K in a $K \times K$ transition matrix P are derived by matrix exponentiation from an instantaneous rate matrix Q having the property that

$$\sum_j Q(i, j) = 0 \tag{4}$$

Because this common formulation requires all changes to be accounted for by substitutions among characters, insertion and deletion events have meaning as character values. To account for insertions and deletions in this framework, the $K \times K$ transition matrix is augmented to a $(K + 1) \times (K + 1)$ matrix to incorporate insertion and deletion events. Third, under the traditional two-step approach, the continuous-time stochastic process is assumed to be globally stationary, hence the equilibrium character frequencies π_K, obtained in the limit $t \to \infty$ for the entries of $\mathbf{P}$ are constant across the tree topology. Fourth, the stochastic process is assumed to be globally reversible, hence

$$\pi_i Q(i, j) = \pi_j Q(j, i) \tag{5}$$

and so insertions and deletions are estimated as a single ("indel") parameter. Lastly, the stochastic process is

assumed to be globally homogeneous, in that the same entries of **P** apply across all branches of the topology.

The main differences between the broad approach and the two-step approach discussed here go beyond ontological formulation of character states. The broad approach separates insertion and deletion events from the character alphabet on the conceptual ground that physical changes in sequence size are not part of the substitution process. The M01 model (McGuire et al., 2001b) embodies the traditional approach by augmenting the substitution matrix to include indels as a fifth state. By contrast, the RE08 model (Rivas and Eddy, 2008) approximates the broad formulation (eqn 2) by modelling insertions and deletions as non-reversible birth/death parameters separate from reversible character substitutions. Both the RE08 and M01 models operate on fixed alignments with atomistic characters, and are hence different from models like TKF91 and TKF92 (Thorne et al., 1991, 1992), which operate as ML alignment methods on initially ungapped sequences. The justification for a non-reversible birth/death indel approach lies in the claim that the traditional approach assumes sequences must have constant size and an indel character frequency that is constant with respect to divergence time. Additionally, under the traditional approach there is a purported "memory effect" (Felsenstein, 2004) among neighbouring columns that disallows insertions of lengths greater than the length of a deletion in a given region. The birth/death approach is thus favoured because of these objections.

However, the problems with a non-reversible insertion/deletion approach are threefold. First, a nonreversible (insertion/deletion) Markov process makes the likelihood computation dependent on the root location, and the likelihood-optimal solution in this context may disagree with the outgroup designation based on other forms of evidence, such as morphology or the stratigraphic record. Second, a zero-divergence indel frequency is biologically practical only at narrow scales of phylogenetic inference, which are rarely the goal of modern large-scale analyses. A root indel frequency is plausible as either an artefact of incomplete sampling or as a true reflection of large-scale sequence divergence. Forcing a zero-divergence across either broad or incomplete sampling is likely to significantly influence the estimated values. Third, an "evolutionary process" birth/death modelling of insertions and deletions necessarily requires estimation of a deletion rate separate from an insertion rate. The analogous macroevolutionary literature on diversity-dependent diversification using molecular phylogenies, which also relies on birth-death modelling to estimate speciation and extinction rates (Nee, 2006), has observed that an extinction (i.e. deletion) "rate" is likely not estimable (Rabosky, 2010) under a globally homogeneous framework when rates vary significantly among lineages. Additionally, a birth-death process complementary to the substitution process requires estimation of insertion and deletion parameters at internal nodes. Hence, the process inherits the problems of ancestral state reconstruction twice, once for the substitution alphabet and once for the insertion/deletion evolutionary process. Furthermore, the process assumes that ancestor-descendant sequences are drawn from a generative distribution of a specific form. Although the RE08 model marginalizes over the ancestral states as in maximum average likelihood (Felsenstein, 1981), it is currently unclear in what ways this approach affects topology estimation, and whether the optimal topology in this context differs significantly from that selected by the reversible indel approach. Lastly, it is also unclear whether globally homogeneous modelling of indels, regardless of the assumed nature of insertion and deletion events, is appropriate, and the relationship among global homogeneity, reversibility, atomistic columns, and insertion/deletion parameterization is an active research area. To further this discussion, we present results that argue strongly in favour of parameterizing indel events, regardless of the specific form.

Theory

In this section, we identify a novel form of trivial alignment possible in likelihood analyses as a result of treating indels as missing data on a fixed alignment. When we refer to ML, we mean maximum average likelihood (Felsenstein, 1981), as opposed to most parsimonious likelihood (Barry and Hartigan, 1987) or maximum evolutionary path likelihood (Farris, 1973). Additionally, we note that throughout this paper we alternate among parsimony (scores closer to zero), conditional likelihood (scores closer to one), and log likelihood (absolute value of scores closer to zero) in discussing optimality.

Consider a single-character dataset, below:

Taxon 1|A
Taxon 2|A
Taxon 3|A
Taxon 4|C
Taxon 5|G
Taxon 6|T

Under a Neyman (1971) model, transition probabilities are given by

$$pr\,(x \rightarrow x) = \frac{1}{r} + \frac{(r-1)}{r} e^{-l_i} \tag{6}$$

$$pr\,(x \rightarrow y) = \frac{1}{r} - \frac{1}{r} e^{-l_i} \tag{7}$$

where l_i is the (non-normalized) length of branch i, and $r = 4$ for the JC69 model (Jukes and Cantor, 1969).

In the examples below, we refer to $p(x \rightarrow x)$ as s_i and $p(x \rightarrow y)$ as d_i, for branch i of the topology.

Case 1.1: Trivial Alignment under standard assumptions

We now introduce a particular (trivial) alignment, which we will henceforth refer to as trivial identity alignment (TIA). In TIA, substitution costs are infinite and indel costs are nonzero such that character columns are compressed into a minimum-length alignment of binary states, with each character of size n_{tax} consisting of any proportion of atomistic indels and a single nucleotide alphabet character. This trivial alignment differs critically from the form of parsimony-optimal trivial alignment identified in a previous paper (Wheeler et al., 1995), in which nucleotide data are aligned completely out of sequence. Whereas the contribution to the likelihood by the term for equilibrium base frequencies ("priors" of Felsenstein, 1981, p. 370) renders this parsimony-optimal form of trivial alignment highly suboptimal under the likelihood criterion by penalizing the overall length of the alignment, the reduced length of the TIA suggests its possible optimality under standard likelihood assumptions, which we show below with a simple example. For the single character example, the TIA is:

Taxon 1|A – – –
Taxon 2|A – – –
Taxon 3|A – – –
Taxon 4| – C – –
Taxon 5| – – G –
Taxon 6| – – – T

Consider the tree in Fig. 1. The conditional likelihood of the row D_1 (the first character of the example TIA) in this example is calculated as

$$p(D_1|T, \theta) = s_1 s_2 (s_3 s_4 + 3d_3 d_4) + 3d_1 d_2 (s_3 d_4 + d_3 (s_4 + 2d_4)) \quad (8)$$

which is maximized by making all s_i values equal to 1, and all d_i values equal to zero. The optimized conditional likelihood for D_1 is therefore $\hat{p}(D_1|T, \theta) = 1$.

The conditional likelihoods of D_2, D_3, and D_4 are equivalent due to the root-independence of reversible stochastic models and are calculated as

$$p(D_j|T, \theta) = \prod_{i=1}^{10} (s_i + 3d_i) \quad (9)$$

where j is each of the characters $j = 2, 3, 4$. Because the individual terms $(s_i + 3d_i)$ sum to 1, the product is also 1, and so each of these characters' conditional likelihoods are 1.

Additionally, due to the indels-as-missing assumption, the character-optimal branch lengths for the example

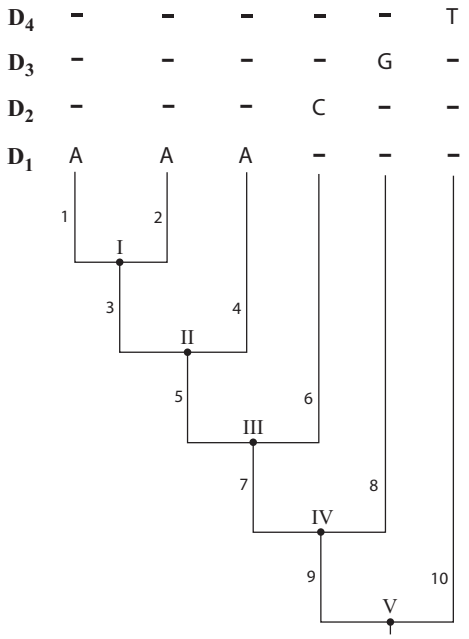

Fig. 1. Tree topology used to calculate likelihoods.

TIA are equal to the optimal lengths shared across the characters (i.e. there is zero homoplasy). The likelihood of the six-taxon TIA example is therefore simply the product of the equilibrium frequencies, which is

$$L_{TIA,4} = \prod_{i=1}^{4} \frac{1}{r} \quad (10)$$

Therefore, the −log likelihood is $-\log L_{TIA,4} = 5.545$.

Case 1.2: Trivial Alignment under standard assumptions is an optimal solution

Since the parsimony score for any TIA is necessarily zero for any topology when indels are treated as missing data, the results of ML mirror those using TS97/No Common Mechanism (Tuffley and Steel, 1997). Under TS97/NCM, ML and parsimony are related by the equation

$$L_T(X) = \prod_{i=1}^{n} r_i^{-l(\chi_i, T)-1} \quad (11)$$

where n is the number of characters, χ_i is character i, X is the set of n character columns, $l(\chi_i, T)$ is the parsimony score of the topology for column i, and r_i is the character alphabet size, which for DNA data under the indels-as-missing assumption is $r_i = 4$. Because n, the number of characters, equals 1, the single non-trivial character example above under TS97/NCM has likelihood

$$L_T(X) = r^{-l(\chi, T)-1} \quad (12)$$

By inspection, the parsimony score of the single-character nontrivial example is $l(\chi,T) = 3$, making the likelihood under TS97/NCM equal to that of the TIA. It is notable that the TIA is optimal whenever the compressibility of the data (the difference between the number of taxa and the number of redundant characters within a single column) exceeds the parsimony score. Although in the example presented, the TIA optimality is equal to that of the single character case under standard assumptions, we later show that the TIA optimality is superior to that of heuristic MSAs on several real datasets.

Case 2.1: Trivial Alignment with indel information

When indels are treated as a fifth state, likelihood calculation is expanded by an additional term at each node. Additionally, unlike in the preceding case, the inclusion of indels in the character alphabet makes character-optimal branch lengths conflict with the optimal branch lengths shared by the characters. Consider the first character of three A, as given in the topology of Fig. 1. The simplified conditional likelihood in this case is obtained by observing that a single change from an indel to an A occurs on branch 5, rendering $d_5 = 1/5$, $s_5 = 1/5$, and all other $d_i = 0$. This simplification results in the conditional likelihood for the first character,

$$p(D_1|T, \theta) = \frac{1}{5}\prod_{\forall i \neq 5} s_i \tag{13}$$

Following the procedure of relative likelihood, each branch is optimized separately. The maximum values taken by s_i occur at $s_i = 1$, when each branch length equals zero. The individually optimal conditional likelihood for character 1 is therefore $\hat{p}(D_1|T, \theta) = \frac{1}{5}$.

The conditional likelihood for D_2 is simplified by observing that a single change occurs on branch six leading to the nucleotide. Therefore, $d_6 = 1/5$, $s_6 = 1/5$, and all other $d_i = 0$, and

$$p(D_2|T, \theta) = \frac{1}{5}\prod_{\forall i \neq 6} s_i \tag{14}$$

As for D_1, the individually optimal conditional likelihood for D_2 is obtained by setting all s_i values to 1, resulting in $\hat{p}(D_2|T, \theta) = \frac{1}{5}$. The results are similar for the character-optimal conditional likelihoods of the third and fourth characters, D_3 and D_4, in which the s_i corresponding to the branch, 8 and either 9 or 10 (there are two equally parsimonious mappings for D_4), respectively, on which the change occurs is removed from the product term:

$$p(D_3|T, \theta) = \frac{1}{5}\prod_{\forall i \neq 8} s_i \tag{15}$$

$$p(D_4|T, \theta) = \frac{1}{5}\prod_{\forall i \neq 10} s_i \tag{16}$$

The conditional likelihoods of these characters are also optimized to 1/5. It should be noted that placing the change for D_4 on either branch 9 or branch 10 has no effect on the character-optimal conditional likelihood. In the best-case scenario, the likelihood of the TIA is calculated under the condition that each character is individually optimized on the topology (TS97/No Common Mechanism, see below). Under these assumptions, branch lengths are separate for each character, and the best likelihood score for the TIA under a fifth-state-indel Neyman model is the product of the individual conditional likelihoods and the equilibrium frequencies, which is $-\log L_{TIA,5} = 12.876$. However, the likelihood score for the TIA when indels are incorporated in the character alphabet is worse than this best estimate. The general computational framework for optimizing branches under these conditions is provided by either Brent's, (1973) method or the Newton–Raphson method in software implementations, and is not in general directly calculable as a closed-form solution. However, because POY5α (http://research.amnh.org/scicomp/research/ projects/invertebrate-zoology/poy?q=projects/poy.php) enables indel-as-character models, the −log likelihood of the example TIA can be calculated directly as $-\log L_{TIA,5} = 20.944$. The likelihoods of alternative single-character datasets and their TIAs are presented in Table 6.

Case 2.2: Trivial Alignment with indel information is not optimal

As in Case 1.2, we consider the optimality of the single-character TIA against the nontrivial single-character alignment, but now consider indels as a fifth state in the character alphabet ($r = 5$).

As before, the single, non-trivial character under TS97/NCM has likelihood of eqn (12). By inspection, the parsimony score for this single character is observed to be $l(\chi,T) = 3$, and so is unchanged by the addition of indels as a state. By contrast, the addition of indels to the character alphabet adds a step to the six-taxon TIA parsimony score ($l(\chi_i,T) = 4$). Therefore, the TS97NCM log likelihoods are $-\log L_{NT} = 6.438$ and $-\log L_{TIA,5} = 12.876$. From these simple examples, it is thus apparent that adding indels to the alphabet results in an order of magnitude drop in the optimality score on the TIA. The contribution to this suboptimality can be divided into effects based on the alphabet alone, and those effects based on increased homoplasy.

Empirical examples

Nucleotide Datasets

Five single-locus datasets were employed to assess the performance of indel-as-missing and indel-as-state analyses: *scop16S*, a 40-taxon 16S rRNA dataset of Scopelomorph fishes (Teleostei; Myctophiformes) assembled from Yamaguchi et al. (2000a,b), with a length difference of 31 between the longest and shortest sequences; *whiting18S*, a 62-taxon 18S rRNA dataset of holometabolous insects (Svenson and Whiting, 2004), with a length difference of 283 between the longest and shortest sequences; small, a 32-taxon dataset of simulated sequences taken from the SATé program folder (Liu et al., 2009), with a length difference of 34 between the longest and shortest sequences; large, a 60-taxon dataset of randomly-selected simulated sequences from the 1000 taxon dataset large.fas in the SATé program folder, with a length difference of 44 between the longest and shortest sequences; and *metazoa18S*, a 60-taxon 18S rRNA unpublished dataset of metazoa, with a length difference of 479 between the longest and shortest sequences. Pre-existing indels were purged from each dataset before alignment.

Heuristic Multiple Sequence Alignments

Each of the five datasets was aligned in each of four commonly-employed multiple sequence alignment programs, with default settings, as follows: CLUSTALw 2.0.12 (Larkin et al., 2007), gap opening: 15, gap extension: 6.66; MAFFT 6.7.13 (Katoh et al., 2005), gap opening 1.53, gap extension 0.00; MUSCLE 3.6 (Edgar, 2004), gap opening −400, gap extension 0.00; PRANK v.100311 (Löytynoja and Goldman, 2008), gap opening 0.025, gap extension, 0.75. Cost parameters not displayed as default output were confirmed by logging a verbose form of the output to a logfile. Individual alignments submitted to PRANK were additionally supplied with a guide tree in the form of the RAxML-optimal topology generated from the corresponding MAFFT alignment (see Tree Search Intensity section), and the "phylogeny-informed" multiple sequence alignment was carried out using the default HKY model of evolution.

Additionally, implied alignments (Wheeler, 2003) were generated for each dataset via direct optimization (Wheeler, 1996) in POY 4.1.2.1 (Varón et al., 2010) under a parsimony weighting scheme that approximates a five-state GTR based on the logarithm of the priors. Five-by-five cost matrices were generated for each unaligned dataset, and priors for indels were calculated via the number of indels over the aligned rows and columns in each dataset. The tree search to generate each implied alignment was carried out by assigning a hundredfold-scaled version of the corresponding cost matrix to each dataset, and then performing 100 random addition sequences, followed by SPR and TBR branch swapping; 50 iterations of parsimony ratchet, reweighting 20% of the characters by a factor of 3; and 200 iterations of tree fusing, followed by an iterative pass with neighbourhood size 2. The resulting implied alignments were used as a fifth multiple sequence alignment, the GTR-Neyman-implied-alignment (GTRNCMIA) for each dataset.

For all resulting alignments, median proportion indels per taxon, defined as the median proportion of per-taxon atomistic indel characters across all taxa in an alignment, were calculated in Mesquite ver. 2.7.2 (Maddison and Maddison, 2010) and the number of parsimony-informative, constant, and autapomorphic characters under both an indels-as-missing and indels-as-state assumption were calculated using a script in C.

Trivial Alignments

For each of the five datasets, TIAs were generated for comparison with those generated using the default parameters of the traditional alignment programs. Because the problem of generating the TIA is likely to be NP-hard based on its similarity to the NP-hard tree-alignment (Sankoff, 1975), binary-character clique (Hamel and Steel, 1996), and strip-packing (Martello et al., 2003) problems, heuristic TIAs were generated in POY4.1.2.1 as implied alignments using a custom 5×5 cost matrix specifying mismatch cost 1000, gap insertion cost 10, and match cost 0. TIAs were generated in POY4.1.2.1 via topology search consisting of 20 random-addition sequences, followed by TBR swapping. The likelihoods of final identity alignments were evaluated in RAxML7.2.6 (-m GTRGAMMA -c4) and PHYML3.0 (-m GTR -s SPR) for indels-as-missing, and as prealigned data in POY5α using a $4 \times 4 + 1$ GTR model in which all nucleotide to gap transitions were estimated as a single parameter. Log likelihoods of the equilibrium frequencies alone were calculated for comparison to TIA likelihoods.

Tree Search Intensity

Topology searches for each multiple sequence alignment used a GTR + Γ4 model, with empirical equilibrium base frequencies, and were divided into two sets: (a) analyses treating indels as missing data, and (b) analyses treating indels as a state. Indels-as-missing analyses were carried out in the commonly used ML software packages RAxML 7.2.6 (Stamatakis, 2006) with commands -m GTRGAMMA -c 4, PHYML 3.0 (Guindon et al., 2005, 2010) with commands -m GTR -s SPR, and TreeFinder (Jobb et al., 2004) with commands GTR[Optimum,Empirical]:G[Optimum]:4 and search depth 2. Additionally, ML analyses were carried out in

POY5α, with base frequencies, rate matrix entries, gamma shape parameter, and branch lengths calibrated to those produced by PHYML. Branch length values on topologies produced by POY5α exhibited an average discrepancy of 0.01 from values estimated in PHYML, presumably due to differences in floating point approximation. Topology searches in POY5α were conducted by specifying prealigned data using the command read (prealigned:(filename,tcm:(1,0))) and then by generating a parsimony topology using the same search intensity as for the implied alignments. The parsimony topology was then transformed to likelihood under exact iterative pass optimization using the command transform(likelihood:(gtr,gamma:(4),estimate)) and subjected to a constrained SPR search using the command swap-(all,spr,sectorial:5) to replicate the search intensity used by RAxML. Cross-validations of tree scores were conducted on a subsample of trees produced by the different programs by re-estimating branch lengths and rate matrices and comparing the re-evaluated scores to the original scores.

Given the paucity of software implementing indel models for ML, two versions of indel-as-state analyses were conducted in POY5α. First, each character-to-gap transition was treated as separate parameters (e.g. $A \leftrightarrow - \neq C \leftrightarrow -$) by adding the condition gap:(independent) to the transform command. Second, nucleotide-to-gap transitions were treated as a single parameter (e.g. $A \leftrightarrow - = C \leftrightarrow -$) by adding the condition gap:(coupled) to the transform command. All analyses were run from the command line on either a 2.4 GHz Intel Core 2 Duo MacBook Pro with 4 GB 667 MHz DDR2SDRAM under OSX ver. 10.5.8 or on the Demeter Cluster (256 × 2.8 GHz Pentium 4 Xeon CPU) at the American Museum of Natural History. All datasets, heuristic multiple sequence alignments, and tree topologies are available in the Supporting Appendix S1.

Results

Sequence alignment properties

Multiple sequence alignments of the five datasets varied in length depending on the alignment program used (Table 1, column 3). Differences in aligned sequence length resulted from differences in the median proportion of indels per taxon. In all cases, default affine CLUSTAL alignments exhibited the fewest median proportion indels per taxon and the shortest aligned lengths, and either PRANK + GT or the GTRNCMIA exhibited the highest median proportion indels per taxon and the longest aligned lengths for all five datasets. The non-affine default MAFFT and MUSCLE alignments exhibited similar median proportion indels per taxon, falling in between those generated by CLUSTAL and those generated by PRANK + GT and GTRNCMIA.

The pattern of inserted indels exhibited significant interaction with nucleotide characters. Treating indels as a state reduced the proportion of constant characters (CS) in all alignments (Table 1, column 6), with the smallest differences in constant characters observed in the alignments of the *scop16S* and *whiting18S* datasets.

Additionally, treating indels as a state increased the proportion of parsimony-informative characters (IS) in all alignments (Table 1, column 7), with the largest increases observed in the alignments of the *whiting18S* and *metazoa18S* datasets. Treating indels as a state also slightly increased the proportion of autapomorphic characters (AS) in most alignments (Table 1, column 8), with the exception of the CLUSTAL alignments of the large and *whiting18S* datasets. The relative ratios of differences in the proportion of informative characters to autapomorphic characters (P/A) between the two treatments of indels (Table 1, column 9) ranged from 0.323 for the MAFFT alignment of the large dataset to 13.002 for the CLUSTAL alignment of the *scop16S* dataset. Most values for the heuristic multiple sequence alignments (MSAs) fell between zero and one, and below 3.5. Among the tree-informed alignment methods (PRANK + GT and GTRNCMIA), GTRNCMIA alignments exhibited larger P/A values for all datasets except the *scop16S* dataset.

Trivial alignment properties

Treating indels as a state decreased the proportion of constant characters in TIAs (Table 1). This effect was an order of magnitude larger for TIAs than for the MSAs of the corresponding datasets, with the simulated dataset (large and small) TIAs exhibiting the largest differences in constant characters between the two indel treatments. Treating indels as a state increased the proportions of both informative and autapomorphic characters by an order of magnitude in all datasets except the *whiting18S* dataset.

Four of the five datasets (*whiting18S*, small, large, and *metazoa18S*) exhibited approximately equal base frequencies across the nucleotide alphabet (Table 2), with *scop16S* exhibiting a slight bias toward adenine and a slight deficiency in cytosine. When indels were treated as missing data, TIAs for all datasets exhibited parsimony scores of 0, as well as likelihood scores closely resembling the contribution by the equilibrium frequencies (Table 2, column 3).

As expected, when indels were treated as a fifth character state, equilibrium base frequencies for all trivial alignments significantly decreased, and the proportion of indels dominated the frequencies by at least three orders of magnitude. TIAs exhibited non-zero unweighted parsimony scores.

Table 1
Differences in character properties of sequence alignments between the indels-as-missing and indels-as-state assumptions

Dataset	Taxa	Length	Alignment	MPI*	ΔCS†	ΔIS‡	ΔAS§	P/A¶
scop16S	40	1315	CLUSTAL	0.049	−0.032	0.030	0.002	13.002
scop16S	40	1320	MAFFT	0.053	−0.037	0.019	0.018	1.042
scop16S	40	1332	MUSCLE	0.061	−0.041	0.019	0.022	0.862
scop16S	40	1379	PRANK + GT	0.093	−0.077	0.028	0.049	0.559
scop16S	40	1397	GTRNCMIA	0.105	−0.087	0.030	0.057	0.532
scop16S	40	4520	TIA	–	−0.865	0.375	0.490	0.765
whiting18S	62	1828	CLUSTAL	0.051	−0.099	0.111	−0.011	9.619
whiting18S	62	1838	MAFFT	0.057	−0.122	0.112	0.010	10.789
whiting18S	62	1853	MUSCLE	0.064	−0.132	0.102	0.030	3.375
whiting18S	62	1909	PRANK + GT	0.092	−0.174	0.112	0.062	1.814
whiting18S	62	1887	GTRNCMIA	0.081	−0.162	0.125	0.037	3.357
whiting18S	62	2324	TIA	–	−0.420	0.210	0.210	1.000
small	32	1109	CLUSTAL	0.095	−0.040	0.020	0.020	1.000
small	32	1144	MAFFT	0.123	−0.051	0.044	0.007	6.250
small	32	1153	MUSCLE	0.130	−0.083	0.044	0.013	3.400
small	32	1337	PRANK + GT	0.249	−0.188	0.058	0.122	0.513
small	32	1331	GTRNCMIA	0.246	−0.180	0.093	0.087	1.069
small	32	8300	TIA	–	−0.998	0.475	0.523	0.910
large	60	1074	CLUSTAL	0.059	−0.005	0.013	−0.008	1.556
large	60	1665	MAFFT	0.393	−0.144	0.035	0.108	0.323
large	60	1529	MUSCLE	0.339	−0.076	0.061	0.014	4.227
large	60	2603	PRANK + GT	0.611	−0.247	0.112	0.136	0.824
large	60	2039	GTRNCMIA	0.504	−0.307	0.204	0.103	1.967
large	60	17 268	TIA	–	−1.000	0.430	0.570	0.754
metazoa18S	60	2504	CLUSTAL	0.298	−0.165	0.111	0.040	2.743
metazoa18S	60	2607	MAFFT	0.326	−0.184	0.119	0.051	2.338
metazoa18S	60	2637	MUSCLE	0.333	−0.179	0.102	0.064	1.592
metazoa18S	60	3372	PRANK + GT	0.478	−0.397	0.136	0.250	0.544
metazoa18S	60	3692	GTRNCMIA	0.524	−0.524	0.184	0.281	0.655
metazoa18S	60	10 018	TIA	–	−0.945	0.427	0.517	0.826

*Median proportion indels per taxon.

†The difference in the proportion of constant characters, calculated as CS(+) − CS(−), where the (+) and (−) indicate indels treated as a state, and as missing data, respectively.

‡The difference in the proportion of parsimony-informative characters, calculated as IS(+) − IS(−).

§The difference in the proportion of autapomorphic characters, calculated as AS(+) − AS(−).

¶The absolute value of the ratio of ΔIS to ΔAS.

Multiple sequence alignment scores

Indel-as-missing. There were two sources of variability in the log likelihood scores of each dataset (Table 3). Most of the variation in scores was due to different alignment methods (Table 3, column 9) which, in all cases, was at least an order of magnitude larger than the largest variation in log likelihood resulting from differences in tree estimation (see Supporting Appendix S1). The *scop16S* and small datasets exhibited the lowest percentage variation in log likelihood scores across tree estimators (Table 3, column 10), and the *metazoa18S* and large datasets exhibited the highest percentage variation in log likelihood scores across tree estimators. The *whiting18S* dataset exhibited the third highest percentage variation in log likelihood score across tree estimators.

Different tree estimators produced different log likelihood scores for a given alignment (Stamatakis, 2008). In general, RAxML and PHYML log likelihood scores were more similar to one another than to scores from TreeFinder. Scores from POY5α were never better than those of RAxML or PHYML, but were better than those of TreeFinder in several cases.

Across tree estimators, the PRANK + GT and GTRNCMIA alignments exhibited the best log likelihood scores and showed little variability in alignment rank. MUSCLE alignments exhibited the worst alignment scores, and showed little variability in alignment rank. CLUSTAL and MAFFT alignments exhibited variation in alignment rank. Despite differences in alignment scores across tree estimators, all produced the same average ranking of alignments. PRANK + GT alignments were optimal, followed by GTRNCMIA, CLUSTAL, MAFFT and MUSCLE alignments, respectively.

Table 2
Trivial identity alignment properties

Alignment	Parsimony	Base frequency*	Likelihood†	*f*(A)‡	*f*(C)	*f*(G)	*f*(T)	*f*(−)
Indels-as-missing								
scopTIA	0	6196.550	6287.073	0.310	0.273	0.219	0.198	–
smallTIA	0	11499.426	11515.130	0.255	0.243	0.238	0.263	–
metazoaTIA	0	13862.078	13905.980	0.256	0.220	0.265	0.259	–
largeTIA	0	23937.828	23941.520	0.254	0.250	0.247	0.249	–
whitingTIA	0	3212.978	3220.505	0.232	0.250	0.286	0.232	–
Indels-as-state								
scopTIA	5538	4375.051	36479.437	0.087	0.076	0.060	0.054	0.724
smallTIA	11528	4458.113	66836.578	0.031	0.029	0.028	0.033	0.879
metazoaTIA	13408	7166.793	87306.429	0.046	0.038	0.048	0.047	0.822
largeTIA	24654	5247.781	151057.720	0.015	0.016	0.014	0.015	0.942
whitingTIA	1272	3704.426	11223.688	0.170	0.184	0.210	0.170	0.266

*The −log likelihood of the equilibrium frequencies alone, calculated as $-\log(f(A)^{N_A} f(C)^{N_C} f(G)^{N_G} f(T)^{N_T} f(-)^{N_-})$.

†Under Indels-as-missing, this score is the best score between searches of similar intensity in RAxML and PHYML under a GTR + Γ4 model. Scores between the two programs were in all cases equivalent to one another to one decimal place. Under Indels-as-state, this score is the likelihood calculated in POY under a five-state GTR + Γ4 model with nucleotide-to-indel transitions coupled.

‡Frequency values of alphabet characters, rounded to three decimal places. Frequency values used in calculations were accurate to five decimal places.

Table 3
−Log likelihood scores and ranks of heuristic MSAs under the indels-as-missing assumption

Dataset	CLUSTAL	MUSCLE	MAFFT	PRANK	GTRNCMIA	Ranks*	SD†	PV‡
RAxML								
scop16S	15534.580	15497.409	15586.875	15298.217	15406.107	4, 3, 5, 1, 2	114.050	0.737
whiting18S	5301.630	5122.640	5149.720	4936.403	4978.695	5, 3, 4, 1, 2	145.855	2.861
small	29377.889	29859.684	29692.456	29207.467	29280.066	3, 5, 4, 1, 2	280.139	0.950
large	65776.105	70561.593	65402.169	63009.494	63906.222	4, 5, 3, 1, 2	2923.235	4.447
metazoa18S	38358.668	39539.582	39037.312	37141.467	34877.445	3, 5, 4, 2, 1	1860.361	4.923
PHYML								
scop16S	15532.120	15499.340	15585.402	15294.497	15405.417	4, 3, 5, 1, 2	114.879	0.743
whiting18S	5301.090	5121.701	5149.115	4935.992	4979.863	5, 3, 4, 1, 2	145.449	2.853
small	29377.741	29860.246	29692.212	29207.273	29279.794	3, 5, 4, 1, 2	280.393	0.951
large	65778.044	70562.991	65401.966	63009.196	63905.974	4, 5, 3, 1, 2	2923.933	4.448
metazoa18S	38362.613	39531.740	39042.111	37141.305	34877.314	3, 5, 4, 2, 1	1859.695	4.921
TreeFinder								
scop16S	15538.020	15501.000	15594.100	15297.650	15406.830	4, 3, 5, 1, 2	116.890	0.756
whiting18S	5323.906	5139.330	5192.912	4942.678	4984.636	5, 3, 4, 1, 2	155.707	3.043
small	29376.770	29856.820	29692.320	29206.460	29278.780	3, 5, 4, 1, 2	279.740	0.949
large	65793.761	70574.200	65403.330	63009.192	63905.900	4, 5, 3, 1, 2	2928.606	4.455
metazoa18S	38306.209	39520.493	39007.630	37104.248	34834.680	3, 5, 4, 2, 1	1866.924	4.945
POY5α								
scop16S	15550.290	15520.777	15604.831	15304.652	15424.412	4, 3, 5, 1, 2	118.341	0.764
whiting18S	5304.617	5123.776	5154.554	4938.872	4978.629	5, 3, 4, 1, 2	146.716	2.877
small	29382.497	29880.961	29709.951	29216.327	29286.360	3, 5, 4, 1, 2	286.800	0.972
large	65808.194	70611.676	65462.583	63014.545	63905.971	3, 5, 4, 1, 2	2941.279	4.473
metazoa18S	38379.246	39590.655	39078.137	37146.145	34879.084	3, 5, 4, 2, 1	1879.748	4.971

*The rank of −log likelihood scores for each heuristic MSA, per dataset. The median ranks, taken with respect to each alignment implementation, were 3, 5, 4, 1, 2 for all likelihood implementations.

†The standard deviation of the −log likelihood scores for each dataset.

‡The per cent variation in the −log likelihood scores, calculated as 100 × (SD/mean).

Indel-as-state. Treating indels as independent, rather than coupled, resulted in better log likelihood scores overall, with few exceptions. For all datasets, treating indels as a state in POY5α (Table 4) increased log likelihood scores by between 20 and 30% relative to POY5α indels-as-missing scores (Table 3), regardless of

Table 4
−Log likelihood scores and ranks of heuristic MSAs under the indels-as-state assumption

Dataset	CLUSTAL	MUSCLE	MAFFT	PRANK	GTRNCMIA	Ranks*	SD	PV
POY indels independent (all transformations unequal)								
scop16S	17502.079	17630.136	17475.334	17918.445	17644.414	2, 3, 1, 5, 4	175.792	0.997
whiting18S	8615.444	8711.837	8728.307	9026.259	8609.090	2, 3, 4, 5, 1	169.933	1.945
small	32054.197	33937.911	32809.508	33904.058	33073.306	1, 5, 2, 4, 3	792.439	2.390
large	69167.188	95314.502	82237.960	91151.178	74924.892	1, 5, 3, 4, 2	10888.660	13.189
metazoa18S	51104.708	53902.119	53149.401	56052.226	53214.680	1, 4, 2, 5, 3	1775.959	3.321
POY indels coupled (all transformations equal)								
scop16S	17507.656	17644.756	17504.662	17932.247	17647.699	2, 3, 1, 5, 4	173.959	0.986
whiting18S	8638.792	8725.563	8847.194	9046.251	8614.815	2, 3, 4, 5, 1	177.090	2.018
small	32057.046	33941.854	32816.774	33905.659	33075.249	1, 5, 2, 4, 3	791.959	2.388
large	69169.077	95246.828	82251.754	91153.548	74924.892	1, 5, 3, 4, 2	10868.657	13.166
metazoa18S	51184.408	53966.430	53262.021	56124.226	53280.379	1, 4, 2, 5, 3	1771.355	3.307

*The rank of −log likelihood scores for each heuristic MSA, per dataset. The median ranks, taken with respect to each alignment implementation, were 1, 4, 2, 3, 5 for both coupled and independent treatment of the indel parameter in POY.

Table 5
Comparison of TIA and heuristic MSA performance under the different indel treatments

		Dataset				
Indel treatment	Alignment type*	*scop16S*	*whiting18S*	small	large	*metazoa18S*
Missing	Mean MSA	15469.126	5103.038	29486.104	65740.155	37787.804
Missing	TIA	6287.080	3220.532	11515.360	23941.527	13905.979
State (coupled)	Mean MSA	17647.404	8774.523	33159.316	82549.220	53563.493
State (coupled)	TIA	36479.437	11223.688	66836.578	151057.720	87306.429
State (independent)	Mean MSA	17634.082	8738.187	33155.796	82559.144	53484.627
State (independent)	TIA	36031.374	10970.053	65458.790	147461.013	86165.294

*Mean −log likelihood scores for each dataset were calculated using values from Table 3 (indels-as-missing) and Table 4 (indels-as-state) for heuristic MSAs, and Table 2 for TIAs.

whether indels were treated as independent or coupled. Treating indels as a state increased the variance of log likelihood scores relative to POY5α indels-as-missing scores in all cases except for the *metazoa18S* dataset, regardless of whether indels were independent or coupled.

Treating indels as a state significantly increased the per cent variation in the small and large dataset likelihood scores relative to the indels-as-missing condition, and also slightly increased the per cent variation in the *scop16S* dataset log likelihoods. By contrast, treating indels as a state decreased per cent variation in likelihood scores for the *whiting18S* and *metazoa18S* datasets.

Treating indels as a state altered the average ranking of alignments relative to the indels-as-missing rankings. CLUSTAL alignments had the best likelihoods, followed by MAFFT and the GTRNCMIA alignments, which differed in rank significantly on the *scop16S* and *whiting18S* datasets. MUSCLE alignments ranked fourth, differing from the indels-as-missing condition in the improved scores on the *metazoa18S* dataset. PRANK + GT alignment rank differed most significantly from the indels-as-missing condition in exhibiting the worst log likelihood scores for the biological datasets *scop16S*, *whiting18S* and *metazoa18S*, and second-worst likelihood scores on the simulated datasets small and large.

Alignment performance under standard assumptions and with indel information. When indels were treated as missing data, TIAs exhibited the best log likelihood scores for each of the five datasets. The mean multiple sequence alignment likelihood scores for each of the five datasets fell well below the TIA likelihoods. In most cases, the TIA log likelihood scores were nearly twice the optimality score of those of the multiple sequence alignments (Table 5).

When indels were treated as a state, TIA likelihoods dropped below those of the multiple sequence alignments across all datasets, regardless of whether indels were treated as independent or coupled (Table 5).

Discussion

We show that a novel form of trivial alignment (heuristic TIA) exists in both parsimony and ML frameworks, and

that TIA remains the optimal solution in ML analyses when indels are treated as missing data. Although the calculated likelihoods of the TIA example provided invoke a Neyman fifth-state indel model, whose assumptions are likely to be extremely simplistic compared to real data and the dynamics of indel events, the relationships among heuristic MSAs and trivial alignments predicted under these simplified assumptions were nonetheless validated when alternative single-character TIAs were examined (Table 6), and when real datasets were analysed using a GTR model with indels explicitly modelled as a separate parameter. These results indicate that the optimality of the TIA under standard ML is a real phenomenon independent of the model complexity employed.

The existence of the TIA as an optimal solution when indels are treated as missing data in the substitution alphabet has two implications. First, and most obviously, it suggests that indel events should be included as a state if one wishes to use the substitution likelihood (eqn 3, left term of right-hand side) to infer an optimal topology from a fixed MSA, to avoid potential bias (''trivialization'') in the placement of otherwise invisible indel events. This approach has been well-motivated in the parsimony literature. Alternatively, it suggests that a substitution likelihood may not be the best method for inferring the likelihood of the tree/alignment pairing of data. This second point has been made before by Thorne et al. (1991), who observe that the inference of evolutionarily relevant parameters from a single alignment is likely to be biased in proportion to the degree of divergence among the sequences. The implications of this problem are illustrated by the iterative SATé procedure (Liu et al., 2009), which divides an initial heuristic MSA into subsets that are re-aligned and merged according to the structure of the tree topology associated with the alignment from the preceding iteration. Because this method operates under standard likelihood assumptions, its optimal solution is a TIA, whether identified or not.

Table 6
Alternate single-character trivial identity alignments

					−Log likelihood*			
Taxa	N_A	N_C	N_G	N_T	Single, 4	TIA, 4	Single, 5	TIA, 5
6	2	2	1	1	5.545	5.545	6.437	20.944
7	3	1	1	2	5.545	5.545	6.437	20.944
8	2	2	2	2	5.545	5.545	6.437	20.944
10	3	3	2	2	5.545	5.545	6.437	20.944
12	3	3	3	3	5.545	5.545	6.437	20.944

*Likelihood scores calculated in POY under the Neyman-4 and Neyman-5 models discussed in the theory section, using the commands read (prealigned:(file,tcm:(1,0))), either transform (likelihood:(jc69)) for Neyman-4 or transform (likelihood:(jc69,gap:(character))) for Neyman-5, then build(), and swap(all).

Tree-based methods are one solution to the problems associated with progressive alignment algorithms. Methods for simultaneous (as opposed to iterative) phylogeny/alignment inference exist (Sankoff, 1975; Wheeler, 1996, 2006; Suchard and Redelings, 2006; Redelings and Suchard, 2007; Varón et al., 2010) under several optimality frameworks, most notably parsimony affine direct optimization (Wheeler et al., 2006; Varón et al., 2008). The use of a consistent optimality criterion during alignment and tree search has been shown (for parsimony and a limited form of most parsimonious likelihood) to produce more optimal results than standard two-step procedures (Whiting et al., 2006). The utility of a tree-based approach is implied in this study by the comparable performance of the PRANK + GT and GTRNCMIA alignments with indels as missing data, and the apparent stability of the GTRNCMIA alignment ranking to the inclusion of indels in the alphabet. However, the two methods approach the problem from different directions. The former penalizes alignments by their proportion of indel events, given a fixed guide tree, and is in many regards similar to the broad approach referenced earlier in the section on this ''indel problem.'' The latter method searches for the best tree/alignment combination using a parsimony cost matrix approximating the P matrix entries of the GTR model.

The largely consistent range in the TIA P/A ratios between 0.75 and 1.0 itself suggests that an increase in the proportion of autapomorphic characters relative to informative characters when indels are parameterized on a fixed alignment may be a significant predictor of an inversion in alignment rank between an indels-as-missing and indels-as-state assumption, and thus a significant predictor of whether a given alignment of real data might be ''trivialized'' in the direction of a TIA. However, this idea is contradicted by the ranks of several MSAs with P/A ratios in this range, among them the MUSCLE *scop16S*, GTRNCMIA *metazoa18S*, and CLUSTAL small alignments. Additionally, the wide variation in P/A ratios and alignment ranks across alignment methods for a single dataset (Table 1) suggests that a more complex combination of aligned length and character patterns interacts with the estimation by ML implementations of shared branch lengths across characters to produce a homoplasy artefact explaining the observed phenomena. This last idea is supported by the extreme difference in rankings of the PRANK + GT and GTRNCMIA alignments, which exhibit similar lengths and changes in character patterns (although in nearly all cases, the GTRNCMIA alignments have higher P/A ratios), but significantly different alignment ranks between the indels-as-missing and indels-as-state assumptions. The suboptimality of the PRANK + GT approach may also be due to, or exacerbated by, the suboptimality of the marginalization approach favoured

by eqn (3) when combined with the treatment of indels as a substitution alphabet character, or due to significant bias by the guide tree used to place indels in the alignment. Guide tree bias has been observed in other studies to constrain the tree found as optimal to resemble the tree used during progressive alignment (Lake, 1991), but the effect may instead be to bias the gap placement in the PRANK alignments. Given these results, it remains unclear what predictors best inform on the suitability of indel placement for ML analyses of sequence datasets.

In a TIA, indels are essentially placeholders for maximizing within-column single-nucleotide similarity. Although this conceptual treatment of indels is questionable at best, it is consistent with a "two-step" cost framework, albeit in an admittedly esoteric way. Under standard ML assumptions, indels have zero parsimony cost and probability 1 during topology search, despite having a cost during alignment. Whereas many authors have worked forward from alignment to topology search and developed methods for coding indels into a parsimony cost matrix to be used during tree search (Simmons and Ochoterena, 2000; Young and Healy, 2003; Müller, 2006; Simmons et al., 2007) or model-based framework (Thorne et al., 1991, 1992; McGuire et al., 2001b; Redelings and Suchard, 2005, 2007; Wheeler, 2006; Rivas and Eddy, 2008), the TIA addresses the converse statement–if indels have no cost during topology search, they have no cost during alignment. If this statement is consistent with the assumption of treating indels as missing data, then the optimality scores resulting from this procedure should be higher than those produced by heuristic MSAs, and even the heuristic TIAs provided here outperform the MSAs considered. Therefore, it is likely that a more rigorous search will produce TIAs of still greater optimality under standard ML.

An immediate question is whether maximizing within-column similarity statements results in a legitimate hypothesis of potential homologues in a two-step framework, the debate on the utility of "similarity" in comparative biology (Rieppel and Kearney, 2002; Rieppel, 2006, and sources therein) notwithstanding. The TIA is presented here only as a numerical optimum and byproduct consequence of current ML assumptions; we do not intend it to be, nor are we advocating its use as, a legitimate MSA, even though it is in the solution space. However, an examination of the phylogenetic hypotheses produced by a TIA under standard ML remains to be conducted. As we show, the TIA is only optimal when indels are treated as missing data and is clearly suboptimal when indels are treated as a state. Hence, the treatment of indels by TIA under standard assumptions is confirmed when indels are incorporated in the model.

The indel models implemented on the biological datasets in this study are not new, as indel models (S_iF_jG), are available in POY3 (Wheeler, 2006; Wheeler et al., 2006), in MAC5 (McGuire et al., 2001a) and in dnamlε (Rivas and Eddy, 2008). The models implemented in the present analysis resemble the POY3 implementations S_6F_5G (GTR + indels coupled) and $S_{10}F_5$ (GTR + indels independent). However, whereas the POY3 implementation of indel models is restricted to dynamic (Needleman and Wunsch, 1970) likelihood (i.e. dominant and total likelihood) calculations, the present implementation has potentially broader utility for increasing the number of indel-related studies by operating on static characters in heuristic MSAs, thus making it accessible to current two-step procedures. Additionally, the indel model implemented here in heuristic MSAs differs from previous continuous-time Markov indel models currently available as the software MAC5 and dnamlε (McGuire et al., 2001a; Rivas and Eddy, 2008), in being a rearrangement-based ML method that is both reversible and stationary, *sensu* Jayaswal et al. (2005, 2011). As mentioned earlier, reversibility has been cited as a problem for indel models for two reasons: first, it assumes a frequency of indel characters that is constant with respect to divergence time, and second, it assumes that the sequences under study have the same expected lengths at all divergence times (Rivas and Eddy, 2008). Additionally, the presently-employed model, like M01 (McGuire et al., 2001b), and RE08 (Rivas and Eddy, 2008), assumes independence among columns, potentially overweighting indels of greater length (Simmons et al., 2007) and biasing the topologies inferred as optimal. However, as has also been noted frequently in both parsimony analyses and a few model-based analyses, the utility of including indel information regardless of the assumptions is better than excluding it altogether (e.g. McGuire et al., 2001b; Bapteste and Philippe, 2002; Vogt, 2002; González et al., 2006; van Rheede et al., 2006; Dwivedi and Gadagkar, 2009), although much more work remains to be done to assess this claim for model-based analyses. The proportion of contiguous indels observed in an analysis will depend significantly on the marker utilized and on the taxonomic sampling employed. In the *scop16S* and *whiting18S* datasets of the present study, inferred contiguous internal indels within a taxon were rarely larger than four, whereas in *metazoa18S* and the simulated datasets, contiguous indels were of much greater length. Future work must design studies with length-variable indels in mind. Additionally, new model-based approaches to length-variable indel events should focus on achieving local homogeneity, rather than global homogeneity, as in TKF92 (Thorne et al., 1992), which only allows insertions and deletions of indel blocks of a constant length over all branches of a topology.

Conclusions

We considered the performance of heuristic TIA against heuristic multiple sequence alignments (MSAs) for five datasets under a traditional ML framework of a globally homogeneous, stationary, reversible GTR + Γ model on static characters. Under this standard ML framework with indels treated as missing data, the heuristic TIA alignment, identified as optimal in both ML and parsimony, outperformed heuristic MSAs by a factor of two in nearly all cases.

By contrast, when indels were modelled as a state using a globally homogeneous, stationary, reversible GTR + Γ model, log-likelihood scores necessarily increased, given the transition from a 4×4 to a 5×5 rate matrix, regardless of whether indels were treated as a single event or as separate events. However, under this GTR + indels model, the performance of the heuristic TIA alignment dropped below those of the heuristic MSAs. The numerical optimality of the TIA under standard ML emphasizes the need for incorporation of indels (however parameterized) into models in ML analyses to eliminate this trivial optimum. Ideally, ML analyses must be conducted using parameterized indel information explicitly. At the very least, caution must be exercised when using new software claiming improvement in likelihood scores under standard assumptions.

Because identifying a potentially trivialized alignment of real data remains difficult, caution must also be exercised when drawing conclusions from alignments based on likelihood scores generated in standard tree search implementations.

Acknowledgements

The authors would like to thank Gonzalo Giribet, Nick Lucaroni, Alexandros Stamatakis, Mike Steel, Diane Stevenson, and two anonymous reviewers for helpful comments and discussion that greatly improved the clarity of the manuscript. This work was supported by the US Army Research Laboratory and the US Army Research Office (W911NF-05-1-0271 to W.C.W.) and a Lerner-Grey Fund for Marine Research (to J.S.S.D.).

References

Bapteste, E., Philippe, H., 2002. The potential value of indels as phylogenetic markers: position of trichomonads as a case study. Mol. Biol. Evol. 19, 972–977.

Barry, D., Hartigan, J., 1987. Statistical analysis of hominid molecular evolution. Stat. Sci. 2, 191–210.

Bishop, M.J., Thompson, E.A., 1986. Maximum likelihood alignment of DNA sequences. J. Mol. Biol. 190, 159–165.

Brent, R.P., 1973. Algorithms for Minimization without Derivatives, Chap. 4. Prentice-Hall, Englewood Cliffs, NJ.

Cartwright, R., 2007. Ngila: global pairwise alignments with logarithmic and affine gap costs. Bioinformatics 23, 1427–1428.

Dessimoz, C., Gil, M., 2010. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. 11, R37.

Dwivedi, B., Gadagkar, S., 2009. Phylogenetic inference under varying proportions of indel-induced alignment gaps. BMC Evol. Biol. 9, 211.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

Egan, A., Crandall, K., 2008. Incorporating gaps as phylogenetic characters across eight DNA regions: ramifications for North American Psoraleeae (Leguminosae). Mol. Phylogenet. Evol. 46, 532–546.

Farris, J.S., 1973. A probability model for inferring evolutionary trees. Syst. Zool. 22, 250–256.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.

Felsenstein, J., 1989. PHYLIP: Phylogeny inference package (version 3.2). Cladistics 5, 164–166.

Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, MA.

Fisher, R.A., 1912. On an absolute criterion for fitting frequency curves. Mess. Math. 41, 155–160.

Giribet, G., Wheeler, W., 1999. On gaps. Mol. Phylogenet. Evol. 13, 132–143.

González, D., Cubeta, M., Vilgalys, R., 2006. Phylogenetic utility of indels within ribosomal DNA and [beta]-tubulin sequences from fungi in the *Rhizoctonia solani* species complex. Mol. Phylogenet. Evol. 40, 459–470.

Gotoh, O., 1982. An improved algorithm for matching biological sequences. J. Mol. Biol. 162, 705–708.

Guindon, S., Lethiec, F., Duroux, P., Gascuel, O., 2005. PHYML Online–a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res. 33, W557.

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

Hamel, A., Steel, M., 1996. Finding a maximum compatible tree is NP-hard for sequences and trees. Appl. Math. Lett. 9, 55–59.

Jayaswal, V., Jermiin, L., Robinson, J., 2005. Estimation of phylogeny using a general Markov model. Evol. Bioinform. Online 1, 62–80.

Jayaswal, V., Jermiin, L., Poladian, L., Robinson, J., 2011. Two stationary nonhomogeneous Markov models of nucleotide sequence evolution. Syst. Biol. 60, 74–86.

Jobb, G., Von Haeseler, A., Strimmer, K., 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol. Biol. 4, 18.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, N.H. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, pp. 21–132.

Katoh, K., Kuma, K., Toh, H., Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33, 511–518.

Kearney, M., 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. Syst. Biol. 51, 369–381.

Lake, J., 1991. The order of sequence alignment can bias the selection of the tree topology. Mol. Biol. Evol. 8, 378–385.

Larkin, M., Blackshields, G., Brown, N., Chenna, R., McGettigan, P., McWilliam, H., Valentin, F., Wallace, I., Wilm, A., Lopez, R., Thompson, J., Gibson, T., Higgins, D., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948.

Liu, K., Raghava, S., Nelesen, S., Linder, C., Warnow, T., 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science 324, 1561–1564.

Löytynoja, A., Goldman, N., 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320, 1632–1635.
Maddison, W., Maddison, D., 2010. Mesquite: a modular system for evolutionary analysis. http://mesquiteproject.org.
Martello, S., Monaci, M., Vigo, D., 2003. An exact approach to the strip-packing problem. INFORMS J. Comput. 15, 310–319.
McGuire, G., Denham, M., Balding, D., 2001a. MAC5: Bayesian inference of phylogenetic trees from DNA sequences incorporating gaps. Bioinformatics 17, 479–480.
McGuire, G., Denham, M., Balding, D., 2001b. Models of sequence evolution for DNA sequences containing gaps. Mol. Biol. Evol. 18, 481–490.
Mott, R., 1999. Local sequence alignments with monotonic gap penalties. Bioinformatics 15, 455–462.
Müller, K., 2006. Incorporating information from length-mutational events into phylogenetic analysis. Mol. Phylogenet. Evol. 38, 667–676.
Nee, S., 2006. Birth-death models in macroevolution. Annu. Rev. Ecol. Evol. Syst. 37, 1–17.
Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. 48, 443–453.
Neyman, J., 1971. Molecular studies in evolution: a source of novel statistical problems. In: Gupta, S.S., Yackel, J. (Eds.), Statistical Decision Theory and Related Topics. Academic Press, New York, pp. 1–27.
Pas′ko, L., Ericson, P., Elzanowski, A., 2011. Phylogenetic utility and evolution of indels: a study in neognathous birds. Mol. Phylogenet. Evol. 61, 760–771.
Rabosky, D., 2010. Extinction rates should not be estimated from molecular phylogenies. Evolution 64, 1816–1824.
Redelings, B.D., Suchard, M.A., 2005. Joint Bayesian estimation of alignment and phylogeny. Syst. Biol. 54, 401–418.
Redelings, B., Suchard, M., 2007. Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. BMC Evol. Biol. 7, 40.
van Rheede, T., Bastiaans, T., Boone, D., Hedges, S., 2006. The platypus is in its place: nuclear genes and indels confirm the sister group relation of monotremes and therians. Mol. Biol. Evol. 23, 587–597.
Rieppel, O., 2006. The merits of similarity reconsidered. Syst. Biodivers. 4, 137–147.
Rieppel, O., Kearney, M., 2002. Similarity. Biol. J. Linn. Soc. 75, 59–82.
Rivas, E., Eddy, S., 2008. Probabilistic phylogenetic inference with insertions and deletions. PLoS Comput. Biol. 4, 1–12.
Sankoff, D.M., 1975. Minimal mutation trees of sequences. SIAM J. Appl. Math. 28, 35–42.
Simmons, M., Ochoterena, H., 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst. Biol. 49, 369–381.
Simmons, M., Müller, K., Norton, A., 2007. The relative performance of indel-coding methods in simulations. Mol. Phylogenet. Evol. 44, 724–740.
Simmons, M., Richardson, D., Reddy, A., 2008. Incorporation of gap characters and lineage-specific regions into phylogenetic analyses of gene families from divergent clades: an example from the kinesin superfamily across eukaryotes. Cladistics 24, 372–384.
Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.
Stamatakis, A., 2008. The RAxML 7.0.4 Manual. The Exelixis Lab, Department of Computer Science, Ludwig-Maximilians-Universität, München, Germany.
Suchard, M.A., Redelings, B.D., 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics 22, 2047–2048.
Svenson, G.J., Whiting, M.F., 2004. Phylogeny of Mantodea based on molecular data: evolution of a charismatic predator. Syst. Ent. 29, 359–370.
Swofford, D.L., 2002. PAUP*: Phylogenetic Analysis Using Parsimony (* and other Methods), Version 4.0b10. Sinauer Associates, Sunderland, MA.
Thorne, J.L., Kishino, H., Felsenstein, J., 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33, 114–124.
Thorne, J.L., Kishino, H., Felsenstein, J., 1992. Inching toward reality: an improved likelihood model of sequence evolution. J. Mol. Evol. 34, 3–16.
Tuffley, C., Steel, M., 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59, 581–607.
Varón, A., Wheeler, W., Bar-Noy, A., 2008. An Efficient Heuristic for the Tree Alignment Problem. Technical report. City University of New York, New York, USA.
Varón, A., Vinh, L.S., Wheeler, W.C., 2010. POY version 4: phylogenetic analysis using dynamic homologies. Cladistics 25, 72–85.
Vogt, L., 2002. Weighting indels as phylogenetic markers of 18s rDNA sequences in Diptera and Strepsiptera. Org. Divers. Evol. 2, 335–349.
Waterman, M.S., Smith, T., Beyer, W., 1976. Some biological sequence metrics. Adv. Math. 20, 367–387.
Wheeler, W.C., 1993. The triangle inequality and character analysis. Mol. Biol. Evol. 10, 707–712.
Wheeler, W.C., 1994. Sources of ambiguity in nucleic acid sequence alignment. In: Schierwater, B., Streit, B., Wagner, G., DeSalle, R. (Eds.), Molecular Ecology and Evolution: Approaches and Applications. Birkhäuser Verlag, Germany, pp. 323–352.
Wheeler, W., 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? Cladistics 12, 1–9.
Wheeler, W.C., 2003. Implied alignment: a synapomorphy-based multiple-sequence alignment method and its use in cladogram search. Cladistics 19, 261–268.
Wheeler, W.C., 2006. Dynamic homology and the likelihood criterion. Cladistics 22, 157–170.
Wheeler, W.C., Gladstein, D.S., 1994. MALIGN: a multiple sequence alignment program. J. Hered. 85, 417–418.
Wheeler, W.C., Gatesy, J., DeSalle, R., 1995. Elision: a method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. Mol. Phylogenet. Evol. 4, 1–9.
Wheeler, W.C., Aagesen, L., Arango, C.P., Faivovich, J., Grant, T., D'Haese, C.A., Janies, D., Smith, W.L., Varón, A., Giribet, G., 2006. Dynamic Homology and Phylogenetic Systematics: A Unified Approach Using POY. American Museum of Natural History, New York, USA.
Whiting, M.F., Carpenter, J.M., Wheeler, Q.D., Wheeler, W.C., 1997. The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. Syst. Biol. 46, 1–68.
Whiting, A., Sites, J. Jr, Pellegrino, K., Rodrigues, M., 2006. Comparing alignment methods for inferring the history of the new world lizard genus *Mabuya* (Squamata: Scincidae). Mol. Phylogenet. Evol. 38, 719–730.
Wiens, J., 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst. Biol. 47, 625–640.
Wiens, J., 2003a. Incomplete taxa, incomplete characters, and phylogenetic accuracy: is there a missing data problem? J. Vertebr. Paleontol. 23, 297–310.
Wiens, J., 2003b. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52, 528–538.
Yamaguchi, M., Miya, M., Okiyama, M., Nishida, M., 2000a. Molecular phylogeny and larval morphological diversity of the

lanternfish genus *Hygophum* (Teleostei: Myctophidae). Mol. Phylogenet. Evol. 15, 103–114.

Yamaguchi, M., Miya, M., Okiyama, M., Nishida, M., 2000b. Molecular phylogeny of the lanternfishes (Pisces: Myctophidae) inferred from mitochondrial 16S rDNA. PhD thesis. University of Tokyo Ocean Research Institute, Minamidai 1-15-1, Nakano-ku, Tokyo 164-8639, Japan.

Young, N., Healy, J., 2003. GapCoder automates the use of indel characters in phylogenetic analysis. BMC Bioinformatics 4, 6.

Zwickl, D., 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD thesis. The University of Texas at Austin, USA.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. All datasets, heuristic multiple sequence alignments and tree topologies.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.