Available online at www.sciencedirect.com
ScienceDirect
Molecular Phylogenetics and Evolution 42 (2007) 817–826
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

How should gaps be treated in parsimony? A comparison of approaches using simulation

T. Heath Ogden [a,b,*], Michael S. Rosenberg [b]

[a] Department of Biological Sciences, Idaho State University, Pocatello, ID 83209, USA
[b] Center for Evolutionary Functional Genomics, The Biodesign Institute, and the School of Life Sciences, Arizona State University, Tempe, AZ 85287-4501, USA

Received 7 February 2006; revised 7 July 2006; accepted 22 July 2006
Available online 22 August 2006

Abstract

Simulation with indels was used to produce alignments where true site homologies in DNA sequences were known; the gaps from these datasets were removed and the sequences were then aligned to produce hypothesized alignments. Both alignments were then analyzed under three widely used methods of treating gaps during tree reconstruction under the maximum parsimony principle. With the true alignments, for many cases (82%), there was no difference in topological accuracy for the different methods of gap coding. However, in cases where a difference was present, coding gaps as a fifth state character or as separate presence/absence characters outperformed treating gaps as unknown/missing data nearly 90% of the time. For the hypothesized alignments, on average, all gap treatment approaches performed equally well. Data sets with higher sequence divergence and more pectinate tree shapes with variable branch lengths are more affected by gap coding than datasets associated with shallower non-pectinate tree shapes.

Keywords: Clustal; DNA sequence alignment; Evolutionary distance; Gap character coding; Homology; Indel; Parsimony; Simulation; Tree reconstruction

1. Introduction

Molecular systematics relies on comparisons of DNA sequence information. Before phylogenetic inference methodologies can be applied, the corresponding (putatively homologous) nucleotide bases (or amino acids) must be associated by multiple sequence alignment, converting sequences of unequal length (although equal length sequences may also require alignment) into sequences of equal length. This is accomplished by proposing insertion and deletion events, more commonly referred to as gaps (or indels). The specific treatment of gaps in phylogenetic analysis can affect the results (Giribet and Wheeler, 1999; Ogden and Whiting, 2003; Simmons and Ochoterena, 2000). Many studies conclude that gaps should be included, in some way, in order to provide additional phylogenetic signal (Freudenstein and Chase, 2001; Graham et al., 2000; Hennequin et al., 2003; Kawakita et al., 2003; Petersen et al., 2004; Simmons et al., 2001; van Dijk et al., 1999; Vogler and DeSalle, 1994, just to name a few). Still, others have argued against the usefulness of insertion deletion events as phylogenetic characters (Ford et al., 1995; Golenberg et al., 1993; Li, 1997).

Notwithstanding many of the studies cited above, many phylogenetic studies perform tree reconstruction treating gaps in only one particular way and do not examine the sensitivity of their conclusions to differential gap treatment. For example, in a 3 year (1993–1996) literature review of four systematic journals (*Systematic Biology*, *Systematic Botany*, *Cladistics*, and *Molecular Biology*), Gonzalez (1996) found that 78% of the phylogenetic studies did not examine multiple ways of treating gaps. Of the remaining cases, essentially half found that gaps were informative and the other half concluded that the coding of gaps was

* Corresponding author. Fax: +1 208 282 4570.
E-mail addresses: ogdet@isu.edu (T.H. Ogden), msr@asu.edu (M.S. Rosenberg).

1055-7903/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2006.07.021

irrelevant. There still seems to be a paucity of information concerning the treatment and exploration of gaps in the literature. Most parsimony studies do not explicitly report the way in which gaps are treated, and one must assume that the program default is being used; e.g., for PAUP* (Swofford, 2002) this would be gaps being treated as missing. Most distance-based analyses and, until recently (Holmes, 2005), most likelihood and Bayesian analyses either treated gaps as unknowns or removed the gap containing column(s) from the analyses for pairs of sequences or for all sequences in an alignment. Given that the specific treatment of gaps may make a difference in the resulting phylogenetic hypotheses, it is important to elucidate any contributing factors and how much of a difference these factors play. In this study, we will focus on the treatment gaps in a parsimony framework.

The treatment of indels has been addressed for a number of years, even if only tangentially (Baldwin et al., 1995; Barriel, 1994; Baum et al., 1994; Crandall and Fitzpatrick, 1996; Freudenstein and Chase, 2001; Hibbett et al., 1995; Kelchner, 2000; Kelchner, 2002; Kretzer et al., 1996). Simmons and Ochoterena (2000) provided a comprehensive discussion of the variety of ways that gaps can be treated during phylogenetic inference via parsimony, ranging from treating them as missing data (including regions that are highly divergent) to coding them as fifth state or separate characters. They presented two formal descriptions of coding insertion and deletion events as separate presence/absence characters. The first, termed "simple indel coding" (SIC) is a more conservative approach and is relatively easy to implement "by coding all gaps that have different 5′ and/or 3′ termini as separate presence absence characters." Essentially, (1) indels in multiple sequences are considered a single character if they start and end at the same position; (2) overlapping indels are considered separate characters; and (3) when a large indel contains smaller indels completely within it, the large indel is coded as unknown/missing for the character associated with the shorter indel, rather than presence/absence. This approach has been implemented in an automated fashion in the program GapCoder (Young and Healy, 2003), which can be used on single datasets. The second approach, termed "complex indel coding," is more difficult because it takes into account additional information concerning the number of steps required to transform from one indel to another; there are also specific cases not explicitly covered by the described rule set. Recently, Muller (2006) also discussed incorporating length mutational events and presented a new program for automated adaptation of the "complex indel coding" (this paper was brought to our attention after completion of our analyses and near the termination of the preparation of this manuscript). Other methods for dealing with regions of unequal length have also been proposed but will not be specifically dealt with in this paper, as we will be dealing with the issue of using (or not using) gaps as information in parsimony analyses. Among these different approaches, are found the recently emerging concurrent analysis (combined analysis) methods (Lunter et al., 2005; Redelings and Suchard, 2005; Wheeler, 1996; Wheeler et al., 2003), coding separate multistate characters (Lutzoni et al., 2000), recoding indels as single events (Crandall and Fitzpatrick, 1996; Manos and Steele, 1997; Swofford, 1993), secondary structure guided inferences (Gillespie et al., 2005; Kjer, 1995), likelihood or Bayesian derived models (Holmes, 2005; Holmes and Bruno, 2001; Knudsen and Miyamoto, 2003; McGuire et al., 2001; Metzler, 2003; Miklos et al., 2004; Mitchison, 1999; Rivas, 2005; Thorne et al., 1991, 1992), and stretch and block coding (Geiger, 2002).

DNA sequence simulation allows investigators to directly compare results (true vs. hypothesized) in order to determine where one method outperforms another and to what extent. Only recently have alignments been simulated that include indels (Blanchette et al., 2004a,b; Hall, 2005; Keightley and Johnson, 2004; Ogden and Rosenberg, 2006; Pollard et al., 2004; Rosenberg, 2005a,b). However, none of these studies examine the effects of gap treatment differently during tree reconstruction.

The primary objective of this study is to investigate, within a parsimony framework, the effect on topological accuracy when gaps are treated as unknowns, fifth state characters, or as separate coded gap characters using SIC. We simulate datasets with indels, which represent the true alignments (all columns are homologous) and are analyzed for each of the gap treatments, eliminating the uncertainty of misalignment. We also align these datasets, after having removed all gaps, using automated techniques to explore how the gap treatments perform when alignment is uncertain. Finally, we examine the roles that tree shape and maximum evolutionary distance may play in conjunction with different treatments of gaps.

2. Materials and methods

2.1. Data simulation

We used the simulated datasets from Ogden and Rosenberg (2006), consisting of seven base 16-taxon topologies, providing sufficient tree shape diversity and complexity. The seven topologies consisted of a balanced tree, a pectinate tree, and 5 random trees (See Fig. 3 in Ogden and Rosenberg, 2006). The relative branch lengths of each topology were set under 11 different conditions: ultrametric equal-branch length, clocklike random branch length (5 sets), and non-clock-like random branch lengths (5 sets). Each of these 11 conditions was scaled such that the maximum evolutionary distance between a pair of sequences was equal to 1.0 or 2.0. Thus, each of the 7 topologies was used to create 22 model trees. All simulations were conducted under identical conditions in the program MySSP (Rosenberg, 2005c). The initial sequence length was 2000 bp. Aside from the different conditions explained above, DNA evolution was simulated under the Hase-

gawa–Kishino–Yano (HKY) model (Hasegawa et al., 1985), with a transition bias (Kappa) of 3.6 (Rosenberg and Kumar, 2003); initial and expected base frequencies of A and T = 0.2; and G and C = 0.3.

Insertion and deletion events were modeled as a Poisson process, following Rosenberg (2005a). Expected numbers of insertions and deletions (modeled separately) for a given branch were determined as a function of the realized number of substitutions (itself a Poisson process) which occurred on that branch. Expected rates were based on observed values from primates and rodents, with one insertion event for every 100 substitutions and one deletion event for every 40 substitutions (Ophir and Graur, 1997). The realized number of insertion and deletion events was drawn from a Poisson distribution with mean equal to the expectation. The actual size of each insertion and deletion event was independently determined from a truncated (so as not to include zero) Poisson distribution with mean equal to four bases, (as observed in primates and rodents) (Ophir and Graur, 1997; Sundstrom et al., 2003).

Each simulation was replicated 100 times. The fate of every insertion and deletion event was tracked throughout the simulations, such that the columns, including those with gaps in the final alignment, represented the true homologies (Rosenberg, 2005a).

2.2. Alignment

These simulations resulted in 15,400 unique datasets (alignments) containing gaps representing either insertion or deletion events during the simulation process, and will be referred to as the True Alignments (TA). Each of the TA were then stripped of their gaps and were realigned via ClustalW version 1.83 (Thompson et al., 1994) using default parameters. We will refer to these alignments as the hypothesized alignments (HA). The HA represent a reasonable and realistic amount of alignment error and were used to investigate the role that alignment accuracy plays, in conjunction with treatment of gaps, in topological accuracy. However, the TA are the major focus of this study. See Ogden and Rosenberg (2006) for further details on methods.

2.3. Tree reconstruction analyses

Each of the datasets (15,400 TA and 15,400 HA) were analyzed under parsimony with gaps treated in three ways: as an unknown, as a fifth state character, and as a separate presence/absence character (Simmons and Ochoterena, 2000). In order to code gaps as presence or absence characters, we automated the SIC method of Simmons and Ochoterena (2000) to allow coding of thousands of input files (program available upon request from M.S. Rosenberg). GapCoder (Young and Healy, 2003) can only recode individual data files, but was used to cross check our implementation of the SIC scheme.

All phylogenetic analyses were performed using PAUP* Version 4.b10 Windows (Swofford, 2002). For each treatment of gaps, the TA and HA were analyzed identically allowing for subsequent direct comparisons. The analyses consisted of 100 random additions with TBR swapping and all other default settings, except for using GapMode = NewState for the gaps as a fifth state character approach. When multiple trees were recovered, the strict consensus of these trees was used as the result. The total alignment accuracy (TAA) is the average accuracy of all pair wise sequence comparisons in the multiple alignment (Ogden and Rosenberg, 2006). Topological accuracy is referred to in terms of the Robinson and Foulds (1981) distance to the known true tree.

3. Results

The average number of total characters and parsimony informative characters for the analyzed data sets for each of the three gap treatment approaches are found in Table 1 (see Supplemental information for details on each of the random trees). Across all tree shapes, the resulting data sets used for unknown and fifth state analyses consisted of matrices with average lengths (number of columns) of 2362 and 2145 for the TA and the HA, respectively. After SIC of the TA, an average of 218 additional characters (columns in the matrix) were added when the maximum evolutionary distance among any two taxa was 1, and an average of 396 characters were added when the distance was 2. Thus, the new data matrices, which include the additional SIC gap characters, resulted in average lengths of 2668 and 2414 for the TA and the HA, respectively. Across all tree shapes, the average increase in number of parsimony informative characters, when gaps are treated as a fifth state was greater in TA (185 additional characters) than in HA (120 additional characters) datasets. However, for the pectinate tree shape, when gaps were treated as a fifth state, there were more characters added for HA (114) than for TA (99).

A pair wise comparison, for the TA, of each of the three approaches (Table 2) of treating gaps indicated that 82% of the time there was no difference between the three treatments. In other words, any of the two contrasting methods performed equally well (or equally poorly as the case may be) in the majority of the analyses. For the balanced tree shape, the percentage of cases with no difference reached 88%. Contrastingly, in the pectinate tree shape, the value decreased to 66% where no difference was seen across the gap treatment comparisons.

TA analyses that treated gaps either as fifth state character or using SIC outperformed treating gaps as unknown (Fig. 1 and Table 2). For example, across all tree shapes, of the 23% fifth vs. unknown cases, where there was a difference between the approaches, treating gaps as fifth state outperformed unknown 83% of the time (2928 out of 3550 cases). For the SIC vs. unknown the results were even more pronounced with SIC recovering more accurate topologies in 92% of the cases (1792 out of 1862) where there was a difference. However, it should be noted that the SIC vs. unknown contrast consisted of more cases where there was

Table 1
Average number of characters for the different analyzed data sets across all tree types and conditions

	TA # characters	TA # parsimony informative characters	HA # characters	HA # parsimony informative characters	Max evolutionary distance	
					Rate = 1	Rate = 2
All tree shapes						
Unknown	2362.02	1644.34	2145.37	1684.92		
Fifth state	2362.02	1828.97	2145.37	1804.47		
Simple indel coding	2668.30	1752.44	2414.08	1799.87		
# Additional characters added after fifth state treatment		184.62		119.55		
# Additional characters added after simple indel coding	306.28	108.09	268.71	114.96	218.11	396.05
Balanced tree shape						
Unknown	2363.39	1726.84	2162.06	1769.61		
Fifth state	2363.39	1925.77	2162.06	1891.62		
Simple indel coding	2673.31	1863.92	2452.79	1916.83		
# Additional characters added after fifth state treatment		198.92		122.01		
# Additional characters added after simple indel coding	309.92	137.07	290.73	147.22	218.25	401.59
Pectinate tree shape						
Unknown	2410.06	1506.67	2159.30	1545.64		
Fifth state	2410.06	1606.13	2159.30	1659.53		
Simple indel coding	2752.69	1547.51	2453.64	1618.69		
# Additional characters added after fifth state treatment		99.46		113.89		
# Additional characters added after simple indel coding	342.64	40.84	294.34	73.05	243.72	441.55

The last two columns are the number of additional characters added to the TA after simple indel coding procedure for the maximum evolutionary distance for each of the different rates used during simulation.

Table 2
Different gap treatment comparisons across the indicated tree shapes and all conditions for the true alignments (TA) and the hypothesized alignments (HA)

	Fifth vs. unknown			SIC vs. fifth			SIC vs. unknown		
	No difference	5th state	Unknown	No difference	SIC	5th State	No difference	SIC	Unknown
TA comparison across all 7 tree shapes									
Total (%)	76.95	19.01	4.04	82.49	5.65	11.86	87.91	11.64	0.45
# of cases	11,850	2928	622	12,704	870	1826	13,538	1792	70
% of cases with difference		82.48	17.52		32.27	67.73		96.24	3.76
TA comparison across balanced tree shape									
Total (%)	83.73	12.82	3.45	87.05	4.36	8.59	92.64	7.14	0.23
# of cases	1842	282	76	1915	96	189	2038	157	5
% of cases with difference		78.77	21.23		33.68	66.32		96.91	3.09
TA comparison across pectinate tree shape									
Total (%)	57.18	34.68	8.14	64.59	12.45	22.95	75.59	22.73	1.68
# of cases	1258	763	179	1421	274	505	1663	500	37
% of cases with difference		81.00	19.00		35.17	64.83		93.11	6.89
HA comparison across all 7 tree shapes									
Total (%)	72.88	12.91	14.21	76.24	13.27	10.49	84.18	8.62	7.20
# of cases	11,224	1988	2188	11,741	2044	1615	12,964	1327	1109
% of cases with difference		47.61	52.39		55.86	44.14		54.47	45.53

The number of cases represents the number of analysis replicates (out of a total 15,400 for all tree shapes and 2200 for the balanced and pectinate tree shapes) where the particular gap treatment (unknown, fifth state, or SIC) outperformed the opposing method. No difference indicates the number of cases where both methods performed equally (well or poor).

no difference, relative to fifth state vs. unknown. So although the percentage was higher, the absolute number of cases where fifth state coding outperformed unknown (as compared to SIC vs. unknown) was greater. Thus while fifth state coding recovered more accurate topologies, SIC was a more conservative method of adding additional characters (unknown only outperformed SIC 70 times, while it outperformed fifth state 622 times). However, the down side of this conservative reliability was that in many analyses the indel coding was not sensitive enough to recover more accurate topologies than the unknown analyses (see below). The same general trends were recovered for each of the separate tree shape breakdowns as well (Supplemental material).

The SIC vs. fifth state comparisons, for TA analyses, indicated that, on average, treating gaps as a fifth state character resulted in more accurate tree reconstruction. For example, across all tree shapes, of the 18% of the cases

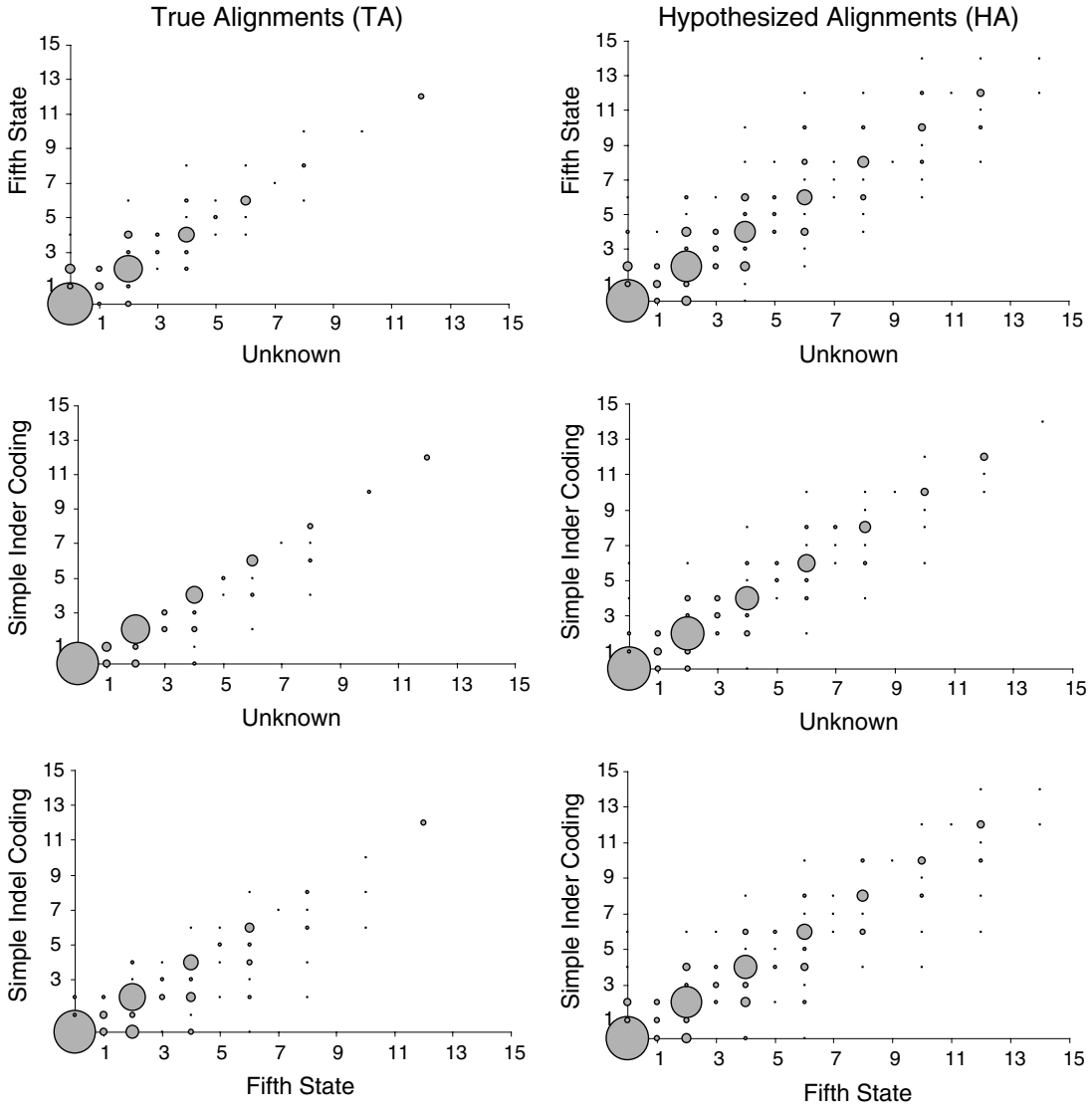

Fig. 1. Bubble plot, representing the relative number of cases, for comparisons of the three gap treatment approaches. The *X* and *Y* axes are distances from the true tree. Points along the diagonal indicate that there was no difference in reconstructed topological accuracy between the two contrasted approaches. Points above and below indicate the relative amount of cases where the gap treatment recovered less accurate topologies.

where there was a difference, treating gaps as fifth state outperformed SIC 68% of the time (Table 1 and Fig. 1). So while SIC was more reliable (because it outperformed unknown gap treatment in a higher percentage than fifth state vs. unknown), there were consistently more cases where treating gaps as a fifth state recovered more accurate topologies, even though the difference may be minimal.

The above results are all based on the true alignments (TA). Using the hypothesized alignments (HA) led to a different pattern. There was no difference between the different comparisons of gap treatment 77% of the time, and when there was a difference all of the methods performed essentially equally well (Table 2 and Figs. 1 and 2). For example, for the cases where there was a difference, treating gaps as unknown recovered slightly more accurate topologies (14% of the cases) than treating gaps a fifth state (13% of the characters). Furthermore, as alignment accuracy decreased, all three comparisons showed that, on average, differential gap treatment did not affect topological accuracy. In other words, contrary to many of the TA, treating gaps as fifth state or SIC did not outperform treating gaps as unknown across the span of generated alignment error in HA. This generalization is based on a moving average across many analyses and any one dataset could be highly affected by any particular gap treatment and alignment inaccuracy.

4. Discussion

4.1. Treatment of gaps

The main question that stimulated this study was: within a parsimony framework, is it better to use gaps as information and, if so, what approach to gap coding leads to more accurately reconstructed phylogenies? While many approaches are now available, we decided to investigate

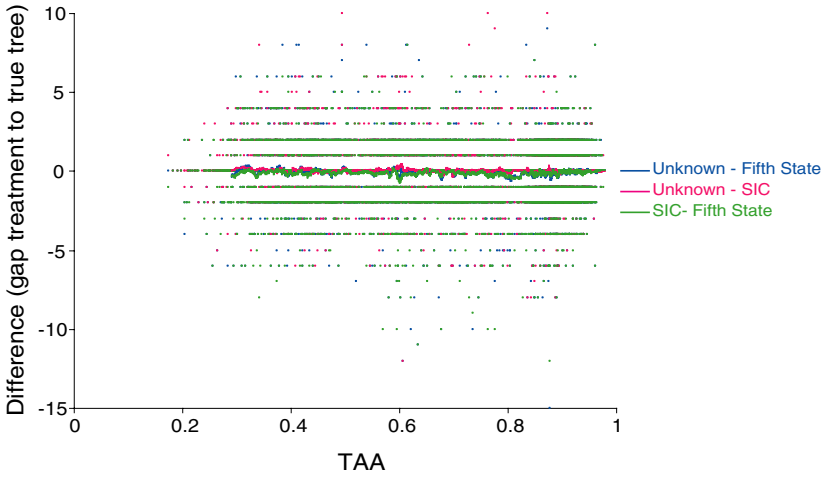

Fig. 2. Relationship, across all HA analyses and tree shapes, of total alignment accuracy (TAA) and the difference of the tree distance between one gap treatment approach vs. true tree and opposing gap treatment approach vs. true tree [e.g., (unknown reconstructed tree vs. true tree)—(5th state reconstructed tree vs. true tree)]. The three comparisons for treating gaps as unknown, fifth state character, and Simple Indel Coding are shown. Points to the far right are the most accurate alignments while points to the left are the least accurate alignments. Points above the 0 line are cases where the first approach recovered less accurate topologies than the contrasted approach, and vice versa. Many points may be superimposed upon one another. The lines are moving averages based on an overlapping sliding window of 100 consecutive points.

this question within the parsimony framework and to limit the analyses to the three most widely used automated methods. The fact that 82% of the time there is no difference, across all tree shapes and conditions for TA, is important. Thus, for the majority of true alignments across a wide range of tree shapes, it does not matter which method is used. To reiterate though, this is in terms of distance to the true tree; the fact that two approached performed equally well, or equally poor, does not mean that the same topology was necessarily recovered. Balanced trees resulted in more cases where no difference was found between alternative gap treatments than pectinate trees. Therefore, it would be preferable to more thoroughly explore the use of differential gap treatments when a recovered topology is more pectinate (Aagesen et al., 2005; Ogden and Whiting, 2003). These generalizations are based on trends and averages across all analyses, and it should again be emphasized that for any one dataset, even if the tree is balanced, gap treatment may have an effect (positive or negative). In contrast to the many cases where no difference was registered, our results indicate that, across our studied conditions, there is a relatively large percentage (e.g., 18% across all tree shapes up to 34% in pectinate tree shape analyses) of cases that did react differently to alternative gap treatments. In the majority of these cases, treating gaps as a fifth state character or as SIC (Simmons and Ochoterena, 2000) recovered more accurate topologies. Hence, we suggest (in parsimony analyses) that gaps should minimally be treated in one of these two ways, as no other automated and objective gap treatment approaches are currently available (although see Muller, 2006). On average, both approaches perform equally well. However, if the particular tree shape and dataset are such that it does make a difference, one is more likely to add accuracy to the phylogenetic reconstruction by coding gaps as a fifth state character.

Initially the superiority of fifth state coding over SIC for TA analyses, as judged by the number of cases where the former recovered more accurate topologies than the latter (Table 2 and Fig. 1), was surprising. However, after further investigation, the number of parsimony informative characters that are added for each approach (Table 1), explains, in part, this phenomenon. Fifth state coding added around 77 more parsimony informative characters than SIC, which could account for the recovery of more accurate topologies, due to the potential additional phylogenetic information content (i.e., synapomorphies). However, as noted above, SIC is a more conservative approach and therefore recovered less accurate topologies only in a small percentage of cases when compared to unknown gap treatment. These results suggest either approach is better than not coding gaps, but based on our data (particularly for more accurate alignments), treating gaps as a fifth state character appears to be the more preferable method, even if only slightly better. The specific effect of indel length on these two methods needs to be examined further (see below).

Because we do not know the actual true alignments with non-simulated data, it was very important to examine the results for the HA analyses. Coding gaps as fifth state or with SIC for HA analyses, on average, did not increase topological accuracy (Table 2 and Figs. 1 and 2) across the span of generated alignment error. However, this does not mean that gap coding will not affect any one particular analysis as many single dataset cases exist where differential gap coding made a large difference (note the point spread above and below the *y*-axis in Fig. 2). Nevertheless, it is readily demonstrated that for alignments that were 97% accurate as measured by TAA all the way down to alignments with less than 20% accuracy (which are almost random alignments), there is no average difference. Why then, does coding gaps as characters for the TA result in topolog-

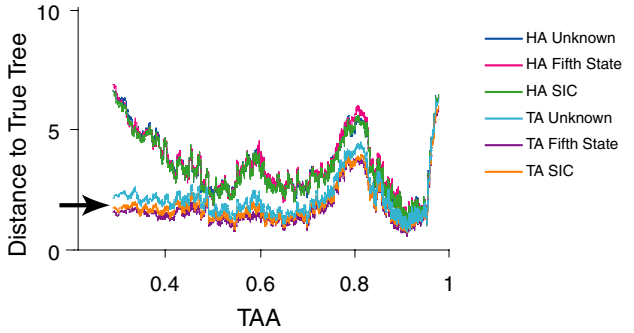

Fig. 3. Moving average lines of treating gaps as unknown, fifth state, and Simple Indel Coding plotted against total alignment accuracy (TAA). For the HA, there is no approach to gap coding, on average, that outperforms the others (top 3 lines). The arrow indicates the area where coding gaps outperforms treating gaps as unknown for TA associated to datasets that resulted in very poor HA (bottom three lines).

ical accuracy differences when coded gaps vs. unknown comparisons are made? In order to elucidate this conundrum, the average lines for HA and TA topological accuracy results, in terms of TAA vs. distance to true tree, were plotted (Fig. 3). While, there is no alignment error for the TA (they are exactly true by definition), the plot allows one to identify data sets that would have been difficult to align. In the region where HA alignments are very inaccurate (less than 50%), TA topologies with gap coding outperform TA topologies with gaps treated as unknown. This is due to datasets (and associated tree types) that result in poor alignments using ClustalW default settings. Thus, for the corresponding HA datasets in this region, when more accurate alignments could be generated (approaching the true alignment), then coding gaps as characters would result, on average, in more accurate topologies than treating gaps as missing. Hence, across datasets that have little possibility for large alignment errors, one is no better or worse off, on average, coding gaps. For datasets where alignment is likely to be problematic, if one can create more accurate alignments, then coding gaps is better than treating them as unknowns, because on average, topological accuracy increases.

Given the above conclusion, how does one identify datasets that are hard to align? This inquiry will require further investigation, but we can mention some of the contributing factors to alignment error in our study. One issue, evolutionary distance, is highlighted by the fact that there were 2579 datasets resulting in alignment accuracy (as measured by TAA) lower than 50%. Of these, 2330 (90%) were datasets simulated with the maximum evolutionary distance of two. Of the other 249 cases, simulated with a maximum evolutionary distance of one, almost all were simulation replicates from ultrametric equal-branch length random tree shapes. However, although these replicates (with the distance of one) contained large amounts of alignment error, they recovered completely accurate topologies, with only a few exceptions. Therefore, 16 taxa datasets containing pairwise comparisons (Rosenberg, 2005a) with a maximum evolutionary distance of 1 or less are much less prone to differential gap treatments affecting topological accuracy. Similarly, of the 2579 datasets with TAA <50%, 2239 were generated under the random branch length condition. Additionally, associated tree shape has also been shown to play a significant role in alignment accuracy and tree reconstruction (Ogden and Rosenberg, 2006). One interesting aspect of tree shape is that although pectinate trees, on average, contain less alignment error than more balanced topologies, these errors have a larger effect on tree reconstruction then the same amount of error in a balanced tree. Although there are relatively few pectinate topologies in the <50% accuracy range, gap coding in the corresponding TA pectinate tree datasets showed a larger difference than in more balanced trees. Therefore, datasets that have large evolutionary distances (2 or more), associated with deep, variable branch length, and certain types of topologies may be considered "hard to align" and gap treatment may play an important role during tree reconstruction. Also worthy of noting, is that Ogden and Rosenberg (2006) showed that, on average, for more balanced tree shapes and shorter branch lengths, alignment error may have little affect on topological reconstruction, and that for more pectinate tree shapes and longer branches the effect is much more pronounced.

4.2. The indel model

One final issue has to do with the biological accuracy of the indel model used and how the modeled number and length of indels may bias our conclusions. Although very simple, the model used is not tremendously unrealistic, particularly for non-coding DNA. Insertions and deletion events were independently modeled through a relatively simple Poisson processes with frequency of occurrence on each branch based (indirectly) on the branch length and general rate parameters obtained from empirical studies (Ophir and Graur, 1997; Sundstrom et al., 2003). Although the decision to model insertion and deletion events separately was likely inconsequential to this study, it could have importance for future work since advances in multiple sequence alignment have found advantages to treating them as separate processes (Löytynoja and Goldman, 2005). While some indel models are based on single base length events (e.g., Thorne et al., 1991), in our simulations individual indel events were not restricted to single base pairs but were drawn from a size distribution. The Poisson distribution we used for indel sizes appears to be a poor fit to empirically derived size distributions estimated from entire genome alignments (Chimpanzee Sequencing and Analysis Consortium); however, it should be noted that this and other empirically determined patterns of indel size (from pairwise comparisons of mammalian genomes) cannot easily be modeled by any standard theoretical distribution. Despite the limitations and simplicity of our model, the produced alignment accuracies are very similar to those found by other researchers using alternate indel models (Keightley and Johnson, 2004; Pollard et al., 2004).

Despite these potential problems with our indel model, we hypothesize that our general conclusions will hold true for a similar constructed study, even if a more "realistic" model were employed. We simulated an expected rate of one insertion event (mean length = 4) for every 100 substitutions and one deletion event (mean length = 4) for every 40 substitutions. These parameters led to alignments (HA) that ranged between 97% and 20% accurate, thus, the simulations produced alignment error across the large majority of the possible alignment space. Our indel rates are intermediate between two possible dataset extremes: no indels and almost all indels. As datasets contain fewer and shorter indels, the effect of gap coding will be less. For example, in 82% of the comparisons no difference was seen between approaches for treating gaps; if one were to simulate shorter and less frequent indel events, the proportion of contrasts with no difference would likely be larger. On the other hand, as indels become longer and more frequent, coding gaps as unknown would introduce more missing data into the matrix and would most likely lead to less accurate tree reconstructions; however, if gaps were treated as a fifth state, there would likely be more contrasts that have a difference, but there would likely also be more homoplasious indel sites, and so the percentage of cases where coding gaps as a fifth state outperforms unknown coding may not increase significantly.

It is important to underscore that the specific distribution of gaps does not matter under parsimony analysis when gaps are treated as unknowns or as a fifth state character. In fact, one could randomly rearrange the columns, and the new dataset, although mixed up in order, would generate the identical phylogenetic estimate. For SIC and other similar coding approaches, however, the distribution of indel lengths may have an affect on phylogenetic accuracy. An array of simulations across a range of modeled indel lengths would need to be generated and analyzed to address this issue (for example, simulating the identical total number of indel sites, but distributed in larger or smaller gap clusters). Thus, while our study provides some small insight into the effect that indel frequency and length can have on alignment and tree reconstruction accuracy, this is an issue that needs to be further scrutinized, particularly for approaches other than parsimony-based ones.

In this study, we have only examined three different treatments of gaps in multiple sequence alignment, and there are still many questions that need to be answered. It is difficult to predict exactly how "complex indel coding" (Simmons and Ochoterena, 2000) and other approaches to gap coding will perform, but undoubtedly in many cases no difference will be seen. In cases where differences are observed, using gaps as information may recover more accurate topologies when the associated "hard to align" alignment is aligned more accurately. Other approaches have largely ignored the issue of treating gaps as phylogenetic information. For example, Bayesian and Likelihood methods either treat each gap as ambiguous data or ignore the gaps by removing the entire column from the analysis. However, Bayesian and Likelihood approaches that include indel formation in the model are becoming more readily available and practical for larger data sets (Holmes, 2005; Holmes and Bruno, 2001; Knudsen and Miyamoto, 2003; McGuire et al., 2001; Metzler, 2003; Miklos et al., 2004; Mitchison, 1999; Rivas, 2005; Thorne et al., 1991, 1992). A similar trend is being seen in concurrent analysis frameworks where the alignment and phylogeny are estimated simultaneously (Fleissner et al., 2005; Lunter et al., 2005; Redelings and Suchard, 2005). The treatment of gaps as information in modeling approaches is an area that needs further study. Approaches for dealing with gaps in distance-based algorithms also need to be explored.

In conclusion, simulation with indels produced alignments where the true homologies were known and these were then analyzed under the three most common approaches of treating gaps in parsimony. For these true alignments, most of the time there was no difference in topological accuracy for the different methods of gap coding. However, in the true alignment cases where a difference was present, coding gaps recovered more accurate topologies than treating gaps as unknowns. Treating gaps as fifth state outperformed simple indel coding in a majority of the true alignment cases where there was a difference. Therefore, our data suggest that using gaps as information is preferable to treating gaps as unknown and that coding gaps as fifth state characters is slightly preferable to simple indel coding. We also showed that data sets with large maximum evolutionary distances, and certain tree shapes may be more affected by differential gap coding approaches.

Acknowledgments

This work was supported by the NIH R03-LM008637 (to M.S.R.) and Arizona State University. We thank Sudir Kumar, Michael Whiting, Karl Kjer, Rod Page, and anonymous reviewers for providing comments on early versions of this manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2006.07.021.

References

Aagesen, L., Petersen, G., Seberg, O., 2005. Sequence length variation, indel costs, and congruence in sensitivity analysis. Cladistics 21, 15–30.

Baldwin, B.G., Sanderson, M.J., Porter, J.M., Wojciechowski, M.F., Campbell, C.S., Donoghue, M.J., 1995. The Its region of nuclear ribosomal dna—a valuable source of evidence on angiosperm phylogeny. Ann. MO Bot. Garden, 247–277.

Barriel, V., 1994. Molecular phylogenies and how to code insertion/deletion events. Comptes Rendus de l'Academie des Sciences—Serie III 317, 693–701.

Baum, D.A., Sytsma, K.J., Hoch, P.C., 1994. A phylogenetic analysis of Epilobium (Onagraceae) based on nuclear ribosomal DNA-sequences. Syst. Bot. 19, 363–388.

Blanchette, M., Green, E.D., Miller, W., Haussler, D., 2004a. Reconstructing large regions of an ancestral mammalian genome in silico. Genome Res. 14, 2412–2423.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F.A., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., Haussler, D., Miller, W., 2004b. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 14, 708–715.

Consortium, T.C.S.A., 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 436, 69–87.

Crandall, K.A., Fitzpatrick Jr., J.F., 1996. Crayfish molecular systematics: Using a combination of procedures to estimate phylogeny. Syst. Biol. 45, 1–26.

Fleissner, R., Metzler, D., Haeseler, A., 2005. Simultaneous statistical multiple alignment and phylogeny reconstruction. Syst. Biol. 54, 548–561.

Ford, V.S., Thomas, B.R., Gottlieb, L.D., 1995. The same duplication accounts for the PgiC genes in *Clarkia xantiana* and *C. lewisii* (Onagraceae). Syst. Bot. 20, 147–160.

Freudenstein, J.V., Chase, M.W., 2001. Analysis of mitochondrial nad1b-c intron sequences in Orchidaceae: utility and coding of length-change characters. Syst. Bot. 26, 643–657.

Geiger, D.L., 2002. Stretch coding and block coding: two new strategies to represent questionably aligned DNA sequences. J. Mol. Evol. 54, 191–199.

Gillespie, J.J., Yoder, M.J., Wharton, R.A., 2005. Predicted secondary structure for 28S and 18S rRNA from Ichneumonoidea (Insecta: Hymenoptera: Apocrita): impact on sequence alignment and phylogeny estimation. J. Mol. Evol. 61, 114–137.

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

Golenberg, E.M., Clegg, M.T., Durbin, M.L., Doebley, J., Ma, D.P., 1993. Evolution of a noncoding region of the chloroplast genome. Mol. Phylogenet. Evol. 2, 52–64.

Gonzalez, D., 1996. Codificación de las inserciones-deleciones en el análisis filogenético de secuencias genéticas. Boletı´n de la Sociedad Botánica de México 59, 115–129.

Graham, S.W., Reeves, P.A., Burns, A.C.E., Olmstead, R.G., 2000. Microstructural changes in noncoding chloroplast DNA: interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. Int. J. Plant Sci. 161, S83–S86.

Hall, B.G., 2005. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. Mol. Biol. Evol. 22, 792–802.

Hasegawa, M., Kishino, K., Yano, T., 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22, 160–174.

Hennequin, S., Ebihara, A., Ito, M., Iwatsuki, K., Dubuisson, J.-Y., 2003. Molecular systematics of the fern genus Hymenophyllum s.l. (Hymenophyllaceae) based on chloroplastic coding and noncoding regions. Mol. Phylogenet. Evol. 27, 283–301.

Hibbett, D.S., Fukumasanakai, Y., Tsuneda, A., Donoghue, M.J., 1995. Phylogenetic diversity in Shiitake inferred from nuclear ribosomal DNA-sequences. Mycologia 87, 618–638.

Holmes, I., 2005. Using evolutionary Expectation Maximization to estimate indel rates. Bioinformatics 21, 2294–2300.

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

Kawakita, A., Sota, T., Ascher, J.S., Ito, M., Tanaka, H., Kato, M., 2003. Evolution and phylogenetic utility of alignment gaps within intron sequences of three nuclear genes in bumble bees (*Bombus*). Mol. Biol. Evol. 20, 87–92.

Keightley, P.D., Johnson, T., 2004. MCALIGN: stochastic alignment of noncoding DNA sequences based on an evolutionary model of sequence evolution. Genome Res. 14, 442–450.

Kelchner, S.A., 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. Ann. MO Bot. Garden 87, 482–498.

Kelchner, S.A., 2002. Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. Am. J. Bot. 89, 1651–1669.

Kjer, K.M., 1995. Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. Mol. Phylogenet. Evol. 4, 314–330.

Knudsen, B., Miyamoto, M.M., 2003. Sequence alignments and pair Hidden Markov models using evolutionary history. J. Mol. Biol. 333, 453–460.

Kretzer, A., Li, Y.N., Szaro, T., Bruns, T.D., 1996. Internal transcribed spacer sequences from 38 recognized species of *Suillus sensu lato*: phylogenetic and taxonomic implications. Mycologia 88, 776–785.

Li, W.-H., 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA.

Löytynoja, A., Goldman, N., 2005. An algorithm for progressive multiple alignment of sequences with insertions. PNAS 102, 10557–10562.

Lunter, G., Miklos, I., Drummond, A., Jensen, J., Hein, J., 2005. Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinformatics 6, 83.

Lutzoni, F., Wagner, P., Reeb, V., Zoller, S., 2000. Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. Syst. Biol. 49, 628–651.

Manos, P.S., Steele, K.P., 1997. Phylogenetic analyses of "higher" Hamamelididae based on plastid sequence data. Am. J. Bot. 84, 1407–1419.

McGuire, G., Denham, M.C., Balding, D.J., 2001. Models of sequence evolution for DNA sequences containing gaps. Mol. Biol. Evol. 18, 481–490.

Metzler, D., 2003. Statistical alignment based on fragment insertion and deletion models. Bioinformatics 19, 490–499.

Miklos, I., Lunter, G.A., Holmes, I., 2004. A "Long Indel" model for evolutionary sequence alignment. Mol. Biol. Evol. 21, 529–540.

Mitchison, G.J., 1999. A probabilistic treatment of phylogeny and sequence alignment. J. Mol. Evol. 49, 11–22.

Muller, K., 2006. Incorporating information from length-mutational events into phylogenetic analysis. Mol. Phylogenet. Evol. 38, 667–676.

Ogden, T.H., Rosenberg, M.S., 2006. Multiple sequence alignment accuracy and phylogenetic inference. Syst. Biol. 55, 314–328.

Ogden, T.H., Whiting, M., 2003. The problem with "the Paleoptera Problem": sense and sensitivity. Cladistics 19, 432–442.

Ophir, R., Graur, D., 1997. Patterns and rates of indel evolution in processed pseudogenes from humans and murids. Gene 205, 191–202.

Petersen, G., Seberg, O., Aagesen, L., Frederiksen, S., 2004. An empirical test of the treatment of indels during optimization alignment based on the phylogeny of the genus *Secale* (Poaceae). Mol. Phylogenet. Evol. 30, 733–742.

Pollard, D., Bergman, C., Stoye, J., Celniker, S., Eisen, M., 2004. Benchmarking tools for the alignment of functional noncoding DNA. BMC Bioinformatics 5, 6.

Redelings, B., Suchard, M., 2005. Joint Bayesian estimation of alignment and phylogeny. Syst. Biol. 54, 401–418.

Rivas, E., 2005. Evolutionary models for insertions and deletions in a probabilistic modeling framework. BMC Bioinformatics 6, 63.

Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.

Rosenberg, M., 2005a. Evolutionary distance estimation and fidelity of pair wise sequence alignment. BMC Bioinformatics 6, 102.

Rosenberg, M.S., 2005b. Multiple sequence alignment accuracy and evolutionary distance estimation. BMC Bioinformatics 6, 278.

Rosenberg, M.S., 2005c. MySSP: non-stationary evolutionary sequence simulation, including indels. Evolutionary Bioinformatics Online 1, 51–53.

Rosenberg, M.S., Kumar, S., 2003. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. Mol. Biol. Evol. 20, 610–621.

Simmons, M.P., Ochoterena, H., 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst. Biol. 49, 369–381.

Simmons, M.P., Ochoterena, H., Carr, T.G., 2001. Incorporation, relative homoplasy, and effect of gap characters in sequence-based phylogenetic analyses. Syst. Biol. 50, 454–462.

Sundstrom, H., Webster, M.T., Ellegren, H., 2003. Is the rate of insertion and deletion mutation male biased? Molecular evolutionary analysis of avian and primate sex chromosome sequences. Genetics 164, 259–268.

Swofford, D.L., 1993. PAUP: Phylogenetic Analysis Using Parsimony, version 3.1.1. Smithsonian Institution.

Swofford, D.L., 2002. PAUP* Phylogenetic analysis using parsimony (* and other methods), version 4.0b10. Sinauer Associates.

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

Thorne, J.L., Kishino, H., Felsenstein, J., 1991. An evolutionary model for the maximum likelihood alignment of sequence evolution. 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.

van Dijk, M.A.M., Paradis, E., Catzeflis, F., Jong, W.W.D., 1999. The virtues of gaps: *Xenarthran* (Edentate) monophyly supported by a unique deletion in alpha A-crystallin. Syst. Biol. 48, 94–106.

Vogler, A.P., DeSalle, R., 1994. Evolution and phylogenetic information content of the ITS-1 region in the tiger beetle *Cicindela dorsalis*. Mol. Biol. Evol. 11, 393–405.

Wheeler, W., 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? Cladistics 12, 1–9.

Wheeler, W.C., Gladstein, D., De Laet, J., 2003. POY version 3.0.11. American Museum of Natural History.

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