Morrison (2009) raises a fundamental question, “Why would phylogeneticists ignore computerized sequence alignment?” Although well aware of the difficulties, he considers the whole issue is a “gaping hole that needs to be filled.” Particularly with the expansion of genomic-scale data, there are many advantages to using automated alignment for phylogenetic analyses, the most obvious being that it is much faster and potentially less prone to experimenter bias. So yes, it is obviously desirable to automate data preparation as far as possible, but the question remains whether we are yet at the stage that automated sequence alignment is sufficiently accurate to permit unbiased phylogenetic inference. The question of automated alignment leads directly into the question of the evaluation of gaps in an alignment, and whether the gaps are an important source of additional phylogenetic information.

In this paper, we use an example shorebird data set to explore three related questions regarding the interplay between alignment and phylogeny estimation: 1) can gap-rich alignments be used for reasonably accurate and unbiased phylogenetic inference? 2) How much phylogenetic information is contained in gap characters as compared with the nucleotides alone? 3) Are models of the insertion/deletion process advantageous and if so at what level of phylogenetic divergence? We report that there is considerable information created by the indel (insertion/deletion) process that is potentially available for phylogenetic inference. Ideally, we should be able to independently obtain the same tree from both nucleotide characters and from gap characters; however, there is still considerable variability in the alignments produced by different programs. We predict that better and more computationally tractable models of the indel process will be required before the information in gaps can be fully exploited for phylogenetic inference.

It is well known that issues of alignment (e.g., Golubchik et al. 2007; Martin et al. 2007; Wong et al. 2008; Liu et al. 2009; Anisimova et al. 2010; Simmons et al. 2010) can significantly affect the tree inferred from the data. Furthermore, we expect that introns could be an increasingly important type of data, in particular for closer divergences, because they contain two classes of data for which trees can be compared quantitatively—the aligned nucleotides, and the position of gaps in the alignment. The first is the standard sequences (DNA or protein) that we all use routinely, but a second allows a separate quantitative test, and is whether the same tree (or subtrees) are found from treating the gaps as characters. We have long emphasized that it is standard scientific approach to have different data sets and to compare the results quantitatively (Penny et al. 1991; Pratt et al. 2009). Given also that we expect the Markov models we use with sequence data to saturate at the deepest divergences (Mossel and Steel 2004), it is important to have alternative data to compare results, and indel data have long been available for such tests (Golenberg et al. 1993).

Following both Kjer et al. (2007) and Morrison (2009), we distinguish between indels (real insertions or deletions of fragments of DNA) and gaps (hypothetical insertions or deletions in a proposed multiple sequence alignment). In principle, as indels have so many possible locations where they could occur, they may be examples of a “rare genomic change” (Rokas and Holland 2000) and such data (e.g., the position of “jumping genes”) have been used effectively for the relationships of marsupials (Nilsson et al. 2010) or with numts for primates (Hazkani-Covo 2009) but see also Han et al. (2011). If indels, and hence gaps in the alignment, are sufficiently unlikely to independently occur at the same position in a sequence along different branches of the phylogenetic tree, then it could be expected that they will provide homoplasy-free phylogenetic information, that is, parsimony might be the maximum likelihood estimator for this class of data (Steel and Penny 2000, 2005). However, the extent to which this is the case is questionable, for instance, Kjer (1995) and Gillespie et al. (2005) found evidence for regions of hypervariability that are more likely to experience indels. For this reason, there have been attempts over the last 20 years to put alignment on a similar model-based footing as for tree estimation.
In the early 1990s, Hein (1990) released a program that performed simultaneous alignment and tree building, and Thorne et al. (1991, 1992) developed two models of sequence evolution that incorporated indels. The first of these models, TKF91, only modeled indels of length 1, TKF92 allowed longer insertions or deletions but inserted fragments were subsequently indivisible. Thatte (2006) showed that the TKF91 model, which allows indels of length 1 only, was a good model mathematically in the sense of being “identifiable,” but the properties of more biologically realistic models that allow longer indels, for example, the TKF92 model, are still unknown. Later, Wheeler (2003) introduced a parsimony based approach to assessing homology and phylogeny simultaneously; this was implemented in the software POY (Wheeler 2003). Bali-Phy (Suchard and Redelings 2006) is a Bayesian-based method for simultaneously estimating the phylogenetic tree and the alignment it allows indels of length > 1 and also avoids the unrealistic constraint that inserted fragments are indivisible, however, it ignores branch lengths in the indel part of the model, meaning that indels are equally likely to occur on short or long edges (i.e., a parsimony like assumption). All the methods described above are in contrast to the more commonly used two-phase approach where an alignment is determined first and then is followed by phylogenetic inference. In most cases, these algorithms are slow compared with two-phase approaches and unfortunately do not always improve accuracy (Liu et al. 2009). These authors (Liu et al. 2009, 2010) recently introduced Simultaneous Alignment and Tree estimation (SATé) and demonstrated that it could have good performance on both simulated and some biological data sets. In order to illustrate the principles of gap placement in an automatic alignment, we evaluate the status of five alignment programs; Clustal (Thompson et al. 1994), MUSCLE (Edgar 2004), T-Coffee (Notredame et al. 2000), MAFFT (Katoh et al. 2002), and SATé (Liu et al. 2009). We briefly describe the similarities and differences between these alignment methods. Clustal performs pairwise sequence alignment to get a distance matrix followed by neighbor joining (NJ) to construct a guide tree. The NJ guide tree is also used to weight the importance of sequences, that is, sequences that are very similar to others are down weighted. Clustal also uses position and residue-specific gap opening and extension penalties (GOPS and GEPs). T-Coffee is also a progressive aligner but it is less greedy than Clustal when aligning a pair of sequences (or pair of groups of sequences) features of the other sequences can influence the decision. MAFFT was designed with speed in mind; it uses a Fast Fourier Transform of sequence characteristic to rapidly find homologous areas so that the size of the problem solved by dynamic programming is significantly reduced. Like T-Coffee it uses context sensitive GOPs and GEPs. MUSCLE uses k-mer libraries to speed up the construction of the initial guide tree; it then uses the Kimura 2 parameter model to find corrected distances from the draft MSA to compute another guide tree and realigns. Finally, MUSCLE allows refinement of the MSA by subdividing the alignment according to edges in the guide tree and realigning these subalignments. Iterative refinement is continued until no further improvement is made. SATé iteratively computes trees and alignments. SATé does not have a model for gaps (it just treats them as missing data in the tree estimation phase). It uses a divide-and-conquer approach in the alignment phase (by default, it farms out the subalignments to MAFFT and remerges them using MUSCLE). How the subalignments get broken up is decided using Center-Tree decomposition that tries to break the tree up into 2 roughly equal subsets around a central edge.

Because we were studying the question raised by Morrison (2009), our aim is not to compare (or evaluate) alignment programs but rather to study any differences between these popular alignment methods in terms of the resulting phylogeny, the gaps inferred, the phylogenetic information content of the gaps, and the extent to which the gaps appear to be examples of rare genomic changes (i.e., homoplasies free). In all cases, we used the default parameter settings of the programs, except for Clustal where we used ClustalX but set the GOP to 10 and the GEP to 0.1 as per the default settings in the online version of ClustalW2 at http://www.ebi.ac.uk/Tools/msa/clustalw2/index.html see also Carroll et al. (2006).

We have a special interest in the deeper avian evolutionary tree because we want to evaluate hypotheses about whether microevolutionary processes are sufficient for macroevolution (Penny and Phillips 2004). An important data set from Hackett et al. (2008) follows the trend of Fain and Houde (2004) of using a large proportion of intron data for evaluating deeper avian evolution. Such intron data are often variable in length, leading to alignments containing many gaps. Our practice (e.g., Morgan-Richards et al. 2008) is to remove sites with gaps from the alignment before doing tree estimation, but with β-fibrinogen intron 7 of Fain and Houde (2004) this approach would delete all sites. Clearly, information is lost by this conservative approach, but we need to know whether gap-rich alignments are capable of returning a reasonable tree.

For the present study, our test data set is from shorebirds (Charadriiformes) because they have a reasonably well-defined tree, good data sets are available and there are sufficient sequences to allow tests at different levels. We use data sets at two evolutionary depths—suborders (3–4 taxa each) and orders (11 taxa). We have four data sets, including protein-coding genes: the alignment of β-fibrinogen intron 7 (~1 kb, a subset of Hackett et al. 2008), the gap-coding alignment of the same data, the RAG-1 nuclear gene (~1kb) from Paton et al. (2003), and also mitochondrial protein-coding genes (~11 kb, Paton and Baker 2006) (All alignments and additional details on the data sets can be found at the Dryad data repository at http://datadryad.org, doi:10.5061/dryad.mg76th0n). Thus, we set out to compare trees from four sets of data (Table 1), find where
TABLE 1. Comparison of trees resulting from different alignment programs

<table>
<thead>
<tr>
<th></th>
<th>β-fibrinogen intron 7</th>
<th>β-fibrinogen intron 7 gap data</th>
<th>RAG1 exon</th>
<th>Mt exons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charadrii</td>
<td>✓</td>
<td>✓</td>
<td>(C,H)(B,F)</td>
<td>Only three taxa available</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Scolopaci</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>(R,P)(A,J)</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>No align</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>(R,P)(A,J)</td>
</tr>
<tr>
<td>Shorebirds</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>No align</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>(R,P)(A,J)</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>(R,P)(A,J)</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>(R,P)(A,J)</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>(R,P)(A,J)</td>
</tr>
</tbody>
</table>

aWithin each cell, the results are ordered Clustal, MUSCLE, T-Coffee, MAFFT, and SATé. The first letter of the genus name given in Figure 1 is used for identification, and a tick indicates the tree is in agreement with the default tree of Figure 1. Differences to this default tree (Fig. 1) are shown; a right arrow → indicates the species to the left of the arrow has moved adjacent to the species on the right; +1 indicates the species has moved 1 node deeper into the tree; two pairs of brackets indicates a different tree. Strict consensus trees of equally scoring maximum parsimony trees are shown for β-fib intron 7 gap data. All other trees were constructing using NJ. “No align” indicates that the program would not handle the large data set. Intron gap data sometimes gave no resolution (a polytomy) and such taxa are given in square brackets []. Details of the data sets are given in supplementary material. An important conclusion from this table is that both the alignment method and the data set can give different trees—it is not just one or the other.

automated alignment is possible, assess the similarity of the alignments, and make quantitative tests about the phylogenetic information content contained in the gap characters.

Figure 1 is the expected unrooted tree for the 11 shorebirds found by both Baker et al. (2007) and Hackett et al. (2008), and we use this as a reference tree. It contains three suborders, namely:

Figure 1. The unrooted default/null tree. The tree is based on Thomas et al. (2004), Paton and Baker (2006), Baker et al. (2007) and shows the three suborders—Lari, Scolopaci, and Charadrii. The first letter of a genus name (in bold) is used for identification in Table 1 except that “F” is used for Phegornis to avoid confusion with Pedionomus.
Lari (sea gulls [Larus], crab plover [Dromas], and button quail [Turnix]);
Scolopaci (turnstone [Arenaria], painted snipe [Rostratula], plains wanderer [Pedionomus], lily trotter [Jacana]); and
Charadrii (plover [Charadrius], sandpiper-plover [Phegornis], oystercatcher [Haematopus], stone curlew/thick knees [Burhinus]).

Both the Scolopaci and the Charadrii have four species with data available to test alignments, and although there are only three unrooted trees for each suborder, we can use them to test the different levels of divergence. For example, we can test the performance of introns (both nucleotide sequences and gaps) and compare them with other sequence data (exon and mitochondrial). Thus, comparisons can be made both within and between suborders. In this case, we have a small number of taxa and so this study is complementary to that of Liu et al. (2010) where they had large numbers of taxa.

As a matter of principle, we should make predictions about our expectations. We expect that intron data (both indels and aligned nucleotides) will be excellent for more recently diverged taxa (here within suborders) because a small number of insertions and deletions should be alignable accurately. Indels within intron data are expected to be useful within suborders and orders where we expect there would be a sufficient number of changes to give some consistency, an example is with Galliformes (Bonilla et al. 2010). If indels are sufficiently rare, then we would expect parsimony to be the maximum likelihood estimator (Steel and Penny 2005), and we can test whether the trees have the minimum possible number of changes—either by using the MinMax squeeze program (Holland et al. 2005b) or (in this case) by using PAUP* (Swofford 2002) to evaluate all trees for the 11 taxa. However, if identical indels can occur in parallel, we may need to include a more formal model of the indel process. At the order level, we expect exons (nuclear and mitochondrial) should still be performing successfully, however, phylogenetic signal from intron sequence and gaps may be weak. Nevertheless, we find significant differences both between alignment methods and between the data sets (Table 1). Thus, although we started with the question of Morrison (2009) and were disappointed that automatic alignment is not yet ready, we find that there is considerable information available in indels and at different levels.

RESULTS

Differences Between Alignment Methods

The five programs differ significantly in the results of their alignments (Table 1) both across genes and across methods. We note that differences across genes cannot be conclusively attributed to errors in alignment or tree estimation as there may be genuine differences due to gene tree–species tree incongruence. However, differences for a single gene indicate that at least one of the alignment methods must be inaccurate. We show an example of the variability in different alignment methods in Figure 2 on alignments for β-fibrinogen intron 7 with the suborder Scolopaci. Pedionomus was interesting in that it had a unique 291 bp insertion starting at base 115, and in preliminary experiments including or removing the insertion in Pedionomus led to different placing of other gaps some distance from the insertion (data not shown). This in itself was a warning about automated alignment, and the insertion was often deleted for

![Figure 2. Differences in alignments of β-fibrinogen intron 7 between Clustal, MUSCLE, T-Coffee, MAFFT, and SATé. For Jacana (Scolopaci). Sideways arrows identify equivalent gaps. In the first gap, MUSCLE, MAFFT, and SATé give a single 23 bp gap starting at position 98; Clustal has two gaps of 6 and 17 bps at positions 98 and 103, and T-Coffee the same combination of a 6 and 17 bp gaps but the second starts at a different location. The apparent single base gaps inserted by T-Coffee, and indicated by a circled 1, are artifacts in that the version of T-coffee used did not recognize “R” and “Y” codes. In the third gap example, each program gives two or three gaps totaling 47 bp in length but at different locations.](https://academic.oup.com/sysbio/article-abstract/61/6/1075/1664484/7)


further analysis. There is a minor issue in that T-Coffee (version 8.9710111) did not handle R and Y coding and so introduces a "gap" of length one nucleotide for every "R" or "Y" character state. Although this is a current problem and occurs here three times in Figure 2 (shown as a "1" within a small circle), we assume that this problem can be fixed and eliminated.

The second of the differences is of more concern and is that gaps are often not in the same location (Fig. 2 and for an alternative visualization see supplementary Fig. S1, available at the Dryad data repository at http://datadryad.org, doi:10.5061/dryad.mg76th0n). This may not be of concern with only a small number of gaps and when only using the nucleotide sequences themselves (and perhaps deleting columns back to a constant site). However, it is of more importance as the numbers of genes increases (such as with genomic data) or if we want to identify homologous gaps in order to use them as characters. At present, we define a homologous gap as starting and stopping at identical points in the sequence, and if there is ambiguity in the placement of the gap, then it is less certain that two gaps are homologous. This is especially a concern when gaps are of different lengths and it is unclear whether two gaps arose independently or whether they arose successively (e.g., with the longer gap perhaps being a secondary extension of the shorter). However, there are potentially more serious issues highlighted in Figure 2 when one program gives a single gap and others give two; or one program gives two and another three gaps. No doubt, some of the difficulties in the number and position of gaps arise from the different algorithms used as well as differences in gap penalties between programs. However, different gap penalties are outside the scope of the present study of exploring Morrison’s 2009 question.

**Phylogenetic Information in Gaps**

Being able to use additional information in sequences (including the position and lengths of gaps) is important, though some qualifications are necessary (Steel 2011). There is considerable information that we are not yet using routinely for phylogenetic estimation. A technique for evaluating additional information in aligned sequences is to calculate the probabilities that particular gap-based characters fit as well as they do on a tree under a null hypothesis that the character states (gap or no gap) occur at random. This is illustrated using 11 taxa from shorebirds, and the tree in Figure 1. Tables 2 and 3 show how an indel, as a rare genomic change, can be used to test hypotheses and is an extension of the work of Steel et al. (1992) for two-state characters (presence or absence of an indel) for up to 10 taxa. If at least two taxa share the same character state, in this case an indel of the same length and starting position, then we can calculate the probability that it will match a split of taxa in a tree already derived from different data. For example, for $n = 11$ taxa, there are 55 splits/combinations of just two taxa sharing a gap. This is calculated from the number of ways of selecting $r$ objects from $n$ objects, namely $\binom{n}{r}$ or $n! / [(n - r)! r!]$. Given the tree of Figure 1, only 3 of the 55 splits will fit the tree precisely (with a single change on the tree). Thus, the probability $P$ for a perfect fit is 0.055 (Table 2); the other 52 combinations require two changes on the same tree. This gives the expected parsimony score of a character representing a split of 9 taxa from 2 taxa as 1.945—three fitting the tree with one change and 52 with two changes.

For three taxa sharing the same character state (an 8:3 split), there are 165 combinations ($\binom{11}{3}$, 11 choose 3) with 3 fitting the tree precisely ($P = 0.018$), 28 requiring a second change ($P = 0.17$), and the remaining 134 requiring three changes (Table 2). Because an alignment can have many gaps the probabilities are multiplied together giving a powerful test (Steel et al. 1992) to reject, or otherwise, the null hypothesis that the gap characters occur at random. The assumption is made that the indel events are independent—but because each indel event occurs at a particular point in time on the tree, this appears reasonable as a starting assumption, though (just as for point mutations) we need to be alert to the possibility of “hotspots” where mutations are more likely to occur; there is some evidence for a bias toward loss (Johnson 2004). This approach of using the position of gaps is one way of using additional information in the sequences, and we come back to it later after evaluating the alignments.

It appears that there is high information content in the position of gaps. Considering for example the Clustal

---

**Table 2.** Numbers of splits on 11 taxa that fit the tree of Figure 1 with a cost of 1, 2, 3, 4, or 5 changes

<table>
<thead>
<tr>
<th>11-rr</th>
<th>Number of splitsa</th>
<th>Cost 1</th>
<th>Cost 2</th>
<th>Cost 3</th>
<th>Cost 4</th>
<th>Cost 5</th>
<th>Expected cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:1</td>
<td>11</td>
<td>11 (1b)</td>
<td>52 (0.945)</td>
<td>143 (0.812)</td>
<td>188 (0.570)</td>
<td>112 (0.242)</td>
<td>1.945</td>
</tr>
<tr>
<td>9:2</td>
<td>55</td>
<td>3 (0.055)</td>
<td>28 (0.170)</td>
<td>134 (0.812)</td>
<td>2.794</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8:3</td>
<td>165</td>
<td>2 (0.006)</td>
<td>25 (0.076)</td>
<td>115 (0.348)</td>
<td>3.482</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7:4</td>
<td>330</td>
<td>0 (0.000)</td>
<td>23 (0.050)</td>
<td>115 (0.249)</td>
<td>3.893</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6:5</td>
<td>462</td>
<td>0 (0.000)</td>
<td>23 (0.050)</td>
<td>115 (0.249)</td>
<td>3.893</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total 1023 19 128 364 400 112

---

aThere are $11C_r$ possible splits with $r$ taxa having the less frequent state.
bProbabilities of different split costs are shown in brackets.
TABLE 3. Summary of gap characters for different alignment methods grouped by the type of split and cost on the tree shown in Figure 1.

<table>
<thead>
<tr>
<th>Type of split</th>
<th>Cost 1</th>
<th>Cost 2</th>
<th>Cost 3</th>
<th>Cost 4</th>
<th>Average cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:1</td>
<td>39</td>
<td>33</td>
<td>44</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>9:2</td>
<td>3</td>
<td>4</td>
<td>1.571</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8:3</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>1.333</td>
<td></td>
</tr>
<tr>
<td>7:4</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1.333</td>
<td></td>
</tr>
<tr>
<td>6:5</td>
<td>None observed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>48</td>
<td>7</td>
<td>0</td>
<td>62 (23)</td>
<td></td>
</tr>
</tbody>
</table>

a Within each cell results are shown for the Clustal, MUSCLE, T-Coffee, and MAFFT intron alignment gap characters in turn (the SATé intron alignment was identical to the MAFFT alignment).

b Bold numbers in the bottom right cell show the parsimony score for each alignment, the number in brackets is the parsimony score for informative characters only.

results and without the long (291 bp) Pedionomus insertion, we find that the alignment has seven occasions where two taxa share the same gap (a 9:2 split). In three cases, the gaps fit the tree exactly (with a single change) and in the other four cases, the gap requires two changes. This gives an average of 1.57 [(3 × 1.0 + 4 × 2.0)/7 = 1.57] changes compared with the expected 1.95 (Tables 2 and 3). The results for the 8:3 splits (where three taxa share a gap) are even more striking. The different alignment methods have average costs of between 1.0 and 1.33, compared with the expected 2.79 changes. For the MAFFT alignment, the three gaps fit the tree exactly (requiring only a single “mutation”), giving an average cost of 1. The gaps that give 7:4 splits are similar with the different programs requiring one or two changes (none require three or four changes). This again gives an average cost of between 1.33 and 2 changes, but this time compared with an expected cost of 3.48 changes. There were no 6:5 splits in any alignment. Note that the 10:1 splits can fit onto any tree with a cost of one.

Are Indel Models Required?

The median network for the set of 42 splits representing the gap characters in the SATé alignment is shown Figure 3. It shows the high level of agreement between the gap characters and the tree in Figure 1 but also demonstrates that on this alignment, gap characters are not homoplasy free. This implies that it may not be a good idea to analyze gaps with parsimony as our results found that, given the alignments used, even within suborders gap characters were already homoplasious.

All the gap-based character matrices are far more congruent with the tree (Fig. 1) than would be expected for chance sets of splits of the same composition; the details for each method were as follows. The parsimony informative gap characters for the Clustal intron alignment have a parsimony score of 23 on the tree shown in Figure 1. In 1000 random draws of 16 splits of the same composition (i.e., seven 9:2 splits, six 8:3 splits, and three 7:4 splits), we found parsimony scores ranging from 29 to 40. The informative MUSCLE gap characters have a parsimony score of 19, 1000 random draws of splits with the same composition had a range of parsimony scores from 19 to 26 with only 24 out 1000 draws having a value ≤19. The informative T-Coffee gap characters have a parsimony score of 17, 1000 random draws of splits with the same composition had a range of parsimony scores from 19 to 26 with only 24 out 1000 draws having a value ≤17. The informative MAFFT (and SATé) gap characters have a parsimony score of 16, 1000 random draws of splits with the same composition had a range of parsimony scores from 16 to 25 with only 15 out 1000 draws having a value ≤16.

A more probabilistic way of using the information in gaps is suggested by the probabilities in Table 2 which, just using the Clustal alignment for an example, gives us $P \approx 1.29 \times 10^{-4}$ for the seven 9:2 splits, $\approx 3.15 \times 10^{-9}$ for the six 8:3 splits, and $\approx 4.15 \times 10^{-6}$ for the three 7:4 splits. If we assume that the splits, just as for the position of single nucleotide changes, are conditionally independent (given the tree), then we can combine these...
probabilities, giving only one chance in \(1.33 \times 10^{-17}\) that 13 randomly selected splits that had the same composition (i.e., seven 9:2 splits, six 8:3 splits, and three 7:4 splits) would fit the tree so well. This suggests a tree selection criterion somewhat analogous to maximum likelihood where you instead search for a tree that gives the minimum likelihood of producing the observed costs for gaps under a model where gaps were randomly allocated. However, this may be unrealistic as we show in Figure 2 that different alignment programs can insert, for example, either a single gap or two gaps. Therefore, we still need to be careful that a program that inserts two gaps (instead of a single gap) does not artificially deflate the probabilities. Nevertheless, the main conclusion of these last few paragraphs is that there is high information in data that is normally omitted. The difficulties inherent in incorporating gap information into a phylogenetic analysis postalignment reinforce the need for improved likelihood models which incorporate the indel process as well as sequence evolution (i.e., extensions to the TKF models) for handling this high information content.

We predict that if we are able to model indel information successfully, then the intron data should still be fairly good for the Order Charadriiformes but not when we get to the deeper Neoaves. An explanation would be that intron data will not be able to be aligned accurately for deeper divergences because there will be too many insertions and deletions. We could go to deeper levels and compare different alignments utility within Neoaves and between Neoaves, Galloanserae (chickens plus ducks) and Paleognaths (ratites and tinamous). Similarly, the experiment could be expanded to investigate shallower divergences within families also. Although indels within introns (both their length and positions) are certainly beyond their “use by date” for avian orders (Table 1), they still seem potentially highly informative for more recent divergences. We note that Löhne and Borsch (2005) also concluded that different classes of introns evolved at different rates. However, we still need to visualize contradictory signals in such data (Holland et al. 2005a).

**DISCUSSION AND CONCLUSIONS**

Our results suggest that further work must be done before we can use gaps to reliably extract phylogenetic information. The position of shared gaps is clearly highly phylogenetically informative, but at the same time, Figure 3 demonstrates that gaps should not be treated as rare genomic changes, indicating that a model-based approach may be necessary. Current models for indels are still unrealistic biologically and are on the edge of computational tractability, so this seems like no easy task. A positive approach is to accept that different data are appropriate at the various levels of divergence. Many unlinked markers are required for population studies, where hybridization, lineage sorting, and introgression may be occurring. Intron sequences (including gaps) should be helpful especially for closer divergences though we have not tried pushing that data back to the separation of avian orders within the Neoaves. Mitochondrial sequences should work further back but are still expected to saturate earlier than exon data and have the problem of bias in nucleotide frequencies (Lockhart et al. 1992). However, the potential of RY coding (see Pratt et al. 2009; Phillips et al. 2010), particularly for the third position of the triplet code, still needs more exploration. Deeper again exons should be outstanding, but we still need to be able to use much more of the information in the data, in the present study indels within exons seemed too few to be really useful, but that could certainly change when genomic data become routine—though again automating sequence alignment would then be essential.

Thus the question raised by Morrison (2009) about the potential need for automatic alignment is fundamentally important even if we are not yet able to rely on it. Automation is required already for very large data sets, and the need for it will only increase as data sets become genomic in size. The problems of automatic alignment just emphasize that there are different classes of data that could be used to give more confidence in phylogenies.

**SUPPLEMENTARY MATERIAL**

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository at [http://datadryad.org](http://datadryad.org), doi:10.5061/dryad.mg76th0n.

**FUNDING**

This work was supported by the Australian Research Council (FT100100031 to B.R.H.).

**ACKNOWLEDGMENTS**

K.S. and B.R.H. acknowledge the Allan Wilson Centre for supporting a summer scholarship. The authors thank the editorial team at Systematic Biology, David Morrison, and one anonymous reviewer for comments that greatly improved the manuscript.

**REFERENCES**


