*Syst. Biol.* 49(2):369–381, 2000

# Gaps as Characters in Sequence-Based Phylogenetic Analyses

MARK P. SIMMONS AND HELGA OCHOTERENA

*L.H. Bailey Hortorium, 462 Mann Library, Ithaca, New York 14853, USA,*
*E-mail: mps14@cornell.edu*

In the analysis of sequence-based data matrices, the use of different methods of treating gaps has been demonstrated to influence the resulting phylogenetic hypotheses (e.g., Eernisse and Kluge, 1993; Vogler and DeSalle, 1994; Simons and Mayden, 1997). Despite this influence, a well-justified, uniformly applied method of treating gaps is lacking in sequence-based phylogenetic studies. Treatment of gaps varies widely from secondarily mapping gaps onto the tree inferred from base characters to treating all gaps as separate characters or character states (González, 1996). This diversity of approaches demonstrates the need for a comprehensive discussion of indel (insertion or deletion) coding and a robust method with which to incorporate gap characters into tree searches. We use the term "indel coding" instead of "gap coding" because the term "gap coding" has already been applied to coding quantitative characters (Mickevich and Johnson, 1976; Archie, 1985). Although "indel coding" undesirably refers to processes that are not observed (insertions and deletions) instead of patterns that are observed (gaps), the term is unambiguous and does not co-opt established terminology. The purpose of this paper is to discuss the implications of each of the methods of treating gaps in phylogenetic analyses, to allow workers to make informed choices among them. We suggest that gaps should be coded as characters in phylogenetic analyses, and we propose two indel-coding methods. We discuss four main points: (1) the logical independence of alignment and tree search; (2) why gaps are properly coded as characters; (3) how gaps should be coded as characters; and (4) problems with a priori weighting of gap characters during tree search.

## LOGICAL INDEPENDENCE OF ALIGNMENT AND TREE SEARCH

Alignment and tree search have been cited as having a common goal and therefore as being dependent procedures (Mindell, 1991a). This common goal has led to the notion that the two processes are inseparable (Mindell, 1991b; Wheeler, 1994). The stance taken by Mindell and Wheeler is that alignment and tree search are so innately linked that alignment parameters should be used to weight characters in phylogenetic analyses.

Although alignment and phylogenetic analyses often have the same ultimate goal, we consider these procedures to be logically independent of one another. In alignment, hypotheses of primary homology are made (de Pinna, 1991:388). On the other hand, in phylogenetic analyses, hypotheses of primary homology are tested by congruence (Patterson, 1982) or by an evolutionary model (Felsenstein, 1973), resulting in the establishment of secondary hypotheses of homology (de Pinna, 1991) that will contain the maximum explanatory power in parsimony-based cladistic analyses (Farris, 1979). These two procedures (detection of similarity as a basis for primary homology, and tree construction to test hypotheses of primary homology) have not been linked in analyses of nonsequence data, and we see no reason why they should be linked in sequence-based analyses. Because the two procedures are not linked, alignment parameters need not be used to weight characters in phylogenetic analyses.

A second reason why it is consistent to use different cost functions in alignment and phylogenetic analyses (contra Wheeler, 1994) is that alignment parameters are gen-

erally selected by using a different criterion from that used to weight characters in phylogenetic analyses. Four of the criteria proposed to assign gap costs in alignment are as follows.

1. Gap costs may be set on the basis of the probability with which gaps are thought to occur (Gu and Li, 1995). Note two potential problems with this criterion: if gaps are thought to be more likely to occur than substitutions, or if gaps of any length are thought to be equally frequent, a trivial alignment in which bases from one sequence are aligned with gaps in all other sequences will result (Wheeler, 1994). If this criterion is used, only one alignment should be performed, with gap costs assigned on the basis of expected indel probabilities. On the other hand, if more than one alignment is performed, with different gap costs used, an alignment based on some other criterion is being sought (as in criteria 2–4).
2. Gap costs may be selected with the goal of maximizing congruence among independent sources of characters (Wheeler, 1995). For example, gap costs may be determined by congruence of the inferred gene tree produced when using those costs with the tree inferred from use of morphological characters (Wheeler, 1995; Whiting et al., 1997), biogeographical data, or paleontological data, or some combination of these (Wheeler et al., 1995).
3. Gap costs may be set to find conserved regions (Hershkovitz and Lewis, 1996).
4. Gap costs may be selected on the basis of the apparent quality of the alignment for the range of sequences sampled (Vogler and DeSalle, 1994).

In contrast with criteria 2, 3, and 4, a priori weighting in phylogenetic analyses (as is done when alignment parameters are incorporated in tree searches; e.g., Wheeler, 1990; Williams and Fitch, 1990; Mindell, 1991a; Knight and Mindell, 1993) can be justified only on the basis of how frequently characters are thought to have changed during the course of evolution. Therefore, only if gap costs are based on the frequency with which gaps are thought to occur (as per Gu and Li, 1995) is there a basis for use of identical cost functions in alignment and phylogenetic analyses.

## Gaps as Characters

### *Alignment*

The unmodified sequences determined by DNA or protein sequencing are not the same as the sequences used in phylogenetic analyses. Unmodified sequences determined by sequencing, to the extent that they are accurate, are direct representations of the actual organismal sequences. Sequences used in phylogenetic analyses, on the other hand, represent organismal sequences that have been adjusted on the basis of comparisons with one another through alignment. Before sequences have been aligned, insertions and deletions represent processes that are impossible to infer from the pattern observed in organismal sequences that are considered independently of one another. For this reason, Wheeler (1996:2) stated: "Nucleotide bases are observable, gaps are not. Hence a certain amount of logical inconsistency is introduced into the analysis since a process (insertion or deletion of bases) could be treated as a pattern (synapomorphy)." This "logical inconsistency" was the basis for Wheeler proposing his "optimization alignment" in which "indels appear not as states but as transformations linking ancestral and descendent nucleotide sequences" (Wheeler, 1996:2). Following Wheeler's point, the same "logical inconsistency" applies to any mutation (including transitions and transversions), because for homologous sequences that differ in length, comparable positions are not observable before sequences have been aligned. To produce comparable positions that provide the basis for the establishment of hypotheses of primary homology, alignment is required. If homologous sequences differ in length before alignment, gaps are required so as to produce comparable patterns for the entire sequence, not just the regions in which the gaps are inserted. Therefore, once the sequences have been aligned, comparisons or homology hypotheses apply to all positions, some of which may contain bases and gaps. That is, gaps have become part of the pattern as much as any nucleotide or amino

acid. The pattern used to code characters for phylogenetic analysis—and consequently the putative recognition of transitions, transversions, and indels in DNA sequences—is the one created by the alignment, not the unaligned pattern that occurs in organisms. Therefore, the "logical inconsistency" suggested by Wheeler (1996:2) does not exist.

After the sequences have been aligned, it is possible to recognize phylogenetically informative characters. In this context, we define gaps that may be treated as characters in phylogenetic analyses as follows: a single position or a contiguous set of positions for which no bases are present in one or more sequences, bounded on either side by aligned base(s) (nucleotides or amino acids). We deliberately exclude leading and trailing gaps, which are generally artifacts of aligning sequences with different 5′ and 3′ termini (e.g., alignment of sequences amplified by using different primers).

Putatively homologous gaps (as a statement of primary homology sensu de Pinna, 1991) are those with identical 5′ and 3′ termini. Gaps with different 5′ and/or 3′ termini are not treated as homologous because at least one indel event must be postulated to transform one gap into another. Therefore, although gaps with different 5′ and/or 3′ termini may result from sequential indel events, it is always equally or more parsimonious to treat the gaps as derived independently of one another.

One complication in the use of gaps as characters in tree searches is the problem of ambiguously aligned gaps (gaps that have more than one equally optimal alignment) that are potentially phylogenetically informative. Ambiguously aligned gaps (whether from a single set or multiple sets of alignment parameters, if a computer program is used) are not different from ambiguously aligned bases because both, as primary homology assessments, are the results of alignment. Gatesy et al. (1993:156) noted that "insertions/deletions may be unambiguous phylogenetic indicators in spite of alignment ambiguity." That is, a gap with more than one equally optimal alignment is not necessarily homoplastic. On the basis of this reasoning, Davis et al. (1998) treated ambiguously aligned gaps found in three taxa as homologous (as a hypothesis of primary homology) because the gaps could be aligned in the same positions in all three sequences.

### *Reliability of Gap Characters*

Because gaps are just as much a part of the aligned pattern as nucleotides are, gaps should be incorporated in the tree search along with base characters. Nevertheless, in recent literature, gaps have often been excluded in tree searches (e.g., Terry et al., 1997; Denton et al., 1998; Diaz-Lazcoz et al., 1998; Mason-Gamer et al., 1998). One reason that many workers do not include gaps in tree searches was stated by Johnson and Soltis (1995:167): "Because apparently identical indels may have multiple origins in unrelated taxa."

Golenberg et al. (1993:62,63) suggested that "superimposed indel mutations occurred more frequently than superimposed substitutions" in a noncoding region of the chloroplast genome, such that gap characters are less reliable than base characters. For coding gaps in their study, Golenberg et al. (1993:55) coded overlapping gap positions as homologous for gaps of different lengths. They considered this coding "conservative" whereas we consider this coding unjustified because there is no basis to assert that gaps that differ in 5′ and/or 3′ termini are homologous. The indel coding used by Golenberg et al. (1993) does not take into account the potential for indels with different termini to be used as evidence that the indels are not homologous (Lloyd and Calder, 1991). Therefore, superimposed indel mutations need not be homoplastic, contra Golenberg et al. (1993).

In contrast to Golenberg et al. (1993), some workers consider gaps to be better than substitution characters. Lloyd and Calder (1991:11) suggested that multiresidue gaps are reliable phylogenetic characters because it is unlikely that indels would be repeated in the exact same position, with the same length and sequence (for insertions); indels of different lengths at the same position are recognized as separate events. Although we agree with this assertion in general, we do not consider it appropriate for insertions to be hypothesized as homologous only if they have an identical sequence. This ignores the potential for

substitutions to occur after the inferred insertion event.

Van Dijk et al. (1999:94) suggested that "because indels are caused by more complex mutational mechanisms than base substitutions, homoplasy by parallel and back mutations—a plague of molecular phylogeny—is less likely to occur." It has even been suggested that only gaps should be analyzed and substitutions should be excluded from analysis (Lloyd and Calder, 1991). Entire studies have been devoted to using single gaps as phylogenetic characters (e.g., Bailey and Doyle, 1997; Lai et al., 1997).

Gaps have been found to be good characters (as measured by levels of homoplasy) in both coding and noncoding regions. Baldwin and Markos (1998) found 12 of 13 informative gaps to map on the substitution-based external-transcribed-spacer nrDNA (nuclear ribosomal DNA) gene tree without homoplasy (12 of the informative gaps were 1 base pair [bp] long). Prather and Jansen (1998) found 13 of 14 gaps to map onto ingroup taxa of the substitution-based internal-transcribed-spacer nrDNA gene tree without homoplasy. Lloyd and Calder (1991) found six of eight gaps >1 bp long to map on the substitution-based $\Psi\eta$-globin pseudogene gene tree without homoplasy. Each of the other two gaps had two parallel origins. Van Ham et al. (1994) found 33 of 34 gaps >2 bp long to map on the substitution-based *trnL-trnF* intergenic spacer gene tree without homoplasy. Johnson and Soltis (1995) found 15 of 16 gaps to map on the substitution-based *matK* gene tree without homoplasy. In none of these studies had the gaps been included as characters used to find the most-parsimonious tree(s) on which the gaps were mapped. Therefore, the phylogenetic signal from gap characters did not affect the tree topologies on which the gap characters were found to have low homoplasy.

### *Problems of a Priori Weighting of Gaps in Tree Search*

Because gaps are part of the aligned pattern that may be phylogenetically informative, to exclude gap characters in tree searches (i.e., to treat gap positions as missing values without scoring the gaps as additional characters, or to exclude aligned positions having gaps in some sequences) is to discard data. Gap costs used in alignment may be incorporated into tree searches by coding gaps as extra character states weighted by using a step matrix (Wheeler, 1994; Janies and Wheeler, 1998; Giribet and Wheeler, 1999).

Unless the gap cost in alignment is identical to all substitution costs, weighting gaps as fifth character states can be accomplished by using a step matrix in which the cost of a change from a base to a gap is equal to the gap cost used in alignment (when adjacent gap positions are treated independently of one another). This method treats adjacent gap positions separately from one another. This method also treats positions at which different gaps (gaps with different 5′ and/or 3′ termini) overlap in separate sequences as homologous to one another. For example, if gap 1 in sequence A is located between aligned positions 100 and 200, and gap 2 in sequence B is located between aligned positions 125 and 225, the gaps would be treated as homologous for positions 125 through 200, even though these are most-parsimoniously interpreted as two different gaps. If the gap:change cost is 2:1, in effect there exists a character supporting sequence A as being more closely related to sequence B than either sequence is related to sequences that have no gaps between positions 125 and 200 with a weight of 150. This results in a very highly weighted, probably homoplasious character.

### *Gaps as Fifth Character States versus Separate Characters*

Gaps may be coded as fifth character states for nucleotides (or 21st states for amino acid sequence data) or as separate presence/absence characters. However, these two treatments are not identical to one another in theory or in practice. To treat gaps that are one position long as fifth states, and gaps more than one position long are treated as separate presence/ absence characters (Barriel, 1994), is inconsistent.

A gap may be considered to represent an alternative condition to any base; that is, the presence of a gap does not contain any additional information compared with any

other state (i.e., base) at an aligned position. In this case, the gap should be treated as a fifth character state. Alternatively, the presence or absence of a gap could be taken into account as an additional source of information for tree searches. In this case, the gap is taken to represent a condition different from that of any base at any given aligned position.

Coding gaps as fifth character states is easily accomplished for one-position-long gaps. Treating gaps longer than one position (one nucleotide for DNA sequences, one amino acid for protein sequences) as fifth states for each position treats adjacent gap positions (putatively caused by a single indel event) as though they were independent of one another (Eernisse and Kluge, 1993). One method for coding gaps as fifth character states without treating adjacent gap positions independently of one another was used by Bena et al. (1998:554): "Single-site gaps were treated as a new state. For gaps longer than one nucleotide, we recoded the first site in the gap as a new state and coded all other sites 'missing data' in order for the gaps to be counted as a single event." This approach has two problems for gaps longer than one position. First, gaps with identical 5′ termini but with different 3′ termini are treated as homologous. Second, the gap is arbitrarily coded for a single position. If the distribution of bases among sequences without gaps at this position differs from the distribution of bases at other positions at which the gap occurs (as is generally the case), then different numbers of steps for any given tree topology may result. Therefore, the arbitrary decision regarding the position for which the gap is coded can partially determine which trees are most parsimonious.

To code gaps as separate characters, an extra presence/absence character for every gap is added to the data matrix. The corresponding aligned position(s) in the sequence where the gap is inferred is then coded as inapplicable. However, coding gaps as separate presence/absence characters can lead to problems caused by the introduction of missing values. Maddison (1993) discussed the potential artifact generated by optimization when, on the cladogram, terminals with missing values were intercalated between two or more groups for which informative character states were coded. He referred to this problem as "long-distance influence" or "leak" of characters. For sequence-based analyses, this optimization artifact occurs only when a paraphyletic group for which the gap is present separates two groups that share one or more bases, and when two or more bases are present in at least one of the groups.

When coding gaps as separate characters, an extra hypothesis of homology is made that is not coded for in gap-as-a-fifth-state coding. Coding gaps as separate characters calls for an explicit hypothesis that all sequences that have any base at a given position had a common ancestor with a base at that position. That is to say, all bases at a given position are treated as homologous to one another in the sense that a base is present. This extra hypothesis of homology restricts the number of equally parsimonious trees, compared with those obtained when coding gaps as fifth states. In other words, for a gap occupying one position, the fifth-state coding is more conservative and less informative.

Aligned positions may be used to code base characters (e.g., base at aligned position one, base at aligned position two) and gap characters (e.g., gap at aligned position three, gap from aligned positions six to ten). Aligned positions need not be directly interpreted as the characters themselves (e.g., state at position one, state at position two), as is generally done. Gaps are inapplicable for base characters, and the type of base (adenine, guanine, proline, or glutamic acid) is irrelevant for gap characters.

We assert that gaps, regardless of length, are appropriately coded as separate presence/absence characters, not as fifth character states for each base character (corresponding to each aligned position). Our basis for this is as follows. Characters should be independent of one another, and character states are alternative forms of a given character (Pogue and Mickevich, 1990). Gaps are an alternative form of an aligned position (or positions, when gaps are more than one position long), but they do not represent alternative forms of a base (nucleotide or amino acid). Organismal sequences contain only bases; gaps and aligned positions are inferred only after sequences have been aligned with one an-

other (Wheeler, 1996). If a base corresponding to an aligned position does not occur in a particular organismal sequence, there is nothing in that sequence that is comparable with, or homologous to, bases corresponding to that aligned position in other organismal sequences. Because there is nothing that can be compared, that aligned position is inapplicable for the sequences that lack the base. Therefore, it is incorrect to code a gap as a fifth state for the base character corresponding to that inapplicable aligned position.

When a deletion eliminates a base corresponding to an aligned position in a given organismal sequence, that aligned position is forever lost in that sequence. Even if an insertion occurs later at the same location, the bases that occur there should not be compared with bases in taxa that lack the deletion followed by the insertion in their lineage. The bases at the positions created by an insertion are de novo and are not homologous to any base in any other lineage without the insertion (in the sense of orthology; they may be paralogous to bases at other positions if the insertion was a duplication of these bases). Not only do organismal sequences that lack a base corresponding to an aligned position not have a comparable character state for the base character corresponding to that aligned position, they do not have *anything* corresponding to that aligned position at all. Gap characters are therefore fundamentally different types of characters than base characters are. This distinction is contradicted by any coding method that treats gaps as alternative states of base characters.

### *Contiguous Gap Positions*

The treatment of contiguous gap positions has been controversial. Some workers treat individual gap positions independently of adjacent gap positions and therefore treat them as separate characters (Wheeler, 1994; Janies and Wheeler, 1998; Wheeler and Hayashi, 1998; Giribet and Wheeler, 1999). When individual gap positions are treated independently of adjacent gap positions in alignment, contiguous gap positions are treated as if they are the result of independent indels (Eernisse and Kluge, 1993). Other workers treat contiguous gap positions as arising from single indels and therefore treat them as single characters (e.g., Lloyd and Calder, 1991, Barriel, 1994; van Dijk et al., 1999). When a gap-opening penalty and a gap-extension penalty that is less than the gap-opening penalty are used in alignment, the investigator is treating all contiguous gap positions as if they were due to a single indel (Giribet and Wheeler, 1999). Contiguous gap positions are generally recognized as caused by a single insertion or deletion, both in coding regions (Pascarella and Argos, 1992) and noncoding regions (Gu and Li, 1995).

Contiguous gap positions are most-parsimoniously interpreted as the result of a single event (indel). Contiguous gap positions can be considered as co-occurring patterns caused by single indel events, just as co-occurring DNA or protein sequences can be interpreted as being caused by single events. The presence of co-occurring base sequences constitutes the main source of evidence to postulate changes in sequences as caused by single events such as duplications (e.g., Baldwin and Markos, 1998), inversions (e.g., Golenberg et al., 1993), and transpositions (e.g., Nishio et al., 1995; Migheli et al., 1999). In contrast, sequences that are not co-occurring provide no indication that contradicts the possibility that the differences resulted from multiple events; these are therefore best treated as independent characters. The presence of co-occurring patterns as evidence for nonindependence is also a criterion for delimiting characters in morphological analyses. For example, five different petal-pubescence characters—one for each petal—would generally not be coded for a five-merous flower. Rather, a single character, "petal pubescence," would be coded if the same type of pubescence were found to occur on all five petals. Likewise, co-occurring sequences and contiguous gap positions are best treated as single characters, putatively representing single events. Therefore, contiguous gap positions should not be coded as multiple separate characters because this would imply independent events for each gap position, hence over-weighting the putative single indel event by an amount of the contiguous gap length minus one.

## CODING GAP CHARACTERS

We propose two methods for coding gaps as presence/absence characters in tree searches. One method, which we term "simple indel coding," is conservative and easy to implement. The second method, which we term "complex indel coding," is more complicated but allows the use of additional information by taking into account the minimum number of steps required for the transformation of one gap to another.

Simple indel coding is implemented by coding all gaps that have different 5′ and/or 3′ termini as separate presence/absence characters. Whenever gaps from different sequences may be a subset of other gaps, sequences that have these longer, completely overlapping gaps (i.e., extending to or beyond both the 5′ and 3′ termini of the gap being coded) are coded as inapplicable for the gap character being coded.

An example of simple indel coding is as follows (Fig. 1a). Gap 1 (from aligned positions 3 to 9) is present in sequences A and B. Gap 2 (from aligned positions 6 to 11) is present in sequences C and D. Gap 3 (from aligned positions 14 to 25) is present in sequences A and B. Gap 4 (from aligned positions 14 to 16) is present in sequences C and D. Gap 5 (from aligned positions 19 to 22) is present in sequences C and D. Sequence E has no gaps. In the data matrix, sequences C and D are coded as absent for gap 1, sequences A and B are coded as absent for gap 2, and sequences A and B are coded as inapplicable for gaps 4 and 5 (Fig. 1b). Using simple indel coding, the presence/absence characters for gaps 4 and 5 are uninformative for the five sequences in this example.

(a)

```
          1   5    10   15   20   25
          |   |    |    |    |    |
Seq. A    GG---①---CCTT----③-------GG
Seq. B    GG---①---CCTT----③-------GG
Seq. C    GGAAA--②---TT-④-AC-⑤--AAAGG
Seq. D    GGAAA--②---TT-④-AC-⑤--AAAGG
Seq. E    GGAAACCCCCCTTCAAACCCCAAAAGG
```

(b)

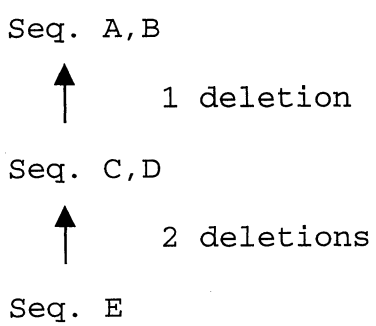
| | gap: | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Seq. A | | 1 | 0 | 1 | - | - |
| Seq. B | | 1 | 0 | 1 | - | - |
| Seq. C | | 0 | 1 | 0 | 1 | 1 |
| Seq. D | | 0 | 1 | 0 | 1 | 1 |
| Seq. E | | 0 | 0 | 0 | 0 | 0 |

FIGURE 1. Example of simple gap coding. (a) Alignment. Circled no. 1 = gap 1 from positions 3 to 9; circled no. 2 = gap 2 from positions 6 to 11; circled no. 3 = gap 3 from positions 14 to 25; circled no. 4 = gap 4 from positions 14 to 16; circled no. 5 = gap 5 from positions 19 to 22. (b) Gap characters in data matrix. "0" = gap absent, "1" = gap present, "-" = inapplicable.

The reason why sequences with longer, completely overlapping gaps are coded as inapplicable for the subset-gap character being coded is because it is impossible to infer absence of the subset gap in these sequences. In the above example, for sequences A and B that have gap 3, there is no way to infer whether they could also have gap 4 or gap 5. Gaps are absence of aligned bases. There is no way for a sequence to lose bases that it does not have to begin with. Because sequences A and B do not have aligned bases 14 to 25, there is no way they can lose a subset of these bases (bases 14 to 16 for gap 4, bases 19 to 22 for gap 5). Likewise, there is no way we can detect whether a subset of these bases was lost before the indel(s) responsible for the longer gap.

If sequences with longer, completely overlapping gaps were not coded as inapplicable for the subset gap, these longer, completely overlapping gaps would in effect be coded as homologous to bases in other sequences that do not have gaps in this region. Using the example above, suppose that sequence E is ancestral to sequence C, which is ancestral to sequence A (Fig. 2a). If the inapplicability rule is not used, the coding for absence/presence of gaps 3, 4, and 5 is as follows: sequence E 0,0,0; sequence C 0, 1,1; sequence A 1,0,0 (Fig. 2b). For gaps 3, 4, and 5 we recognize that the minimum number of steps from sequence E to sequence C is two (two deletions), and from sequence C to sequence A is one (a third "superimposed" deletion). However, this coding results in five steps: two steps from sequence E to sequence C, and three steps from sequence C to sequence A. The reason for this is that the 10-bp gap in sequence A is treated as homologous to the contiguous bases in sequence E for gap 4. However, there is no basis with which to make this homology assessment

(a)




(b)

| | gap: | 3 | 4 | 5 |
|---|---|---|---|---|
| Seq. | A,B | 1 | 0 | 0 |
| Seq. | C,D | 0 | 1 | 1 |
| Seq. | E | 0 | 0 | 0 |

(c)

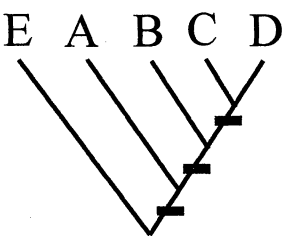



FIGURE 2. The example shown, using alignment from Figure 1a for gaps 3–5, demonstrates the necessity of inapplicable coding. (a) Evolutionary history of sequences A, C, and E. (b) Gap characters 3, 4, and 5 in data matrix without coding "Seq. A" as inapplicable for gaps 4 and 5. (c) Three alternative optimizations of gap character 3 on a given tree when sequences with longer, completely overlapping gaps are coded as inapplicable for the subset gap.

when one cannot detect the presence or absence of gap 4 in taxon A in the first place.

Although simple indel coding is easy to implement and incorporates gap characters into the data matrix, it does not utilize all available information and therefore is less informative than it could be because it can imply fewer steps than are biologically possible (given a correct alignment). For instance, in the example from Figure 1, it costs one step to change from a sequence with gap 3 to a sequence with gap 5; actually, however, it would require at least two evolutionary transformations (two insertions, or one insertion followed by one deletion). Finally, this method adds missing data (in the form of cells scored as inapplicable) to the data matrix, which can result in ambiguous optimizations. In the example from Figure 2a, b, there is one step in gap character 4 between sequence E and sequences C and D. However, if sequences A and B are a paraphyletic group, that is "intermediate" between sequence E and sequences C and D, there are three alternative optimizations of this character-state change (Fig. 2c).

Complex indel coding is implemented by coding all gaps that have different 5′ and/or 3′ termini as presence/absence characters. To do this, we follow six rules for coding, as described in the following text and Table 1. The basis for each of these rules and examples are then given. Different gaps may be treated in a single character to account for the possibility that they represent sequential evolutionary events for which coding these gaps as separate presence/absence characters would result in illogical homology assessments or an artificial increase or decrease in the minimum number of steps to change from one sequence to another.

*Rule 1.*—Gaps that partially overlap (the aligned region spanned by one gap is not entirely a subset of the aligned region spanned by any other gap) but share neither a common 5′ nor 3′ terminus should be coded as separate characters. This rule can be used to treat gaps in regions where many gaps of different sizes and degree of overlap occur (and then apply rules 2 through 6). Figure 3a presents an example illustrating the application of rule 1. Gap 1 (from aligned positions 3 to 6) is present in sequence A. Gap 2 (from aligned positions 3 to 14) is present in sequence B. Gap 3 (from aligned positions 10 to 14) is present in sequence C. Gap 4 (from aligned positions 12 to 20) is present in sequence D. Gap 5 (from aligned positions 18 to 20) is present in sequence E. Applying rule 1, we should treat gaps 1, 2, and 3, separately from gaps 4 and 5, even though gaps 2 and 3 partially overlap with gap 4, because these two groups

TABLE 1. Implementation of complex indel coding following six rules.

| Gaps to be coded | How gaps are coded |
|---|---|
| Gaps partially overlap but share neither a common 5′ nor 3′ terminus. | Rule 1. Code gaps as separate characters. |
| Gaps overlap and share a common 5′ or 3′ terminus. | Rule 2. Code gaps as a single unordered multistate character (but see rule 3 for implementation of step matrices). |
| One gap shares a common 5′ terminus with another gap and a common 3′ terminus with a third gap. | Rule 3. Code all three gaps as a single character in a symmetrical step matrix (two steps to change from gaps that share neither a common 5′ or 3′ end, one step for every other change). |
| Gaps have different 5′ and 3′ termini and one gap is entirely a subset of another gap. | Rule 4. Code gaps as a single character in an asymmetrical step matrix (two steps to change from the longer gap to the shorter gap but one step to change from the shorter gap to the longer gap). |
| Gaps have different 5′ and 3′ termini and more than one gap in different sequences is entirely a subset of another gap. | Rule 5. Code gaps as a single character in an asymmetrical step matrix (as in 4, and two steps to change from one subset gap to another subset gap). |
| Gaps have different 5′ and 3′ termini and more than one gap in the same sequence is entirely a subset of another gap. | Rule 6. Code gaps as a single character in an asymmetrical step matrix (as in rules 4 and 5). All subset gaps in each sequence are coded together as a separate character state in the step matrix, and the minimum number of steps between this character state and all other character states is determined and coded. |

of gaps do not share a common 5′ or 3′ terminus.

The only reason to include gaps of different lengths into single characters is to account for potential "nesting" of the gaps. However, if the gaps have no common 5′ or 3′ terminus and only partially overlap, potential nesting is irrelevant. It is equally parsimonious (two steps) to lose one gap and then gain the other gap, as it is to change the length of one end of the gap and then change the length of the other end of the gap. Hence, these two steps may be accounted for by two separate presence/absence characters. Furthermore, one can detect the absence of either gap in any given sequence. Therefore, both gap characters can be coded for all sequences.

*Rule 2.*—Gaps that overlap and share a common 5′ or 3′ terminus (such that the region spanned by one gap is a subset of the region spanned by another gap) should be coded as a single unordered multistate character (but see rule 3 for implementation of step matrices). Based on this rule, gaps 1, 2, and 3 in the example from Figure 3a should be treated as one unordered, multistate character, and gaps 4 and 5 should be treated as another unordered, multistate character. If gaps 4 and 5 were coded as two separate characters, an illogical homology assessment and an extra step would be inferred if the larger gap was derived from the smaller gap (as detailed above in the discussion of simple indel coding). Alternatively, if both gaps were coded as a single presence/absence character (coding the two gaps as homologous), then loss of information is guaranteed—loss of at least one step, and would be made an unsupported homology assessment. However, both of these problems are avoided by coding both gaps in a single multistate, unordered character.

*Rule 3.*—When one gap shares a common 5′ terminus with another gap and a common 3′ terminus with a third gap, all three gaps should be coded as a single character for which a symmetrical step matrix is implemented (such that two steps are required to change from gaps that share neither a common 5′ or 3′ end with one another, and one step is required for every other change). Based on this rule, gaps 1, 2, and 3 in the example from Figure 3a should be treated as one character in a step matrix. The minimum number of steps to change from sequence A to sequence C is two. The minimum number of steps to change from sequence B to either sequence A or C is one. If the gaps were coded as a single character without the step matrix (Fig. 3b), only one step would be required to change from gap 1 to gap 3, which would underestimate the

(a)

```
          1   5     10    15    20
          |   |     |     |     |
Seq. A    TT-①--GGGAAAACCTTTGGCC
Seq. B    TT----②-------CTTTGGCC
Seq. C    TTAAAGGGG--③--CTTTGGCC
Seq. D    TTAAAGGGGAA---④-----CC
Seq. E    TTAAAGGGGAAAACCTT-⑤-CC
```

(b)

| | absent | gap 1 | gap 2 | gap 3 |
|---|---|---|---|---|
| Absent | - | 1 | 1 | 1 |
| gap 1 | 1 | - | 1 | 2 |
| gap 2 | 1 | 1 | - | 1 |
| gap 3 | 1 | 2 | 1 | - |

FIGURE 3. Example of complex gap coding. (a) Alignment. Circled no. 1 = gap 1 from positions 3 to 6; circled no. 2 = gap 2 from positions 3 to 14; circled no. 3 = gap 3 from positions 10 to 14; circled no. 4 = gap 4 from positions 12 to 20; circled no. 5 = gap 5 from positions 18 to 20. Based on rule 1, gaps 1, 2, and 3, are treated as one character and gaps 4 and 5 should be treated as a second character. Based on rule 2, gaps 1, 2, and 3 should be treated as one unordered, multistate character, and gaps 4 and 5 should be treated as another unordered, multistate character. (b) Step matrix (direction of change is from row to column) for gap character that codes for gaps 1, 2, and 3 applying rule 3.

minimum number of steps. This example can be extended to address multiple gaps with common termini.

*Rule 4.*—When gaps have different 5′ and 3′ termini and one gap is entirely a subset of another gap, these gaps should be coded as a single character for which an asymmetrical step matrix is then implemented (such that two steps are required to change from the longer gap to the shorter gap but one step is required to change from the shorter gap to the longer gap). An asymmetrical step matrix is required because the minimum number of steps to change from a longer gap to a shorter gap (in which both gaps have different 5′ and 3′ termini) is two (one insertion on either end or one long insertion followed by one short deletion), whereas the minimum number of steps to change from a shorter gap to a longer gap is one (one superimposed deletion that extends beyond the original gap at both 5′ and 3′ termini).

*Rule 5.*—When gaps have different 5′ and 3′ termini and more than one gap in different sequences is entirely a subset of another gap, all of these gaps should be coded as a single character for which an asymmetrical step matrix is implemented (as in rule 4, in addition to the two steps to change from one subset gap to another subset gap).

*Rule 6.*—Finally, when gaps have different 5′ and 3′ termini and more than one gap in the same sequence is entirely a subset of another gap, all of these gaps should be coded as a single character for which an asymmetrical step matrix is implemented (as in rules 4 and 5). In addition, all subset gaps in each sequence are treated together as a separate character state in the step matrix, and the minimum number of steps between this character state and all other character states is determined and coded. If rules 4, 5, and 6 are all applicable for any given aligned region, a single asymmetrical step matrix in which these rules are applied is implemented.

Figure 4a illustrates the usage of rules 4, 5, and 6. No gap is present in sequence A. Gap 1 (from aligned positions 6 to 8) is present in sequence B and sequence E. Gap 2 (from aligned positions 14 to 16) is present in sequence C and sequence E. Gap 3 (from aligned positions 3 to 19) is present in sequence D. Based on rules 4 and 5, gaps 1, 2, and 3 should be treated as one character in an asymmetrical step matrix (Fig. 4b). Based on rule 6, the co-occurrence of gaps 1 and 2 in sequence E should be treated as a separate character state (Fig. 4b). This example can be extended to address multiple such subset gaps and does not violate the triangle inequality (Farris, 1981).

It might be suggested that the two completely separate (no overlap and no common 5′ or 3′ termini) gaps in sequence E from the example in Figure 4a are being treated nonindependently of one another. However, this apparent lack of independence is irrelevant because the gaps co-occur in the observed pattern of the aligned sequences. This aligned pattern, in which the gaps co-occur, is all that needs to be coded for. Using this coding method, one does not assert that these gaps always co-occur in nature.

(a)

```
          1   5    10   15   20
          |   |    |    |    |
Seq. A    TTAAATTTCCCCCTTTCCCGG
Seq. B    TTAAA-①-CCCCCTTTCCCGG
Seq. C    TTAAATTTCCCCC-②-CCCGG
Seq. D    TT-------③---------GG
Seq. E    TTAAA-①-CCCCC-②-CCCGG
```

(b)

| | absent | gap 1 | gap 2 | gap 3 | gaps 1,2 |
|---|---|---|---|---|---|
| Absent | - | 1 | 1 | 1 | 2 |
| gap 1 | 1 | - | 2 | 1 | 1 |
| gap 2 | 1 | 2 | - | 1 | 1 |
| gap 3 | 1 | 2 | 2 | - | 3 |
| gaps 1,2 | 2 | 1 | 1 | 1 | - |

FIGURE 4. Example of complex gap coding. (a) Alignment. Circled no. 1 = gap 1 from positions 6 to 8; circled no. 2 = gap 2 from positions 14 to 16; circled no. 3 = gap 3 from positions 3 to 19. Based on rules 4 and 5, gaps 1, 2, and 3 should be treated as one character in an asymmetrical step matrix. Based on rule 6, the co-occurrence of gaps 1 and 2 in sequence E should be as a separate character state. (b) Step matrix (direction of change is from row to column) for gap character that codes for gaps 1, 2, and 3 applying rules 4, 5, and 6.

## CONCLUSIONS

Alignment (making hypotheses of primary homology) and tree searches (testing hypotheses of primary homology) are logically independent steps in phylogenetic analysis. Although both approaches may be incorporated into a single step (Wheeler, 1996), they need not be. Although gaps do not occur in organisms and therefore cannot be directly observed, gaps are as much a part of the pattern of aligned sequences as bases are. Because this aligned pattern is used to code characters for tree searches, the informative variation from gaps should be incorporated along with base characters into tree searches. We assert that gaps are properly coded as separate presence/absence characters (not as fifth character states for nucleotides or 21st character states for amino acids) because although gaps are an alternative form of an aligned position (or positions), they do not represent alternative forms of bases. In addition to evidence that contiguous gap positions originate as single indel events, the parsimony criterion favors the interpretation of coding contiguous gap positions as single characters because of the co-occurring pattern.

Two methods are proposed by which gaps coded as characters can be implemented in tree searches. Simple indel coding is easy to implement but does not utilize all available information and can cause ambiguous optimizations of gap characters. Complex indel coding is more difficult to implement but allows all available information to be utilized. Simple and complex indel coding are justified on both theoretical and methodological bases.

## ACKNOWLEDGMENTS

We thank Donovan Bailey, Jerry Davis, Jeff Doyle, Melissa Luckow, Kevin Nixon, and participants of the Doyle Lab Group for reviewing the manuscript and for helpful discussions. We also thank Richard Olmstead, Brengt Oxelman, and an anonymous reviewer for their constructive criticisms.

## REFERENCES

ARCHIE, J. W. 1985. Methods for coding variable morphological features for numerical taxonomic analysis. Syst. Zool. 34:326–345.

BAILEY, C. D., AND J. J. DOYLE. 1997. The chloroplast *rpl2* intron and ORF184 as phylogenetic markers in the legume tribe Desmodieae. Syst. Bot. 22:133–138.

BALDWIN, B. G., AND S. MARKOS. 1998. Phylogenetic utility of the external transcribed spacer (ETS) of 18S-26S rDNA: Congruence of ETS and ITS trees of *Calycadenia* (Compositae). Mol. Phylogenet. Evol. 10:449–463.

BARRIEL, V. 1994. Molecular phylogenies and how to code insertion/deletion events. Life Sci. 317:693–701.

BENA, G., J.-M. PROSPER, B. LEJEUNE, AND I. OLIVIERI. 1998. Evolution of annual species of the genus *Medicago:* A molecular phylogenetic approach. Mol. Phylogenet. Evol. 9:552–559.

DAVIS, J. I., M. P. SIMMONS, D. W. STEVENSON, AND J. F. WENDEL. 1998. Data decisiveness, data quality, and incongruence in phylogenetic analysis: An example from the monocotyledons using mitochondrial *atp*A sequences. Syst. Biol. 47:282–310.

DENTON, A. L., B. L. MCCONAUGHY, AND B. D. HALL. 1998. Usefulness of RNA polymerase II coding sequences for estimation of green plant phylogeny. Mol. Biol. Evol. 15:1082–1085.

DE PINNA, M. C. C. 1991. Concepts and tests of homology in the cladistic paradigm. Cladistics 7: 367–394.

DIAZ-LAZCOZ, Y., J.-C. AUDE, P. NITSCHKÉ, H. CHIAPELLO, C. LANDÈS-DEVAUCHELLE, AND J.-L. RISLER. 1998. Evolution of genes, evolution of species: The case of aminoacyl-tRNA synthetases. Mol. Biol. Evol. 15:1548–1561.

EERNISSE, D. J., AND A. G. KLUGE. 1993. Taxonomic congruence versus total evidence, and Amniote phylogeny inferred from fossils, molecules, and morphology. Mol. Biol. Evol. 10:1170–1195.

FARRIS, J. S. 1979. The information content of the phylogenetic system. Syst. Zool. 28:483–519.

FARRIS, J. S. 1981. Distance data in phylogenetic analysis. Pages 3–22 *in* Advances in cladistics: Proceedings of the first meeting of the Willi Hennig Society (V. A. Funk and D. R. Brooks, eds.). New York Botanical Garden, New York.

FELSENSTEIN, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22:240–249.

GATESY, J., R. DESALLE, AND W. WHEELER. 1993. Alignment-ambiguous nucleotide sites and the exclusion of systematic data. Mol. Phylogenet. Evol. 2:152–157.

GIRIBET, G., AND W. C. WHEELER. 1999. On gaps. Mol. Phylogenet. Evol. 13:132–143.

GOLENBERG, E. M., M. T. CLEGG, M. L. DURBIN, J. DOEBLEY, AND D. P. MA. 1993. Evolution of a noncoding region of the chloroplast genome. Mol. Phylogenet. Evol. 2:52–64.

GONZÁLEZ, D. 1996. Codificación de las inserciones-deleciones en el análisis filogenético de secuencias génicas. Bol. Soc. Bot. Mex. 59:115–129.

GU, X., AND W.-H. LI. 1995. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. J. Mol. Evol. 40:464–473.

HERSHKOVITZ, M. A., AND L. A. LEWIS. 1996. Deep-level diagnostic value of the rDNA-ITS region. Mol. Biol. Evol. 13:1276–1295.

JANIES, D., AND W. WHEELER. 1998. MALIGN version 2.7 (documentation). Retrieved April 2, 1999 by ftp from ftp://ftp.amnh.org/pub/people/wheeler/malign/malign.txt.

JOHNSON, L. A., AND D. E. SOLTIS. 1995. Phylogenetic inference in Saxifragaceae sensu stricto and *Gilia* (Polemoniaceae) using *mat*K sequences. Ann. Mo. Bot. Gard. 82:149–175.

KNIGHT, A., AND D. P. MINDELL. 1993. Substitution bias, weighting of DNA sequence evolution, and the phylogenetic position of Fea's viper. Syst. Biol. 42: 18–31.

LAI, M., J. SCEPPA, J. A. BALLENGER, AND J. J. DOYLE. 1997. Polymorphism for the presence of the *rpL2* intron in chloroplast genomes of *Bauhinia* (Leguminosae). Syst. Bot. 22:519–528.

LLOYD, D. G., AND V. L. CALDER. 1991. Multi-residue gaps, a class of molecular characters with exceptional reliability for phylogenetic analyses. J. Evol. Biol. 4:9–21.

MADDISON, W. P. 1993. Missing data versus missing characters in phylogenetic analysis. Syst. Biol. 42: 576–581.

MASON-GAMER, R. J., C. F. WEIL, AND E. A. KELLOGG. 1998. Granule-bound starch synthase: Structure, function, and phylogenetic utility. Mol. Biol. Evol. 15:1658–1673.

MICKEVICH, M. F., AND M. S. JOHNSON. 1976. Congruence between morphological and allozyme data in evolutionary inference and character evolution. Syst. Zool. 25:260–270.

MIGHELI, Q., R. LAUGÉ, J.-M. DAVIÈRE, C. GERLINGER, F. KAPER, T. LANGIN, AND M.-J. DABOUSSI. 1999. Transposition of the autonomous *Fot1* element in the filamentous fungus *Fusarium oxysporum*. Genetics 151:1005–1013.

MINDELL, D. P. 1991a. Aligning DNA sequences: Homology and phylogenetic weighting. Pages 73–89 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford University Press, New York.

MINDELL, D. P. 1991b. Similarity and congruence as criteria for molecular homology. Mol. Biol. Evol. 8:897–900.

NISHIO, H., P. E. M. GIBBS, P. P. MINGHETTI, R. ZIELINSKI, AND A. DUGAICZYK. 1995. The chimpansee α-fetoprotein-encoding gene shows structural similarity to that of gorilla but distinct differences from that of human. Gene 162:213–220.

PASCARELLA, S., AND P. ARGOS. 1992. Analysis of insertions/deletions in protein structures. J. Mol. Biol. 224:461–471.

PATTERSON, C. 1982. Morphological characters and homology. Pages 21–74 *in* Problems of phylogenetic reconstruction (K. A. Joysey and A. E. Friday, eds.). Academic Press, London.

POGUE, M. G., AND M. F. MICKEVICH. 1990. Character definitions and character state delineation: The bête noire of phylogenetic inference. Cladistics 6:319–361.

PRATHER, L. A., AND R. K. JANSEN. 1998. Phylogeny of *Cobaea* (Polemoniaceae) based on sequence data from the ITS region of nuclear ribosomal DNA. Syst. Bot. 23:57–72.

SIMONS, A. M., AND R. L. MAYDEN. 1997. Phylogenetic relationships of the creek chubs and the spine-fins: An enigmatic group of North American cyprinid fishes (Actinopterygii: Cyprinidae). Cladistics 13:187–206.

TERRY, R. G., G. K. BROWN, AND R. G. OLMSTEAD. 1997. Phylogenetic relationships in subfamily Tillandsioideae (Bromeliaceae) using *ndhF* sequences. Syst. Bot. 22:333–346.

VAN DIJK, M. A. M., E. PARADIS, F. CATZEFLIS, AND W. W. DE JONG. 1999. The virtues of gaps: *Xenarthran* (Edentate) monophyly supported by a unique deletion in aA-crystallin. Syst. Biol. 48:94–106.

VAN HAM, R. C. H. J., H. HART, T. H. M. MES, AND J. M. SANDBRINK. 1994. Molecular evolution of noncoding regions of the chloroplast genome in the Crassulaceae and related species. Curr. Genet. 25:558–566.

VOGLER, A. P., AND R. DESALLE. 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 12:1–10.

WHEELER, W. C. 1990. Combinatorial weights in phylogenetic analysis: A statistical parsimony procedure. Cladistics 6:269–275.

WHEELER, W. C. 1994. Sources of ambiguity in nucleic acid sequence alignment. Pages 323–352 *in* Molecular ecology and evolution: Approaches and applications (B. Schierwater, G. P. Wagner, and R. DeSalle, eds.). Birkhäuser Verlag, Basel.

WHEELER, W. C. 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. Syst. Biol. 44:321–331.

WHEELER, W. C., J. GATESY, AND R. DESALLE. 1995. Elision: A method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. Mol. Phylogenet. Evol. 4:1–9.

WHEELER, W. C., AND C. Y. HAYASHI. 1998. The phylogeny of the extant chelicerate orders. Cladistics 14:173–192.

WHITING, M. F., J. C. CARPENTER, Q. D. WHEELER, AND W. C. WHEELER. 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.

WILLIAMS, P. L., AND W. M. FITCH. 1990. Phylogeny determination using dynamically weighted parsimony method. *In* Molecular evolution: Computer analysis of protein and nucleic acid sequences (R. F. Doolittle, ed.). Methods Enzymol. 183:615–627.

*Received 28 June 1999; accepted 10 September 1999*
*Associate Editor: R. Olmstead*