Molecular Phylogenetics and Evolution
Vol. 16, No. 3, September, pp. 317–330, 2000
doi:10.1006/mpev.2000.0785, available online at http://www.idealibrary.com on IDEAL®

# REVIEW

# Multiple Sequence Alignment in Phylogenetic Analysis

Aloysius Phillips,[1] Daniel Janies, and Ward Wheeler

*Department of Invertebrates, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024-5192*

Received September 28, 1999; revised January 27, 2000

**Multiple sequence alignment is discussed in light of homology assessments in phylogenetic research. Pairwise and multiple alignment methods are reviewed as exact and heuristic procedures. Since the object of alignment is to create the most efficient statement of initial homology, methods that minimize nonhomology are to be favored. Therefore, among all possible alignments, the one that satisfies the phylogenetic optimality criterion the best should be considered the best alignment. Since all homology statements are subject to testing and explanation this way, consistency of optimality criteria is desirable. This consistency is based on the treatment of alignment gaps as character information and the consistent use of a cost function (e.g., insertion–deletion, transversion, and transition) through analysis from alignment to phylogeny reconstruction. Cost functions are not subject to testing via inspection; hence the assumptions they make should be examined by varying the assumed values in a sensitivity analysis context to test for the robustness of results. Agreement among data may be used to choose an optimal solution set from all of those examined through parameter variation. This idea of consistency between assumption and analysis through alignment and cladogram reconstruction is not limited to parsimony analysis and could and should be applied to other forms of analysis such as maximum likelihood.** © 2000 Academic Press

## INTRODUCTION AND BACKGROUND

Like all things phylogenetic, DNA sequence alignment has sparked debate about the proper methodology with which to analyze these data. Sequence information has become a fundamental tool of not just systematic evolutionary research but also of ecology, bioconservation, disease control, viral origins, and even HIV demographics and the legal intricacies of transmission. To date, most of the focus in phylogenetics has been placed on cladogram construction. However, all analyses of relationships derived from sequence data are fundamentally based upon alignment. Morrison and Ellis (1997) have recently examined the effects of different sequence alignment methods on phylogenetic topology. They conclude that variation in the resulting phylogeny is more dependent on the mode of alignment than on the method of phylogenetic reconstruction. This is not surprising, since the data being analyzed are not simply "theory neutral" observations, but the outcome of the alignment process.

Here, we discuss the basic methodology of pairwise sequence alignment and its extensions to multiple sequence alignment with respect to homology assessment and phylogenetic analysis. We also discuss several issues that arise from the interdependence of sequence alignment and phylogenetic reconstruction.

## HYPOTHESIS TESTING AND PRIMARY HOMOLOGY

The initial step in any phylogenetic analysis is to establish provisional (putative or primary) homology statements across taxa. Molecular sequence alignment is, in essence, a procedure by which we can recognize and describe potential homology among nucleotide or amino acid positions. Multiple sequence alignment algorithms create potential homologies (in the form of columns of bases in the data matrix). Primary homology (sensu dePinna, 1991) or topographic identity (sensu Brower and Scharawoch, 1996) is generally established through the computation of a pairwise similarity cost function. These putative homologies are then subjected to some form of phylogenetic analysis.

In a parsimony framework a logical means of assessing the quality of homology statements is cladistic character congruence (Kluge, 1989). Character congruence argues that among all competing hypotheses, the ones that are defended by the greatest number of independent congruent characters are the best sup-

[1] Current address: Department of Biological Sciences, Columbia University, New York, NY 10027.

1055-7903/00 $35.00

ported. The degree of character congruence in any data set is based on its phylogenetic topology. Logically then, the most parsimonious cladogram resulting from an alignment should be derived from the same set of assumptions that were used to generate that alignment. If not, the data set is generated under one set of assumptions and analyzed under another set of assumptions. Within this paradigm the best alignment is that which yields the most parsimonious cladogram. The hypothesis, which satisfies Occam's razor, requires the fewest ad hoc hypotheses (homoplasies); therefore, the alignment(s) that yield the most parsimonious cladogram(s) best satisfies our desire to maximize homology.

The *sine quae non* of sequence alignment are gaps. When sequences differ in length, insertion–deletion events (indels) are postulated as required to explain the variation and their places held by gap characters. Operationally, if a cost is not assigned to the insertion of gaps during alignment a trivial alignment will result where both sequences will have gaps at each position were there is a potential mismatch with an alignment cost of zero. During cladogram construction, insertion–deletion events are frequently (if implicitly) assigned a cost of zero (Swofford and Olsen, 1990; Giribet and Wheeler, 1999). The operation used to insert gaps during alignment should also be reflected in the phylogenetic analysis of that alignment. Gaps should therefore be treated as a 5th character state in nucleotide data sets (or a 21st state in amino acid data sets) with the cost of transformation between a gap and the other states in the cladogram determined by the assumptions of the alignment process. If a gap cost (or penalty) of "two" is assigned during alignment, indels (insertions or deletions) should also cost "two" in judging phylogenetic trees based on that alignment.

Alignment and phylogenetic analysis, no matter which algorithms or optimality criteria are used, are sensitive to the choice of cost functions (Fitch and Smith, 1983). Various weights must be assigned a priori to alignment parameters (assumptions) such as nucleotide mismatch cost (including any transition–transversion bias) and gap cost. Since homology assessments are sensitive to parameter variation the outcome of the phylogenetic analysis is dependent upon these values. Using sensitivity analysis (Wheeler, 1995), the effect of one's choice of parameter values can be explored by examining many cost function combinations. The numerous results produced by different parameter sets can be assessed via character congruence to assay their ramifications for homology.

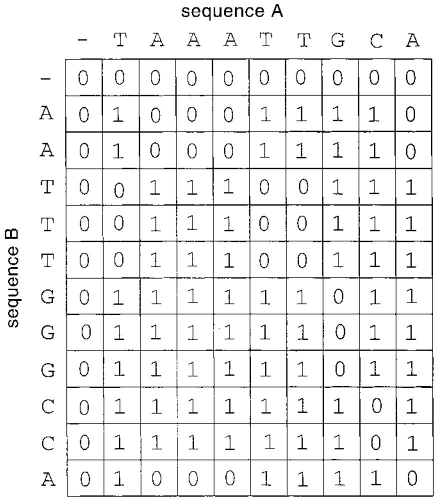
## PAIRWISE ALIGNMENT

The initial step in nearly all methods of sequence analysis is pairwise alignment. Most sequence alignment methods seek to optimize the criterion of similarity. There are two modes of assessing this similarity, local and global. Local methods try to determine if subsegments of one sequence (A) are present in another (B). These methods have their greatest utility in data base searching and retrieval (e.g., BLAST, Altschul *et al.*, 1990). Although they may be of utility in detecting sequences with a certain degree of similarity that may or may not be homologous, in phylogenetic analysis it is assumed that the sequences being compared are orthologous. Global methods make comparisons over the entire lengths of the sequences; in other words, each element of sequence A is compared with each element in sequence B. Global comparison is the principal method of alignment for phylogenetic analysis.

The crux of similarity maximization is the calculation of the minimum edit distance between two sequences. The edit distance is the number of operations (substitutions, insertions, or deletions) required to convert sequence A into sequence B. Each operation must be given a cost (or penalty). The aggregate optimal cost will be a measure that reflects the similarity of the two sequences. The fundamental method of pairwise sequence alignment was first described by Needleman and Wunsch (1970). The Needleman–Wunsch (N-W) method was initially intended for proteins but applies to any pairwise edit distance problem. The procedure seeks to maximize a similarity measure between two sequences. Smith *et al.* (1981) have shown that the Sellers' metric (1974) which minimizes a distance metric is equivalent to the N-W algorithm. These algorithms are an example of dynamic programming (Bellman, 1957) which permits a larger problem to be resolved by solving smaller subproblems recursively and assembling them into a final global result.

The N-W method can be thought of as proceeding through four basic steps: laying out the alignment matrix; initializing the matrix; "wave front" updating the matrix elements; and the trace back. When laying out the matrix, the two sequences define the axes of a two-dimensional array (Fig. 1). As an example we will consider sequence A "TAAATTGCA" and sequence B "AATTTGGGCCA." The top left-hand corner and bottom right-hand corner of the matrix correspond to the 5′ and 3′ end of the sequences, respectively. To allow for leading gaps, the first cell of the matrix (0, 0) is a null cell where column 1 refers to the first base of sequence A and row 1 corresponds to the first base of sequence B. Hence, in this case we have a matrix that has 120 elements.

In order to initialize the matrix elements, the N-W algorithm starts from the beginning of both sequences (top left corner) and traces its way through the matrix to the end of both sequences (bottom right), logging a mismatch value to each cell (Fig. 1). In the simplest scheme of distance minimization, a specified value (for instance 0) is placed in a cell whenever there is a


sequence A
sequence B

|  | - | T | A | A | A | T | T | G | C | A |
|---|---|---|---|---|---|---|---|---|---|---|
| - | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| A | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| T | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| T | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| T | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| G | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| G | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| G | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| C | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| C | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| A | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |

**FIG. 1.** An initialized matrix of a pairwise nucleotide sequence comparison with an assigned mismatch cost of 1.

matching state between sequences, regardless of position. All nonidentical states are assigned a value of one. More elaborate schemes of mismatch-scoring functions can be instituted by referring to a predefined mismatch-scoring matrix.

During the wavefront update of the matrix, each cell in the matrix is assigned a new value (Fig. 2a–2d). This value results from a comparison of three neighbors of matrix cell ($i$, $j$): the cell to the immediate left ($i$, $j$ − 11), directly above ($i$ − 1, $j$), and diagonal above and to the left ($i$ − 1, $j$ − 1). A diagonal path implies a correspondence between sequence elements whether there is a match or a mismatch. A gap is inserted into the alignment by moving across a row or down a column. Gaps are assigned a cost (>0) or a trivial alignment will be generated with a gap at every potential mismatch (total score = 0). The new value of cell ($i$, $j$) will be the minimum path cost of the three possible routes from its neighbors (Fig. 2). The value of these routes is calculated by adding the cell value of the previous cell ($i$, $j$ − 1; $i$ − 1, $j$; or $i$ − 1, $j$ − 1) plus the gap cost in the case of a gap or the mismatch cost in the case of a correspondence. When a gap is instituted the mismatch costs are ignored since no base correspondence is implied:

$$\begin{aligned} D(i, j) = \min\{&d(i - 1, j - 1) + \text{mismatch cost}, \\ &d(i - 1, j) + \text{gap cost}, \\ &d(i, j - 1) + \text{gap cost}\}, \end{aligned}$$

given that $d(i0, j0) = 0$.

In the example here (Figs. 1–3), the gap cost is set to 10 and the mismatch cost is set at 1. The update begins in the null cell of the top left-hand corner (Fig. 2a). The initial value of this cell is zero. The operation continues across the row to the right. Cell (0, 1) has only one neighbor cell, (0, 0). Therefore, cell (0, 1) is assigned a value of 10, that being the gap cost of 10 plus the value of the previous cell (0, 0), which is zero. Cell (0, 2) is assigned a value of 20, a gap cost of 10 plus the cell value of its only neighbor (0, 1) also 10. This process continues across the row, sequentially adding the gap cost to the next cell. The procedure is repeated for column 0 as each cell in that column only has one neighbor. Cell (1, 1) is the first cell where the optimal determination of path cost is performed (Fig. 2a). For cell (1, 1), the path cost from cell (0, 1) is 20, 10 from the gap cost and 10 from the cell value of (0, 1). The same

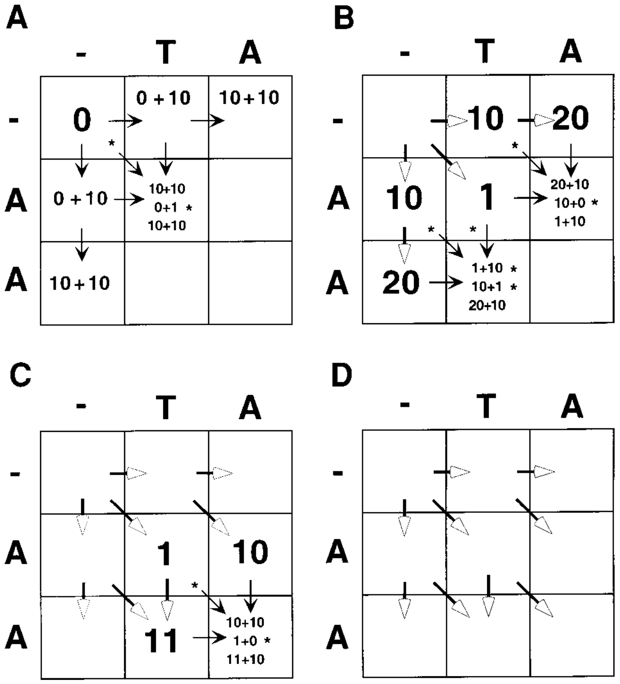
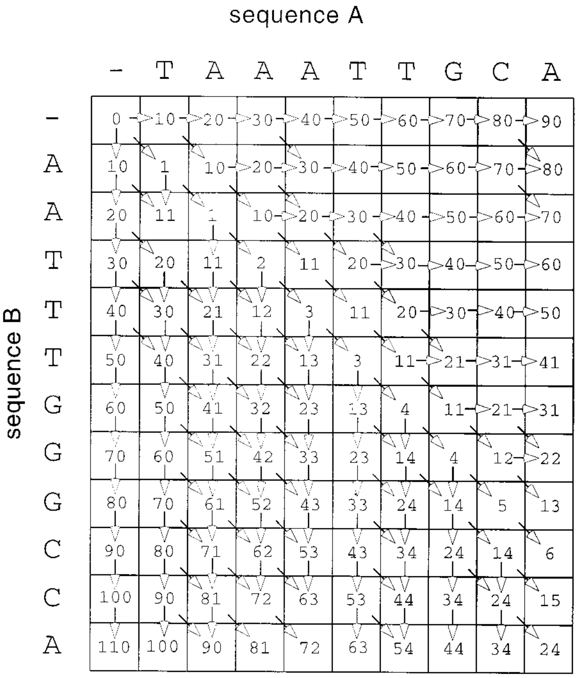


**FIG. 2.** Wavefront updating of the matrix elements corresponding to the first two nucleotides in sequences A and B from Fig. 1. Each cell in the matrix has its value reassigned based on an optimal pathcost calculation. (a) The leading cell of the matrix at the top left (0, 0) has a default value of zero. Each horizontal or vertical path to a cell is assigned a gap cost of 10 and the previously assigned mismatch costs are abandoned. A diagonal path to a cell withholds the cell's mismatch value (1). Each path, whether horizontal, vertical, or diagonal, brings with it the value of the cell from which that path originated. The cells in the first row and column of the matrix continuously accrete a gap cost of 10. Cell (1, 1) is the first cell in the matrix that must discriminate which of the three paths is optimal. An asterisk (*) designates which are the lowest path costs. In this instance it is the diagonal path from cell (0, 0). These optimal paths are retained in memory. (b) Once the optimal path(s) have been retained the value of the cell from which the path originated is no longer necessary and is not reported. The large arrows represent optimal paths to each cell retained in memory. The process is repeated for cells (2, 1) and (1, 2). The optimal path for cell (2, 1) is on the diagonal. Cell (1, 2) has two optimal paths, one on the diagonal and one from directly above. (c) The final cell (2, 2) undergoes the optimization procedure with a single optimal path on the diagonal. (d) A fully updated matrix with the optimal paths retained.



| | - | T | A | A | A | T | T | G | C | A |
|---|---|---|---|---|---|---|---|---|---|---|
| - | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
| A | 10 | 1 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 |
| A | 20 | 11 | 1 | 10 | 20 | 30 | 40 | 50 | 60 | 70 |
| T | 30 | 20 | 11 | 2 | 11 | 20 | 30 | 40 | 50 | 60 |
| T | 40 | 30 | 21 | 12 | 3 | 11 | 20 | 30 | 40 | 50 |
| T | 50 | 40 | 31 | 22 | 13 | 3 | 11 | 21 | 31 | 41 |
| G | 60 | 50 | 41 | 32 | 23 | 13 | 4 | 11 | 21 | 31 |
| G | 70 | 60 | 51 | 42 | 33 | 23 | 14 | 4 | 12 | 22 |
| G | 80 | 70 | 61 | 52 | 43 | 33 | 24 | 14 | 5 | 13 |
| C | 90 | 80 | 71 | 62 | 53 | 43 | 34 | 24 | 14 | 6 |
| C | 100 | 90 | 81 | 72 | 63 | 53 | 44 | 34 | 24 | 15 |
| A | 110 | 100 | 90 | 81 | 72 | 63 | 54 | 44 | 34 | 24 |

**FIG. 3.** A fully updated pairwise matrix of the complete sequences with the optimal paths. The cell values are retained only for illustration; they represent local edit distances. The terminal cell (10, 12) represents the global edit distance between sequences A and B.

is true of the path cost from cell (1, 0), 10 from the gap cost and 10 from the cell value. The path cost from (0, 0) to (1, 1) is 1; there is no gap cost since it is on the diagonal and the cell value of cell (0, 0) is 0, but a mismatch cost is applied since cell (1, 1) represents a correspondence of an A and a T (Fig. 2a). The new cell value entered into cell (1, 1) is 1. At this time, the path from which that value was derived is logged. If two path costs are equal, an arbitrary choice is usually made, but both paths can be retained in memory. Each cell in turn undergoes this process until all cells are updated (Figs. 2b–2d, 3).

All possible alignments, whether optimal or suboptimal, are represented as pathways through the array. The traceback begins at the terminal (bottom right) element of the matrix. Any previously retained lowest path cost notation that can be connected consecutively through $(i - 1, j)$, $(i, j - 1)$, or $(i - 1, j - 1)$ is traced back through the matrix (Fig. 4). This trajectory represents a sequence of edit operations which transforms sequence A into sequence B. This edit path forms the alignment. An uninterrupted diagonal through the array would represent no gap assignments. There may be more than one optimal pathway (Fig. 5); with this particular parameter set there are six possible pathways through the matrix. The traceback procedure merely recognizes any series of retained cell to cell paths that are contiguous through the entire matrix. These contiguous cell to cell paths are forged during the wavefront update and it is at this point that positional homology is established.

*Cost functions.* The gap cost and mismatch cost associated with the N-W algorithm are in a dynamic relationship; increasing mismatch cost will create more gaps in the alignment and increasing gap cost will increase the number of mismatches. Accordingly, an alignment may only be optimal for a particular combination of mismatch and gap costs. Alter these values and the optimal alignment may alter as well yielding a different phylogenetic data set. How, then, does one decide which combination of parameter sets to use? In general these choices are arbitrary. The following is a discussion of various implementations of cost functions in the N-W algorithm. There is a myriad of variations on the implementation of cost functions. Most of these implementations are attempts to mimic biological processes or constraints, which are thought to regulate the evolution of DNA or protein sequences.

*Mismatch cost functions.* There are many variations on the type of mismatch costs one can assign when laying out the N-W matrix. Aside from binary cost functions (0 = nucleotide match or 1 = mismatch), a transformation matrix of substitution costs can be instituted which will assign a separate penalty for each class of mismatches observed. Nucleotide sequence alignment has six types of mismatches in a symmetri-

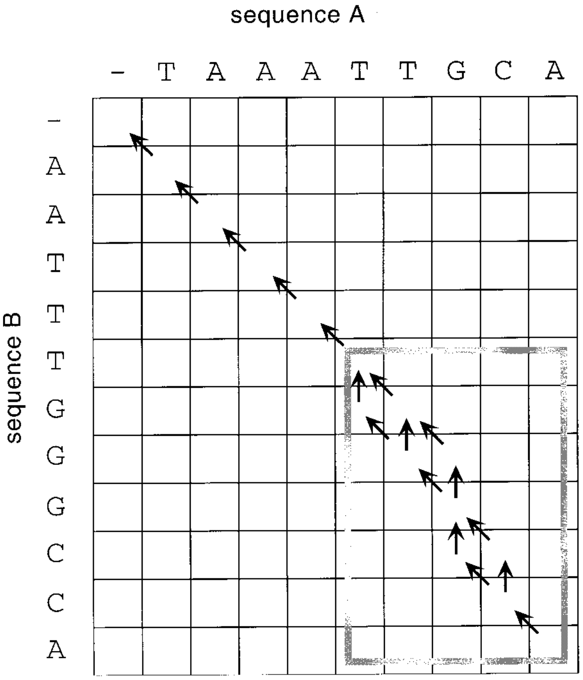


**FIG. 4.** The traceback procedure begins at the terminal cell (bottom right corner) in the matrix and tracks a path back through the matrix following all retained optimal paths until the top left cell (0, 0) is reached.

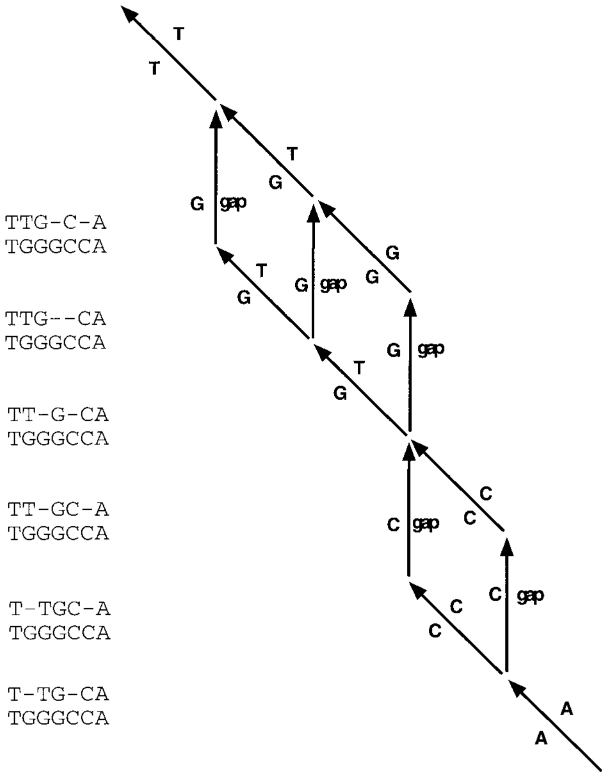


**FIG. 5.** An edit graph representation of the traceback operation in the highlighted region of Fig. 4. Positions in sequence A are represented above the diagonal arrows or to the right of the vertical arrows, positions from sequence B are below the diagonal arrows or to the left of the vertical arrows. This particular parameter set (gap = 10, mismatch = 1) yielded six different optimal paths through the matrix. Each of the six equally optimal alignments can be reconstructed by following every possible route through the edit graph.

cal transformation matrix. One could assign values to these transformations in a mismatch cost matrix based on many different criteria, for instance, observed nucleotide bias. These transformations are often grouped into two classes, transitions and transversions, but they are certainly not limited to these. Amino acid alignments can have much more complicated transformation matrices and there are many different types of amino acid substitution matrices in use.

The minimum mutation distance matrix (Fitch, 1966) is based on the minimum number of nucleotides which must be changed in order to convert the codon for one amino acid to the codon of another amino acid. The most common type of transformation table is the log odds matrix. These log odds matrices contain the relative frequencies with which amino acids are assumed to replace one another over time. The odds ratio is the ratio of the number of times residue *X* is replaced by residue *Y* in a pairwise alignment divided by the number of times residue *X* would be expected to replace residue *Y* is replacements occur at random. Positive values in the matrix indicate a replacement rate greater than expected by chance. Negative scores indicate a replacement rate less than expected by chance. It is assumed these values roughly correspond to conservative and nonconservative replacements, respectively. These log odds values can easily be converted into mismatch cost transformation matrices.

The most prevalent of the log odds matrices is the PAM matrix (Dayhoff *et al.*, 1978). The PAM matrix (allowed point mutations) is constructed by pairwise comparison of 72 sequence families consisting of more than 1300 sequences. A PAM**x** matrix is calculated from the original PAM1 by multiplying the PAM1 matrix by **x** times with itself, thus giving the probability of **x** pam1 mutations. Low PAM**x** matrices are used with closely related sequences, while high PAM**x** sequences are to be used for distantly related sequences (Dayhoff *et al.*, 1983). Blosum matrices (Henikoff and Henikoff, 1992) are based on well-conserved blocks of multiply aligned sequence segments, or motifs, that represented the most conserved regions of aligned families. Various Blosum matrices differ by the way the clustered sequences are built. Blosom 62 contains blocks that are at least 62% similar with one or more of the blocks in the cluster. These substitution probabilities are averages derived from pairwise alignments over a wide range of evolutionary distances. All log odds matrices are of course dependent upon the alignment parameters used in their construction.

*Gap penalty functions.* Gap coding in phylogenetic analysis is not necessarily straightforward. Gaps are not observations but are constructs in the alignment representing implied insertions or deletions. The decision to institute a gap in the alignment is a result of the path cost calculation during the wavefront update of the matrix elements. The simplest mode of inserting gaps into the alignment is to consider each cell in the matrix that implies a gap to be independent of any contiguous cell that also implies a gap. This is the unitary or simple gap cost. In this mode, five contiguous gap symbols contribute to the alignment cost as much as five dispersed gap symbols. This is the method used in the N-W algorithm.

More complicated gap dependencies have been proposed (Waterman *et al.*, 1976; Gotoh, 1982; Waterman, 1984; Miller and Meyers, 1988). These methods treat each contiguous string of gap characters as a single event instead of the sum of "*k*" events, where "*k*" is the length of the string. These procedures take into consideration the number of adjacent gap symbols in a string and are commonly known as affine gap costs. These penalties are composed of two parts, the gap initiation cost and the length-dependent gap extension cost. The gap extension cost need not be a linear function. With concave gap functions, the value of each increment in the gap extension cost decreases with each additional gap character to the string. Gu and Li

(1995) analyzed gap length in processed pseudogenes and found the length distribution to be a log function. This implies a gapping function of $w_g = a + b \ln k$ where $a$ is the gap initiation cost, $b$ is the gap extension cost and $k$ is the gap length. Although many algorithms which incorporate concave gap costs have been described in the literature (Knight and Meyers, 1995; Miller and Meyers, 1988; Allison, 1993), few computer applications actually include them. An additional modification in gapping is to allow gaps at the beginning and end of a sequence to be free of any cost. Cost-free end gaps should be used with caution as this enters the realm of local alignment.

One of the most problematic areas in phylogenetic sequence alignment is the influence of long gaps. Dynamic programming cannot look forward into the matrix. The wavefront update, which establishes the optimal path costs, is unidirectional and can only base its decision to institute a gap or correspondence based on the observed cost function up to the present point. As a gap becomes increasingly long it may become more economical to begin to apply mismatches than to extend the gap even though there may be a high scoring contiguous string of matches later on in the matrix. This is apparent in many 18S ribosomal data sets (see Whiting *et al.,* 1997). Long gaps are particularly a problem when memory-saving modifications of the N-W-algorithm are applied (see below). Gotoh (1982) designed an algorithm to deal with this issue which assigned a constant cost for any gap exceeding a specified length. Another approach also developed by Gotoh (1990) uses a series of linear functions of decreasing slope, which approximate a concave function. These more complex gapping functions contribute to the computational complexity and memory requirements of the operation because all (or many) cell values in column $i$ and row $j$ must be considered during the update of cell $(i, j)$ (Fig. 6).

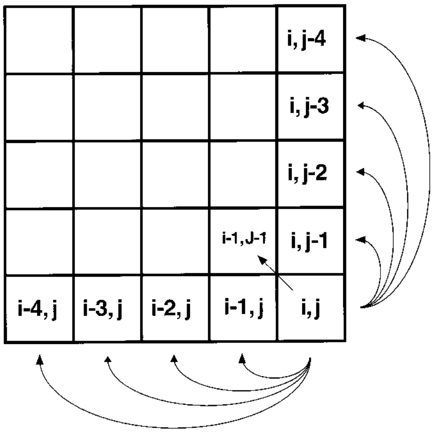


**FIG. 6.** Affine gap costs consider not only the three cells $(i, j - 1; i - 1, j; i - 1, j - 1)$ adjacent to cell $(i, j)$ but every cell in row $i$ and every cell in column $j$. The cell value of each of these cells represents the lowest edit distance required to reach that cell from the first cell (0, 0) in the matrix.

*Affine gap costs and character coding.* Characters in a phylogenetic analysis of sequence data are essentially columns in the alignment. One of the assumptions in phylogenetics analysis is that these columns are independent of one another. They are not, however. The positional homology assessments (the decision to institute a gap or a correspondence) are entirely context dependent and are based on the determination of the minimum path costs during the wavefront. When assigning affine gap costs, the neighborhood of influence expands even further back into the matrix (Fig. 6). The decision to treat a series of contiguous gaps as a singular event may be biologically founded but it is not logically consistent with how we designate what is a character and the phylogenetic analysis of those characters. If a long gap is a singular evolutionary event, it will be represented in the phylogenetic analysis many times since it occupies several columns in the matrix. This results in a problem: how does one code the gap character transformation cost in the vertical column of the data matrix if it is dependent on other positions? Furthermore, what of the nucleotide variation which corresponds with the gap positions? Treating gaps as missing data is not a solution, since these gaps are not in any way “missing” but the results of a specific mode of sequence change. The effect of missing data can be dubious during the character-state optimization procedure, missing data can be construed as an unobserved nucleotide or amino acid and the character state which provides the shortest length tree is inserted (Nixon and Davis, 1991; Platnick *et al.,* 1991). This is not desirable when gaps are implied. Since gap distribution is an integral part of the positional homology assignment process there is no justification for excluding gap positions from the character analysis.

We do not reject the utility of biological constraints when performing alignments, we submit that it creates a complexity in the delimition of characters and state transformation costs which is not usually carried over into the phylogenetic analysis. A restriction of using independent positional homology is that columns as characters do not apply with gap positions which are derived using affine gap functions since the gap characters are interrelated. A tentative solution would be to use unitary (independent, linear) gap costs in the alignment. Alternatively, the entire gap could be the character with the different character states determined by length differences and corresponding nucleotide variation although alternative solutions may lie outside the realm of alignment (Wheeler, 1996; Wheeler, 1999).

*Methods for saving computational effort and memory.* There have been many attempts at conserving computational effort and memory requirements for dynamic programming in general. One can assume that it is not necessary to explore the entire pairwise matrix to find the optimal edit distance. Using a diagonal path through the matrix as the null optimal, one need only consider a portion of the matrix a certain distance from the diagonal. Ukkonen (1985) utilizes the notion that the optimal solution is somewhere near the diagonal by defining a boundary ($2k + 1$) around the diagonal based on the number of gaps present in the sequence so far. Meyers and Miller (1988) devised a method based on Hirschberg (1975) which does not require that the matrix be retained permanently as in the N-W algorithm. The edit distance score is used in a divide-and-conquer procedure that recursively bisects the matrix until there is a series of smaller alignments which require less computational effort. The alignments are subsequently concatenated. Gotoh (1990) then modified the traceback procedure to include all possible optimal alignments. These methods are generalizable to the multiple sequence alignment problem (Carillo and Lipman, 1988) described below.

## MULTIPLE SEQUENCE ALIGNMENT

The N-W algorithm was originally defined for two sequences. In principle, the procedure can be extended to any number of sequences, thereby defining an $N$-dimensional matrix. However, the addition of sequences opens up an immense computational problem. The number of cells in a true simultaneous multiple alignment matrix are exponentially related to the number of taxa and sequence length. Even using the methods that save storage and effort, simultaneous alignment of more than a few sequences is computationally intractable.

Sankoff and Cedergren (1983) proposed a tree-based multiple alignment method within an $N$-dimensional N-W framework. Their method requires, however, that the cladogram of relationships is known a priori and alignments are performed with the cost of each cell in the alignment space determined by the "known" cladogram. Initially, a space is created with $\Pi_0^n L_n + 1$ cells ($n$ is the number of the sequences and $L$ their lengths). A pass is made through this space as with the two-dimensional method, updating each cell. The updated cost of each cell would be the minimum of the cost of each adjacent cell incremented by the cost of the current cell, as determined by the known tree. For $n$ taxa there are $2^n - 1$ cells to be examined (all the combinations, indels, and base matches and mismatches) to determine the cost of each cell.

This is a straightforward extension of the two-dimensional N-W case. As with the 2-d case, the cost of each cell is determined by the minimum of the sum of the subpath to each adjacent cell and the added cost of the path to the current cell. Where the cost of cell ($i, j$) for sequences A and B and transformation cost set "$w$",

$$d_{ij} = \min\left\{\begin{array}{l} d_{i,j-1} + w(\text{“—”}, B_j) \\ d_{i-1,j} + w(A_i, \text{“—”}) \\ d_{i-1,j-1} + w(A_i, B_j) \end{array}\right\}$$

for two sequences becomes

$$d_{ij\cdot\cdot\cdot k} = \min\left\{\begin{array}{l} d_{i,j-1,\cdot\cdot\cdot k} + w(\text{“—”}, B_j, \ldots, c_k) \\ d_{i-1,j\cdot\cdot\cdot k} + w(A_i, \text{“—”}, \ldots, c_k) \\ \quad\vdots \end{array}\right\}$$

$$d_{i,j,\cdot\cdot\cdot k-1} + w(A_i, B_j, \ldots, \text{“—”}),$$

when "$k$" sequences are included. The minimization occurs over all possible combinations of gaps and matches for a total of $2^k - 1$ path to cell $i, j, \ldots, k$. In this general case, the "$w$" set of transformation costs would be based on the predetermined tree. The parsimoniously reconstructed (but not searched) length of the a priori tree is used for the "$w$" costs.

If the tree of relationships were not known ahead of time (the most likely case), the alignment procedure could be repeated for each possible scheme of relationships or phylogenetic searching could be performed for each cell. Given that the number of cells in the alignment space is exponentially related to the number of taxa, and the number of trees is combinatorially dependent on the number of taxa, this type of approach, though exact, would be intractable for all but the smallest data sets.

## HEURISTIC MULTIPLE ALIGNMENT

As a result of this combinatorial complexity, simultaneous alignment of all sequences is rarely attempted. Instead, a series of pairwise alignments are performed and these subalignments amalgamated into a multiple alignment. The intermediate pairwise alignments are added together following a tree-like pattern. The order by which this is carried out is determined by a "guide tree" (Feng and Doolittle, 1987). Each node of the tree represents a separate pairwise alignment. Mindell (1991) advocated using known phylogenies to guide alignments but the required phylogenetic information is often unavailable. In most evolutionary studies, the object of performing a multiple alignment is to allow phylogenetic analysis with a set of putative homologies unbiased by initial assumptions of relationship. Preconceived notions of relationships will bias the analysis.

The phylogeny resultant from analysis of a multiple alignment is obviously dependent on the order in which the sequences are accreted. However, if gap assignment is unambiguous, many different guide tree topologies will lead to the same phylogeny. Thus, the guide tree topology is separate and distinct from the topology of the phylogenetic tree derived from the alignment.

Several methods are currently in use to progressively align sequences via pairwise accretion. The differences among them center on two areas: (1) techniques of establishing an alignment topology or topologies (guide tree), and (2) how the aligned sequence positions at the nodes are combined to create the complete multiple alignment. The guide tree can be established using a pairwise distance-based approach or by choosing from many guide trees in a parsimony framework. The results of each pairwise alignment in the guide tree can produce a consensus sequence which is resolved later, or the character state can be resolved as soon as possible within the alignment process. The decision to choose one mode over another tends to be based on computational effort, the methods that iterate through multiple guide trees being the most consumptive of computational resources.

*Distance-based guide trees.* Initially, guide trees were determined based on distance methods (Feng and Doolittle, 1987, 1990; adapted by Higgins and Sharp, 1988, 1989; Higgins *et al.*, 1992; Thompson *et al.*, 1994). Thompson *et al.* (1994) described the procedure as follows. A similarity score is calculated from a pairwise alignment between every possible pair of sequences (Wilbur and Lipman, 1983). A distance matrix composed of these scores is used to calculate a dendogram using the UPGMA method of Sneath and Sokal (1973) as an alignment topology. This topology is used to direct the alignment of the most similar sequences in a N-W procedure. The CLUSTAL alignment program (Thompson *et al.*, 1994) (ftp://ftp-igbmc.u-strasbg.fr/pub/) aligns the most similar sequences first. A consensus sequence is substituted for the sequence pair. Consensus sequences incorporate only bases present in all sequences or use partial (75%) consensus. Gaps inserted in any alignments are preserved throughout the alignment. Clustal progressively aligns the next most similar sequence to the consensus of the growing cluster or the next two most similar sequences to each other (Fig. 7). Only a single multiple alignment is constructed. However, an industrious user could specify various alignment topologies and run the program repeatedly to explore the relationship between alignment topology and phylogenetic results.

There are other methods based on sequence similarity (Hein, 1989a,b, 1990; Konings *et al.*, 1987). These methods differ in how the distance tree is determined and how the tree interacts with the actual alignment. Hein (1989a,b, 1990) extended the process by adding a parsimony step. In TREEALIGN (Hein, 1989a,b), (ftp://ftp.ebi.ac.uk/pub/software/unix/treealign.tar.Z), pairwise distances are used to construct an alignment topology and an initial alignment with observed sequences at the edges of the tree. The alignment topology is converted to a parsimony tree and used to direct an alignment algorithm. During alignment, potential ancestral sequences are created at each node in the alignment topology using parsimony (Fig. 8). Although a parsimony score is attached to the alignment topology, no other trees are constructed for comparison. The alignment topology is subjected to nearest neighbor interchanges until all branches are swapped or a user-defined number of swapping cycles is reached. Sequences are aligned along the resultant tree via a graph comparison algorithm similar to that of Sankoff and Cedergren (1983). Ancestral sequences are determined for each node via dynamic programming. Base mismatches (e.g., A in sequence 1 and C in sequence 2) are incorporated in the ancestral sequences by the union of the bases (A + C). A choice between the alternatives is postponed until evidence higher up in the tree points to either A or C (Fig. 8). If nothing favors A or C an arbitrary choice is made (J. Hein, pers. commun.).

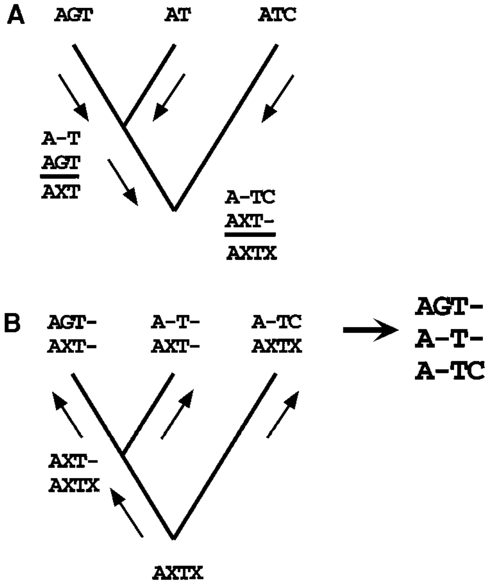


**FIG. 7.** Tree-based depiction of multiple alignment strategy in CLUSTAL. Let all base changes and gaps cost 1. A distance tree is created by UPGMA. In the downpass terminal sequences are related by a dendrogram. (A) A pairwise alignment of two most "closely related" terminal sequences. An unknown residue *X* is used to placehold any disagreement. Then a pairwise alignment is performed between the consensus produced at the previous node with next most "closely related" sequence. (B) In the up pass, progressive alignment of the consensus sequences and terminal sequences is used to resolve any ambiguities and to introduce gaps.

*Parsimony-based alignment topologies.* Any single addition order can lead to a result that is not globally optimal. This is one of the most severe problems of nonexact solutions. In MALIGN (Wheeler and Glad-

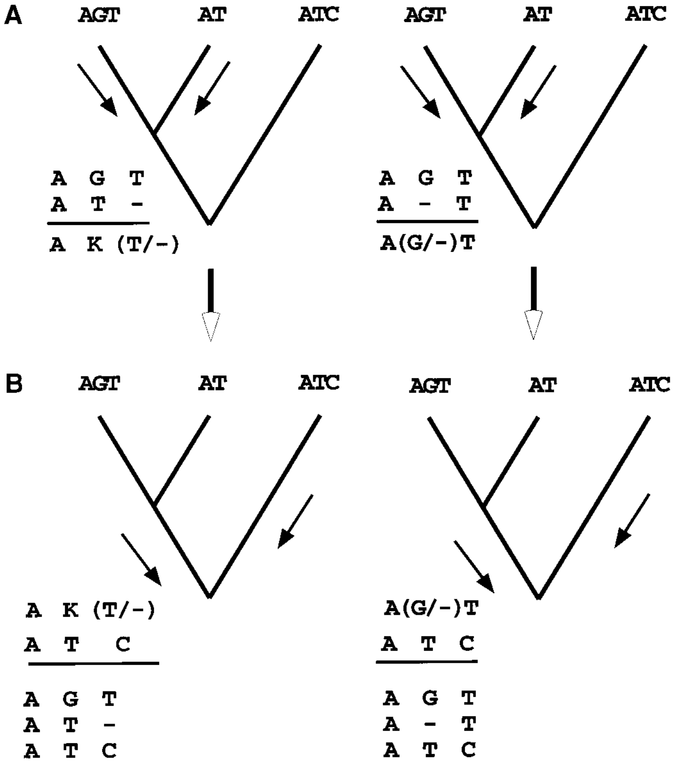


**FIG. 8.** Tree based depiction of the multiple alignment strategy in TREEALIGN as described by Hein (1989b). A distance tree is constructed and subjected to nearest neighbor interchanges. In the down pass terminal sequences are related by a parsimony guide tree that has the same topology as the distance tree. (A) Two subalignments of AGT and AT and implicit ancestral sequences (IUPAC coding used here for ambiguous nucleotides, K = G or T). (B) Selection of most parsimonious ancestral sequence by comparison to ATC. Rearrangements are iterated on the parsimony guide tree to improve the alignment cost. With a mismatch cost of 1 and a gap cost of 1 both alignments are equivocal. TREEALIGN makes an arbitrary choice between the two.

stein, 1994) (ftp://ftp.amnh.org/pub/people/wheeler/malign/), many alignment topologies are used to explore multiple alignments. Various alignment topologies can be constructed via random addition of sequences and branch-swapping. As sequences are accreted alignment topologies (subtrees) are produced by adding sequences to the branch that produces the least costly alignment. Minimization occurs by searching for the least costly path through a N-W matrix determined by the sequences and costs associated with accepting nucleotide mismatches or inserting gaps. Subtrees are then combined to produce a complete multiple alignment.

Alignment cost can be improved through branch swapping (Fig. 9). MALIGN performs branch swapping by removing taxa and adding them back to all the other possible addition points on the cladogram or alignment hierarchy. The cost of the most parsimonious alignment topology is then assigned as the multiple alignment cost. In the heuristic option termed “build,” alignments are constructed by adding taxa to an alignment topology at each of the possible addition points and alignment orders available. Sequences are stored at subtrees as partial multiple alignments. As sequences are accreted, alignment topologies (subtrees) are produced by adding sequences to the branch that produces the least costly alignment. These procedures are analogous to heuristic cladogram construction algorithms (e.g., Farris, 1970).

## MAXIMUM LIKELIHOOD ALIGNMENT

The criterion of maximum likelihood (ML) can also be used to create multiple alignments via a route analogous to parsimony/minimization approaches. In the simple pairwise case, the same sort of N-W dynamic programming matrix is created and costs assigned to each cell in the matrix. Those costs, however, would be calculated in a slightly different manner. In the place of the relatively simple substitution-indel costs (even if more complex base substitution costs are specified) of the basic N-W, a more complex evolutionary model is used to assign the relative likelihoods of different sorts of base substitutions and indels. Unlike minimization procedures, where the minimum path to a cell (see Fig. 2) is chosen and dependent cells are updated based on that single minimum cost path, all paths are included in the likelihood calculations. That is, all possible paths to a given cell contribute to the likelihood value (Bishop and Thompson, 1986).

Other than this rather trivial difference, the similarity in process may continue through tree-based multiple alignment. Mitchison and Durbin (1995) suggest using ML trees to asses the quality of (i.e., optimize) multiple alignments. ML alignment procedures, however, are not widely used even by the proponents of likelihood in phylogeny reconstruction. We presume that this is due only to the computational costs involved in analyzing real data and not any reservation about evolutionary models. The methodologies are clear; the procedures can be made seamless from alignment through phylogeny estimation. The only barrier is implementation (see Thorne, Kishino, and Felsenstein, 1992).

Like all other forms of analysis, ML results are dependent on assumptions. These assumptions are encapsulated in the evolutionary models employed. The models specify not only aspects such as transition–transversion ratios and gap models, but also the shape and rates of sequence evolution. In the same way that parsimony based analysis can be subjected to variation in parameter assumptions, likelihood results should be examined for their behavior under variation of assumptions. With the likelihood value as an optimality criterion (like minimum length), different combinations of parameter and evolutionary model can be

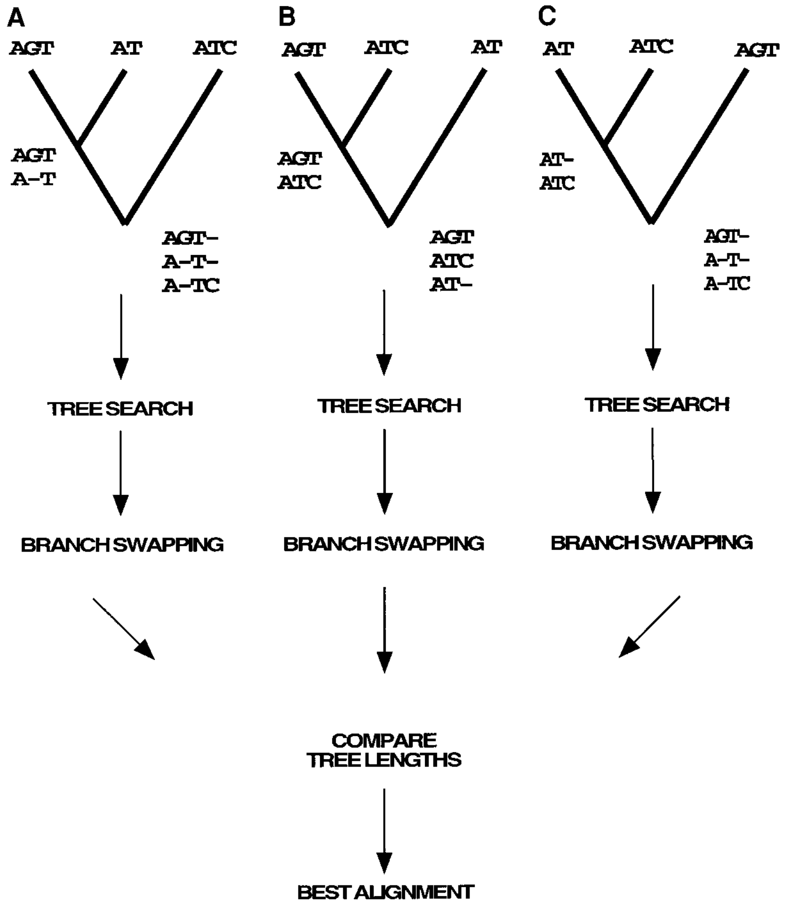


**FIG. 9.** Tree based depiction of an example multiple alignment strategy in MALIGN. Let all base changes = 1 and gaps cost = 2. In each "build" step, terminal sequences are related by a cladogram that results from a random or user-specified addition sequence. For these taxa there are three possible guide trees, A, B, and C. After the tree search is conducted, each alignment is then subjected to branch swapping to potentially improve optimization of character changes. The alignment that produces the shortest tree is the best alignment. In this simple example alignment topology A and C produce the same optimal alignment.

tested to maximize the criterion of choice. These operations would be identical to those for minimization/parsimony, but with a different criterion of optimality.

## PARAMETER SENSITIVITY

The phylogenetic analysis of nucleic acid sequences, as with other data, is unavoidably based on explicit and implicit assumptions. Results of multiple alignment and phylogenetic analysis, no matter the algorithms, are sensitive to choice of evolutionary model. At the fore are character transformation models. For example, various weights must be assigned to parameters such as transitions, transversions, and insertion–deletion events. There are no known means of determining, a priori, which alignment parameters are appropriate for recovering evolutionary relationships.

Simple homogeneous weighting does not avoid the issue of arbitrary, yet crucial, assumptions. As an example, transversion–transition ratio and gap costs are generally not directly measurable. These values are statements of process and they can be inferred appropriately only from a predetermined phylogenetic pattern. The interaction between the specification of values a priori and their inference a posteriori is a general and central problem in molecular phylogenetic analysis.

One of the benefits of likelihood techniques is that the method can estimate the values of its own parameters by simultaneously varying parameters until global maximum likelihood is achieved. In the case of alignment, the likelihood of an alignment based on one set of parameters (e.g., indel cost, transversion ratio) can be compared to that based on another. Unlike the numerical values derived from parsimony analyses, a likelihood of 0.1 for an alignment with a gap cost of twice that of base changes is superior to a likelihood of 0.01 based on gaps costing four times base changes. Continuing this logic, the maximum likelihood alignment over all (or some heuristic subset) of analysis parameters gives both the alignment and the maximum likelihood estimate of each of the components of the model. The costs of alignments based on weighted parsimony are not comparable in this way. A cost (or derived cladogram length) of two is not necessarily superior to a length of four. Each solution is most parsimonious for its own set of parameters and not comparable from parsimonious solution to parsimonious solution.

*Alignment space and congruence.* In order to estimate the multiple alignment parameters, both a model and a space are posited. The model determines the general means of calculating likelihoods based on both the form of the model and the assumed values of its parameters. To perform the likelihood estimates, the likelihood is calculated for each point in the parameter space. The point with the maximum value gives the likelihood estimate of the alignment, and the points along each axis give the parameter values.

If several sources of information are to be used, each data set may require a unique model. It is unclear how to assess the ensemble phylogenetic conclusions. The use of external criteria offers a way of accommodating such results. If an external criterion can be defined, the behavior of each solution can be calculated and compared to those of other solutions. One such external criterion is congruence. As used by Wheeler (1995), congruence measures can be posited and all solutions assayed. The set of analysis parameters which maximizes (or minimizes) this value is then optimal. The potential problem with this approach is that different optimality criteria may be used within and among models. Those solutions, which are optimal by one criterion, may not be with another.

*Sensitivity analysis.* Even though the basic alignment parameters of transversion–transition and gap–change cost ratios are unmeasurable in the absence of a predetermined phylogeny, it is possible to estimate their values through appeal to an external optimality criterion. The most reasonable for phylogenetic analysis must be congruence (whether taxonomic [Nelson, 1979] or character based [Mickevich and Farris, 1981], but see Miyamoto [1981, 1985]). Without any way of objectively measuring the accuracy of reconstruction, only the agreement among data can be used to arbitrate among competing hypotheses.

In order to estimate the sensitivity of an analysis to variation in parameter values, the range of each of the parameters must be determined. This establishes the "analysis space" of the problem. In this space, all possible combinations of parameter values are present; hence all analytical conclusions are implied. These combinations of values are sampled and their analytical consequences determined (see Fitch and Smith, 1983; and Vingron and Waterman, 1994). This would, in the most general case, involve an *N*-dimensional space with each of the *N* parameters defining an axis bounded by the parameter ranges. The two parameters of insertion–deletion cost and transversion–transition ratio would constitute the axes of a simple analysis space (Fig. 10). A sampling regime would consist of taking parameter pairs (transversion–transition ratio, gap–change cost ratio) from this space, aligning the sequences, erecting hypotheses of relationship based on these values, and assaying congruence with an external data set.

Even with only two parameters, the available analysis universe is infinite. Each of the parameters can, at least numerically, achieve any positive real value. Realistic sampling of such a space may be difficult. These values are not boundless and may be constrained by the logic of the triangle inequality as formulated for character analysis (Wheeler, 1995).

Within these theoretical limits, a residuum of possible values exists for the analysis parameters. With two parameters, a plane bounded on two adjacent sides is defined (Fig. 10). Since any and all combinations of parameter values which fall in this plane are possible at least logically, they must all be examined (or at least some sample). To accomplish this sampling, alignment and phylogeny reconstruction must be performed with sufficient combinations of possible values to represent the behavior of the entire space. This is a relatively straightforward procedure (if time-consuming). For each point (a combination of transversion–transition and gap–change value ratios) to be sampled, the sequences are aligned and phylogeny reconstructed. Both alignment and phylogeny reconstruction are performed using the same combination of parameter values. At each of these points, some measure of congruence is calculated with respect to some external data set, the variation of which can be used to assay both the most appropriate values for the unmeasurable parameters and the effects of variation in these parameter values on the overall conclusions of the analysis.

If some congruence measure is plotted with respect to the parameter values, a "congruence surface" is generated, the relief in this surface denoting the areas of relative congruence and incongruence. This surface can be used to estimate the values of the analytical parameters. As with statistical inference, two types of decision (estimate of parameter values) can be made—"best" and "robust." A "best" decision is made by choosing the set (or sets) of parameter values at which the optimality criterion is maximized. According to this type of decision, the set of values for transversion–transition ratio and gap–change ratio that maximize congruence would be chosen. On the other hand, a "robust" decision selects a range of parameter values rather than settling on a single set. This range defines a subset of the analysis space in which some statement is supported. For example, an area might be specified

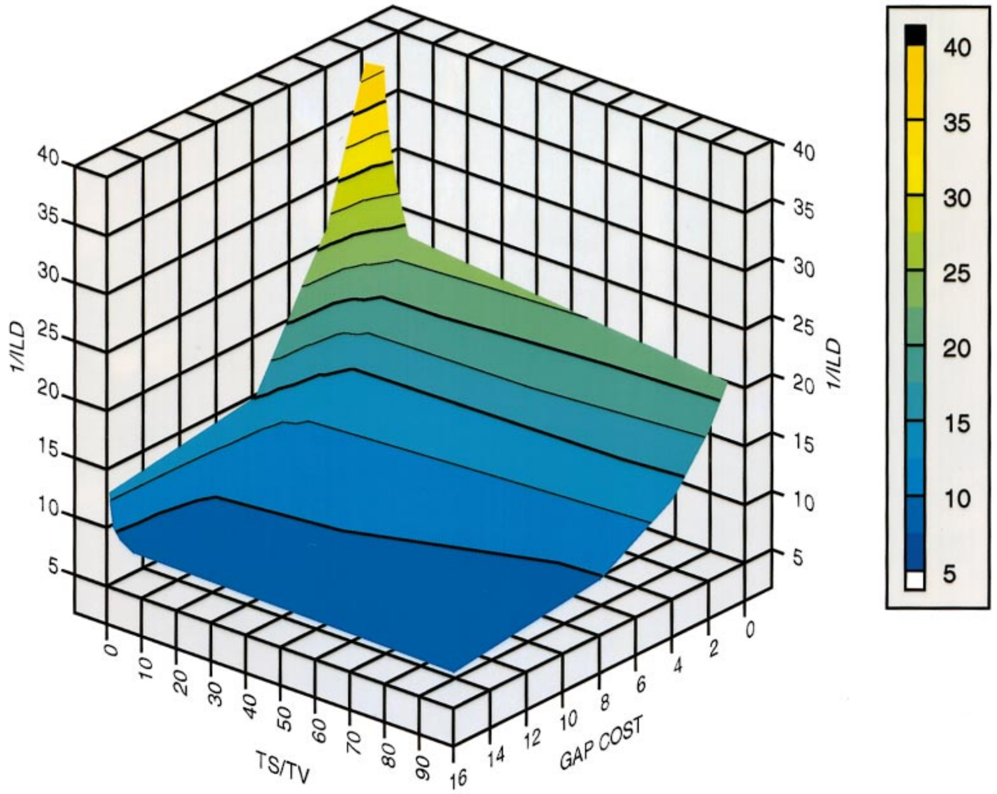


**FIG. 10.** Graphical representation of an alignment parameter sensitivity analysis of a data set consisting of the 12S, 16S and cytochrome B genes from 35 carnivore taxa. The incongruence length difference (ILD) was calculated by subtracting the treelengths of the individual datasets from the treelength of the combined data set and dividing that value by the treelength of the combined data set {ILD = (treelength combined − treelength 12S − treelength 16S − treelength cytB)/treelength combined}. The alignment parameters explored were a range of gap costs including 1, 2, 4, 8, and 16, and transition/transversion ratios of 1, 2, 4, 8, and transitions only. The alignment parameters that yielded the least internal data conflict were gap cost of 2 and a transition/transversion ratio of 1 with an ILD of 0.02569.

in which some particular group was monophyletic, but this clade was not supported generally.

The Mickevich and Farris (1981) measure of congruence seeks to assess the degree of character conflict among multiple data sets. The statistic of Mickevich and Farris (1981) quantifies the degree of character conflict by measuring the number of extra steps forced upon the individual data sets when they are combined. In this way, the additional conflict created by the combination of the data is assessed separately from that derived from internal character conflict. The value generated is simply the length of the most parsimonious cladogram(s) derived from the combined data minus the sum of the lengths of the cladograms from the constituent data sets. This number of steps is normalized through division by the length of the combined data. A value of zero implies complete character congruence, while higher values denote increasing degrees of character conflict between the data sets. No topology statement is implied or required. In fact, data sets, which have zero taxonomic congruence, can have 100% character congruence. This can occur if one data set yields an unresolved bush and the second yields one of its many potential resolutions.

## CONCLUSIONS

In many ways, alignment is where phylogenetic analysis was 20 years ago. Many investigators still advocate creating alignments "by hand," asserting that the human brain is better at determining homology, that computer analysis is not "biological." Computer programs for performing alignments are in their infancy and users are often unfamiliar with the numerical and methodological assumptions made. Presentations at conferences may cite alignment software, but leave crucial information such as gap costs or search algorithms undescribed. Clearly, an increase in analytical sophistication and clarity is warranted in this primary stage of phylogenetic analysis.

We advocate three aspects of phylogenetic alignment-reconstruction: (1) the definition of an optimality criterion for alignment, (2) the consistent use of analysis assumptions in both alignment and phylogeny reconstruction, and (3) the examination of these assumptions through sensitivity analysis to examine the robustness of conclusions.

Since alignment seeks to minimize nonhomology, parsimony seems the most logical of optimality criteria for alignment. That alignment which implies a cladogram of minimal length will, by definition, minimize nonhomology. This is not, of course, the only criterion, maximum likelihood being another. Whatever is chosen, though, that logic must be followed through the entire process from alignment to phylogenetic reconstruction. Whatever cost regime (indels and base substitutions), however defined, must be applied consistently. Without this connection, alignments might well imply other cladograms than the analyses generate, or conversely, resultant cladograms might imply different alignments. The final point, that of testing assumptions through sensitivity analysis, is based on and requires the first two. If analysis parameters are applied objectively and consistently, assumptions can be tested. By "tested," we mean that the effects variations in these assumptions have on phylogenetic conclusions can be assayed. It may be that a result is entirely dependent on specific values of indels cost or transition–transversion ratio. Only by varying these values and examining their perturbations can we assay the robustness of our conclusions.

These points are not specific to parsimony-based methods. Although we favor this approach, likelihood methods could equally well meet these three requirements. Our arguments are more for consistency, repeatability, and transparency in analysis than for any particular optimality criterion or epistemological creed.

## REFERENCES

Allison, L. (1993). Normalization of affine gap costs used in optimal sequence alignment. *J. Theor. Biol.* **161:** 263–269.

Altschul, S. W., Gish, W., Miller, W., Meyers, E. W., and Lipman, D. J. (1990). Basic local alignment tool. *J. Mol. Biol.* **215:** 403–410.

Bellman, R. E. (1957). "Dynamic Programming," Princeton Univ. Press, Princeton, NJ.

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

Brower, A. V. Z., and Scharawoch, V. (1996). Three steps of homology assessment. *Cladistics* **12:** 265–272.

Carillo, H., and Lipman, D. (1988). The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.* **48:** 1073–1082.

Dayhoff, M. O., Barker, W. C., and Hunt, L. T. (1983). Establishing homologies in protein sequences. *Methods Enzymol.* **91:** 524–545.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. *In* "Atlas of Protein Sequence and Structure" (M. O. Dayhoff, Ed.), Vol. 5, Suppl. 3, pp. 345–352, Natl. Biomed. Res. Found., Washington, DC.

Farris, J. S. (1970). A method for computing Wagner trees. *Syst. Zool.* **19:** 83–92.

Feng, D., and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25:** 351–360.

Feng, D., and Doolittle, R. F. (1990). Progressive alignment and phylogenetic tree construction of protein sequences. *Methods Enzomol.* **183:** 375–387.

Fitch, W. M. (1966). An improved method of testing for evolutionary homology. *J. Mol. Biol.* **16:** 9–16.

Fitch, W. M., and Smith, T. F. (1983). Optimal sequence alignments. *Proc. Natl. Acad. Sci. USA* **80:** 1382–1386.

Giribet, G., and Wheeler, W. C. (1999). On gaps. *Mol. Phylogenet. Evol.* **13:** 132–143.

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

Gotoh, O. (1990). Optimal sequence alignment allowing for long gaps. *Bull. Math. Biol.* **52:** 359–373.

Gu, X., and Li, W.-H. (1995). The size distribution of insertion and deletions in human and rodent pseudogenes suggests a logarithmic gap penalty for sequence alignment. *J. Mol. Evol.* **40:** 464–473.

Hein, J. (1989a). A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when a phylogeny is given. *Mol. Biol. Evol.* **6:** 649–668.

Hein, J. (1989b). A tree reconstruction method that is economical in the number of pairwise comparisons used. *Mol. Biol. Evol.* **36:** 396–405.

Hein, J. (1990). Unified approach to alignment and phylogenies. *Methods Enzomol.* **183:** 626–644.

Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89:** 10915–10919.

Higgins, D. G., Bleasby, A. J., and Fuchs, R. (1992). CLUSTAL V: Improved software for multiple sequence alignment. *CABIOS* **8:** 189–191.

Higgins, D. G., and Sharp, P. M. (1988). CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. *Gene* **73:** 237–244.

Higgins, D. G., and Sharp, P. M. (1989). Fast and sensitive multiple sequence alignments on a microcomputer. *CABIOS* **5:** 151–153.

Hirschberg, D. S. (1975). A linear space algorithm for computing maximal common subsequences. *Commun. ACM* **18:** 341–343.

Kluge, A. J. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Syst. Zool.* **38:** 7–25.

Knight, J. R., and Meyers, E. W. (1995). Approximate regular expression pattern-matching with concave gap penalties. *Algorithmica* **14:** 85–121.

Konings, D. A., Hogweg, P., and Hesper, B. (1987). Evolution of the primary and secondary structures of the E1a mRNAs of the adenovirus. *Mol. Biol. Evol.* **4:** 300–314.

Meyers, E., and Miller, W. (1988). Sequence comparison with concave weighting functions. *Bull. Math. Biol.* **50:** 97–120.

Meyers, E., and Miller, W. (1988). Optimal alignments in linear space. *CABIOS* **4:** 11–17.

Mickevich, M. F., and Farris, J. S. (1981). The implications of congruence in Menidia. *Syst. Zool.* **30:** 351–370.

Miller, W., and Meyers, E. W. (1988). Sequence comparison with concave weighting functions. *Bull. Math. Biol.* **50:** 97–120.

Mindell, D. (1991). Aligning DNA sequences: homology and phylogenetic weighting. *In* "Phylogenetic Analysis of DNA Sequences" (M. J. Miyamoto and J. Cracraft, Eds.), pp. 73–89. Oxford University Press, New York.

Mitchison, G., and Durbin, R. (1995). Tree-based maximal likelihood matrices and hidden Markov models. *J. Mol. Evol.* **41:** 1139–1151.

Miyamoto, M. M. (1981). Congruence among character sets in phylogenetic studies in the frog genus Leptodactylus. *Syst. Zool.* **321:** 43–51.

Miyamoto, M. M. (1985). Consensus cladograms and general classifications. *Cladistics* **1:** 186–189.

Morrison, D. A., and Ellis, J. T. (1997). Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18S rDNAs of Apicomplexa. *Mol. Biol. Evol.* **14:** 428–441.

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48:** 443–453.

Nelson, G. J. (1979). Cladistic analysis and synthesis: Principles and definitions, with a historical note on Adanson's Familles des Plantes (1763–1764). *Syst. Zool.* **28:** 1–21.

Nixon, K. C., and Davis, J. I. (1991). Polymorphic taxa, missing values and cladistic analysis. *Cladistics* **7:** 233–241.

de Pinna, M. C. C. (1991). Concepts and tests of homology in the cladistic paradigm. *Cladistics* **7:** 367–394.

Platnick, N. I., Griswold, C. E., and Coddington, J. A. (1991). On missing entries in cladistic analysis. *Cladistics* **7:** 337–343.

Sankoff, D. D., and Cedergren, R. J. (1983). Simultaneous comparison of three or more sequences related by a tree. *In* "Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison" (D. Sankoff and J. B. Kruskal, Eds.), pp. 253–264, Addison-Wesley, Reading, MA.

Sellers, P. H. (1974). On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.* **26:** 787–793.

Smith, T. F., Waterman, M. S., and Fitch, W. M. (1981). Comparative biosequence metrics. *J. Mol. Evol.* **18:** 38–46.

Sneath, P. H. A., and Sokal, R. R. (1973). "Numerical Taxonomy," Freeman and Company, San Francisco.

Swofford, D. L., and Olsen, G. J. (1990). Phylogeny reconstruction. *In* "Molecular Systematics" (D. M. Hillis and C. Moritz, Eds.), 1st ed., pp. 411–501, Sinauer, Sunderland, MA.

Thompson, J. D., Higgins, D. G., and 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., and Felsenstein, J. (1992). Inching towards reality: An improved likelihood model of sequence evolution. *J. Mol. Evol.* **34:** 3–16.

Ukkonnen, U. (1985). Finding approximate patterns in strings. *J. Alg.* **6:** 132–137.

Vingron, M., and Waterman, M. S. (1994). Sequence alignment and penalty choice, Review of concepts, case studies and implications. *J. Mol. Biol.* **235:** 1–12.

Waterman, M. S., Smith, T. F., and Beyer, W. A. (1976). Some biological sequence metrics. *Adv. Math.* **20:** 367–387.

Waterman, M. S. (1984). General methods of sequence comparison. *Math. Biol.* **46:** 473–500.

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

Wheeler, W. C. (1996). Optimization alignment: The end of multiple sequence alignment in phylogenetics? *Cladistics* **12:** 1–9.

Wheeler, W. C. (1999). Fixed character states and the optimization of molecular sequence data. *Claudistics* **15:** 379–385.

Wheeler, W. C., and Gladstein, D. L. (1994). MALIGN, version 1.93. American Museum of Natural History, New York.

Whiting, M., Carpenter, J., Wheeler, Q., and Wheeler, W. (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.

Wilbur, J., and Lipman, D. (1984). The context dependent comparison of biological sequences. *SIAM J. Appl. Math.* **44:** 557–567.