3 Optimization Alignment: Down, Up, Error, and Improvements

Ward C. Wheeler

Contents

1 Introduction

Optimization Alignment (OA) is a method for taking unaligned sequences and creating parsimonious cladograms without the use of multiple alignment. The method consists of two parts. First, a "down-pass" that moves "down" the tree from the terminal taxa (tips) to the root or base of the cladogram and, second, an "up-pass" which moves back up from the base to the tips. The down-pass creates preliminary (i. e., provisional) hypothetical ancestral sequences at the cladogram nodes and generates the cladogram length as a weighted sum of the character transformations (nucleotide substitutions and insertion-deletion events) required by the observed (terminal) sequences. The up-pass takes the information from the down-pass and creates the "final" estimates of the hypothetical ancestral sequences. From these the most parsimonious synapomorphy scheme can be derived to show which character transformation events characterize the various lineages on the tree. The combination of these two procedures allows phylogenetic searches to take place on unaligned sequence data, resulting in improvements in execution time and quality of results. This process differs from multiple alignment procedures (such as that of Sankoff and Cedergren [1]) in that OA attempts to determine the most parsimonious cost of a

Methods and Tools in Biosciences and Medicine
Techniques in molecular systematics and evolution, ed. by Rob DeSalle et al.

cladogram directly, whereas multiple alignment procedures generate column-vector character sets, which are then analyzed phylogenetically in a separate operation.

This OA method was proposed to allow the direct optimization of unequal length sequences on a cladogram [2]. The determination of the length, or cost, of the cladogram is accomplished given the observed sequences and a cost matrix that specifies the costs of all the transformations among nucleotides and insertion-deletion (indel) events. In this sense, the method is a generalization of Sankoff and Rousseau [3], or matrix optimization of character states, but allows for the insertion and deletion of characters. Sankoff and Rousseau [3] expanded the realm of optimization allowing for unequal transformation costs among character states, but still relied on a preexisting alignment to know which states to compare. OA enlarges the world of transformation events that can be optimized on a cladogram, including the creation and destruction of characters. In doing this, OA obviates the need to perform multiple sequence alignment, creating unique, topology-specific homology regimes for each scenario of historical relationship. The method yields better testing of phylogenetic hypotheses since provisional homologies are optimized for each cladogram individually, not *a priori* and universally as with the static homologies of multiple alignment. Furthermore, by treating all sequence variation within the context of topology-specific synapomorphy, hypotheses of molecular variation can be seamlessly integrated with other character variation to yield simultaneous or total evidence analysis. The OA integration of molecular and other character information frequently generates more parsimonious cladograms than multiple sequence alignment [4]; these results often show greater congruence among data sets [5].

Wheeler [2] defined and illustrated OA for a simple case of short sequences, determining the length of a cladogram in terms of nucleotide transformations and indels. Here, I review the method in more detail. First the "down-pass" or initial tree length determination is described in detail, and then the "up-pass" or internal-node sequence reconstruction procedure. Since these procedures yield approximations of the minimal length cladogram (the exact solution is thought to be NP-complete), errors and approximations are introduced and their behavior and techniques for accommodating them are described.

2 Going Down to Get Tree Length

The initial step of any phylogenetic optimization procedure is the "down-pass." The procedure begins at the tips of the tree with the terminal taxa, and moves down through the hypothetical ancestral nodes to the base or root of the tree. This initial or preliminary traverse through the cladogram yields both preliminary character state assignments to the internal (i. e., ancestral) nodes and the cladogram length or cost.

Whether the characters are morphological additive or non-additive characters, unordered or matrix molecular variants [3, 6, 7], the basic operation begins by choosing an internal node whose two descendants are terminal taxa (e. g., T1 and T2; Fig. 1). The down-pass character state assignments (preliminary of node A4) are determined from the two descendants and minimize the amount of change between the ancestor and its two descendants (T1, T2 and A4). This process is repeated until all the internal nodes have been visited (A4-A1) using the relevant ancestral preliminary states as descendants for more basal ancestors (e. g., A4 and T2 for A3). The overall cost of the cladogram is the sum of the costs incurred in determining each ancestral sequence.

Optimization procedures differ in the determination of the ancestral state reconstructions and how their costs are determined. For simple binary or non-additive multistate characters, union and intersection operations are performed on the descendant character states. If the two descendants states agree (identical) or have common states (non-empty intersection), no cost is incurred and the nodal reconstructed state is the identical or overlapping state in the two descendants. If the descendant states disagree (empty intersection), the ancestral state receives the union of the descendant states and the cost of that reconstruction is increased by one (e. g., T1 {A} and T2 {C} to yield A4 {M}; Fig. 2).

The general case of multiple states, linked by arbitrary (but metric) transformation cost matrices, can be determined by trying (at least implicitly) all possible states at all internal nodes and choosing the combination which yields the most parsimonious result (Fig. 3). As with the simple method mentioned above, the cost of the determination of ancestral nodes at each node is summed to get the length of the entire cladogram for that character. For the Sankoff procedure to function, however, all possible internal states must be known and defined *a priori*, which allows them all to be considered. In most situations (e. g., nucleotide data), this is straightforward, with only five (A, C, G, T and gap) states possible. Methods have been proposed where the number of possible states can be arbitrarily large (see [8]), but these character states can still be optimized via the same process as for the five states of nucleotide data. OA generalizes this

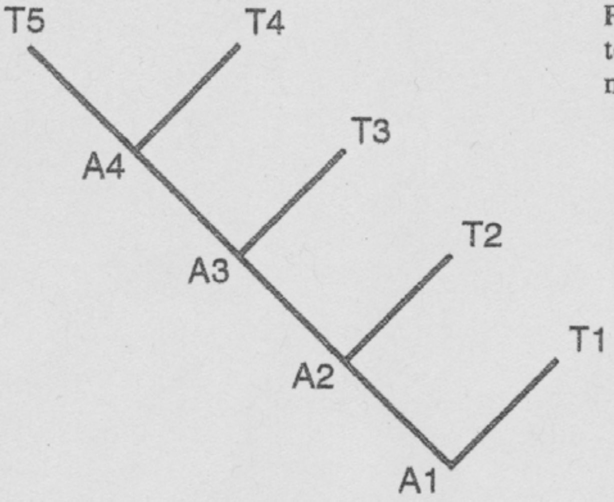


Figure 1 Example cladogram with five terminal taxa (T1-T5) and four internal nodes (A1-A4).

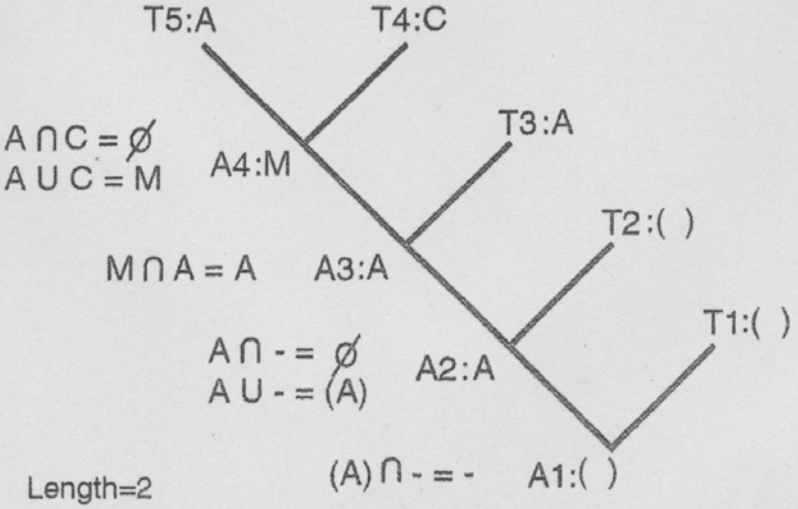


Figure 2 Fitch [7] down-pass for non-additive, or unordered DNA characters with 5 states. IUPAC codes are used to represent nucleotide ambiguity. Parentheses denote the absence of nucleotides at that position (i. e., "gaps").

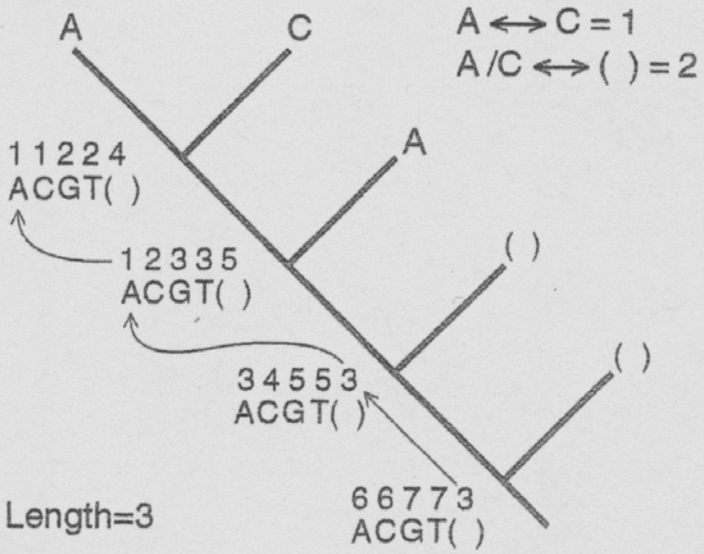


Figure 3 Sankoff and Rousseau [3] optimization for general character transformation matrices. IUPAC codes and parentheses as in figure 2.

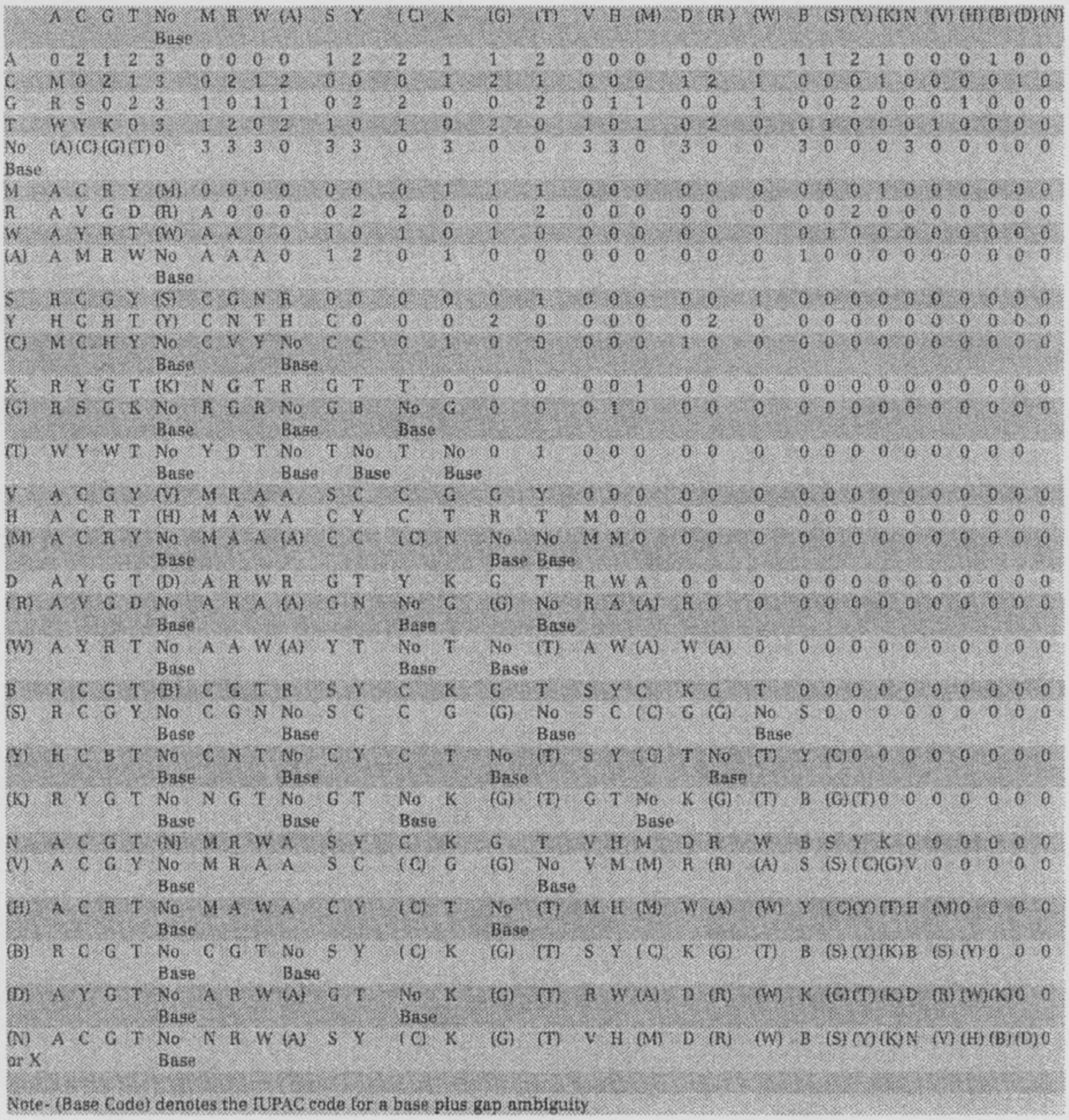
procedure by allowing the creation and destruction of characters, but employing a heuristic character optimization algorithm (Heuristic Sankoff Cost (HSC) or Sankoff procedure) to make the calculations tractable.

The Sankoff procedure, though exact, is time-consuming. If there are "s" states, "n" taxa and "m" characters, the cost of determining the length of a cladogram via the exact procedure would be proportional to $2s^2m(n-1)$ (but short-cuts exist; see [9, 10]). A non-additive character has no dependence on the number of states (i. e., would depend only on m and n), much less its square, and can be implemented with much more efficient bitwise operations. An approximate solution can be found, however, by making a simplifying assumption and performing a simple weighted version of non-additive optimization. The simplifying assumption is that we only have to worry about the immediate descendant states. If this is the case, then we can precalculate the outcome of all possible descendant state pairs. For five states (A, C, G, T, and gap) there are 31 possible combinations of states a descendant can exhibit (five single states and 26 combinations). Hence, there are only 961 possible events that can occur at an ancestral node. Furthermore, there are 31 cases where the two descendants are identical; the remainders are not order-dependent so there are only 465 calculations that are required. Each of these would result in a preliminary ancestral state assignment and a cost (the minimum cost transformation implied by the descendant states). These are calculated and stored in a 31 × 31 table with the descendant states as indices before any optimization or search takes place (Fig. 4). During optimization, the results of every ancestral optimization would just be looked up in the table (Fig. 5). The method is approximate, it ignores intermediate, locally sub-optimal solutions that might be globally more parsimonious later, but can reduce execution time considerably. Solutions can be checked by performing a complete Sankoff down-pass on those rare occasions during a search where a candidate tree is thought to be equally or more parsimonious than the current best [10]. This heuristic procedure was first used in MALIGN [11, 12] and is used in PHAST [10] and POY [13].

The OA procedure relies on a combination of HSC and the Needleman and Wusch (NW) pairwise alignment procedure [14]. The three types of optimization discussed above (Farris, Fitch, and Sankoff), assume that the descendant characters to be compared or optimized are known. In other words, previous optimization schemes assume that it is known which "A" on one sequence corresponds to which "C" in another. This is generally not the case with nucleotide sequence data sets, since the sequences of terminal taxa can vary in length. In other words, these optimization procedures require pre-aligned sequences. However, if a means can be found to create parsimonious preliminary ancestral sequences from two descendant sequences and determine the cost of that creation, the coupling of homology assessment and testing can be made seamless rendering multiple alignment unnecessary.

Determination of preliminary ancestral sequences is approached in the same way that Sankoff optimization looks at all the possible state assignments and chooses the most parsimonious, such that OA looks at all the potential correspondences between the nucleotides in the descendant sequences to determine which scheme of correspondences and transformations yields the most parsimonious preliminary ancestral sequence. This is done in a manner akin to the pair-wise alignment procedure of NW. In this case, the NW procedure is modified from one that maximizes sequence similarity to one that minimizes the cost of the ancestral sequence as approximated by the HSC.

Figure 4 Look-up table including hypothetical ancestral states and cost used in the Heuristic Sankoff Cost (HSC) procedure. IUPAC codes and parentheses as in Figure 2. An IUPAC code with parentheses denotes ambiguity with respect to that IUPAC nucleotide, or ambiguity, and the absence of base or "gap." These costs and states are based on a cost of 3 for indels, 2 for transversions and 1 for transitions.



	A	C	G	T	No Base	M	R	W	(A)	S	Y	(C)	K	(G)	(T)	V	H	(M)	D	(R)	(W)	B	(S)	(Y)	(K)	N	(V)	(H)	(B)	(D)	(N)
A	0	2	1	2	3	0	0	0	0	1	2	2	1	1	2	0	0	0	0	0	0	1	1	2	1	0	0	0	1	0	0
C	M	0	2	1	3	0	2	1	2	0	0	0	1	2	1	0	0	0	1	2	1	0	0	0	1	0	0	0	0	1	0
G	R	S	0	2	3	1	0	1	1	0	2	2	0	0	2	0	1	1	0	0	1	0	0	2	0	0	0	1	0	0	0
T	W	Y	K	0	3	1	2	0	2	1	0	1	0	2	0	1	0	1	0	2	0	0	1	0	0	0	1	0	0	0	0
No Base	(A)	(C)	(G)	(T)	0	3	3	3	0	3	3	0	3	0	0	3	3	0	3	0	0	3	0	0	0	3	0	0	0	0	0
M	A	C	R	Y	(M)	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
R	A	V	G	D	(R)	A	0	0	0	0	2	2	0	0	2	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
W	A	Y	R	T	(W)	A	A	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
(A)	A	M	R	W	No Base	A	A	A	0	1	2	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
S	R	C	G	Y	(S)	C	G	N	R	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Y	H	C	H	T	(Y)	C	N	T	H	C	0	0	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
(C)	M	C	H	Y	No Base	C	V	Y	No Base	C	C	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
K	R	Y	G	T	(K)	N	G	T	R	G	T	T	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
(G)	R	S	G	K	No Base	R	G	R	No Base	G	B	No Base	G	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(T)	W	Y	W	T	No Base	Y	D	T	No Base	T	No Base	T	No Base	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
V	A	C	G	Y	(V)	M	R	A	A	S	C	C	G	G	Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	A	C	R	T	(H)	M	A	W	A	C	Y	C	T	R	T	M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(M)	A	C	R	Y	No Base	M	A	A	(A)	C	C	(C)	N	No Base	No Base	M	M	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	A	Y	G	T	(D)	A	R	W	R	G	T	Y	K	G	T	R	W	A	0	0	0	0	0	0	0	0	0	0	0	0	0
(R)	A	V	G	D	No Base	A	R	A	(A)	G	N	No Base	G	(G)	No Base	R	A	(A)	R	0	0	0	0	0	0	0	0	0	0	0	0
(W)	A	Y	R	T	No Base	A	A	W	(A)	Y	T	No Base	T	No Base	(T)	A	W	(A)	W	(A)	0	0	0	0	0	0	0	0	0	0	0
B	R	C	G	T	(B)	C	G	T	R	S	Y	C	K	G	T	S	Y	C	K	G	T	0	0	0	0	0	0	0	0	0	0
(S)	R	C	G	Y	No Base	C	G	N	No Base	S	C	C	G	(G)	No Base	S	C	(C)	G	(G)	No Base	S	0	0	0	0	0	0	0	0	0
(Y)	H	C	B	T	No Base	C	N	T	No Base	C	Y	C	T	No Base	(T)	S	Y	(C)	T	No Base	(T)	Y	(C)	0	0	0	0	0	0	0	0
(K)	R	Y	G	T	No Base	N	G	T	No Base	G	T	No Base	K	(G)	(T)	G	T	No Base	K	(G)	(T)	B	(G)	(T)	0	0	0	0	0	0	0
N	A	C	G	T	(N)	M	R	W	A	S	Y	C	K	G	T	V	H	M	D	R	W	B	S	Y	K	0	0	0	0	0	0
(V)	A	C	G	Y	No Base	M	R	A	A	S	C	(C)	G	(G)	No Base	V	M	(M)	R	(R)	(A)	S	(S)	(C)	(G)	V	0	0	0	0	0
(H)	A	C	R	T	No Base	M	A	W	A	C	Y	(C)	T	No Base	(T)	M	H	(M)	W	(A)	(W)	Y	(C)	(Y)	(T)	H	(M)	0	0	0	0
(B)	R	C	G	T	No Base	C	G	T	No Base	S	Y	(C)	K	(G)	(T)	S	Y	(C)	K	(G)	(T)	B	(S)	(Y)	(K)	B	(S)	(Y)	0	0	0
(D)	A	Y	G	T	No Base	A	R	W	(A)	G	T	No Base	K	(G)	(T)	R	W	(A)	D	(R)	(W)	K	(G)	(T)	(K)	D	(R)	(W)	(K)	0	0
(N) or X	A	C	G	T	No Base	N	R	W	(A)	S	Y	(C)	K	(G)	(T)	V	H	(M)	D	(R)	(W)	B	(S)	(Y)	(K)	N	(V)	(H)	(B)	(D)	0

Note- (Base Code) denotes the IUPAC code for a base plus gap ambiguity

Consider four sequences T1:"AA", T2:"A", T3:"GG" and T4:"G". T1 is defined *a priori* as the outgroup and the indel cost set to 2 and the base change cost to 1 (transition cost = transversion cost = 1). At least initially, assume a candidate tree (T1 (T2 (T3 T4))) (Fig. 6). As mentioned above, the process starts with a node, both of whose descendants are terminals. Here, that is the node with descendants T3 and T4 (A3). In order to determine the lowest cost preliminary hypothetical ancestral sequence, a NW-type procedure is performed with a cost matrix based on the analysis parameters mentioned above (Fig. 7). The NW

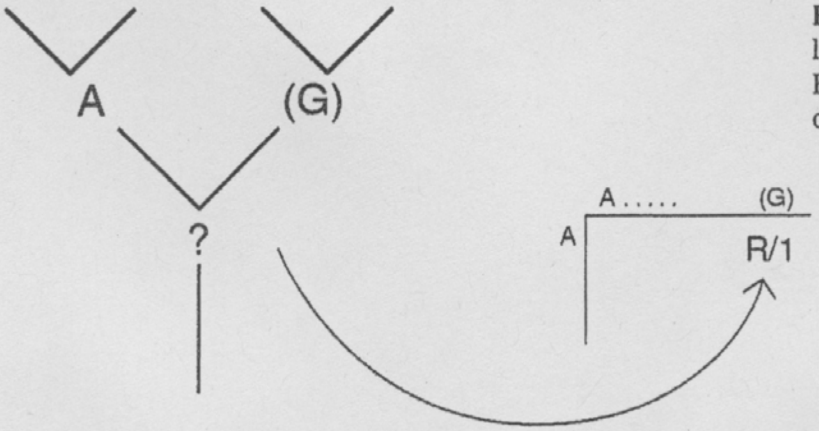


Figure 5 The use of the look-up table (Fig. 4) in the Heuristic Sankoff Cost procedure.

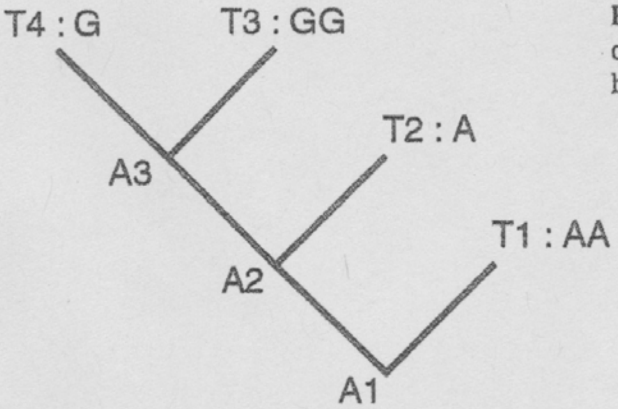


Figure 6 A simple example of four sequences of one to two bases, each related by a cladogram.

	A	C	G	T	( )
A	0	1	1	1	2
C	1	0	1	1	2
G	1	1	0	1	2
T	1	1	1	0	2
( )	2	2	2	2	0

Figure 7 Cost matrix of transformations used to diagnose the cladogram and sequences of Figure 6.

procedure minimizes the cost (in this case) of the nodal sequence by implicitly determining the cost of all possible preliminary reconstructions through a dynamic programming procedure. A matrix is set up and updated via "wavefront" optimization [14–16 and references therein]. This case requires a six cell node (n+1 by m+1, where n and m are the lengths of the descendant sequences) to consider the five possible homology schemes (Fig. 8). The result of this procedure is that there are two possible reconstructions of cost 2 (a single indel) using the HSC. The HSC is used not only to determine the cost of the reconstruction but the state as well. In this simple case, the preliminary ancestral sequence would be ambiguous as to the length (one or two bases) but one of those bases would be a "G" and the ambiguous length would be either a "G" or nothing. The "G" or nothing, represented as (G), is the result of the

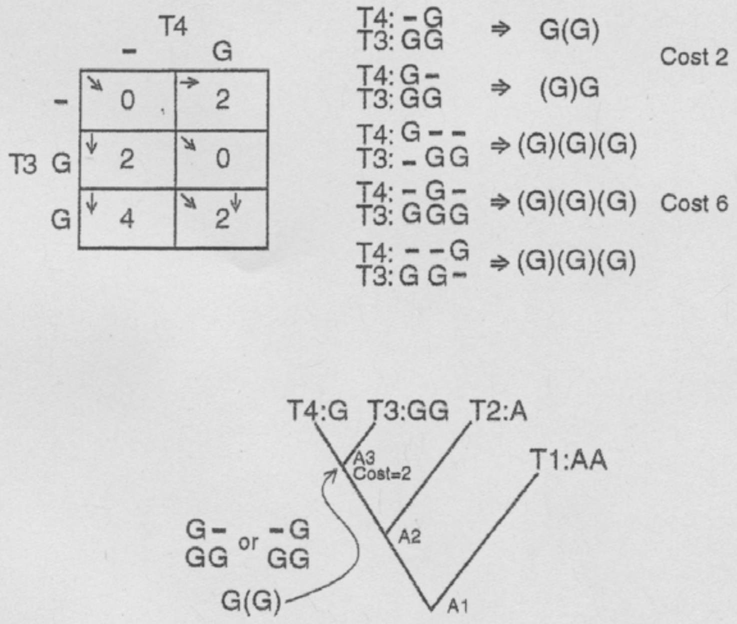


Figure 8 The determination of the preliminary sequence for node A3 of the cladogram in Figure 6. There are five possible paths through the matrix and five possible preliminary ancestral sequences.

lowest cost union (via HSC) between the corresponding descendant states of "G" and nothing (a gap if this were an alignment). An exact solution would require that we follow both of these possibilities (and all their multiplicative derivatives), but in the current implementation and description of the method a single preliminary hypothetical ancestral sequence is chosen.

After this node, the process is repeated for all the other unoptimized nodes. The node (A2) has descendants A3 and T2. The process is repeated as above with the descendant sequences (G)G and A. In this case, preliminary node reconstruction would yield "R" ambiguous with respect to A and G, but unambiguous as to sequence length (1) (Fig. 9). The NW and HSC yield a cost of 1 for this node (3 so far after A3 and A2 have been visited) and the ambiguity of length in A2 is resolved since the length of T1 was also 1. The final node (root node) is determined by comparing its descendants A1 ("R") and T1 ("AA"). Following the same NW-HSC process yields an ambiguous preliminary root node assignment of "A(A)" and "(A)A" (there were two paths to get there) with a local cost of 2 and a total cladogram cost of 5 steps (Fig. 10).

This completes the down-pass. Here the total cost (based on an indel cost of two and base change cost of one) was five with two indels and a single base change. In a search, other cladograms would be optimized and another solution of equal cost would have been found. For example, the cladogram (T1 (T3 (T2 T4))) also requires 5 steps, but with three base changes and a single indel.

An obvious conclusion of this dependence on the indel and base change costs is that the preferred (i. e., most parsimonious) cladogram may vary with the parameter values. For this example if the indel cost is increased by one to three, only the cladogram linking T2 and T4 is chosen (length 6), minimizing the now more costly indels. Alternatively, if the gaps cost is reduced by one to one, the cladogram linking T3 and T4 is favored (length 3), minimizing base changes.

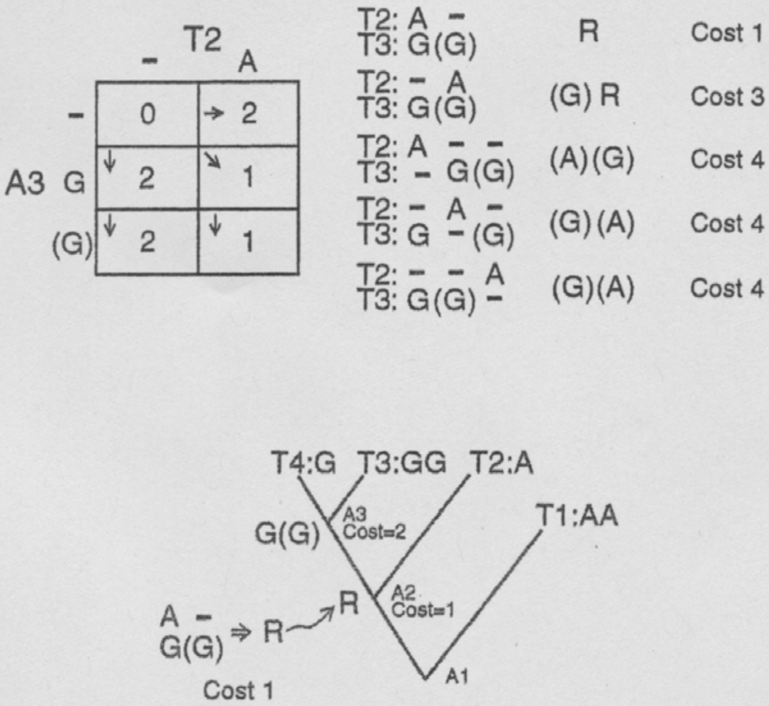


Figure 9 The determination of the preliminary sequence for node A2 of the cladogram in Figure 6. There are five possible paths through the matrix and five possible preliminary ancestral sequences.

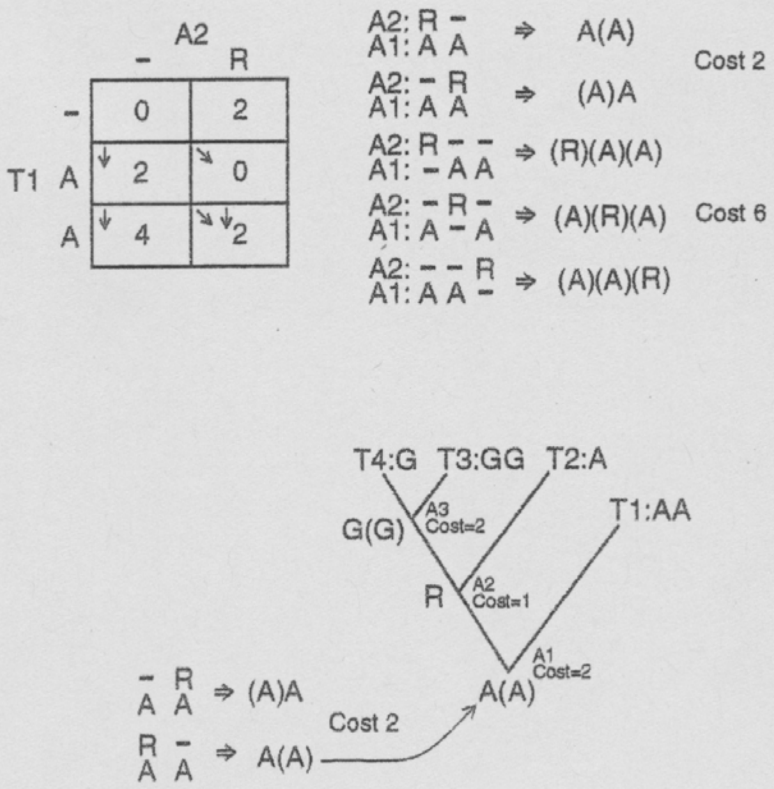


Figure 10 The determination of the preliminary sequence for node A1 of the cladogram in Figure 6. There are five possible paths through the matrix and five possible preliminary ancestral sequences.

3 Going Up to Get Ancestral States

In order to reconstruct a parsimonious set of character states at the internal (hypothetical) nodes, a second or up-pass is required. This process moves from the root of the cladogram "up" to the tips, incorporating the information from a node's ancestor as well as its descendants. As originally described [2] and implemented in POY [13], the process is extremely simple, basically trying all possible states (based on the down-pass homologies) in turn and keeping those with the lowest cost.

More specifically, the starting point is the root node. There is no true up-pass for this node, since it has no ancestor. The final states for the root node are simply assigned from the preliminary states or the final states of the outgroup taxon. The descendants of this node are then visited. These nodes are the first ones with both descendants and ancestors. The preliminary homologies among the preliminary down-pass states and the two descendant sequences are known (saved) from the down-pass step and the correspondences with the final states of its ancestor are determined by the same NW-HSC process used in the down-pass, but between the preliminary states of the node and the final states of its ancestor. For each position then, the two descendant and ancestral states are known. Each state is tried in turn as the final state, and the most parsimonious set taken as the final state set for that node (Fig. 11). The process then moves on to the next node and so on until final state sets are determined for each non-terminal node. The final state sets for the terminals and the root node are of course identical to the preliminary states.

Given that the final states are based on the same "greedy" short-sighted simplifications of the down-pass (locally lowest cost reconstructions), the final state sets are approximations as well. In addition to erroneously estimating

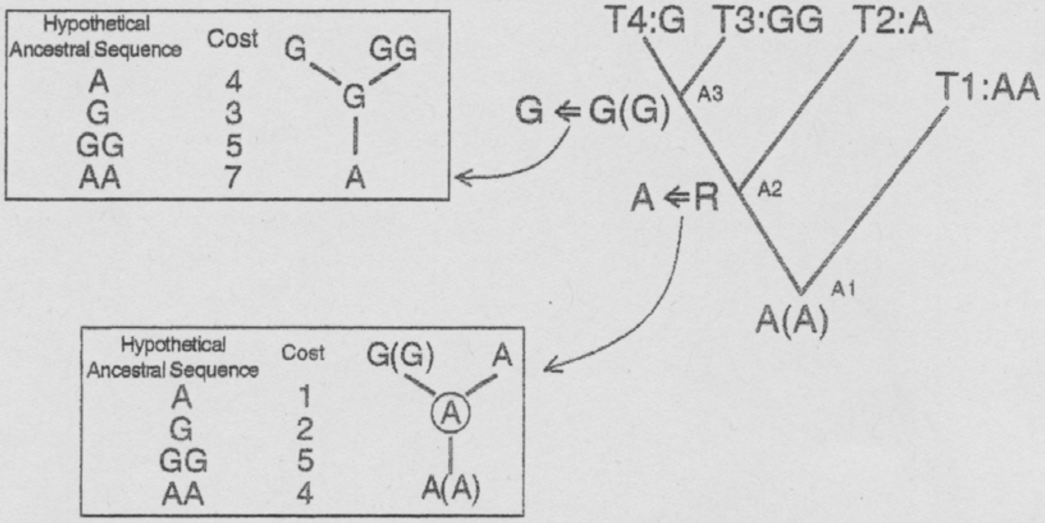


Figure 11 The determination of the final hypothetical ancestral sequences *via* an up-pass.

hypothetical ancestral sequences, this can cause problems with many phylogenetic search shortcuts, which rely on these hypothetical ancestral sequences as a surrogate for the information within their descendant clade.

4 Short-cuts and Errors

This form of sequence optimization makes two sorts of errors. The first concerns the down-pass tree lengths. Since the method determines local node costs based only on the descendant sequence information, the estimated cost must be equal to, or more likely greater than the minimal cost (Fig. 12). This effect can be compounded by the fact that the optimization regime is simultaneously determining homology relationships among the nucleotides as well from this same restricted (only descendant) sequence set. These two factors together ensure that the down-pass tree length is an upper bound on the minimal cost.

The second source of error comes from the establishment of the set of final (up-pass) states for the hypothetical ancestral sequences. Since the preliminary (down-pass) sequences are constructed in a myopic manner, these errors are carried along into the up-pass phase. This can result in both the inclusion of nucleotide states that should not be there, as well as missing those nucleotide states that should be there. When these reconstructions are used as part of short-cut methods during phylogenetic searches, this can cause the short-cuts to over- or underestimate candidate tree lengths. In general, these errors are

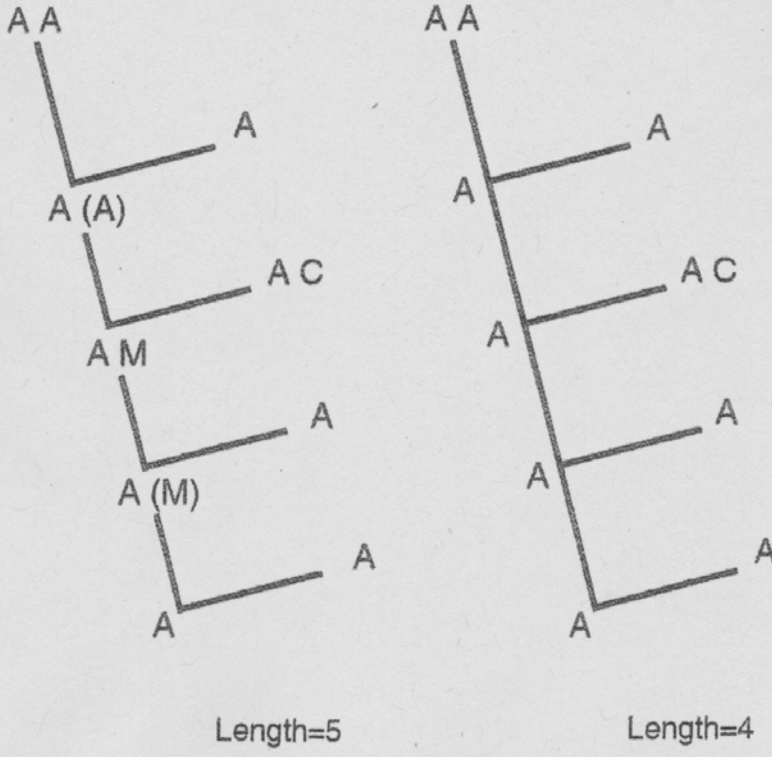


Figure 12 The down-pass calculation of length (left) showing how the value can be overestimated compared to the actual minimum length (right).

not large (less than 1%) and can be accommodated by checking tree length calculations with full down-pass optimization on candidate trees. In POY, the options "slop" and "checkslop" allow the verification of trees within a specified difference in tree length from the current best. Although this slows things down, it increases confidence that the search is proceeding on verified shortest trees.

Another expression of this effect is the apparent dependence of cladogram length on root position. This would seem to be counterintuitive, given that the parameters (indel cost, transition-transversion ratio, etc.) are symmetrical. As mentioned before [2], this is also due to the heuristic tree length procedures. In the example of Figure 12, if rooted at sequence "AA" instead of sequence "A", the minimal tree length of 4 is produced (Fig. 13).

5 Improvements

Many improvements could be made to these procedures. Three general classes would involve multiple solutions, sub-optimal solutions and character-specific virtual roots.

5.1 Multiple solutions

During the down-pass, the consideration of multiple equally parsimonious preliminary sequences would be an obvious step. Since each node would be likely to generate its own multiple solutions which should be multiplied down the cladogram, extremely large potential sets of preliminary sequences could be generated. Although it might not be practical to maintain multiple solutions throughout the entire down-pass, these solutions (or a set of them) could be maintained and considered at the optimization of the next (parent) node (or some other range down the cladogram). The NW-HSC process would be

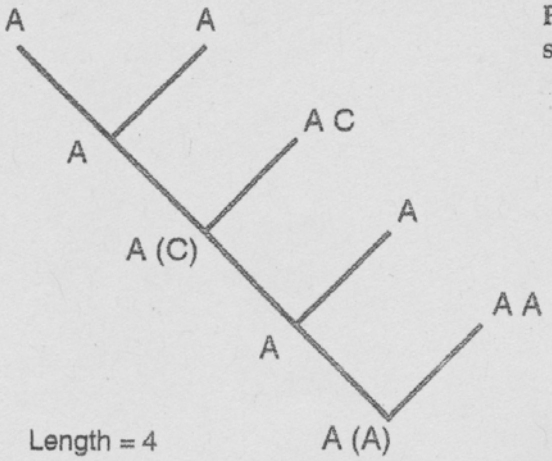


Figure 13 Rerooted diagnosis of Figure 13 showing the achievement of minimum length.

performed on each combination of the candidate preliminary sequences for the two descendants of a node. This would generate a new set of equally costly preliminary sequences for this node and the process would be repeated until the entire cladogram was optimized.

The up-pass determination of the final hypothetical ancestral sequences could also be improved by maintaining multiple equal cost solutions. Here would be the product of the multiple preliminary solutions and those of the final assignment of the parent node, which would be more computationally intensive, but surely tractable. The same decisions would be required on the size of the set of solutions to be maintained (this could be large, or a smaller random sample could be stored) and the reach over which to hold and test these multiple solutions.

5.2 Sub-optimal solutions

One of the principal reasons for the myopia of the down-pass (and up-pass) methods is that they ignore locally sub-optimal solutions which might turn out to be globally optimal. In the example of Figure 12, the assignment of preliminary sequence "A" is globally optimal, but ignored as too expensive initially. If some set of sub-optimal solutions were considered, they might prove useful in subsequent optimization stages. Unfortunately, there are many sub-optimal solutions and many of them are sub-optimal for good reasons. It is unclear whether this notion would prove practicable.

5.3 Virtual roots

As the examples of Figures 12 and 13 show, rooting can affect the behavior of the down- and up-pass algorithms. Certain characters might well behave better (i. e., generate shorter cladogram lengths) given certain roots. Unfortunately, it is unlikely that these roots will be identical for all characters. Perhaps a system of "virtual" roots could be erected with each character having its own "best" root, and then optimized on that basis. The overall cladogram would then be more explicitly unrooted and presented in a rooted fashion only at the end of the analysis.

6 Remarks and Conclusions

This explicit discussion of the procedures involved in the diagnosis of sequence data on cladograms shows both the strengths and weaknesses of this approach. Not requiring *a priori* sequence alignment and generating cladogram-specific homology schemes would seem to be strengths. The heuristic nature of the cladogram length and ancestral sequence reconstruction would seem to be weaknesses. These can be improved, however, as described above. Although the problem is unlikely to be solved exactly, improvements along the lines suggested here could well bring incremental benefits and, combined with ideas of others, generate more satisfactory methods and more reliable results.

Acknowledgments

I would like to acknowledge and thank the following for help with this manuscript: James Carpenter, Rob DeSalle, Rasmus Hövmoller, Gonzalo Giribet, Pablo Goloboff, Matthew Gruell, Daniel Janies, Susanne Schulmeister and Steven Thurston.

References

1 Sankoff DD, Cedergren RJ (1983) Simultaneous comparison of three or more sequences related by a tree. In D Sankoff, JB Kruskal (eds): *Time Warps, String Edits, and Macromolecules: the Theory and Practise of Sequence Comparison.* Addison-Wesley, Reading, MA, 253–264

2 Wheeler WC (1996) Optimization Alignment: the end of multiple sequence alignment in phylogenetics? *Cladistics* 12: 1–9

3 Sankoff DD, Rousseau P (1975) Locating the vertices of a Steiner tree in arbitrary space. *Math. Prog.* 9: 240–246

4 Wheeler WC (2000) Heuristic Reconstruction of Hypothetical-Ancestral DNA Sequences: Sequence Alignment versus Direct Optimization. In: R Scotland, RT Pennington (eds) *Homology and Systematics.* Systematics Society, London, 217 106–113

5 Giribet G (2001) Exploring the behavoir of POY; a program for direct optimization of molecular data. *Cladistics* 17: 560–570

6 Farris JS (1970) Methods for computing Wagner trees. *Syst. Zool.* 19: 83–92

7 Fitch WM (1971) Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.* 20: 406–416

8 Wheeler WC (1999) Fixed Character States and the Optimization of Molecular Sequence Data. *Cladistics* 15: 379–385

9 Goloboff PA (1994) Character optimization and calculation of tree lengths. *Cladistics* 9: 433–436

10 Goloboff PA (1998) Tree searches under Sankoff parsimony. *Cladistics* 14: 229–237

11 Wheeler WC, Gladstein DS (1994) MALIGN: A multiple sequence alignment program. *J. Hered.* 85: 417–418

12 Wheeler WC, Gladstein DS (1991–1998), *Malign. Program and documentation.* New York, NY. Documentation by Daniel Janies and Ward Wheeler

13 Gladstein DS, Wheeler WC (1997) "*POY: The Optimization of Alignment Characters." Program and Documentation.* New York, NY. Available at "ftp.amnh.org" / pub/molecular

14 Needleman SB, Wunsch CD (1970) A general method applicable to the search

for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48: 443–453

15 Kruskal JB and Sankoff D (1983) An anthology of algorithms and concepts for sequence comparison. In: D Sankoff and JB Kruskal (eds) *Time Warps, String Edits, and Macromolecules: the Theory and Practise of Sequence Comparison.* Addison-Wesley, Reading, MA, 265–310

16 Phillips A, Janies D and Wheeler WC (2000) Multiple sequence alignment in phylogenetic analysis. *Mol. Phyl. Evol.* in press