Cladistics **13**,153–159 (1997)
WWW http://www.apnet.com

Forum

Uneconomical Diagnosis of Cladograms: Comments on Wheeler and Nixon's Method for Sankoff Optimization

David L. Swofford[1] and Mark E. Siddall[2]

[1]*Laboratory of Molecular Systematics, MRC-534, Smithsonian Institution, Washington, DC|20560, U.S.A., and*
[2]*Museum of Zoology, University of Michigan, Ann Arbor, Michigan|48109, U.S.A.*

Accepted 19 March 1997

We provide three simple examples demonstrating that Wheeler and Nixon's method of recoding "stepmatrix" characters can fail to yield most parsimonious reconstructions of character evolution under specified cost (transformation-weight) schemes. These examples variously indicate undercounting or overcounting of tree lengths due to an inappropriate assumption of independence among the recoded characters. Their method is therefore not equivalent to Sankoff's dynamic programming algorithm, contrary to their claim. © 1997 The Willi Hennig Society

INTRODUCTION

Sankoff and coworkers (Sankoff and Rousseau, 1975; Sankoff et al., 1976) described a dynamic programming algorithm for optimizing the length of a tree under the parsimony criterion that generalized earlier methods by allowing arbitrary costs for changes between character states (see Swofford and Maddison (1992) for an introductory presentation of this algorithm). This method is sometimes referred to as "generalized parsimony" or "Sankoff optimization" and the matrix specifying the cost associated with each possible character-state change is called a "stepmatrix" or a "cost matrix". The most common application of the method is in the analysis of nucleotide sequence data wherein different relative costs are assigned to transversions, transitions, and (less frequently) insertion-deletion events (indels).

Sankoff's algorithm finds an exact solution to the problem of minimizing the length required by a particular tree under general cost schemes, but requires much more computational effort than the algorithms developed by Farris (1970) and Fitch (1971) for restricted cases. Recently, Wheeler and Nixon (1994) described a method that partitions transformations into seemingly non-redundant binary and unordered multistate characters in an attempt to obviate the use of Sankoff's algorithm as implemented in the software packages MacClade (Maddison and Maddison, 1987, 1992) and PAUP (Swofford, 1993). Wheeler and Nixon stated that their method "can accommodate the vast majority of cases in which these Sankoff characters are required" and that "the only effect of the recoding procedure is to accelerate the search for parsimonious solutions. The length, number and topologies of these solutions are unchanged". Wheeler and Nixon provided no proof for these claims; in this note, we show

0748-3007/97/010153+07/$25.00/0/cl970037

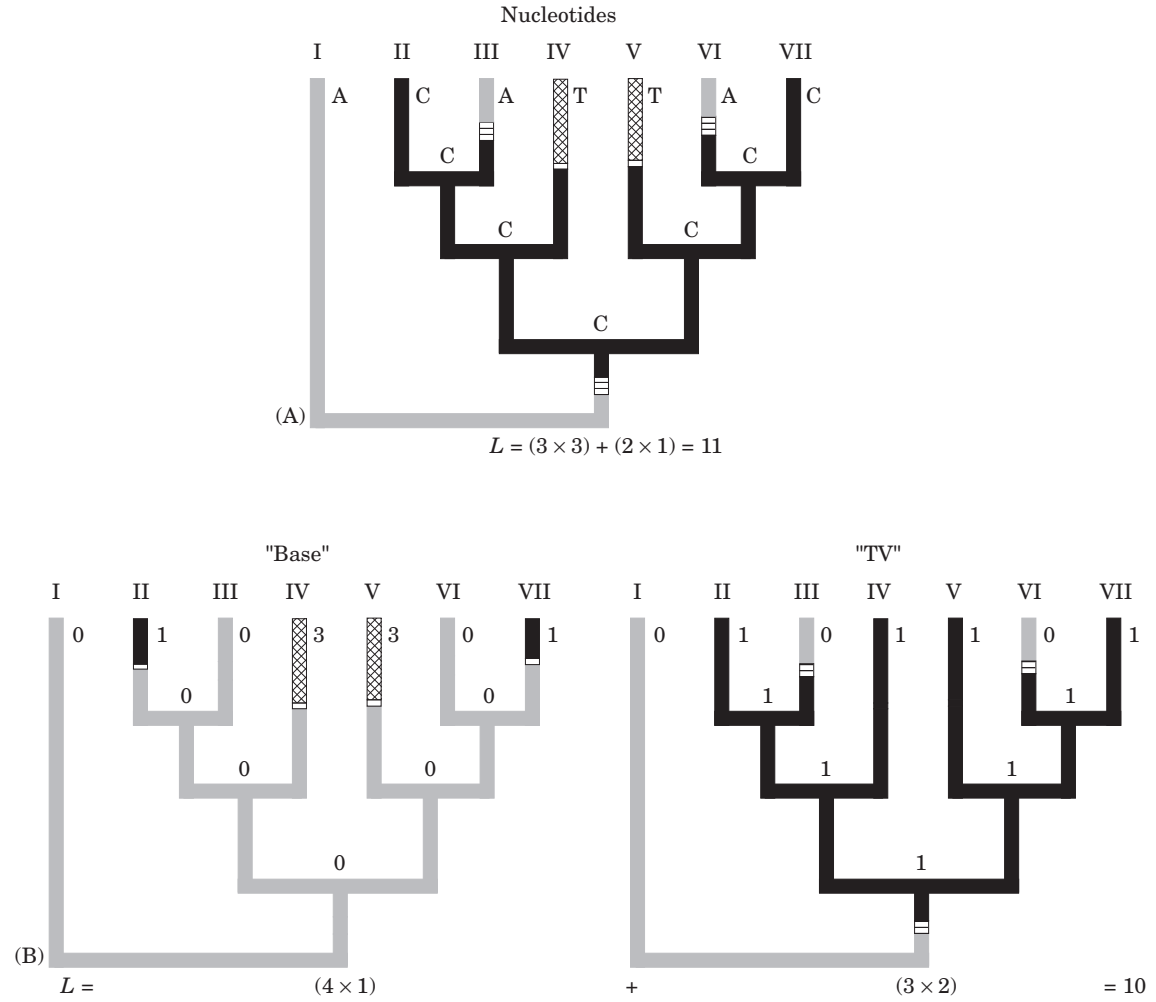

FIG. 1. Example showing nonequivalence of Sankoff optimization and Wheeler–Nixon recoding method (transversions and transitions are assigned costs of 3 and 1 respectively). (A) One most-parsimonious Sankoff optimization requiring three transversions and two transitions, (B) Wheeler and Nixon's method implies four transversions in the "base" character and three transversions in the "TV" character (and no transitions).

three simple counterexamples that demonstrate their invalidity.

For the simplest case, we assign a cost of 1 to transitions (changes between two purines or two pyrimidines) and a greater cost (v) to transversions (changes between a purine and a pyrimidine). The corresponding cost matrix is:

	A	C	G	T
A	0	v	1	v
C	v	0	v	1
G	1	v	0	v
T	v	1	v	0

Wheeler and Nixon's coding strategy for this situation reduces to the following:

State	Base	TV
A	0	0
C	1	1
G	2	0
T	3	1
Weight	1	v-1

That is, an additional binary character (TV) is created that takes one state (0) for purine nucleotides and the other state (1) for pyrimidine nucleotides. The "base"

(A)

Taxon	12	345	678	11 901	111 234	1 5
A	CC	GGG	GGG	AAA	AAA	G
B	AA	GGG	GGG	AAA	AAA	G
C	TT	AAA	GGG	AAA	AAA	G
D	TT	AAA	AAA	AAA	GGG	G
E	AA	AAA	AAA	GGG	GGG	A
F	CC	AAA	AAA	GGG	GGG	A
G	AA	AAA	AAA	AAA	AAA	A

(B)

Taxon	1v	2v	345	678	11 901	111 234	1 5
A	11	11	222	222	000	000	2
B	00	00	222	222	000	000	2
C	31	31	000	222	000	000	2
D	31	31	000	000	000	222	2
E	00	00	000	000	222	222	0
F	11	11	000	000	222	222	0
G	00	00	000	000	000	000	0
(Weights)	12	12	111	111	111	111	1

(C)

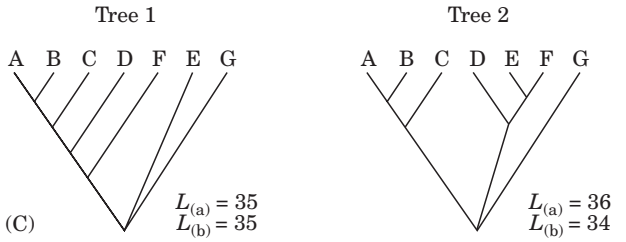

FIG. 2. Example showing that Wheeler and Nixon's coding method can lead to selection of an incorrect tree when transversions are assigned a cost of three times that of transitions. (A) Aligned nucleotide sequences. (B) Recoded matrix and weights (constant characters resulting from recoding of characters 3–15 are not shown). (C) Trees resulting from analysis of these data sets. Note that Tree 1 is more parsimonious than Tree 2 for the original data, whereas Tree 2 is more parsimonious for the recoded data set.

character is treated as an ordinary unordered multistate character. These characters are optimized using a fast algorithm such as that of Fitch (1971). (The transformation of A, C, G, and T to integer values is necessary for Hennig86 [Farris, 1988], which does not accept alphabetical states.) In many situations, this procedure produces the desired results, because transitions require a single step for the base character (with no steps for the purine-pyrimidine character), whereas transversions require a total cost of v (one step for the base character plus v-1 steps for the purine-pyrimidine change).

However, it is easy to find cases for which the Wheeler–Nixon method fails. An example of this failure is shown in Fig. 1, where the tips of the tree are labeled with the nucleotides observed for seven taxa (Fig. 1A). If we let v=3, a most-parsimonious reconstruction for the data requires three transversions and two transitions, for a total of 11 steps (Fig. 1A). If we recode this character into two characters as suggested by Wheeler and Nixon, only 10 steps are required: four steps for the "base" character plus six steps for the purine-pyrimidine character (three changes times a weight of two) (Fig. 1B). Thus, Wheeler and Nixon's claim that their method always yields tree lengths identical to those obtained from Sankoff's algorithm is incorrect. Note that independent optimization of the two recoded characters yields reconstructions that are inconsistent (Fig. 1B); the base character implies state "A" (base=0) at each internal node, but the purine-pyrimidine character implies a pyrimidine (TV=1; C or T) at all of these nodes. It is this inappropriate treatment of the two recoded characters as if they are independent that leads to undercounting of the number of steps required according to the assumed cost matrix.

This single-character example can be extended to full data sets for which use of the Wheeler–Nixon method leads to selection of an incorrect tree; one such example is shown in Fig. 2. Again using a transversion cost (v) of three, the correct length of Tree 1 is 35 steps, equal to the length obtained under the Wheeler–Nixon coding. This tree is the single most parsimonious tree, as determined using the branch-and-bound algorithm of PAUP. However, this tree is less parsimonious than Tree 2 under the Wheeler–Nixon coding (35 vs 34 steps) due to undercounting of the number of changes required by characters 1 and 2, which correspond to the Fig. 1 example. Under correct Sankoff optimization, Tree 2 is less parsimonious (36 steps) than Tree 1. Wheeler and Nixon's conclusion that the topologies of solutions obtained using their coding method are always equivalent to those from Sankoff's algorithm is therefore also incorrect.

Wheeler and Nixon also suggested that their method was more generally applicable, including the ability to handle cases such as different costs for the two types of transitions (A↔G vs C↔T), indicated by the following cost matrix:

	A	C	G	T
A	0	v	i_{ag}	v
C	v	0	v	i_{ct}
G	i_{ag}	v	0	v
T	v	i_{ct}	v	0

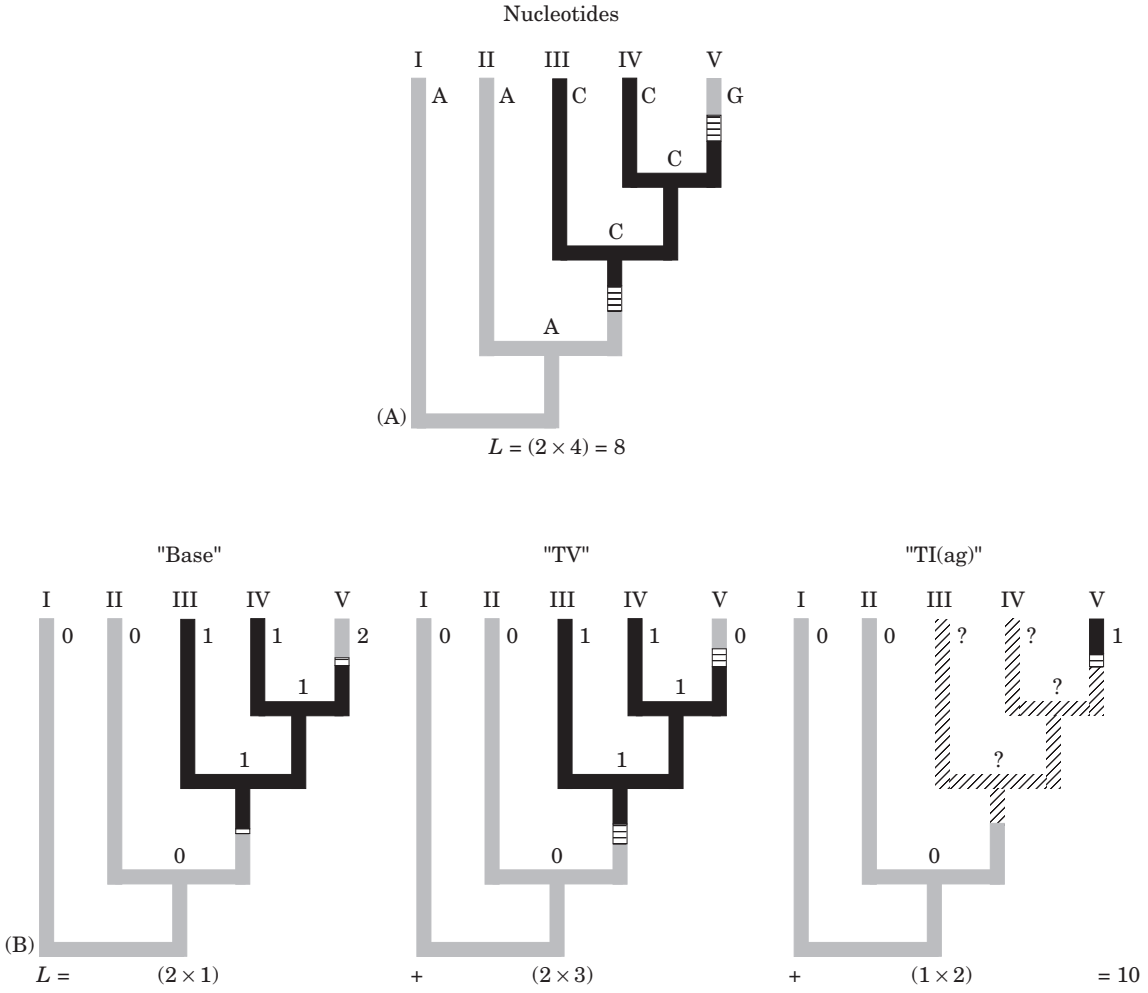

FIG. 3. Example showing nonequivalence of Sankoff optimization and Wheeler–Nixon recoding method when the two kinds of transitions are assigned different costs (transversions, purine transitions, and pyrimidine transitions are assigned costs of 4, 3, and 1, respectively). (A) Most parsimonious reconstruction requires two transversions. (B) Wheeler and Nixon's method requires an additional two steps for a transition in the TI(ag) character that is not needed to explain the data.

In this case, their coding method can lead to overcounting the number of steps (unlike the preceding one in which steps were undercounted). To allow for unequal costs for transversions (v), transitions between purines (i_{ag}), and transitions between pyrimidines (i_{ct}), each character is recoded as follows:

State	Base	TV	TI(ag)	TI(ct)
A	0	0	0	?
C	1	1	?	0
G	2	0	1	?
T	3	1	?	1
Weight	1	v-1	i_{ag}-1	i_{ct}-1

In the example shown in Fig. 3, we set v=4, i_{ag}=3, and i_{ct}=1. Correct Sankoff optimization dictates two transversions (Fig. 3A) for a total length of eight steps. Under the Wheeler–Nixon coding, these two transversions are accommodated by the base and TV characters, but an additional length of two steps is contributed by the TI(ag) character, resulting in an overcounting of the number of steps. As for the first example, incorrect calculation of the tree length can lead to selection of an unparsimonious tree under the intended cost scheme (Fig. 4).

Sankoff's method can also be used to assign different

(A)

Taxon	1234
A	ATTT
B	ATTT
C	CTTC
D	CCCC
E	GCCC

(B)

Taxon	1	*v*	*i*	234
A	0	0	0	333
B	0	0	0	333
C	1	1	?	331
D	1	1	?	111
E	2	0	1	111
(Weights)	1	3	2	111

(C)

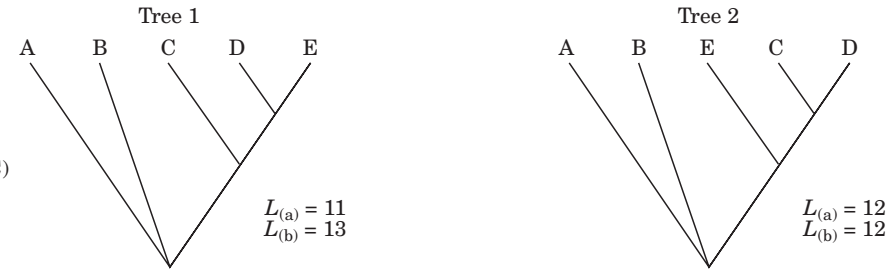

FIG. 4. Example showing that Wheeler and Nixon's coding method can lead to selection of an incorrect tree when transversions, transitions between purines, and transitions between pyrimidines are assigned different costs (4, 3, and 1, respectively). (A) Aligned nucleotide sequences. (B) Recoded matrix and weights (constant, 0-weighted characters resulting from recoding of characters are not shown). (C) Trees resulting from analysis of these data sets. Note that Tree 1 is more parsimonious than Tree 2 for the original data, whereas Tree 2 is more parsimonious for the recoded data set.

costs to indels (g) from those of base substitutions. Wheeler and Nixon recommended the following recoding for this case:

State	Base	Indel
A	0	0
C	1	0
G	2	0
T	3	0
—	4	1
Weight	1	*g*-1

Considering a single character across taxa in Fig. 5A shows that optimization of the two Wheeler–Nixon characters can lead to inconsistent reconstructions. Independent treatment of these two characters has the undesirable property of simultaneously suggesting that all hypothetical ancestors had a gap ("base" character) and yet did not have a gap ("indel" character) (Fig. 5B). As for the previous examples, the resulting incorrect length calculations can lead to selection of a suboptimal hypothesis for multicharacter data sets.

Our examples clearly indicate that Wheeler and Nixon's claim for the equivalence of their method with Sankoff's dynamic programming algorithm is false. In many cases, the Wheeler–Nixon method may yield the same tree(s) as true parsimony. Indeed, in our limited experience, it often does, but its general behavior is unknown. To the extent that it is an approximation of parsimony methods, it may find its greatest use as a device for obtaining starting trees for a true parsimony search that then uses the Sankoff algorithm for exact calculations. Unless a modification can be found that overcomes the problems we have outlined, however, we do not recommend its routine application.

ACKNOWLEDGEMENTS

We thank David Maddison for comments that improved the clarity of the manuscript.

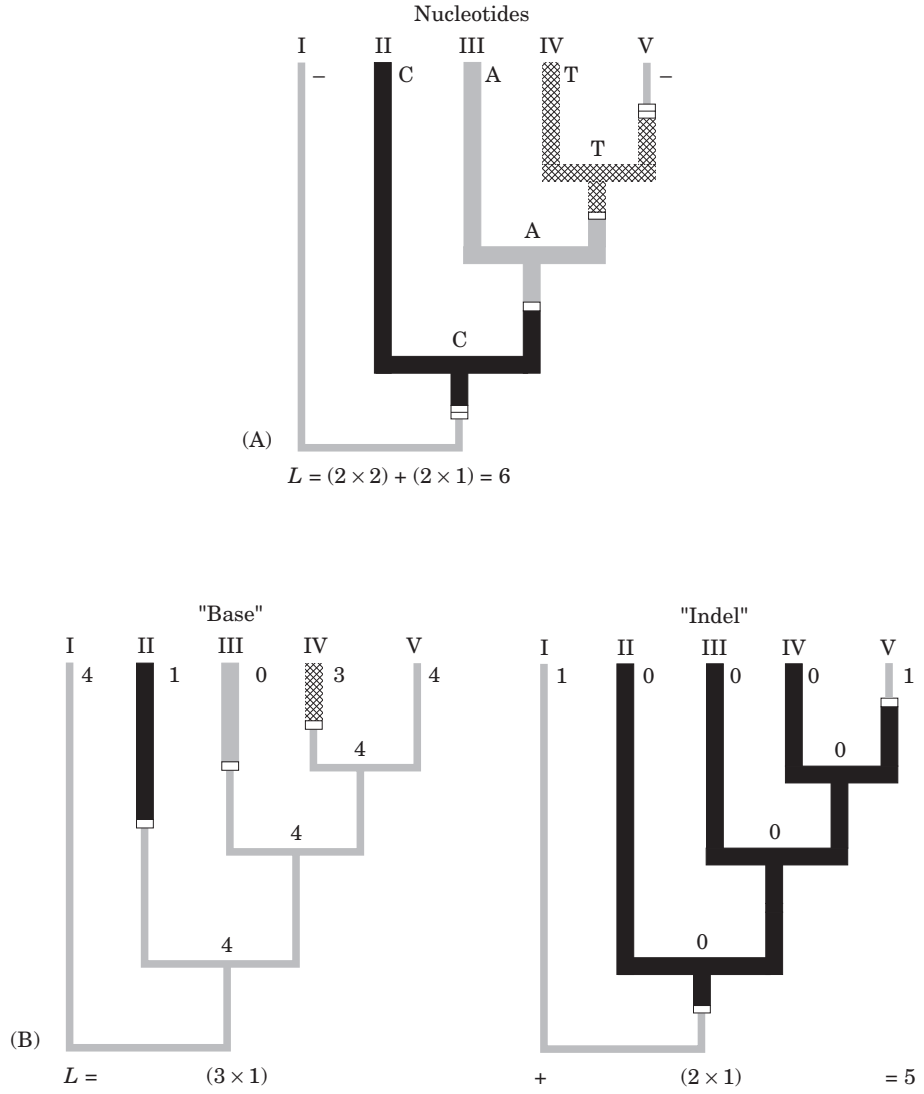

FIG. 5. Example showing nonequivalence of Sankoff optimization and Wheeler–Nixon recoding method when indels have costs exceeding 1 (indels are assigned a cost of 2 here). (A) One of the most parsimonious reconstructions requiring two insertion-deletion events. (B) Wheeler and Nixon's method yields inconsistent optimizations with all ancestors either possessing gaps ("base" character) or lacking gaps ("indel" character).

REFERENCES

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

Farris, J. S. (1988). Hennig86, version 1.5. Distributed by the author, Port Jefferson Station, N.Y.

Fitch, W. M. (1971). Toward defining the course of evolution: minimal change for a specific tree topology. *Syst. Zool.* **20**, 406–416.

Maddison, W. P., and Maddison, D. R. (1987). MacClade 2.1. Distributed by the authors, Cambridge, Massachusetts.

Maddison, W. P., and Maddison, D. R. (1992). MacClade: Analysis of phylogeny and character evolution. Version 3.0. Sinauer Associates, Sunderland, Massachusetts.

Sankoff, D., and Rousseau, P. (1975). Locating the vertices of a Steiner tree in an arbitrary metric space. *Math. Prog.* **9**, 240–246.

Sankoff, D., Cedergren, R. J., and Lapalme, G. (1976). Frequency of insertion-deletion, transversion, and transition in the evolution of 5S ribosomal RNA. *J. Mol. Evol.* **7**, 133–149.

Swofford, D. L. (1993). PAUP. Phylogenetic Analysis Using Parsimony. Version 3.1.1. Formerly distributed by Illinois Natural History Survey, Champaign, Illinois.

Swofford, D. L., and Maddison, W. P. (1992). Parsimony, character-state reconstructions, and evolutionary inferences. *In* "Systematics, Historical Ecology, and North American Freshwater Fishes" (R. L. Mayden, Ed.), pp. 186–223. Stanford University Press, Stanford.

Wheeler, W. C., and Nixon, K. (1994). A novel method for economical diagnosis of cladograms under Sankoff optimization. *Cladistics* **10**, 207–214.