Syst. Biol. 42(4):581–587, 1993

Parsimony Can Be Consistent!

MICHAEL A. STEEL,[1] MICHAEL D. HENDY,[2] AND DAVID PENNY[3]

[1]*Department of Mathematics, University of Canterbury, Christchurch, New Zealand*
[2]*Department of Mathematics, Massey University, Palmerston North, New Zealand*
[3]*School of Biological Sciences, Massey University, Palmerston North, New Zealand*

A desired property of any method for reconstructing evolutionary trees is that it be consistent, i.e., as sequences become longer the method will recover the correct tree with probability tending to 1. In an important development, Felsenstein (1978) showed that the popular parsimony criterion could, with a simple model of evolution of two-state characters, converge to an incorrect tree as the sequences became longer. This problem with parsimony also applies to the compatibility criterion for selecting optimal trees.

The problem with parsimony was originally thought to be limited to cases where lineages had markedly different rates of evolution (Felsenstein, 1978), but the problem was later found under wider sets of conditions. With five or more taxa it can be a problem, even with constant rates of evolution (Hendy and Penny, 1989; Zharkikh and Li, 1993). With six or more taxa with both constant and arbitrarily low rates of evolution (Hendy and Penny, 1989) and with a large but unspecified number of taxa, the standard parsimony criterion may fail to converge to the correct tree even when all edges of the tree have the same expected number of changes (Steel, 1989). Similar examples can be found with four-state character models, such as Kimura's 3ST model (Hendy and Charleston, 1993).

We report here that the original conclusion is too sweeping in that the prob-

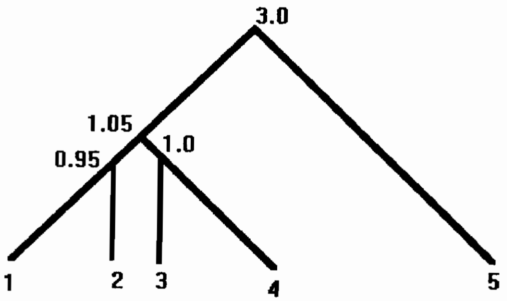

FIGURE 1. A tree of five taxa, 1, 2, 3, 4, and 5, with bifurcation times given as multiples of 10^7 years. The Cavender (1978) model for two-state character sequences with the molecular clock hypothesis assumes a constant rate of change, λ, for each site. With $\lambda = 10^{-8}$ changes per year, the expected frequencies of each of the parsimony sites for sequences of length 1,000 are given in Table 1, and the histogram of parsimony lengths are given in Figure 2. The calculations are determined using spectral analysis (Hendy and Penny, 1993).

TABLE 1. Apparent and actual support for parsimony sites.

Parsimony partition[a]	Actual support[b]	Apparent support[c]
{1, 2}, {3, 4, 5}	10	12.0
{1, 3}, {2, 4, 5}	0	6.5
{2, 3}, {1, 4, 5}	0	6.5
{1, 2, 3}, {4, 5}	0	22.3
{1, 4}, {2, 3, 5}	0	6.5
{2, 4}, {1, 3, 5}	0	6.5
{1, 2, 4}, {3, 5}	0	22.3
{3, 4}, {1, 2, 5}	5	11.0
{1, 3, 4}, {2, 5}	0	21.1
{2, 3, 4}, {1, 5}	0	21.1

[a] The 10 parsimony sites for two-state characters for five taxa.

[b] The expected numbers of changes on the edges of the tree *T* of Figure 1 for sequences of length 1,000.

[c] The expected numbers of changes observed in the sequences generated by the Cavender model from the parameters given for sequences of length 1,000 on tree *T*. The high numbers of "observed" support for some of the partitions with no actual support is a consequence of the long-edges-attract syndrome (Penny et al., 1987).

lem is not with the parsimony optimality criterion itself but rather with the implementation of the criterion. With either the two-state model of Cavender (1978) or the Kimura three-parameter model (Kimura, 1981) for four-state characters, all selection procedures, including all distance methods, are inconsistent unless they incorporate an appropriate nonlinear transformation of the data. Many criteria, including parsimony and compatibility, are consistent after appropriate nonlinear transformations that adjust for multiple hits (Penny et al., 1993).

The importance of adjustments for multiple hits is known for algorithms such as neighbor joining (Saitou and Nei, 1987), which are inconsistent when applied to observed distances but consistent when used with an appropriate correction for multiple changes (Studier and Keppler, 1988). Under Cavender's model (1978) with four taxa, neighbor joining and parsimony fail under identical conditions (Penny et al., 1992b). What appears not to have been recognized is that adjustments for multiple hits also allow parsimony and compatibility to be consistent. Corrections for multiple hits are made at the level of distances (path lengths), but the development of spectral analysis (Hendy and Penny, 1989, 1993) using discrete Fourier (Hadamard) transforms makes it a routine matter to extend this correction process to sequences.

To illustrate this point, consider the tree *T* of five taxa, 1, 2, 3, 4, and 5, as illustrated in Figure 1, with the times as shown for the internal vertices, and assume a constant rate of change λ on all edges of *T* of two-state character sequences. Thus, this tree satisfies the molecular clock hypothesis. This example is similar to many studies where the relationships among taxa 1, 2, 3, and 4 are uncertain but where taxon 5 is known to be an outgroup. Using spectral analysis, we can calculate the frequency of the 10 parsimony partitions; the results for the tree in Figure 1 are given in Table 1 for sequences of length 1,000. Given these values, we then determine the corresponding "corrected" numbers of changes supporting each of these partitions.

Using the parsimony criterion on the sequence data, we find that on the original tree *T* (Fig. 1) the expected number of duplicate changes is 112.8, whereas four other binary trees of these four taxa have only 101.6 duplicate changes and hence would have been selected ahead of *T*. In Figure 2, we give the expected lengths of the 15 binary trees for five taxa; 12 tree lengths are less than 112.8. In contrast, using the corrected data, the original tree T has no duplicate changes, whereas each of the other 14 trees has at least five changes. Thus parsimony and compatibility now select the correct tree.

Inconsistency under Any Linear Transformation

The two basic theorems reported here apply to Cavender's two-character-state model (1978) and Kimura's three-parameter model for four-character-state data (Kimura, 1981) such as nucleotides. We use the terminology of Penny et al. (1992a), including the use of a model to include three components: a tree, a mechanism of character change, and the edge lengths on the tree.

The statement of the theorem is as follows. Under either Cavender's model for two-state characters or Kimura's three-parameter model for four-state characters, even for four taxa, no method can always be consistent that (1) omits singleton sites (i.e., uses only the parsimony "informative" sites, with or without constant sites included) or (2) selects the tree that optimizes any linear function of all sites.

Proofs (given in the Appendix) are for four taxa but can easily be extended to larger numbers. The form of the proof is to construct two distinct trees with the same expected frequencies of parsimony sites. Hence, using only parsimony sites, no transformation of the data, linear or nonlinear, can consistently recover the true tree. A procedure that uses all the sites in the data may be consistent with a suitable nonlinear transformation. However, using only linear transformations (which includes no alteration), the procedure cannot always be consistent. (The linear combinations of the frequencies of the nonparsimony sites would be the same for each tree, and we have shown that the parsimony sites cannot always recover the true tree.) Distances between pairs of taxa are a linear transformation of observed sequences, and thus no method that optimizes a linear function of the observed distances can be consistent, again under the models being considered here. This generalizes an observation (Penny et al., 1992b: 177) that some of the better distance methods for four taxa fail precisely under the same conditions as parsimony and compatibility. Quadratic functions of distances can be consistent (Penny et al., 1992b:173).

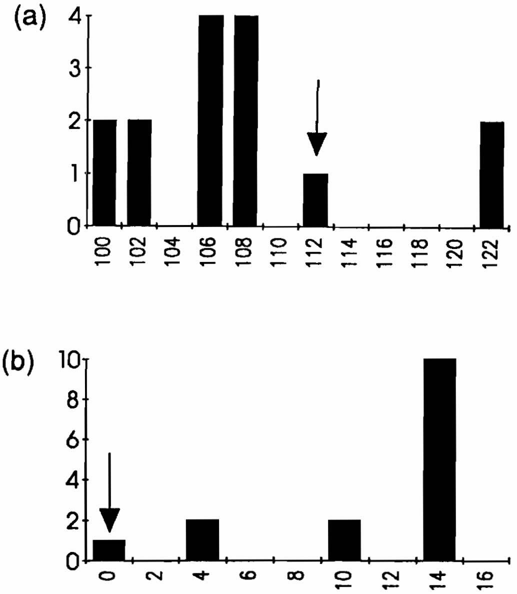

FIGURE 2. The histograms of the frequencies of the parsimony lengths of the 15 binary trees of five taxa. In (a), the lengths have been derived from the parsimony sites in the sequences, and in (b) the lengths are derived from the corrected lengths. In each case, the generating tree T is indicated by the arrow, showing 12 of the trees having a parsimony length shorter than that of T, whereas with the corrected data, T is clearly the shortest.

Consistency under Nonlinear Transformations

Appropriate corrections for multiple hits allow many distance methods to be consistent, including neighbor joining (Studier and Keppler, 1988), neighborliness (Sattath and Tversky, 1977), and several "compare distance" methods that optimize the fit between corrected pairwise distances and the equivalent path lengths on the tree (see Swofford and Olsen, 1990).

The development of spectral analysis has helped clarify the role of nonlinear trans-

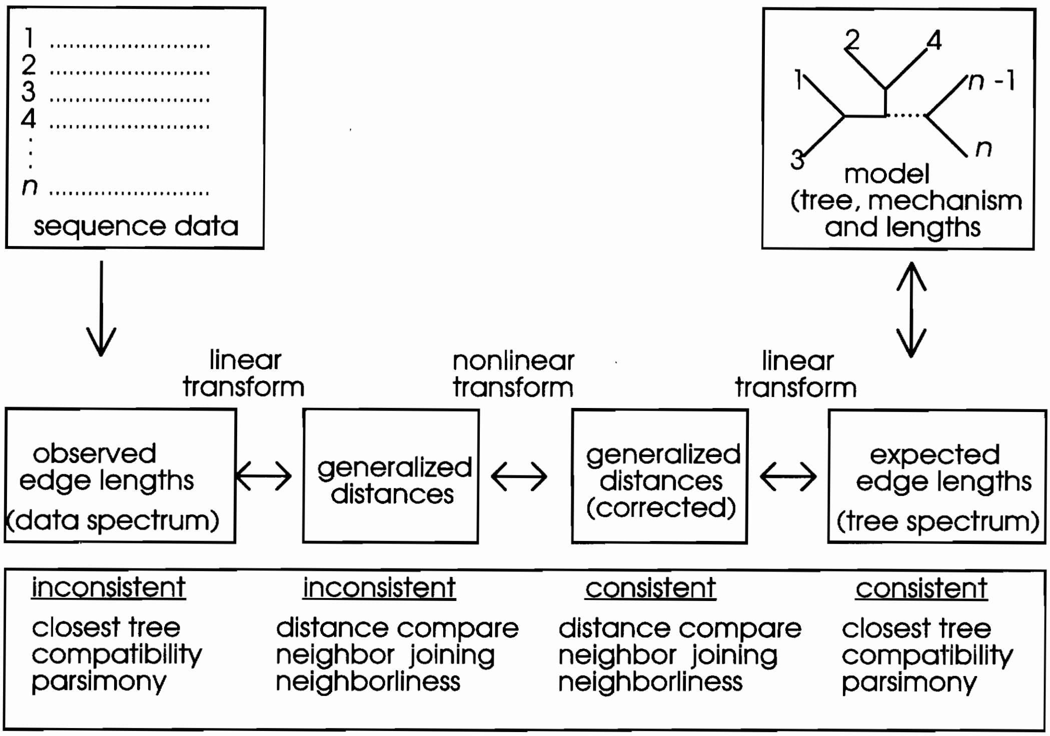

FIGURE 3. Invertible transformations between data and model. The calculations may start with either the data or the model. Sequence data are converted to the frequencies of "observed edge length" patterns (middle left). The Hadamard transform is a linear transformation between these and a generalized set of distances (a set of distances for all subsets with an even number of taxa, i.e., including pairs, quartets, sextets, etc.). Transformation between observed and corrected generalized distances (sets of path lengths) is nonlinear; it is an exponential/logarithmic transformation under both the Cavender model and the Kimura three-parameter model where all sites are equally likely to change. The inverse Hadamard matrix gives the linear transformation between these corrected distances and "expected edge lengths." It inverts back from corrected generalized distances to properties of individual edges of a tree. The tree is derived directly from the expected edge lengths (or indirectly from corrected distances). All transformations are invertible so the model can be used to predict properties of sequences. Each tree selection procedure (applied to information at that point) is identified as consistent or inconsistent for the Cavender model or the Kimura three-parameter model. "Distance compare" procedures optimize the fit between pairwise distances derived from data and path lengths on the tree. Pairwise distances are a subset of the generalized distances. [Based on Penny et al., 1993.]

formations that adjust for multiple hits. The overall relationships among the original data, linear and nonlinear transformations, and how the point of selection of a tree affects consistency are shown in Figure 3. Under the first linear transformation, sequences are equivalent to "generalized distances" (lengths of path sets of Hendy and Penny [1989, 1933]), and observed pairwise distances are a subset of these. The second transformation (nonlinear) corrects for multiple hits, and the third transformation (linear) recovers trees in the following sense. For two-state characters, correcting sequences directly for multiple hits by spectral analysis provides a weighting of each split (bipartition of the taxa set) (Hendy and Penny, 1993). A phylogenetic tree is just a set of compatible splits, each split corresponding to an edge of the tree (Buneman, 1971). Under Cavender's model of substitution, splits not corresponding to edges of the tree generating the sequences will be given a weighting that, with probability 1, tends to zero as the length of the sequence grows. Consequently, any tree-building method that gives a tree, T, whenever each of the weighted splits corresponding to edges of T "dominate" all other weighted splits not in the generating tree must be statistically consistent when applied to these corrected data, even though the method may have been inconsistent when applied to uncorrected data. For example, a sufficient condition for the consistency of a method for corrected data based on splits σ of weight $\omega(\sigma)$ is the following condition. The method returns tree T whenever (for some constant $k > 0$)

$$\min\{\omega(\sigma):\sigma \in T\} > k \sum_{\sigma \notin T} |\omega(\sigma)|.$$

This is satisfied by parsimony, compatibility (for $k = 1$), and closest tree.

For four-state characters that have evolved according to Kimura's three-parameter model, spectral analysis will also identify the splits in the true tree for sufficiently long sequences (see Steel et al., 1992; Evans and Speed, 1993). However, an additional feature of the four-state model is that a subset of the corrected values always tends to zero with increasing sequence length, regardless of which tree generated the data. As with Cavender's two-state model, selection criteria such as parsimony, compatibility, and closest tree will be consistent for the corrected data, even though they may be inconsistent for uncorrected data (Hendy and Charleston, 1993).

Discussion

The results reported here are relevant to the two models discussed. With four-state characters, such as nucleotides, the theorems in the Appendix do not apply to either the one-parameter model (Jukes-Cantor, all nucleotide changes equally likely) or Kimura's two-parameter model (transitions having a different rate from transversions). Linear invariants (involving only linear transformations) are known for these models (Lake, 1987; Sidow and Wilson, 1990; Felsenstein, 1991; Fu and Li, 1992). However, the theorems apply to more complex transition matrices that can be reduced to the three-parameter model.

An apparent problem appears in the use of the "appropriate" correction for multiple hits; a pattern cladist may not wish to "assume" a mechanism. In principle, the assumption of a mechanism is not a problem because the invertibility of all transformations allows a test of the fit between the model and the data. For example, a chi-square or a G-test can be made between the observed and predicted patterns in sequences (Penny et al., 1987; Lockhart et al., 1992). This, in principle, allows the data to reject the model, an essential point for a scientific theory (Penny et al., 1992a). By testing alternative mechanisms, it may be possible to accept that a model (tree, edge lengths, and mechanism of change) explains the observed data. But with many sets of real data, more powerful methods are needed to handle complexities in the data such as unequal GC content between sequences (Lockhart et al., 1992) or nontree models that allow recombination, hybridization, and gene conversion.

Minimal evolution (Rzhetsky and Nei, 1992) is an approximation of the full spectral analysis in that adjustments for multiple hits are for pairwise distance values and their matrix **A** for converting from paths to individual edge lengths is a submatrix of **K** from Hendy and Penny (1989) and Penny et al. (1987). The matrix **K** itself is a linearly transformed submatrix of **H** (Hendy and Penny, 1993). The advantage of using **H** rather than **A** or **K** is that **H** is constant for all trees and its inverse is known, whereas separate transformations and inversions based on **A** or **K** must be made for every tree.

We are not in any way advocating the use of parsimony or compatibility for transformed data; we are just pointing out that these criteria will always be consistent after appropriate nonlinear transformations. If anything, we have a preference for the closest-tree optimality criterion, which is very fast to calculate (Hendy and Penny, 1989). But again, it is not our purpose here to advocate the use of this procedure, because more testing is required. Compatibility is faster to calculate on corrected edge lengths than is parsimony, but we do not know how powerful (in the sense of Penny et al., 1992a) the criterion is compared with parsimony or closest tree. To make appropriate corrections for multiple hits, it is necessary to use all sites in the data, even though parsimony and compatibility use only a subset of sites for the final selection of the optimal tree. We prefer the term parsimony sites rather than informative sites because the information from all sites must be used for correcting for multiple hits, otherwise the methods cannot be consistent.

We must separate a "method" into at least two components: the tree selection procedure and any transformations of the data. We cannot consider parsimony or compatibility a method; rather they are optimality criteria whose properties depend on the information they are applied to, including transformations of the original data. Better transformations depend on a better understanding of the underlying mechanisms of evolution.

REFERENCES

BUNEMAN, P. 1971. The recovery of trees from measures of dissimilarity. Pages 387-395 *in* Mathematics in the archeological and historical sciences (F. R. Hodgson, D. G. Kendall, and P. Tautu, eds.). Edinburgh Univ. Press, Edinburgh.

CAVENDER, J. A. 1978. Taxonomy with confidence. Math. Biosci. 40:270-280.

EVANS, S. N., AND T. P. SPEED. 1993. Invariants of some probability models used in phylogenetic inference. Ann. Stat. 21:355-377.

FELSENSTEIN, J. 1978. Cases in which parsimony and compatibility will be positively misleading. Syst. Zool. 27:401-410.

FELSENSTEIN, J. 1991. Counting phylogenetic invariants in some simple cases. J. Theor. Biol. 152:357-376.

FU, Y.-X., AND W.-H. LI. 1992. Construction of linear invariants in phylogenetic inference. Math. Biosci. 109:201-228.

HENDY, M. D., AND M. A. CHARLESTON. 1993. Hadamard conjugation: A versatile tool for modelling nucleotide sequence evolution. N.Z. J. Bot. 31:231-237.

HENDY, M. D., AND D. PENNY. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297-309.

HENDY, M. D., AND D. PENNY. 1993. Spectral analysis of phylogenetic data. J. Classif. 10:5-24.

KIMURA, M. 1981. Estimation of evolutionary sequences between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA 78:454-458.

LAKE, J. A. 1987. A rate independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. J. Mol. Evol. 24:167-191.

LOCKHART, P. J., D. PENNY, M. D. HENDY, C. J. HOWE, T. J. BEANLAND, AND A. W. D. LARKUM. 1992. Controversy on chloroplast origins. FEBS Lett. 301:127-131.

PENNY, D., M. D. HENDY, AND I. M. HENDERSON. 1987. The reliability of evolutionary trees. Cold Spring Harbor Symp. Quant. Biol. 52:857-862.

PENNY, D., M. D. HENDY, AND M. A. STEEL. 1992a. Progress with evolutionary trees. Trends Ecol. Evol. 7:73-79.

PENNY, D., M. D. HENDY, AND M. A. STEEL. 1992b. Testing the theory of descent. Pages 155-183 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, New York.

PENNY, D., R. E. HICKSON, P. J. LOCKHART, AND E. E. WATSON. 1993. Some recent progress with evolutionary trees. N.Z. J. Bot. 31:275-288.

RZHETSKY, A., AND M. NEI. 1992. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. 9:945-967.

SAITOU, N., AND M. NEI. 1987. The neighbor-joining method: A new method for constructing evolutionary trees. Mol. Biol. Evol. 4:406-425.

SATTATH, S., AND A. TVERSKY. 1977. Additive similarity trees. Psychometrika 42:319-345.

SIDOW, A., AND A. C. WILSON. 1990. Composition statistics: An improvement in evolutionary parsimony and its application to deep branches in the tree of life. J. Mol. Evol. 31:51-68.

STEEL, M. A. 1989. Distributions on bicolored evolutionary trees. Ph.D. Thesis, Massey Univ., Palmerston North, New Zealand.

STEEL, M. A., M. D. HENDY, L. A. SKÉKELY, AND P. L. ERDÖS. 1992. Spectral analysis and a closest tree method for genetic sequences. Appl. Math. Lett. 5:63-67.

STUDIER, J. A., AND K. J. KEPPLER. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. Mol. Biol. Evol. 5:729-731.

SWOFFORD, D. L., AND G. J. OLSEN. 1990. Phylogeny reconstruction. Pages 411-501 *in* Molecular systematics (D. M. Hillis and C. Moritz, eds.). Sinauer, Sunderland, Massachusetts.

ZHARKIKH, A., AND W.-H. LI. 1993. Inconsistency of the maximum-parsimony method: The case of five taxa with a molecular clock. Syst. Biol. 42:113-125.

Received 5 March 1993; accepted 15 July 1993

APPENDIX

Any tree building procedure that either (1) omits singleton sites or (2) selects the tree that optimizes any linear function of all sites can be inconsistent, even for four taxa, under either Cavender's model for two-state characters or under Kimura's three-parameter model for four-state characters. The result described above is a more general conclusion.

Parsimony and compatibility for observed (uncorrected data) satisfy both (1) and (2), and for these

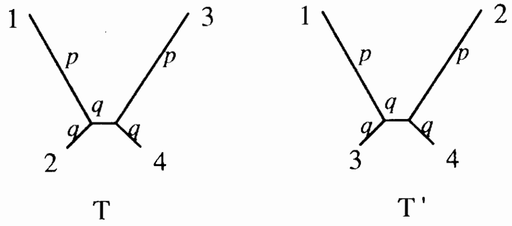

FIGURE 4. Two unrooted binary trees of four taxa, which induce the same expected frequencies for the three parsimony patterns (and the constant pattern) under Cavender's model when $p^2 = q(1 - q)$.

procedures inconsistency has been shown by Felsenstein (1978). Here we prove the more general results described above, the first of which follows from a slight modification of Felsenstein's (1978) argument. Because Cavender's model (1978) is a submodel of Kimura's 3ST model (1981) (by setting the two transversion rates to zero on each edge of the generating tree), it suffices to consider only the Cavender model. The arguments apply equally to rooted and unrooted trees.

Let T and T' be the trees shown in Figure 4, where p and q denote the probabilities of a net change of state across the edges shown. For a pattern π of states (e.g., where 0, 0, 1, 0 is regarded as the same pattern as 1, 1, 0, 1), let f_π (or f'_π) be the probability of generating the pattern π on the leaves of T (or T') under Cavender's model. (For four taxa there are eight distinct patterns: three parsimony patterns [each supporting a different tree], four singleton patterns, and one pattern where all taxa are constant.)

By choosing $p^2 = q(1 - q)$, we see that $f_\pi = f'_\pi$ for the three parsimony patterns and the constant pattern. In particular, if $x = 1 - 2p$ and $y = 1 - 2q$, then

$$s_{1,2} = s_{1,3} = (1 - y^4)/8$$
$$s_{1,4} = [1 - 4xy(1 - y) - y^4]/8$$
$$s_0 = [1 + 4xy(1 + y) - y^4]/8.$$

Thus, two different trees can generate the same expected frequencies of these four patterns, so no method based on just these four patterns can consistently recover the original tree. This establishes (1).

For (2), suppose a method selects a tree T that maximizes the sum

$$L_T = \Sigma_\pi\, w(\pi, T) n_\pi,$$

where n_π is the number of sites indicating pattern π and w is any positive weighting function dependent on T and on π.

We show that such a method can be inconsistent by demonstrating a contradiction: if the method were always consistent it would depend only on the parsimony sites, and therefore by part (1) it would be inconsistent. Thus, suppose the method is consistent. First consider a tree T of four taxa in which all the edges have a probability ϵ of a change of character state. Under Cavender's model, L_{T^*} has expected value

$$c[w(\pi_0, T^*) + h(\epsilon)],$$

where π_0 is the constant pattern,

$$\lim_{\epsilon \to 0} h(\epsilon) = 0,$$

and c is the sequence length. For the method to recover each of the three trees T for all $\epsilon > 0$, we must have

$$w(\pi_0, T^*) \text{ is equal for each } T^*. \qquad (1)$$

Next, consider any tree T on which the edge incident with taxon i (=1, 2, 3, 4) has probability $p > 0$ of a change of state, whereas all the remaining edges have probability ϵ. Under Cavender's model, the expected value of L_{T^*} for sequences of length c generated by T_i is

$$c[pw(\pi_i, T^*) + (1 - p)w(\pi_0, T^*) + h'(\epsilon)], \qquad (2)$$

where π_i is the singleton pattern with taxon i distinguished and

$$h'(\epsilon) = h'(\epsilon, p, i) \to 0 \text{ as } \epsilon \to 0.$$

Applying Equation (1) to Equation (2), we see that consistency of the method for each of the three trees T and all $\epsilon > 0$ implies that $w(\pi_i, T^*)$ is the same for all T^*. Repeating the argument for all i, we see that the method is independent of all singleton sites. Thus by part (1) of the theorem it can be inconsistent. Thus both parts (1) and (2) are established for both the Cavender model and the Kimura three-parameter model.