Point of View

Syst. Biol. 59(4):477–485, 2010

DOI:10.1093/sysbio/syq028
Advance Access publication on May 31, 2010

The Akaike Information Criterion Will Not Choose the No Common Mechanism Model

MARK T. HOLDER[1,*], PAUL O. LEWIS[2], AND DAVID L. SWOFFORD[3,4]

[1]*Department of Ecology and Evolutionary Biology, University of Kansas, 1200 Sunnyside Avenue, Lawrence, KS 66045, USA;* [2]*Department of Ecology and Evolutionary Biology, University of Connecticut, 75 North Eagleville Road, Unit 3043, Storrs, CT 06269-3043, USA;* [3]*Institute for Genome Sciences and Policy Center for Evolutionary Genomics, Duke University, Durham, NC 27708, USA; and* [4]*National Evolutionary Synthesis Center, 2024 W. Main Street, Durham, NC 27705, USA;*

[*]*Correspondence to be sent to: Department of Ecology and Evolutionary Biology, University of Kansas, 1200 Sunnyside Avenue, Lawrence, KS 66045, USA; E-mail: mtholder@ku.edu.*

Received 12 May 2009; reviews returned 6 July 2009; accepted 21 January 2010
Associate Editor: Olivier Gascuel

Recent statistical arguments for parsimony have centred on the demonstration by Tuffley and Steel (1997) that maximum likelihood (ML) under a "no common mechanism" (NCM) model will produce the same ranking of trees as parsimony. We will argue that the NCM model does not provide a compelling justification for using parsimony and will present a proof that if one were to consider the NCM model as a basis for ML inference of trees, then model selection criteria would reject the NCM in favor of simpler models.

Statistical arguments defending parsimony are surprisingly diverse and several distinct "parsimony-equivalent" models have been identified. We will use the phrase "parsimony-equivalent" to refer to models that are guaranteed to cause ML to estimate the same tree as parsimony (with the caveat that employing these models in a Bayesian context may not lead to the same tree as parsimony). We will briefly discuss what these parsimony-equivalent models tell us about the strengths and weaknesses of parsimony. Of particular interest is whether these models have identified general conditions in which parsimony should be preferred to ML inference under one of the commonly used models of character evolution. In our view, statistical arguments for parsimony have been unpersuasive because they have failed to identify general scenarios of speciation, extinction, or character change that would result in data sets in which parsimony's performance is superior to ML inference under standard models.

The simplest model for discrete character data assumes that all substitutions occur at the same rate. We will refer to this model as the Cavender-Farris-Neyman (CFN) model (Jukes and Cantor 1969; Neyman 1971; Farris 1973; Cavender 1978). When evaluating the probability of a transition across a branch in the CFN model, all characters are assumed to use the same branch length. The NCM model of Tuffley and Steel generalizes the CFN model by allowing *each character to have a separate set of branch lengths* in the likelihood calculations.

Because the number of parameters in the NCM model is always larger than the number of data points, many researchers (e.g., Sanderson and Kim 2000; Kolaczkowski and Thornton 2004; Kim and Sanderson 2008) have used the result of Tuffley and Steel to argue that parsimony is best viewed as a nonparametric estimator of phylogeny. This interpretation seems appropriate, but it is unclear whether it is an endorsement of parsimony. Nonparametric estimators are usually thought of as conservative because they sacrifice power for a wider range of applicability. However, Spencer et al. (2005) point out that parsimony's well-known cases of inconsistency (Felsenstein 1978) even under mild assumptions and common conditions (Hendy et al. 1994; Huelsenbeck and Lander 2003) demonstrate that parsimony cannot be viewed as a conservative, robust estimator of trees (see also discussion in Kim and Sanderson 2008).

A different attempt to use the NCM model as a justification for parsimony relies on the reputation of ML as a well-behaved framework for estimation. This form of argument treats the generality of the formulation of the NCM model (no branch length parameters are *forced* to be equal) as a reason to believe that the model will lead to sound inference under very general conditions. For example, Farris (2008) wrote:

> Tuffley and Steel (1997) introduced a model called No Common Mechanism (NCM), in which characters may—but are not required to—vary their relative rates independently, both within and between branches. Because the independent variation is taken only as a possibility, not as a requirement, NCM would apply to almost any situation, and so may be accepted as realistic. This is useful because Tuffley and Steel also showed that maximum likelihood under NCM selects the same trees as does parsimony. With the realistic NCM in the background, then, most

parsimonious trees have greatest power to explain available observations.

Unfortunately, such arguments neglect the importance of model selection in ML inference. We will not review all the arguments against NCM-based justifications of parsimony here (see Steel 2005; Huelsenbeck et al. 2008, for helpful perspectives on the issues) but will focus on the perspective that model selection tools bring to the debate between parsimony and likelihood (see also Sober 2004).

Analyzing data under more complex models is intuitively appealing, but when the number of parameters to be estimated exceeds the information in the data, our inferences can be unsound. Fortunately, there is a well-developed literature on model selection to provide guidance for when a model has too many parameters for the inference to be reliable. The Akaike information criterion (AIC hereafter, Akaike 1973) is a commonly used tool for choosing between alternative models. The AIC is usually expressed as follows:

$$\mathrm{AIC} = 2k - 2\max\ln(L), \tag{1}$$

where k is the number of parameters in a model and $\max\ln(L)$ is the maximum log likelihood under the model for our data. Lower AIC scores are preferred.

In Appendix A, we prove the following theorem:

Theorem 1. *For any $r \geq 2$, the CFN model will have a lower AIC score than the NCM model for any tree and any character matrix comprising 3 or more r-state characters.*

When there is only one character in the matrix, the CFN and NCM models are identical. The proof presented here does not exclude the possibility that the AIC will favor the NCM model over CFN on data sets with 2 characters. Clearly, the CFN model is not rich enough to handle the complexity of most data sets. Our results do not show that the CFN model is the most reasonable model to use, merely that it is superior to the NCM model (according to the AIC).

The current proof relies on several "worst-case" scenarios for the CFN model. The proof uses the likelihood under the NCM model but only a lower bound on the likelihood under the CFN model. This bound was calculated using just one labeling of interior nodes. We also assumed that the proportion of characters that change across each branch in the tree is optimized to generate the largest likelihood ratio in favor of the NCM model. This proportion of characters, $\frac{r-1}{2r}$, corresponds to a large number of changes per branch. When the number of changes on each branch is high, many internal node labelings will contribute significantly to the likelihood. Thus, the lower bound that we employed for the likelihood under the CFN model will be loose. The proof demonstrates a preference for the CFN model over the NCM. Because the bounds used are not tight, the preference for the CFN model will probably be very strong for most real data sets.

Overparameterization of the NCM Model

The NCM model is clearly overparameterized—the number of parameters is always greater than the number of cells in the data matrix. When introducing the NCM model, Tuffley and Steel (1997) pointed out that the large number of parameters in the model makes it statistically inconsistent. This overparameterization is also the focus of the Steel (2005) paper. In this light, it is unsurprising that statistical model selection will prefer a simpler model. Indeed, a model in which the number of parameters exceeds the number of data points is so overparameterized that the asymptotic assumptions underlying the justifications for the AIC (and ML inference in general) do not hold. Far from giving us confidence about the statistical status of parsimony, these considerations merely reinforce the point that the NCM model is too overparameterized to be trusted.

Here, we have attempted to give the benefit of the doubt to the argument of Farris (2008) about the generality of the NCM model. For the present purposes, we will assume that the NCM model is a legitimate and potentially useful model for ML inference. Even if we use the AIC—a model selection criterion that has been criticized for favoring models that are too parameter rich (Katz 1981; Kass and Raftery 1995)—we will not prefer the NCM model over the simplest model used in standard ML tree inference. The theorem proven here shows that it is not even necessary to go through the exercise of formal model selection—the NCM model can simply be rejected without looking at the data set. It is striking that this result can be demonstrated using a loose lower bound on the score of the CFN model and assuming no properties of the data set at hand.

If one accepts the criticisms that the AIC chooses models that are too parameter rich (Katz 1981; Kass and Raftery 1995), then model choice criteria which do not share this deficiency would also prefer the CFN model over the NCM model. For example, the "penalty" for parameters in the Bayesian information criterion (the BIC; Schwarz 1978) is $k\ln M$, where k is the number of parameters and M is the number of characters. The penalty term in the AIC is $2k$. Note that $k\ln M > 2k$ whenever $M \geq 8$, so the BIC penalizes parameter-rich models more severely than the AIC when $M \geq 8$. Thus, a corollary of Theorem 1 is that model selection using the BIC will favor the CFN model over the NCM model for data sets of 8 or more characters.

The Status of Parsimony as a Statistical Approach to Tree Inference

The central message of the previous section is that the generality of the formulation of the NCM model does not provide a compelling argument for the use of parsimony because the NCM model would never be chosen by model selection criteria. This result should not be taken as a statement that parsimony can never be defended—merely that this defense of parsimony is lacking.

As Sober (2004) points out, it is inappropriate to take the result of Tuffley and Steel as the "only" way to justify parsimony. We cannot equate every property of the NCM model with parsimony because there may be other ways to understand parsimony. In addition to the existence of other parsimony-equivalent models, important theoretical results have revealed cases in which parsimony is guaranteed to behave well. Felsenstein (1973) showed that if the probability of change in character state is very small, then parsimony will agree with a tree estimate from ML. Steel (2000) identified sufficient conditions for parsimony to be a consistent estimator; in general, these conditions boil down to the probabilities of change being "sufficiently small and not too unequal across the tree" (see Steel 2000, for explicit description of conditions). In these restricted domains, parsimony will be a consistent estimator of the tree, but ML under the CFN model will also be consistent. It is unclear which method will perform better in terms of power.

Parsimony-equivalent models seem to offer deeper insight into the behavior of parsimony because they reveal what assumptions must be put into a model to make ML act like parsimony for any data set—these models do not just make predictions about parsimony's behavior when the amount of character divergence is low. What do these models imply about parsimony?

The Absence of Branch Length Heterogeneity in Parsimony-Equivalent Models

Goldman (1990) proved that the most parsimonious tree is identical to the tree preferred by ML under the CFN model if we force all branches to have the same length and we simultaneously infer ancestral character states for each internal node. At first glance, the Goldman (1990) model and NCM model seem to be at opposite extremes in terms of what they assume about branch lengths. In the Goldman approach, every branch is forced to have the same length. In the NCM model, each character has its own set of branch lengths.

Unlike virtually every other model used in ML phylogenetics neither the Goldman (1990) model nor the NCM model consider the possibility that different branches should have different expected probabilities of change across all characters. The Goldman model explicitly ignores branch length heterogeneity, but how is it possible to say that the NCM model does something similar? When maximizing likelihood for a character, the NCM model will not prefer to put changes onto branches that appear to be long (based on other characters) because none of the branch length parameters in the model affect multiple characters. Loosely speaking, the NCM model embraces *abundant* branch length heterogeneity (with the same branch being inferred to have a length of 0 for some characters and a length of ∞ for other characters) but not *meaningful* branch length heterogeneity. Suspicion about the meaningfulness of branch length parameters under NCM is aroused by the realization that when more than one most parsimonious reconstruction exists for a character, ML estimates for the same branch length can vary between 0 and ∞. This branch length agnosticism of the NCM model is even more obvious when one integrates over the branch length parameters rather than using a maximum likelihood estimate (MLE) of the length. This Bayesian form of the NCM model results in inference identical to the CFN model with the same branch length applied to every branch (Goloboff 2003; Huelsenbeck et al. 2008).

The refusal to introduce an across character branch length parameter seems crucial to achieving parsimony-equivalent behavior. This assumption is embedded within the procedure parsimony uses to score trees. When calculating a parsimony score, the number of changes in other sites does not alter the cost of adding a change on a branch. This failure to account for meaningful branch length heterogeneity by parsimony-equivalent models clearly has some negative side effects such as the well-known susceptibility to long-branch attraction (Felsenstein 1978). Can this property be a benefit over the standard ML approach of using all the characters to infer a different length for each branch in the tree? At first glance, it seems that the answer would be "probably not." In general, it seems safer to allow for the possibility of important branch length heterogeneity affecting multiple sites than refusing to recognize a "multicharacter branch length effect" when it does occur. However, there are contexts in which it can be better to ignore branch length information (see the discussion of the results of Kolaczkowski and Thornton 2004 below).

Overparameterization of Parsimony-Equivalent Models

The models of Farris (1973), Tuffley and Steel (1997), and Goldman (1990) establish general links between parsimony and ML. All these models are overparameterized (see Felsenstein 1973, for a discussion of the model of Farris 1973). As a result, they display the unusual property that the ML score for a character can be calculated from a single (most parsimonious) character-state reconstruction. Note that, for data sets with multistate characters, it is often possible to assign branch lengths in the NCM model such that multiple ancestral state reconstructions *do* contribute to the ML score. But even in these cases, it is possible to reassign branch lengths to the tree in the NCM model such that the ML score can be obtained from a single reconstruction. Thus, the conclusion that the ML score can be calculated from one reconstruction is valid. Under most models used for ML phylogenetics, all ancestral character-state histories contribute some to the likelihood.

If two trees have the same minimum number of steps, but 1 has more most parsimonious character-state reconstructions than the other, then parsimony will judge the trees to be equally good. Mimicking this property in ML seems to require a procedure that is equivalent to the inference of ancestral character states—either explicitly inferring states (as in the Goldman approach) or as a side effect of zero length branches (as in the NCM

model). The overparameterization inherent in doing this, as well as negative results about the prospects for inferring ancestral states on deep trees (Mossel 2003), imply that this practice is not an appealing aspect of parsimony-equivalent models.

Standard ML under the single-branch-length CFN model would sum over all possible ancestral character states and would not be overparameterized. This form of inference will *not* always infer the same tree as parsimony. The CFN model with one length for all branches evaluates the probabilities of character-state change in a way that is exactly analogous to how parsimony scores changes. But in order to get parsimony-equivalent behavior, Goldman (1990) had to make inference under the CFN overparameterized by estimating ancestral character-state assignments for all internal nodes.

The overparameterization in the Goldman (1990) approach is not simply a mathematical restriction to make it easier to prove that a single-branch-length CFN model is parsimony equivalent. Cases are known in which single-branch-length CFN model and parsimony will disagree on the tree. Steel (1989) showed that parsimony can be an inconsistent estimator of the tree even when branch lengths are equal and characters evolve according to the CFN model (see also Kim 1996). ML inference under the single-branch-length CFN model, on the other hand, is a consistent estimator of the tree if the data are generated on a tree with the same length assigned to every branch. However, model conditions with "unequal" branch lengths can be found for which ML inference under the equal branch length CFN model appears to be inconsistent, whereas parsimony remains consistent (e.g., the tiny region of parameter space in figure 10 of Huelsenbeck et al. 2008).

Sober (2004) and Goloboff (2003) are correct to point out that other connections between ML and parsimony may be demonstrated in the future. Thus, we cannot "know" that properties that are common to all known parsimony-equivalent models are actually prerequisites for parsimony-equivalent behavior. Nevertheless, we conjecture that any general parsimony-equivalent models developed in the future will also be characterized by overparameterization. This conjecture is based on two points from the discussion above. The first is the nature of how parsimony scores a character in the face of multiple most parsimonious reconstructions. The second is the fact that the single-branch-length CFN model agrees with parsimony about how to assess the probability of different events but can only be made equivalent to parsimony if inference of ancestral character states is performed. By "general parsimony-equivalent model," we mean a model for which ML and parsimony agree on any data set when the parameters of the model are not arbitrarily constrained. Steel and Penny (2004) proved that CFN is parsimony-equivalent if the number of states is large enough relative to the number of taxa and characters; but we would not consider this to be a general parsimony equivalence because for a fixed number of character states, data sets can be created for which parsimony and ML disagree. Similarly, a single-branch-length CFN can be forced to be parsimony equivalent by constraining the branch length to be sufficiently small (M. Steel, personal communication; see Kim 1996, for a proof that the single-branch-length CFN model is parsimony equivalent when the MLE of the per-branch transition probability is small), but the general form of the single-branch-length CFN model is not always parsimony equivalent.

Implications of Parsimony-Equivalent Models

It is unwise to take the apparent generality of the NCM model at face value and interpret the entire parameter space of NCM as a realm in which parsimony will perform well. To do this is to adopt an extremely optimistic parametric worldview in which including a parameter in a model to account for a phenomenon guarantees the model will be resistant to noise introduced by that phenomenon. We are familiar with cases in which adding a parameter does make a model more robust. For example, the Jukes–Cantor model can perform poorly if there is transition–transversion bias, but adding a parameter to account for this bias results in a model that is not confused by a high rate of transitions. However, when we are dealing with a model as overparameterized as the NCM model, it is not appropriate to think that we have made our inference widely applicable by adding more parameters. Well-known examples of the inconsistency of parsimony show that it does not perform well for all combinations of parameters that are allowed under the NCM model (Felsenstein 1978; Hendy et al. 1994; Huelsenbeck and Lander 2003). The theorem proven here shows that if we are interested in ML tree estimation and using the AIC or BIC for model selection, we will never prefer the NCM model over even the simplest independent, identically distributed model, the CFN.

Even as a heuristic tool for understanding parsimony, the parametric interpretation of the result of Tuffley and Steel does not seem particularly helpful. On the basis of the NCM model's assumptions, one might assume that parsimony would outperform standard ML models when data are generated such that there are no correlations between the branch lengths from one character to the next. Interestingly, data generated by such a process will show no strong pattern of branch length heterogeneity and, in fact, will be indistinguishable from data generated by a one length for all branches model. If the average amount of change across branches is not too high, then these data will be easy for almost any character-based tree reconstruction method. If the average amount of change is high, then the data can be difficult to analyze. Ironically, it is parsimony, and not the CFN model, that will become an inconsistent estimator of the tree for these data sets (the inconsistency of parsimony when branch lengths are equal but long was shown by Steel 1989; Kim 1996).

Interpreting the Tuffley and Steel (1997) result as justification for parsimony as a nonparametric inference

tool, on the other hand, causes us to consider contexts in which the assumptions of standard ML tree inference might be particularly problematic (instead of looking for cases in which the data might correspond to the assumptions of parsimony-equivalent model). The case of parsimony outperforming ML when data are generated as a mixture of tree models with different branch lengths (Kolaczkowski and Thornton 2004) provides an interesting example. When data from some tree shapes are mixed together, the common branch length assumption made by standard ML methods is so grossly incorrect that it actually misleads us during tree inference. In this case, parsimony's refusal to treat some branches as longer than others seems to be better than estimating one common set of branch lengths. The standard ML approach detects branch length heterogeneity in these data but would do better if it ignored this information.

Fortunately, for advocates of the standard ML approach to phylogenetics, this result does not appear to be general. Spencer et al. (2005) showed that inference under a simple ML model outperforms parsimony on average if we examine a wider range mixture models of the type that Kolaczkowski and Thornton (2004) studied. Gaucher and Miyamoto (2005) showed that the preference for parsimony when data are generated on a mixture of tree models is also sensitive to the exact mixing proportion. Averaging over a wider set of conditions, Gaucher and Miyamoto (2005) found that the performance of ML was superior to the performance of parsimony. Spencer et al. (2005) also demonstrated that using a wider set of mixture models for inference can avoid the poor performance of ML methods even in the difficult mixtures originally reported by Kolaczkowski and Thornton (2004).

Conclusions

The speed and intuitive appeal of parsimony make it a useful tool for exploring data, understanding existing methods and developing new methods. The proof by Tuffley and Steel (1997) stands out as an exceptional and enlightening result, but it should not be taken as a statistical justification for the use of parsimony as a reliable ML estimator of trees. A good model for phylogenetic inference must be rich enough to deal with sources of noise in the data, but ML estimation conducted using models that are clearly overparameterized can lead to drastically wrong conclusions. The NCM model certainly falls in the realm of being too parameter rich to serve as a justification of the use of parsimony based on it being an ML estimator under a general model.

Funding

M.T.H. was supported by a National Science Foundation grant (DEB-0732920); P.O.L. and D.L.S. were supported by EF-03-31495. D.L.S. also received support from the National Evolutionary Synthesis Center (NESCent; NSF EF-0905606).

Acknowledgment

The authors wish to thank Mike Steel, Jack Sullivan, Olivier Gascuel, David Posada, and two anonymous reviewers for comments that improved the manuscript.

References

Akaike H. 1973. Information theory as an extension of the maximum likelihood principle. In: Petrov B.N., Csaki F., editors. Second International Symposium on Information Theory. Akademiai Kiado. New York: Academic Press. p. 267–281.

Alon N., Chor B., Pardi F., Rapoport A. 2010. Approximate maximum parsimony and ancestral maximum likelihood. IEEE/ACM Trans. Comput. Biol. Bioinform. 7:183–187.

Cavender J.A. 1978. Taxonomy with confidence. Math. Biosci. 40: 271–280.

Farris J.S. 1973. A probability model for inferring evolutionary trees. Syst. Zool. 22:250–256.

Farris J.S. 2008. Parsimony and explanatory power. Cladistics. 24: 825–847.

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

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Gaucher E., Miyamoto M. 2005. A call for likelihood phylogenetics even when the process of sequence evolution is heterogeneous. Mol. Phylogent. Evol. 37:928–931.

Goldman N. 1990. Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. Syst. Zool. 39:345–361.

Goloboff P.A. 2003. Parsimony, likelihood, and simplicity. Cladistics. 19:91–103.

Hendy M.D., Penny D., Steel M.A. 1994. A discrete Fourier analysis for evolutionary trees. Proc. Natl. Acad. Sci. USA. 91:3339–3343.

Huelsenbeck J.P., Ane C., Larget B., Ronquist F. 2008. A Bayesian perspective on a non-parsimonious parsimony model. Syst. Biol. 57:406–419.

Huelsenbeck J.P., Lander K. 2003. Frequent inconsistency of parsimony under a simple model of Cladogenesis. Syst. Biol. 52: 641–648.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. Maryland Heights (MO): Academic Press. p. 21–132.

Kass R.E., Raftery A.E. 1995. Bayes factors. J. Am. Stat. Assoc. 90: 773–795.

Katz R. 1981. On some criteria for estimating the order of a Markov chain. Technometrics. 23:243–249.

Kim J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing number of taxa. Syst. Biol. 45:363–374.

Kim J., Sanderson M.J. 2008. Penalized likelihood phylogenetic inference: bridging the parsimony-likelihood gap. Syst. Biol. 57: 665–674.

Kolaczkowski B., Thornton J.W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature. 431:980–984.

Mossel E. 2003. On the impossibility of reconstructing ancestral data and phylogenies. J. Comput. Biol. 10:2003.

Neyman J. 1971. Molecular studies of evolution: a source of novel statistical problems. In: Gupta S.S, Yackel J., editors. Statistical decision theory and related topics. New York: Academic Press. p. 1–27.

Sanderson M., Kim J. 2000. Parametric phylogenetics? Syst. Biol. 49:817–829.

Schwarz G. 1978. Estimating the dimension of a model. Ann. Stat. 6:461–464.

Sober E. 2004. The contest between parsimony and likelihood. Syst. Biol. 53:644–653.

Spencer M., Susko E., Roger A. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol. Biol. Evol. 22:1161–1164.

Steel M. 1989. Distributions on bicoloured evolutionary trees [dissertation]. [Palmerston North (New Zealand)]: Massey University.

Steel M. 2000. Sufficient conditions for two tree reconstruction techniques to succeed on sufficiently long sequences. SIAM J. Dis. Math. 14:36–48.
Steel M. 2005. Should phylogenetic models be trying to 'fit an elephant'. Trends Genet. 21:307–309.
Steel M., Penny D. 2004. Two further links between MP and ML under the Poisson model. Appl. Math. Lett. 17:785–790.
Tuffley C., Steel M. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59:581–607.

APPENDIX A

Proof of theorem

The proof will be formulated for a generic alphabet size of r states (for DNA data, $r = 4$). The transition probability across an edge in the CFN and NCM models are a function of the edge length, ν. Under both these models, the conditional probability $\Pr(i \to i|\nu)$ is the same regardless of which of the r states is represented by i. This probability of remaining in the same state will be denoted $p_s(\nu)$.

Furthermore, the probability $\Pr(i \to j|\nu)$ is the same for all pairs of states such that $i \neq j$. This probability of a different state across an edge will be denoted $p_d(\nu)$. The probability of a change to a particular different state cannot exceed $1/r$ in either model. Thus,

$$0 \leq p_d(\nu) \leq 1/r. \tag{A.2}$$

By the law of total probability, we have

$$p_d(\nu) = \frac{1 - p_s(\nu)}{r - 1}. \tag{A.3}$$

Consider an aligned data matrix, X, consisting of M characters (columns in the data matrix). Tuffley and Steel (1997) demonstrated that the maximized likelihood under the NCM model (denoted L_N, here) can be calculated from a single most parsimonious labeling of the internal vertices of the tree with character states for each character. There may be many most parsimonious labelings, but, without loss of generality, we can choose any one of them in the following discussion. We will denote this labeling as m.

$$\ln L_N(T, \widehat{\mathcal{W}}|X) = -M \ln(r) - M \ln(r) \sum_{b \in E} (1 - s_{b,m}), \tag{A.4}$$

where $\widehat{\mathcal{W}}$ denotes the edge length parameters for the NCM model when the parameter values are chosen to maximize the likelihood, E is the set of all edges in the tree T and $s_{b,m}$ is the proportion of characters that retain the same state across branch b in the labeling m.

The result of Tuffley and Steel is usually expressed in terms of the parsimony score rather than a summation over a most parsimonious labeling. Note that by definition of a most parsimonious labeling, $M \sum_{b \in E}(1 - s_{b,m})$ is equal to the parsimony score of the tree.

Equation A.4 relies on the fact that under the NCM model all characters that do not change across an edge can be explained by assigning the corresponding edge a length of 0. Such site $\times$ edge combinations will contribute factors of 1.0 to the likelihood. For every site $\times$ edge combination that involves a site changing across an edge in m, the ML estimate of the length of edge can be set to ∞, which gives a transition probability of $1/r$. By the definition of a most parsimonious labeling, the number of such site $\times$ edge combinations is equal to the parsimony score.

Note that the fact that all MLEs of branch lengths go to 0 or ∞ is a property of the NCM on a particular labeling and not true of the NCM in general. When $r > 2$, some of the edge length estimates of some site $\times$ edge combinations may be nonidentifiable. Here, we are calculating the ML score under the NCM from one labeling, so we assume that all the branch lengths will go to 0 or ∞. The likelihood calculation (given these branch lengths) is still a sum of probabilities over all possible ancestral character state labelings, but in this context, all other labelings will have a probability of 0, so we can simply calculate the likelihood on the labeling used to infer branch lengths.

A lower bound on the CFN likelihood as a function of one labeling.—The maximized likelihood under the CFN model, L_C, is the sum of likelihoods over all possible labelings:

$$L_C(T, \widehat{\mathcal{V}}|X) = \sum_l L_C(\widehat{\mathcal{V}}|l), \tag{A.5}$$

where $\widehat{\mathcal{V}}$ is the vector of ML estimates of the edge lengths for the CFN model, $L_C(\widehat{\mathcal{V}}|l)$ is the probability of a particular labeling given these edge lengths, and the summation is performed over the set of all possible labelings.

Rather than marginalize over all labelings, we will focus on the likelihood of one most parsimonious labeling. Because $L_C(\mathcal{V}|m)$ is just one of the terms that are summed to obtain $L_C(T, \mathcal{V}|X)$, and because these likelihoods of labelings are all positive numbers, we know that $L_C(T, \mathcal{V}|X) \geq L_C(\mathcal{V}|m)$.

We will denote the set of edge lengths for the CFN model that maximizes the likelihood under a particular labeling as $\widehat{\mathcal{V}}_l$. If we choose edge lengths for the CFN model that maximize $L_C(\widehat{\mathcal{V}}_m|m)$, then we may not be choosing the parameter values that maximize $L_C(T, \mathcal{V}|m)$. In other words, $\widehat{\mathcal{V}}_m$ is not necessarily equal to $\widehat{\mathcal{V}}$ (the set of edge lengths that maximize the likelihood under the CFN model). Nevertheless, the likelihood of a most parsimonious labeling under the CFN using $\widehat{\mathcal{V}}_m$ is a lower bound on the ML score of a tree under the CFN:

$$L_C(T, \widehat{\mathcal{V}}|X) \geq L_C(T, \widehat{\mathcal{V}}_m|X) \geq L_C(\widehat{\mathcal{V}}_m|m). \tag{A.6}$$

We can express the log likelihood of one labeling under the CFN model as:

$$\ln L_C(\widehat{\mathcal{V}}_m|m) = -M \ln r + \sum_{b \in E} (M s_{b,m} \ln(p_s(\widehat{\nu}_b)) + M(1 - s_{b,m}) \ln p_d(\widehat{\nu}_b)). \tag{A.7}$$

Here, $\widehat{\nu}_b$ denotes the member of $\widehat{\mathcal{V}}_m$ that is assigned to edge b.

The events that occur on different branches are assumed to be independent of each other. Thus, when examining the effect of a single labeling, we can find the length for each edge that maximizes the likelihood by considering just the factors in the likelihood equation that refer to that edge. So, for any edge b, the MLE of length will result in the following transition probabilities:

$$p_{\mathrm{s}}(\widehat{\nu}_b) = \max\left(s_{b,m}, \frac{1}{r}\right), \tag{A.8}$$

$$p_{\mathrm{d}}(\widehat{\nu}_b) = \min\left(\frac{1-s_{b,m}}{r-1}, \frac{1}{r}\right). \tag{A.9}$$

In other words, the MLEs of the edge lengths will result in the probability of an $i \to i$ transition that exactly matches the proportion of characters that do not change state across the edge in the most parsimonious labeling (subject to the constraint that in the CFN $\Pr(i \to i) \geq \left(\frac{1}{r}\right)$). Although we do not need to be using MLEs for the CFN model in this proof (because we are just establishing a lower bound for the ML score under CFN), using ML branch lengths makes our bound tighter. Appendix B shows that these branch lengths are indeed the MLEs in this context.

An upper bound on the log-likelihood ratio in favor of the NCM model.—If we evaluate the likelihood under the NCM model and the lower bound for the likelihood of the CFN model on the same most parsimonious labeling, then we can bound the improvement in likelihood that can be obtained by adopting the NCM model. Examination of Equations A.4 and A.7 reveals that the formula for the upper bound on the log of the likelihood ratio, R, between NCM and CFN is

$$R(m) = \ln L_{\mathrm{N}}(\widehat{\mathcal{W}}_m|m) - \ln L_{\mathrm{C}}(\widehat{\mathcal{V}}_m|m) \tag{A.10}$$

$$= \sum_{b\in E} M\left(s_{b,m}\ln\left(\frac{1}{p_{\mathrm{s}}(\widehat{\nu}_b)}\right) + (1-s_{b,m})\ln\left(\frac{1}{rp_{\mathrm{d}}(\widehat{\nu}_b)}\right)\right). \tag{A.11}$$

For notational convenience, we will drop the explicit dependence on a labeling for the remainder of the proof. In all cases, the upper bound on the log-likelihood ratio will be referred to as simply R, despite the fact that it is a function of a most parsimonious labeling m. We can consider each edge independently:

$$R_b = M\left(s_{b,m}\ln\left(\frac{1}{p_{\mathrm{s}}(\widehat{\nu}_b)}\right) + (1-s_{b,m})\ln\left(\frac{1}{rp_{\mathrm{d}}(\widehat{\nu}_b)}\right)\right), \tag{A.12}$$

$$R = \sum_{b\in E} R_b. \tag{A.13}$$

Without knowing the number of changes across each branch in the tree, we cannot calculate the exact contribution of each branch to the log-likelihood ratio between models. However, we can find the upper bound for the contribution to the total R made by a single branch, b. This upper bound, R_b^*, will be achieved when the proportion of characters that are constant across a branch is optimal in terms of favoring the NCM model over the CFN model. This proportion of characters will be denoted by $s_{b,m}^*$ in the remainder of the proof:

$$R_b^* = M\left(s_{b,m}^*\ln\left(\frac{1}{p_{\mathrm{s}}(\widehat{\nu}_b)}\right) + (1-s_{b,m}^*)\ln\left(\frac{1}{rp_{\mathrm{d}}(\widehat{\nu}_b)}\right)\right). \tag{A.14}$$

Solving for the worst-case proportion of sites changing across a branch.—We can solve for $s_{b,m}^*$ by finding the point where the derivative of R_b with respect to $s_{b,m}$ is 0. If such a point does not exist or the second derivative is positive, then we must evaluate the end points.

When $\frac{1}{r} \leq s_{b,m} \leq 1$, we can use the equality $p_{\mathrm{s}}(\widehat{\nu}_b) = s_{b,m}$ to transform Equation A.12 as follows:

$$R_b = M\left(s_{b,m}\ln\left(\frac{1}{s_{b,m}}\right) + (1-s_{b,m})\ln\left(\frac{r-1}{r(1-s_{b,m})}\right)\right), \tag{A.15}$$

$$\frac{\mathrm{d}R_b}{\mathrm{d}s_{b,m}} = M\ln\frac{r(1-s_{b,m})}{(r-1)s_{b,m}}, \tag{A.16}$$

$$\frac{\mathrm{d}^2R_b}{\mathrm{d}s_{b,m}^2} = -M\left(\frac{1}{1-s_{b,m}} + \frac{1}{s_{b,m}}\right). \tag{A.17}$$

We note that the second derivative is always negative for the range that we are considering, so finding a point at which the derivative is 0 will give us a maximum at

$$s_{b,m}^* = \frac{r}{2r-1}. \tag{A.18}$$

When $s_{b,m} \leq \frac{1}{r}$, then the edge length that maximizes the likelihood for the CFN will be infinite, and $p_{\mathrm{s}}(\widehat{\nu}_b) = \frac{1}{r}$. The characters that change state across an edge will not affect the likelihood ratio because both the NCM model and the CFN model assign these transitions a probability of $1/r$. Equation A.12 becomes

$$R_b = Ms_{b,m}\ln(r). \tag{A.19}$$

Clearly, for this part of the range of $s_{b,m}$, the log-likelihood ratio is a linear function of $s_{b,m}$. Higher values of $s_{b,m}$ result in larger values of R_b, so the maximum for this part of the range occurs at the largest value of $s_{b,m}$.

Thus, the upper bound on the log-likelihood ratio when the fraction of characters with no change of state is $\leq \frac{1}{r}$ occurs when $s_{b,m} = \frac{1}{r}$ exactly, and the log-likelihood ratio value is $\frac{M \ln(r)}{r}$. Note that this value of $s_{b,m}$ was also considered in the range $1/r \leq s_{b,m} \leq 1$ above. The value of $s^*_{b,m}$ given by Equation A.18 results in a higher upper bound of the log-likelihood ratio. Thus, we can ignore the range $s_{b,m} \leq \frac{1}{r}$ in our calculation of the upper bound.

Bounding the difference in AIC between the models.—Substituting the value of $s^*_{b,m}$ given by Equation A.18 into Equation A.14 and performing substitutions using Equation A.9 and $p_s(\widehat{\nu}_b) = s_{b,m}$ gives us the upper bound on the per-edge log-likelihood ratio between NCM and CFN:

$$R^*_b = M \ln\left(2 - \frac{1}{r}\right). \tag{A.20}$$

Thus, if each branch in the tree has a length that matches this "worst case" for the CFN compared with the NCM model, then

$$R^* = |E| M \ln\left(2 - \frac{1}{r}\right), \tag{A.21}$$

where $|E|$ is the number of edges in the tree.

The NCM model has $|E|M$ parameters, whereas the CFN model has $|E|$ parameters. Thus, we can find the conditions in which we have proven that AIC will favor CFN over NCM by testing for when the difference in the AIC score for NCM model and the AIC for the CFN model is less than 0. Our upper bound on R^* allows us to err on the side of the NCM model—we calculate a lower bound on the AIC difference, $\Delta\text{AIC}^*_{\text{NC}}$, between the NCM and the CFN models:

$$\Delta\text{AIC}^*_{\text{NC}} = 2\left(|E|M - |E| - R^*\right) \tag{A.22}$$

$$= 2|E|(M - 1 - R^*_b). \tag{A.23}$$

The AIC can only prefer the NCM over the CFN model if $\Delta\text{AIC}^*_{\text{NC}}$ is negative. Thus, we must look for cases in which the following condition holds:

$$2|E|(M - 1 - R^*_b) < 0, \tag{A.24}$$

which is equivalent by Equation A.20 to:

$$M - 1 - M \ln\left(2 - \frac{1}{r}\right) < 0. \tag{A.25}$$

Proof of theorem for $M > 3$.—We will prove the theorem by contradiction. Suppose that $M > 3$. Equation A.25 simplifies to:

$$\frac{M-1}{M} < \ln\left(2 - \frac{1}{r}\right). \tag{A.26}$$

This condition is most easily satisfied when r is effectively infinite, and the constraint becomes approximately $\frac{M-1}{M} < 0.6931472$. However, if $M > 3$, then the lowest value that $\frac{M-1}{M}$ can attain is $3/4$ (which occurs when $M = 4$). The statement that $3/4 < 0.6931472$ is a contradiction, showing that $\Delta\text{AIC}^*_{\text{NC}}$ cannot be negative when $M > 3$.

Proof of theorem for $M = 3$.—To prove the case when $M = 3$, we have to find a tighter bound for the log-likelihood ratio in favor of the NCM model than the one given in Equation A.21. Note that $s_{b,m}$ cannot assume a continuum of values between 0 and 1 but will be constrained to have fractional values in which the denominator is M and the numerator is an integer in the range $[0, M]$. For $M = 3$, we can see that that $s_{b,m} \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$. With these 4 possible values of $s_{b,m}$ in mind, we can return to Equation A.15 and examine all possible cases.

When $s_{b,m} \in \{0, 1\}$, the CFN model and NCM model can achieve the same likelihood for the transitions across branch b. When $s_{b,m} = \frac{1}{3}$, then

$$R_b = M\left(s_{b,m} \ln\left(\frac{1}{s_{b,m}}\right) + (1 - s_{b,m}) \ln\left(\frac{r-1}{r(1-s_{b,m})}\right)\right) \tag{A.27}$$

$$= M\left(\ln(3) + \frac{2}{3} \ln\left(\frac{r-1}{2r}\right)\right). \tag{A.28}$$

The value of R_b resulting from $s_{b,m} = \frac{1}{3}$ cannot exceed approximately $0.63651416M$ even when r is large.

When $s_{b,m} = \frac{2}{3}$, then

$$R_b = M\left(\ln(3) - \frac{2\ln(2)}{3} + \frac{\ln\left(\frac{r-1}{r}\right)}{3}\right). \tag{A.29}$$

Once again the likelihood ratio becomes largest for large r and will not exceed $\approx 0.63651416M$.

Substituting these tighter bounds for the log-likelihood ratio into inequality Equation A.24 yields the condition:

$$|E|\left(M - 1 - 0.63651416M\right) < 0. \tag{A.30}$$

When $M = 3$, this evaluates to the statement that $0.090457495|E| < 0$. This contradiction shows that inequality Equation A.24 cannot be met for $M = 3$ and thus the AIC will not prefer NCM model over the CFN model for data sets of 3 characters.

APPENDIX B

Proof of Equation A.8

Examination of Equation A.7 reveals that the contribution of each branch to $L_C(\widehat{\mathcal{V}}_m|m)$ can be considered independently. Here, we show that, for each branch, b, the MLE of the branch length, $\widehat{\nu}_b$, is the length that results in $p_s(\widehat{\nu}_b) = s_{b,m}$. Similar results for the 2-state model were shown by Alon et al. (2010).

The likelihood factor associated with a branch, b, will be denoted L_b. Note that, with this notation, the probability of any $i \rightarrow j$ transition for which $i \neq j$ is $\frac{1-p_s(\widehat{\nu}_b)}{r-1}$. The proportion of characters that change over branch b is simply $1-s_{b,m}$. Equation A.7 and these notational simplifications lead to

$$\ln L_b = M[s_{b,m} \ln p_s(\widehat{\nu}_b) + (1 - s_{b,m}) \ln(1 - p_s(\widehat{\nu}_b)) - (1 - s_{b,m}) \ln (r - 1)], \quad \text{(B.1)}$$

$$\frac{\partial \ln L_b}{\partial p_s(\widehat{\nu}_b)} = M \left(\frac{s_{b,m} - p_s(\widehat{\nu}_b)}{p_s(\widehat{\nu}_b)(1 - p_s(\widehat{\nu}_b))} \right) \quad \text{(B.2)}$$

$$\frac{\partial^2 \ln L_b}{\partial p_s(\widehat{\nu}_b)^2} = M \left(\frac{2p_s(\widehat{\nu}_b)s_{b,m} - p_s(\widehat{\nu}_b)^2 - s_{b,m}}{p_s(\widehat{\nu}_b)^2(1 - p_s(\widehat{\nu}_b))^2} \right). \quad \text{(B.3)}$$

Clearly, Equation B.2 becomes 0 when $p_s(\widehat{\nu}_b) = s_{b,m}$. This point is a maximum because the second derivative at the MLE is negative for the range $0 < p < 1$.