TWO FURTHER LINKS BETWEEN MP AND ML UNDER THE POISSON MODEL

Mike Steel and David Penny

Allan Wilson Centre for Molecular Ecology and Evolution
Department of Mathematics and Statistics
University of Canterbury
Private Bag 4800
Christchurch, New Zealand

Report Number: UCDMS2003/12 AUGUST 2003

TWO FURTHER LINKS BETWEEN MP AND ML UNDER THE POISSON MODEL

Mike Steel[1] and David Penny[2]
Allan Wilson Centre for Molecular Ecology and Evolution
[1] *University of Canterbury*
Christchurch, New Zealand
[2] *Massey University*
Palmerston North, New Zealand

ABSTRACT. Maximum parsimony and maximum likelihood are two contrasting approaches for reconstructing phylogenetic trees from sequence and character data. We establish analytic links between these methods (extending connections reported earlier) under the simple Poisson model of substitutions in two settings. First, we show that if the underlying state space is sufficiently large then the maximum likelihood estimate phylogenetic tree is always a maximum parsimony tree for the data. Second, we show that a sufficiently dense sampling of sequences ensures that the most parsimonious likelihood tree is always a maximum parsimony tree.

Key words. Trees, maximum likelihood estimation, maximum parsimony method, stochastic models.

AMS subject classifications. 05C05, 92D15

1. PRELIMINARIES

Evolutionary relationships in biology are typically represented by trees, for which some set X of present-day species appear as a subset of the vertices. A central problem in molecular systematics is how to infer such trees from character data - that is, from functions from the set X into some set of states. In this paper, we establish new links between two such tree reconstruction methods - one of which (maximum likelihood) is based explicitly on an underlying Markov model for the evolution of characters on a tree, while the other (maximum parsimony) is based on a minimality principle.

We begin by recalling some background and definitions that are required to state our results. Throughout this paper, X will denote a set of n extant species or individuals. A *character (on X, over a set R of character states)* is any function χ from X into some finite set R. Throughout this paper, we let r denote the size of R.

Suppose we have a tree $T = (V, E)$. We say that T is a *tree on* X if X is a subset of V, and all vertices of T of degree 1 or 2 are contained in X. If, in addition, X is precisely the set of leaves of T we say that T is a *phylogenetic X–tree*, and if, furthermore, every vertex of T has degree 3 we say that T is *fully resolved*. Two phylogenetic X–trees are regarded equivalent if the identity mapping from X to X induces a graph isomorphism between the two trees. Further background and mathematical details concerning phylogenetic trees can be found in [9].

The maximum parsimony method for reconstructing a tree on X from a collection of characters on X can be described as follows. Suppose we have a tree $T = (V, E)$ on X, and a function $\overline{\chi} : V \to R$. Let $\mathrm{ch}(\overline{\chi}, T) := |\{e = \{u, v\} \in E : \overline{\chi}(u) \neq \overline{\chi}(v)\}|$. Given a character $\chi : X \to R$, the *parsimony* score of χ on T, is defined by

$$l(\chi, T) := \min_{\overline{\chi}: V \to R, \overline{\chi}|X = \chi} \{\mathrm{ch}(g, T)\},$$

where $\overline{\chi}|X$ denotes the restriction of g to X. Suppose we are given a sequence $\mathcal{C} = (\chi_1, \ldots, \chi_k)$ of characters on X. The *parsimony* score of $\mathcal{C}$ on T, denoted $l(\mathcal{C}, T)$, is defined by

$$l(\mathcal{C}, T) := \sum_{i=1}^{k} l(\chi_i, T).$$

Any tree T on X that minimizes $l(\mathcal{C}, T)$ is said to be a *maximum parsimony* (MP) tree for $\mathcal{C}$, and the corresponding l–value is the *parsimony* or "MP" score of $\mathcal{C}$.

We now consider the simplest tree-based model for the evolution of characters over a set R, which we will refer to here simply as the *Poisson model on R (with parameters (T, p))*. In this model, one has a tree T on X. Let us select any element $x_0 \in X$, as a reference vertex and direct all edges of T away from x_0. We will regard the value from R assigned to vertex x_0 as being given (it would make little difference to the arguments below if we allowed the state at x_0 to be random). The model then assigns states from R recursively to the remaining vertices of the tree according to the following scheme: if $e = \{u, v\}$ is an edge of T directed from u to v and u has been assigned state α, then, with probability $1 - p(e)$ we assign v state α, otherwise, with probability $p(e)$ we select uniformly at random one of the other $r - 1$ states (different to α) and assign this state to v. The assignments are made independently across edges, and the value $p(e)$ is called the *substitution probability* associated to edge e. It is natural to constrain $p(e)$ to lie in the half-open interval $[0, \frac{r-1}{r})$ - the reason for the upper bound is that, if we realise this model by a continuous-time Markov process, then the probability of a net substitution over any period of time is always less than $\frac{r-1}{r}$. We will say that the mapping $e \to p(e)$ is *admissible* if the $p(e)$ values all lie within this allowed interval.

When $r = 4$, this model is essentially the same as what is often referred to as the Jukes-Cantor model [11]. For general values of r, this model has more recently been studied by Paul Lewis [7] as a starting framework for likelihood analysis for certain morphological characters. It has also been referred to in the bioinformatics literature as the 'Neyman r–state model' and the 'Cavender-Farris-Neyman model'.

Given the pair (T, p) where $T = (V, E)$ is a tree on X, and p is an admissible assignment of transition probabilities, and given a map $\overline{\chi} : V \to R$, let

$\mathbb{P}(\overline{\chi}|T,p)$ denote the probability that the vertices in T take values specified by $\overline{\chi}$ under the Poisson model on R with parameters (T,p). More formally, $\mathbb{P}(\overline{\chi}|T,p) = \mathbb{P}(\cap_{v\in V-\{x_0\}}\{\eta(v) = \overline{\chi}(v)\}|)$, where $\eta(v)$ is the random variable state assigned to v under the model. By the assumptions of the model, we have

$$\mathbb{P}(\overline{\chi}|T,p) = \prod_{\{u,v\}\in E:\overline{\chi}(u)\neq\overline{\chi}(v)} \frac{p(e)}{r-1} \prod_{\{u,v\}\in E:\overline{\chi}(u)=\overline{\chi}(v)} (1-p(e)). \tag{1}$$

Given a sequence $\mathcal{C} = (\chi_1,\ldots,\chi_k)$ of characters on X, we put

$$\mathbb{P}(\mathcal{C}|T,p) = \prod_{i=1}^{k} \sum_{\overline{\chi_i}\in c(i)} \mathbb{P}(\overline{\chi_i}|T,p), \qquad \mathbb{P}(\mathcal{C}|T,p)_{\mathrm{mp}} = \prod_{i=1}^{k} \max(\mathbb{P}(\overline{\chi_i}|T,p) \,|\, \overline{\chi_i}\in c(i))$$

$$L(T|\mathcal{C}) = \sup_p(\mathbb{P}(\mathcal{C}|T,p)), \qquad L_{\mathrm{mp}}(T|\mathcal{C}) = \sup_p(\mathbb{P}(\mathcal{C}|T,p)_{\mathrm{mp}})$$

where $c(i) := \{\overline{\chi_i} : V \to R : \overline{\chi_i}|X = \chi_i\}$, and the supremum is taken over all admissible choices of p. Recall that $L(T|\mathcal{C})$ is referred to as the *likelihood* or "ML" score, and $L_{\mathrm{mp}}(T|\mathcal{C})$ as the *most-parsimonious likelihood* or "MPL"score, of T given $\mathcal{C}$ (cf. [3],[10]). Note that $\mathbb{P}(\mathcal{C}|T,p)$ is the probability of generating the k characters by independent and identical evolution under a Poisson model with parameters (T,p). A tree T on X is said to be a *maximum likelihood* (ML) tree or a *most-parsimonious likelihood* ($\mathrm{MP}_\perp$) tree for $\mathcal{C}$ if $L(T|\mathcal{C}) \geq L(T'|\mathcal{C})$ or $L_{\mathrm{mp}}(T|\mathcal{C}) \geq L_{\mathrm{mp}}(T'|\mathcal{C})$, respectively, holds for all other trees T' on X. The problem of finding an MPL tree given only $\mathcal{C}$ was recently shown to be NP-hard in [1] (where the method is referred to as 'ancestral maximum likelihood'.) We say that an MP, ML or MPL tree for $\mathcal{C}$ is *irreducible* if we cannot collapse any edge of T to obtain another such tree for $\mathcal{C}$.

2. Link One: Large state space

Maximum parsimony has already shown to be a maximum likelihood estimator for phylogenetic trees under a 'no-common mechanism' model in which each character evolves independently under a Poisson model on R but where p in the parameter pair (T,p) for this model can vary freely between the characters (for details, see [12] which extended the result for $r = 2$ that was described by [8]). In this section, we describe quite a different link. In contrast to the aforementioned link we consider the 'common-mechanism' setting - here the two methods are in general quite different (they may select different trees, as Felsentein [6] showed). However when the number of states is sufficiently large, then once again maximum likelihood trees are always MP trees. This may be relevant to the use of certain genomic data (such as gene order) for inferring phylogenies, as in this case the underlying state space may be very large.

Theorem 2.1. *Suppose $\mathcal{C} = (\chi_1,\chi_2,\ldots,\chi_k)$ is a sequence of k characters on X over a state space R of size $r \geq 4^{nk}$. Under the model in which the characters evolve independently according to the same Poisson model on R, any ML tree for $\mathcal{C}$ is then an MP tree for $\mathcal{C}$.*

Proof. Suppose that T, but not T' is an MP tree for $\mathcal{C}$. We will show that ML will not select T' since T has a larger ML score than T' given $\mathcal{C}$. By assumption we can

write $l(\mathcal{C},T') = l + \delta$, where $l = l(\mathcal{C},T)$ and $\delta \geq 1$, and we may assume (without loss of generality) that $T' = (V', E')$ is a fully resolved phylogenetic X–tree. We now invoke a key result from [12]: for any character $\chi : X \to R$, we have

$$\sup_{p'} \sum_{\overline{\chi} \in c(\chi)} \mathbb{P}(\overline{\chi}|T', p') = r^{-l(\chi,T')},$$

where $c(\chi) = \{\overline{\chi} : V' \to R : \overline{\chi}|X = \chi\}$ and the supremum is over all admissible p'. Consequently, we have the following upper bound on the ML score of T'

$$L(T'|\mathcal{C}) = \sup_{p'} \mathbb{P}(\mathcal{C}|T', p') \leq r^{-l-\delta} \tag{2}$$

We show that T' cannot be an ML tree because this upper bound (given by (2)) on $L(T'|\mathcal{C})$ is strictly less than $\mathbb{P}(\mathcal{C}|T,p)$ for a particular p that sets $p(e) = \lambda$ for all edges e of T (we will determine λ shortly). Let $T = (V, E)$ and, for each $i \in \{1, \ldots, k\}$, let us select a map $\overline{\chi}_i : V \to R$ for which $\mathrm{ch}(\overline{\chi}_i, T) = l(\chi_i, T)$. Let $l_i := l(\chi_i, T)$ and let $\epsilon = |E|$. By (1), we have

$$\mathbb{P}(\overline{\chi}_i|T,p) = (\frac{\lambda}{r-1})^{l_i}(1-\lambda)^{(\epsilon - l_i)},$$

and, since $\mathbb{P}(\mathcal{C}|T,p) = \prod_{i=1}^{k} \sum_{\overline{\chi} \in c(i)} \mathbb{P}(\overline{\chi}|T,p)$, we have

$$\mathbb{P}(\mathcal{C}|T,p) \geq \left(\frac{\lambda}{(1-\lambda)(r-1)}\right)^{\sum_{i=1}^{k} l_i} (1-\lambda)^{\epsilon k}.$$

Thus, since $l = \sum_{i=1}^{k} l_i$, we have

$$\mathbb{P}(\mathcal{C}|T,p) \geq \left(\frac{\lambda}{(1-\lambda)(r-1)}\right)^{l} (1-\lambda)^{\epsilon k}.$$

Now let us set $\lambda = \frac{r-1}{2r-1}$, so that $\frac{\lambda}{(1-\lambda)(r-1)} = r^{-1}$, and $1 - \lambda = \frac{r}{2r-1} > \frac{1}{2}$. Thus,

$$\mathbb{P}(\mathcal{C}|T,p) \geq r^{-l} 2^{-\epsilon k}. \tag{3}$$

Comparing (2) and (3), and noting that $\delta \geq 1$, we see that

$$L(T|\mathcal{C}) \geq \mathbb{P}(\mathcal{C}|T,p) > \sup_{p} \mathbb{P}(\mathcal{C}|T',p') = L(T'|\mathcal{C})$$

provided $r > 2^{\epsilon k}$ and this certainly holds if $r \geq 4^{nk}$ (since $\epsilon \leq 2n - 3 < 2n$), as required. This completes the proof.

□

3. Link Two: Dense sampling of sequences

Let $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ be a collection of aligned sequences of length k on $r \geq 2$ states. Equivalently, we may view $\mathcal{S}$ as a sequence $\mathcal{C}_\mathcal{S} = (\chi_1, \ldots, \chi_k)$ where χ_i is an r–state character on X. If we write S_i as $S_i(1), \ldots S_i(k)$, then $S_i(l) = \chi_l(i)$ for all $i \in \{1, \ldots, n\}$ and $l \in \{1, \ldots, k\}$. Let d_H denote the Hamming metric on $\mathcal{S}$, that is, $d_H(S_i, S_j) = |\{l : S_i(l) \neq S_j(l)\}|$. We will suppose that the sequences in $\mathcal{S}$ are distinct - that is, $d_H(S_i, S_j) > 0$ for all $i \neq j$. Let $G_\mathcal{S}$ be the graph with

vertex set $\mathcal{S}$ and with an edge connecting any two sequences that differ in exactly one co-ordinate. Equivalently, $G_{\mathcal{S}} = (\mathcal{S}, E)$ where

$$E = \{(S_i, S_j) : d_H(S_i, S_j) = 1\}.$$

In the context of molecular genetics, $G_{\mathcal{S}}$ is the 'haplotype graph' described, for example, in [5].

Definition We say that $\mathcal{S}$ is *ample* if $G_{\mathcal{S}}$ is connected.

The following lemma follows easily from the definitions.

Lemma 3.1. *If $\mathcal{S}$ is an ample collection of sequences then the set of spanning trees of $G_{\mathcal{S}}$ is precisely the set of irreducible MP trees for $\mathcal{C}_{\mathcal{S}}$. Consequently, $\mathcal{C}_{\mathcal{S}}$ has MP score $n-1$.*

We now show that when $\mathcal{S}$ is ample, then any spanning tree for $\mathcal{C}_{\mathcal{S}}$ is also an MPL tree for $\mathcal{C}_{\mathcal{S}}$ under this model. That is, we cannot improve the MPL score by introducing additional "Steiner points" (hypothetical ancestral sequences). As an aside, this result provides another case where a particular instance of an NP-hard problem has a simple, polynomial-time solution. We note also that the Buneman complex [4] or, equivalently, the median network [2] of a collection of X–splits provides natural examples of ample sets of sequences.

Theorem 3.2. *Suppose that $\mathcal{S}$ is ample. Then, under the model in which the characters evolve independently under the same Poisson model on R, the MP trees and the MPL trees for $\mathcal{C}_{\mathcal{S}}$ coincide.*

Proof. It suffices to show that the set of spanning trees for $\mathcal{C}_{\mathcal{S}}$ equals the set of irreducible MPL trees for $\mathcal{C}_{\mathcal{S}}$ (by Lemma 3.1, and the observation that the set of MP (resp. MPL) trees for $\mathcal{C}_{\mathcal{S}}$ is simply the set of all resolutions of the irreducible MP (resp. MPL) trees for $\mathcal{C}_{\mathcal{S}}$).

Suppose that $T = (\mathcal{S}, E)$ is a spanning tree of $G_{\mathcal{S}}$. Then,

$$L_{\mathrm{mp}}(T|\mathcal{C}) = \mathbb{P}(\mathcal{C}_{\mathcal{S}}|T, p) = \prod_{e \in E} \left[\frac{p(e)}{r-1}(1-p(e))^{k-1}\right], \tag{4}$$

for some map $p : E \to [0, \frac{r-1}{r}]$. It is easily checked that the map p that maximizes the expression on the right hand side of (4) assigns the values $p(e) = \frac{1}{k}$ for all $e \in E$ (and this is admissible, since we may assume $k \geq 2$). In view of $|E| = n-1$, this implies

$$L_{\mathrm{mp}}(T|\mathcal{C}_{\mathcal{S}}) = \left(\frac{1}{k(r-1)}(1-\frac{1}{k})^{k-1}\right)^{n-1}. \tag{5}$$

Now, suppose that $T' = (V', E')$ is any irreducible MPL tree for $\mathcal{C}_{\mathcal{S}}$ under a Poisson model on R, select maps $\overline{\chi}_i$ that extend χ_i $(i = 1, \ldots, k)$ so that $L_{\mathrm{mp}}(T'|\mathcal{C}) = \sup_{p'} \prod_{i=1}^{k} \mathbb{P}(\overline{\chi}_i|T', p')$ (which is possible as there are only finitely many such $\overline{\chi}_i$), and put $\overline{\chi}(v) := (\overline{\chi}_i(v))_{i=1,\ldots,k}$ for each $v \in V'$. Then, $\mathcal{S} \subseteq \{\overline{\chi}(v) : v \in V'\}$. Write $E' = \{e_1, e_2, \ldots, e_m\}$ where $m := |E'|$. For an edge $e_i' = (u, v) \in E'$, let

$$y_i = \frac{d_H(\overline{\chi}(u), \overline{\chi}(v))}{k}, \text{ and } \lambda_i = \min\{y_i, \frac{r-1}{r}\}.$$

These λ_i values provide the optimal admissible substitution probabilities for maximizing the MPL score of T' given $\mathcal{C}$. Thus if we let $L = -\log(L_{\rm mp}(T'|\mathcal{C}_S))$, then

$$L = k\sum_{i=1}^{m} f(\lambda_i), \tag{6}$$

where, for $\lambda \in [0, \frac{r-1}{r}]$, $f(\lambda) = -\log\left((\frac{\lambda}{r-1})^{\lambda}(1-\lambda)^{1-\lambda}\right)$.

Note that f is an increasing function on $[0, \frac{r-1}{r}]$. So, $\lambda_i > 0$ (by the irreducibility assumption on T') and the definition of λ_i implies that $\lambda_i \geq \frac{1}{k}$ for each i and so $\sum_{i=1}^{m} f(\lambda_i) \geq mf(\frac{1}{k}) \geq (n-1)f(\frac{1}{k})$. Thus, $L_{\rm mp}(T'|\mathcal{C}_S) \leq L_{\rm mp}(T|\mathcal{C}_S)$ with equality precisely if $m = n-1$ and $\lambda_i = \frac{1}{k}$ for all i. Yet, $m = n-1$ implies that T' is a spanning tree for G_S, and $\lambda_i = \frac{1}{k}$ implies $L_{\rm mp}(T'|\mathcal{C}_S) = L_{\rm mp}(T|\mathcal{C}_S)$ in view of (5), so T' has the same MPL score as any irreducible MPL tree. This completes the proof. □

Acknowledgments We thank the New Zealand Marsden Fund for its support, Daniel Huson for kind hospitality in Tübingen where this paper was written, and Andreas Dress for several helpful comments on an earlier version of this manuscript.

References

[1] L. Addario-Berry, B. Chor, M. Hallett, J. Lagergren, A. Panconesi and T. Wareham, Ancestral Maximum Likelihood of Phylogenetic Trees is Hard. Proceedings of WABI 2003 (Workshop on Algorithms in Bioinformatics, 2003).

[2] H.-J. Bandelt, P. Forster, B.C. Sykes, and M.B. Richards. Mitochondrial protraits of human populations using median networks. Genetics 141 743-753 (1995).

[3] D. Barry and J.A. Hartigan, Statistical analysis of hominoid molecular evolution. Stat. Sci. 2 191-210 (1987).

[4] P. Buneman. The recovery of trees from measures of dissimilarity. In Mathematics in the archaeological and historical sciences (Edited by F. R. Hodson, D. G. Kendall, and P. Tautu), pp.387–395. Edinburgh University Press (1971).

[5] L. Excoffier and P.E. Smouse, Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: Molecular variance parsimony. Genetics 136 343-359 (1994).

[6] J. Felsenstein, Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology, 27 401–410 (1978).

[7] P. O. Lewis, A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology 50 913-925 (2001).

[8] D. Penny, P.J. Lockhart, M.A. Steel, and M.D. Hendy, The role of models in reconstructing evolutionary trees. In Models in Phylogeny Reconstruction (Edited by R.W. Scotland, D.J. Siebert, and D.M. Williams), Systematics Association Special Volume No. 52, pp. 211-230, Clarendon Press, Oxford (1994).

[9] C. Semple and M. Steel, Phylogenetics. Oxford University Press (2003).

[10] M. Steel and D. Penny, Parsimony, likelihood and the role of models in molecular phylogenetics. Molecular Biology and Evolution 17 839–850 (2000).

[11] D.L. Swofford, G.J. Olsen, P.J. Waddell and D.M. Hillis, Phylogenetic inference. In Molecular Systematics (2nd edn.), (Edited by D. M. Hillis, C. Moritz, B. K. Marble), pp. 407–514 Sinauer, Sunderland, U.S.A. (1996).

[12] C. Tuffley and M. Steel, Links between maximum likelihood and maximum parsimony under a simple model of site substitution Bull. Math. Biol. 59 (3) 581-607 (1997).