LINKS BETWEEN MAXIMUM LIKELIHOOD AND MAXIMUM PARSIMONY UNDER A SIMPLE MODEL OF SITE SUBSTITUTION

CHRIS TUFFLEY and MIKE STEEL
Biomathematics Research Centre,
Department of Mathematics and Statistics,
University of Canterbury,
Christchurch, New Zealand
(E.mail: m.steel@math.canterbury.ac.nz)

Stochastic models of nucleotide substitution are playing an increasingly important role in phylogenetic reconstruction through such methods as maximum likelihood. Here, we examine the behaviour of a simple substitution model, and establish some links between the methods of maximum parsimony and maximum likelihood under this model. © 1997 Society for Mathematical Biology

1. Introduction. Stochastic models for nucleotide substitution are becoming increasingly important as a foundation for inferring phylogenetic trees from genetic sequence data. Such models allow for tree reconstruction through either maximum likelihood-based approaches or the fitting of transformed functions of the data to trees (see Swofford et al. (1996) for a recent survey). The models are also useful for analysing the performance of other, more conventional tree reconstruction methods, which are not explicitly based on such models, such as the popular maximum parsimony method (see, for example, Fitch (1971)). Such methods will indeed perform well (be "statistically consistent") for sequences that evolve under simple models with certain constraints (see, for example, Hendy and Penny (1989)), although without these constraints, the methods may be misled (Felsenstein, 1978). Thus, for certain data sets, maximum parsimony and maximum likelihood will agree, and in other cases, they will disagree. In this paper, we carry this analysis a little further for a simple model on any number of states, in which the rate of substitution is the same between any two states. In particular, we establish, for any tree and any number of states, an inequality between the probability of a character at a site and a function of the character's parsimony score on the underlying tree (Theorem 1). This bound becomes an equality for certain choices of parameters in the underlying model, and we completely characterise these choices when $r$ (the number of states) is 2 (Theorem 3).
We then use these results in four applications to the theory of phylogenetic analysis in Section 6. We establish three further cases in which maximum parsimony will agree with certain versions of maximum likelihood in the selection of trees and the reconstruction of ancestral states on a given tree. One of these results, Theorem 5, extends a result of Penny et al. (1994) from 2 to r states; another offers insight into the observations in Lockhart et al. (1996). We also generalise the example of Steel (1994) to an arbitrary number of species, and thereby show that the maximum likelihood function can be maximised at many points in the underlying parameter space.

2. Preliminaries.

Definitions 1 (Phylogenetic trees, characters). A phylogenetic tree is a tree $T = (V(T), E(T))$ having no vertices of degree 2, and such that each leaf (degree 1 vertex) is given a unique label from $\{1, \ldots, n\}$, where $n$ is the number of leaves of $T$. We say that $T$ is a tree on $n$ leaves, and write $[n]$ for $\{1, \ldots, n\}$. Where convenient, we identify each leaf with its label. If every internal (non-leaf) vertex of $T$ has degree 3, we say that $T$ is binary. In the case of rooted trees, we allow the root to have degree 2.

A function $\chi : [n] \to R$, where $R$ is a set of $r$ states, is an (r-state) character. When $r = 2$, $\chi$ is said to be binary. A function $\hat{\chi} : V(T) \to R$ is called a state function for $T$; if $\hat{\chi}$ is such that $\hat{\chi}|_{[n]} = \chi$ (that is, $\hat{\chi}$ agrees with $\chi$ on the leaves of $T$), then $\hat{\chi}$ is called an extension of $\chi$ (on $T$).

With each character $\chi$ and phylogenetic tree $T$ on $n$ leaves, we may associate a non-negative integer (the "length" of $\chi$ on $T$) as follows.

Definitions 2 (Length of $\chi$ on $T$, minimal extensions). If $\hat{\chi} : V(T) \to R$, then the changing number of $\hat{\chi}$, $ch(\hat{\chi})$, is the number of edges $\{u, v\}$ such that $\hat{\chi}(u) \neq \hat{\chi}(v)$. We say that a change occurs across $\{u, v\}$ under $\hat{\chi}$.

If $\chi : [n] \to R$, then the length of $\chi$ on the phylogenetic tree $T$, $l(\chi, T)$, is the minimum of $ch(\hat{\chi})$ over all extensions $\hat{\chi}$ of $\chi$ on $T$. An extension of minimal changing number is called a minimal extension of $\chi$ (on $T$).

Figure 1 illustrates these definitions.

In practical applications, the length of a character on a given tree is found using Fitch's algorithm, which is an order $n$ process for determining $l(\chi, T)$ and finding a minimal extension (Fitch (1971)). However, for theoretical purposes, $l(\chi, T)$ is usefully given in the two-state case by the following corollary of Menger's Theorem, a result that will be of use to us later. Although this is an often-quoted result (for example, Erdős and Székely (1993); Steel (1993b)), we include a proof as it does not follow directly from Menger's Theorem, and we believe a proof has yet to appear in the literature.