The Performance of Neighbor-Joining Methods of Phylogenetic Reconstruction

K. Atteson

Abstract. We analyze the performance of the popular class of neighbor-joining methods of phylogeny reconstruction. In particular, we find conditions under which these methods will determine the correct tree topology and show that these perform as well as possible in a certain sense. We also give indications of the performance of these methods when the conditions necessary to show that they determine the entire tree topology correctly, do not hold. We use these results to demonstrate an upper bound on the amount of data necessary to reconstruct the topology with high confidence.

Key Words. Phylogenetic reconstruction, Neighbor-joining, Evolutionary trees.

1. Introduction. The phylogenetic reconstruction problem is to determine the evolutionary relationships between a set of species typically from information contained in biomolecular sequence data. These evolutionary relationships may be represented by a phylogenetic tree, that is, a tree in which the leaves represent extant species and the internal nodes represent possibly extinct common ancestors of the extant species. Besides being of scientific interest, methods of phylogenetic reconstruction can have important applications to human health as, for instance, in the choice of drugs for targeting particular parasites [KDD]. Particular evolutionary relationships are often debated among biologists and different relationships can be obtained by the multitude of different phylogenetic reconstruction methods. In recent years, the growth of large-scale DNA sequencing has begun to provide a wealth of data for phylogenetic reconstruction.

While over the last several decades, many methods for phylogenetic reconstruction have been proposed, there have been few proven performance guarantees for these methods until recently. One such performance guarantee is given in [ABF], which demonstrates a method which outputs an additive distance matrix (tree distance) which is within a factor of 3 of the additive distance matrix which is closest under the $l_{\infty}$ norm on distance matrices (see Section 3.1 for definitions of these terms). Assuming the Cavender–Farris stochastic model of evolution, Farach and Kannan [FK] demonstrate sample-size bounds for obtaining a tree which is nearby the true model tree with respect to the variational distance between distributions defined by these trees. However, the performance guarantees of these works are difficult to interpret in terms of finding the tree which represents the actual evolutionary relationship between the species. Here, as in [ESSW], we take the view, prevalent among biologists, that the primary goal of phylogenetic reconstruction

1 This work was supported by NSF Grant Number BIR 9413215 while the author was at the University of Pennsylvania.

2 Yale University, Ecology and Evolutionary Biology, New Haven, CT 06520, USA. atteson@peapant.biology.yale.edu.

Received June 1, 1997; revised March 16, 1998. Communicated by D. Gusfield and M.-Y. Kao.
is to reconstruct all or some of the edges of the true tree. We give conditions under which
the neighbor-joining methods, some of the most popular of computationally efficient
methods, will do so. In particular, we find the radius around the true tree, for a certain
metric, in which the observed distances must be in order to guarantee that these methods
reconstruct all or some of the edges of the tree. These conditions yield upper bounds on
the sequence length needed for these methods to reconstruct all or some of the edges of
the tree. In fact, these methods do the best possible at reconstructing the entire tree, that
is, no method can be guaranteed to reconstruct the tree for observed distance matrices
in a larger radius around the true tree. When the observed distance matrix is not within
the radius mentioned above, so that we cannot show that the topology can be determined
completely, a slight modification of one of the neighbor-joining methods can be shown
to do the best possible at reconstructing some of the edges of the true tree.

In the next section we introduce some notation. In Section 3 we discuss the details
of the results of the paper and their significance. The subsequent sections of the paper
present the proofs of these results.

2. Some Notation. As mentioned previously, we represent evolutionary relationships
by trees, which we now define. We assume the reader is familiar with the basic concepts
of graph theory, see, e.g., [Bo]. Since we are trying to determine the topology of the
tree relative to the extant species which are represented as leaves, evolutionary trees
are leaf-labeled trees, that is, two evolutionary trees are the same if they have the same
topology relative to the leaves.

\begin{definition} A tree is a connected acyclic graph. We write \( V(T) \) and \( E(T) \) for the
vertex set and edge set, respectively, of a tree \( T \). A leaf of a tree is a node of degree 1. We
write \( L(T) \) for the set of leaves of tree \( T \). When the tree \( T \) is implicitly understood, we
write \( V, E, \) and \( L \) for the vertex, edge, and leaf sets of \( T \), respectively. Two trees \( T \) and \( T' \)
are (leaf-labeled) isomorphic, written \( T \sim T' \) if there is a bijection \( f: V(T) \rightarrow V(T') \)
which preserves adjacency, that is, \( E(T') = \{(f(v), f(v')) : (v, v') \in E(T)\} \), and
which preserves leaves, that is, \( f(v) = v \) for all leaves \( v \in L(T) \). Isomorphism is an
equivalence relation and we define the topology of a tree as the equivalence class of trees
isomorphic to it. Isomorphic trees are trees which are the same for our purposes and so
we sometimes blur the distinction between isomorphic trees and trees which are equal.
A rooted tree is a tree along with a special node called the root. A binary tree is a tree
in which every internal node has degree 3. A rooted binary tree is a tree having a single
node of degree 2, called the root, and such that every other internal node has degree 3.

For a tree \( T \) and an edge \( e \in E(T) \), the graph \( T - e \) is the graph obtained by removing
\( e \) from \( T \), that is, if \( T - e = (V, E - \{e\}) \). Note that \( T - e \) has exactly two components
and so partitions the set of leaves into two components. For \( k \in V \), we use the notation
\( L_k(T - e) \), or just \( L_k(e) \) if the tree \( T \) is implicitly understood, for the set of leaves
in the component of \( T - e \) containing \( k \) (see Figure 1). Let \( s(T - e) = \{L_k(T - e),
L - L_k(T - e) \} \) which we refer to as the split of \( T \) generated by \( e \). Let \( S(T) \) denote
the set of all splits of \( T \), that is, \( S(T) = \{s(T - e) : e \in E(T)\} \). Note that \( S(T) = S(T') \) if
and only if \( T \sim T' \) (see, e.g., [BD]).