Correcting large distances for phylogenetic reconstruction

Daniel H. Huson
Applied and Computational Mathematics
Princeton University, Princeton NJ 08544-1000
e-mail: huson@math.princeton.edu
Kelly Ann Smith
Department of Computer and Information Science
University of Pennsylvania, Philadelphia PA 19104
e-mail: kellyann@gradient.cis.upenn.edu
Tandy J. Warnow
Department of Computer Science
University of Arizona, Tucson AZ 85721
e-mail: tandy@cs.arizona.edu

May 28, 1999

A major computational problem in biology is the reconstruction of evolutionary trees from biomolecular sequences. Most polynomial time methods are *distance-based*, in that they estimate the evolutionary distances between every pair of taxa in the dataset, and then they use these estimates to infer the tree. Unfortunately, when the dataset contains pairs of sufficiently distantly related taxa, these techniques for estimating interleaf distances are undefined, and hence cannot be used. This is a real problem in phylogenetic data analysis, and so biologists have developed several ad-hoc methods for estimating distances under these conditions. However, there has been little study of these alternative techniques and how each affects the topological accuracy of the trees reconstructed using popular distance methods such as Neighbor Joining, Weighbor, and BIONJ.

In this poster, we describe the results of an extensive experimental performance study investigating these current techniques for computing distances, as well as some new techniques which we have developed. We simulate sequence evolution on 60 different model trees, and compare the topological accuracy of the trees computed by the best distance methods when given distance matrices calculated using our new techniques against the trees computed by the same methods when given the traditionally computed matrices. We show conclusively that our new techniques for computing distances greatly improve the accuracy of phylogenetic distance methods, often reducing the topological error rate by a factor of 10. This experimental result has tremendous implications for the practicing systematic biologist, who must analyze large and highly divergent datasets.