

Performance Analysis of Phylogeny Reconstruction Packages

Yamini Sridharan
3rd yr B.Tech (Hons)
Instrumentation Engineering, IIT Kharagpur

Guided by
David A Bader
Associate Professor and Regents' Lecturer
Electrical and Computer Engineering & Computer Science Department
University of New Mexico

Abstract

The project titled "Performance Analysis of Phylogeny Reconstruction Packages" is related to the field of computational biology. It addresses the concern that high-performance computing needs a revolutionary step to meet the growing high-end computing requirements of the new 21st century problems in computational science and biology.

We collected a suite of programs that solve the problem of Maximum Parsimony (MP) in computational biology and compared the performance with an open source program developed in our lab and optimized it further. A phylogeny is a representation of the evolutionary history of a collection of organisms or genes (known as taxa). The basic assumption of process necessary to phylogenetic reconstruction is repeated divergence within species or genes. A phylogenetic reconstruction is usually depicted as a tree, in which modern taxa are depicted at the leaves and ancestral taxa occupy internal nodes, with the edges of the tree denoting evolutionary relationships among the taxa. Reconstructing phylogenies is a major component of modern research programs in biology and medicine (as well as linguistics). Naturally, scientists are interested in phylogenies for the sake of knowledge, but such analyses also have many uses in applied research and in the commercial arena.

Existing phylogenetic reconstruction techniques suffer from serious problems of running time (or, when fast, of accuracy). The problem is particularly serious for large data sets: even though data sets comprised of sequence from a single gene continue to pose challenges (e.g., some analyses are still running after two years of computation on medium-sized clusters), using whole-genome data (such as gene content and gene order) gives rise to even more formidable computational problems, particularly in data sets with large numbers of genes and highly-rearranged genomes.

To date, almost every model of speciation and genomic evolution used in phylogenetic reconstruction has given rise to NP-hard optimization problems. Three major classes of methods are in common use. Heuristics (a natural consequence of the NP-hardness of the problems) run quickly, but may offer no quality guarantees and may not even have a well-defined optimization criterion, such as the popular neighbor-joining heuristic.

Optimization based on the criterion of maximum parsimony (MP) seeks the phylogeny with the least total amount of change needed to explain modern data. Finally, optimization based on the criterion of maximum likelihood (ML) seeks the phylogeny that is the most likely to have given rise to the modern data.

We have focused on the Maximum Parsimony (MP) problem, and the well-known codes such as Hennig86, PAUP, Phylip, and TNT. We have compared their performance and limitations on a variety of challenging data sets, and optimized our open source code, including designing my own implementation of the neighbor-joining technique.