



New Phylogeny Reconstruction Code, LosLobos Cluster, Mean Speedy Solution to Computational Problem

released December 12, 2000

Contact

David A. Bader
University of New Mexico
dbader@eece.unm.edu
505.277.6724

ALBUQUERQUE, NM — Using the largest open, production Linux supercluster, LosLobos, researchers at The University of New Mexico's **Albuquerque High Performance Computing Center** have achieved a nearly one-million-fold speedup in solving the computationally-hard phylogeny reconstruction problem for the family of twelve Bluebell species (scientific name: Campanulaceae) from the flowers' chloroplast gene order data. (The problem size includes a thirteenth plant, Tobacco, used as a distantly-related outgroup). Phylogenies derived from gene order data may prove crucial in answering some fundamental open questions in biomolecular evolution. Yet very few techniques are available for such phylogenetic reconstructions.

One method is breakpoint analysis, developed by Blanchette and Sankoff for solving the "breakpoint phylogeny." Earlier experiments with various techniques for reconstructing phylogenies from gene order data suggested that Sankoff and Blanchette's implementation of BPAnalysis is much too slow. On a collection of Campanulaceae with 13 genomes of 105 gene segments, it is estimated that Sankoff and Blanchette's BPAnalysis would take well over 200 years to complete—an estimate based on the average number of trees processed by the code per unit time and extended to the 13,749,310,575 tree topologies on 13 leaves.

Professors David A. Bader (EECE, UNM) and Bernard M.E. Moret (Computer Science, UNM) along with Professor Tandy Warnow (Computer Science, University of Texas - Austin) have solved the Campanulaceae problem in about one hour and forty minutes on the 512-processor UNM/Alliance LosLobos supercluster, an IBM Netfinity cluster with 512 733MHz Pentium III processors, interconnected with four 64-way Myrinet 2000 switches, and running the Linux 2.2 operating system. The parallelization uses MPI and includes techniques for concurrently evaluating candidate trees and sharing improved upper and lower bounds by the processors.

The new phylogeny reconstruction code, GRAPPA, is freely available as open source from <http://www.cs.unm.edu/~moret/GRAPPA/> and shows the power of algorithmic engineering techniques. Run on a single processor, GRAPPA performs 2500 times faster than the BPAnalysis code, using similar algorithms, but with the implementation paying close attention to the detail of cache-sensitive and efficient data structures and algorithms. GRAPPA is able to take full advantage of parallel computers like the LosLobos supercluster, for additional speedups. Hence, the total speedup for our solution (going from an estimated 200 years down to 100 minutes) is one million—nearly 6 orders of magnitude.

This research is part of Moret and Bader's NSF Information Technology Research project "Algorithms for Irregular Discrete Computations on SMPs" based at the High Performance Computing, Educational and Research Center (HPCERC), a strategic center at The University of New Mexico.

[Headlines Archive](#)

