A FLOWER'S FAMILY TREE

03.02.01

by J. William Bell, NCSA Senior Science Writer

HPCwire

Gene data allow researchers to recover the evolutionary history of plants, but even the smallest dataset can require impossibly large computations. Using an Alliance Linux cluster and newly designed software, a team from the University of New Mexico and the University of Texas have increased the speed of the process millionfold for one family of plants.

If you're looking for extreme diversity, consider bluebells, officially the Campanulaceae family. The 2,000 species of these plants with bell-shaped flowers show amazing variety. They live everywhere on Earth except the Sahara, Antarctica, and the northern extremes of Greenland. Some are annuals, and others are perennials. Despite their common name, the flowers can be blue, purple, red, or yellow. The bluebells of North America typically grow close to the ground, while Asian species can grow as tall as eight feet.

As fascinating as that diversity is, it's not the sort of thing that computational scientists usually get excited about. Uncovering how that diversity came to be has captured the attention of a team of researchers at Alliance partner University of New Mexico and the University of Texas, though. Using the 512-processor LosLobos Linux Pentium III supercomputing cluster at the Albuquerque High Performance Computing Center, the team has created a phylogeny reconstruction -- or evolutionary history -- of 12 bluebell species, predicting all of the steps that take these species back to a single common ancestor. To meet the challenge, they created a whole new piece of software known as GRAPPA.

"In our context, we can answer the question of why we do this work by saying, 'We just want to know, and we like a challenge.' But phylogeny reconstruction has very significant implications to pharmaceutical design and in other industries," says Bernard Moret, a computer science professor at the University of New Mexico.

Before beginning to reconstruct the species' evolutionary history, the team had to decide which model of evolution would inform their project. "There are a number of models out there that talk about how evolution occurs. They represent the different general principles that evolution might have followed," says David A. Bader, an electrical and computer engineering professor at the University of New Mexico.

Some models figure the likelihood of given events within the species' history, such as duplications, deletions, and insertions of genes, and use that information to create possible histories. These methods are computationally expensive, and gathering the data necessary to run them is difficult.

Rather than taking a statistical approach, however, the model used by Bader and Moret's team was built on an idea known as parsimony. "By saying parsimony, we're basically saying that nature is efficient," says Moret. "It gives rise to new species through the least amount of change. Parsimony is founded on the same principle as Occam's razor: the simplest explanation is the best. Here, the shortest evolutionary path is the best."

The team arrived at this theory by following the lead of biologists like their colleague Robert Jansen, chairman of the University of Texas at Austin's Section of Integrative Biology, who studies evolution and collects the gene data used by the team. The gene data are taken from the chloroplast of each species of Campanulaceae. Chloroplasts provide energy for the plant cells and do not occur independently outside of cells. They also make excellent candidates for phylogenic reconstruction because each chloroplast has a single chromosome. The genes on the chloroplasts' chromosomes have been sequenced, and biologists like Jansen hypothesize that evolution occurs through a mechanism known as inversion. Inversion contributes to evolution by changing the order and orientation of a sequence of genes within a genome. At times inversions may even undo themselves, returning the order of genes on the chloroplast chromosome to its former state.

Parsimony may imply efficiency, but, the efforts necessary to actually build reconstructions, clearly leave simplicity far behind.

Each phylogenetic reconstruction, called a tree, represents one possible history of the species. The Campanulaceae project begins with 12 modern species of bluebell and a single species of tobacco. Tobacco is used as an outgroup, a species that is clearly very distant from the others, and is used to identify the root of the tree. To predict the evolutionary history, almost 14 billion trees must be built and compared to one another. Bader, Moret and their colleague Tandy Warnow of the computer science department at the University of Texas at Austin go far beyond constructing the underlying tree and its eventual outcome, also calculating gene order for each predicted ancestor within the trees. That means a whopping 100 billion genomes must be reconstructed.

"We need to build as complete a picture of the ancestry as possible," says Moret. "Computing ancestral gene orders also enables us to put differing evolutionary models to the test."

For these computer scientists, the reconstructions amount to a massive optimization problem -- a game of creating all the possible evolutionary scenarios that could have occurred and then narrowing down those billions of options to a single best solution. The process begins with the raw gene data taken from the chloroplast of each species of Campanulaceae. The bluebell chloroplasts' single chromosomes are each made up of 105 gene segments. The genes within the individual genomes are always the same. That is, all 13 genomes have identical genes and identical lengths, but in different species the genes appear along the chromosome in a different order or orientation.

The team's code generates trees one at a time. Once a tree is generated, its internal nodes -- the intervening ancestors that come between the individual species and their final common ancestor -- are labeled by the software.

Labels, which consist of the gene order data for a given node, are derived through a complex optimization process based on the notion of breakpoints.

A breakpoint occurs any time two genes are adjacent in one genome but are not adjacent in a genome to which the first is compared. An internal node's label is derived by finding the gene order that minimizes the number of breakpoints between a node and its three closest neighbors. "This is where the parsimony criterion comes in," say Moret. "We find a label that minimizes the amount of change at this place in the tree." A travelling salesperson problem solver -- a common, if expensive, mathematical method of solving optimization problems -- is used to find the median, calculating the hypothesized gene order data for each node.

In the initial labeling of the tree, nodes that represent known data are separated by many intervening nodes. As a result, a great deal of approximation is used in the early stages. Multiple passes at a tree refine the approximation. Once the initial labels are assigned, each node has closer neighbors that can be used to find the breakpoint median. The code recalculates the nodes' labels based on these new data, repeating this process again and again and recalculating any node that saw changes in one of its neighbors in the previous pass, until the tree stabilizes.

Labeling and refining the trees is by far the most challenging step. "Computing a single median is intractable in itself, and we solve these over and over. It's a very computationally intensive procedure," Moret says. With this step complete, each tree is scored, using inversions as the metric. These scores show researchers which trees are most parsimonious. Thus, the scores also show which evolutionary history best fits the model and give a plausible snapshot of each genome within that history.

"In completing so many trees, you see very good and very bad ones. Ones that match up very closely and ones that don't," says Bader. "What we look for is commonality. A consensus tree that we can dig in deep on and that will give us a foothold to larger phylogenies."

But mostly, the team lets biologists who use their data look for commonality, consensus, and the like. Bader, Moret, Warnow, and their students are in it for the computational challenge. And the Campanulaceae phylogeny reconstruction certainly gives them that.

At the beginning of the project, they wanted nothing more than to improve BPAnalysis, a code used for breakpoint phylogeny research. BPAnalysis would have required over 200 years to generate, label, and score the nearly 14 billion trees represented in the Campanulaceae problem.

"We started with the goal of reimplementing BPAnalysis from the ground up. It was simply much too slow. We wanted to gain efficiency and speed by improving the algorithm and by parallelizing the code. Just focusing on one of those wouldn't have given us the kinds of improvements we've seen," says Moret.

BPAnalysis relabels every internal node each time it refines a tree; GRAPPA recalculates labels for only those nodes that could possibly show a change. BPAnalysis looks at identical strings of genes over and over again, even those matching other gene fragments that have already been analyzed; GRAPPA identifies common subsequences and condenses them, leaving fewer genes to be considered. BPAnalysis runs on only one processor; GRAPPA scales linearly to hundreds of processors running in parallel. GRAPPA leaves a scant memory footprint of only 1.6 megabytes and can work almost entirely in a computer's cache memory thanks to a working set of less than 0.5 megabytes. GRAPPA is also modular, allowing different methods of calculating the nodes' labels to be swapped in and out easily.

When recently tested on the Campanulaceae problem on Albuquerque's LosLobos computing cluster, GRAPPA showed incredible results. The massive reconstruction was completed in one hour and 40 minutes on the machine's 512 733-MHz Pentium III processors -- a 1,000,000-fold speedup over BPAnalysis.

Now that's the sort of thing a computational scientist can get excited about -- an example of cluster computing in full bloom.

Relevant URLs:

--Access story:
http://access.ncsa.uiuc.edu/CoverStories/phylogeny/>http://access.ncsa.uiuc.edu/CoverStories/phylogeny/

--GRAPPA homepage:
http://thelma.cs.unm.edu/~moret/GRAPPA/>http://thelma.cs.unm.edu/~moret/GRA

--David A. Bader's homepage:
http://www.eece.unm.edu/~dbader/>http://www.eece.unm.edu/~dbader/

****