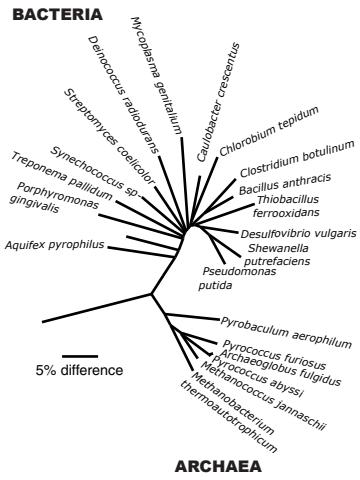
Tim Lilburn
Ribosomal Database Project II, Center for Microbial Ecology
294 Giltner Hall
353-3842; lilburn@pilot.msu.edu

The calculation of large phylogenetic trees



Phylogenetic trees are graphical representations of inferred evolutionary relationships between the organisms (or genes). The inferences are most often based on aligned nucleic acid or amino acid sequence data. The overview of the evolutionary history of a group of organisms provided by a phylogenetic tree can inform studies in paleontology, ecology epidemiology, biogeography and in many other fields. At the Ribosomal Database Project (RDP), phylogenetic trees are also used as a database interface for users, allowing them to select and download groups of sequences that are evolutionarily related to the sequence(s) of interest. The current RDP phylogenetic tree for the prokaryotes includes 6,205 sequences, most of which represent individual, named prokaryotes. However, this tree has not been generated using standard phylogenetic methods but instead by, in essence, accreting new sequences to a pre-existing tree. The topology of the tree has become unstable and the placement of new sequences on the tree is uncertain in some cases. Hence, the need for a recalculated tree has arisen.

Historically, most research on the generation of phylogenies has concentrated on relatively small datasets involving as few as four taxa, far smaller than the RDP is envisioning. Some studies involving datasets of several hundred sequences have been done, but they employed methods known either to be limited by the search method used to find the best solution or to be subject to error when highly divergent data are used. Maximum likelihood estimation offers reliable results that are also amenable to statistical evaluation, but it cannot be used for datasets of more than about 120 sequences. We are thus faced with two possible courses: firstly, increase the computational resources applied to the maximum likelihood method as it exists, or, secondly, develop new stategies using maximum likelihood estimation.

To investigate the first possibility, we are collaborating with researchers at the Lawrence Berkeley National Laboratory. This group will be optimizing the parallel implementation of the maximum likelihood algorithm used to generate the current tree, and using the improved algorithm to generate large trees on a Cray T3E supercomputer. The second possibility, the development of new strategies for the application of maximum likelihood methods, is an attractive one. The use of a genetic algorithm to search the sample space could speed up phylogenetic estimation, especially if the search could be conducted in steps. In the first step, the objective function for scoring phylogenetic estimates would be an optimality criterion from a faster, less reliable phylogenetic method. The best estimates from this step could then serve as a starting point for a search that uses the maximum likelihood optimality criterion as the objective function.