Inferring Phylogenies with RAxML-VI-HPC

Alexandros Stamatakis
Swiss Federal Institute of Technology Lausanne
School of Computer & Communication Sciences
Alexandros.Stamatakis@epfl.ch

Abstract: Randomized Axelerated Maximum Likelihood version VI for High Performance Computing (RAxML-VI-HPC) is an efficient program for phylogenetic analyses with thousands of taxa under the popular Maximum Likelihood (ML) criterion. The software demonstration will cover the basic features of RAxML-VI-HPC and will show how to deploy it on sequential and parallel computer architectures for real-world phylogenetic studies. To the best of the authors knowledge RAxML-VI-HPC has been used to infer trees on the two largest data matrices analyzed under ML to date: a 25,057-taxon alignment of protobacteria (1,463 bp) and a 2,182-taxon alignment of mammals (51,089 bp). The program is available as open-source code at `diwww.epfl.ch/~stamatak` (software frame).

1 RAxML Software Demonstration

RAxML-VI-HPC [SLM05] is a program for large-scale phylogenetic analyses under the ML criterion. The program is currently the featured application on the CIPRES (Cyber-Infrastructure for Phylogenetic RESearch, www.phylo.org) project web site. It allows for inference of huge ML trees under sufficiently complex substitution models within reasonable times. For example, on a medium-sized cluster with 136 CPUs it was feasible to compute 1,000 non-parametric bootstraps on a 1,500 taxon alignment. The author has recently given a series of invited talks about RAxML at various research institutes and international conferences:

- Invited Talk: "RAxML-VI: A Program for Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa: How it works and how to use it", Botanic Garden of Munich, Germany, May 2006.
- Invited Talk: "RAxML-VI: A program for large-scale Maximum Likelihood-based phylogenetic analyses; How it works and how to use it", Max-Planck-Institute for Developmental Biology, Tübingen, Germany, February 2006.
- Invited Talk: "Using RAxML in practice", CIPRES (Cyberinfrastructure for Phylogenetic Research) project All Hands Meeting 2006, University of Texas at Austin, Texas, February 2006.
- Invited Talk: "Computing Huge Trees with Maximum Likelihood: An HPC Perspective", Workshop on "The Problems of Phylogenetic Analysis of Large Datasets", Mathematical Biosciences Institute, Columbus, Ohio, December 2005.
- Tutorial: Joint half-day tutorial with David Bader and Usman Roshan on "Computational Grand Challenges in Assembling the Tree of Life: Problems & Solutions". Presented at 18th IEEE/ACM Supercomputing Conference (SC2005), Seattle, Washington, November 2005.

Some of the largest published ML-based analyses to date have been conducted with RAxML [RHJP05] [LBT+05] [LHW+06]. On-going work includes the computation of a backbone tree for Bacteria with approximately 9,000 taxa (Pace Laboratory, University of Colorado at Boulder), a phylogeny for Acer with 582 taxa (Guido Grimm, Universität Tübingen), and the analysis of a mammalian multi-gene alignment comprising 2,182 sequences (Olaf Bininda-Emonds, Technische Universität München). The program is also part of the greengenes project [DHL+06] (greengenes.lbl.gov).

A recent performance study on real world datasets $\geq$ 1,000 taxa reveals that it is able to find better trees in less time and with lower memory consumption than other current programs (IQPNNI, PHYML, GARLI). Moreover, RAxML-VI-HPC has been parallelized with MPI (Message Passing Interface) for LINUX PC clusters to enable parallel non-parametric bootstrapping. In addition, it has been parallelized with OpenMP [SOL05] to accelerate inferences on large memory-intensive multi-gene alignments.

The presentation will provide guidelines under what circumstances which version should be used and how alignment-dependent search parameters are best determined. The usage of the GTR+CAT approximation [Sta06] in RAxML which can be used as a replacement for the computationally more intensive GTR+Γ model will also be explained. Finally, several new program options of RAxML, such as the possibility to infer trees under mixed/partitioned models and to specify various forms of constraint trees, will be presented.

References

[DHL+06] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl. Environ. Microbiol.*, 72(7):5069–5072, 2006.

[LBT+05] R.E. Ley, F. Backhed, P. Turnbaugh, C.A. Lozupone, R.D. Knight, and J.I. Gordon. Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31):11070–11075, 2005.

[LHW+06] R. E. Ley, J. K. Harris, J. Wilcox, J. R. Spear, S. R. Miller, B. M. Bebout, J. A. Maresca, D. A. Bryant, M. L. Sogin, and N. R. Pace. Unexpected Diversity and Complexity of the Guerrero Negro Hypersaline Microbial Mat. *Appl. Envir. Microbiol.*, 72(5):3685 – 3695, May 2006.

[RHJP05] C.E. Robertson, J.K. Harris, J.R.Spear, and N.R. Pace. Phylogenetic diversity and ecology of environmental Archaea. *Current Opinion in Microbiology*, 8:638–642, 2005.

[SLM05] A. Stamatakis, T. Ludwig, and H. Meier. RAxML-III: A Fast Program for Maximum Likelihood-based Inference of Large Phylogenetic Trees. *Bioinformatics*, 21(4):456–463, 2005.

[SOL05] A. Stamatakis, M. Ott, and T. Ludwig. RAxML-OMP: An Efficient Program for Phylogenetic Inference on SMPs. In *Proc. of PaCT05*, pages 288–302, 2005.

[Sta06] A. Stamatakis. Phylogenetic Models of Rate Heterogeneity: A High Performance Computing Perspective. In *Proc. of IPDPS2006*, Rhodos, Greece, 2006.