Bioinformatics Advance Access published August 23, 2006

BIOINFORMATICS

RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models

Alexandros Stamatakis [a*]

[a]Swiss Federal Institute of Technology Lausanne, School of Computer & Communication Sciences, Lab Prof. Moret, STATION 14, CH-1015 Lausanne, Switzerland

Associate Editor: Keith A Crandall

ABSTRACT

Summary: RAxML-VI-HPC (Randomized Axelerated Maximum Likelihood for High Performance Computing) is a sequential and parallel program for inference of large phylogenies with Maximum Likelihood (ML). Low-level technical optimizations, a modification of the search algorithm, and the use of the GTR+CAT approximation as replacement for GTR+Γ yield a program that is between 2.7 and 52 times faster than the previous version of RAxML. A large-scale performance comparison with GARLI, PHYML, IQPNNI, and MrBayes on real data containing 1,000 up to 6,722 taxa shows that RAxML requires at least 5.6 times less main memory and yields better trees in similar times than the best competing program (GARLI) on datasets up to 2,500 taxa. On datasets $\geq$ 4,000 taxa it also runs 2-3 times faster than GARLI. RAxML has been parallelized with MPI to conduct parallel multiple bootstraps and inferences on distinct starting trees. The program has been used to compute ML trees on two of the largest alignments to date containing 25,057 (1,463 bp) and 2,182 (51,089 bp) taxa respectively.

Availability: diwww.epfl.ch/~stamatak

Contact: Alexandros.Stamatakis@epfl.ch

1 INTRODUCTION

Phylogenetic inference with the ML method is NP-hard (Chor and Tuller, 2005). Despite the algorithmic complexity and the high computational cost of ML, significant progress has been achieved with the release of fast and accurate programs such as PHYML (Guindon and Gascuel, 2003), IQPNNI (Minh *et al*., 2005), MrBayes (Ronquist and Huelsenbeck, 2003), GARLI (Zwickl, 2006), and RAxML (Stamatakis *et al.*, 2005). Most of these programs allow for inference of 1,000 taxon trees on a single CPU in less than 24 hours.

This paper describes the new version of RAxML (RAxML-VI-HPC (v2.0.1)) which is significantly faster than the previous versions of RAxML due to simple, yet very efficient technical optimizations and a slight alteration of the search algorithm. In addition, RAxML has been parallelized with MPI to enable parallel bootstrapping and multiple inferences on distinct starting trees on PC clusters. Moreover, it implements bifurcating and multifurcating constraint trees and the capability to assign and estimate separate model parameters [1] for individual genes of multi-gene alignments (mixed/partitioned models).

The main focus is on the computation of huge trees ($\geq$ 1,000 taxa) for real-world data and the comparative performance study with GARLI, IQPNNI, MrBayes, and PHYML. Since the efficiency of the novel optimizations in RAxML-VI-HPC increases with the number of taxa, less significant performance improvements will be observed on smaller datasets. Performance comparisons of RAxML with other popular ML programs on smaller datasets, including simulated alignments, can be found in Hordijk and Gascuel (2005); Stamatakis *et al.* (2005) and Zwickl (2006). Finally, the experimental study also shows that the GTR+CAT approximation (see Stamatakis (2006) for a detailed description) can be efficiently deployed as a replacement for the significantly more compute- *and* memory intensive GTR+Γ model.

Some of the largest published ML-based analyses to date have been conducted with RAxML (Robertson *et al.*, 2005; Ley *et al.*, 2005, 2006). On-going work includes the computation of a backbone tree for Bacteria with approximately 9,000 taxa, a phylogeny for Acer with 582 taxa, and the analysis of a mammalian multi-gene alignment comprising 2,182 sequences.

2 OPTIMIZATIONS OF RAXML

A detailed description of the optimizations listed below is provided in the on-line supplement. The main improvements cover:

- An efficient mechanism to store and re-store topologies and branch lengths via rearrangement descriptors
- A consequent re-use of partial likelihood vectors
- A dynamic adaptation of the rearrangement distance
- Low-level optimization of the GTR+CAT and GTR+Γ likelihood functions
- An efficient re-implementation of Maximum Parsimony starting tree computations

An important and generally applicable insight from those optimizations is that storing and re-storing an unrooted tree topology with $2n-3$ branch lengths and $2n-2$ nodes can become a major performance bottleneck for trees with more than 1,000 taxa. It is thus important to store alternative topologies rather as a sequence of topological changes applied to the current topology rather than as complete data object. Only the consequent avoidance of storage operations reveals the actual power of the Lazy Subtree Rearrangement (LSR) mechanism introduced in Stamatakis *et al.* (2005).

Another issue which becomes important for huge trees is to determine a "good" rearrangement distance, i.e. re-insertion radius for the LSR moves. In RAxML-VI the algorithm initially determines the best rearrangement distance by applying distances of 5, 10,..., 25 for one iteration of LSRs, to the starting tree. The minimum rearrangement distance which yields the best likelihood improvement on the starting tree is then selected for the inference. Despite the extra computations which are performed, a "good" rearrangement distance pays off in terms of likelihood units for huge alignments with

*to whom correspondence should be addressed

[1] CAT and Γ can not be used simultaneously in the same analysis.

large evolutionary diameters (e.g. the 6,722 and 7,769 taxa alignments, see Table 2 in the on-line supplement).

3 RESULTS & DISCUSSION

The exact experimental setup as well as the results are described in detail in the on-line supplement. Table and Figure numbers also refer to the on-line supplement.

Results in Table 2 show that RAxML-VI-HPC clearly outperforms RAxML-V in terms of inference times. In addition, due to the usage of a "good" rearrangement setting it also yields significantly better Log Likelihood values on the larger and more diverse datasets $\geq 4,000$ taxa. Figure 3 shows the significant computational advantages of the GTR+CAT over the GTR+Γ implementation in RAxML-VI.

Tables 3 through 6 indicate that RAxML-VI-HPC outperforms other current sequential phylogeny programs, on huge datasets with respect to inference times, memory consumption, as well as final Log Likelihood values. In addition, the performance advantage with respect to run-times increases with growing alignment size (Table 5). Another important result is that the GTR+CAT approximation (Table 3) can be used to significantly reduce memory consumption and still yield significantly better GTR+Γ likelihood values (Table 4) than competing programs.

GARLI terminated within approximately the same time as RAxML-VI-HPC on the 6 smaller datasets and yielded the second-best likelihood score in all cases. This is an astonishing achievement for several reasons: GARLI implements a genetic search algorithm and was executed under GTR+Γ. Moreover, it maintains a whole population of trees in memory, including *some* intelligently selected (Zwickl, 2006) partial likelihood vectors as well as *all* tree topologies. Thus, it is expected to be slower than the RAxML hill-climbing algorithm. This extraordinary performance is due to the sophisticated implementation of the likelihood function and promising algorithmic ideas (Zwickl, 2006) such that the forthcoming publication about GARLI is surely something to look forward to. Note that, the parallel genetic search algorithm of GARLI performs a distinct and more thorough search, that yields e.g. better final trees on the 1,000 taxon alignment (Zwickl, 2006). However, the focus of the current study is on the strictly sequential versions of all programs.

The performance of the new version of MrBayes is also remarkable. Given, that it has to maintain four distinct Markov chains, the relatively low memory consumption in combination with acceptable likelihood values after 60 hours under GTR+Γ, the performance is quite impressive. Because Bayesian inference conceptually differs from pure ML-based inference, a comparison based on likelihood scores is certainly not fair since it uses MrBayes as an ML heuristic. MrBayes has mainly been included due to its popularity.

IQPNNI and PHYML both suffer from a relatively inefficient technical implementation. The high memory consumption of IQPNNI and PHYML is due to a different memory organization which uses 2 likelihood vectors per branch ($3n - 6$ vectors) instead of 1 per inner node ($n - 2$ vectors).

Moreover, PHYML uses NNI moves which only exploit a very small fraction of the search space. A solution to this problem has been proposed by Hordijk and Gascuel (2005). However, the respective program is currently only available as proof-of-concept implementation (Hordijk, Gascuel, personal communication) and can not be used for large trees due to numerical problems.

In the final analysis, it can be stated that technical implementation aspects are becoming increasingly important and can yield significant performance improvements. In addition, in all programs there exist excellent algorithmic ideas which in the optimal case could significantly advance the field, when merged into one program.

4 CONCLUSION & FUTURE WORK

The new version VI of RAxML has been presented, which incorporates efficient technical optimizations, parallel OpenMP- and MPI-based implementations, and a mixed model implementation. A thorough experimental study on large real-world datasets shows that RAxML can find better trees with a significantly lower memory consumption within similar or less time than the best competing program.

Future work will mainly cover the development of new methods for rapid bootstrapping. Despite the fact, that RAxML and GARLI allow for inference of huge trees with ML in reasonable times, conducting a full biological analysis still requires at least 100 or 1,000 bootstraps which places the computational burden much higher than for the inference of a single ML tree.

ACKNOWLEDGMENTS

The author would like to thank Derrick Zwickl, Wim Hordijk, Olivier Gascuel, B.Q. Minh, L.S. Vinh, and Bret Larget for useful discussions on experimental setup and their programs. He would also like to thank Usman Roshan, Charles Robertson, Josh Wilcox, Robin Gutell, and Daniel Dalevi for providing the alignment data.

REFERENCES

Chor,B. and Tuller,T. (2005) Maximum likelihood of evolutionary trees: hardness and approximation. *Bioinformatics*, **21** (1), 97–106.

Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52** (5), 696–704.

Hordijk,W. and Gascuel,O. (2005) Improving the effi ciency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics*, **21** (24), 4338–4347.

Ley,R., Backhed,F., Turnbaugh,P., Lozupone,C., Knight,R. and Gordon,J. (2005) Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America*, **102** (31), 11070–11075.

Ley,R.E., Harris,J.K., Wilcox,J., Spear,J.R., Miller,S.R., Bebout,B.M., Maresca,J.A., Bryant,D.A., Sogin,M.L. and Pace,N.R. (2006) Unexpected diversity and complexity of the guerrero negro hypersaline microbial mat. *Appl. Envir. Microbiol.*, **72** (5), 3685 – 3695.

Minh,B.Q., Vinh,L.S., von Haeseler,A. and Schmidt,H.A. (2005) pIQPNNI: parallel reconstruction of large maximum likelihood phylogenies. *Bioinformatics*, **21** (19), 3794–3796.

Robertson,C., Harris,J., J.R.Spear and Pace,N. (2005) Phylogenetic diversity and ecology of environmental Archaea. *Current Opinion in Microbiology*, **8**, 638–642.

Ronquist,F. and Huelsenbeck,J. (2003) Mrbayes 3: bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Stamatakis,A. (2006) Phylogenetic models of rate heterogeneity: a high performance computing perspective. In *Proc. of IPDPS2006*, Rhodos, Greece.

Stamatakis,A., Ludwig,T. and Meier,H. (2005) Raxml-iii: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, **21** (4), 456–463.

Zwickl,D. (2006). *Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biologiical Sequence Datasets under the Maximum Likelihood Criterion*. PhD thesis, University of Texas at Austin.