Chapter 3
High-Performance Phylogenetic Inference

David A. Bader and Kamesh Madduri

Abstract Software tools based on the maximum likelihood method and Bayesian methods are widely used for phylogenetic tree inference. This article surveys recent research on parallelization and performance optimization of state-of-the-art tree inference tools. We outline advances in shared-memory multicore parallelization, optimizations for efficient Graphics Processing Unit (GPU) execution, as well as large-scale distributed-memory parallelization.

Keywords Phylogenetic tree inference · Maximum likelihood · Bayesian inference · Parallel algorithms · Algorithm engineering

3.1 Introduction

Computational phylogenetics is an active research area. A variety of algorithms and software tools exist for the compute-intensive task of tree inference. Early methods were based on distance-based similarity clustering [18, 43, 46] and on the Maximum Parsimony principle [17, 20]. These simple methods are now subsumed by more sophisticated algorithms. Probabilistic approaches, specifically Maximum Likelihood (ML)-based [15] methods and Bayesian inference methods [25, 42], currently dominate the landscape of tree inference software. As of October 2018, the OMICtools website [39] lists 266 software tools in the Phylogenetic Inference category. Felsenstein's Phylogeny Programs web page [14] lists more than 90 ML-based methods and more than 25 Bayesian inference methods. The Cyberinfrastructure for Phylogenetic Research (CIPRES) Science Gateway Version 3.3 [9, 30] currently supports 15 parallel programs for tree inference and sequence alignment. Phylogeny.fr [10] is another long-running web portal for phylogenetic analysis.

D. A. Bader (✉)
Georgia Institute of Technology, Atlanta, GA, USA
e-mail: bader@cc.gatech.edu

K. Madduri
Pennsylvania State University, University Park, PA, USA
e-mail: madduri@cse.psu.edu

© Springer Nature Switzerland AG 2019
T. Warnow (ed.), *Bioinformatics and Phylogenetics*, Computational Biology 29,
https://doi.org/10.1007/978-3-030-10837-3_3

Popular, free, and open-source tools include PHYLIP [13], RAxML [47, 48], PhyML [22, 23], MrBayes [42], and BEAST 2 [6]. Nearly all of these tools support some form of parallelism.

Moret played a seminal role in establishing the research area of high-performance computational phylogenetics by leading the development of GRAPPA and associated algorithms [21, 32–34]. GRAPPA is a maximum parsimony-based suite of programs for phylogeny reconstruction using genome rearrangements. For breakpoint phylogeny reconstruction, using efficient data structures and optimizations, GRAPPA was engineered to perform nearly 2500 times faster than the original Sankoff–Blanchette algorithm [44] on a single processor. When executed on a 512-processor cluster, GRAPPA achieved an awe-inspiring million-fold speedup [5]. GRAPPA is a significant milestone in the areas of algorithm engineering and parallel phylogenetic inference. Many of the current probabilistic inference methods take aligned sequences, typically DNA or amino acid sequences, as input. The quality of multiple sequence alignment will thus directly impact the quality of trees generated. The methods also assume a model for site evolution and estimate model parameters. The Generalized Time Reversible (GTR) model [51] is a commonly used model for inference on DNA and amino acid sequences. For additional background on statistical methods, please refer to [24, 52]. Current software tools support a wide variety of evolutionary models.

Likelihood calculations [48] constitute a significant fraction of the overall running time of both ML and Bayesian inference methods. We first discuss performance optimizations and parallelization strategies to speed up likelihood calculations. In Sect. 3.3, we discuss miscellaneous execution time-reducing implementation changes and approaches to improve multi-node performance. (See also the chapters by Stamatakis and Guindon & Gascuel in this book for more about this subject.).

3.2 Faster Likelihood Calculations

ML-based tree reconstruction has been shown to be an NP-hard optimization problem under various assumptions [8, 41]. An exponential number of tree configurations need to be evaluated in order to find the optimal solution, and this is intractable with even a modest number of organisms. Thus, software tools employ a variety of heuristics to reduce the search space. For each tree topology, evaluating the likelihood function involves postorder tree traversal and propagating likelihood values from the tips to the root according to Felsenstein's pruning algorithm [15]. Likelihood computations also appear in Bayesian inference methods. These computations are both floating-point operation and memory-intensive, and take up a dominant fraction of the running time in state-of-the-art programs.

Fortunately, there is abundant fine-grained parallelism to exploit in these likelihood calculations. The partial likelihood scores at each site can be computed independent of other sites. Since the number of sites can vary from thousands to millions, the multiple sequence alignment output can be further split into partitions that can be evaluated independently. Likelihood calculations are also prone to floating-point rounding errors and need to be evaluated carefully. The community is moving away from monolithic codes and transitioning to using library-based approaches. Bio++ [12] is an early example of a C++ library with optimized implementations of key phylogenetic primitives. BEAGLE (Broad-platform Evolutionary Analysis General Likelihood Evaluator) [4, 50] is a library and an application programming interface for parallel likelihood calculations. BEAGLE routines can be used in both ML-based inference methods and Bayesian methods. In addition to partitioning of alignment sites, fine-grained data parallelism is possible across rate categories and state values. BEAGLE includes SSE implementations for CPUs, as well as CUDA and OpenCL implementations of routines for GPUs.

BEAGLE also provides interfaces to the inference tools BEAST 2 [6], BEAST [11], MrBayes [42], and GARLI [54]. It is shown that the library-based approaches outperform the standalone implementations, and that the GPU-based approach delivers a significant performance boost over a CPU implementation. Recent work by Ayres and Cummings [3] explores additional tuning opportunities to further improve the performance of BEAGLE routine.

Phylogenetic Likelihood Library (PLL) [19] is another open-source library inspired by Bio++ and BEAGLE. PLL is used by ExaML [29] and RAxML-NG [28], two recent and modern implementations of RAxML, and also interfaces with IQ-TREE [37], a recent ML-based inference package. PLL has a backend for the Intel Xeon Phi accelerator, Python bindings, includes many SIMD implementations, and also supports MPI parallelization. It is shown to be 1.9–4× faster [19] than BEAGLE on benchmarks.

3.3 Performance Optimizations and Multi-node Parallelism

Bayesian methods [7, 52] approximate the posterior distribution of evolutionary parameters using Bayes' theorem. The methods rely on sampling approaches such as the Metropolis-coupled Markov chain Monte Carlo (MCMC) algorithm, give probability distributions for model parameters, and allow incorporation of prior assumptions. Altekar et al. [2] discuss shared-memory and distributed-memory parallelization of the sampling scheme used in MrBayes. ExaBayes [1] also uses distributed Metropolis-coupled chains, and further proposes chain swaps using nonblocking communication messages. This nonblocking communication optimization is shown to reduce running time by up to 19% [1]. ExaBayes also includes a memory-saving technique by recomputing partial results on-demand. ExaML uses a similar recomputation optimization to reduce inter-node communication. When likelihood calculations are parallelized based on partitions, the P_i matrix calculations are redundantly performed by every process. Kobert et al. [27] formulate a bi-criterion data distribution problem to determine the optimal distribution of partitions and sites to processes, and show that their new implementation is up to 3× faster than the implementation with the prior data distribution scheme. Other notable multi-node parallelizations include the master–worker strategy to parallelize the IQPNNI approach [31] and the Java-based DPRml [26] method. I/O optimizations and checkpointing are other important considerations in parallel environments. ExaML and Beast 2 include support for periodic disk-based checkpointing. ExaML converts the text-based input file to a binary format to permit parallel I/O.

In addition to parallelism, algorithmic changes also contribute to significant speedups. For instance, FastTree [40] employs several novel optimizations and is shown to be two orders of magnitude faster than RAxML version 7. A recent evaluation by Zhou et al. [53] shows that FastTree continues to be faster than recent versions of RAxML/ExaML, PhyML, and IQ-TREE, while also producing trees that are more dissimilar to trees generated using the other tool.

3.4 Conclusions

We have witnessed dramatic advances since early work on parallel phylogenetic inference [16, 45, 49]. Software development for computational phylogenetics is thriving [36], and performance optimization continues to be a focal area. It is now possible to achieve significant performance improvements for phylogenetic likelihood function calculations by leveraging modern libraries such as BEAGLE and PLL. Moret et al. [35] review methods for phylogenetic inference from rearrangement data, and describe an ML-based method that is competitive with approaches for sequence data. For the problem of supertree estimation, Nguyen et al. [38] show that Matrix Representation with Likelihood (MRL), an ML-based approach, is fast and outperforms leading alternative supertree methods (see chapter by Warnow in this book for more about MRL and supertree methods). Parallel algorithms and optimizations to improve scaling of these recent ML-based methods could be a promising future research direction.

Acknowledgements This work is supported in part by the National Science Foundation awards #1339745, #1439057, and #1535058.

References

1. Aberer, A.J., Kobert, K., Stamatakis, A.: ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. Mol. Biol. Evol. **31**(10), 2553–2556 (2014). https://doi.org/10.1093/molbev/msu236
2. Altekar, G., Dwarkadas, S., Huelsenbeck, J.P., Ronquist, F.: Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics **20**(3), 407–415 (2004). https://doi.org/10.1093/bioinformatics/btg427
3. Ayres, D.L., Cummings, M.P.: Rerooting trees increases opportunities for concurrent computation and results in markedly improved performance for phylogenetic inference. In: Proceedings of the 2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), pp. 247–256 (2018). https://doi.org/10.1109/IPDPSW.2018.00049

4. Ayres, D.L., Darling, A., Zwickl, D.J., Beerli, P., Holder, M.T., Lewis, P.O., Huelsenbeck, J.P., Ronquist, F., Swofford, D.L., Cummings, M.P., Rambaut, A., Suchard, M.A.: BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Syst. Biol. **61**(1), 170–173 (2012). https://doi.org/10.1093/sysbio/syr100
5. Bader, D.A., Moret, B.M.E.: GRAPPA runs in record time. HPC Wire **9**, 47 (2000)
6. Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J.: BEAST 2: a software platform for Bayesian evolutionary analysis. PLOS Comput. Biol. **10**(4), 1–6 (2014). https://doi.org/10.1371/journal.pcbi.1003537
7. Box, G.E.P., Tiao, G.C.: Bayesian Inference in Statistical Analysis, vol. 40. Wiley (2011)
8. Chor, B., Tuller, T.: Maximum likelihood of evolutionary trees: hardness and approximation. Bioinformatics **21**(suppl1), i97–i106 (2005). https://doi.org/10.1093/bioinformatics/bti1027
9. CIPRES Cyberinfrastructure for Phylogenetic Research. http://www.phylo.org/. Accessed Oct 2018
10. Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.M., Gascuel, O.: Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. **36**(suppl2), W465–W469 (2008). https://doi.org/10.1093/nar/gkn180
11. Drummond, A.J., Rambaut, A.: BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. **7**(1), 214 (2007). https://doi.org/10.1186/1471-2148-7-214
12. Dutheil, J., Gaillard, S., Bazin, E., Glémin, S., Ranwez, V., Galtier, N., Belkhir, K.: Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. BMC Bioinform. **7**(1), 188 (2006). https://doi.org/10.1186/1471-2105-7-188
13. Felsenstein, J.: PHYLIP version 3.697. http://evolution.genetics.washington.edu/phylip.html. Accessed Oct 2018
14. Felsenstein, J.: Phylogeny programs. http://evolution.genetics.washington.edu/phylip/software.html. Accessed Oct 2018
15. Felsenstein, J.: Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**(6), 368–376 (1981). https://doi.org/10.1007/BF01734359
16. Feng, X., Buell, D.A., Rose, J.R., Waddell, P.J.: Parallel algorithms for Bayesian phylogenetic inference. J. Parallel Distrib. Comput. **63**(7), 707–718 (2003). https://doi.org/10.1016/S0743-7315(03)00079-0
17. Fitch, W.M.: On the problem of discovering the most parsimonious tree. Am. Nat. **111**(978), 223–257 (1977). https://doi.org/10.1086/283157
18. Fitch, W.M., Margoliash, E.: Construction of phylogenetic trees. Science **155**(3760), 279–284 (1967)
19. Flouri, T., Izquierdo-Carrasco, F., Darriba, D., Aberer, A., Nguyen, L.T., Minh, B., Von Haeseler, A., Stamatakis, A.: The phylogenetic likelihood library. Syst. Biol. **64**(2), 356–362 (2015). https://doi.org/10.1093/sysbio/syu084
20. Foulds, L.R., Graham, R.L.: The Steiner problem in phylogeny is NP-complete. Adv. Appl. Math. **3**(1), 43–49 (1982)
21. GRAPPA genome rearrangements analysis under parsimony and other phylogenetic algorithms. https://www.cs.unm.edu/~moret/GRAPPA/. Accessed Oct 2018
22. Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O.: New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. **59**(3), 307–321 (2010). https://doi.org/10.1093/sysbio/syq010
23. Guindon, S., Gascuel, O.: Recent computational advances in maximum-likelihood phylogenetic inference. In: Warnow, T. (ed.) Bioinformatics and Phylogenetics—Seminal Contributions of Bernard Moret. Springer International Publishing AG (2018)
24. Holder, M., Lewis, P.O.: Phylogeny estimation: traditional and Bayesian approaches. Nat. Rev. Genet. **4**(4), 275–284 (2003)
25. Huelsenbeck, J.P., Ronquist, F., Nielsen, R., Bollback, J.P.: Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294**(5550), 2310–2314 (2001). https://doi.org/10.1126/science.1065889

26. Keane, T.M., Naughton, T.J., Travers, S.A.A., McInerney, J.O., McCormack, G.P.: DPRml: distributed phylogeny reconstruction by maximum likelihood. Bioinformatics **21**(7), 969–974 (2005). https://doi.org/10.1093/bioinformatics/bti100
27. Kobert, K., Flouri, T., Aberer, A., Stamatakis, A.: The divisible load balance problem and its application to phylogenetic inference. In: Brown, D., Morgenstern, B. (eds.) Algorithms in Bioinformatics, pp. 204–216. Springer, Berlin Heidelberg (2014)
28. Kozlov, A.: amkozlov/raxml-ng: RAxML-NG v0.6.0 BETA (2018). https://doi.org/10.5281/zenodo.1291478
29. Kozlov, A.M., Aberer, A.J., Stamatakis, A.: ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics **31**(15), 2577–2579 (2015). https://doi.org/10.1093/bioinformatics/btv184
30. Miller, M.A., Schwartz, T., Pfeiffer, W.: User behavior and usage patterns for a highly accessed science gateway. In: Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale, pp. 46:1–46:8. ACM (2016). https://doi.org/10.1145/2949550
31. Minh, B.Q., Vinh, L.S., von Haeseler, A., Schmidt, H.A.: pIQPNNI: parallel reconstruction of large maximum likelihood phylogenies. Bioinformatics **21**(19), 3794–3796 (2005). https://doi.org/10.1093/bioinformatics/bti594
32. Moret, B.M., Tang, J., Wang, L.S., Warnow, T.: Steps toward accurate reconstructions of phylogenies from gene-order data. J. Comput. Syst. Sci. **65**(3), 508–525 (2002). https://doi.org/10.1016/S0022-0000(02)00007-7
33. Moret, B.M., Wang, L.S., Warnow, T., Wyman, S.K.: New approaches for reconstructing phylogenies from gene order data. Bioinformatics **17**(suppl1), S165–S173 (2001). https://doi.org/10.1093/bioinformatics/17.suppl_1.S165
34. Moret, B.M.E., Bader, D.A., Warnow, T.: High-performance algorithm engineering for computational phylogenetics. J. Supercomput. **22**(1), 99–111 (2002). https://doi.org/10.1023/A:1014362705613
35. Moret, B.M.E., Lin, Y., Tang, J.: Rearrangements in phylogenetic inference: compare, model, or encode? In: Chauve, C., El-Mabrouk, N., Tannier, E. (eds.) Models and Algorithms for Genome Evolution, pp. 147–171. Springer, London (2013). https://doi.org/10.1007/978-1-4471-5298-9_7
36. Nekrutenko, A., Galaxy Team, Goecks, J., Taylor, J., Blankenberg, D.: Biology needs evolutionary software tools: let's build them right. Mol. Biol. Evol. **35**(6), 1372–1375 (2018). https://doi.org/10.1093/molbev/msy084
37. Nguyen, L.T., Schmidt, H.A., von Haeseler, A., Minh, B.Q.: IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. **32**(1), 268–274 (2015). https://doi.org/10.1093/molbev/msu300
38. Nguyen, N., Mirarab, S., Warnow, T.: MRL and SuperFine+MRL: new supertree methods. Algorithms Mol. Biol. **7**(1), 3 (2012). https://doi.org/10.1186/1748-7188-7-3
39. OMICtools: phylogenetic inference software tools. https://omictools.com/phylogenetic-inference-category?tab=software&page=1. Accessed Oct 2018
40. Price, M.N., Dehal, P.S., Arkin, A.P.: FastTree 2 approximately maximum-likelihood trees for large alignments. PLOS ONE **5**(3), 1–10 (2010). https://doi.org/10.1371/journal.pone.0009490
41. Roch, S.: A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. IEEE/ACM Trans. Comput. Biol. Bioinform. **3**(1), 92 (2006). https://doi.org/10.1109/TCBB.2006.4
42. Ronquist, F., Huelsenbeck, J.P.: MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**(12), 1572–1574 (2003). https://doi.org/10.1093/bioinformatics/btg180
43. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**(4), 406–425 (1987)
44. Sankoff, D., Blanchette, M.: The median problem for breakpoints in comparative genomics. In: Jiang, T., Lee, D.T. (eds.) Computing and Combinatorics, pp. 251–263. Springer, Berlin, Heidelberg (1997)
45. Snell, Q., Whiting, M., Clement, M., McLaughlin, D.: Parallel phylogenetic inference. In: Proceedings of the 2000 ACM/IEEE Conference on Supercomputing. IEEE Computer Society (2000)

46. Sokal, R.R., Michener, C.D.: A statistical method for evaluating systematic relationship. Univ. Kansas Sci. Bull. **28**, 1409–1438 (1958)
47. Stamatakis, A.: RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**(9), 1312–1313 (2014). https://doi.org/10.1093/bioinformatics/btu033
48. Stamatakis, A.: A review of approaches for optimizing phylogenetic likelihood calculations. In: Warnow, T. (ed.) Bioinformatics and Phylogenetics—Seminal Contributions of Bernard Moret. Springer International Publishing AG (2018)
49. Stewart, C.A., Hart, D., Berry, D.K., Olsen, G.J., Wernert, E.A., Fischer, W.: Parallel implementation and performance of fastDNAml: a program for maximum likelihood phylogenetic inference. In: Proceedings of the 2001 ACM/IEEE Conference on Supercomputing. ACM (2001). https://doi.org/10.1145/582034.582054
50. Suchard, M.A., Rambaut, A.: Many-core algorithms for statistical phylogenetics. Bioinformatics **25**(11), 1370–1376 (2009). https://doi.org/10.1093/bioinformatics/btp244
51. Tavaré, S.: Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. **17**(2), 57–86 (1986)
52. Yang, Z.: Computational Molecular Evolution. Oxford University Press (2006)
53. Zhou, X., Shen, X.X., Hittinger, C.T., Rokas, A.: Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. Mol. Biol. Evol. **35**(2), 486–503 (2018). https://doi.org/10.1093/molbev/msx302
54. Zwickl, D.J.: Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. thesis, The University of Texas at Austin (2006)