Evolutionary Bioinformatics

SHORT COMMENTARY

Snapshots of Tree Space

Zheng Wang[1], Francesc López-Giráldez[1] and Jeffrey P. Townsend[1,2]

[1]Department of Ecology and Evolutionary Biology and [2]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT06520, USA. Email: wang.zheng@yale.edu or jeffrey.townsend@yale.edu

Abstract: Achievement of the best balance between the accuracy and efficiency is always an important issue when searching a tree space of large data sets. In the 5th issue in 2009, Rodrigo et al used bootstrapped topologies as fixed genealogies to distribute an MCMC analysis across a cluster of computers, resulting in an efficiency yielding results 37 times faster than in the standard MCMC methods. Tree searches can seldom be guided with certainty, so that such snapshots sampling partial but "more-likely" tree space facilitated by parallel programs on computer clusters may provide great promise among a few choices that are computationally affordable when tree space is large and complicated.

Keywords: large phylogeny, tree sampling, computer cluster

Evolutionary Bioinformatics 2009:5 133–136

This article is available from http://www.la-press.com.

The technical task underlying the reconstruction of phylogeny and evolutionary history centers on the location and evaluation of the globally optimal or best-scoring tree(s) according to an optimality criterion. This search occurs within a tree space composed of every possible relationship among sampled genes or organisms. In theory, one could simply perform an exhaustive evaluation of every tree in the corresponding tree space. However, the number of trees increases tremendously as the number of sampled taxa increases, rapidly making exhaustive enumeration of a tree space computationally expensive. For more than trivial levels of taxon sampling, exhaustive search is a mission impossible for even the most powerful computers.[1] Thus, the development of methods for analysis of large data sets is a challenge within computer science, but it is much more than that: it is vital for the future progress of taxonomy, for predicting gene function and functional residues, and for classifying environmental DNA sequences.

Methods arising from diverse philosophies (e.g. parsimony, likelihood, and Bayesian) have featured strengths and suffered from weaknesses in stochastic algorithmic searches for the best-scoring trees in tree space. Even for small data sets, use of traditional and Bayesian approaches to inferring phylogeny have distinct algorithmic consequences.[2] For inference of phylogeny from large data sets, one of the core algorithmic challenges faced has been balancing the need to arrive at the tree that best reflects the evolutionary history with the speed to find the best score or to estimate robustness of a phylogeny in the shortest time. Methods such as the Parsimony Ratchet,[3] FastTree,[4] MrBayes,[5] PHYML,[6] fastDNAml,[7] and RaxML[8,9] that are still themselves evolving, have become popular to infer phylogeny from data sets incorporating extensive taxon sampling, and performance of some of these methods has been evaluated on both synthetic and real data alignments.[7,8] As the gathering of data accelerates, the memory and CPU requirements of previous approaches are becoming prohibitive; current Maximum Likelihood (ML)-based tree reconstruction programs can make reasonable tree topology inferences from data sets encompassing up to thousands of sequences of a few genes.[9] With the current rapid improvement in sequencing throughput, larger and larger data sets with thousands of distinct sequences are imminent (Table 1). For example, in the iPlant collaborative (http://iptol.iplantcollaborative.org), a phylogeny of approximately 500,000 organisms is targeted. This size of analysis will require at the very least an order of magnitude improvement in the throughput of current computational methods. How might we possibly achieve these goals?

The most promising methods will take advantage of both algorithmic finesse and technical advances in parallel computing. Some programs for phylogenetic inference support parallel computing.[5,7,8] Among them, Bayesian Markov chain Monte Carlo (MCMC) sampling as in MrBayes has become increasing popular as a rapid method that both estimates an optimal topology and supplies measures of confidence in individual nodes in comparatively short computation time. While Bayesian posterior probabilities have come to be interpreted as upper bounds of nodal support, bootstrap proportions are commonly interpreted as lower bounds of node reliability. One recent hybrid algorithm to evaluate the robustness of large phylogenies supplied bootstrapped data matrices to MrBayes, yielding a "Bayesian bootstrap proportion" support value for each branch.[10] Although no research yet has established the mapping between the distribution of the most likely topologies based on an empirical dataset and the distribution of optimal topologies from bootstrapped character matrices, several studies have revealed strong correlations between posterior probabilities,

Table 1. The largest taxon-sampling phylogeny deposited at TreeBase each year from 1999 to 2009 (data kindly provided by Dr. William Piel at www.treebase.org). Size of data alignments (number of characters) was provided correspondingly.

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number of sampled taxa	131	160	318	154	233	329	330	295	400	574	375
Alignment length (nchar)	897	3453	1434	317	357	1711	142	829	781	844	4714

Bayesian bootstrap proportion, and ML bootstrap support.[11,12]

On page 97–105 in the 5th issue in 2009, Rodrigo et al[13] introduced an alternative way to combine bootstrapping and Bayesian MCMC—an approach using bootstrapped topologies as fixed genealogies to distribute the MCMC analysis with coalescent models across a cluster of computers. Instead of performing a heuristic MCMC search of the tree space, a set of Maximum Likelihood bootstrap trees were generated and parallelized onto a number of CPUs as fixed topologes for MCMC sampling of posterior probability based parameters against the data set. They described a computational efficiency yielding results 37 times faster than in the standard MCMC methods for achieving stationarity, facilitating estimation of the target posterior probability for large sample sizes. In the paper this idea was credited to Felsenstein, who actually argued that the bootstrap histories are not guaranteed to be the "more likely" histories in any technical sense, and pointed out zero-length branch issues raised by the bootstrapping approach.[14] It would be interesting to see how well the approach functions on data sets simulated with slow, fast, or mixed evolutionary rates, particularly assessing the extent of the diversity of bootstrap trees generated. In this context, tree space is being explored by evaluating the bootstrap trees, but in principle, trees should be accepted only by the evaluation of optimality criteria on data. In practice, data consisting of characters that lead to unresolved phylogenies with low-confidence or "not likely" bootstrap histories usually are of little interest for phylogenetics. However, bootstrap support has served as the most common estimator of node confidence in published phylogenies, and bootstrap trees have been successfully specified as starting trees for maximum likelihood analyses on large datasets. For example, a recent plant phylogeny analysis with more than 4000 species in taxon sampling and over 100 megabases in sequence data was aided by this approach.[15] One hopes that it is safe to say that bootstrap trees constitute an assemblage of "more likely" histories—at least for informative data. Informative data features signal for the historical epoch under study without also imposing a burden of noise. Many phylogenetic studies select their data based on personal or second-hand experiences and referenced studies. More recently, direct measures of data quality or potential phylogenetic informativeness are beginning to become available.[16] However, good data will still not guarantee that bootstrap sampled trees locate preferentially to peaks of likelihoods in the tree space—Bootstrap trees can be seen as hurried—yet probably still not quick enough—snapshots of the tree space, from a platform floating up and down as columns of the data set are being resampled and replaced.

Aside from selecting informative data and having access to a computational cluster, using the approach of Rodrigo et al on real data sets will require further decisions regarding some issues of technique, such as the heuristic searching criteria to be used for bootstrapping, the number of bootstrap trees to be generated, and how to deal with short or zero branches in bootstrap trees. An alternative would be to separate the tree-generation and parameter-MCMC and analyze in parallel the posterior trees generated with Bayesian approaches, but such an approach is unlikely to be realistic for large data sets, considering the uncertainty in the amount of time to reach the stationarity.

As the authors pointed out, there is always a tradeoff between efficiency and accuracy when researchers deal with large data sets. Their approaches took good advantage of cluster computation, greatly improving computational efficiency without sacrificing much accuracy of estimation. A similar approach might also increase the efficiency of analyses of multilocus data by parallelization of STEM, BEST and other recently developed gene trees to species trees methods.[17–19] Each gene set could be analyzed in parallel, then compiled into a set of fixed topologies to be analyzed as concatenated data on computer clusters for species trees. In photography, a snapshot camera is typically programmed to achieve a deep depth of field and high shutter speed so that as much of the image is in focus as possible, providing well-resolved images that nevertheless depict only part of the story underlying large and chaotic scenes. As the phylogeny can seldom be guided with certainty and Bayesian methods can in principle integrate over a suite of genealogies to recover a posterior probability distribution, snapshots of tree space followed by MCMC parallel analyses may provide great promise among a few choices that are computationally affordable when tree space is large and complicated.

Disclosures

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors report no conflicts of interest.

References

1. Felsenstein J. The number of evolutionary trees. *Systematic Zoology*. 1978; 27:27–33.
2. Holder MT, Lewis PO. Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics*. 2003;4:275–84.
3. Nixon KC. The parsimony Rachet, a new method for rapid parsimony analysis. *Cladistics*. 1999;15:407–14.
4. Price MN, Dehal PS, et al. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*. 2009;26(7):1641–50.
5. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*. 2001;17(8):754–5.
6. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*. 2003;52: 696–704.
7. Stewart C, Hart D, et al. Parallel implementation and performance of fastDNAml—a program for maximum likelihood phylogenetic inference. *Proceedings of 14th IEEE/ACM Supercomputing conference (SC2001)*, Denver, CO. May 18, 2001.
8. Stamatakis A, Ludwig T, et al. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*. 2005;21(4):456–63.
9. Stamatakis A. RAxML-VI-HPC: Maximum Likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 2006;22(21):2688–90.
10. Lutzoni F, Kauff F, et al. Assembling the fungal tree of life: progress, classification, and evolution of subcellular traits. *American Journal of Botany*. 2004;91(10):1446–80.
11. Alfaro ME, Zoller S, et al. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in asscessing phylogenetic confidence. *Molecular Biology and Evolution*. 2003;20(2):255–66.
12. Douady CJ, Delsuc F, et al. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution*. 2003;20(2):248–54.
13. Rodrigo AG, Tsai P, et al. One the use of bootstrapped topologies in coalescent-based Bayesian MCMC inference: a comparison of estimation and computational efficiencies. *Evolutionary Bioinformatics*. 2009;5:97–105.
14. Felsenstein J. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetic Research*. 1992;59(2):139–47.
15. Smith SA, Donoghue MJ. Rates of molecular evolution are linked to life history in flowering plants. *Science*. 2008;322(5898):86–9.
16. Townsend JP. Profiling phylogenetic informativeness. *Systematic Biology*. 2007;56(2):222–31.
17. Liu L, Pearl DK. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology*. 2007;56(3):504–14.
18. Kubatko LS, Carstens BC, et al. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*. 2009; 25(7):971–3.
19. Gegnan JH, DeGiorgio M, et al. Properties of consensus methods for inferring species trees from gene trees. *Systematic Biology*. 2009. DOI:10.1093/sysbio/syp008.