Rapidly updating tree topologies using ancestral state reconstruction

Emily Jane McTavish, University of Kansas, Lawrence KS, USA
∗ E-mail: mctavish@ku.edu

Introduction

Phylogenetic trees describe the evolutionary relationships among groups of organisms. Phylogenetics is not only important for taxonomy and species classification. Understanding these relationships is essential making inferences about biological processes. Phylogenies are the basis for examining trait evolution though time (e.g. [1]), have been applied to tracking disease transmission and spread (e.g. [2]), and are used to inform conservation decisions (e.g. [3]), and understand processes of community assembly [4]. Phylogenetics is used to answer questions about human origins [5]. Researchers on the Open Tree of life project are curating and combining hundreds of phylogenies to develop an accessible and reusable tree of life for all species [6]. All of these downstream analyses rely on the availability of accurate phylogenetic estimates for the species of interest [7].

Phylogenetic reconstruction

Tree-like reconstructions of species relationships have been applied to species taxonomy since Lamarck in 1809 [8], but over the past few decades the inference of phylogenetic relationships from DNA sequence data has expanded rapidly. Whereas 100 years ago inferring evolutionary relationships required coding of morphological characters across many individual specimens, today millions of DNA base pairs can be sequenced a few hours. Next generation sequencing technologies are continuing to rapidly increase the availability of sequence data for thousands of species that have not previously been included in phylogenies. Phylogenies with 13,000 or more tips have been published [9], [10]. As the data available for phylogenetic inference has increased, so has the complexity of the models used to infer these phylogenies. Maximum likelihood and Bayesian estimators can take advantage of models that include biological knowledge about sequence evolution, such as differences in rates of nucleotide substitutions (e.g. [11]) or codon based models of evolution [12]. These methodological advances have greatly increased the accuracy with which researchers are able to infer evolutionary relationships, even in challenging cases [13]. However, the complexity of appropriate models for biological sequence data means that these analyses can be computationally expensive. Even with recent advances in processing power and the availability of high performance computing resources, the computational time for inferring these phylogenies under maximum likelihood can be significant, even using the most efficient available approaches [14]. Thus, for very large trees researchers often use less optimal approaches such as distance metrics or approximately-maximum-likelihood approaches such as FastTree [15, 16].

Current approaches for updating phylogenies

Maximum likelihood and Bayesian approaches to phylogeny use heuristics to optimize the sequence evolution parameters and topology for a sequence alignment of sampled taxa. If a new taxon is sampled, it cannot be included in the tree without aligning it and performing a full reanalysis. Re-analysis of phylogenies in the face of new taxa is important, as additional data can add information that can affect relationships even in other clades as well as confidence in existing relationships. Therefore researchers generally wait until several new taxa have been sequenced, and then perform a full reanalysis, and publish an updated phylogeny. This process requires not only significant compute time but also non-trivial researcher person hours, as phylogenetic pipelines are largely not automated. Several pipelines to automate phylogenetic analyses with respect to data have been developed (e.g mor [17]; WASABI [18]; HAL [19]). While these pipelines streamline determining orthology and aligning sequences, they do not leverage the information gained by previous phylogenetic analyses. This creates a significant time lag between the generation of sequence data for a taxon and the ability to make evolutionary inferences using that taxon.

An alternative to re-analysis is using placement techniques. These approaches add taxa onto trees by comparison to taxa already included in the tree. Pplacer [20] is a maximum likelihood and Bayesian approach to placing samples on a phylogenetic tree. EPA is an evolutionary placement algorithm developed in the Stamatakis lab which rapidly places short reads on a phylogeny [21] Placement methods are common for taxonomic assignment of bacterial samples or reads from meta-genomic studies (e.g. Phylosift [22]). These placement allow rapid classification of new samples based on an existing backbone phylogeny, even for reads with little phylogenetic signal. As these methods do not require re-inference of existing branches, it can be very fast. However, these placement based algorithms do not update existing branches on the tree with respect to information contained in the added taxa. While these placement algorithms are useful for classification of new sequences given an existing tree, they do not affect the underlying phylogeny.

New approaches

Some novel methods are currently being developed to combine the advantages of these two approaches. Matsen and colleagues `http://matsen.github.io/talks/online_evolution2013.html#/` have been building on Bouchard-Cote and colleagues partially ordered set sequential monte carlo [23] to re-use portions of previous Bayesian MC runs to speed up the addition of taxa in new analyses.

Dr. Alexandros Stamatakis of the Heidelberg Institute for Theoretical Studies (HITS), and his lab, with whom I propose to collaborate, have been developing novel approaches to address this problem. They have developed several very efficient software packages for phylogenetics including RAxML [24], [25] [14] for maximum likelihood estimation, EAxML (Exascale Maximum LIkelihood) for very large datasets [26] , and are working on ExaBayes, for rapid Bayesian phylogenetics on very large data sets. They recently published a perpetually-updating-tree pipeline [27]. This pipeline combines PHLAWD [9] and EaxML [26] to automate the updating of alignments with new sequences as they are submitted to public databases, such as NCBI GenBank. Once several new taxa have been added re-estimate the phylogeny using the

previous topology as a starting tree. This method efficiently automate the process researchers have previously applied to their taxon of interest. In addition, by using the previous phylogeny as a starting tree for the new search, the search is faster than inferring phylogenies from scratch, and finds trees as good as those inferred from scratch. They are currently maintaining a perpetually updated phylogeny for the green plants using the commonly sequenced chloroplast gene rbcL.

Dr. Mark Holder, my current postdoctoral advisor at the University of Kansas, has applied for a Humboldt Research Fellowship for Experienced Researchers to collaborate with Dr. Stamatakis on further developing this perpetual-updating-tree pipeline, in particular with respect to the aspect of adding new sequences to trees following alignment. While the current perpetual tree updating pipeline speeds up phylogenetic inference by using as a starting tree the previous ML tree, and adding the new sequences by random stepwise addition, better starting tree estimates would further speed up this process. By collaborating with Dr. Stamatakis and Dr. Holder at HITS I will build upon these approaches by utilizing ancestral sequence estimation to better place new taxa on phylogenetic trees and update the relationships in those trees. We have some funding via the NSF-funded Open Tree of Life award (Holder is one of the principal investigators on that award); that would include some funds for me to visit Germany briefly during Dr. Holder's sabbatical. This proposal would allow me to stay in Germany for a longer period of time and collaborate more extensively with the Stamatakis lab.

Aims

Our aims are to improve on currently existing methods for updating phylogenies. Specifically, to develop a method to add new taxa to trees and rapidly evaluate the likelihood of the new tree that is faster than full re-analysis and more accurate than placement based methods.

Methods

We have begun developing a pipeline to rapidly and accurately update phylogenetic trees with new taxa. In brief, the pipeline consists of 1) generating a maximum likelihood tree for the starting set of sequences, 2) generating a set of stochastic mutational histories across the tree, 3) searching the ancestral states generated by those mutational histories for similarities to our test taxon, 4) Adding the test taxon to that node in the tree, and 5) re-evaluating relationships within the region of the tree impacted by the tip addition. This process is summarized in figure 1, and described below.

Maximum Likelihood Phylogenetics

First we perform a full maximum likelihood estimation of the phylogeny for the alignment of the starting set of taxa, using RAxML [24], (Figure 1.1). While in the preliminary pipeline we move forward with only the maximum likelihood tree found in our search, we will test whether using a distribution of trees improves results. RAxML, developed by the Stamatakis lab the most computationally efficient approach

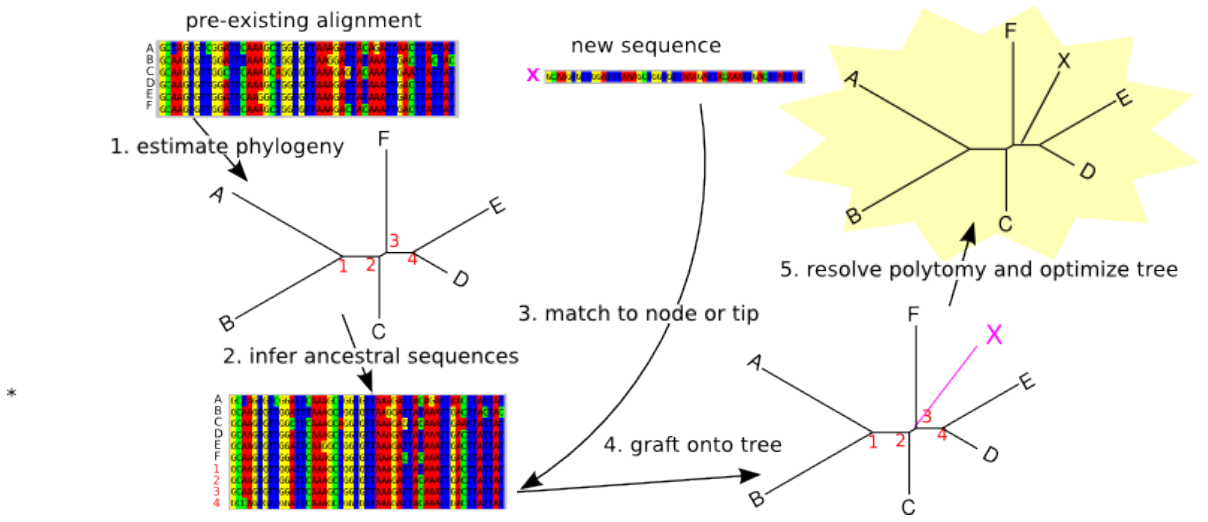

Figure 1. Schematic of tree updating plan.
1) Estimate maximum likelihood phylogeny for starting set of taxa (A,B,C,D,E). 2) Create distributions of potential ancestral sequences at each node (numbered 1-4 here). Only one potential reconstruction pictured. 3) Compare new taxon (X) to ancestral sequences. 4) Graft new taxon onto existing phylogeny, and add to alignment. 5) Update phylogeny under ML. Working with the Stamatakis lab at HITS we will be able to optimize all of these steps and explore alternative options for tree updating.

to estimating maximum likelihood phylogenies available today. It has been optimized for use on high performance computing clusters [14] as well as for very large phylogenies [25].

Stochastic mapping of ancestral states

From our starting tree we then generate a distribution of putative ancestral states for all nodes in the phylogeny, across all branches, using stochastic simulation of mutational histories. Stochastic simulation of mutations on phylogenies captures possible mutational histories along phylogenies [28] [29] and has been implemented both in a standalone GUI [30] and as an R package [31]. Felsenstein's pruning algorithm [32], upon which ML phylogenetic methods are based, calculates the probability of each base at each site at each node in the tree. The Stamatakis lab's Phylogenetic Likelihood Library (PLL) have a function to calculate these marginal ancestral sequences `http://www.libpll.org/`. By using stochastic mapping we can derive a joint distribution of character states at each site across the tree, and combine those to form full joint distributions of putative alignments. Although generating these distributions is a computationally slow step, these alignments of putative ancestral states can be updated as new tips are added and do not need to be generated anew. Alternatively, Pupko [33] [34] has an algorithm for efficiently sampling the ML joint reconstruction at nodes. We can modify that algorithm to just sample from the posterior at nodes rather than along branches. We will investigate the relative accuracy and efficiency of these different approaches.

Local alignments

Following inference of ancestral states at nodes, we then take a new taxon sequenced for the same region, and perform BLAST [35] or Smith-Waterman [36] searchers across each node within each stochastic iteration of ancestral state mappings. BLAST is an rapid search algorithm for DNA sequences that allows us to rapidly provisionally place new sequences at nodes in the tree. We will add the novel sequence to the multiple sequence alignment used to build the tree. We will compare alternative alignment techniques to determine the best approach to adding sequences, including profile based alignment [37] [9] and tracking insertion deletion histories using automata theory [38]. PaPaRa [39] is a fast phylogeny informed method for accurately adding taxa to an alignment developed in the Stamatakis lab.

Re-optimization

We then calculate the likelihood score for a new tree with this additional branch grafted on at the at location or locations suggested by the BLAST step. By using this new starting tree informed by ancestral state distributions, and building upon previous ML estimates of parameter values, we will be able to rapidly generate trees as good or better than those estimated from scratch.

Testing

We will thoroughly investigate several potential issues, including the impact of the addition order of taxa as we rebuild the tree, and whether global tree rearrangements are required on top of the process we propose. We will compare the speed and accuracy of out taxon addition procedure to alternative taxon addition procedures as well as to full re-analyses.

Software

When we have tested and determined an effective procedure for updating phylogenies with new sequences, we will develop an open source software package to be used by researchers in the field.

Advantages of collaboration with Stamatakis lab

By collaborating with the Stamatakis lab I will be able to improve my skills in high performance computing and low level program optimization. They have developed a parallelized software library, Phylogenetic Likelihood Library (PLL) which we hope to leverage to speed up our calculations. In addition, I can work with the postdoc Dr. Tomas Flouri to incorporate the stochastic mapping of ancestral states into PLL. By working directly with the Stamatakis lab I will be able to move forward without replicating work they have already accomplished. I will be able to build upon their work on ExaBayes [26] , EPA [21], and PaPaRa [39] .

Additional potential

If development of tools for phylogeny updating is successful within the time frame of the project, there are many potential directions for expansion. While many phylogenies are currently built using information from a single gene, incorporating information from multiple loci makes inferences more robust [40]. Rapid phlyogenetics at the single locus level can then be extended to genome scale analyses. By speeding up updating of gene-trees, we can incorporate multiple gene trees into species phylogenies. Ideally the software we develop for rapid phylogenetic updating can be incorporated into open online phylogeny resources, such as The Open Tree of Life project. This project, which Dr. Holder and myself are involved in, is building a synthetic phylogeny incorporating s many species as possible from across the tree of life. By integrating rapid phylogeny update tools with this online phylogeny repository we will be able to add taxa to this tree as data are processed, rather than the current lag time of moths or even years. Phylogenetic updating also has potential for rapid response to disease spread. Incorporating pathogen isolates into phylogenies can inform public health officials about the potential source [2]. In addition, as genomic sequencing is becoming common across wide taxonomic breadth, imputed internal sequences can assist in homology searching. Although PSI-BLAST [35] tries to exploit phylogenies, but using distributions of joint ancestral sequence samples could facilitate better detection of homology in distant comparisons.

Time schedule w/ milestones

Requested fellowship length: 9 months

- Months 1-2 Familiarize self with RAxML code and PLL library, with assistance of Stamatakis Lab. Develop C coding skills.
- Month 3 Develop brute-force computational approaches and test if these methods are accurate using simulated data.
- Month 4-6 Optimize and test code. Develop for distribution.
- Months 7-8 Write manual and manuscript.
- Month 9 Submission and revisions of manuscript.

Conclusions

Much computational development will be necessary in order for evolutionary analysis to keep pace with the staggering rate of data generation. Although maximum likelihood methods estimating phylogenetic relationships are useful and are widely applied, new approaches are needed to build on large trees. The Stamatakis lab has been at the forefront of development of efficient tools for phylogenetic analysis, and by working with them I will be able to build upon their expertise. We propose to develop a better approach

for adding new taxa onto existing phylogenies that combines the advantages of using complex models of sequences evolution and maximum likelihood analysis with the speed needed to keep up with the stream of new biological sequence data.

References

1. O'Meara BC, An C, Sanderson MJ, Wainwright PC (2006) Testing for different rates of continuous trait evolution using likelihood. Evolution 60: 922933.

2. Timme RE, Pettengill JB, Allard MW, Strain E, Barrangou R, et al. (2013) Phylogenetic diversity of the enteric pathogen salmonella enterica subsp. enterica inferred from genome-wide reference-free SNP characters. Genome Biology and Evolution 5: 2109–2123.

3. Isaac NJ, Turvey ST, Collen B, Waterman C, Baillie JE (2007) Mammals on the EDGE: conservation priorities based on threat and phylogeny. PLoS ONE 2: e296.

4. Emerson BC, Gillespie RG (2008) Phylogenetic analysis of community assembly and structure over space and time. Trends in Ecology & Evolution 23: 619–630.

5. Endicott P, Ho SY, Stringer C (2010) Using genetic evidence to evaluate four palaeoanthropological hypotheses for the timing of neanderthal and modern human origins. Journal of Human Evolution 59: 87–95.

6. Drew BT, Gazis R, Cabezas P, Swithers KS, Deng J, et al. (2013) Lost branches on the tree of life. PLoS Biology 11: e1001636.

7. Stoltzfus A, Lapp H, Matasci N, Deus H, Sidlauskas B, et al. (2013) Phylotastic! making tree-of-life knowledge accessible, reusable and convenient. BMC Bioinformatics 14: 158.

8. Lamarck JB (1873) Philosophie zoologique: ou, Exposition des considrations relatives l'histoire naturelle des animaux. F. Savy.

9. Smith SA, Beaulieu JM, Donoghue MJ (2009) Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. BMC Evolutionary Biology 9: 37.

10. Smith SA, Beaulieu JM, Stamatakis A, Donoghue MJ (2011) Understanding angiosperm diversification using small and large phylogenetic trees. American Journal of Botany 98: 404–414.

11. Jukes T, Cantor C (1969) {Evolution of protein molecules}. In: Munro M, editor, Mammalian protein metabolism, Academic Press, volume III. pp. 21–132.

12. Shapiro B, Rambaut A, Drummond AJ (2006) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Molecular Biology and Evolution 23: 7–9.

13. Kuhner MK, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Molecular Biology and Evolution 11: 459–468.

14. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690.

15. Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Molecular Biology and Evolution 26: 1641–1650.

16. Price MN, Dehal PS, Arkin AP (2010) FastTree 2 approximately maximum-likelihood trees for large alignments. PLoS ONE 5: e9490.

17. Hibbett D, Nilsson R, Snyder M, Fonseca M, Costanzo J, et al. (2005) Automated phylogenetic taxonomy: An example in the homobasidiomycetes (mushroom-forming fungi). Systematic Biology 54: 660–668.

18. Kauff F, Cox CJ, Lutzoni F (2007) WASABI: an automated sequence processing system for multi-gene phylogenies. Systematic Biology 56: 523–531.

19. Robbertse B, Yoder RJ, Boyd A, Reeves J, Spatafora JW (2011) Hal: an automated pipeline for phylogenetic analyses of genomic data. PLoS Currents 3.

20. Matsen F, Kodner R, Armbrust EV (2010) pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC bioinformatics 11: 538.

21. Berger SA, Krompass D, Stamatakis A (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. Systematic Biology 60: 291–302.

22. Darling AE, Jospin G, Lowe E, Matsen FA, Bik HM, et al. (2014) PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ 2: e243.

23. Bouchard-Ct A, Sankararaman S, Jordan MI (2012) Phylogenetic inference via sequential monte carlo. Systematic Biology 61: 579–593.

24. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics : btu033.

25. Stamatakis A, Aberer AJ, Goll C, Smith SA, Berger SA, et al. (2012) RAxML-Light: a tool for computing terabyte phylogenies. Bioinformatics 28: 2064–2066.

26. Stamatakis A, Aberer A (2013) Novel parallelization schemes for large-scale likelihood-based phylogenetic inference. In: 2013 IEEE 27th International Symposium on Parallel Distributed Processing (IPDPS). pp. 1195–1204. doi:10.1109/IPDPS.2013.70.

27. Izquierdo-Carrasco F, Cazes J, Smith SA, Stamatakis A (2014) PUmPER: phylogenies updated perpetually. Bioinformatics : btu053.

28. Nielsen R (2002) Mapping mutations on phylogenies. Systematic Biology 51: 729–739.

29. Huelsenbeck JP, Nielsen R, Bollback JP (2003) Stochastic mapping of morphological characters. Systematic Biology 52: 131–158.

30. Bollback JP (2006) SIMMAP: stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics 7: 88.

31. Revell LJ (2012) phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3: 217223.

32. Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17: 368–376.

33. Pupko T, Pe I, Shamir R, Graur D (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. Molecular Biology and Evolution 17: 890–896.

34. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, et al. (2012) FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic acids research 40: W580–584.

35. Altschul SF, Madden TL, Schffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25: 3389–3402.

36. Rognes T (2011) Faster smith-waterman database searches with inter-sequence SIMD parallelisation. BMC Bioinformatics 12: 221.

37. Loytynoja A, Vilella AJ, Goldman N (2012) Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. Bioinformatics 28: 1684–1691.

38. Westesson O, Lunter G, Paten B, Holmes I (2012) Accurate reconstruction of insertion-deletion histories by statistical phylogenetics. PLoS ONE 7: e34572.

39. Berger SA, Stamatakis A (2011) Aligning short reads to reference alignments and trees. Bioinformatics 27: 2068–2075.

40. Edwards S, Bensch S (2009) Looking forwards or looking backwards in avian phylogeography? a comment on zink and barrowclough 2008. Molecular Ecology 18: 29302933.