Learning, using, and extending variational distributions of phylogenetic trees

Frederick A Matsen IV,* **Michael Karcher**
Fred Hutchinson Cancer Research Center

Mathieu Fourment
University of Technology Sydney

Andrew Magee
University of Washington

Christiaan Swanepoel
University of Auckland

Cheng Zhang
Peking University

Abstract

Although Bayesian phylogenetic inference is a powerful tool for learning about evolution, its applicability is limited by computational expense. The current state-of-the-art consists of random-walk Markov chain Monte Carlo (MCMC) and counting posterior samples to estimate tree structure support. Random-walk MCMC converges slowly in the large and complex space of phylogenetic trees, and counting estimators are noisy and cannot extend posterior mass to unsampled trees. In this paper, we describe recent progress on using parameterized posterior approximations to phylogenetic trees in order to better estimate posterior support and infer Bayesian posterior distributions by direct variational inference.

1 Introduction

The inference of evolutionary history from sequence data is inherently uncertain. Correspondingly, researchers frequently use a Bayesian approach to infer evolutionary history in the form of a phylogenetic tree. Trees consist of a discrete tree structure, called a *topology*, and *branch lengths* that express the amount of evolution that has happened along that edge. The Bayesian approach to phylogenetic inference targets a posterior distribution on trees given an aligned collection of sequence data.

Bayesian phylogenetic inference is central to modern biology, with applications from inferring risk factors for recent HIV transmission [1] to reconstructing ancient nucleotide usage patterns [2]. Current Bayesian phylogenetic practice proceeds as follows:

- run a Markov Chain Monte Carlo sampler that uses random tree modification proposals
- estimate topological support as the frequency with which a given tree was in the sample
- if the data or model changes in some way, start over from scratch.

Each of these points have drawbacks. In typical settings with real data, random tree proposals are likely to either be timid or have low acceptance rates [3]. The simple frequency-based estimator, while unbiased, need not be optimal and cannot extend support to unsampled trees [4–6]. It is clearly disavantageous to have to throw away computationally expensive previous inferences when the data or model changes in some way.

In this paper, we will describe recent progress on representations of phylogenetic posterior distributions that allow for more flexible and efficient algorithms. This framework:

- enables more efficient variational inference and adaptive tree proposals

*ematsen@gmail.com; http://matsen.group/

14th Conference on Machine Learning in Computational Biology (MLCB 2019), Vancouver, Canada.

- gives more accurate topology support estimators by leveraging tree structure
- lays a solid framework for re-use of previous inferences when data or model changes.

By establishing these representations and tying them into modern machine-learning and probabilistic programming packages, we hope to move Bayesian phylogenetic inference into the modern era.

2 The challenge of Bayesian phylogenetic inference

The goal of molecular phylogenetic inference is to reconstruct the evolutionary history of a group of entities from some biological sequence data (i.e. DNA, RNA, or amino acid sequence). The inference is formalized as a phylogenetic tree with the observed sequences at the leaves of the tree. Statistical phylogenetic inference, which now dominates the field, treats phylogenetics as a statistical inference problem in which one picks among phylogenetic tree models by how well they explain the observed data [7]. Here the sequence mutation process given a tree is modeled as a continuous-time Markov chain (CTMC) that are allowed to run the duration of the tree branch lengths. It is very common to assume that the CTMC is defined per-nucleotide-position and that the sites evolve independently.

Bayesian phylogenetic inference targets the posterior distribution $p(\mathbf{z} \mid D)$ on structures $\mathbf{z}$ consisting of phylogenetic trees along with associated model parameters including branch lengths. To characterize the posterior distribution by Bayes' rule $p(\mathbf{z} \mid D) \propto p(D \mid \mathbf{z})\, p(\mathbf{z})$ we need the likelihood $p(D \mid \mathbf{z})$, which can be calculated quickly using dynamic programming [7], and the prior $p(\mathbf{z})$, which can be as simple as a uniform prior on topologies with exponentially-distributed branch lengths, or a more complex descriptor of demographic history [8].

Bayesian phylogenetic inference is computationally challenging. The corresponding maximum-likelihood inference problem is known to be NP-hard [9]. These posteriors are difficult to compute because of the super-exponential number of phylogenetic trees, the intertwined discrete and continuous nature of the space, and the frequent presence of multiple modes [3].

Random-walk Markov chain Monte Carlo (MCMC) is currently the only technique used in practice. In these algorithms, a tree modification is proposed by cutting off part of the tree and randomly attaching it to another location, which is then accepted or rejected according to the Metropolis-Hastings ratio. Because the high-density region of the posterior is focused on a very small subset of the volume of possible trees, these uninformed tree modifications have low acceptance rates. Once such an algorithm is run, one uses the sample as an approximation of the phylogenetic posterior distribution. Practically speaking, important phylogenetic inferences can take a month of compute [10], which limits the amount of data researchers are willing to input into these algorithms.

3 Parametrized distributions on phylogenetic tree topologies

We can alleviate some of the difficulties of Bayesian phylogenetic inference by developing structured approximations to phylogenetic posterior distributions. One cannot apply a continuous relaxation of a discrete embedding [11, 12] because tree space, using natural notions of similarities between trees, is inherently non-euclidean [13].

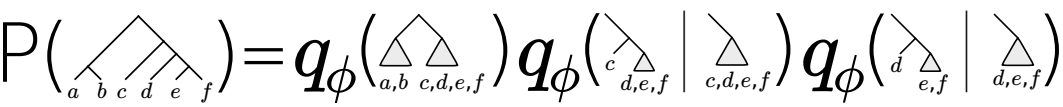

Fig. 1. The conditional clade decomposition (CCD) model for a probability distribution on a phylogenetic tree.

Rather, recent work has developed approximations based on tree decompositions [4–6, 14]. In this setup, the probability of a rooted tree is approximated as a product of conditional probability terms concerning the next set of leaves to branch off of given some information about the rest of the tree (Fig. 1). In the *conditional clade probability* case [4, 5, 14], one conditions on a group of leaves forming a *clade*, or subtree of unspecified structure. This is an approximation because real posterior distributions need not have conditional independence of tree substructures.

Several years later these approaches were generalized to enable more flexible sets of conditional dependences by formulating the procedure as inference on a graphical model [6]. In this *subsplit Bayesian network* (SBN) formulation, one realizes a phylogenetic tree as a series of *subsplits*, which are bipartitions of subsets of the leaves. If the Bayesian network has no connections, one recovers the previous CCD model (Fig. 1). If, on the other hand, the subsplit is conditioned on a "parent" subsplit that it refines, then one obtains something like CCD but where one also conditions on the sister clade (Fig. 2). However, one could easily add additional edges to the network to encode more complex dependences, such as on "aunt" clades [6].

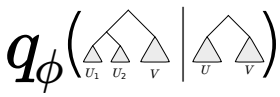

Fig. 2. Conditioning on a sister clade for the next split.

Training these distributions q_ϕ given a collection of trees does not pose a substantial challenge. In simple cases one only has to count up the number of subsplits and normalize in order to get the maximum-likelihood estimator [5]. To estimate conditional subsplit distributions on unrooted trees, marginalizing out the position of the root, one uses expectation-maximization [6] in a form that will be familiar to those accustomed to estimation of graphical models. One can also include conjugate Dirichlet priors to regularize estimation with no asymptotic additional cost.

Investigation of posteriors from real data shows that the additional flexibility offered by the SBN model (Fig. 2) is required to accurately capture the shape of posteriors [6]. Furthermore, these models give better posterior estimates than simply using the frequency of occurrence of the tree in a posterior sample, which is the only posterior probability estimator used in practice. In fact, on a standard panel of benchmark data sets [15], this currently-used simple estimator was best only for a very simple data set with 7 trees in the posterior. One can attribute this performance gain to both the methods smoothing out Monte Carlo error by using nearby trees and by extending the distribution to unsampled trees with nonzero posterior mass.

4 Variational Bayes phylogenetic inference

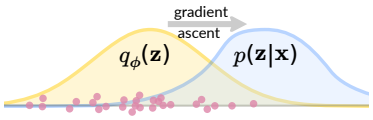

Fig. 3. Variational inference by stochastic gradient ascent.

Given how well the structured approximations work to capture complex phylogenetic posteriors, one might wish to forego random-walk MCMC and instead directly infer a posterior distribution using variational inference (VI). In variational inference, one posits a variational distribution $q_\phi(\mathbf{z})$ and fits it to the unknown posterior distribution using gradient ascent. Specifically, one samples from the current approximation q_ϕ (pink points in Fig. 3) and uses that sample to take an optimization step on the parameters ϕ to improve the fit as described in more detail below. Early in the fitting procedure, this will involve increasing the probability of generating samples $\mathbf{z}$ that had a high un-normalized posterior and decrease the probability of generating those that did not.

Returning to phylogenetics, the variational distribution is a distribution on phylogenetic trees. To achieve this full VI procedure, one must complement the variational distribution on tree topologies (Fig. 1, 2) with a variational distribution on branch lengths. Variational distributions on continuous parameters is a much more classical setting, and previous research has identified families of distributions useful for modeling phylogenetic branch lengths [16, 17].

However, this previous work was done under the assumption of a fixed phylogenetic tree topology, whereas a full phylogenetic VI algorithm requires the tree to be dynamically sampled. In principle one could have one branch length distribution for every branch of every tree, but this would certainly have too many parameters to infer accurately. On the other hand one could consider each edge as a bipartition of the leaf set and have only one parameter for each such bipartition. Such a parameterization runs the risk of being insufficiently flexible, as the distribution of branch lengths for a given edge may depend on the edges that surround it.

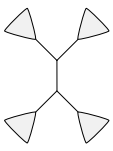
Fig. 4. A fat quartet.

The optimal solution appears to be to incorporate some local branching structure in the tree. The strategy of [18] is to use a parameterization in terms of what one might call *fat quartets* (Fig. 4): five compatible splits around a single edge. Any edge with the same fat quartet will have the same variational distribution. In fact, each fat quartet gets broken up into its *primary subsplit pair*, of the split given by the edge and the two subsplits on either side of that edge; each parameter of the variational parameterization for that fat quartet is the sum of parameters for these three structures individually.

Once the variational approximation is established, it remains to devise algorithms to fit it to the (unknown) posterior distribution. This is typically done by minimizing the Kullback-Liebler divergence $\mathrm{KL}(q_\phi(\mathbf{z}) \parallel p(\mathbf{z} \mid D))$ of the variational approximation q_ϕ to the posterior distribution. As in other applications of variational inference, one can avoid problems with the normalizing constant of the posterior by instead optimizing a related quantity called the ELBO [19], which only differs from the KL divergence by a constant independent of ϕ.

One can then use the stochastic gradient of the ELBO for gradient ascent. This requires some care: a naive approach will have too high of variance to obtain an accurate variational approximation. Thus [18] uses multi-sample gradient estimators [20, 21]. Putting all of these components together gives a rapidly-converging algorithm that has low divergence from the ground-truth posterior distribution, even in a preliminary implementation [18]. Furthermore, one can obtain accurate marginal likelihoods over trees with a small amount of additional computational effort, a task which classical requires many runs of MCMC [22, 23].

5 Conclusion and opportunities for future work

The current state-of-the-art for phylogenetic inference lags significantly behind the more sophisticated algorithms now available for posteriors on continuous parameters [19, 24]. In addition, the discrete aspect of phylogenetic posterior distributions keeps the phylogenetic community from capitalizing on the tremendous progress in machine learning frameworks [25, 26] and probabilistic programming [27–29]. Here we have described how continuously-parameterized variational distributions on phylogenetic trees can accurately fit phylogenetic posteriors, opening phylogenetic inference to a whole suite of new approaches and tools. Many opportunities remain.

There are opportunities to improve the core variational inference algorithm. First, although the SBN parameterization does greatly reduce the number of parameters that are required to specify a phylogenetic posterior distribution, there are still in principle an exponential number. For this reason, the existing implementation [18] uses rapid phylogenetic bootstrapping [30] to generate a collection of trees then uses this collection to define the set of parent-child subsplit pairs that are used for variational inference. Although this ad-hoc approach works for benchmark data sets, a more principled approach is desirable. Second, other variational parameterizations of branch length distributions should be developed and tested.

There are opportunities to improve implementation. The original implementation was in Python, and although this was enough to establish the viability of the approach, further work is needed to develop an efficient implementation. On the other hand, Python now boasts a very strong collection of machine learning and probabilistic programming packages [25–29]. For this reason we are developing a Python-interface C++ library available at `https://github.com/phylovi/libsbn` such that users can formulate interesting parametrized phylogenetic prior distributions in Python, yet have efficient SBN routines and likelihood calculation using the BEAGLE C++ library [31–33].

The general concept of fitting an approximation to phylogenetic posteriors opens up opportunities for creative ways of building and using them. While MCMC requires rerunning an analysis from scratch when anything in the model or data changes, one can use an existing variational starting inference as a starting point for a slightly different analysis. For example, we are currently pursuing methods to combine variational inferences on overlapping data subsets in order to get an inference on their union. This will enable the first divide-and-conquer approach to Bayesian phylogenetics, as well as online algorithms that add sequences to an existing posterior distribution by combining the previous variational approximation with one inferred from the new sequence and the sequences in its close neighborhood.

Finally, in order to obtain both the asymptotic correctness of MCMC and the efficiency of variational approaches, one can use the variational distribution as a proposal distribution. This returns to the original objectives of these parameterizations [4]. However, we are going a step further than this previous work: rather than use the parameterization to guide local modifications, we are using it to propose entirely new phylogenetic tree structures to obtain an independence sampler. Even a modest acceptance probability for such a sampler would make it a valuable tool to erase autocorrelation in phylogenetic MCMC chains.

Acknowledgments

This work was supported by National Science Foundation grant CISE-1564137, as well as National Institutes of Health grant R01-GM113246. The research of Frederick Matsen was supported in part by a Faculty Scholar grant from the Howard Hughes Medical Institute and the Simons Foundation.

References

[1] Erik M Volz and Simon D W Frost. Inferring the source of transmission with phylogenetic data. *PLoS Comput. Biol.*, 9(12):e1003397, December 2013.

[2] Bastien Boussau, Samuel Blanquart, Anamaria Necsulea, Nicolas Lartillot, and Manolo Gouy. Parallel adaptations to high temperatures in the archaean eon. *Nature*, 456(7224):942–945, November 2008.

[3] Chris Whidden and Frederick A Matsen, IV. Quantifying MCMC exploration of phylogenetic tree space. *Syst. Biol.*, 64(3):472–491, May 2015.

[4] Sebastian Höhna and Alexei J. Drummond. Guided tree topology proposals for Bayesian phylogenetic inference. *Syst. Biol.*, 61(1):1–11, January 2012.

[5] Bret Larget. The estimation of tree posterior probabilities using conditional clade probability distributions. *Syst. Biol.*, 62(4):501–511, July 2013.

[6] Cheng Zhang and Frederick A Matsen, IV. Generalizing tree probability estimation via Bayesian networks. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1449–1458. Curran Associates, Inc., 2018.

[7] J Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–376, 1981.

[8] Vladimir N Minin, Erik W Bloomquist, and Marc A Suchard. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.*, 25(7):1459–1471, July 2008.

[9] Sebastien Roch. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 3(1):92–94, January 2006.

[10] Gytis Dudas, Luiz Max Carvalho, Trevor Bedford, Andrew J Tatem, Guy Baele, Nuno R Faria, Daniel J Park, Jason T Ladner, Armando Arias, Danny Asogun, Filip Bielejec, Sarah L Caddy, Matthew Cotten, Jonathan D'Ambrozio, Simon Dellicour, Antonino Di Caro, Joseph W Diclaro, Sophie Duraffour, Michael J Elmore, Lawrence S Fakoli, Ousmane Faye, Merle L Gilbert, Sahr M Gevao, Stephen Gire, Adrianne Gladden-Young, Andreas Gnirke, Augustine Goba, Donald S Grant, Bart L Haagmans, Julian A Hiscox, Umaru Jah, Jeffrey R Kugelman, Di Liu, Jia Lu, Christine M Malboeuf, Suzanne Mate, David A Matthews, Christian B Matranga, Luke W Meredith, James Qu, Joshua Quick, Suzan D Pas, My V T Phan, Georgios Pollakis, Chantal B Reusken, Mariano Sanchez-Lockhart, Stephen F Schaffner, John S Schieffelin, Rachel S Sealfon, Etienne Simon-Loriere, Saskia L Smits, Kilian Stoecker, Lucy Thorne, Ekaete Alice Tobin, Mohamed A Vandi, Simon J Watson, Kendra West, Shannon Whitmer, Michael R Wiley, Sarah M Winnicki, Shirlee Wohl, Roman Wölfel, Nathan L Yozwiak, Kristian G Andersen, Sylvia O Blyden, Fatorma Bolay, Miles W Carroll, Bernice Dahn, Boubacar Diallo, Pierre Formenty, Christophe Fraser, George F Gao, Robert F Garry, Ian Goodfellow, Stephan Günther, Christian T Happi, Edward C Holmes, Brima Kargbo, Sakoba Keïta, Paul Kellam, Marion P G Koopmans, Jens H Kuhn, Nicholas J Loman, N'faly Magassouba, Dhamari Naidoo, Stuart T Nichol, Tolbert Nyenswah, Gustavo Palacios, Oliver G Pybus, Pardis C Sabeti, Amadou Sall, Ute Ströher, Isatta Wurie, Marc A Suchard, Philippe Lemey, and Andrew Rambaut. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature*, April 2017.

[11] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. November 2016.

[12] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. November 2016.

[13] Chris Whidden and Frederick A Matsen. Ricci–Ollivier curvature of the rooted phylogenetic subtree–prune–regraft graph. *Theor. Comput. Sci.*, 699(Supplement C):1–20, November 2017.

[14] Gergely J Szöllősi, Wojciech Rosikiewicz, Bastien Boussau, Eric Tannier, and Vincent Daubin. Efficient exploration of the space of reconciled gene trees. *Syst. Biol.*, 62(6):901–912, November 2013.

[15] Clemens Lakner, Paul van der Mark, John P Huelsenbeck, Bret Larget, and Fredrik Ronquist. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. *Syst. Biol.*, 57(1):86–103, February 2008.

[16] Andre J Aberer, Alexandros Stamatakis, and Fredrik Ronquist. An efficient independence sampler for updating branches in Bayesian Markov chain Monte Carlo sampling of phylogenetic trees. *Syst. Biol.*, 65(1):161–176, January 2016.

[17] Brian C Claywell, Vu Dinh, Mathieu Fourment, Connor O McCoy, and Frederick A Matsen, IV. A surrogate function for one-dimensional phylogenetic likelihoods. *Mol. Biol. Evol.*, 35(1):242–246, January 2018.

[18] Cheng Zhang and Frederick A Matsen, IV. Variational Bayesian phylogenetic inference. In *International Conference on Learning Representations (ICLR)*, 2019.

[19] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.*, 112(518):859–877, April 2017.

[20] Jörg Bornschein and Yoshua Bengio. Reweighted Wake-Sleep. June 2014.

[21] Andriy Mnih and Danilo J Rezende. Variational inference for Monte Carlo objectives. February 2016.

[22] Nicolas Lartillot and Hervé Philippe. Computing Bayes factors using thermodynamic integration. *Syst. Biol.*, 55(2):195–207, April 2006.

[23] Wangang Xie, Paul O. Lewis, Yu Fan, Lynn Kuo, and Ming-Hui Chen. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.*, 60(2):150–160, March 2011.

[24] Matthew D Homan and Andrew Gelman. The No-U-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, January 2014.

[25] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[26] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017.

[27] Dustin Tran, Alp Kucukelbir, Adji B Dieng, Maja Rudolph, Dawen Liang, and David M Blei. Edward: A library for probabilistic modeling, inference, and criticism. October 2016.

[28] Dustin Tran, Matthew D Hoffman, Rif A Saurous, Eugene Brevdo, Kevin Murphy, and David M Blei. Deep probabilistic programming. January 2017.

[29] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 2018.

[30] Diep Thi Hoang, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Le Sy Vinh. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.*, 35(2):518–522, February 2018.

[31] Daniel L Ayres, Aaron Darling, Derrick J Zwickl, Peter Beerli, Mark T Holder, Paul O Lewis, John P Huelsenbeck, Fredrik Ronquist, David L Swofford, Michael P Cummings, Andrew Rambaut, and Marc A Suchard. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.*, 61(1):170–173, January 2012.

[32] Daniel L Ayres, Michael P Cummings, Guy Baele, Aaron E Darling, Paul O Lewis, David L Swofford, John P Huelsenbeck, Philippe Lemey, Andrew Rambaut, and Marc A Suchard. BEAGLE 3: Improved performance, scaling, and usability for a High-Performance computing library for statistical phylogenetics. *Syst. Biol.*, April 2019.

[33] Xiang Ji, Zhenyu Zhang, Andrew Holbrook, Akihiko Nishimura, Guy Baele, Andrew Rambaut, Philippe Lemey, and Marc A Suchard. Gradients do grow on trees: a linear-time O(N)-dimensional gradient for statistical phylogenetics. May 2019.