Using Parsimony-Guided Tree Proposals to Accelerate Convergence in Bayesian Phylogenetic Inference

Seminar Report

Luise Häuser

ufziw@student.kit.edu

Abstract. In this report, we examine the paper *Using Parsimony-Guided Tree Proposals to Accelerate Convergence in Bayesian Phylogenetic Inference* published in 2020 by Chi Zhang, John P. Huelsenbeck and Frederic Ronquist [14]. They propose improved topological moves as proposal mechanisms for the Metropolis-Hastings-Algorithm. We introduce these moves and retrace the ideas behind them. Furthermore, we consider the experiments the authors carried out on empirical data. Additionally, we analyze the promising results presented in the publication.

Keywords: Bayesian Phylogenetics · Parsimony

1 Introduction

Working on the tree inference problem in phylogenetics, the Markov chain Monte Carlo (MCMC) method realized in the Metropolis-Hastings-Algorithm [10] [4] is a common approach. The paper *Using Parsimony-Guided Tree Proposals to Accelerate Convergence in Bayesian Phylogenetic Inference* [14] published by Zhang et al. proposes improvements to it. The MCMC method's aim is to find a set of assignments for $\theta \coloneqq \{\text{tree topology, branch lengths, model parameters}\}$ overall admitting a high posterior probability $f(\theta \,\middle|\, A)$ for the given alignment A. Samples are drawn iteratively to approximate the posterior. The next sample point θ_{i+1} is determined based on the current sample point θ_i using a Markov chain with a *proposal distribution* $Q(\theta_{i+1} \,\middle|\, \theta_i)$. This distribution is composed of several proposal mechanisms such as the topological moves introduced in Section 2. To decide whether a proposed data point should get sampled next, the *acceptance ratio* R is determined (see Equation 1) and then the proposal is accepted with probability $p \coloneqq \min(1, R)$.

$$R \coloneqq \frac{f(A \,|\, \theta_{i+1})}{f(A \,|\, \theta_i)} \cdot \frac{f(\theta_{i+1})}{f(\theta_i)} \cdot \frac{Q(\theta_i \,|\, \theta_{i+1})}{Q(\theta_{i+1} \,|\, \theta_i)} \tag{1}$$

In the formula, $f(A \,|\, \theta_{i+1})/f(A \,|\, \theta_i)$ denotes the ratio of the data points' likelihoods, computed as introduced in [3]. $f(\theta_{i+1})/f(\theta_i)$ is the ratio of the prior probabilities and the last factor $Q(\theta_i \,|\, \theta_{i+1})/Q(\theta_{i+1} \,|\, \theta_i)$ is known as *Hastings ratio* [4].

Both the convergence speed and the quality of the results depend on the proposal mechanisms used. Hence, this part of the method admits a high potential for optimizations. Considering the acceptance ratio (see Equation 1), we observe that for $Q(\theta_{i+1} \,|\, \theta_i) = f(A \,|\, \theta_{i+1}) \cdot f(\theta_{i+1})$, $R = 1$ always holds which means that every proposal gets accepted. Zhang et al. propose to approximate the likelihoods using parsimony as introduced in [3]. The parsimony score $\mathrm{Par}(T)$ of a tree T can be computed significantly faster than its likelihood [11] which the authors exploit to develop improved proposal mechanisms. They focus on the topological moves Subtree-Pruning and Regrafting (SPR), and Tree Bisection and Reconnection (TBR) which we introduce in Section 2. The parsimony-guided versions of these moves we subsequently consider in Section 3.

Zhang et al. implemented the introduced move types in the framework MrBayes [7] [12] and evaluated them on different datasets (see Section 4). Although there exist links between parsimony and probabilistic phylogenetics [6] [3], likelihood and parsimony are still different metrics, hence mismatches occur. In their experiments, the authors were able to obtain promising results indicating that the possibility to generate a preview predominates these mismatches (see Section 5). Nevertheless, there are some points which can be discussed further, some we finally mention in Section 6.

2 Topological Moves

In the following, we explain the topological moves *Subtree Pruning and Regrafting* (SPR) and *Tree Bisection and Reconnection* (TBR) as introduced in [3].

Applying a SPR move to a phylogenetic tree, we prune a subtree out of it and reinsert this subtree at some other point. The detailed procedure is illustrated in Figure 1. For using SPR in the MCMC method, the choice of r needs to be randomized. For this purpose, the *extending SPR* (eSPR) is introduced in [9].

For a TBR move, we start by randomly selecting a branch a from our phylogenetic tree. In each of the subtrees next to a, a pendant branch q_1, q_2 respectively, is picked. Subsequently, the three branches a, q_1, and q_2 are cut out of the tree. This produces two disconnected subtrees. In each of them, we choose a reconnection point r_1, r_2 respectively. Next, the subtrees are joined again by inserting the pendant branches next to the reconnection points and like this a new tree topology is obtained. As a proposal mechanism for the MCMC method, *extending TBR* (eTBR) is used [9].

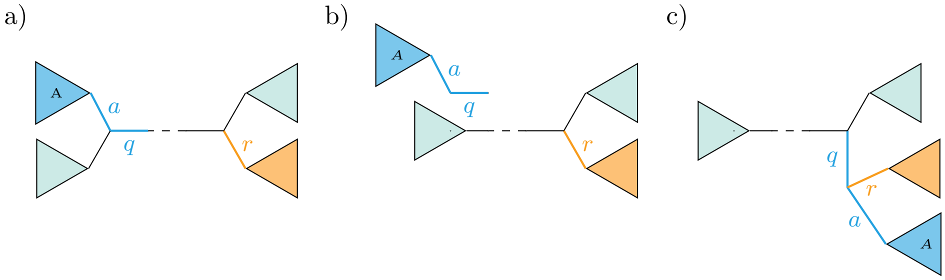

Fig. 1: Subtree-Pruning and Regrafting (SPR) a) Initial tree topology. A branch a is selected at random, with A we label a subtree next to it. In the other subtree, a branch r is selected for regrafting. By q we denote the pendant branch which is the branch next to a on the path to r. b) Together with the branches a and q, the subtree A is pruned out of the tree c) Resulting tree topology after reinserting the pruned subtree next to r.

3 Parsimony-Guided Moves

Zhang et al. introduce parsimony-guided versions of SPR and TBR. *Parsimony-guided SPR* (pSPR) uses parsimony to determine the regrafting point. Thus, they consider the topology of the phylogenetic tree after the pruning of the subtree A together with the branches a and q. Let $\mathcal{B}$ be the remaining subtree and $E(\mathcal{B})$ its branch set. The branch adjacent to a which is left back, is denoted by b (see Figure 2).

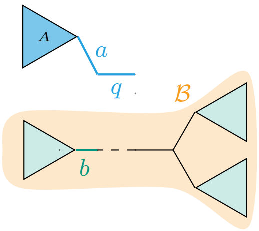

Fig. 2: Tree topology after pruning the subtree $A+a+q$ at SPR. b denotes the branch adjacent to a which is left back. All branches in $\mathcal{B}$ except b are considered as potential regrafting points.

All branches in $E(\mathcal{B})$ except b are taken into account as potential regrafting points. b is excluded because it would produce the same tree topology as before. For each branch i in $E(\mathcal{B}) \setminus b$, let T_i denote the tree which is obtained, when choosing i for regrafting. Further, let the score S_i be defined as:

$$S_i := \mathrm{Par}(T_i) - (\mathrm{Par}(A + a + q) + \mathrm{Par}(\mathcal{B})) \tag{2}$$

Determining these scores for all relevant branches is computationally convenient as it does not require a separate parsimony calculation for every branch. Instead, Zhang et al. once compute the intermediate results for the root of $A+a+q$ and for the inner nodes of $\mathcal{B}$. Subsequently, they can finish the computations for every branch by considering the involved nodes only. (For details concerning parsimony computation see [3]).

Regarding the choice of the regrafting point r, the authors compute a weight $\omega_i := \rho_i^{\epsilon S_i}$ for every relevant branch. ϵ is the so called *wrap factor*. It can be adapted to adjust the influence of the parsimony score on the weights. In their implementation, the authors used $\epsilon = 0.5$. ρ_i is the *base factor*. For its choice, Zhang et al. introduce different schemes, pSPR_1 and pSPR_2. The probability for i to get proposed for regrafting is determined as $\omega_i / \sum_{j \in \mathcal{B}/b} \omega_j$.

Applying pSPR_1, the authors set $\rho_i = e^{-1}$ for every branch. For pSPR_2, they introduce a more complex estimation. Let c be the length of the sequences in the phylogenetic tree. s_i denotes the parsimony score associated with the branch i. Further, let x and y be the sequences associated with the nodes incident to i. In the case of a tip node, the respective sequence can be inferred directly from the alignment. For inner nodes, we consider the sequences constructed during parsimony computation (see [3]). In order to understand the idea behind pSPR_2, we ignore the ambiguous characters these sequences can contain. Thus, we assume that s_i corresponds to the number of sites differing in x and y. Zhang et al. estimate the branch length of i with $\upsilon_i := s_i/c + \eta$. η is a small positive number to avoid $\upsilon_i = 0$. Further, let $p_0(\upsilon_i)$ the probability for a site to be the same in x and y, and $p_1(\upsilon_i)$ the probability for a site to change. Hence, the probability for x evolving into y along i is given as $p_i^* = p_0(\upsilon_i)^{c-s_i} \cdot p_1(\upsilon_i)^{s_i}$. As for short branches, $p_0(\upsilon_i) \approx 1$, it follows that $p_i^* \approx p_1(\upsilon_i)^{s_i}$. For $\rho_i = p_1(\upsilon_i)$, ω_i is thus an approximation of p_i^*. Subsequently, Zhang et al. set $\rho_i = \frac{3}{4}(1 - e^{-\frac{4}{3}\upsilon_i})$, which is the corresponding value for $p_1(\upsilon_i)$ under the Jukes-Cantor-Model[8]. Using this approach, they obtain a larger base factor for longer branches. In return, the parsimony score has a lower impact on the weights. Likelihood and parsimony tend to be more apart when long branches are involved. With pSPR_2, the authors aim to counteract this effect.

Long-branch-attraction (LBA) is a phenomenon affecting parsimony. It describes the fact that tree topologies in which long branches are in the same subtree, admit better scores [2]. Because of this bias, Hastings correction is required for pSPR. To understand the corresponding ratio introduced by Zhang et al., we consider configurations θ_i and θ_{i+1}, as illustrated in Figure 3. We note that the subtree $\mathcal{B}$ is the same in both situations. To propose θ_{i+1} given θ_i, r is selected for regrafting with a probability of $Q(\theta_{i+1} \mid \theta_i) = \omega_r / (\sum_{j \in \mathcal{B}/b} \omega_j)$. For the reversed proposal, b is picked as a regrafting point. $Q(\theta_i \mid \theta_{i+1}) = \omega_b / (\sum_{i \in \mathcal{B}/r} \omega_i)$ is the probability for this to occur. The ratio of these two probabilities then corresponds to the Hastings ratio introduced for pSPR by Zhang et al.:

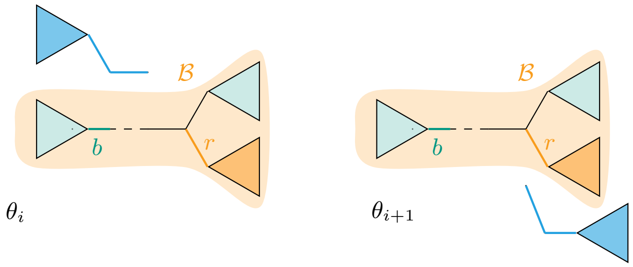

Fig. 3: Tree topologies after the pruning of $A + a + q$ at θ_i (θ_{i+1} resp.). Going from θ_i to θ_{i+1}, all branches in $\mathcal{B}$ except b are considered. Going the other way around, all branches in $\mathcal{B}$ except r are potential regrafting points.

$$Q(\theta_i \,|\, \theta_{i+1})/Q(\theta_{i+1} \,|\, \theta_i) = \frac{\omega_b}{\sum_{i\in\mathcal{B}/r} \omega_i} / \frac{\omega_r}{\sum_{j\in\mathcal{B}/b} \omega_j} \tag{3}$$

The pSPR move also changes some branch lengths. This affects the branches a, b and q and their lengths are updated via a scaler mover algorithm [5]. Equivalently to pSPR$_1$ and pSPR$_2$, Zhang et al. define the moves pTBR$_1$ and pTBR$_2$, applying the described mechanisms for determining the two reconnection points. Changes of branch lengths concern the branches a, q_1, and q_2.

For a pSPR move on a n-taxa tree, $2n$ parsimony scores must be calculated. Applying pTBR, there are two reconnection points and all possible combinations need to be taken into account. Therefore, the number of required parsimony computations grows quadratically with the number of taxa. As this can make the move computationally costly, Zhang et al. introduce an additional parameter δ. They only consider branches with a distance $\leq \delta$ to the bisection point as possible reconnection points. Thus only $2^{2(\delta+1)}$ parsimony scores must be computed. In their implementation, the authors set $\delta = 5$.

4 Experiments

Zhang et al. implemented the introduced move types in the framework MrBayes [7] [12] available on GitHub. For their experiments on empirical data, the authors began working on 20 empirical datasets. At first, they made a reference run with three heated chains and 20 Million generations. The purpose of this was to ensure convergence and to produce reference results. For every dataset they ran six independent chains. Then they determined the average standard deviation of split frequencies (ASDSF) [1] for the resulting tree sets. In the following, they only kept working on the six datasets for which this score was below 0.02.

In the main experiment, no heated chains were used and the algorithm ran for 10 Million generations only. Zhang et al. aimed to make the problem more challenging to assess the performance of the different move types. They executed the algorithm with three different setups, one for the extending move types and one for each of the two parsimony-guided schemes. SPR and TBR were always applied in a ratio of 2:1. For every setup, they ran 16 independent chains.

The authors provide a visualization of the tree spaces revealed for the six datasets used in the main experiment (see Figure 4). Using multidimensional scaling, the tree space computed for the respective dataset is projected into the plane according to the trees SPR distances. To compute these distances, the authors made use of RSPR, an open-source tool developed for this purpose. It implements a fixed-parameter algorithm admitting a runtime which grows exponentially with the resulting distance but only linearly regarding the size of the trees [13]. As considering the complete tree space was still too time consuming, the analysis got restricted to a subset of 4000 trees (10%).
The visualizations show that the tree sets vary in their distribution. In some, one or more islands can be observed, in others the data points are spread more evenly across the tree space. Hence, the authors evaluated their implementation on input data with different properties.

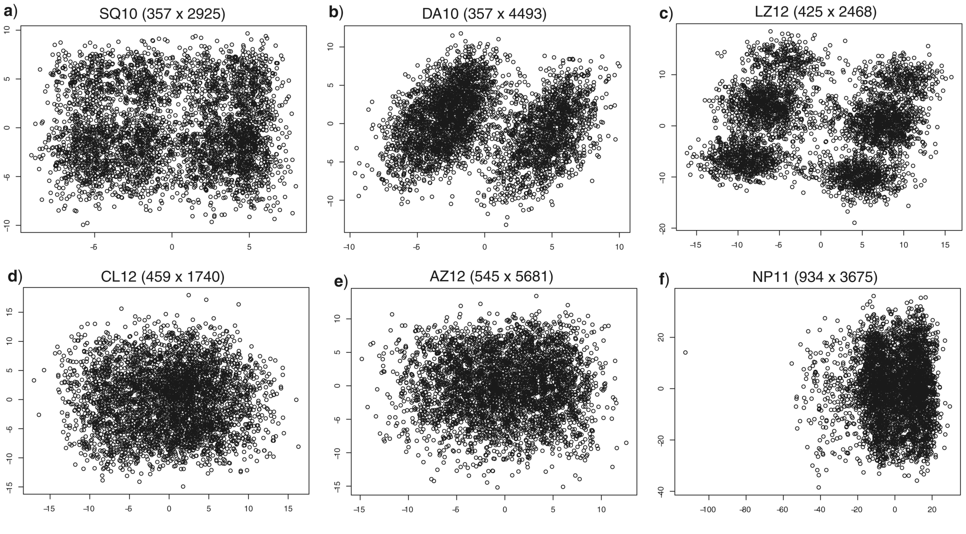

Fig. 4: Each diagram corresponds to one dataset with the given abbreviation, number of taxa and number of size. The cloud diagrams are produced by multidimensional scaling based on the SPR distances of 4000 sampled trees for the respective dataset (adapted from [14]).

5 Results

In their first analysis, Zhang et al. considered how the likelihood of the sampled tree evolves with the number of generations. The results for the dataset SQ10 are depicted in Figure 5. For the parsimony-guided moves, the likelihoods increase faster and reach better final results. Sometimes, stair-like structures occur in the diagrams, indicating that the respective chain got stuck in a local maximum. However, this is less likely to happen if parsimony-guided moves are used. Secondly, the authors focused on the ASDSF scores for the tree set in the current generation compared to the reference results. Smaller values are reached for the chains using parsimony-guided moves, hence This analysis reveals as well that faster convergence occurs, when parsimony-guided moves are used. Further, the Also in this analysis, faster convergence can be ob The outcome supports what was also discovered in the analysis of the likelihoods: Using parsimony-guided moves leads to faster convergence and moreover to results close to the reference (ASDSF < 0.002). The results for the other five datasets differ slightly but the major observations are the same.

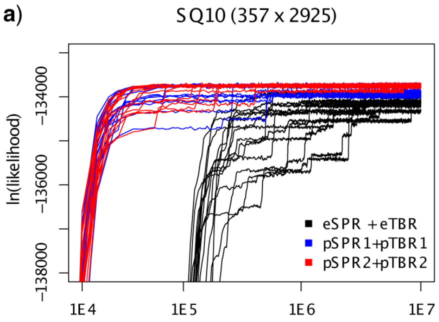

Fig. 5: Variation of the likelihood of the current sample with the number of generations. The x-axis gives the generations in a logarithmic scale, and the y-axis indicates the log-likelihood. Each line refers to an independent run. A different color is used for each setup (adapted from [14]).

Considering the results for the SQ-10 dataset, $pSPR_2$ and $pTBR_2$ seem to work better than $pSPR_1$ and $pTBR_1$. Taking the other datasets into account, no clear difference between the two schemes of can be observed.

Additionally, Zhang et al. mention that mixing parsimony-guided and extending proposals turned out to be helpful to improve convergence in certain datasets. Overall, they read from their results that the links between parsimony and likelihood predominate the occurring mismatches which makes the approximation of likelihood a with parsimony a powerful technique to improve convergence of Bayesian phylogenetic methods.

6 Discussion

Although the findings sound very promising, there are some points which can be discussed. In their main experiments, Zhang et al. worked on six datasets only. Further, these datasets admit a rather low variance regarding the number of taxa (ranging from 357 to 935) and the number of sites (ranging from 1740 to 5681). Using more data here would make the results more convincing.

In their analysis, the authors focus solely on improvements concerning the number of generations, disregarding runtime. They state that the parsimony-guided moves are not significantly slower than the extending ones, but do not provide evidence supporting this claim.

The promising results are however a clear motivation to keep working on parsimony-guided moves in order to improve convergence of Bayesian phylogenetic methods.

References

1. Aberer A.J. 2015. Algorithmic Advancements and Massive Parallelism for Large-Scale Datasets in Phylogenetic Bayesian Markov Chain Monte Carlo [dissertation]. Karlsruher Institute für Technologie.
2. Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Biol. 27:401410.
3. Felsenstein J. 2003. Inferring phylogenies. Sunderland (MA): Sinauer Associates.
4. Hastings W.K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 57:97109.
5. Holder M.T., Lewis P., Swofford D., Larget B. 2005. Hastings ratio of the LOCAL proposal used in Bayesian phylogenetics. Syst. Biol. 54:961965.
6. Huelsenbeck J.P., Ane C., Larget B., Ronquist F. 2008. A Bayesian perspective on a non-parsimonious parsimony model. Syst. Biol.57:406419.
7. Huelsenbeck J.P., Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754755.
8. Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N., editor. Mammalian protein metabolism. New York:Academic Press. p. 21132.
9. Lakner C., van der Mark P., Huelsenbeck J.P., Larget B., Ronquist F. 2008. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. Syst. Biol. 57:86103.
10. Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E. 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:10871092.
11. Ronquist F. 1998. Fast fitch-parsimony algorithms for large data sets. Cladistics. 14:387400.
12. Ronquist F., Huelsenbeck J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:15721574.
13. Whidden C., Matsen F.A. 2015. Quantifying MCMC exploration of phylogenetic tree space. Syst. Biol. 64:472491.
14. Zhang C., Huelsenbeck J.P., Ronquist F. 2020. Using Parsimony-Guided Tree Proposals to Accelerate Convergence in Bayesian Phylogenetic Inference. Syst. Biol. 69:10161032.