Evaluation of proposal distributions on clock-constrained trees in Bayesian phylogenetic inference

Sebastian Hoehna
Department of Computer Science
University of Auckland
Auckland, New Zealand
shhn001@ec.auckland.ac.nz

Alexei J. Drummond
Department of Computer Science
University of Auckland
Auckland, New Zealand
alexei@cs.auckland.ac.nz

ABSTRACT

Bayesian Markov chain Monte Carlo (MCMC) has become one of the principle methods of performing phylogenetic inference. Implementing the Markov chain Monte Carlo algorithm requires the definition of a proposal distribution which defines a transition kernel over the state space. The precise form of this transition kernel has a large impact on the computational efficiency of the algorithm. In this paper we investigate the efficiency of a number of different proposal distributions for clock-constrained phylogenetic trees (i.e. constrained by a strict or relaxed molecular clock). Clock-constrained trees have become increasingly important in phylogenetic inference, especially in the context of divergence time estimation and their constraints require substantially different proposal algorithms to unrooted phylogenetic trees.

We investigated the efficiency of seven proposal moves on clock-constraint trees first on a small data set and then on six additional data sets. In contrast to the results for the case of MCMC on unconstrained phylogenetic trees we found that subtree swapping moves perform better than subtree prune and regraft algorithms and moderate proposals dominate bold proposals. However, the results varied with the data set we analyzed and the intermediate subtree swap proposal distribution which we introduce in this paper was the only one with a continuous high level of efficiency.

Categories and Subject Descriptors

G.2.2 [**Discrete Mathematics**]: Graph Theory—*Trees*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Markov Processes*; G.4 [**Mathematics of Computing**]: Mathematical Software—*Algorithm design and analysis, Efficiency, Reliability and robustness*

Keywords

Clock-constraint phylogenetic inference, Bayesian MCMC, Mixing in Treespace

NZCSRSC 2008, April 2008, Christchurch, New Zealand.

1. INTRODUCTION

In the last decade Bayesian MCMC has been established as the dominant technique for phylogenetic inference. In the context of this paper phylogenetic inference is the evolutionary relation between various species. The relations are represented in a tree structure where the species are the external nodes and the common ancestors are the internal nodes. Reconstructing the optimal phylogenetic tree from DNA sequences is NP-Hard and therefore not feasible for more than 30 sequences[5]. Although no algorithm that guarantees the computation of the optimal phylogenetic tree under realistic probabilistic models for trees with hundreds of tips has been developed, Markov chain Monte Carlo (MCMC) algorithms provide a good approximation of the posterior distribution over phylogenetic trees given a multiple alignment of genetic sequences [8]. The key idea of the MCMC algorithm is collecting samples from a hypothesis space and conclude from them how the overall hypothesis space looks like and where the good hypothesis are located. To retrieve the samples a new hypothesis is proposed and is stochastically accepted or rejected proportional to its relative probability compared with the current state. Therefore the algorithm tends to produce more hypothesis with a high probability over time[14, 7].

The proposal kernel has a large impact on the overall performance of the MCMC run, so that good proposal kernels can lead to much faster convergence. A run is considered to have converged if it does not change the result (within a threshold) when further iterations or different initial trees were used. In contrast to this a poor proposal mechanism can lead to a MCMC algorithm that fails to correctly estimate the posterior distribution even for small data sets. The proposal kernels or algorithms, which we call moves in the rest of this paper, can be applied together with weights which specify the proportion of commitment of this move. In this research we focus on the behaviour and mixing in tree space of the most used tree proposal moves for clock-constrained trees and further improvements.

Phylogenetic trees can be represented as rooted or unrooted, whereby rooted trees are referred to ultrametric when all of the external nodes (tips) are contemporaneous (see Figure 4). More generally they are known as clock-constrained when tips are fixed at different times. Either way, clock-constrained trees have $n-1$ node heights for n tips whereas unrooted trees have $2n-3$ branch lengths. Modifications to the standard tree proposals on unrooted trees have to be made to accommodate clock-constraints because rooted

trees have more restrictions than unrooted trees. In particular, the clock constraint forces each parent node in a phylogenetic tree to be older than both of its children. Some of the most often used moves such as the Tree-Bisection-and-Reconnetion (TBR), Subtree-Pune-and-Regraft (SPR) and Nearest-Neighbor-Interchange (NNI) moves violate this constraint. To maintain the order of the nodes the SPR and NNI moves can be modified to change the node height too, whereas for the TBR move this is not feasible without more complex computations. Following we will only consider clock-constraint phylogenetic trees shortly called trees if not other specified.

In this paper we will first discuss several metrics to evaluate the tree proposal moves. This is followed by a discussion of the tree proposal moves implemented in BEAST [4] and other MCMC software packages for clock trees. Finally, we present a comparison for the single moves and the improvement which can be achieved by using the best moves with the best weights compared to the current default settings in BEAST.

2. RELATED WORK

Tree distance metrics are studied over the last twenty years in several papers. Bourque[1] proposed the symmetric difference and Kuhner and Felsenstein[9] the Branch-Score metric. These two metrics are tested as representative examples for the current available metrics and their capabilities to measure the mixing in tree space.

Lakner et al.[10] introduced a framework for measuring efficiency of tree proposals for non-clock trees with the phylogentic inference software MrBayes [16]. The idea behind their approach is to measure the number of steps required for the chain to converge. Therefore they assumed that the time a move consumed is equal for all operators. A second value they measured was the percentage of converged runs for each single tree proposal operator. A run has converged in their definition if it reaches an average standard deviation of split frequencies below 0.01 compared to a reference run. The reference run is retrieved ahead by several independent Metropolis Coupled MCMC runs (MCMCMC). Finally they present their most striking results stating that bold topology change proposals with a bias preferring more local rearrangements perform best.

In our results we will show similar behaviour for clock trees. The tree proposal moves Narrow Exchange, Wide Exchange, Wilson-Balding and Subtree-Slide are taken from Drummond et al.[3] and the NNI move from Felsenstein et al.[5].

3. METHODS

3.1 Convergence Diagnostics of MCMC runs

A review of convergence diagnosis algorithms for Markov chain Monte Carlo algorithms is given by Cowles and Carlin[2]. Their investigation showed 13 different convergence diagnostics and concluded that no algorithm is known which meets the demands of detecting convergence without failures. They compared diagnostics which do not use any additional knowledge as how the optimal distribution should be. This is due to the huge amount of time a single run takes and hence usually no previous runs are created which could present this optimal distribution. This is caused by the fact that a run with the optimal result requires to be longer for being more precise. Contrary to this we can use the prior knowledge of the optimal distribution as we analyse the performance of the MCMC algorithm instead. In our evaluation the same data set is used for the different moves and the time to obtain the desired distrubtion is more important than the distribution itself. So the data set can be chosen where the optimal distribution is known. A useful prior knowledge for the convergence is a reference run which can be used as a benchmark for the runs in the test. Next, it indicates whether a run has converged to the right posterior distribution if the test run has stabilized close to this "golden run".

Recall that a higher performance and faster convergence of MCMC algorithms can be achieved by good mixing in tree space. The more an algorithm obtains samples distributed over the whole space the better the spamles represent the distribution in the space. Tree comparison metrics or tree distance metrics are defined as a measurement of the relation between two trees according to their topology and branch length. This characteristic could be used for measuring the mixing in tree space as how often trees with different distances to a fixed tree are sampled. The more trees with a high distance and all distance in between are sampled the higher is the chance that the whole space was explored. However, non of the tree metrics we used could give an indication of the performance of the MCMC run. Therefore we focused on two further ideas. Namely the Split Swap Rate which we newly introduce in this paper and the convergence time measured by the maximal deviation of split frequency.

3.1.1 Convergence time by the maximal deviation of split frequency

The state of the art technique (i.e. implemented in AWTY[18]) to evaluate two MCMC runs is comparing the standard deviation of split frequencies. The splits are defined as groups of species (taxa) and the split frequency defines the frequency in percent how often this group of species occured in the samples. This means how often a subtree was present. To obtain a measure how related the two trees are we used the maximal difference of split frequencies.

Since MCMC algorithms are probabilistic algorithms the performance can vary for the different runs. This leads to a deviation of the performance, i.e. the maximal difference of split frequencies is different after the same amount of steps. Hence the mean of multiple runs is necessary for a reliable evaluation.

In our studies we used two different means. First we calculated the integral of the maximal difference of split frequencies over the chain length. Second we calculated the mean of the maximal difference of split frequencies after a fixed number of steps. The integral of the maximal difference of split frequencies is assumed to give a better result because it uses the past of the run too. But either way of calculating the mean seems to be justifiable because both diagnostics agree with each other in the performance of the different moves.

3.1.2 Split Swap Rate

We believe that topology changes can be obtained by observing the partitions (splits) of a data set. Therefore we invented the Split Swap Rate to measure efficiency comparable between different tree proposal distributions. Recall that measuring the efficiency of tree proposal moves depends on the time needed to converge for a MCMC run and the convergence is defined by the deviation of split frequencies. Then the frequencies are approximately close enough to their actual values if there is a high rate of swaps of presence and the maximal time a run spent in one state (absent or present) is comparable small to the length of the run or the number of swaps. The focus for this test lies on the splits which have a frequency between 5-95% of the golden run. Every other split will have a neglecting small swap rate and is assumed to be similar over all operators.

Let r_{SS} denote the split swap rate, s_i the swaps for split i, $\bar{s_i}$ the average swaps of all operators for split i and σ_i the standard deviation of swaps of all operators for split i .

$$r_{SS} = \frac{\sum_{i=1}^{n} \frac{(s_i - \bar{s_i})}{\sigma_i}}{n} \quad (1)$$

The split swap rate turned out to be good for detailed analysis where the tree proposal operators mix well and hence it is a good estimator for convergence. Furthermore we could verify in our tests that clades exist which are hard to swap for local operators and prune and regraft operators with this metric. This was obvious by a low split swap rate for all these operators compared to the Wide Exchange in the affected clades (see figure 2). Mossel and Vigoda[15] stated that these tree valleys exist which are hard to traverse for prune and regraft operators. Ronquist et al. [17] contradicted that these valleys are unlikely to occur in real data sets. Lakner et al.[10] also failed to show these valleys. However, in our results we can clearly present them (see figure 4).

3.2 Tree proposal moves

The kernel of the MCMC algorithm for phylogenetic inference are tree proposal moves. The tree proposal moves can be classified in two categories: Branch change moves and subtree rearrangement moves. Branch change moves focus on changing the branch length of the tree with topology changes as side products of it. Further, the moves can be divided into global and local regarding to their dimension. A local version is the LOCAL move from [12, 13] and [11] which is similar to the Subtree-Slide move in BEAST from Drummond et al.[3]. The local move apply changes in a small, local area of the tree whereas the global move makes topology changes which can affect wider parts of the tree. Therefore the local moves are considered to be moderate and the global moves bold. We will consider the Subtree-Slide move as the only branch length move since the performance for the global move has shown to be poor.

The tree rearrangement moves can be further grouped into subtree swapping moves and subtree prune and regraft moves. The subtree swapping moves exchange two subtrees either locally or globally. This global exchange of two subtrees is called Wide Exchange and the local one Narrow Exchange respectively. The Narrow and Wide Exchange we used do not modify the branch length but could be modified to change the branch length too. However, shorting the branch length could demand to change the branch length of further nodes in the subtree which leads to a lower posterior probability of the proposal in most cases and therefore is unlikely to produce good proposals. The other alternative, extending the branch length, is not necessary because it does not violate the constraint for root clock-constraint trees if the branch length remain the same and changing the branch length could prevent good topology proposals from being accepted. Additionally to these two existing swapping moves, a mixture between Narrow and Wide Exchange was developed to obtain a move which is in the middle between moderate and bold.

The subtree prune and regraft moves listed from moderate to bold are: NNI, SPR and TBR. The NNI can be considered as a subtree swapping move too where the Narrow Exchange is the counterpart from the group of subtree swapping moves. Hence, we extended the Narrow Exchange move to obtain the NNI in its original meaning. This offers the possibility to compare the local subtree rearrangement move with and without branch length changing. The SPR move is more difficult for clock-constraint phylogenetic trees than for unconstraint or unrooted trees. It can not be implemented without further modification for the heights of the nodes. Wilson and Balding proposed one alternative [19] for the SPR move which is implemented in BEAST [3]. Instead of allowing a subtree to be attached on every branch it is restricted to reattach the subtree as a child of a node which has to be higher in the tree so that the height does not have to be decreased. So the disadvantage of decreasing further nodes in the subtree is eliminated. To compare the SPR with an alternative without changing the branch length, a new move which we call FNPR (Fixed Nodeheight Prune and Regraft) was developed. Since the only case of the TBR which does not need a modification of the node order for rooted trees is the SPR, the TBR is assumed to not give any further improvement to the SPR.

Following the implementation of the 7 used moves are described more in detail. All moves are reversible. A reversible move is defined as a move move which can undo the change by another move of the same algorithm. The ratio of how likely it is to obtain the new proposal is called Hastings ratio. Let P_F define the probability of proposing the new tree (forward probability) and P_B the probability of undoing the move (backward probability). Then, the Hastings ratio is $hr = \frac{P_B}{P_F}$[7]. The Hastings ratio is necessary to remove the bias from the chain with these moves to prefer trees with a high probability of being proposed.

3.2.1 Subtree-Slide

The Subtree-Slide move [3] is similar to the LOCAL move proposed by [12, 13] and [11]. The purpose of this move is rather changing the branch length than proposing a new topology. A node is randomly chosen and drawn either on a path towards the root or leafs. For each branch the path is chosen randomly on the way down. Instead of using the idea for sliding a distance on a path down as implemented in BEAST, the LOCAL move changes the node height. [12, 13] proposed to select randomly a canonical representation of the tree. The canonical representation defines the order or in this context the path to choose if a node height is

changed. The topology is generated from the representation by taking the oldest node to the right and left as the children recursively. The length of the slide is computed randomly between $-\Delta$ and Δ whereby the sign indicates the direction. The Hastings ratio is defined as follows with n as the number of nodes passed on the path:

$$hr = \begin{cases} \frac{1}{2^n} & ;if \Delta > 0 \\ 1 & ;otherwise \end{cases} \quad (2)$$

3.2.2 Narrow Exchange

Drummond et al.[3] implemented for their phylogenetic inference software BEAST [4] a subtree swap move called Narrow Exchange. The Narrow Exchange move is similar to the NNI move and can be considered as a subset of it. A node is randomly chosen and exchanged with the sibling of the parent node if this exchange does not violate the requirements of a clock-constraint phylogenetic tree (if the sibling of the parent has a lower molecular time than the parent node). Therefore the molecular time of the nodes has not to be adjusted.

Every node is chosen randomly and swapped with the sibling of its parent. Hence the chance for taking this proposal is $\frac{1}{n}$ where n is the number of nodes. Since the topological changes affect only the two swapped subtrees and this proposal is reversible, the probability for a backward proposal is the same $\frac{1}{n}$. The Hastings ratio follows to be 1.

3.2.3 Wide Exchange

The Wide Exchange move is a generalization of the Narrow Exchange where the second node is chosen randomly too [3]. Both nodes can be arbitrarily far away from each other and thus this subtree swap move is global.

Let i and j denote two arbitrary chosen nodes. Further, iP is the parent of i and jP the parent of j. The two nodes i and j are swapped if $height(i) < height(jP)$ and $height(j) < height(iP)$, otherwise it fails.

Both nodes are selected randomly without any additional information which nodes to prefer and is reversible since the only constraint for this move belongs to the node heights and these are not changed. This leads to the Hastings ratio of 1.

3.2.4 Intermediate Exchange

The Intermediate Exchange move is newly introduced in this paper and obtained as a mixture between Narrow and Wide Exchange. It can be considered as a Wide Exchange where the second node is not chosen totally arbitrary. The selection process is given a bias to prefer local nodes with a higher chance. The probability of a node to get selected as second depends on the path length to the first node. Let i denote the first node, j the second node and l_{ij} the path length between them in the current tree, then the probability of choosing j after i was chosen is:

$$P_i(j) = \frac{l_{ij}}{\sum_{k=1}^{n_i} l_{ik}} \quad (3)$$

where n_i are all possible nodes to swap from i considering the node height constraint, and n_j respectively.

The Hastings ratio results from the equation:

$$Hr = \frac{\frac{1}{\sum_{k=1}^{n'_i} l_{ik}} + \frac{1}{\sum_{k=1}^{n'_j} l_{jk}}}{\frac{1}{\sum_{k=1}^{n_i} l_{ik}} + \frac{1}{\sum_{k=1}^{n_j} l_{jk}}} \quad (4)$$

where n'_i is the set of nodes i can swap to after the proposal and n'_j respectively.

3.2.5 NNI

The Nearest Neighbor Interchange (NNI) is one of the standard operators for tree rearrangements [5]. [6] showed an implementation for clock-constraint phylogenetic trees. A branch $P \rightarrow C$, where P is the parent of C, is selected randomly and one child of C is swapped with the other child of P. Further, B is the second child of P and G the parent of P (grandparent of C). The Narrow Exchange is a subset of the NNI where the only difference is that the node heights are changed to allow changes where the chronological order could be disarranged. Therefore the node height of P is set randomly between the height of G (h_G) and $max(h_U, h_B)$ where h_U and h_B is the height of U and B respectively. To derive the Hastings ratio hr let h_I denote the height of i, h_G the height of the grandparent of i and h_B the height of the sibling of i.

$$hr = min(1, \frac{h_G - max(h_U, h_B)}{h_G - max(h_I, h_B)}) \quad (5)$$

3.2.6 Wilson-Balding

The Subtree Prune and Regraft move [5] is another standard operator for phylogenetic inference. The SPR move selects randomly a non leaf and non root node i. The subtree rooted at iP is pruned by breaking the edges between the parent and the other child of iP and connecting these two together. We used a variation of the SPR move which just changes the height of iP because clock-constraint trees have stronger constraints and changing the heights of many node decrease the likelyhood of the tree to be accepted. The reattachment point is found by a node which has a height between the height of i and the root. So the height of iP is set arbitrary between the height of i and the height of the root as proposed by [19]. This move is a subset of the original SPR. Let i and j denote two arbitrary chosen nodes. iP and jP are the parent nodes of i and j respectively. The moves fails if $height(i) > height(jP)$. The range for the new height of iP is $height(jP) - max(height(i), height(j))$. The subtree belonging to i is pruned and reattached above j. jP is the pruning node and used as the reattachment node too.

$$hr = \frac{height(iG) - max(height(i), height(iB))}{height(jP) - max(height(i), height(j))} \quad (6)$$

, where iG denote the grandparent of i and iB the brother of i.

3.2.7 FNPR

Another alternative for the SPR move is to fix all node heights. We name this move FNPR (Fixed Nodeheight Prune and Regraft). Thus a subtree is pruned and reattached without changing any node heights. The FNPR is a smaller subset of the Wilson-Balding move. First, a node i is picked randomly and a second node j is selected arbitrary from all nodes. If $height(j) > height(iP)$ or $height(iP) > height(jP)$

then the move fails because iP does not fit between j and jP without changing the height. iP and jP are the parent nodes of i and j respectively. iP is pruned from its original place and reattached above j.

The Hastings ratio for this move is 1 because no node heights are changed and both nodes are chosen arbitrary. The possibility to make the backward proposal remains the same as the forward probability.

3.3 Estimating the weights for the moves

The combination of the operators is as important as the choice of the operators. Several of the described operators are working in different neighbourhoods which complement each other. The Wide Exchange and the Narrow Exchange as well as the NNI, Wilson-Balding and FNPR move share some but not all tree rearrangements. Defining a general weighting scheme which moves and how often they should be used is very complicated because the moves and the combinations perform different on the diverse data sets. Further the dimensions of this optimization problem growth with the amount of available moves. We approach this problem with a Genetic Algorithm (GA) which optimizes the weights regarding the average distance which the MCMC algorithm had with these values. The MCMC algorithm was run on the Anolis data set with a chain length of 1 million. We assume that a good performance on this data set will give a good performance on most other data sets too. The results of the GA can be used as a further indication which moves should be used together.

4. TESTS

The tests for this research were done with a developer's version of BEAST 1.4 where additional moves and metrics were implemented. We divided the tests into two phases, the optimization phase and validation phase. First, a smaller sample data set was used to evaluate and optimize the moves. Second, the results of the first phase were validated with more tests on other data sets.

4.1 Evaluation of the tree proposal distributions

The data set for this first evaluation was extracted from a set of 19 Anolis species which were reduced to the 950 interesting nucleotides. For retrieving the reference run called "golden run" we used the standard BEAST moves (Narrow Exchange, Wide Exchange, Subtree-Slide and Wilson-Balding) and run this chain for a length of 100 million. This length is much longer than usually applied for data sets of this size. A second run from a different starting point was used to validate the result. With the online tool AWTY (Are We There Yet [18]) we could show that this golden run had converged because the maximal difference of split frequencies were very low. First we tested each algorithm separately on its performance. In addition to the tree proposal moves an internal branch length change operator and root height change operator was used because of the nature of some moves which do not change the branch length. This was added in all runs to give all operators the same amount of proposals during the run. A second approach would be to include this move to the moves without branch length changes. Then it would not be possible to analyze if moves with or without branch length changes perform better. For all operators we can show that a separation of the topology and branch length proposal improves the performance.

The performance of each operator was evaluated in two ways. For every move ten runs were executed for a chain length of 10 million and from independent initial trees to observe the behaviour for a move. The split swap rate for all clades which are between 5-95% of the golden run was measured and compared to the swap rate of all other runs. Second the maximal difference of the split frequencies in the current run compared to the golden run was measured to observe how close they are. The integral of these values was calculated and compared with the integrals of the other moves.

4.2 Validation of the results

Data set	No. of taxa	No. of sites	Type of data	TreeBase matrix acc. no.
1	27	1,949	rRNA, 18s	M336
2	29	2,520	rDNA, 18s	M501
3	36	1,812	mtDNA, COII (1 - 678), cytb (679 - 1,812)	M1510
4	41	1,137	rDNA, 18s	M1366
5	43	1,660	rDNA, 18s	M932
6	50	378	Nuclear protein coding, *wingless*	see Acc. Numbers

Figure 1: The 6 first data sets used in the research by Lakner et al. [10] which we used to confirm our theses.

Lakner et al.[10] used for their research a set of 12 different data sets between 27 to 71 taxa. To get comparable test results we used the same data sets to validate our research from the Anolis data set but we performed the tests only on the first 6 data sets (see figure 1). However, from their conclusion we could assume that the first 6 data sets are sufficient for a qualitative conclusion since they could not verify any difference of the performance for the moves related to the size of the data set. Finally, the same framework was used as described before and additional attention was paid to the patterns observed in the first test phase.

4.3 Results

4.3.1 Phase 1

In the first phase we obtained a height map for the split swap rate (see Figure 2) and the convergence time (Figure 3) for our test data set. Focusing just on the swap rate the

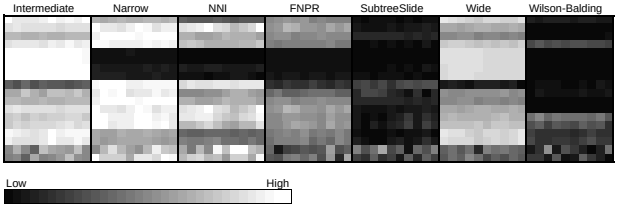

Figure 2: The split swap rate for the 7 different moves. Each move was run from 10 different starting trees and the split swap rate measured for the 18 clades between 5-95%. This height map represents the greyscale where white is the highest split swap value for this clade. Black means low respectively.

Narrow Exchange outperforms the global counterpart Wide Exchange in many clades according to its higher mean swap rate. However, the convergence (shown in Figure 3) is worse.

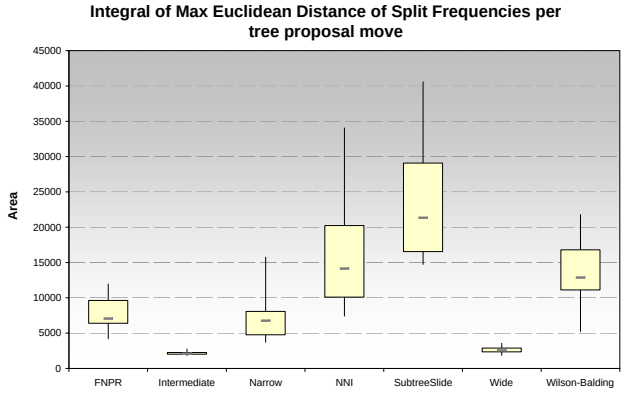

Figure 3: This box plot shows the integral over the distance curves. The distance was measure by the maximal difference of the split frequencies. This was done for 10 independent runs for each single tree proposal move on the Anolis data set. The length of the MCMC run was 10 million.

This is due to the very poor mixing of four out of the 18 observed clades which leads to a worse overall performance in contrast to the global moves. These 4 clades in our data set are very hard to swap for the local and subtree prune and regraft moves. The clades can be represented in trees which are only separated by one Wide Exchange move. Since these clades are crucial for the final result the convergence statistic was poor. In figure 4 we can show one pair of clades which

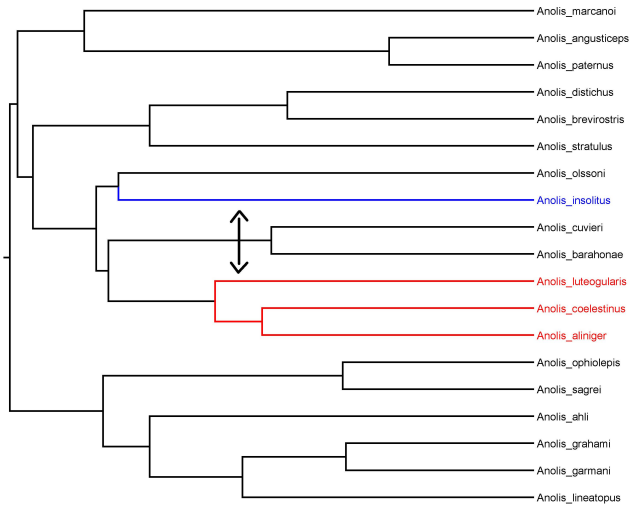

Figure 4: A phylogenetic tree showing the swap done by the Wide Exchange to traverse the tree valley. This was can not be done by any other move in one step.

can be swapped by one Wide Exchange move. Obviously it needs more than just one Narrow Exchange or any other prune and regraft step to swap these clades. This verifies the statement that real data sets exist for which a global subtree swapping move is essential for fast convergence.

The Wilson-Balding move and the FNPR move can be both considered as variations of the original SPR move. Both moves prune and regraft a subtree globally. The main difference is that our implementation of the FNPR move does not change a branch length. The Wilson-Balding move can change the branch length of the root of the pruned subtree randomly between the root height and the height of its remaining child. An obvious result of our test shows that the FNPR move performs much better than the Wilson-Balding move. Further, the acceptance rate for the FNPR move was $P_{Acc} \approx 0.0084$ which is approximately twice as high as the acceptance rate for the Wilson-Balding move which was $P_{Acc} \approx 0.0046$. The acceptance rate is the percentage of accepted proposals. The higher the posterior probability (fitness) of a proposal the higher the chance it gets accepted.

The same behaviour is observable for the Narrow Exchange and NNI. Both indicate that changing the branch length simultaneously decreases the acceptance rate. Next it shows that just a broad range of topology proposals and a higher acceptance rate are crucial for a high swap rate and fast convergence.

For our small test data set we can confirm the observation from [10] that the branch length move namely Subtree-Slide performs worse compared to the branch rearrangement moves. Eventually the Subtree-Slide operator can give good result for a good starting tree. In most runs it was not the case when we took random starting trees.

Summarized we observed that the global moves perform better than the local moves and the tree rearrangement moves better than the branch change moves. Further, changing the node heights simultaneously expands the neighbourhood but worsen the convergence due to the low acceptance of new proposals.

4.3.2 *Phase 2*

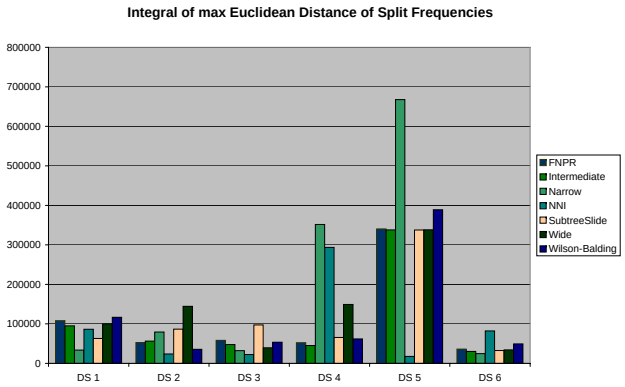

Figure 5: The integral of the maximal difference of split frequencies measured for each of the seven moves on the six different data sets. This plot shows the mean of three runs of a length of 100 million.

In phase two we aimed to validate these assumptions on 6 further data sets (see figure 1). The same framework of tests was performed for this data set as before for the Anolis data set. Surprisingly most of the before observed characteristics can not be affirmed in general (see figure 5). The best move

according to the mean over the maximal differences of split frequencies is the Narrow Exchange. The majority of the results seem to be close together and therefore it is more likely that the Narrow Exchange has converged faster to the golden run than the other moves. Regardless this good attitude the Narrow Exchange has some enormous outliers which show it can get stuck in a different area of the tree space. Since this outlier is exactly 100% different from the golden run, it is a clade which is always present (or absent) in the test run but is never present (or absent) in the golden run.

But we can show in our result that the branch change move namely Subtree-Slide does not perform much worse than all the branch rearrangement moves and has furthermore the ability produce less outliers than the other moves. This is contrary to the main conclusion of [10].

The efficiency of the seven moves varied for the six data sets (see figure 5). In particular, data set 5 shows unexpected performance for all moves. The measurement we have taken for this data set might have been to susceptible. A similar behaviour is observed for data set 4. These data sets are suitable for further studies to observe the efficiency of seven different moves for difficult data sets for clock-constraint trees. These data sets can be studied further to conclude how they differ from the other data sets and what a good approach for these data sets could be. If these two data sets are disregarded there is a strong tendency of moderate tree proposal distributions to dominate the bolder ones. Further, the subtree swapping algorithms perform better than the subtree prune and regraft algorithms.

4.4 Total performance improvement

	Narrow Exchange	Wide Exchange	SubtreeSlide	Wilson-Balding	Intermediate	FNPR	NNI
Current Default Setting	15	3	15	3	-	-	-
New Proposed Setting	10	3	10	3	5	3	10

Figure 6: The weights for the combination of the tree proposal distributions for the current default setting and the new proposed ones.

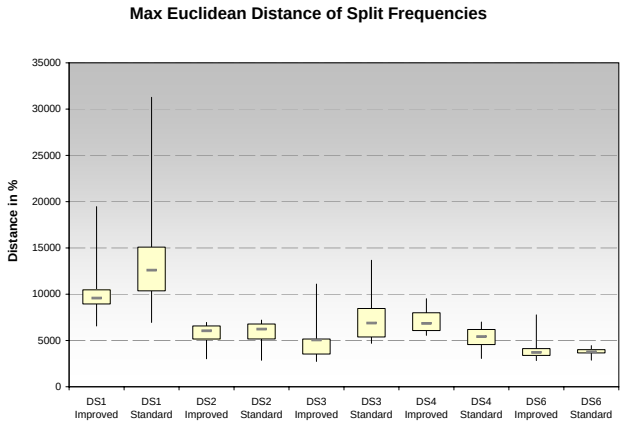

Figure 7: This box-plot compares the standard setting for BEAST to the new proposed settings with the new moves. Each setting was run 10 times for a chain length of 10 million on each data set.

The benefit of this research is improving the efficiency of Bayesian phylogenetic inference using MCMC algorithm. The overall performance of the MCMC algorithm can be achieved by a combination and weighting scheme which performs better than the current default settings in BEAST. To achieve this goal 10 runs with the default settings and 10 runs with the new proposed settings were performed on the 6 data sets. The new proposed settings were obtained as an observation for the test results of the single tree proposal distributions and the GA performance. These weightings are presented in figure 6.

Since data set 5 is expected to mislead clear results we excluded this dataset from the comparison since both settings perform equally worse and the difference is rather an outcome of luck than from better combination of proposal distributions. Finally we observe a performance improvement of 10% for the data sets (shown in figure 7) and 20% if we exclude data set 4 too which is reasonable since it has some unexpected results compared to all the other data sets.

5. CONCLUSION

The preliminary result of our study is that there is no easy metric described in the literature to date for measuring the mixing in tree space. This mixing in tree space is hardly connected to the convergence of multiple runs from random starting trees. The state of the art technique is the deviation of posterior split probabilities as implemented in AWTY [18]. The swap rate we defined in our research could be used to support this hypothesis of convergence. Further the swap rate gives more detailed diagnosis of the performance. It is not dependent on whether the run has converged or not to give a comparable statement.

In our test on the Anolis data set we conclude that the moves without changing the branch length simultaneously perform better. Particularly it means that the Narrow Exchange performs better than the NNI operator and the FNPR better than the Wilson-Balding operator. Hence for clock-constrained phylogenetic trees a combination of moves changing the topology and the branch length in another move is preferable. This conclusion is not confirmed by all other data sets since the results varied much between each of them.

For the data set of the 19 anolis with 950 nucleotides it was obvious that there exists a valley between two groups of good trees. These trees (like Figure 4 shows one example) are hard to pass for all local and prune-and-regraft moves. To prevent getting stuck in these tree valleys the Wide Exchange operator should always be used additionally if other moves are also used.

Evaluating the moves by the average swap rate over all clades can mislead in the conclusion of how fast a move will converge. For example the Narrow Exchange move had a better average swap rate but converged slower compared to the Wide Exchange move. The Narrow Exchange could not converge very fast because it performed very badly in 4 clades of the Anolis test set. This low swap rate for particular clades did not influence much the average for the swap rate. Nevertheless the convergence was determined by the maximal Euclidean distance of the split frequencies. If one clade does not swap very often it is very likely to have a high distance for this split. Hence, fast convergence is just

approachable if non of the clades has a low swap rate. However, it gives a good estimation of how good a tree proposal distribution performs and the results of this metric were continuous. The moderate tree proposal distributions (Narrow Exchange and NNI) performed best and this can be verified for the majority of the convergence analysis of all data sets.

A further improvement for the moves is achieved by mixing between moderate and bold. We could observe this effect for a bias that prefers moderate proposal from a bold move in the Intermediate Exchange. The bias forces the move to prefer local changes over global ones. This combines the strength of both moves as having a low convergence time and being robust for more data sets. We used a simple selection mechanism which weighted the nodes by their path distance in the tree to each other. In future research different selection mechanism (i.e. ones which are used in GA's) could be evaluated as how other biases could affect the performance of the tree proposal moves.

A supplementary study could measure the impact of moves which change from bold to moderate over time. We believe that the more moderate moves are essential for fast convergence because the Narrow Exchange showed the best performance in approximately half of the controlled clades in the anolis data set. But the bolder moves are more robust so that they can overcome the shortcoming of the moderate moves to get stuck. This could be achieved by a single move which has a bias from moderate to bold which can be changed by a parameter. In our studies we could show that usually the Intermediate Exchange has a performance between the Narrow and Wide Exchange but in some cases it performed better than both of them. However, we have not tested yet if one moves which can be adjusted in its bias performs better than two moves which are selected with a bias. Thus in phylogenetic inference software like BEAST several moves are applied they can be easily modified to change their weights to select the moves over time. The selection of the moves and the calculation for the weights of them have the advantage to be computational less expensive since it has not to be calculated as often as the bias inside a move. The question remains if both have the same performance.

6. ACKNOWLEDGEMENT

SH thanks Michael Defoin Platel for helpful discussions on genetic algorithms and Clemens Lakner for providing us with their manuscript and data sets.

7. REFERENCES

[1] M. Bourque. *Arbres de Steiner et reseaux dont varie lŠemplagement de certains sommets*. PhD thesis, Ph. D. diss., Universite de Montreal, Quebec, Canada, 1978.

[2] M. K. Cowles and B. P. Carlin. Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904, ???? 1996.

[3] A. J. Drummond, G. K. Nicholls, A. G. Rodrigo, and W. Solomon. Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. *Genetics*, 161(3):1307–1320, 2002.

[4] A. J. Drummond and A. Rambaut. Beast, a program for bayesian mcmc of evolution & phylogenetics using molecular sequences, 2007.

[5] J. Felsenstein et al. *Inferring phylogenies*. Sinauer Associates Sunderland, Mass., USA, 2004.

[6] V. Gowri-Shankar and M. Rattray. A Reversible Jump Method for Bayesian Phylogenetic Inference with a Nonhomogeneous Substitution Model. *Mol Biol Evol*, 24(6):1286–1299, 2007.

[7] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, apr 1970.

[8] J. Huelsenbeck, F. Ronquist, R. Nielsen, and J. Bollback. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. *Science*, 294(5550):2310, 2001.

[9] M. Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates [published erratum appears in Mol Biol Evol 1995 May;12(3):525]. *Mol Biol Evol*, 11(3):459–468, 1994.

[10] C. Lakner, P. van der Mark, J. P. Huelsenbeck, B. Larget, and F. Ronquist. Efficiency of mcmc tree proposals in bayesian phylogenetics. School of Computational Science, Florida State University, Tallahassee, Florida USA. To be published, 2007.

[11] B. Larget and D. Simon. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol*, 16(6):750–759, 1999.

[12] B. Mau and M. A. Newton. Phylogenetic inference for binary data on dendograms using markov chain monte carlo. *Journal of Computational and Graphical Statistics*, 6(1):122–131, mar 1997.

[13] B. Mau, M. A. Newton, and B. Larget. Bayesian phylogenetic inference via markov chain monte carlo methods. *Biometrics*, 55(1):1–12, mar 1999.

[14] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computer machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[15] E. Mossel and E. Vigoda. Phylogenetic mcmc algorithms are misleading on mixtures of trees. *Science*, 309(5744):2207–2209, 2005.

[16] F. Ronquist and J. P. Huelsenbeck. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, 2003.

[17] F. Ronquist, B. Larget, J. P. Huelsenbeck, J. B. Kadane, D. Simon, and P. van der Mark. Comment on "phylogenetic mcmc algorithms are misleading on mixtures of trees". *Science*, 312(5772):367, 2006.

[18] J. Wilgenbusch, D. Warren, and D. Swofford. AWTY: A system for graphical exploration of MCMC convergence in Bayesian phylogenetic inference, 2007.

[19] I. J. Wilson and D. J. Balding. Genealogical Inference From Microsatellite Data. *Genetics*, 150(1):499–510, 1998.