Coupled MCMC in BEAST 2

Nicola F. Müller[a,b,1] and Remco Bouckaert[c,d,1]

[a]ETH Zurich, Department of Biosystems Science and Engineering, 4058 Basel, Switzerland
[b]Swiss Institute of Bioinformatics (SIB), Switzerland
[c]University of Auckland, New Zealand
[d]Max Planck Institute for the Science of Human History, Germany

[1]Corresponding authors
Contacts: nicola.felix.mueller@gmail.com;r.bouckaert@auckland.ac.nz

Abstract

Coupled MCMC has long been used to speed up phylogenetic analyses and to make use of multi-core CPUs. Coupled MCMC uses a number of heated chains with increased acceptance probabilities that are able to traverse unfavourable intermediate states more easily than non heated chains and can be used to propose new states. While more and more complex models are used to study evolution, one of the main software platforms to do so, BEAST 2, was lacking this functionality. Here, we describe an implementation of the coupled MCMC algorithm for the Bayesian phylogenetics platform BEAST 2. This implementation is able to exploit multiple-core CPUs while working with all models and packages in BEAST 2 that affect the likelihood or the priors and not directly the MCMC machinery. We show that the implemented coupled MCMC approach is exploring the same posterior probability space as regular MCMC when MCMC behaves well. We also show our implementation is able to retrieve more consistent estimates of tree distributions on a dataset where convergence with MCMC is problematic.

Introduction

Phylogenetic method are being increasingly used to study complex population dynamics by using ever larger datasets. These analyses however also require an increasingly large amount of computational resources. Tree likelihood calculations (Suchard and Rambaut, 2009) often assume independent evolutionary processes on different branch and nucleotide site and can be easily parallelised (Suchard and Rambaut, 2009). In contrast to that, it can be very complex or even impossible to for example parallelise tree prior calculations to make use of multi-core CPUs. As a results, Markov chain Monte Carlo (MCMC) runs can be very time consuming, which limits the datasets that can be studied and the complexity of models that can be used to do so. Alternatively, coupled Markov Monte Carlo, also called parallel tempering, Metropolis coupled MCMC, or MC3, can be used in Bayesian phylogenetics Altekar et al. (2004). This approach is based on running multiple MCMC chains, each at a different "temperature", which effectively flattens the posterior probability space. This leads to less favourable moves being accepted more often, and in turn increases the chance to travel between local optimas. After some amount of iterations, two chains are randomly exchanged in what is essentially an MCMC move. In such a move, the parameters of the two chains are exchanged, but each chain keeps its temperatures. While the heated chains do not explore the true posterior probabilities, the one cold chain does.

In BEAST 2 (Bouckaert et al., 2014), where a lot of novel Bayesian phylogenetic model development takes place (Bouckaert et al., 2018), this approach is currently missing. Here, we provide such an implementation of the coupled MCMC algorithm of Altekar et al. (2004) in BEAST 2. This implementation makes use of multiple CPU cores, allowing virtually any analyses in BEAST 2 to be performed on multi-core machines increasing the size of datasets that can be analysed and the complexity of models that can be used to do so.

We first show the correctness of our implementation of the coupled MCMC by comparing summary statistics of multi type tree distributions sampled under the structured coalescent (Vaughan et al., 2014) to the summary statistics received when using regular MCMC. Additionally, we validate that the inference between regular MCMC and our implementation of coupled MCMC match, when applying both to infer the past population dynamics of Hepatitis C in Egypt (Ray et al., 2000; Pybus et al., 2003). We then compare MCMC with coupled MCMC using different levels of heating on two different datasets. First, we apply it to the Hepatitis C dataset, where we do not expect regular MCMC to be stuck in local optimas. Then, we apply it to a dataset which has been described to be easily stuck in local optimas (Lakner et al., 2008; Höhna and Drummond, 2011).

Methods and Material

Background

Coupled MCMC makes use of running n different chains $i = 1, ..., n$ at different temperature (Geyer, 1991; Gilks and Roberts, 1996; Altekar et al., 2004). Each of the different chains works similar to a regular MCMC chain. In regular MCMC, a parameter space is explored as follows: Given that the MCMC is currently at state x, we propose a new state x' from a proposal distribution $g(x'|x)$ given the current state. At this new state,

we calculate the likelihood $P(D|x')$ of the data D given the state and the prior probability of the new state $P(x')$ and compare it the to old state. The acceptance probability of accepting this new state is then calculated as follows:

$$R = \min\Big[1, \frac{P(D|x')P(x')}{P(D|x)P(x)} \frac{g(x|x')}{g(x'|x)}\Big] \tag{1}$$

If R is greater than a randomly drawn value between $[0, 1]$, the new state x' is accepted as the current state, otherwise it is rejected and we remain in the same state. If we keep proposing new states x' and accept these using (1), we eventually explore parameter space with the frequency at which values of a parameter are visited being its marginal probability (Geyer, 1991).

One of the issues of using this approach is that acceptance probabilities can be quite low, which makes it hard to move between different states in parameter space. Alternatively, an MCMC chain can be heated by using a temperature scaler $\beta_i = \frac{1}{1+(i-1)\Delta t}$, with i being the number of the chain (Altekar et al., 2004). Heating of an MCMC chain changes its acceptance probability R_{heated} to:

$$R_{heated} = min\Big[1, \Big(\frac{P(D|x')P(x')}{P(D|x)P(x)}\Big)^{\beta_i} \frac{g(x|x')}{g(x'|x)}\Big]$$

For a heated chain however, the frequency at which a value of a parameter is visited does not correspond to its marginal probability any more. However, heated chains can be used as a proposal to update the non heated chain by using what essentially is an MCMC move. This move proposes to swap the current states of two random chains i and j with the temperature β_i and β_j such that $\beta_i < \beta_j$. Exchanging the states of chains i and j is accepted with an acceptance probability R_{ij} of:

$$R_{ij} = min\Big[1, \frac{P(x_i|D)^{\beta_j} P(x_j|D)^{\beta_i}}{P(x_i|D)^{\beta_i} P(x_j|D)^{\beta_j}}\Big]$$

As for a regular MCMC move, swapping the states of the two chains is accepted when a randomly drawn uniformly distribution value in $[0, 1]$ is smaller than R_{ij}.

Implementation

In our implementation of the coupled MCMC, we run n different MCMC chains, with each chain $i \in [1, \ldots, n]$ running at a temperature $\beta_i = \frac{1}{1+(i-1)\Delta t}$. Chain number 1 is therefore the only cold chain and explores the state space like a regular MCMC chain.

Upon initialisation, we first sample at random at which iteration the states of two chains with which number are proposed to be exchanged. We then initialise each chain to be run in its own Java thread using multiple CPU cores, if available. Each chains is then run for as many iterations until it reaches the next time an exchange of states with another chain is proposed. This means than every chain runs independently of each other until an iteration at which it actually participates in a proposed exchange, minimising the crosstalk between threads Altekar et al. (2004). If the exchange of states between different chains is accepted, we exchange the temperature of the two chains instead of the states themselves. The states can be quite large and exchanging them across different chains is potentially quite time consuming. Alongside exchanging the states, we

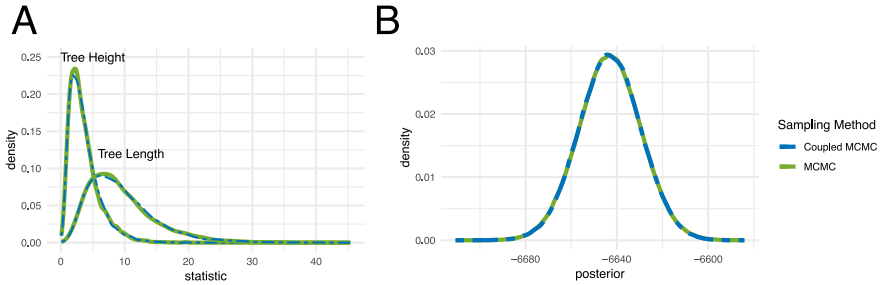

Figure 1: **Comparison of inference between coupled and regular MCMC**. **A** Comparison of the distribution of tree heights and tree lengths sampled under the structured coalescent using MultiTypeTree (Vaughan et al., 2014). The inferred distribution of tree heights and tree lengths match up between MCMC and the cold chains in coupled MCMC. **B** Comparison of the distribution of posterior probability estimates of a Bayesian Coalescent Skyline (Drummond et al., 2005) analysis of Hepatitis C in Egypt (Ray et al., 2000).

exchange the operator specifications and logger. We exchange the operator specifications such that the step size of operators can be optimized to run at specific temperatures. The loggers are exchanged such that each heated chain logs its states to the log file that corresponds to its temperature and not the number of the chain.

We implemented the coupled MCMC algorithm such that finished runs to be resumed. In case that chains did not fully convergence just yet, it is not necessary to restart the analysis scratch, which is of great practical value.

Usually, a graphical user interface called BEAUti is used to set up BEAST 2 analyses. Setting up analyses with coupled MCMC works differently depending on whether a BEAUTi template is needed to set up an analysis as required for some packages. If no such template is needed, an analysis can be set up to run with coupled MCMC directly in BEAUTi and we provide a tutorial on how to do this on `https://taming-the-beast.org/tutorials/CoupledMCMC-Tutorial/` (Barido-Sottani et al., 2017).

Data Availability and Software

The BEAST 2 package coupledMCMC can be downloaded by using the package manager in BEAUti. The source code for the software package can be found here: `https://github.com/nicfel/CoupledMCMC`. The XML files used for the analysis performed here can be found in `https://github.com/nicfel/CoupledMCMC-Material`. All plots were done using ggplot2 (Wickham, 2016) in R (Team et al., 2013).

Validation

Similar to the validation of MCMC operators, we can sample under the prior to validate the implementation of the coupled MCMC approach. To do so, we sampled typed trees with 5 taxa and two different states under the structured coalescent using MultiTypeTree (Vaughan et al., 2014).

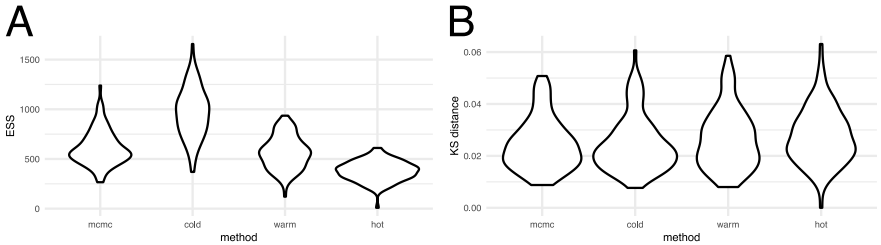

Figure 2: **Convergence of coupled MCMC and regular MCMC using posterior ESS values and Kolmogorov Smirnov distances. A** Here, we show the distribution of posterior ESS values after $4 * 10^7$ for regular MCMC and after $1 * 10^7$ for coupled MCMC with 4 chains. The cold scenario uses coupled MCMC, but does not use any heating. The warm scenario uses slightly heated chains and the hot scenario relatively hotter chains. **B** Here we show the distribution of Kolmogorov Smirnov distances between individual runs and the concatenation of all individual runs. We assume that all 400 runs concatenated describe the true distribution of posterior values and then take the KS distance as a measure of how good an individual run approximates that distribution. The smaller a KS value, the better the true distribution was approximated.

We did this sampling once using regular MCMC and once using coupled MCMC. If the implementation of the coupled MCMC algorithm explores the same parameter space as regular MCMC, parameters sampled using both approaches should match. We ran coupled MCMC proposing to exchange states between chains every 1000 iterations. In figure 1A, we compare the distribution of different summary statistics of typed trees between MCMC and coupled MCMC. For all the summary statistics considered here, the distributions are the same.

Next, we validate that the coupledMCMC package estimates the same parameters in a Bayesian coalescent skyline (Drummond et al., 2005) analysis of Hepatitis C in Egypt (Ray et al., 2000). To do so, we analysed the Hepatitis C dataset once using coupled MCMC with 4 chains and once using regular MCMC. We find that the inferred posterior probability density is the same between the two approaches(see figure 1B).

Results

The effect of heating on exploring the posterior

In order to explore how heating affects exploring the posterior probability space, we first compare effective sample size (ESS) values between regular and coupled MCMC at different temperatures on a dataset where we do not expect any problems in exploring the posterior space caused by several local optimas. To do so, we ran the Bayesian coalescent skyline (Drummond et al., 2005) analysis of Hepatitis C in Egypt (Ray et al., 2000) for $4 * 10^7$ iterations using regular MCMC in 100 replicates. Additionally, we performed 100 replicates using coupled MCMC on 4 different chains for $1 * 10^7$ iterations using 3 different temperature scalers referred to as cold, warm and hot. The different chains lengths are chosen such that the overall number of iterations over the cold and heated chains is the same for coupled as for regular MCMC. In the cold scenario, we did not use any

heating and exchanges between chains were accepted with a probability of about 100%. In the other two scenarios, we used heating such that exchanges between chains were accepted with around 50% in the warm and with about 25% in the hot scenario. After running all 4 times 100 analyses, we computed the ESS values of the posterior probability estimates using loganalyser in BEAST 2 (Bouckaert et al., 2014).

As shown in figure 2, the average ESS values are highest for the cold scenario when using coupled MCMC and drop the stronger the temperature scaler becomes. Regular MCMC gets in average slightly lower ESS values when using 4 times longer chains. The trends of ESS values are the same when calculating ESS values using coda (Plummer et al., 2006) (see figure S1).

In order to assess if coupled MCMC approximates the true distribution of posterior values better than regular MCMC, we compared Kolmogorov-Smirnov (KS) distances between individual runs and the true distribution of posterior values. Since we can not directly calculate the true distribution of posterior values, we concatenated the 400 regular and coupled MCMC runs and used the concatenated distribution of posterior values as the true distribution. Figure 2 shows the distribution of KS distances between individual runs using regular and coupled MCMC to what we assume to be the true distribution. In contrast to the comparison of ESS values, we find that the distribution of KS distances is fairly comparable across all methods. This indicates that in this analysis, coupled MCMC with 4 individual chains performs equally well as regular MCMC run for 4 times as long.

We next compare the inference of trees on a dataset DS1 that has proved problematic for tree inference using MCMC (Lakner et al., 2008; Höhna and Drummond, 2011; Maturana Russel et al., 2018). This dataset has many different tree island, transitioning between which is highly unlikely due to very unfavourable intermediate states (Höhna and Drummond, 2011).

We ran the dataset using regular MCMC for $5 * 10^7$ iteration and coupled MCMC for $5 * 10^7$ with 4 different chains. We ran coupled MCMC without heating (cold) with a maximum temperature of 0.2 (warm) and the maximum temperature being 1.0 (hot). MCMC converges to different optimas, resulting in differences between inferred clade credibilities across different runs (see figure 3). The clade credibilities are more comparable when using multiple chains but no heating (cold). The increased consistency of clade credibilities across runs is in this case due to the main chain essentially being an average over 4 MCMC runs. When using heating (warm and hot), the heated chains are able to more easily cross the unfavourable intermediate states in tree space, resulting in a better consistency of clade credibilities across different runs for the warm scenario and essentially the same clade credibilities across different runs in the hot scenario.

Conclusion

Next generation sequencing has made ever larger datasets of genetic sequence available to researcher. To study these, more and more complex models are developed, many of which are implemented in the Bayesian phylogenetic software platform BEAST 2 (Bouckaert et al., 2014). Parallelising these models can often be hard or even impossible and MCMC

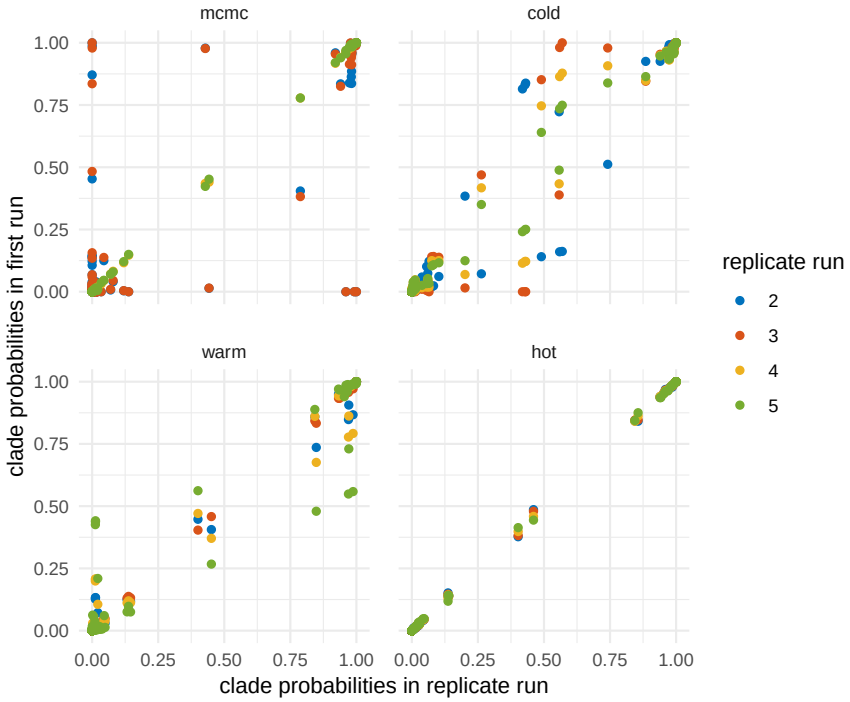

Figure 3: **Inferred clade probabilities between different replicate runs.** Here we compare inferred clade credibilities between one run (y-axis) and four replicates from different starting points (x-axis) using MCMC and coupled MCMC run at different temperature increments.

analyses often have to be run on single CPU cores. Alternatively, coupled MCMC can make use of multiple cores, but a full featured version was so far not available in BEAST 2. Here, we provide an implementation of the coupled MCMC algorithm for BEAST 2.5 (Bouckaert et al., 2018). We showed that this implementation explores the same posterior space as regular MCMC and we give an example for when the heating of chains can drastically improve convergence. While ESS values are higher on coupled MCMC runs with 4 chains and no heated than on regular MCMC runs that are run for 4 times longer, the distribution of posterior probability values was not better approximated by those runs. This indicates that convergence statistics like the scale reduction factor (Brooks and Gelman, 1998), might be better suited to assess convergence than ESS values. Since the coupled MCMC runs required 4 times less iterations of the cold chain to approximate the distribution of posteriors values as well, coupled MCMC can help speed up analysis by a factor that is approximately proportional to the number of CPU's used. This implementation is compatible with other BEAST 2 packages, so works with any model that works with MCMC.

Acknowledgement

NFM is funded by the Swiss National Science foundation (SNF; grant number CR32I3_166258).

Authors contribution

NFM and RB implemented the code, NFM performed the analyses and NFM and RB wrote the paper.

References

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., and Ronquist, F. (2004). Parallel metropolis coupled markov chain monte carlo for bayesian phylogenetic inference. Bioinformatics, **20**(3), 407–415.

Barido-Sottani, J., Bošková, V., Plessis, L. D., Kühnert, D., Magnus, C., Mitov, V., Müller, N. F., PečErska, J., Rasmussen, D. A., Zhang, C., et al. (2017). Taming the beast?a community teaching material resource for beast 2. Systematic biology, **67**(1), 170–174.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. (2014). Beast 2: a software platform for bayesian evolutionary analysis. PLoS computational biology, **10**(4), e1003537.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchene, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kuhnert, D., De Maio, N., et al. (2018). Beast 2.5: An advanced software platform for bayesian evolutionary analysis. BioRxiv, page 474296.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics, **7**(4), 434–455.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular biology and evolution, **22**(5), 1185–1192.

Geyer, C. J. (1991). Markov chain monte carlo maximum likelihood.

Gilks, W. R. and Roberts, G. O. (1996). Strategies for improving mcmc. Markov chain Monte Carlo in practice, **6**, 89–114.

Höhna, S. and Drummond, A. J. (2011). Guided tree topology proposals for bayesian phylogenetic inference. Systematic biology, **61**(1), 1–11.

Lakner, C., Van Der Mark, P., Huelsenbeck, J. P., Larget, B., and Ronquist, F. (2008). Efficiency of markov chain monte carlo tree proposals in bayesian phylogenetics. Systematic biology, **57**(1), 86–103.

Maturana Russel, P., Brewer, B. J., Klaere, S., and Bouckaert, R. R. (2018). Model selection and parameter inference in phylogenetics using nested sampling. Systematic biology, **68**(2), 219–233.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: convergence diagnosis and output analysis for mcmc. R news, **6**(1), 7–11.

Pybus, O., Drummond, A., Nakano, T., Robertson, B., and Rambaut, A. (2003). The epidemiology and iatrogenic transmission of hepatitis c virus in egypt: a bayesian coalescent approach. Molecular biology and evolution, **20**(3), 381–387.

Ray, S. C., Arthur, R. R., Carella, A., Bukh, J., and Thomas, D. L. (2000). Genetic epidemiology of hepatitis c virus throughout egypt. The Journal of infectious diseases, **182**(3), 698–707.

Suchard, M. A. and Rambaut, A. (2009). Many-core algorithms for statistical phylogenetics. Bioinformatics, **25**(11), 1370–1376.

Team, R. C. et al. (2013). R: A language and environment for statistical computing.

Vaughan, T. G., Kühnert, D., Popinga, A., Welch, D., and Drummond, A. J. (2014). Efficient bayesian inference under the structured coalescent. Bioinformatics, **30**(16), 2272–2279.

Wickham, H. (2016). ggplot2: elegant graphics for data analysis. Springer.