Subject Section

MASCOT: Parameter and state inference under the marginal structured coalescent approximation

Nicola F. Müller$^{1,2,*}$, David Rasmussen$^{1,2,3,4}$ and Tanja Stadler$^{1,2,*}$

$^1$ETH Zurich, Department of BioSystems Science and Engineering, 4058 Basel, Switzerland
$^2$Swiss Institute of Bioinformatics (SIB), Switzerland
$^3$Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, North Carolina, USA
$^4$Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA

* To whom correspondence should be addressed.

Abstract

**Motivation:** The structured coalescent is widely applied to study demography within and migration between sub-populations from genetic sequence data. Current methods are either exact but too computationally inefficient to analyse large datasets with many sub-populations, or make strong approximations leading to severe biases in inference. We recently introduced an approximation based on weaker assumptions to the structured coalescent enabling the analysis of larger datasets with many different states. We showed that our approximation provides unbiased migration rate and population size estimates across a wide parameter range.

**Results:** We extend this approach by providing a new algorithm to calculate the probability of the state of internal nodes that includes the information from the full phylogenetic tree. We show that this algorithm is able to increase the probability attributed to the true sub-population of a node. Furthermore we use improved integration techniques, such that our method is now able to analyse larger datasets, including a H3N2 dataset with 433 sequences sampled from 5 different locations.

**Availability:** The presented methods are part of the BEAST2 package MASCOT, the Marginal Approximation of the Structured COalescenT. This package can be downloaded via the BEAUti package manager. The source code is available at https://github.com/nicfel/Mascot.git.

**Contact:** nicola.mueller@bsse.ethz.ch, tanja.stadler@bsse.ethz.ch

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

1 Introduction

Phylogenies contain information regarding the history of a population and can be used to quantify demographic parameters. This has been widely done to study the spread of pathogens (Pybus et al., 2001; Russell et al., 2008), the speciation dynamics of extant species or the migration pattern of humans to name but a few. Forwards in time birth-death and backwards in time coalescent models allow us to elucidate population dynamics from trees by calculating the probability of a phylogeny $T$ given a set of demographic parameters $\Theta$. To do so they classically rely on the assumption of well mixed populations, meaning that the rate at which any two pairs of lineages coalesce is the same. In most empirical applications this assumption of well mixed populations is however violated.

To address this model violation, so-called structured methods have been developed that consider birth-death processes in heterogeneous populations (Stadler and Bonhoeffer, 2013). In the backward-in-time coalescent framework, the structured coalescent (Takahata, 1988; Hudson, 1990; Notohara, 1990) describes a coalescent process in sub-populations between which individuals can migrate. Such coalescent methods however typically require the state (or location) of any ancestral lineage in the phylogeny at any time to be inferred (Beerli and Felsenstein, 2001; Ewing et al., 2004; Vaughan et al., 2014). Inferring lineage states is computationally expensive, as it normally requires MCMC based sampling, and limits the complexity of scenarios that can be analysed. As the number of different states is increased, convergence of the MCMC chains becomes a severe issue for inference under the structured coalescent when the sampling of migration histories is needed (Vaughan et al., 2014;
De Maio et al. 2015). This essentially limits the number of different states that can be accounted for to three or four.

We addressed this limitation recently by introducing a new approximation of the structured coalescent that avoids this MCMC sampling of lineage states by integrating over all possible migration histories using a set of ordinary differential equations (Müller et al., 2017). In contrast to previous approximations that treat the movement of one lineage completely independently of all other lineages (Volz, 2012; De Maio et al., 2015), we explicitly include information about the location of other lineages and their probability of coalescing when modelling the movement of a lineage. This avoids the strong biases resulting from this independence assumption (Müller et al., 2017). We showed that this approximation is able to infer coalescent and migration rates well in various scenarios. However, this approach currently lacks the possibility to estimate the ancestral state of any internal nodes except the root.

Here, we introduce a new algorithm to calculate the probability of internal nodes being in any state that incorporates information from the entire tree using a forwards/backwards approach (Pearl, 1982). We additionally make improvements of the current BEAST2 (Bouckaert et al., 2014) implementation of Müller et al. (2017) in terms of calculation speed, allowing us to analyse datasets with more states and lineages. We do so, we need to make the following approximation:

\[
P_i(L_i = l_i, L_j = l_j) = P_i(L_i = l_i | T) P_i(L_j = l_j | T) P_i(L_k = l_k | T)
\]

In other words, we assume that lineages \( i, j \) and \( k \) and their states \( l_i, l_j \) and \( l_k \) are pairwise independent.

2.2 The probability of a lineage being in a state

As described in Müller et al. (2017), we seek to calculate the probability of every lineage being in any state jointly with the probability of having observed the coalescent history \( T \) up to time \( t \). We previously denoted this probability as \( P_i(L_i = l_i, T) \). Calculating these terms over time for increasing \( t \) leads to ever smaller values, eventually causing numerical issues. To avoid this, we can calculate \( P_i(L_i = l_i, T) = P_i(L_i = l_i | T) P_i(T) \) instead. The expression for \( P_i(L_i = l_i | T) \) can be directly derived from \( \frac{d}{dt} P_i(L_i = l_i, T) \) (see Supplement) and can be written as:

\[
\frac{d}{dt} P_i(L_i = l_i | T) = \sum_{a \in \mathbb{A}} \left( \mu_{a,i} P_i(L_i = a | T) - \mu_{i,a} P_i(L_i = l_i | T) \right) + P_i(L_i = l_i | T) \sum_{a \in \mathbb{A}} \lambda_a P_i(L_i = a | T) \sum_{k \neq i, j} P_i(L_k = a | T) - P_i(L_i = l_i | T) \lambda_i \sum_{k \neq i, j} P_i(L_k = l_k | T)
\]

where \( \mu_{a,i} \) denotes the backwards in time rate at which lineages migrate from state \( a \) to state \( l_i \) and \( \lambda_i \) denotes the rate of coalescence in state \( a \).

To calculate \( P_i(T) \), i.e. the probability of having observed the coalescent history \( T \) up to time \( t \), the following differential equation has to be solved (see supplement for derivation):

\[
\frac{d}{dt} P_i(T) = -P_i(T) \sum_{a \in \mathbb{A}} \sum_{j \neq a, i} \lambda_a P_i(L_i = a | T) P_i(L_j = a | T)
\]

At coalescent events between lineages \( i \) and \( j \), we update \( P_i(T) \) by multiplication with the probability of the coalescent event:

\[
P^\text{after}(T) = P^\text{before}(T) \sum_{a \in \mathbb{A}} \lambda_a P_i(L_i = a | T) P_i(L_j = a | T)
\]

Integrating these equations from the present to the root of a phylogeny, allows us to calculate \( P_i(T) \), that is the probability density of a phylogeny \( T \) under the MASCO approximations of the structured coalescent.

2.3 The probability of a node being in a state given the whole phylogeny

2.3.1 Backwards calculation of node states conditional on the sub-trees

Integrating equation 1 allows us to calculate the probability of each lineage being in any state given the coalescent history between the lineage and the
present. However, for applications, it is much more interesting to calculate the probability of each lineage at time \( t \) being in any state given the whole phylogenetic tree between the time of root and the present. At coalescent events between lineage \( i \) and \( j \) at time \( t \), the probability of the parent lineage \( \pi(t) \) in state \( \alpha \) can be calculated as follows:

\[
P^{bw}_{\lambda}(L_\alpha | \pi(T)) = \frac{\sum_{\ell=1}^{\lambda_i} \lambda_i P^{bw}_{\ell}(L_\alpha | \pi(T)) P^{bw}_{\ell}(L_\alpha | \pi(T))}{\sum_{\ell=1}^{\lambda_i} P^{bw}_{\ell}(L_\alpha | \pi(T))},
\]

with \( P^{bw}_{\lambda}(L_\alpha | \pi(T)) \) denoting the probability of the daughter lineage \( i \) being in state \( \alpha \) just before the coalescent event at time \( t \) calculated in the backwards step using equation 1. Since \( P^{bw}_{\lambda}(L_\alpha | \pi(T)) \) denotes the probability of the parent node of lineages \( i \) and \( j \) calculated in the backwards step at time \( t \), it includes only information up to the time of coalescence and does not include information from the full phylogeny. We introduce the label \( bw \) to differentiate from the forward probabilities introduced below. To additionally incorporate information from the phylogeny between the time of the root and time \( t \) of the coalescent event, one has to deploy a “backwards/forwards” approach that is related to Pearl (1982).

For convenience, we now change to vector notation. We define \( P^{bw}_{\lambda}(t) \) as the vector of the parent lineage \( \pi \) with entries \( P^{bw}_{\lambda}(L_\alpha | \pi(T)) \) in position \( \alpha \) that includes information from time 0 up to time \( t \). \( P^{bw}_{\lambda}(t) \) is the vector with entries \( P^{bw}_{\lambda}(L_i | \pi(T)) \) for every lineage \( i \).

### 2.3 Calculation of transition probabilities

Going through the tree backwards in time, we also seek to calculate the probability that, given lineage \( i \) was in state \( \alpha \) at the last coalescent event \( t \), where lineage \( i \) was the parent and given the tree \( T \), the lineage \( i \) is in state \( \alpha \) at the time of the next coalescent event \( nc \). We denote this probability as \( P_{\lambda\alpha}(L_i | \beta(L_{\alpha} | \pi(T))) \). The matrix \( M_t \) with entries \( P_{\lambda\alpha}(L_i | \beta(L_{\alpha} | \pi(T))) \) in positions \((a,b)\) now denotes the matrix for which the following equation holds:

\[
P^{bw}(t = nc) = P^{bw}(t = lc) \cdot M_t,
\]

with \( P^{bw}(t = nc) \) being the vector with the state probabilities of lineage \( i \) just before the next coalescent event \( nc \). To calculate the entries of the matrix \( M_t \), we solve the following differential equation between the two coalescent events \( t \) and \( nc \): involving lineage \( i \).

\[
\frac{d}{dt}(P_{\lambda\alpha}(L_i = b | L_{\alpha}(t = \infty) = a)) = \sum_{c \in \mathcal{C}} \lambda_i P_c(L_i = c | L_{\alpha}(t = \infty) = a) \cdot \left( \begin{array}{c} \text{in case of } \delta \rightarrow 0 \end{array} \right)
\]

\[
= \mu_{\alpha\beta} P_{\lambda\beta}(L_i = b | L_{\alpha}(t = \infty) = a) + P_{\lambda\alpha}(L_i = b | L_{\alpha}(t = \infty) = a) \cdot \sum_{c \in \mathcal{C}} \lambda_i P_c(L_i = c | L_{\alpha}(t = \infty) = a)
\]

\[
+ \sum_{k \in \mathcal{K}} \lambda_k P_{\lambda\alpha}(L_i = b | L_{\alpha}(t = \infty) = a) \cdot P_{\lambda\alpha}(L_k = c | T) - P_{\lambda\alpha}(L_i = b | L_{\alpha}(t = \infty) = a) \cdot \lambda_i \cdot \sum_{k \in \mathcal{K}} \lambda_k P_{\lambda\alpha}(L_k = b | T),
\]

The entries in positions \((a,b)\) of the matrix \( M_t \) are then the solution of the above differential equation at time \( nc \) with initial values \( 1 \) if \( b = a \) and 0 otherwise. In other words, equation 6 describes the same equation as 1 with lineage \( i \) starting in \( a \) (rather than in any state as in equation 1), assuming that all other lineages, other than \( i \), evolve according to equation 1.

### 2.3.3 Forwards calculation of node states including all information in the phylogeny

Based on section 2.2, we now know the probabilities of every internal node being in any state. Based on Section 2.3.2, we know how these probabilities change between coalescent events. Going backwards, we calculate \( P^{bw}_{\lambda}(L_i = a | \pi(T)) \), which only includes information up to time \( t \). At the root however \( P^{bw}_{\lambda}(L_{root} = a | \pi(T)) \) includes information from the full phylogenetic tree from time 0 up to the time of the root. We hence write \( P^{bw}_{\lambda}(L_{root} = a | \pi(T)) = P^{bw}_{\lambda}(L_{root} = a | \pi(T)) \), that is the forwards probability of the root being in any state. The forwards probabilities denote the probability of a lineage being in a state that includes information from the full phylogenetic tree. We use the forwards probability \( P^{bw}_{\lambda}(L_{root} = a | \pi(T)) \) at the root as a starting point to calculate \( P^{bw}_{\lambda}(L_i = a | \pi(T)) \) for every lineage \( i \). From the root, we proceed backwards in the tree to calculate \( P^{bw}_{\lambda}(L_i = a | \pi(T)) \) for every internal node at the time \( t \) of the coalescent event for which lineage \( i \) was the parent lineage. \( P^{bw}_{\lambda}(L_i = a | \pi(T)) \) could be calculated at other times as well, but here we focus on the state of nodes. This we do as follows:

\[
P^{bw}_{\lambda}(t = lc) = \frac{P^{bw}_{\lambda}(t = nc) \cdot M(T) \cdot P^{bw}_{\lambda}(t = lc)}{||P^{bw}_{\lambda}(t = nc) ||_1 \cdot ||P^{bw}_{\lambda}(t = lc) ||_1}
\]

2.4 Integration of the differential equations

To integrate equation 1, we use a second order Taylor method with third order step size estimation. This integration technique is similar to the very basic Euler integration, but makes use of the second derivative as well:

\[
y_{t+1} = y_t + y_t \Delta t + \frac{1}{2} y_t \Delta t^2 + O(h)
\]

\[
O(h) \text{ stands for derivatives higher than second order. The error that is made by only considering the first and second derivative can be calculated as follows:}
\]

\[
y_{t+1 \text{ est}} = y_{t+1 \text{ true}} = O(h) = \epsilon
\]

With \( y_{t+1 \text{ est}} \) being the updated term using equation 8 and \( y_{t+1 \text{ true}} \) being the hypothetical true value if all derivatives would be considered. We now assume that the Taylor term of every derivative higher than the third are zero. The error \( \epsilon \) we introduce at every step can therefore be approximated as:

\[
\epsilon \approx \frac{1}{6} y_t \Delta t^3
\]

For every integration step, we now choose \( \Delta t \) such that the absolute value of \( \epsilon \) is smaller than a specified value. While we calculate the second derivative exactly (see supplement), we approximate the third derivative, assuming that the sum of probability mass in each state and that the sum of the derivatives of lineage \( i \) coalescing in any state is constant (see
We then use this to update equation 1 between sampling and coalescent events. At the root of the tree, the probability of the tree under the approximate structured coalescent is then calculated by solving equation 3 at the root. This is covered in more depth in Müller et al. (2017).

2.5 Software
The method above is implemented into our BEAST 2 package MASCOT (Marginal Approximation of the Structured CrossCoalescent) and the analyses were done using version 1.0.0 and BEAST v2.5.0. (Bouckaert et al., 2014) Simulations were performed using a backwards in time stochastic simulation algorithm of the structured coalescent process using MASTER 5.0.2 (Vaughan and Drummond, 2013) and BEAST 2.4.7. MultiTypeTree analyses were performed using version 6.3.1 (Vaughan et al., 2014) and BEAST 2.4.7. Script generation and post-processing were performed in Matlab R2015b. Plotting was done in R 3.2.3 using ggplot2 (Wickham, 2009) and igraph 1.1.2 (Csardi and Nepusz, 2005). Tree plotting and tree height analyses were done using ape 3.4 (Paradis et al., 2004) and phytools 0.5-10 (Revell, 2012). Effective sample sizes (ESS) for MCMC runs were calculated for the posterior probability after a burn-in of 10% using coda 0.18-1 (Plummer et al., 2006). Parameter and state inference of the simulated data were only used if the posterior had an ESS of at least 100 and discarded elsewise.

2.6 Data availability
The source code for MASCOT is available at https://github.com/nicfel/Mascot.git. All scripts for performing the simulations and analyses presented in this paper are available at https://github.com/nicfel/Mascot-Material.git, including the MASCOT xml file of the H3N2 analysis. Output files from these analyses, which are not on the github folder, are available upon request from the authors. A tutorial is available through the Taming the BEAST project (Barido-Sottani et al., 2017) on how to use MASCOT and its BEAuTi interface is available at https://github.com/nicfel/Mascot-Tutorial.git.

3 Results
3.1 Inference of migration rates, effective population sizes, and internal node states
First, we tested how well effective population sizes and migration rates are inferred using MASCOT: We simulated 1000 trees with MASTER (Vaughan and Drummond, 2013) using randomly sampled effective population sizes from LogNormal Distribution ($\mu=0.125, \sigma=0.5$) and migration rates from an exponential distribution with mean=$0.5$. We used 1000 tips and 6 different states. In order to have scenarios of under- and over-sampling of states, we randomly sampled the number of tips in each state. The number of samples per state was randomly drawn to be a value between 20 and 1000 in increments of 20, conditional on the overall number of samples in each state being exactly 1000. We then inferred the effective population size of every state and the migration rates between states, using MASCOT from fixed phylogenies. The results of these simulations are summarized in figure 2.

Both effective population sizes and migration rates are inferred well. Population size estimates are however much more precise than estimates of migration rates, see figure 2. This is expected since there are typically many fewer migration events in a phylogeny than coalescent events. Additionally, the number of migration rate parameters estimated (30) is much larger than the number of effective population size parameters (6). The estimates are well correlated with the truth, only at lower migration rates do estimates become worse. This is also to be expected since a low migration rate automatically means less events which will put the estimates closer to the prior (exponential with mean 1). The coverage is 95% for both migration rate estimates and effective population size estimates. We further inferred the state of each internal node with and without the backwards/forwards algorithm. The backwards/forwards algorithm reduces the probability mass that is attributed to the wrong node states in this scenario (figure 2C).

To estimate how the CPU time varies with the number of lineages and states, we used the same framework but with varying numbers of states and lineages. The CPU time per million samples depends approximately linearly on the number of sampled sequences. With an increasing number of states, the calculation time per million sample increases approximately quadratically.

We next compared the convergence properties of MASCOT with MultiTypeTree version 6.3.1 (Vaughan et al., 2014), which is based on the exact structured coalescent model without approximation, but requires MCMC sampling of lineage states. To do so, we compared the ESS per hour to MultiTypeTree (Vaughan et al., 2014) using fixed trees. MASCOT shows much higher ESS values per hour in each scenario, demonstrating the drastically improved convergence properties originating from not having to sample migration histories. While these estimates can vary based on the parameters under which simulations were performed and based on the MCMC operator setup used, they show the benefits of integrating over migration histories.

3.2 Application to H3N2
We then applied MASCOT to 433 Influenza A/H3N2 sequences sampled between 2000 and 2003 from Australia, Hong Kong, New York, New Zealand and Japan. We ran 3 independent chains each for 120 hours using an HKY+$\Gamma_4$ site model with a fixed clock rate of $5 \times 10^{-3}$ substitutions per site and year. We fixed the clock rate due to a lack of temporal information from the sequences collected for this short amount of time. We then inferred the phylogenetic tree as well as the effective population sizes of every location, the migration rates between them, as well as the additional parameters from the HKY+$\Gamma_4$ model.

Figure 3 shows the maximum clade credibility tree with the different colors indicating the maximum posterior location estimate of each node. The pie charts indicate the probability of the marked nodes being in any possible location inferred with and without the backwards/forwards algorithm. These probabilities are the average over all the node state probabilities for each tree in the posterior containing that clade. We inferred New York to be a source location mainly for strains in Australia and New Zealand. Strains from Japan were inferred to originate mainly from Hong Kong and New York. The root of the phylogeny was inferred to be most likely in New York. The lack of samples near the root however makes the inference of its location unreliable.

4 Discussion
We provide a new algorithm to calculate the state of any node in a phylogeny under the marginal approximation of the structured coalescent (Müller et al., 2017). This algorithm entirely avoids the sampling of migration histories. Additionally, we improve the calculation time of our previously introduced approximation to allow for the analysis of phylogenies with more samples and more states.

We have shown on simulated data that our approach is able to infer migration rates and effective population sizes reliably even when many different states (6) are present. This is a case where exact methods that sample migration histories are currently not able to reach convergence. Even though MASCOT is an approximation, we reach a coverage of 95% for migration rates and effective population size estimates.

We also showed on simulated data that adding a backwards/forwards approach for the calculation of node states improves the inference of
Fig. 2. Inference of effective population sizes, migration rates and node states. A Inferred effective population sizes on the y-axis vs. true effective population sizes on the x-axis. The effective population sizes for the tree simulations were sampled from a lognormal (-0.125, 0.5) distribution. The coverage of migration rate estimates was 95.5% and for effective population size estimates 94.9%. B Inferred migration rates on the y-axis vs. true migration rates on the x-axis. The migration rates between states were sampled from an exponential distribution with mean=0.5. C Inferred node states using MASCOT with and without the backwards/forwards algorithm. D Median CPU time per mega sample depending on the number of lineages and the number of different states. The CPU time was taken from 100 replicates of the simulation scenario used. E Median posterior ESS per hour from 100 replicates from MASCOT and MultiTypeTree (dashed lines). The different colors indicate the different number of states. Dashed lines show median ESS per hour values for MultiTypeTree.

Fig. 3. MASCOT analysis of globally sampled Influenza A/H3N2 viruses. A Here we show the maximum clade credibility tree inferred from H3N2 sequences from Australia, Hong Kong, New York, New Zealand and Japan. The colour of each branch indicates the most likely state of its daughter node. The pie charts indicate the probability of chosen nodes being in any of the possible states. The left pie chart is the probability inferred using the backwards/forwards algorithm and the right pie chart without using the backwards/forwards algorithm. Since at the root, these probabilities are the same, only one chart is shown. The node heights are the median node heights. B The median inferred immigration rates as indicated by the width of the arrow. The wider an arrow into a location, the more likely it is that a lineage in the destination originated from the source location of that arrow. The different source locations are denoted by the colours of the arrows. A wider arrow from light blue to green than from red to green shows that lineages in New Zealand are more likely to have originated from New York than from Hong Kong. The dot sizes are proportional to the median inferred effective population sizes of that state.

We use the backwards/forwards to calculate the state of every internal nodes in a way that is consistent with the complete phylogeny, which is not given by the backwards step alone. To estimate the probability of a node being in any possible state given a set of parameters we therefore do not need to average over many MCMC samples of migration histories. Whereas for some nodes the difference between with and without backwards/forwards is small, it is especially large for nodes where the difference in where the node is inferred to be compared to the parent node is large. This is due to conflicting information of the state of a node from the backwards and forwards step. This can for example be the result of a sampling event that adds information of where a lineage was further in the past. Future extensions could include an explicit sampling of migration histories or of the number of state changes. To do so, an algorithm similar to that of (Minin and Suchard, 2008a,b) could be deployed, but lineage states...
would need to be sampled in a way that is probabilistically consistent with our equations for the forward line state probabilities. Finally, we applied MASCOT to a globally sampled H3N2 dataset where we inferred the phylogenetic tree and associated parameters. Our approach is able to reach convergence, even when a large number of sequences and different locations is present. The calculation time still causes challenges in the analysis of very large datasets. These could be circumvented by a further approximation of $\sum_{k=1}^{\infty} k \approx \sum_{k=1}^{\infty} 1$ in equation 1 when many lineages are present. This would allow every lineage to have the same transition probabilities and would therefore reduce the number of ODEs that have to be solved.

MASCOT still requires all migration rates and effective population sizes to be inferred. Especially the number of migration rates ($\text{states} + (\text{states} - 1)$) can become problematic relatively fast. Future additions could however reduce the parameter space by for example deploying Bayesian Search Variable Selection (Lemey et al., 2009) or by making use of generalized linear models (Lemey et al., 2014) to describe migration rates as a combination of different covariates and hence only require the parameters of the GLM model to be inferred.

Acknowledgements

We would like to thank Remco Bouckaert and Tim Vaughan for helpful comments on the implementation of MASCOT. We would also like to thank three anonymous reviewers for the helpful comments on the manuscript. NM and TS are supported in part by the Swiss National Science foundation (SNF; grant number CR32I3_166258). DR is funded by the ETH Zürich Commission (PhyPD: grant agreement number 335529). DM is supported in part by the European Research Council under the Seventh Framework Programme of the European Commission (PhyPD: grant agreement number 335529).

References