Agemo: an open-source library for Laplace transformed coalescence time distributions.

Gertjan Bisschop *
Institute of Evolutionary Biology, University of Edinburgh, EH9 3FL Edinburgh, Scotland

Abstract

`agemo` is an open-source tool with a python API that allows users to generate the Laplace transform of the coalescence time distribution of a sample under any structured coalescent model. In addition, `agemo` provides ways to efficiently query that distribution, by using the fact that its generating function can be represented most simply as a directed graph with all possible ancestral states of the sample as nodes. Past implementations have not made full use of this, relying on computer algebra systems instead of graph traversal to process these recursive expressions. `agemo` can be used to compute the probabilities of the joint site frequency spectrum for blocks of a given size, under the specified model. Calculating these probabilities requires repeated differentiation of the generating function which suffers from an explosion in the number of terms when implemented naively. Using a closed-form expression for the coefficients of a series expansion of the equations associated with each edge, we can efficiently propagate these coefficients through the graph avoiding redundant operations.

1 Introduction

Laplace transforms have been used extensively in coalescent theory since the introduction of the structured coalescent (Takahata, 1988; Notohara, 1990; Barton and Wilson, 1995). Initially it was a

*g.bisschop@sms.ed.ac.uk

formal tool to analyse continuous-time Markov chains (Takahata, 1988; Griffiths, 1991; Wilkinson-Herbots, 1998). But as noted by Wilkinson-Herbots (1998) the Laplace transforms of the distributions of coalescence times are also interesting in their own right. The Laplace transform of a random variable has a clear probabilistic interpretation. One can think of that random variable as describing the length of an interval. If a Poisson marking process with intensity ω marks the interval, then the Laplace transform $f^*(\omega)$ is the probability of not observing any marks in the considered interval (Råde, 1972). Let the random variable be the stochastic process modelling the (sum of the) exponentially distributed waiting times describing the ancestry of a random sample, then the marking process can describe the arrival of mutations. Formulated as such, the connection to the coalescent (Kingman, 1982; Hudson, 1983; Tajima, 1983) becomes obvious, making the Laplace transform a natural choice (Weissman and Hallatschek, 2017).

Moreover, used in mathematical fields like queueing theory, the Laplace transform often provides us with an alternative description of the behaviour of a system, often simplifying its analysis in the process. In essence, this is what the generating function (GF) method as described in Lohse et al. (2011) does. Using the probabilistic interpretation of the Laplace and the fact that the transform turns convolution into multiplication (see 2.1), one can come up with a much simpler, recursive and automated, description of the distribution of the (sum of) interevent times given a set of samples and a (structured coalescent) model. Phrased differently, being given a flow diagram or coalescent state space graph describing all possible transitions during the coalescent process (see fig. 2.1.2), one can readily write down all associated expressions in the Laplace domain. This is not the case for the time domain.

This framework has been used to develop likelihood-based inference methods both for IM-type demographic histories (Lohse et al., 2011; Bunnefeld et al., 2015; Lohse et al., 2016) as well as estimating sweep parameters (Bisschop et al., 2021). Given that the GF relies on generating the coalescent state space graph, and that this state space graph will grow superexponentially with sample size, there are limitations on the sample size and number of events for which the GF can be generated and more importantly evaluated within reasonable (memory) space and time. These two bottlenecks, the ability to make mathematically tractable descriptions of a problem as well the ability to efficiently implement the solution hold for all inference problems. The first issue is generally solved by making simplifying assumptions about recombination, either ignoring the information contained within closely linked variants completely (Gutenkunst et al., 2009; Excoffier et al., 2013), or by only ignoring recombination within very short blocks (Yang, 2002; Hey, 2004). Additionally, one can limit the description to a small number of lineages or only consider very basic underlying (demographic) models.

The second issue is often very closely connected to the type of mathematical description used.

In general, one will benefit from the potential to implement a particular solution in terms of data structures for which efficient algorithms exist for each of the required basic operations. For example, preserving matrix structure up to the point of evaluation, as used in phase-type theory (Hobolth et al., 2019), will tend to have a positive impact on performance due to the existence of many efficient algorithms for linear algebra operations.

So far, all inference approaches using the GF, rely on deriving the probability of observing each of the different block-wise site frequency spectrum (bSFS) vectors. These combinations of mutation configurations along the distinguished branch types within blocks of a certain length summarise the joint distribution of linked polymorphisms (Bunnefeld et al., 2015). However, depending on both the number of observed mutations and the number of distinguished branch types, they require repeated differentiation of the entire GF. As a consequence, any naive implementation of the GF with higher-order derivatives using a computer algebra system (CAS) will quickly run into computational bottlenecks, essentially limiting the extendability of the framework.

Previous efforts to make the framework more efficient have targeted the first bottleneck category by limiting the number of distinguishable branchtypes, limiting the number of occurrences of particular events, and using symmetries within the recursive description (Lohse et al., 2016). No attempts have been made to tackle the second bottleneck and describe the problem in a way that allows for more efficient computation.

Here we present the first open-source implementation of the GF framework and lay out the key algorithms that allow us to efficiently calculate the bSFS by using the correspondence between the recursive formulation of the Laplace transform and the coalescent state space graph. The focus on the bSFS is only superficial however. Any future extension of the library will make use of the concepts outlined here.

First, we will outline how the GF can be represented more succinctly. Secondly, we will lay out the graph traversal algorithm allowing us to efficiently evaluate the GF. We will then apply this to the bSFS by using the algorithm to propagate the coefficients of a truncated Taylor series through the graph.

2 Methods

2.1 Recursive description of the Laplace transform

2.1.1 single population

Given a sample of n uniquely labelled lineages $\Omega = \{a, ..., n\}$ coming from a single population, we can represent all possible coalescent histories of that sample as a series of trees. In each of these ranked topologies nodes represent the lineage ancestral to the subtended nodes. Alternatively, the same information can be captured by a single rooted directed graph describing the state space of all (coalescence) events affecting the history of the sampled lineages. In this graph, each node is uniquely labelled by one of the partitions of Ω. As lineages coalesce, we move through the graph from the source node, representing the set of all sampled lineages, to the root or most recent common ancestor (mrca). In the case of the neutral coalescent and a single population each path through the graph is equivalent to a single ranked topology. However, in general the state space graph will contain information on all events that affect the lineage configuration, not only those that join to lineages.

For k uncoalesced lineages, the rate of coalescence is $\binom{k}{2}$ (in units of $2N_e$ generations). Because each step is conditionally independent of the previous and as touched upon in the introduction, the Laplace transform of the sum of inter-event times or the time to the most recent common ancestor (t_{mrca}) is a simple product of all Laplace transforms of the random variable describing the waiting time to go from k to $k-1$ lineages. As such, we can associate each edge of the directed graph with a single equation describing this transition. The Laplace transform of the t_{mrca} for a single (ranked) topology can then be retrieved by multiplying the equations associated with a single path of the graph. Likewise, the entire GF is the sum of all expressions retrieved by a depth-first traversal of the graph (see 2.1.2).

Knowing that the equation associated with a single edge in the coalescent state space graph as outlined above, gives the probability of observing the coalescence event prior to any other event happening at rate ω, for any two out of k lineages, the Laplace transform is $f^*(\omega) = \frac{1}{\binom{k}{2}+\omega}$. Note that in general we can label each branch connecting two lineages in a coalescent tree by the samples that will be affected by a mutation arising during the time represented by that branch. And in general, we will associated each of these branch types with a unique dummy variable. Replacing ω by $\sum_{i=1}^{b} \omega_i$, we can retrieve the coalescence time distribution for each of the b distinguished branch types (Lohse et al., 2011).

2.1.2 general form

So far, we have dealt with a single population, tracking only coalescence events. In a structured population, we require at least one type of (migration) event to ensure lineages eventually all end up in the same population to allow for coalescence. These events can be incorporated in the state space graph by adding a node for each of these events and labelling that node by their impact on the tracked set of lineages. Looking at any node in the graph, the set of all possible events is determined by all outgoing edges and the set of all observable branchtypes during the time represented by each edge will be given by the label of the node from which the edges are leaving. The LT associated with a single edge still describes the probability of observing a particular event prior to all other events. So as long as the events that we have added in are associated with exponentially distributed waiting times and given that $min(X, Y) \sim exp(\omega_1 + \omega_2)$ when both $X \sim exp(\omega_1)$ and $Y \sim exp(\omega_2)$, we can still readily associate each edge with the correct equation.

$$f[\boldsymbol{\omega}] = \frac{c}{\sum c_i \kappa_i + \sum l_j \lambda_j + \sum o_k \omega_k} \tag{1}$$

Here, Roman letters represent integers counting the number of ways a certain (coalescence or other) event can take place (c_i, l_j) given the current state space, or the number of branches of a particular type (o_i). The Greek letters represent the rate of the associated competing processes. Note that in the case of coalescence with multiple populations, rates are relative and given by $\kappa_i = N_{e_i}/N_{e_{ref}}$.

Each equation can be encoded as a matrix with two rows containing the integer coefficients (c, c_i, l_j, o_k). The first row represents the numerator and will only contain a single non-zero value. The second row then represents the denominator. Storing the equations this way ensures we can efficiently substitute in parameter values by taking the dot product with a vector representing a point in parameter space ($\kappa_i, \lambda_j, \omega_k$) once the Laplace transform needs to be evaluated. Also, storing equation coefficients in array form allows us to efficiently perform operations on the equations (see 2.1.4). A minimal representation of the GF therefore consists of an array containing all unique equations, and an array of arrays with equation indices describing all paths through the graph.

2.1.3 adding in events

In a structured population, we require at least one type of migration event to ensure lineages eventually all end up in the same population to allow for coalescence. Currently, two types of migration events have been implemented: uni-directional migration and mass migration. Other events can readily be incorporated by adding in a method that provides a description of what the event does to the configuration of lineages. In the state space graph, nodes are labelled as partitions

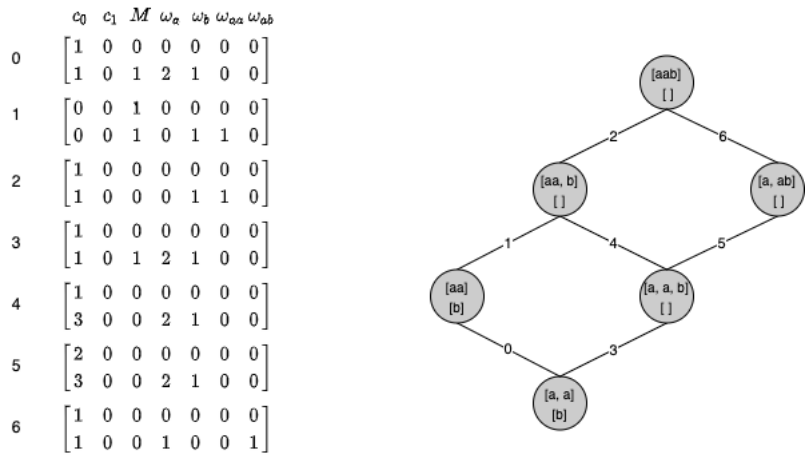

Figure 1: **From coalescent state space to equation array:** coalescent state space graph for two populations A and B with 2 and 1 unphased sample(s) respectively. The demographic model assumes a single mass migration event from population B to A back in time.

of the samples present in each of the subpopulations. One could extent this description and therefore allow for events, like mutations, that do not affect the configuration of lineages. However, there are more efficient ways of dealing with mutations (see 2.3)

Note that to limit state space size and in line with Lohse et al. (2016) migration is indeed limited to being uni-directional. Also, although both migration events are implemented as continuous exponential processes, mass migration is obviously a discrete event.

2.1.4 discrete events

Initially treating discrete events as a competing exponential process we recover the GF parametrised by the discrete time T of the event, by taking the inverse transform of the GF divided by its associated dummy variable (δ). This has been used to incorporate mass migration and bottlenecks (Lohse et al., 2011; Bunnefeld et al., 2015) as well as hard sweeps (Bisschop et al., 2021).

Previous implementations have relied on a CAS to get an analytic solution to the inverse transform. However, we can use the fact that when it comes to inverting with respect to δ each path in the graph corresponds to a product of rational functions of the form $c/(d+\delta)$ with c, d constants. And, limiting the GF to a single discrete event, all equations associated with paths departing from the node representing that event are constants with respect to delta.

Having stored all equation coefficients as an array, we can formulate a closed-form solution to

the inverse Laplace that allows for efficient substitution of all parameter values. Note, this means that the inverse transform does not happen until the GF is being evaluated. However, one could recover the analytical expression by providing symbolic variables to the parameter vector. This would require using a CAS like `SageMath`.

The result is a sum of terms of the form

$$C(\boldsymbol{\omega}) \sum (-1)^i \frac{e^{-P_i(\boldsymbol{\omega})T}}{\prod Q_i(\boldsymbol{\omega})} \tag{2}$$

With P_i equalling denominator of factor i of the path, and $Q_i(\boldsymbol{\omega})$ equalling product of pairwise differences of i with all other denominators. And $C(\boldsymbol{\omega})$ the part of the path following the discrete event. The outlined correspondence between the state space graph and the GF no longer holds for the inverted GF unless we transform the graph and collapse each section of all path leading up to the discrete event node to a new node (see fig. 2.1.4). Once this operation has been performed, each path of the graph represents a product of factors again. Given that inverting is not performed until evaluation, we can simply store the indices of the equations and associate them to the edge leading up to the discrete event. Under the hood however, we perform one additional operation while collapsing the graph, and make the (sets of) equations the actual nodes of the graph (see fig. 2.1.4 C). This simplifies any algorithm only dealing with the equations once generated (see 2.3). Finally, note that so far, this approach is only compatible with a single discrete event. Adding in more discrete events would require the use of a CAS to determine the inverse transform.

2.2 graph traversal

Given an equation graph (as described in fig. 2.1.4) and a dependency sequence describing the evaluation order of the graph, a general algorithm to propagate any evaluation of the equations associated with each node is given by alg. 1. It relies on the fact that implicitly the edges of the graph represent multiplication. Once both multiplication and addition have been defined for the propagated values we can rely on the commutative property to efficiently traverse the graph towards the root. Especially in the case where addition is a less costly operation than multiplication (as is the case for polynomials, see 2.3), it will pay off to add the values associated with the children of a node prior to multiplication.

2.3 blockwise site frequency spectrum

The block-wise site frequency spectrum or bSFS is the vector of site frequency spectrum counts in short blocks of a fixed length (Bunnefeld et al., 2015). It is the probability of seeing k_i mutations on each of the i distinguished branchtypes. For an unphased sample of size 3 the bSFS would be

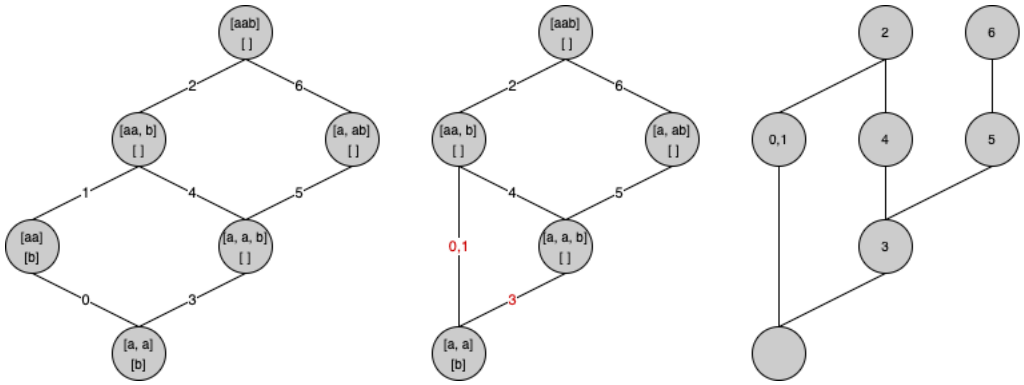

Figure 2: **From coalescent state space to equation graph:** setup identical as fig. 2.1.2. A shows the initial coalescent state space graph, while B shows the collapsed form grouping all parts of each path that require inverting with respect to the dummy variable associated with the discrete event. C takes this logic one step further turning the (sets of) equations previously associated with the edges into nodes, simplifying algorithms that do not require state space information.

Algorithm 1: Propagate values through graph.

```
function PROPAGATE;
  Input: adjacency list of graph, node_values, evaluation_order
foreach parent in evaluation_order do
    children = graph[parent];
    temp = 0;
    if children then
        foreach child in children do
            temp+=node_values[child];
        end
        if parent not root then
            node_values[parent] = PRODUCT(temp, node_values[parent]);
        else
            return temp;
        end
    end
end
```

a vector of the form $\{k_1, k_2\}$ with k_1 and k_2 representing the number of singletons and doubletons respectively. Each count is an element within the interval $[0, 1, 2, ..., k_{max} + 1]$ where $k_{max} + 1$ is used to bin all mutation configurations with more than k_{max} mutations.

Deriving the probability for each of the k^c_{max+2} possible bSFS configurations, with c the length of the bSFS vector, has thus far been the way in which the recursively generated Laplace transform of coalescence time distributions has been used. The probability for configuration $\boldsymbol{k}$ is equal to $(-\theta/2)^{\sum \boldsymbol{k}}$ times the coefficient of $\boldsymbol{\omega}^{\boldsymbol{k}}$ in a truncated multivariate Taylor series of the GF (see eq. (1) in Lohse et al. (2011) for details).

Any naive approach, based on calculating all higher order derivatives using a CAS, will suffer from an explosion in the number of terms due to the Leibniz or product rule when differentiating. Generally a CAS will fail to take into account the fact that the same partial derivatives are computed multiple times. This problem has been well-studied for the purpose of automatic differentiation algorithms (Neidinger, 1992, 1995; Griewank et al., 2000; Bettencourt et al., 2019). One possible approach to overcome this issue consists of implementing recurrence relations on how to combine truncated Taylor series. Thus essentially building higher-order derivatives from the ground up by defining the series coefficients for all elementary functions and defining the recurrence relations for arithmetic operators as well as function compositions (Neidinger, 2013). Here, we will not build up the derivatives from the elementary functions but will make use of the algorithm defining the product of two truncated Taylor series (see alg. (8) in Neidinger (2013) and, see, also adapt font). Note that adding two truncated Taylor series simply amounts to the pairwise addition of all corresponding coefficients.

Algorithm 2: Product of two truncated Taylor series (Neidinger, 2013).

```
function SERIES_PRODUCT;
  Input  : Two same-size arrays of floats a and b
  Output: array c of same size as a and b
foreach multi-index k do
    sum = 0;
    foreach multi-index j ≤ multi-index k do
        sum = ADD(sum, a[j] * b[k − j]);
    end
    c[k] = sum;
end
return c[ ];
```

Because in our case the equations associated with each of the nodes of the equations graph are well characterised and fall in one of two categories depending on whether an inversion step was needed, we will benefit from defining closed-form implementations of the derivatives with respect to each of the distinguished branchtypes. Again, equations will either take the form of a rational function with a constant numerator (branchtype variables never appear in the numerator) and a multi-variate polynomial of order 1 as denominator (see eq. 1). Or will be a sum of exponentials divided by higher order polynomials (see eq. 2). And for both cases, algorithms **??** and **??** respectively show how operations on the integer coefficient matrix allow us to efficiently compute all the coefficients defining a truncated Taylor series expansion of these equations ready to perform the graph traversal algorithm.

Note however that when binning the probability of seeing more than k_{max} mutations for each branchtype, a single pass through the graph does not suffice. In fact, $2^{|\boldsymbol{k}|}$ passes are required. This is due to the fact that calculating the binned residual probabilities requires us to marginalise the GF over the l variables for which the residual probability is determined. This amounts to setting those l dummy variables to 0, compute all derivatives for the remaining variables and propagate the coefficients of the new truncated Taylor series in $|\boldsymbol{k}| - l$ variables. Given that without having to calculate marginals a single pass would suffice, not binning probabilities, but instead calculating the probabilities of all observed mutation configuration might be more efficient in specific cases.

Also remark that alg. 2 contains an explicit ADD function. This refers to the fact that care needs to be taken when summing (a subset of) the coefficients of a Taylor series. Because these will be both positive and negative, catastrophic cancellation might occur leading to accuracy loss. We have implemented the compensated summation algorithm of Ogita-Rump-Oishi (Ogita et al., 2005) as an efficient way to bound the potential loss in accuracy. Another advantage of using Taylor series coefficients rather than the derivatives is that the coefficients will always be smaller by a factor $(\sum \boldsymbol{k})!$ leading to less cumulative roundoff error.

2.4 simplification

Following Lohse et al. (2016), we can simplify the GF by taking away root and/or phase information. removing phase info: replacing all unique labels by a single label determined by the population they were sampled from. practically: do this when providing sample list.

removing root: practically: pass branchtype dict, mapping branchtypes to their corresponding label. How do we enforce a particular order? Otherwise, fully labelled

3 Results and Discussion

In summary, the work presented here constitutes a CAS-independent, open-source implementation of the GF. A general outline has been given on how to rely on the correspondence between the coalescent state space graph and the GF to query the distribution of Laplace transformed coalescence times efficiently. In particular, algorithms have been laid out to efficiently calculate the bSFS making this package ideal as a backend for likelihood calculations.

Currently, `agemo` is implemented in `python 3`, relying on `numba` jit-compilation to speed up the critical parts of the code. Compiling the code using `numba` has a few consequences. Firstly, compilation happens each time the code is run and will require a few seconds. This is generally not an issue given that usually many points in parameter space will be evaluated for example when calculating the bSFS. Secondly, a few of the `numpy` inbuilt numerical precision algorithms ... no longer being used following jit-compilation. This has ... us to implement a compensated sum algorithm. A better/faster solution would be to temporarily increase machine precision for the evaluation of particular sums. This is not possible using `numba` however and will require us to transfer part of the code to `C`.

Because we have opted to no longer depend on a CAS to calculate derivatives, no longer placing the evaluation order of any expression out of user's reach allowing us to alleviate potential floating point precision issues. This was not possible previously forcing any implementation to use rationals rather than floats increasing computation time. However, there might be a point whenever one might want to analyse more complex models (e.g. containing more than one discrete event) that the need for relying on symbolic algebra or automatic differentiation becomes necessary. This does not change the general idea outlined in this paper. It will still pay off to evaluate the building block equations associated with the coalescent state space graph first, and then rely on graph traversal to get to the correct results. This could be achieved by putting in place a function to propagate a user-provided array (with first dimension equal to the number of equation nodes) through the graph using user-defined addition and multiplication rules. This way, we would not have the additional burden of having a CAS as a direct dependency.

There are more efficient algorithms to multiply (truncated) multivariate polynomials. The fast Fourier transform (FFT) would allow us to perform the operation at $\mathcal{O}(nlog(n))$ with n the order of the polynomial instead of $\mathcal{O}(n^2)$ for the current multiplication. However, gains would only be noticeable for a large number of distinguished branchtypes, or a really large k_{max} values.

These algorithms will not suddenly make it possible to extent the usage of the GF to any arbitrary model and sample size. The fact remains that the number of sample configurations grows superexponentially with sample size (Lohse et al., 2016).

Any incorporate more demographic events (bottleneck,) and incorporate hard sweep approximations (Bisschop et al., 2021). Although most people still process vcfs when it comes to mutation data. There is the more succinct treesequence format containing both topology information as well as mutations. It would therefore make a lot of sense to add on the ability to calculate the likelihood of a mutation configuration under a particular coalescent model as given by a marginal tree from a treesequence. This boils down to restricting the GF for that particular model to the observed topology, count the number of mutations of each distinguished branchtype and compute its probability.

acknowledgements

This work was supported by an ERC starting grant (ModelGenomLand, 757648)

References

Barton, N. H. and Wilson, I. (1995). Genealogies and geography. *Philosophical Transactions of the Royal Society B*, 349(1327):49–59.

Bettencourt, J., Johnson, M. J., and Duvenaud, B. D. (2019). Taylor-Mode Automatic Differentiation for Higher-Order Derivatives in JAX. 10.

Bisschop, G., Lohse, K., and Setter, D. (2021). Sweeps in time: Leveraging the joint distribution of branch lengths. *Genetics*, 219(2).

Bunnefeld, L., Frantz, L. A., and Lohse, K. (2015). Inferring bottlenecks from genome-wide samples of short sequence blocks. *Genetics*, 201(3):1157–1169.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., and Foll, M. (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, 9(10).

Griewank, A., Utke, J., and Walther, A. (2000). Evaluating Higher Derivative Tensors by Forward Propagation of Univariate Taylor Series Source : Mathematics of Computation , Jul ., 2000 , Vol . 69 , No . 231 (Jul ., 2000), pp . 1117- Publi. *Mathematics of Computation*, 69(231):1117–1130.

Griffiths, R. C. (1991). The Two-Locus Ancestral Graph. *Selected proceedings of the Sheffield symposium on applied probability*, 18(1991):100–117.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genetics*, 5(10):e1000695.

Hey, J. (2004). Multilocus Methods for Estimating Population Sizes, Migration Rates and Divergence Time, With Applications to the Divergence of Drosophila pseudoobscura and D. persimilis. *Genetics*, 167(2):747–760.

Hobolth, A., Siri-Jégousse, A., and Bladt, M. (2019). Phase-type distributions in population genetics. *Theoretical Population Biology*, 127:16–32.

Hudson, R. R. (1983). Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, 37(1):203–217.

Kingman, J. F. (1982). The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248.

Lohse, K., Chmelik, M., Martin, S. H., and Barton, N. H. (2016). Efficient strategies for calculating blockwise likelihoods under the coalescent. *Genetics*, 202(2):775–786.

Lohse, K., Harrison, R. J., and Barton, N. H. (2011). A general method for calculating likelihoods under the coalescent process. *Genetics*, 189(3):977–987.

Neidinger, R. D. (1992). An Efficient Method for the Numerical Evaluation of Partial Derivatives of Arbitrary Order. *ACM Transactions on Mathematical Software (TOMS)*, 18(2):159–173.

Neidinger, R. D. (1995). Computing multivariable Taylor series to arbitrary order. In *Proceedings of the international conference on Applied programming languages - APL '95*, pages 134–144, New York, New York, USA. ACM Press.

Neidinger, R. D. (2013). Efficient recurrence relations for univariate and multivariate Taylor series coefficients. *Conference Publications*, pages 587–596.

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. *Journal of mathematical biology*, 29:59–75.

Ogita, T., Rump, S. M., and Oishi, S. (2005). Accurate sum and dot product. *SIAM Journal on Scientific Computing*, 26(6):1955–1988.

Råde, L. (1972). On the use of generating functions and laplace transforms in applied probability theory. *International Journal of Mathematical Education in Science and Technology*, 3(1):25–33.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2):437–460.

Takahata, N. (1988). The coalescent in two partially isolated diffusion populations. *Genetical Research*, 52(3):213–222.

Weissman, D. B. and Hallatschek, O. (2017). Minimal-assumption inference from population-genomic data. *eLife*, 6:1–21.

Wilkinson-Herbots, H. M. (1998). Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology*, 37(6):535–585.

Yang, Z. (2002). Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, 162(4):1811–1823.

Supplementary Information

closed-form derivatives

$$\frac{\partial f(\mathbf{x})^{-1}}{\partial^{k_i} x_i} = (-1)^s s! \frac{c_i^{k_i}}{f(\mathbf{x})^2} \tag{3}$$

with $s = \sum_{i=1} k_i$ and $f(\mathbf{x}) = c_i x_i + b$

$$\frac{\partial g(\mathbf{x})}{\partial^{k_i} x_i} = c^s c_i^{k_i} g(\mathbf{x}) \tag{4}$$

with $s = \sum_{i=1} k_i$ and $g(\mathbf{x}) = e^{c * \sum c_i x_i + b}$