Coalescent simulation in continuous space

Citation for published version:
Kelleher, J, Etheridge, AM & Barton, NH 2014, ‘Coalescent simulation in continuous space: Algorithms for large neighbourhood size’ Theoretical population biology. DOI: 10.1016/j.tpb.2014.05.001

Digital Object Identifier (DOI):
10.1016/j.tpb.2014.05.001

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Final published version

Published In:
Theoretical population biology

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Coalescent simulation in continuous space: Algorithms for large
neighbourhood size

J. Kelleher\textsuperscript{a,}*\textsuperscript{1}, A.M. Etheridge\textsuperscript{b}, N.H. Barton\textsuperscript{c}

\textsuperscript{a} Institute of Evolutionary Biology, University of Edinburgh, Kings Buildings, West Mains Road, Edinburgh EH9 3JT, UK
\textsuperscript{b} Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK
\textsuperscript{c} Institute of Science and Technology, Am Campus I, A-3400 Klosterneuburg, Austria

\textbf{ARTICLE INFO}

\textbf{Article history:}
Received 28 October 2013
Available online xxxx

Keywords:
Isolation by distance
Coalescent simulation

\textbf{ABSTRACT}

Many species have an essentially continuous distribution in space, in which there are no natural divisions between randomly mating subpopulations. Yet, the standard approach to modelling these populations is to impose an arbitrary grid of demes, adjusting deme sizes and migration rates in an attempt to capture the important features of the population. Such indirect methods are required because of the failure of the classical models of isolation by distance, which have been shown to have major technical flaws. A recently introduced model of extinction and recolonisation in two dimensions solves these technical problems, and provides a rigorous technical foundation for the study of populations evolving in a spatial continuum. The coalescent process for this model is simply stated, but direct simulation is very inefficient for large neighbourhood sizes. We present efficient and exact algorithms to simulate this coalescent process for arbitrary sample sizes and numbers of loci, and analyse these algorithms in detail.

\textsuperscript{1} Corresponding author.
E-mail addresses: jerome.kelleher@ed.ac.uk (J. Kelleher), etheridg@stats.ox.ac.uk (A.M. Etheridge), n.barton@ist.ac.at (N.H. Barton).

1. Introduction

As Wright noted (1978, p. 54), many species, such as the dominant plants in grasslands, have an essentially continuous and uniform distribution in space. In these populations there are no divisions between discrete units, within which mating occurs randomly and between which migrants are exchanged. Despite this obvious observation, the majority of methods available to simulate the history of such populations require this arbitrary structure to be imposed. The typical approach is to use a stepping stone model in which a user must specify a deme size and migration rates between these demes. This approach models many evolutionary scenarios very well, but it can hardly be described as a natural representation of a continuously distributed population.

This arbitrary grid of demes is required because of the lack of a well-defined model to capture the dynamics of continuously distributed populations. The classical model of isolation by distance (Wright, 1943; Malécot, 1948) suffers from severe technical problems, most notably a lack of local density regulation leading to clumps of arbitrarily high density (Felsenstein, 1975). The Wright–Malécot model is also inconsistent with some important biological observations such as large-scale patterns, correlations across loci, and lower diversity than expected from census numbers (Barton et al., 2010b). Etheridge (2008) introduced a model that can describe extinction and recolonisation over a range of scales and provides a simple and effective means of modelling the evolution of a continuously distributed population. In this model, each individual occupies a fixed location in space during their lifetime, and all movement and reproduction happens as a consequence of extinction/recolonisation events. Events span a range of scales, from the steady process of local reproduction within neighbourhoods to large-scale demographic shifts affecting a substantial fraction of the population. The model solves the long-standing technical problems mentioned above, and opens the way for mathematically rigorous analytical and inference methods based on a true continuum.

This model of extinction and recolonisation can be viewed as a framework for describing population models, in which different realisations of the framework share a few essential characteristics but have a great deal of freedom in the details of how the replacement mechanisms work. We focus on one instance of the model here, known as the disc model (generalisations of the methods are discussed in Section 5). Suppose that we have a population of individuals distributed uniformly at random on a square torus of diameter $L$ with some density $\rho$. The evolution of...
the population is driven by extinction/recolonisation events that occur at rate $\lambda$. At an event, we choose a point $z$ uniformly at random on the torus, and this event only affects the individuals within distance $r$ of $z$. Within this disc centred on $z$, a small number of parents $v$ are chosen uniformly, and each individual has a probability $u$ of dying. (The parameters $r$ and $u$ are often referred to as the radius and impact of an event, respectively.) A Poisson number of children with mean $\mu r^2 u$ are then thrown down uniformly within the disc, and the children choose parents uniformly from the pool chosen earlier, completing the event.

Moving to the limit of high density, we then obtain the spatial Fleming–Viot process, for which many analytical results have been derived (Barton et al., 2013b), and the coalescent process is simple to describe. We begin with a sample of $n$ lineages and proceed backwards in time event-by-event until one of these events overlaps at least one lineage. Then, this lineage has probability $u$ of having been born in this event (since we are proceeding backwards in time); if it was, it jumps to the location of a parental lineage whose location is chosen uniformly at random from the same disc. Eventually, two lineages will be overlapped by the same event and with probability $u^2/\nu$ a coalescence event occurs and both lineages jump to the location of their parent. We may also have many lineages descending from a single parent in a process. This process can be formalised as an algorithm very simply (Kelleher et al., 2013); unfortunately, however, such a direct implementation is very inefficient if we wish to simulate the history of populations with large neighbourhood size.

Wright’s neighbourhood size $\Lambda$ (1943; 1946) is widely regarded as the most important parameter describing populations evolving in a two-dimensional continuum, and determines the relative rates of genetic drift and gene flow. Wright (1943) identified the approximate magnitudes of the neighbourhood sizes we might expect to observe in natural populations. Small values, with $\Lambda < 100$, correspond to populations with short range dispersal and a high degree of differentiation. Large neighbourhood sizes, with $\Lambda > 10^3$, correspond to populations that are “substantially equivalent to panmixia throughout a range of any conceivable size”. Empirical estimates of $\Lambda$ have supported Wright’s analysis, with mobile species such as Drosophila having neighbourhood size greater than $10^4$ and estimates for plant species being as little as 10 or less (Crawford, 1984). These estimates are also consistent with the range of observed $F_{ST}$ values (Morjan and Rieberg, 2004).

In the disc model, neighbourhood size is given very simply as $\Lambda = \nu/\mu$ (Barton et al., 2013a), where $\nu$ is the number of potential parents and $u$ is the probability an individual dies in an event. Thus, assuming a single parent and a modest neighbourhood size of 100, we arrive at an impact of $u = 1/100$. Unfortunately, these parameters present serious difficulties to a direct simulation of the coalescent process, in which we expect to generate $1/u$ events that intersect with a lineage before it jumps. For large neighbourhood size, this represents a heavy computational burden.

In this article we develop algorithms to simulate the coalescent process efficiently for large neighbourhood and range sizes. In Section 2 we begin by defining an algorithm to simulate the important special case of the ancestry of a sample of size two. In this case it is simple to calculate the exact distribution of the time that elapses between events in which a lineage jumps, allowing us to simulate only these events. We then generalise these ideas in Section 3 to allow us to simulate the history of a sample of $n$ lineages. Because of the inherent geometric difficulty of the problem we cannot directly generalise the pairwise algorithm. By tessellating the torus into square pixels of edge $s$ we can approximate the areas we require very quickly and then apply a correction to precisely cancel the errors entailed using rejection sampling. We then model the behaviour of the algorithm and derive an optimal value for the pixel size $s$ to minimise the computational effort required for a lineage jump. Section 4 continues by generalising this algorithm so that we can simulate the history of a sample of individuals with a large number of loci. We describe a general algorithm to distribute genetic material among ancestors over an arbitrary number of loci, and show that this method is efficient when recombination is fast. In particular, we show that the approach is much more efficient than the method used by the classical ma program (Hudson, 2002).

A Python interface to an efficient implementation of this multilocus algorithm is available at https://pypi.python.org/pypi/discsim under the terms of the GNU General Public License.

In this article we study algorithms in much more detail than is customary in the population genetics literature. We provide detailed and unambiguous algorithm listings in a well established format (Knuth, 1997a, Section 1.1), which is important for several reasons. Firstly, it is impossible to establish the correctness of an algorithm that is stated in an ambiguous natural language format. Given the centrality of coalescent algorithms in modern population genetics and the difficulty of detecting errors in computer programs with stochastic results, it is imperative that we are certain the underlying algorithm is correct. Secondly, since stochastic programs are prone to subtle errors, it is important that a diversity of implementations of a given algorithm exist. If we are dependent on a single (possibly opaque) implementation of a given algorithm, then it is very difficult to verify the results of this program. With a detailed algorithm listing, implementation is routine and allows for multiple independent implementations. Finally, a rigorous description of an algorithm allows us to analyse the properties of this algorithm using mathematical techniques, allowing us to improve performance without resorting to approximations.

2. Pairwise coalescent

Simulating the ancestry of a sample of size two is an important special case, as we are often interested in pairwise statistics. In the extinction/recolonisation continuum model this involves tracing the history of two lineages as they move around the range before meeting and, eventually, coalescing. Events fall uniformly at random across the range, but it is only events that fall within a disc of radius $r$ around the location of at least one of the two lineages that can result in a lineage jumping to a new location. Similarly, the lineages can only coalesce in a given event if the centre of the event falls in the intersection of the discs around these lineages: the centre must be within distance $r$ of both lineages or they cannot both be born in the event.

Remark 1. A lineage is defined by a point $x$ on the torus, as this is the location of the ancestral individual in question. To simplify the following discussions, however, we refer to a lineage as the disc of radius $r$ around this point. For example, when we refer to the intersection of two lineages this is a shorthand for the intersection of the discs of radius $r$ centred on the locations of the lineages.

The idea behind the pairwise coalescent algorithm is straightforward—we only simulate events that result in a lineage jumping, and calculate the distribution of the time that elapses between these events. Since events fall in a Poisson process with rate $\lambda$, we can “thin” this process to generate only the events that result in jumps. Specifically, if events of a particular type occur with probability $p$, we know that these events constitute an independent process with rate $p \lambda$. In the pairwise algorithm we must thin the Poisson process twice: once to remove the events whose centres fall outside of the union of the two lineages, and again to remove the events that fall within this area but do not result in a lineage jumping.

To calculate the rate at which at least one lineage jumps, we must calculate the area covered by exactly one lineage, $\alpha_1$, and the area covered by exactly two lineages, $\alpha_2$. These are given...
by the symmetric difference and the union of the two lines,
respectively. (The symmetric difference of two sets A and B is
(A \cup B) \setminus (A \cap B).) Let \( A(x) \) denote the area of the intersection
of two discs of radius \( r \) with centres distance \( x \) apart, i.e.,
\[
A_r(x) = 2r^2 \arccos \left( \frac{x}{2r} \right) - \frac{x}{2} \sqrt{4r^2 - x^2}
\]
(1)
for \( x < 2r \) and \( A_r(x) = 0 \) otherwise. We also define
\( \bar{u}_k = 1 - (1 - u)^k \)
in the interest of brevity.

**Lemma 1.** Given two lineages separated by distance \( x \), the rate at which at least one lineage jumps is
\[
\Omega_2 = \frac{\lambda}{L^2} (\alpha_1 u + \alpha_2 \bar{u}_2)
\]
where \( \alpha_2 = A_r(\alpha_1) \) and \( \alpha_1 = 2\pi r^2 - 2\alpha_2 \).

**Proof.** Events fall uniformly at rate \( \lambda \) and hit the area covered by
exactly one lineage at rate \( \lambda \alpha_1/L^2 \). A single lineage subsequently
jumps with probability \( u \), and so the rate at which jump events
fall in the symmetric difference of the two lineages is \( \lambda \alpha_1 u/L^2 \).
Similarly, events fall in the intersection of the two lineages at rate
\( \lambda \alpha_2/L^2 \). Then, since there are two lineages that may jump in this
region, the probability that at least one jumps is \( \bar{u}_2 \). \( \square \)

**Lemma 2.** Conditional on at least one lineage jumping, the probabil-
ity that the centre of the event is in the intersection of the two lines
is \( 1 - \alpha_1/(\alpha_1 - \alpha_2(u - 2)) \).

**Proof.** The rate at which jump events fall in the intersection of the
lines is \( \alpha_2/\bar{u}_2 \); therefore, the probability that a given
jump event is of this class is \( \alpha_2/\Omega_2 \). \( \square \)

**Lemma 3.** Conditional on the centre of a jump event falling in the
intersection of the two lineages, the probability that both lineages
jump and coalesce is \( u/(2 - u) \).

**Proof.** The probability that both lineages are hit is \( u^2 \), and the probability of at least one lineage being hit is \( \bar{u}_2 \). Therefore, the probability of both being hit conditional on at least one being hit is \( u^2/\bar{u}_2 = u/(2 - u) \). \( \square \)

To describe the algorithm concisely, we require some notation.
Let \( D_j(z) \) define a disc of radius \( r \) centred at \( z \) on a two-dimensional
torus of diameter \( L \), and let \( |x| \) be the Euclidean norm of the vector \( x \)
on such a torus (we suppress the dependence on \( L \) for simplicity).
We also require notation to describe sampling values from random
variables drawn from a variety of distributions. Let \( R_A(\xi_1, \ldots, \xi_k) \)
define a single independent sample from a random variable
with distribution \( \Delta \) and parameters \( \xi_1, \ldots, \xi_k \). (Note that each
instance of \( R_A(\xi_1, \ldots, \xi_k) \) within an algorithm listing represents an
independent random sample from the specified distribution.)
Using this notation, we define \( R_{U_2}(A) \) to be an element of the set
\( A \) chosen uniformly at random, and \( R_{U}(\lambda) \) as a sample from an
exponentially distributed random variable with rate \( \lambda \).

**Algorithm P (Pairwise Coalescent).** Simulate the coalescence time
t of two lineages sampled at distance \( x \) under a model in which
events with radius \( r \) and impact \( u \) occur at rate \( \lambda \) on a two-
dimensional torus of diameter \( L \).

**P1.** [Initialisation.] Set \( y_1 \leftarrow (0, 0), y_2 \leftarrow (0, x) \) and then set
\( t \leftarrow 0 \).

**P2.** [Event.] Set \( \omega_2 \leftarrow A_r(||y_1 - y_2||) \) and \( \alpha_1 \leftarrow 2\pi r^2 - 2\omega_2 \).
Then, set \( \Omega \leftarrow \lambda(\alpha_1 u + \alpha_2 \bar{u}_2)/L^2 \) and \( t \leftarrow t + R_{\Omega}(\Omega) \).
If \( R_{\Omega}((0, 1)) < 1/2 \), set \( j \leftarrow 1 \); otherwise set \( j \leftarrow 2 \).
Finally, if \( R_{\Omega}((0, 1)) < 1 - \alpha_1/(\alpha_1 - \alpha_2(u - 2)) \), go to P4.

**P3.** [Symmetric difference.] Set \( k \leftarrow (j \ mod\ 2) + 1 \) and then set \( z \leftarrow R_{\Omega}(D_j(y_1) \setminus D_j(y_2)) \) and go to P5.

**P4.** [Intersection.] Set \( z \leftarrow R_{\Omega}(D_j(y_1) \cap D_j(y_2)) \). Then if \( R_{\Omega}((0, 1)) < u/(2 - u) \), terminate the algorithm.

**P5.** [Jump.] Set \( y_j \leftarrow R_{\Omega}(D_j(z)) \) and go to P2. \( \blacksquare \)

The algorithm sets up two lineages on the torus separated by
distance \( x \) and simulates their history jump-by-jump until the
lineages coalesce, returning their coalescence time \( t \). In step P2
we first determine the distance between the two lineages and compute \( \alpha_1 \) and \( \alpha_2 \). We then calculate the rate at which jumps occur using **Lemma 1** and increment \( t \) accordingly.

After this, we choose a lineage \( j \) to jump, and decide if the centre of the event \( z \) is in the symmetric difference of the two lineages or their intersection according to **Lemma 2**. If the centre falls in the symmetric difference of the two lineages, we generate \( u \) uniformly in \( D_j(y_1) \setminus D_j(y_2) \) in P3 and then proceed immediately to step P5.

If, on the other hand, \( z \) falls in the intersection of the two
lineages we proceed to step P4, where we throw \( z \) down uniformly
within this area. Then, according to **Lemma 3**, both lineages jump
with probability \( u/(2 - u) \) and so a coalescence occurs and we
terminate the algorithm. Otherwise, lineage \( j \) jumps to a new
location in the disc centred on \( z \), and we return to P2.

Given Lemmas 1-3 it is straightforward to show that **Algorithm P** simulates the coalescent process correctly. In **Fig. 2** we compare the results of calculating the probability of identity in state using **Algorithm P** with numerical estimates (outlined in the **Appendix**). Under the infinitely many alleles model, the probability
of identity in state for two genes, \( F \), is the probability that no
mutations have occurred since the lineages diverged. We therefore
have \( F = \exp(-2\mu t) \) for mutation rate \( \mu \) and coalescence time \( t \).
By taking many replicates we can estimate \( F \) from simulations, and
we see an excellent agreement between the results of **Algorithm P**
and numerical methods in **Fig. 2**.

The algorithm is clearly very efficient, since each lineage jump
requires a constant number of algorithm steps, and each of these
steps is straightforward. Indeed, other than generating points
uniformly within the symmetric difference and intersection of the
lineages, which can be achieved via standard methods (**Knuth**, 1997b, Section 3.4), the time required for each jump is constant.

It is not difficult to generalise **Algorithm P** to simulate several
classes of event occurring at different rates. Suppose we have \( k \)
classes of event, occurring at rate \( \lambda_j \) with radius \( r_j \) and impact \( u_j \). To
do this, we must modify step P2. In this new step, we first calculate
the rate at which successful events are happening for all of these
classes by setting \( \omega_j \leftarrow \lambda_j (\alpha_1 u_j + \alpha_2 \bar{u}_2)/L^2 \) for \( 1 \leq j \leq n \),
where \( \omega_{j, 2} = A_r(||y_j - y_i||) \) and \( \alpha_1 = 2\pi r_j^2 - 2\omega_{j, 2} \), as before.
Having calculated the rates at which each of these classes of event
are occurring, we increment time by setting \( \Omega \leftarrow \omega_1 + \cdots + \omega_k \)
and \( t \leftarrow t + R_{\Omega}(\Omega) \). We then choose an event class \( j \) with probability
\( \omega_j/\Omega \), and set \( r \leftarrow r_j \) and \( u \leftarrow u_j \). After this, we can proceed
with P2 as before, choosing the location of the event in either the
union or symmetric difference of the lineages. The remaining steps of
the algorithm are unchanged.

**3. Single locus coalescent**

**Algorithm P** provides a satisfactory method to simulate the
history of two genes. The basic idea of the algorithm is to partition
the torus into regions in which zero, one or two lineages intersect,
which can be done quite simply. Although we can generalise the
**Lemmas 1**–**3** to \( n \) lineages easily, it is not so simple to actually
calculate the areas involved. Specifically, calculating the area in
which \( k \) out of a given \( n \) discs intersect is not a trivial problem.
We approach the problem instead by calculating these areas
approximately using a tessellation of the torus into square pixels.
and then correct for the errors introduced precisely using rejection sampling.

The output of the algorithm is the simulated history of the sample (π, r), where π is an oriented tree and r the corresponding node times. An oriented tree (Knuth, 2011, p. 461) is a sequence π1, π2, ..., such that πj is the parent of node j and j is a root if πj = 0. For example, the oriented trees

```
1 4 5
2 3
```
correspond to the sequences 44550, 4440 and 54450, respectively. Oriented trees provide an elegant means of describing genealogies, since only parent–child relationships are encoded and the order of children at a node is not important. Oriented trees also have several advantages over more traditional approaches such as nested parentheses. The principal advantage for our purposes here is that we can describe coalescent algorithms in terms of oriented trees concisely and without ambiguity.

3.1. Algorithm N

The pairwise coalescent, Algorithm P, is almost entirely defined by Lemmas 1–3. Although we cannot generalise the algorithm in a direct way because of geometric difficulty, we still require the corresponding results for a sample of n lineages. The following results generalise the pairwise results, and form the basis of the subsequent algorithm.

**Lemma 4.** Given a sample of n lineages, the rate at which at least one lineage jumps is

\[ \Omega_n = \frac{\lambda}{2} \sum_{k=1}^{n} \alpha_k \bar{u}_k \]

where \( \alpha_k \) is the area covered by the intersection of exactly k lineages.

**Proof.** The centre of an event \( x \) causing a lineage to jump must fall in the intersection of one or more lineages, and an event falls in the intersection of k lineages with probability \( \alpha_k/2^k \). If \( z \) falls in the intersection of k lineages, then at least one lineage jumps with probability \( \bar{u}_k \). Summing over all the possible intersections of lineages, we obtain an overall jump rate of \( \lambda \sum_{k=1}^{n} \alpha_k \bar{u}_k/2^k \) as required. □

**Lemma 5.** Conditional on at least one lineage jumping, the probability that the centre of the event is in the intersection of exactly k lineages is \( \alpha_k \bar{u}_k / \sum_{j=1}^{n} \alpha_j \bar{u}_j \).

**Proof.** The rate at which the centre of jump events fall in the intersection of k lineages is \( \alpha_k \bar{u}_k / 2^k \); therefore, the probability that a given jump event is of this class is \( \alpha_k \bar{u}_k / \Omega_n \). □

**Lemma 6.** Given that \( z \) falls in the intersection of k lineages and that at least one of them jumps, the probability that exactly j jumps is

\[ \beta(k, j) = \binom{k}{j} \frac{w^j(1-w)^{k-j}}{\bar{u}_k}. \]

**Proof.** Exactly j lineages jump with probability \( w^j(1-w)^{k-j} \), and there are \( \binom{k}{j} \) ways this can happen. Then, since we have conditioned on at least one lineage jumping, we divide by \( 1-w \), to give us \( \beta(k, j) \) as required. □

Calculating the area in which k lineages intersect is nontrivial, but we can use Lemmas 4–6 to derive an efficient simulation algorithm. In this algorithm we proceed by dividing our torus into pixels of edge s, and then use this tessellation to quickly calculate an over-estimate of the area in which k lineages intersect. The errors introduced by this over-estimate can then be precisely cancelled out by discarding potential jump events with a certain probability using rejection sampling.

The key idea of Algorithm N is to divide the range into pixels of edge s; we assume that L/s is an integer so that pixels do not overlap. Let a pixel \( v \in [1, \ldots, L/s]^2 \) define a \( s \times s \) square on the torus such that a point \( x \) is in the pixel \( v \) if \( 0 \leq x_i - x_i < s \) and \( 0 \leq x_2 - x_2 < s \) (i.e., the top–right corner of pixel \( v \) is \( sv \) in torus coordinates). Let \( D^i_s(z) \) be the set of pixels that intersect with \( D_i(z) \), i.e.,

\[ D^i_s(z) = \{ ([x_1/s], [x_2/s]) | x \in D_i(z) \}. \]

We also require some further notation to describe sampling from random variables. Firstly, we extend the notation \( R_0(A) \) for sampling an element of a set \( A \) uniformly, by letting \( R_0(A, k) \) be a \( k \)-subset of \( A \) chosen uniformly at random (that is, \( R_0(A, k) \) is a random sample of \( k \) elements chosen from \( A \) without replacement). Secondly, we let \( R_0(p) \) be a random sample from a discrete distribution with probability mass function defined by the vector \( p \), such that if we set \( j \leftarrow R_0(p) \), then \( j = k \) with probability \( p_k \). Also, let \( \varnothing \) be the null point.

**Algorithm N (Single Locus Coalescent).** Simulate the ancestry \( (\tau, x) \) of individuals sampled at locations \( x_1, \ldots, x_n \) under a model in which events with radius \( r \) and impact \( u \) occur at rate \( \alpha \) on a two-dimensional torus of diameter \( L \) using pixel size \( s \), such that \( L/s \in \mathbb{N} \).

N1. [Initialisation.] Set \( \tau_j \leftarrow 0 \), \( \eta_j \leftarrow 0 \) and \( y_j \leftarrow \varnothing \) for \( 1 \leq j < 2n \) and then set \( \kappa \leftarrow n \), \( \eta \leftarrow n + 1 \), \( t \leftarrow 0 \) and \( h^* \leftarrow 0 \). Set \( P_v \leftarrow \varnothing \) for \( v \in [1, \ldots, L/s]^2 \) and set \( Q \leftarrow \varnothing \) for \( 1 \leq j \leq n \). Then, for \( 1 \leq j \leq n \) set \( y_j \leftarrow x_j \) and for each \( v \in D^j_s(x_j) \), set \( P_v \leftarrow P_v \cup \{ j \} \), \( \bar{P}_v \leftarrow \bar{P}_v \cup \{ j \} \), \( h^* \leftarrow \max(h^*, h) \), and, if \( h^* > 1 \), set \( Q_{j-1} \leftarrow Q_{j-1} \cup \{ v \} \).

N2. [Rate.] When \( Q_{j-1} = \varnothing \), set \( h \leftarrow h^* - 1 \). Then, set \( w \leftarrow 0 \) and for \( 1 \leq j \leq h^* \), set \( \bar{P}_j \leftarrow \{ \bar{P}_j \} \bar{u}_j \) and \( w \leftarrow w + \bar{u}_j \). Afterwards, set \( \bar{P}_j \leftarrow \bar{P}_j \cup w \) for \( 1 \leq j \leq h^* \).

N3. [Location.] Set \( t \leftarrow t + R_0(u \lambda s^2 \bar{D}^2_s) \), \( h \leftarrow R_0(p) \), \( v \leftarrow R_0(Q_0) \), and \( z \leftarrow (v \leftarrow R_0(0, 1)^2) \)). Then set \( s \leftarrow 0 \) and, for each \( v \in P_v \), if \( z \in D_i(y_i) \), set \( s \leftarrow s \cup \{ j \} \). Finally, if \( R_0(0, 1) \leq 1 - u \bar{u}_j / h \), return to N3.

N4. [Choose children.] Set \( b_j \leftarrow B(\bar{S}, \bar{S}) \) for \( 1 \leq j \leq |S| \), then set \( j \leftarrow R_0(b) \) and \( C \leftarrow R_0(S, j) \). Then, if \( |C| = 1 \), set \( k \leftarrow C \); otherwise, set \( k \leftarrow \eta_j \).

N5. [Remove children.] For each \( j \in C \), and for each \( v \in D^j_s(y_j) \), set \( h \leftarrow h \leftarrow \bar{P}_j \), \( \bar{P}_j \leftarrow \{ \} \), \( Q_0 \leftarrow Q_0 \cup \{ v \} \) and, if \( h^* > 1 \), set \( Q_{j-1} \leftarrow Q_{j-1} \cup \{ v \} \).

N6. [Insert parent.] Set \( y_j \leftarrow R_0(D_i(z)) \) and then for each \( v \in D^j_s(y_j) \), set \( P_v \leftarrow P_v \cup \{ k \} \), \( h \leftarrow h + \bar{P}_j \), \( \bar{P}_j \leftarrow \{ \} \), \( h^* \leftarrow \max(h^*, h^* + \bar{P}_j \), and, if \( h^* > 1 \), set \( Q_{j-1} \leftarrow Q_{j-1} \cup \{ v \} \). Finally, if \( |C| = 1 \) go back to N2.

N7. [Coalesce.] For each \( j \in C \), set \( \tau_j \leftarrow \eta_j \). Then, set \( \eta_j \leftarrow \eta_j + 1 \) and \( \kappa \leftarrow \kappa - |C| + 1 \). Finally, if \( \kappa > 1 \) go back to N2.

As illustrated in Fig. 1, Algorithm N maintains a spatial index via the matrix \( P \). Each entry in \( P \) corresponds to a pixel, and is the set of lineages that intersect with that pixel. The list \( Q \) is used to keep track of the pixels with a given occupancy (defined as the number of lineages that intersect with it), such that \( Q_k \) is the set of pixels with occupancy \( k \). These data structures serve several purposes. Firstly, \( Q \) allows us to quickly determine the number of pixels with
was born in this event and so we choose the exact number born according to Lemma 6. The variable $k$ is used represent the parental lineage. If exactly one child is born in the event, then there is no coalescence, and the lineage simply jumps; thus, we set $k$ to be the child lineage (through a slight abuse of notation). Otherwise, when more than one child is born in the event, a coalescence occurs, and $k$ is set to the new lineage $\eta$.

Step N5 then removes the child lineages from $P$ and $Q$. Afterwards, in step N6 we choose the location of the parent lineage, and then update $P$ and $Q$ to reflect the insertion of this new lineage, revising the maximum occupancy $h^*$ upwards, as required.

If $C$ contains more than one lineage, we then proceed on to step N7, where these lineages coalesce and we update the ordered tree and node time structures to reflect this. We set $\tau_j \leftarrow \eta$ for each $j \in C$ to signify that the parent of lineage $j$ is the new lineage $\eta$, and also set $\tau_j \leftarrow t$ to record the fact that lineage $\eta$ entered the sample at time $t$. Finally, we calculate the number of remaining lineages, and if this is greater than 1, return to step N2.

In the previous section we discussed the process by which the pairwise coalescent can be extended to multiple event classes. Unfortunately, it is not so simple to incorporate such events into Algorithm N. To fully generalise the algorithm we must maintain a spatial index for each class of event, and use this to calculate the rates at which events of the various classes are currently happening. While this is fairly straightforward to implement, the memory requirements and the costs of updating the indexes soon become a burden. In the most common case that we wish to model, however, there is a simpler and more efficient approach. If our model consists of small frequent events interspersed with rare large events (Barton et al., 2010a), then we can simply calculate the probability of a large event occurring without conditioning on at least one lineage jumping. Given that such large events are rare, this approach is very efficient since the cost of the large events that we simulate in which no lineages are affected is negligible when amortised over a large number of reproduction events.

3.2. Correctness

Unlike Algorithm P, it is not immediately obvious that Algorithm N correctly simulates the coalescent process given the rates that lineages jump. The algorithm works by over-estimating the area covered by the intersection of $k$ lineages and then discarding potential jumps with a certain probability, chosen to precisely cancel the error in jump rate that this introduces. The following lemmas prove that this process of rejection sampling results in the correct jump and coalescence rates.

Lemma 7. Lineages jump at the correct rate in Algorithm N.

Proof. The rate at which lineages jump as the result of events falling in the intersection of $k$ lineages is $\lambda \alpha_i \frac{\bar{u}_k}{L^2}$. Therefore, we must show that this rate holds for Algorithm N. Suppose we choose a pixel $v$ with occupancy $h$. Then, let $\alpha_i^v$ be the area of this pixel covered by the intersection of $k$ lineages and then discard potential jumps with a certain probability, chosen to precisely cancel the error in jump rate that this introduces. The following lemmas prove that this process of rejection sampling results in the correct jump and coalescence rates.
Cancelling terms, we see that the rate at which an event falls in the intersection of $k$ lineages and a lineage jumps in a pixel of occupancy $h$ is $\lambda \alpha_k h^2 / L^2$. Since pixels are disjoint, we can find the total rate by summing over all pixels with occupancy $h$ and over all occupancy values. This gives us

$$\frac{\lambda \alpha_k}{L^2} \sum_{h=1}^{\hat{n}} \sum_{v \in Q_h} a^v_h = \frac{\lambda \alpha_k h^2}{L^2}$$

as required. □

Lemma 7 proves that the rejection sampling step of N4 correctly adjusts the rate at which lineages jump to account for the approximated areas, resulting in an exact algorithm to simulate the spatial coalescent. The remainder of the algorithm can be easily verified; for example, we can immediately see that step N4 chooses a subset of the available lineages according to Lemma 6.

Fig. 2 shows the probability of identity in state calculated via an implementation of Algorithm N and via numerical methods outlined in the Appendix. There is an excellent agreement between the results. Although identity in state is a pairwise measure, numerical methods still provide a good means of verifying the correctness of the implementation. We begin with a sample of regularly spaced locations $x_2, x_3, \ldots, x_n$, simulate the history of the sample, and then calculate the probability of identity between the pairs $(x_1, x_2), (x_1, x_3), \ldots, (x_1, x_n)$. These probabilities are then aggregated over many replicates to obtain an estimate of the mean probability of identity in state over these distances.

3.3. Analysis

The previous subsection assures us that Algorithm N simulates the coalescent process correctly for any pixel size $s$, but does not give us any indication of what the value of this parameter should be. It is clear that $s$ has an important part to play in the amount of computational effort that is required to simulate the coalescent. If $s$ is too large we will generate many events that miss all of the lineages, and so spend a great deal of time looping around step N3. On the other hand, if $s$ is too small, we spend our time in steps N5 and N6 updating the indexing structures $P$ and $Q$.

One approach to choosing the value of $s$ is to simply run Algorithm N for a variety of $s$ values and make a choice based on the value which minimises the running time. While this is an effective method, it is unsatisfactory for several reasons. The most obvious problem is that there is no indication of the generality of the results obtained: the optimum value of $s$ may depend on the model parameters, and a great deal of computer time might be invested to explore the effects of $L, u$ and $r$. A more subtle problem is that this approach can only find the optimum value of $s$ for a given implementation of the algorithm. Ideally, each implementation should repeat this time consuming process to determine its own optimal pixel sizes for varying model parameters.

A much more satisfactory method is to analyse the algorithm, using theoretical methods to estimate the optimum value of $s$ and assess the effects of model parameters on this optimum. In principle, the analysis of an algorithm involves counting the number of primitive operations (such as assignments, arithmetic operations, etc.) incurred during its execution. Using this detailed model of the computational cost in terms of its input parameters, we can then derive the value of $s$ that minimises this function. In practice, however, we are almost never interested in such detailed information, and an approximate broad-brush stroke model of the running time of the algorithm provides all the information that we need.

In this section we apply this methodology to analyse Algorithm N. We derive a simple approximation to quantify the expected computational cost of the algorithm as a function of the pixel size $s$, which provides us with a useful prediction for its optimum value. The analysis of the algorithm is slightly complicated by the stochastic nature of the coalescent process. The most common means of analysing an algorithm is to derive an expression to approximate its total running time as a function of the input parameters. This would be very difficult in this case, as it would first involve deriving the coalescence time of a sample of size $n$ in the extinction/recolonisation model. We are not interested in the overall running time however, just minimising the cost of simulating the process. Since the basic unit of work in the process is a single lineage jump, we are therefore interested in the value of $s$ that minimises the computational effort required to effect one lineage jump. Coalescence can be ignored in this analysis, as there are at most $n - 1$ coalescences, and the number of lineage jumps is much, much larger than the sample size.

Throughout this analysis we assume a moderately sized sample of $n$ lineages on a large torus of diameter $L \gg r$, and this torus is tessellated into pixels of edge $s \ll L$ such that $L/s$ is an integer. We also assume a large neighbourhood size, so $u \ll 1$. These assumptions are reasonable, since they reflect the biological reality that we wish to model. The following lemmas provide us with the key quantities for our analysis.

Lemma 8. Let $z$ be a point chosen uniformly at random on a torus of diameter $L$ tessellated into pixels of edge $s$. Let $\sigma_r(s)$ be the expected number of pixels that a disc of radius $r$ centred at $z$ intersects.

Then,

$$\sigma_r(s) = \frac{s^2 + 4rs + \pi r^2}{s^2}$$

Proof. Let $N = (L/s)^2$, and consider the disc of radius $r$ centred on $z$. The area in which $z$ can fall such that it intersects with a given pixel is $s^2 + 4rs + \pi r^2$, as shown in Fig. 3. Therefore, the probability that the disc intersects with a given pixel is $(s^2 + 4rs + \pi r^2)/(N\pi s^2)$ since the total area of the torus is $N\pi s^2$. Then, summing this probability over all $N$ pixels gives us the expected number of pixels the disc intersects with, and the required result. □

Lemma 9. Let $z$ be a point chosen uniformly at random on a torus of diameter $L$ tessellated into pixels of edge $s$, and let $V$ be the set of pixels that a disc of radius $r$ centred at $z$ intersects with. Then, let $v$ be a pixel chosen uniformly from $V$ and $x$ be a point chosen uniformly from the pixel $v$. Let $\psi_r(s)$ be the probability that $x$ is within the disc $D_r(z)$.

Then,

$$\psi_r(s) = \frac{\pi r^2}{s^2 + 4rs + \pi r^2}$$

Please cite this article in press as: Kelleher, J., et al., Coalescent simulation in continuous space: Algorithms for large neighbourhood size. Theoretical Population Biology (2014), http://dx.doi.org/10.1016/j.tpb.2014.05.001
The expected number of covered pixels is $\sigma(s)$, and therefore the area of these pixels is $s^2 \sigma(s)$. The area of the disc is $\pi r^2$, and therefore the probability that a point falling uniformly at random in a covered pixel is within the disc is $\geq \pi r^2/(s^2 \sigma(s))$ by Jensen’s inequality.

We analyse Algorithm $N$ by counting the approximate number of basic operations required per lineage jump over a large number of jumps. This gives us an expression $T_N(s)$ that should be proportional to the time required by an implementation to simulate a single jump when averaged over a large number of lineage jumps.

Lemma 10. Over a large number of lineage jumps, the expected computational cost of a single jump in Algorithm $N$ is

$$T_N(s, n) \propto n \sigma(s) \left( 2\sigma(s) + \frac{1}{\psi(s)} \right).$$

Proof. We proceed by considering the expected contribution of each algorithm step when averaged over a large number of lineage jumps. Over long timescales, lineages are uniformly distributed over the torus, regardless of the initial state of the sample (Barton et al., 2010a, Lemma 6.9). Therefore, the probability that two lineages occupy the same pixel in a given event is negligible and we therefore know that the expectation of the maximum occupancy $h^*$ is equal to 1.

Step N2 is executed once for each jump, and since $\mathbb{E}[h^*] = 1$, the total contribution of this step is constant.

Step N3 loops a certain number of times, and involves non-constant time operations. Since $\mathbb{E}[h^*] = 1$, we know that $\mathbb{E}[|Q_n|] = n \sigma(s)$, since there are $n$ lineages covering $\sigma(s)$ pixels each. Assuming a balanced tree data structure (Knuth, 1998, Section 6.2.3) for $Q_n$, each pass through N3 requires $\log_2(n \sigma(s))$ time. Since the expected occupancy of a pixel is 1, we therefore have $|S| = 1$ with probability $\psi(s)$, and consequently expect to repeat this step 1/$\psi(s)$ times for each lineage jump. Therefore, we have a contribution of $\log_2(n \sigma(s))/\psi(s)$ per jump for this step.

Upon reaching N4, the expected value of $|S| = 1$, and this step therefore requires a constant amount of time per lineage jump. In step N5 we remove each lineage in $C$ from the $\sigma(s)$ pixels that it covers. Since $\mathbb{E}[h^*] = 1$, the time required to update $P_{\sigma}$ is constant, and the time required to update $Q_{\sigma}$ is $\log_2(n \sigma(s))$ as before. Then, since $\mathbb{E}[|C|] = 1$, the contribution of this step is $\sigma(s) \log_2(n \sigma(s))$. Similarly, step N6 adds the parental lineage to $\sigma(s)$ pixels, and therefore also requires $\sigma(s) \log_2(n \sigma(s))$ time.

Step N7 is executed at most $n - 1$ times, and requires a constant number of operations. It does not contribute significantly to the running time over a large number of jumps.

Summing the significant contributions from the steps above gives us (5), as required.

Lemma 10 provides us with a prediction for the expected time required to perform a single lineage jump in Algorithm $N$ in terms of the sample size $n$. It is more useful, however, to consider the limit of large sample sizes, and to derive a prediction that is independent of $n$. In general, we expect $s \gg r$, and so $\sigma(s)$ is between 1 and 10; hence, for realistic sample sizes, the dominant term is $\log_2(n)(2\sigma(s) + 1/\psi(s))$. As a result, the optimal pixel size is approximately independent of $n$ and can be found by minimising $2\sigma(s) + 1/\psi(s)$. Performing this minimisation numerically we find a minimum at $s \approx 2.24r$, providing us with a useful prediction for the optimal pixel size in Algorithm $N$. Fig. 4 shows an excellent agreement between the predicted and observed optimal pixel size over a wide range of sample sizes.

4. Multilocus coalescent

There are two practical approaches to including the effects of recombination in the coalescent algorithm. The first, pioneered by Hudson (1983), is to track the state of $m$-locus individuals and build $m$ genealogies as we proceed backwards in time, generating common ancestor and recombination events and applying the effects to the ancestral population. Recombination events increase the size of the ancestral population, but do not increase the amount of ancestral material; common ancestor events reduce the size of the ancestral population, and potentially reduce the amount of ancestral material present through coalescence. Coalescence occurs when two or more ancestors trace back to a single parent, and both have ancestral material at one or more loci. Each time a coalescence occurs we update the genealogy at the loci involved, and the simulation terminates when all genealogies are complete. This approach is the basis of the classical ms program (Hudson, 2002).

The second approach, introduced by Wiuf and Hein (1999), involves first generating the genealogy for the left-most locus and then moving rightwards along the sequence, generating recombination breakpoints. At each breakpoint the current genealogy is modified, and the algorithm terminates when the rightmost point of the sequence has been reached. Wiuf and Hein’s algorithm is considerably more complicated than Hudson’s, but the time complexity of the methods is similar (Wiuf and Hein, 1999). The approach of working along the genome from left to right, however, has led to an approximation that makes simulating the history of large genomic regions much more efficient. In the sequentially Markov coalescent (McVean and Cardin, 2005), the coalescent with recombination is approximated by assuming that the genealogy at a breakpoint depends only on the immediately previous genealogy. This approximation has been used and extended in many simulation algorithms (Marjoram and Wall, 2006; Chen et al., 2009; Excoffier and Foll, 2011), as the time and space requirements of simulating the history of a sample are greatly reduced by this simplification. We do not use the sequentially Markov coalescent here, as the necessary theory to extend the method to a two-dimensional continuum has not been developed.

In this section we extend the single locus coalescent algorithm of Section 3 by adapting Hudson’s approach to the setting of a spatial continuum. We incorporate recombination by letting each
individual in the sample consist of \(m\) linearly arranged loci, and extending the model such that we have \(v\) parents at each event. At an event, there is a probability \(\rho\) that a recombination event occurs between locus \(\ell\) and \(\ell+1\), and a child derives from different parents at these loci. See Etheridge and Véber (2013) and Barton et al. (2013a) for more details on this model of recombination.

It may seem unusual to allow a child lineage to descend from an arbitrary number of parents \(v\). In events modelling the steady reproduction of individuals in a sexually reproducing species we have \(v = 2\) as one might expect. We must also consider events that model large-scale demographic shifts, however, where several generations may elapse before the local population returns to the stationary distribution. In this scenario, the eventual children of the event may descend from a number of founders, with recombination mixing the contributions of each parent arbitrarily. More sophisticated models of recombination may also be required for these large-scale events, which can be readily incorporated. In the interest of simplicity, however, we assume that all events follow the recombination process outlined above.

In this section we describe Algorithm M, the extension of Algorithm N to simulate the ancestry of \(m\)-locus individuals. The algorithms share the same overall structure, and, besides the process of transferring ancestral material from offspring to parents, do not differ in significant ways. Rather than writing out the process of transferring ancestral material from offspring to parents, with the ancestral material we insert it into the sample using an arbitrary number of parents \(v\). In events modelling the steady reproduction of individuals in a sexually reproducing species we have \(v = 2\) as one might expect. We must also consider events that model large-scale demographic shifts, however, where several generations may elapse before the local population returns to the stationary distribution. In this scenario, the eventual children of the event may descend from a number of founders, with recombination mixing the contributions of each parent arbitrarily.

In this section we describe Algorithm M, the extension of Algorithm N to simulate the ancestry of \(m\)-locus individuals. The algorithms share the same overall structure, and, besides the process of transferring ancestral material from offspring to parents, do not differ in significant ways. Rather than writing out the process of transferring ancestral material from offspring to parents, with the ancestral material we insert it into the sample using an arbitrary number of parents \(v\). In events modelling the steady reproduction of individuals in a sexually reproducing species we have \(v = 2\) as one might expect. We must also consider events that model large-scale demographic shifts, however, where several generations may elapse before the local population returns to the stationary distribution. In this scenario, the eventual children of the event may descend from a number of founders, with recombination mixing the contributions of each parent arbitrarily.

Algorithm M operates in the same manner as Algorithm M, where we have \(n\) individuals sampled on the torus and simulate until the history of the sample is complete. Each locus \(\ell\) has an oriented tree \(\pi,\tau\) and list of node times \(\tau_i\). We also have a sequence \(\eta\) such that \(\eta_i\) is the next available node in the oriented tree \(\pi\). The state of the ancestral population is most simply represented as set of tuples \((\alpha, a_1, \ldots, a_m)\), where \(\alpha\) is the location of an individual and the sequence \(a_1, \ldots, a_m\) records the node the individual occupies in the genealogy at each locus. If \(a_i = 0\) (i.e., the null node), then there is no ancestral material present in this individual at locus \(\ell\), and so we are not interested in tracking its history. Termination of the algorithm is controlled by letting \(k\) track the total amount of ancestral material in the sample; initially, \(k = nm\), and we know that all loci have coalesced when \(k = n\).

The same method of spatial indexing via the pixels \(P\) and occupancy \(Q\) applies in Algorithm M, and so steps N2–N5 are essentially unchanged. Once we have selected the set of children, however, we must decide how the ancestral material of these individuals is distributed among the parents of the event, taking into account the effects of possible recombination and coalescence. This process is described in detail in Algorithm G. Afterwards, we have \(v\) parents with the ancestral material from the children shared among them via their node mappings, and for each parent that has ancestral material we insert it into the sample using a similar method to N6. Parents without ancestral material are discarded.

For notational convenience, we let \(a\) be a matrix of node assignments for the children of an event, such that \(a_{j,\ell}\) is the node that the \(j\)th child occupies in the genealogy at locus \(\ell\). Similarly, we let \(p\) be the matrix of node assignments for the parents such that \(p_{k,\ell}\) is the node that the \(k\)th parent occupies in the genealogy at locus \(\ell\). Initially, parents have nonancestral material at every locus (and so \(p_{k,\ell} = 0\) for all \(k\) and \(\ell\)). Then, as the node assignments in \(a\) are distributed among the parents, taking into account the effects of coalescence and recombination, some of these loci are assigned ancestral material.

Algorithm G (Generate Parents), Given the matrix of node mappings of \(n\) child nodes \(a\), generate the node mappings of \(v\) parental individuals \(p\), and update \(\pi,\tau,\eta\) and \(k\) to record coalescence events. Recombination occurs with probability \(\rho\) for each locus between adjacent parents.

G1. [Initialisation.] Set \(p_{k,\ell} \leftarrow 0\) for \(1 \leq j \leq v\) and \(1 \leq \ell \leq m\). Then set \(c_\ell \leftarrow 0\) for \(1 \leq \ell \leq m\), and \(j \leftarrow 1\).

G2. [Choose first parent] Set \(k \leftarrow R_1(\{1, \ldots, v\})\) and set \(\ell \leftarrow 1\).

G3. [Ancestral?] Set \(\alpha \leftarrow a_{k,\ell}\). If \(\alpha = 0\) go to G7.

G4. [Inheritance.] If \(p_{k,\ell} = 0\), set \(p_{k,\ell} \leftarrow \alpha\) and go to G7. Otherwise, if \(c_\ell \neq 0\), go to G6.

G5. [Single coalescence] Set \(c_\ell \leftarrow 1\), \(\beta \leftarrow p_{k,\ell}\) and \(\eta \leftarrow \eta + 1\). Then set \(\pi,\tau,\eta \leftarrow \gamma, \tau_j,\gamma \leftarrow \tau_j,\gamma \leftarrow \gamma + 1\). 

G6. [Multiple coalescence.] Set \(\pi,\alpha \leftarrow \eta - 1\) and \(k \leftarrow k - 1\).

G7. [Next locus.] If \(R_2(\{0,1\})\), set \(r \leftarrow R_2(\{1, \ldots, v\} \setminus \{k\})\). Then set \(\ell \leftarrow \ell + 1\) and if \(\ell \leq m\) go to G3.

G8. [Next child.] Set \(j \leftarrow j + 1\), and if \(j \leq n\) go to G2.

Algorithm G proceeds by considering each child \(j\) and each locus \(\ell\) in turn. If child \(j\) has ancestral material at locus \(\ell\) we transfer this ancestral material to a parent. Each time such an assignment is made we test for coalescence, which occurs when more than one child descends from a particular parent at a given locus, and update the data structures to record this event.

For each child \(j\) we first choose a parent uniformly at random for the first locus in step G2. We then consider each locus \(1 \leq \ell \leq m\) in turn; if \(p_{k,\ell} = 0\) then there is ancestral material and we proceed on to G4. In G4 we test to see if there has already been ancestral material assigned to this parent \(k\) at locus \(\ell\). If this is the case, then a coalescence has occurred at this locus, and we must record this fact. Because we can have more than two children in an event that have ancestral material at a given locus, there may be several children descending from a given parent. The first time we encounter a coalescence at a locus, we update \(\pi,\tau,\eta\) to register this event in G5. For subsequent coalescence events at this locus we skip directly to G6, where we record that the parent of node \(\alpha\) at locus \(\ell\) is \(\eta - 1\) and decrement \(k\), accounting for the loss of one piece of ancestral material. Recombination between adjacent loci occurs in step G7, where we choose a new parent for locus \(\ell + 1\) with probability \(\rho\).

Algorithm G is a simple and effective method of transferring ancestral material from children to parents for small numbers of loci \((m < 10, \text{say})\). It requires only one pass through the set of children and transfers ancestral material directly from children to parents, without requiring any intermediate data structures. It is, however, extremely inefficient for large \(m\), wasting large amounts of time and space by storing and iterating over the large tracts of non-ancestral material that enter the sample as we progress backwards in time.

This inefficiency can be easily resolved by changing the representation of the node mappings for an individual from the ‘dense’ sequence \(a_1, \ldots, a_m\) to a sequence of pairs \((\ell, \alpha)\) mapping loci to nonzero nodes. The first benefit of this ‘sparse’ representation is that the amount of memory required to store the state of the ancestral population is far smaller. Using the dense representation the amount of memory required grows with the size of the ancestral population. Since each individual requires \(O(m)\) space to store its node mappings and the ancestral population can grow up to a limit of \(mn\), we require \(O(nm^2)\) space. For large \(m\), this is prohibitive. The sparse representation, on the other hand, requires \(O(nm)\) space, since the amount of ancestral material in the sample is nonincreasing as we proceed backwards in time.
This approach may seem wasteful, however, when modelling lower levels of recombination, where we expect to see many adjacent loci sharing the same ancestry. In this case, storing the ancestry for each locus in each individual and storing a tree for each locus is unnecessary. A more compact approach is to store a tuple holding the first and the last locus of each segment, and to only store trees for each unique genealogy. Unfortunately, this method has some major pitfalls. Without the use of specialised data structures, maintaining the segments and calculating overlaps becomes a serious burden and negates any advantages of a more compact representation.

The ms program (Hudson, 2002) uses this segment approach. It maintains a list of segments representing the distinct genealogies that have been created through recombination, and each ancestor maintains a list of segments mapping their ancestral material to these trees. Each recombination event results in a new segment being created, and each common ancestor event conducts a linear scan of these segments to detect overlaps between the segments the individuals carry and the segments defining the extant trees.

The expected number of recombination events in a simulation of the standard coalescent with a sample of size $n$ is $R_{\text{H}} - 1$ (Hudson and Kaplan, 1985), where $H_n$ is the $n$th Harmonic number and $R = 4N_e(m - 1)$ is the scaled recombination rate. ($N_e$ is the effective population size, $m$ the number of loci and $\rho$ the between-locus recombination probability.) Since $H_n \approx \ln(n + y)$, where $y$ is the Euler–Mascheroni constant, the expected number of segments is approximately $R$ for large $R$ and small $n$.

Therefore, in the worst case, each common ancestor event in ms costs $O(R)$ time. For large $R$, this represents an extremely heavy cost. If we follow the example of Chen et al. (2009) and assume $N_e = 12,500$ and $\rho = 1.2 \times 10^{-8}$, we have a scaled recombination rate of $R = 12,000$ for a 20 Mb region. Profiling ms under these parameters with a sample of two individuals reveals that over 90% of the simulation time is spent performing common ancestor events, compared to around 2% of the time executing recombination events. Since the number of these events is similar (otherwise the simulation would quickly end, or never finish), this demonstrates a major flaw in ms, at least in the case of large $R$ and $m$.

A thorough analysis of the different approaches to maintaining ancestry in coalescent simulations is beyond the scope of this article. The comparison with ms is intended to illustrate that, although our approach is simple, it is effective when the expected amount of ancestral material per ancestor is small. For low levels of recombination, the ms approach is very efficient. It is not clear what type of approach is suitable for large genomic regions with intermediate levels of recombination. This is an important question for the efficiency of coalescent simulation.

Ideally, we would like to extend the analysis of Section 3.3 and derive the optimal pixel size $s$ for a given neighbourhood size and recombination rate. Such an analysis, however, would require a detailed understanding of the spatial distribution of ancestors and the ancestral material they carry. Since very little is currently known about these distributions, we must defer a full analysis of the multilocus algorithm to future work. Before this analysis is performed, we can make some basic recommendations about the choice of pixel size for multilocus simulations. Firstly, the optimal value for $s$ derived for the single locus case in Section 3.3 is certainly an upper bound on the value of $s$ that should be used for multilocus simulations. Any simulations in which we expect many ancestors to be in close proximity for extended periods of time should have $s < 2.24r$. On the other extreme, for $s < 1$ memory requirements for maintaining the spatial indexes increase sharply. In this case, any potential advantages of faster simulation may be overwhelmed by excessive memory usage, preventing replication over available CPU resources.
5. Conclusion

A central goal of the spatial Λ-Fleming–Viot continuum model is to provide a means of incorporating events over different scales, ranging from the steady process of local reproduction to large-scale demographic shifts. Without such events, we cannot explain the patterns that we observe in nature over large spatial scales since the process of diffusion is too slow (Barton et al., 2010b). The continuum model incorporates these fluctuations in a flexible and elegant manner, by allowing events of different radii and impacts to occur at different rates. For example, the effect of rare large events modelling demographic shifts along with frequent small events modelling reproduction is analysed in detail by Barton et al. (2010a).

In the interest of simplicity we have given detailed algorithm listings for a single class of event occurring at a fixed rate, and outlined the changes required to generalise to an arbitrary number of event classes from the disc model. Incorporating events from different models, however, is not so straightforward, as our methods here depend entirely on the geometry of the disc model. Including events from the Gaussian replacement model (Barton et al., 2010b), for example, would require a different approach. This is not a major drawback, however. It is only the most frequent reproduction events that must be from the disc model; rarer events can have any geometry that we wish, provided we follow the method outlined in Section 3.1 for incorporating large-scale events. Since these events are rare with respect to regular reproduction, there is little point in conditioning on the events affecting individuals and is in fact more efficient to simulate their effects directly.

We have discussed and analysed the algorithms in this article in a great deal more detail than is customary in the literature. Given the centrality of coalescent simulation in modern population genetics, it is important that the algorithms used are both correct and as efficient as possible, and this can only be achieved via detailed algorithm listings and analyses. The consequences of hiding these important details are well illustrated by the ms program. As our brief analysis in Section 4 revealed, ms is very efficient for small numbers of loci, but scales extremely poorly for large genomic regions. However, this fact is not widely known and ms is usually regarded as being very efficient (Carvajal-Rodríguez, 2008; Chen et al., 2009). The performance of ms is in effect taken to be a measure of the inherent difficulty of simulating the coalescent with recombination. Consequently, researchers have abandoned the full coalescent and resorted to various approximations (Li et al., 2007; Padhukasahasram et al., 2008; McVean and Cardin, 2005; Chen et al., 2009) in order to simulate larger genomic regions. An analysis of the algorithm and the use of the appropriate data structures would make simulating the full coalescent many times faster than is possible with ms without the need for approximation.

Substantial efforts have been made in this article to avoid approximation and to ensure that the process that we simulate is precisely equal to the well-defined coalescent process of the continuum model. A sceptical reader might argue that this is wasted effort, since some approximation to a model which is (at best) a cartoon of reality can do little harm. However, this ignores one of the most important applications of stochastic simulations: assessing the accuracy of analytical methods. Without exact simulations, it is very difficult to measure the accuracy of putative analytical approximations; deviations from simulation results may be due to either approximations of the model in simulations or to errors in the approximate calculations. Exact simulations also allow us to be confident that the data we observe are a result of interesting phenomena arising from the underlying model, and not some unexpected consequence of an ad-hoc approximation.

Acknowledgments

We would like to thank Amandine Véber for discussions and comments on the manuscript.

The first author’s work was supported by EPSRC grant EP/I013091/1. The second author’s work was supported in part by EPSRC grant EP/I01361X/1. The third author’s work was supported by European Research Council grant 250152.

Appendix. Calculating identity

Let $A_r(x, z)$ be the area of the intersection of three discs of radius $r$, centred at $(-x/2, 0), (x/2, 0)$ and $z$, and let $h = \sqrt{r^2 - x^2/4}$. If we consider $z$ in the upper half plane, the domain of $A_r(x, z)$ is the union of four sets, indicated by the solid arcs:

\[
\begin{align*}
(0, -h) & \quad (x/2, 0) \\
(x/2, 0) & \quad (x/2, 2h) \\
(0, h) & \quad (-x/2, 0)
\end{align*}
\]

If $z$ is in the region bounded by the $x$-axis and the arc of radius $r$ centred at $(0, -h)$, $A_r(x, z)$ is constant and equal to $A_r(x)$.

Otherwise, if $z$ falls in the region bounded by the arc of radius $r$ centred on $(0, h)$, $A_r(x, z)$ is the area of a circular triangle (Fewell, 2006). Finally, if $z$ falls within the region bounded by the arc of radius $2r$ centred on $(-x/2, 0)$, $A_r(x, z) = A_0(\sqrt{z_1 + x/2}^2 + z_2^2)$; a similar rule applies to the region bounded by the arc centred on $(x/2, 0)$. For all $z$ outside of these regions, $A_r(x, z) = 0$.

Let $F_r(x)$ be the probability of identity in state of two genes sampled at distance $x$, under the infinitely many alleles model with mutation at rate $\mu_1$ (Barton et al., 2010b, 2013a). After rephrasing Eq. (A.8) of (Barton et al., 2013a) in terms of the $A_r$ functions, converting to polar coordinates and simplifying, we obtain

\[
\phi(x)F_r(x) = \frac{2\mu}{\Lambda + 2\mu r^2} - \frac{2\mu A_r(x)}{r}
\]

where $\phi(x) = 2\mu/\Lambda - 2\mu r^2 - u^2 A_r(x)$, the kernel $K(x, y)$ is given by

\[
K(x, y) = f\left(\frac{y}{x}\right)\left(1 - \frac{1}{x^2}A_r(x)\right)
\]

\[
+ \frac{4\mu y}{r^2}\int_0^\pi \int_0^\pi \frac{Q_1(x, y, \theta)}{\cos \theta + \sin \theta} d\theta d\theta,
\]

with

\[
Q_1(x, y, \theta) = A_r(\sqrt{\Lambda - 2xy \cos \theta + y^2})
\]

\[
Q_2(x, y, \theta) = A_r(x - x/2 + y \cos \theta, y \sin \theta)
\]

Here, $f(x)$ is the probability density function of the distance between two points sampled independently and uniformly at random within the unit disc (Alagarsamy, 1976).

\[
f(x) = \frac{4}{\pi}\left(4 \arccos(x/2) - \sqrt{4 - x^2}\right), \quad x \in [0, 2],
\]

with $f(x) = 0$ for $x > 2$. The value of $\phi(x)$ is found by observing that $\int_{x/2}^{\infty} A_r(z) dz = \pi r^2$, and $\int_{x/2}^{\infty} A_r(z) dz = \pi r^2 A_r(x)$.

Eq. (A.2) is a Fredholm equation of the second kind, and can be solved numerically using a number of methods (Atkinson, 1997), with the Nyström method being most convenient. Solving
the equation in this manner requires the repeated evaluation of
\[ \int_0^{\pi} Q_1(x, y, \theta) - uQ_2(x, y, \theta) \, d\theta, \]
which can be time consuming and inaccurate if not done with care. In particular, integrating
\[ Q_2(x, y, \theta) \]
can be troublesome. Performing the integration from 0 to \( \pi \) piecewise, however, as the arc of radius \( y \) centred at \((-x/2, 0)\) intersects with the arcs defining the domain of \( A(x, z) \), as indicated in (A.1), gives us a fast and accurate numerical method.

References