Journal of Computational Physics 229 (2010) 3777–3791

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

A retrodictive stochastic simulation algorithm

T.G. Vaughan [a,*], P.D. Drummond [a], A.J. Drummond [b]

[a] Centre for Atom Optics and Ultrafast Spectroscopy, Swinburne University of Technology, Melbourne VIC 3122, Australia
[b] Department of Computer Science, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

ARTICLE INFO

Article history:
Received 7 August 2009
Received in revised form 4 December 2009
Accepted 20 January 2010
Available online 28 January 2010

Keywords:
Stochastic process
Simulation algorithm
Retrodiction
Genetics

ABSTRACT

In this paper we describe a simple method for inferring the initial states of systems evolving stochastically according to master equations, given knowledge of the final states. This is achieved through the use of a retrodictive stochastic simulation algorithm which complements the usual predictive stochastic simulation approach. We demonstrate the utility of this new algorithm by applying it to example problems, including the derivation of likely ancestral states of a gene sequence given a Markovian model of genetic mutation.

1. Introduction

Models based on continuous-time birth/death Markov (memory-less) processes have been used to describe the stochastic dynamics of a truly vast array of natural systems. Well-known examples include collisional dynamics and the progress of chemical reactions in physics and chemistry [1,2], the dynamics of microscopic or macroscopic populations of organisms and the evolutionary dynamics of gene populations [3]. Such models are most conveniently expressed in terms of master equations which describe the temporal evolution of a probability distribution over the state space of the system. However, explicit solutions to these equations are often difficult or impossible to obtain for state spaces of even quite modest dimension, except in special cases such as the steady-state limit. This difficulty remains despite the fact that the state space is discrete, meaning matrix diagonalization is a possible strategy. The central problem here is the exponential complexity of the equations when there are large numbers of modes, chemical species, genotypes and/or particles.

For this reason, these models are often treated stochastically using an approach originally developed to treat master equations describing chemical reactions [4,5]. This approach, known generally as the stochastic simulation algorithm (SSA), involves directly simulating the time-evolution of a system as a series of discrete transitions at (pseudo-)random times. By averaging over the results of many of these simulations, one can reconstruct the stochastic dynamics of any property of the system to any precision one desires. However, the primary advantage of the SSA is that it allows one to significantly reduce the computational complexity of the problem by sacrificing unnecessary precision. In this approach the precision is determined simply by the random sampling error. This can be reduced to any required level by increasing the number of time-evolved samples or "trajectories" used for averaging.

Unfortunately, however, the efficiency gains afforded by the SSA do not readily allow one to obtain answers to *all* questions which may be asked of birth/death Markov processes. In particular, performing Bayesian inference on previous states of the system given information about the current state is at best difficult to achieve using the standard forward-time SSA and is usually practically impossible. Such inference is often desirable in the context of biological evolutionary systems, where

* Corresponding author. Tel.: +61 3 9214 8465; fax: +61 3 9214 5160.
E-mail address: tvaughan@swin.edu.au (T.G. Vaughan).

0021-9991/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2010.01.027

one might only have direct access to samples from the present population. It is also necessary in the context of studying the genetic dynamics of within-host infection where it is usually impossible to obtain samples of the microbial population immediately following infection, meaning that such information must be inferred from later samples (see, for example, [6]). Similar problems of efficiency arise when tackling first-passage time problems that arise frequently in many applications, from physics to economics.

In this paper, we present an effective systematic means of performing such inference. We firstly demonstrate that recent work in the context of backwards-time inference on quantum dynamical systems [7,8] is equally applicable to classical birth/death Markov systems. In particular, we find that the "retrodictive" quantum master equation (so named in contrast to the more usual "predictive" quantum master equation) has a classical analogue which is similar to the backwards master equation commonly used to solve first-passage problems. We then go on to develop a "retrodictive" stochastic simulation algorithm (RSSA) capable of providing exact numerical solutions to either type of time-reversed master equation. Finally, we demonstrate the use of this algorithm in inferring previous states of various example continuous-time Markov processes. Our examples include basic birth/death processes, the continuous-time Moran model in genetics, and a simple model of genetic mutation in a population of haploid organisms.

The central result of this paper is the development of an efficient algorithm for inferring which states a system subject to a Markovian birth/death process has occupied at earlier times, given an observed final state.

2. Retrodictive master equations

Traditional master equations describe the Markovian dynamics of a conditional probability for the current state $\vec{n}$ of a system at time t, given some initial state $\vec{n}_i$ at an earlier time $t_i < t$. Here the state $\vec{n}$ will be regarded as discrete, and described by a vector of integers. For a set of independent time-homogeneous Markov processes, the master equation can be written in the following general form:

$$\frac{\partial}{\partial t}P(\vec{n},t|\vec{n}_i,t_i) = \sum_k [T_k(\vec{n}-\vec{v}_k)P(\vec{n}-\vec{v}_k,t|\vec{n}_i,t_i) - T_k(\vec{n})P(\vec{n},t|\vec{n}_i,t_i)], \tag{1}$$

where $T_k(\vec{n})dt$ is the probability that at a system in state $\vec{n}$ will undergo process k and be transferred to the state $\vec{n}+\vec{v}_k$ within the interval dt. (In the special case of a chemical master equation, the $T_k(\vec{n})$ are the usual combinatoric factors and the vectors $\vec{v}_k$ are columns of the stoichiometric matrix for the reacting system.)

The above conditional probability can be interpreted either using the so-called "objective Bayesian" viewpoint [9–11] or the frequency interpretation. In the first of these, the probability represents a quantitative measure of the degree to which the fact that the system is known to occupy the initial state $\vec{n}_i$ logically implies that it will occupy a certain other state $\vec{n}$ at some later time. This interpretation, which is validated by Cox's theorem [10,11], is functionally equivalent, in this context, to the interpretation that the probability represents the limit of the relative frequency with which this event occurs in an infinitely large ensemble of identical experimental trials.

As logical implication is not constrained to act in the direction of physical causation, we can sensibly ask with what probability a given *final* state of the system logically implies the system was in a particular state at some *initial* time, given what we know of the forward-time dynamics of the system. This question is answered directly through a simple application of the rules of probability theory, which yield a special form of Bayes' theorem:

$$P(\vec{n}_i,t_i|\vec{n}_f,t_f) = \frac{P(\vec{n}_f,t_f|\vec{n}_i,t_i)P(\vec{n}_i,t_i)}{\sum_{\vec{m}}P(\vec{n}_f,t_f|\vec{m},t_i)P(\vec{m},t_i)}. \tag{2}$$

Here $\vec{n}_f$ is the state of the system at some final time $t_f > t_i$, $P(\vec{n}_f,t_f|\vec{n}_i,t_i)$ are the conditional probabilities provided by the stochastic model through the master equation and $P(\vec{n}_i,t_i)$ is the *a priori* probability distribution over the initial states. (The *a priori* distribution contains the state of knowledge of states of the system in the absence of the measurement of $\vec{n}_f$.)

The traditional master equation is clearly well-suited to the direct calculation of the predictive conditional probabilities $P(\vec{n}_f,t_f|\vec{n}_i,t_i)$ corresponding to a single $\vec{n}_i$, as this can be posed as a straight-forward initial value problem. If an analytical solution to this problem exists, one can apply Eq. (2) to obtain the retrodictive conditional probabilities $P(\vec{n}_i,t_i|\vec{n}_f,t_f)$ directly from the predictive master equation. However, as the majority of systems of practical interest are described by master equations lacking known analytical solutions, one must resort to some form of numerical integration of the initial value problem. In such cases, using the forward equation Eq. (2) to obtain the retrodictive probabilities is exponentially more complex than obtaining the predictive probabilities, as the former necessitates numerically evaluating $P(\vec{n}_f,t_f|\vec{n}_i,t_i)$ for every possible initial state $\vec{n}_i$.

Fortunately, there is a more direct approach. We consider the Chapman–Kolmogorov equation

$$P(\vec{n}_f,t_f|\vec{n}_i,t_i) = \sum_{\vec{n}}P(\vec{n}_f,t_f|\vec{n},t)P(\vec{n},t|\vec{n}_i,t_i), \tag{3}$$

where $t \in (t_i,t_f)$, which is simply a precise statement of the Markovian nature of the dynamics. Taking a derivative with respect to the interim time t yields

$$\sum_{\vec{n}}\frac{\partial}{\partial t}P(\vec{n}_f,t_f|\vec{n},t)P(\vec{n},t|\vec{n}_i,t_i) = -\sum_{\vec{n}}P(\vec{n}_f,t_f|\vec{n},t)\times\frac{\partial}{\partial t}P(\vec{n},t|\vec{n}_i,t_i) \tag{4}$$

which, after incorporating Eq. (1) and performing a variable substitution, reduces to the well-known backward form of the master equation,

$$\frac{\partial}{\partial t}P(\vec{n}_f, t_f|\vec{n}, t) = \sum_k T_k(\vec{n})[P(\vec{n}_f, t_f|\vec{n}, t) - P(\vec{n}_f, t_f|\vec{n} + \vec{v}_k, t)], \tag{5}$$

where $t < t_f$. This form, known also as the *Kolmogorov backward equation*, has various applications including finding solutions to first-passage problems and systems with absorbing boundaries [1]. In our context, however, it provides a direct route to inferring past states of the system by allowing one to calculate all of the necessary predictive conditional probabilities in Eq. (2) by solving a single initial value problem.

To emphasise its connection to retrodictive inference, we note that the backward master equation can be regarded as the equation of motion for the *likelihood* [11] $L(\vec{n}, \tau|\vec{n}_f) \equiv P(\vec{n}_f, t_f|\vec{n}, t_f - \tau)$ for an earlier state given a final state:

$$\frac{\partial}{\partial \tau}L(\vec{n}, \tau|\vec{n}_f) = \sum_k T_k(\vec{n})[L(\vec{n} + \vec{v}_k, \tau|\vec{n}_f) - L(\vec{n}, \tau|\vec{n}_f)] \tag{6}$$

which we have expressed here in terms of the reversed time $\tau \equiv t_f - t$. The difference between this likelihood and the predictive conditional probability is simply that the latter is considered a normalised distribution over $\vec{n}_f$ parameterised by $\vec{n}$, while the former is considered an *unnormalised* distribution over $\vec{n}$ parameterised by $\vec{n}_f$.

In analogy to the retrodictive quantum mechanical density operator [7,8], we now define the normalised distribution

$$R(\vec{n}, \tau|\vec{n}_f) = \frac{L(\vec{n}, \tau|\vec{n}_f)}{\sum_{\vec{m}} L(\vec{m}, \tau|\vec{n}_f)} = \frac{P(\vec{n}_f, t_f, |\vec{n}, t_f - \tau)}{\sum_{\vec{m}} P(\vec{n}_f, t_f|\vec{m}, t_f - \tau)} \tag{7}$$

which is related to the retrodictive conditional probability through Eq. (2) in the same way as the likelihood. That is,

$$P(\vec{n}, t_f - \tau|\vec{n}_f, t_f) = \frac{R(\vec{n}, \tau|\vec{n}_f)P(\vec{n}, t_f - \tau)}{\sum_{\vec{m}} R(\vec{m}, \tau|\vec{n}_f)P(\vec{m}, t_f - \tau)}. \tag{8}$$

In contrast to the likelihood however, $R(\vec{n}, \tau|\vec{n}_f)$ is precisely equivalent to the retrodictive conditional probability in the special case that $P(\vec{n}, t_f - \tau)$ is uniform (i.e. constant), which is often appropriate in the absence of relevant *a priori* information.

Combining its definition in Eq. (7) with Eq. (6), we find that the normalised likelihood distribution satisfies the following equation of motion

$$\frac{\partial}{\partial \tau}R(\vec{n}, \tau|\vec{n}_f) = \sum_k T(\vec{n}_k)\left[R(\vec{n} + \vec{v}_k, \tau|\vec{n}_f) - R(\vec{n}, \tau|\vec{n}_f)\right] - \frac{\partial \Omega(\tau)}{\partial \tau}R(\vec{n}, \tau|\vec{n}_f), \tag{9}$$

where

$$\frac{\partial \Omega(\tau)}{\partial \tau} = \sum_{\vec{n}} \sum_k T(\vec{n}_k)[R(\vec{n} + \vec{v}_k, \tau|\vec{n}_f) - R(\vec{n}, \tau|\vec{n}_f)] \tag{10}$$

and $R(\vec{n}, \tau|\vec{n}_f)$ is subject to the initial condition $R(\vec{n}, 0|\vec{n}_f) = \delta_{\vec{n},\vec{n}_f}$.

This equation will be referred to from here on as the *retrodictive master equation* (RME), it being the classical analogue of the retrodictive quantum master equation considered by Pegg et al. in their treatment of open quantum systems [8]. In the same way that the solutions to the traditional master equation yield predictive conditional probabilities for later states given earlier states, solutions to the RME yield the retrodictive conditional probabilities for earlier states given later states either directly (in the case of a uniform *a priori* distribution) or through Eq. (8) above.

2.1. Retrodictive master equations and the inverse problem

In terms of the frequency-based definition of probability, the most natural interpretation of the master equation given as Eq. (1) is that it describes how the frequency distribution over states of a large ensemble of identically conditioned but statistically independent experiments can be expected to evolve. For such an ensemble with an initial frequency distribution $f(\vec{n}, t_i)$ at the initial time t_i, the frequency distribution at some later time t_f can be written in terms of the convolution

$$f(\vec{n}_f, t_f) = \sum_{\vec{n}_i} P(\vec{n}_f, t_f|\vec{n}_i, t_i)f(\vec{n}_i, t_i), \tag{11}$$

where $P(\vec{n}_f, t_f|\vec{n}_i, t_i)$ are the conditional probabilities provided by the master equation.

It therefore appears from this perspective that the general backward time problem to solve is that of recovering $f(\vec{n}_i, t_i)$ given knowledge of the later frequency distribution $f(\vec{n}_f, t_f)$. This is an inverse problem similar in form to the deconvolution problem of object reconstruction in the fields of image processing, geophysical analysis, inverse scattering theory, etc., and is similarly ill-posed due to the general one-to-many mapping of the matrix elements $P(\vec{n}_f, t_f|\vec{n}_i, t_i)$—meaning that its solution necessitates the inversion of a singular matrix.

Interestingly however, the problem in this precise mathematical form is rarely one that actually needs to be addressed, as in practice one can only ever obtain a finite number of samples of the frequency distribution $f(\vec{n}_f, t_f)$, which consequently can never be precisely known. Here we focus on the simplest problem of this type, which is the problem of identifying the probability that given a *single* member of the hypothetical ensemble, the system occupied a particular state at some earlier time given its observed final state.

This problem—which is the subject of the current paper—can be dealt with unambiguously using modern 'Bayesian' probability techniques, without recourse to other more complex methods (such as Tikhonov regularization [12]) usually invoked to deal with the singular nature of the former. A comparison between the two approaches to solving inverse problems exists elsewhere [13] and so we will not discuss this further.

3. Retrodiction via stochastic simulation

While in principle one can use the RME directly to determine the probability distribution over past states for any system subject to a birth/death process, this is usually at best difficult in practice due to the fact that the possible number of configurations (the size of the state space) can be huge, even for systems composed of a moderate number of individuals. For example, a system composed of five sub-populations each containing a maximum of just 99 individuals can be in any one of 10^{10} states. For larger systems, direct numerical integration of the backwards master equation (5), or its normalised equivalent (9), quickly becomes impossible due to the prohibitively large computational time and/or memory requirements.

Fortunately, there is an alternative. Just as stochastic simulations algorithms permit the calculation of moments of random variables subject to birth/death processes without necessitating explicit integration of the master equation, the *retrodictive* stochastic simulation algorithm (RSSA) presented here allows one to extract detailed information about the probability distribution over previous states using random sampling methods.

To understand the RSSA, we first summarise the approach of the usual forward-time SSA, then derive an appropriate retrodictive form.

3.1. Predictive SSA

The forward-time 'predictive' SSA involves generating stochastic trajectories (random walks) through the system state space such that the relative frequency for an ensemble of these trajectories at any time $t > t_i$ approaches $P(\vec{n}, t|\vec{n}_i, t_i)$, the solution to the master equation. In order to determine the stochastic trajectory motion which results in the required convergence, we will use a slightly different line of reasoning to that found in Gillespie's original paper [5], as this better facilitates the explanation of our retrodictive algorithm.

Firstly, consider that for a short-time increment $\Delta \ll 1$ the solution to Eq. (1) can be written

$$\begin{aligned} P(\vec{n}', t+\Delta|\vec{n}, t) &\simeq P(\vec{n}', t|\vec{n}, t) + \Delta \sum_k [T_k(\vec{n}' - \vec{v}_k)P(\vec{n}' - \vec{v}_k, t|\vec{n}, t) - T_k(\vec{n}')P(\vec{n}', t|\vec{n}, t)] \\ &\simeq \delta_{\vec{n}',\vec{n}} \exp[-\Delta \sum_k T_k(\vec{n})] + \Delta \sum_k \delta_{\vec{n}',(\vec{n}+\vec{v}_k)} T_k(\vec{n}). \end{aligned} \tag{12}$$

Secondly, note that due to the Markovian nature of the dynamics, the predictive conditional probability $P(\vec{n}_f, t_f|\vec{n}_i, t_i)$ can be expanded in terms of these short-time solutions through repeated application of the Chapman–Kolmogorov equation:

$$P(\vec{n}_f, t_f|\vec{n}_i, t_i) = \sum_{\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_{N-1}} \prod_{j=0}^{N-1} P(\vec{s}_{j+1}, t_i + (j+1)\Delta|\vec{s}_j, t_i + j\Delta), \tag{13}$$

where $\vec{s}_0 = \vec{n}_i$, $\vec{s}_N = \vec{n}_f$ and N is defined so that $t_f = t_i + N\Delta$. By incorporating Eq. (12) and re-expressing each configuration of intermediate states $\{\vec{s}_0, \vec{s}_1, \ldots, \vec{s}_N\}$ in terms of a sequence of runs of M states $\{\vec{m}_1, \ldots, \vec{m}_M\}$ and the associated run lengths $\vec{\delta t} = \{\delta t_1, \ldots, \delta t_M\}$ and taking the limit as $\Delta \to 0$, this becomes

$$P(\vec{n}_f, t_f|\vec{n}_i, t_i) = \sum_M \sum_{\vec{m}_2, \ldots, \vec{m}_M} \int_V d^M \vec{\delta t} e^{-\delta t_M T_0(\vec{m}_M)} \prod_{j=1}^{M-1} [e^{-\delta t_j T_0(\vec{m}_j)} \sum_k \delta_{\vec{m}_{j+1},(\vec{m}_j+\vec{v}_k)} T_k(\vec{m}_j)], \tag{14}$$

where $T_0(\vec{m}_j) = \sum_k T_k(\vec{m}_j)$ and the integration volume V is understood to be the simplex defined by the constraint $\sum_{j=1}^{M} \delta t_j = t_f - t_i$.

Finally, noting that only those configurations composed of states which satisfy $\vec{m}_{j+1} = \vec{m}_j + \vec{v}_k$ for some reaction k actually contribute to the above sum, we sum instead over the connecting reactions defined by the elements of the vector $\vec{k} = \{k_1, \ldots, k_{M-1}\}$ to find the following expression for the predictive conditional probability:

$$P(\vec{n}_f, t_f|\vec{n}_i, t_i) = \sum_{M,\vec{k}} \delta_{\vec{m}_M,\vec{n}_f} \int_V d^M \vec{\delta t} e^{-\delta t_M T(\vec{m}_M)} \prod_{j=1}^{M-1} e^{-\delta t_j T_0(\vec{m}_j)} T_{k_j}(\vec{m}_j), \tag{15}$$

where for a given $\vec{k}$ we have defined $\vec{m}_1 = \vec{n}_i$ and $\vec{m}_j = \vec{n}_i + \sum_{l=1}^{j-1} \vec{v}_{k_l}$ for $j > 1$. This is essentially the path summation (cf. path integral) form of the master equation used recently by Sun [14], which expresses the conditional probability $P(\vec{n}_f, t_f|\vec{n}_i, t_i)$ as a sum of the probabilities of each of the possible paths the system can take between the initial and final states.

One can therefore stochastically sample this probability distribution by assembling trajectories[1] (specified by the time interval and reaction sequences $\vec{\delta t}$ and $\vec{k}$) according to the probability densities given by

$$p_{\text{path}}(M, \vec{\delta t}, \vec{k}) = e^{-\delta t_M T_0(\vec{m}_M)} \prod_{j=1}^{M-1} p(\delta t_j, k_j | \vec{m}_j) \tag{16}$$

(where M is fixed by the sequence lengths) and recording the final state of each trajectory so generated. Here we have defined

$$p(\delta t, k | \vec{n}) = e^{-\delta t T_0(\vec{n})} T_k(\vec{n}), \tag{17}$$

which is Gillespie's *reaction probability density function* and specifies the joint probability density that a simulated system will remain in state $\vec{n}$ for a time δt and then be subject to reaction k. The extra exponential factor in Eq. (16) specifies the probability for the final time increment, after which the system is sampled and no reaction occurs.

Generating an SSA trajectory then simply involves taking the initial system state and time pair $(\vec{n}_i, t_i)$ and iteratively modifying it by randomly choosing time increments and processes according to Eq. (17) until the time exceeds t_f. This process will yield a particular trajectory with the probability density $p_{\text{path}}(M, \vec{\delta t}, \vec{k})$, meaning that the relative frequency with which states $\vec{n}_f$ at the later time t_f are occupied by the members of an ensemble of such trajectories must converge to $P(\vec{n}_f, t_f | \vec{n}_i, t_i)$.

3.2. Retrodictive SSA

We now present a backward time 'retrodictive' stochastic simulation algorithm (RSSA), which involves randomly generating trajectories backwards in time in such a way that the relative frequency distribution of an ensemble of these trajectories over states converges to the solution to the retrodictive master equation given in Eq. (9).

Formulating such an algorithm is not so straight-forward as in the predictive case, however, due to the fact that the RME is in general non-linear. We therefore use a slightly different approach to deriving the appropriate motion for the RSSA trajectories, which involves firstly considering the evolution of the likelihood $L(\vec{n}, \tau | \vec{n}_f)$ given by Eq. (6), which although linear does not in general preserve the normalisation of the distribution.

Just as in the previous section, we begin by calculating the short-time solution, which is in this case

$$L(\vec{n}', \Delta | \vec{n}) \simeq \delta_{\vec{n}', \vec{n}} \exp[-\Delta T_0(\vec{n})] + \Delta \sum_k \delta_{\vec{n}', (\vec{n} - \vec{\nu}_k)} T_k(\vec{n} - \vec{\nu}_k), \tag{18}$$

where as before $T_0(\vec{n}) = \sum_k T_k(\vec{n})$. As the likelihood is functionally equivalent to $P(\vec{n}_f, t_f | \vec{n}, t_f - \tau)$, it can be expanded in terms of the short-time solutions in the same way as the predictive conditional probability. That is,

$$L(\vec{n}_i, \tau_i | \vec{n}_f) = \sum_{\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_{N-1}} \prod_{j=0}^{N-1} L(\vec{s}_{j+1}, \Delta | \vec{s}_j), \tag{19}$$

where $\tau_i = t_f - t_i$, $\vec{s}_0 = \vec{n}_f$, $\vec{s}_N = \vec{n}_i$ and N is defined so that $\tau_i = N\Delta$. Upon incorporation of the short term solutions, taking the limit as $\Delta \to 0$ and following the same procedure as that used in the predictive case, one obtains

$$L(\vec{n}_i, \tau_i | \vec{n}_f) = \sum_{M, \vec{k}} \delta_{\vec{m}_M, \vec{n}_i} \int_{\mathcal{V}} d^M \vec{\delta\tau} e^{-\delta \tau_M T_0(\vec{m}_M)} \prod_{j=1}^{M-1} e^{-\delta \tau_j T_0(\vec{m}_j)} T_{k_j}(\vec{m}_j - \vec{\nu}_{k_j}). \tag{20}$$

Here we have again expressed the intermediate state configurations in terms of length M sequences of runs of states $\{\vec{m}_1, \ldots, \vec{m}_M\}$ and associated run lengths $\vec{\delta\tau} = \{\delta\tau_1, \ldots, \delta\tau_M\}$ and again we have restricted these intermediate states to those which actually contribute to the summation, which in this case are those satisfying $\vec{m}_{j+1} = \vec{m}_j - \vec{\nu}_{k_j}$ with $\vec{m}_1 = \vec{n}_f$ for some reaction sequence $\vec{k} = \{k_1, \ldots, k_{M-1}\}$. The integration volume $\mathcal{V}$ is the simplex defined by the constraint $\sum_{j=1}^{M} \delta\tau_j = \tau_i$.

Besides the obvious reversal of the reaction directions, the key difference between the path summation expansion for the likelihood and that of the predictive conditional probability given in Eq. (15) is that while the latter is an expansion of a true probability and the expansion coefficients can therefore be regarded directly as path probabilities, the former is not strictly a probability and cannot be treated as such due to the fact that its evolution does not preserve the normalization of the distribution. We therefore factorise the integration/summation kernel of Eq. (20) in the following manner

$$L(\vec{n}_i, \tau_i | \vec{n}_f) = \sum_{M, \vec{k}} \delta_{\vec{m}_M, \vec{n}_i} \int_{\mathcal{V}(\vec{\delta\tau})} d^M \vec{\delta\tau} \tilde{p}_{\text{path}}(M, \vec{\delta\tau}, \vec{k}) q_{\text{path}}(M, \vec{\delta\tau}, \vec{k}), \tag{21}$$

where the factor

$$\tilde{p}_{\text{path}}(M, \vec{\delta\tau}, \vec{k}) = e^{-\delta \tau_M \mathcal{T}_0(\vec{m}_M)} \prod_{j=1}^{M-1} \tilde{p}(k_j, \delta\tau_j | \vec{m}_j) \tag{22}$$

[1] Here we make a distinction between 'paths', which we define as the possible routes between the initial and final states occurring in the path summation expansion, and 'trajectories', which we use to refer to members of a stochastically simulated ensemble that *follow* such paths.

with

$$\tilde{p}(\delta\tau, k|\vec{n}) = e^{-\delta\tau\mathcal{T}_0(\vec{n})}\mathcal{T}_k(\vec{n}) \tag{23}$$

is of the same form as Eqs. (16) and (17) and is therefore a true norm-conserving probability, while

$$q_{\text{path}}(M, \vec{\delta\tau}, \vec{k}) = e^{\delta\tau_M\mathcal{T}_0(\vec{m}_M)\pi^{(c)}(\vec{m}_M)}\prod_{j=1}^{M-1} q(k_j, \delta\tau_j|\vec{m}_j) \tag{24}$$

with

$$q(\delta\tau, k|\vec{n}) = e^{\delta\tau\mathcal{T}_0(\vec{n})\pi_k^{(c)}(\vec{n})}(1 - \pi_k^{(d)}(\vec{n})) \tag{25}$$

accounts for the contribution of a path to any variation in the norm. Here we have made use of the following definitions:

$$\mathcal{T}_k(\vec{n}) = \max[T_k(\vec{n}), T_k(\vec{n} - \vec{\nu}_k)], \tag{26}$$

$$\pi_k^{(c)}(\vec{n}) = \frac{1}{\mathcal{T}_k(\vec{n})}\max[T_k(\vec{n} - \vec{\nu}_k) - T_k(\vec{n}), 0], \tag{27}$$

$$\pi_k^{(d)}(\vec{n}) = \frac{1}{\mathcal{T}_k(\vec{n})}\max[T_k(\vec{n}) - T_k(\vec{n} - \vec{\nu}_k), 0] \tag{28}$$

and as before we have set $\mathcal{T}_0(\vec{n}) = \sum_k \mathcal{T}_k(\vec{n})$. Just as the reaction probability density function given by Eq. (17) is the cornerstone of the predictive SSA, the two functions given by Eqs. (23) and (25) above together form the foundation of the new algorithm.

The first of these functions is the retrodictive reaction probability density function $\tilde{p}(\delta\tau, k|\vec{n})$, which provides the joint probability density that the next state in a simulated trajectory will be $\vec{n} - \vec{\nu}_k$ and that this transition will occur after a backwards-time increment $\delta\tau$, given that the current trajectory state is $\vec{n}$. Trajectories generated in this fashion will occur with the probability density $\tilde{p}_{\text{path}}(M, \vec{\delta\tau}, \vec{k}|\vec{n}_f)$.

The second of these functions $q(\delta\tau, k|\vec{n})$ is not a probability at all, but specifies how a particular path increment must alter the normalisation of the likelihood in order that the path as a whole contribute an amount determined by $q_{\text{path}}(M, \vec{\delta\tau}, \vec{k})$ to that normalisation. By noting that $L(\vec{n}, \tau|\vec{n}_f)$ is a positive semi-definite function and that therefore we can equate the expected absolute number distribution of trajectories $\mathcal{N}(\vec{n}, \tau|\vec{n}_f)$ with the likelihood in the following fashion:

$$\mathcal{N}(\vec{n}, \tau|\vec{n}_f) = L(\vec{n}, \tau|\vec{n}_f)\mathcal{S}_0, \tag{29}$$

where $\mathcal{S}_0$ is an initial number of trajectories, it is clear that we can represent dynamical variations in the normalisation of $L(\vec{n}, \tau|\vec{n}_f)$ as variations in the absolute number of trajectories present in an ensemble. We can therefore interpret $q(\delta\tau, k|\vec{n})$ as the fraction of the number of trajectories that were present in such an ensemble at state $\vec{n}$ at the start of the interval which must remain at the end of the interval to account for the required variation in normalisation. From this point of view, the exponential term in $q(\delta\tau, k|\vec{n})$ corresponds to trajectory creation due to a breeding process occurring at a fraction $\pi_k^{(c)}(\vec{n})$ of the total reaction rate $\mathcal{T}_0(\vec{n})$, while the linear term corresponds to the deletion of a fraction $\pi_k^{(d)}(\vec{n})$ of the trajectories at the end of the increment.

The trajectory creation/deletion can be combined with the path generation by iteratively modifying trajectories according to the reaction probability density function $\tilde{p}(\delta\tau, k|\vec{n})$ in the same way as the predictive SSA, but after each trajectory state increment $(\vec{n}, \tau) \to (\vec{n} - \vec{\nu}_k, \tau + \delta\tau)$, either

- creating a new trajectory in state $\vec{n}$ with probability $\pi_k^{(c)}(\vec{n})$ or
- deleting the original trajectory with probability $\pi_k^{(d)}(\vec{n})$.

This strategy, which we call the RSSA process, is summarised in Fig. 1 where it is compared to the SSA process and we have made use of the additional definition for the probability of a pure jump:

$$\pi_k^{(j)}(\vec{n}) = 1 - (\pi_k^{(c)}(\vec{n}) + \pi_k^{(d)}(\vec{n})). \tag{30}$$

It results in the required probability density of paths, along with the necessary modifications to the absolute number of trajectories to ensure that the equivalence given in Eq. (29) holds.

Finally, since the retrodictive probability distribution $R(\vec{n}, \tau|\vec{n}_f)$ can be obtained by normalising the likelihood distribution $L(\vec{n}, \tau|\vec{n}_f) = \mathcal{N}(\vec{n}, \tau|\vec{n}_f)/\mathcal{S}_0$, the expected relative frequency distribution at time τ of an ensemble of trajectories initialised in the state $\vec{n}_f$ and iteratively modified according to the process described above will yield the required stochastic solution to the RME.

3.3. Implementation

So far, we have described in formal terms a scheme capable of generating solutions to the RME using stochastic trajectories. We now turn to the issues surrounding the practical implementation of this scheme.

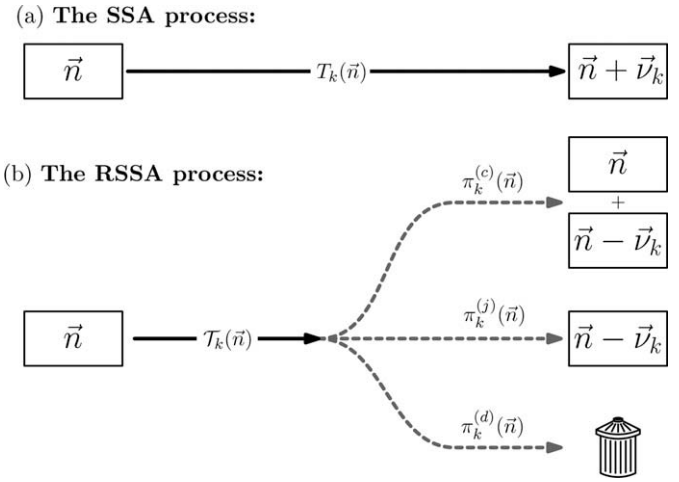

Fig. 1. Comparison of the implementation of stochastic processes in (a) the SSA and (b) the RSSA. The SSA implementation involves a simple jump between the states $\vec{n}$ and $\vec{n}+\vec{\nu}_k$ at rate $T_k(\vec{n})$, while that of the RSSA occurs at rate $\mathcal{T}_k(\vec{n})$ and involves randomly selecting from either a simple jump from $\vec{n}$ to $\vec{n}-\vec{\nu}_k$, creation of a new trajectory at $\vec{n}-\vec{\nu}_k$ or deletion of the original trajectory at $\vec{n}$, with probabilities $\pi_k^{(j)}(\vec{n})$, $\pi_k^{(c)}(\vec{n})$ and $\pi_k^{(d)}(\vec{n})$, respectively.

3.3.1. Sampling from the reaction PDF

The reaction probability density function defined in Eq. (23) dictates exactly when and how stochastic trajectories must be modified in order that their mean dynamics is equivalent to a solution of the RME. In order to propagate these trajectories in practice, we therefore need to be able to efficiently draw process number and time increment pairs $(k,\delta\tau)$ from $\tilde{p}(k,\tau+\delta\tau|\vec{n},\tau)$. The most direct way of achieving this is by following the prescription used by the original SSA, which involves selecting two pseudo-random numbers r_1 and r_2 from a uniform distribution on $[0,1]$, setting

$$\delta\tau = \frac{1}{\mathcal{T}_0(\vec{n})}\log\left[\frac{1}{r_1}\right], \tag{31}$$

and identifying k such that, if $r_2\mathcal{T}_0(\vec{n}) > \mathcal{T}_1(\vec{n})$,

$$\sum_{l=1}^{k-1}\mathcal{T}_l(\vec{n}) < r_2\mathcal{T}_0(\vec{n}) \leqslant \sum_{l=1}^{k}\mathcal{T}_l(\vec{n}) \tag{32}$$

but otherwise setting $k=1$.

This approach, the computational complexity of which scales as $O(K)$ where K is the number of concurrent birth/death processes, is perfectly adequate for simple models for which K is small. However, when performing retrodictive simulations on more complex models, one should be aware that the subtler methods which provide significant efficiency gains to the SSA can also be used to dramatically improve the efficiency of the RSSA. In particular, the approach proposed by Gibson and Bruck [15] yields a scaling of $O(\log[K])$ while, more recently, Slepoy et al. [16] reported an approach with a complexity completely independent of K.

3.3.2. Explicit resampling algorithm

While application of the algorithm described in Section 3.2 will result in a stochastic trajectory distribution which rigorously converges to the solution of the RME, the frequent addition and subtraction of trajectories from the simulation can easily lead to trajectory pools which either become impossibly large or else diminish to extinction in a very short time. This therefore presents a large impediment to any real calculation.

Fortunately there are practical solutions to this problem. Arguably the most obvious of these is to control the ensemble size via an explicit periodic resampling procedure such as the following:

1. Set a time τ_{sync} at which to perform the resample,
2. iteratively update the simulation pool until the times of all trajectories pass τ_{sync} and record their states as they pass, and
3. create a new pool using random samples taken from the recorded states, setting the new times to τ_{sync}.

As far as the actual sampling technique is concerned, a sensible method is as follows:

- When the number of trajectories in the pool is greater than desired, sample without replacement in order to ensure the maximum number of unique trajectories are included in the new pool.

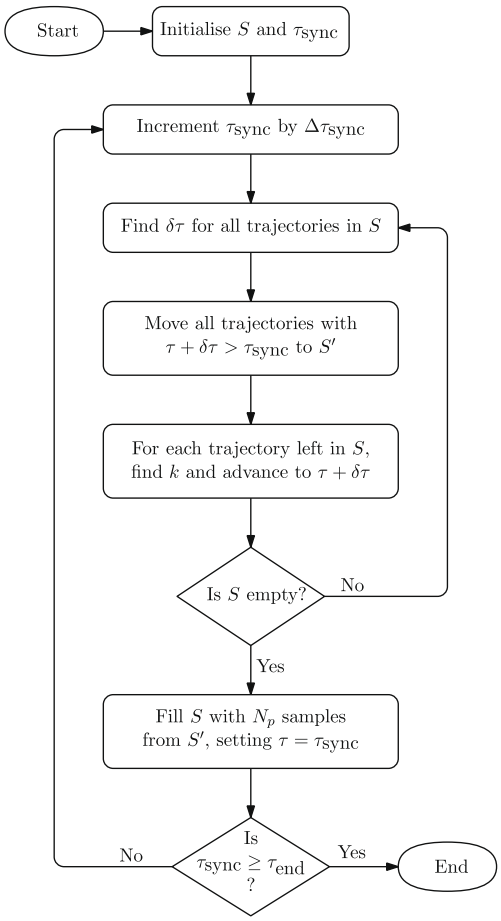

Fig. 2. Flow-chart illustrating the algorithm for the explicit resampling of a trajectory pool. The set S contains the pairs $(\vec{n}, \tau)$ which specify the state and time of each trajectory in the pool, while the set S' is used to hold the synchronised trajectory states at each τ_{sync}.

- When the present pool is smaller than desired, partially fill the new pool with the contents of the old and then add enough new trajectories by sampling from the old pool with replacement until the desired size is achieved.

By applying this resampling procedure periodically as outlined by the flow-chart shown in Fig. 2, one can reduce the variation in trajectory ensemble size to a manageable level. (In practice, it is best to adaptively modify the resampling period $\Delta\tau_{\text{sync}}$ to keep the fluctuations in the ensemble size below some tolerance.)

3.3.3. Implicit resampling algorithm

Another means of countering the depletion and/or unbounded growth of the RSSA trajectory pool is to update only one member of the ensemble at a time and to perform any resampling at that point. The calculation can then proceed using a fixed number of trajectories, eliminating the computational burden of dealing with a continually changing digital memory footprint, which is a side-effect of the explicit resampling algorithm presented in the previous subsection.

To implement this alternative approach we consider the state $\vec{V}$ of an entire simulation trajectory pool, which is at any given time identified by the tensor product of the $\mathcal{S}$ constituent trajectory state vectors $\vec{n}^{(i)}$ at that time. Then, instead of treating the dynamics of each trajectory independently, we consider the stochastic dynamics of the composite state. This can be done by regarding the stochastic modification of the ith trajectory in this pool, which proceeds at the rate $\mathcal{T}_0(\vec{n}^{(i)})$, as being instead the modification of the composite state $\vec{V}$ by the ith of $\mathcal{S}$ RSSA processes acting on independent subsets of the pool state space. From this perspective, the stochastic evolution of the pool as a whole is governed by the following trajectory/reaction probability density function:

$$\tilde{p}_{\mathcal{S}}(i, k, \delta\tau|\vec{V}, \tau) = \exp[-\mathbb{T}(\vec{V})\delta\tau]\mathcal{T}_k(\vec{n}^{(i)}), \tag{33}$$

where we have defined $\mathbb{T}(\vec{V}) \equiv \sum_{i=1}^{\mathcal{S}} \mathcal{T}_0(\vec{n}^{(i)})$. This specifies the probability that the ith trajectory will be modified by RSSA process k at time $\delta\tau$, given that the pool is in state $\vec{V}$ at time $\delta\tau$.

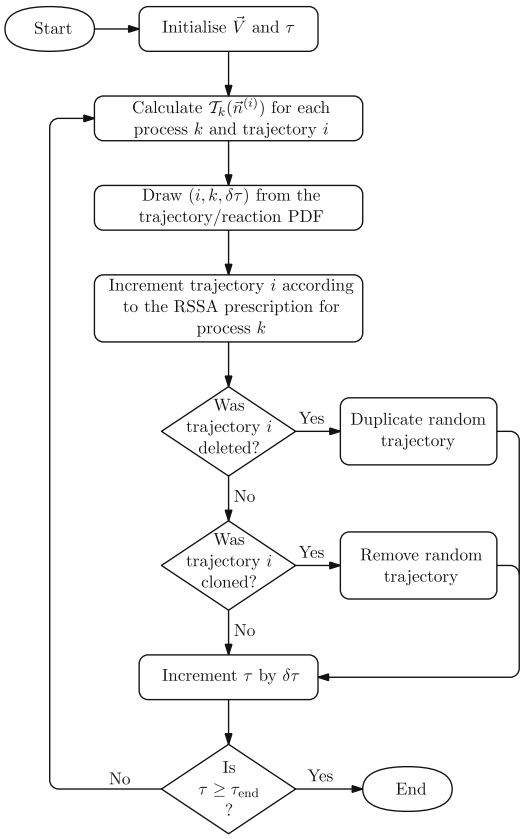

Fig. 3. Flow-chart illustrating the implicit resampling algorithm for stochastically evolving the composite state $\vec{V}$ of an RSSA trajectory pool of fixed size $\mathcal{S}$ backwards in time.

Collectively evolving a trajectory pool backwards in time therefore involves repeatedly drawing $(i, k, \delta\tau)$ triplets from the trajectory/reaction PDF, modifying the state of trajectory i according to the RSSA process k and performing any necessary resampling in order to maintain a fixed pool size. The resampling can be accomplished in the following way:

- If the chosen process k happens to result in the creation of a new trajectory, remove a random member of the pool to compensate. (This can be achieved practically by replacing the state of the randomly selected trajectory with the state of the trajectory to be created.)
- If the chosen process happens to result in the deletion of trajectory i, compensate by duplicating a random member of the pool. (That is, replace the state of trajectory i by that of a randomly chosen pool member.)

This algorithm is laid out in the flow-chart shown as Fig. 3.

Despite its clear structural simplicity, implementation of the implicit resampling algorithm requires some care if it is to be done in a computationally efficient manner. This is due to the fact that by treating the individual trajectory evolutions as separate stochastic processes, we are always operating in a regime where the effective number of processes, given by $\mathcal{S}K$, is large. This essentially mandates the use of the Gibson–Bruck scheme [15] mentioned previously in order to reduce the computational complexity of drawing from the trajectory/reaction PDF.

3.3.4. Dealing with systematic errors due to resampling

In applying the algorithms described in the previous sections, one needs to be aware that resampling does not come without a cost. Although each resample, either explicit or implicit, of the trajectory pool does preserve the relative frequency distribution over system states in the mean, it inevitably introduces correlations between trajectories in the pool which may cause systematic deviations from the exact result. The time necessary for a problematic fraction of the total ensemble to become correlated in this way will depend on the ensemble size, the frequency and severity of the resampling procedure and the particular details of the physical process being simulated.

Fortunately, one can guard against errors resulting from the resampling procedure by repeating a given calculation using larger and larger ensembles of trajectories, until no change in result is observed (or the change is below some reasonable

tolerance). This is analogous to the standard way in which one deals with finite time-step errors in numerical integration algorithms.

3.4. Comparison with existing Monte Carlo techniques

Although, as far as we have been able to determine, the particular numerical techniques discussed above represent a completely novel means of inferring the past state of systems subject to Markovian dynamics, they possess similarities to members of a class of techniques known broadly as Quantum Monte Carlo methods which are used to numerically determine exact ground states of quantum many-body systems. A particular example is the Green's Function Monte Carlo (GFMC) method of Kalos and Ceperly [17] which relies upon the fact that the ground state can be obtained as the limit of repeatedly multiplying an arbitrary state vector by a projection operator: $|\Psi_0\rangle = \lim_{n\to\infty} \widehat{O}^n |\phi\rangle$. One probabilistic interpretation of this iterative operation can be obtained in terms of stochastic trajectories that are subject to cloning and deletion.

That this branching form of GFMC is reminiscent of the algorithm described in this paper is not surprising when one considers that the discrete-time form of the backwards master equation (the unnormalised RME) can be written in the vector form $\mathbf{P}_{n+1} = M\mathbf{P}_n$, yielding solutions obtained through repeated application of the projection matrix: $\mathbf{P}_n = M^n\mathbf{P}_0$. The problem of identifying the long-time limit of solutions to the discrete-time backwards master equation is therefore mathematically equivalent to the problem of identifying ground states of quantum many-body systems.

4. Example applications

Up until this point we have dealt only with the development of our stochastic algorithm for retrodiction, the RSSA. We will now demonstrate its utility by calculating the likely ancestral states of a some systems evolving under elementary birth/death processes that commonly occur in realistic models of actual physical processes.

4.1. Simple birth/death models

We will firstly consider two extremely simple models: the basic death process and a constant-rate birth process. Using the standard notation for describing chemical reactions, we write the first of these as

$$X \xrightarrow{r} 0,$$

which describes a "reaction" in which individual reactants of type X are annihilated at rate r. The size of a well-mixed population of this reactant will evolve stochastically such that the probability of finding the system size to be n_X given that it was at some point n_{X0} is given by the solution to the predictive master equation

$$\frac{\partial}{\partial t} P(n_X, t|n_{X0}, t_0) = r[(n_X + 1)P(n_X + 1, t|n_{X0}, t_0) - n_X P(n_X, t|n_{X0}, t_0)]. \tag{34}$$

Through comparison with Eq. (1) one can see that the total transition rate for the death process is $T(n_X) = rn_X$, implying that the corresponding RME is

$$\frac{\partial}{\partial \tau} R(n_X, \tau|n_{Xf}) = rn_X[R(n_X - 1, \tau|n_{Xf}) - R(n_X, \tau|n_{Xf})] - R(n_X, \tau|n_{Xf})\frac{\partial \Omega(\tau)}{\partial \tau}. \tag{35}$$

Fig. 4(a) and (b) demonstrates, for $r = 1$, the agreement between the exact numerical integration of Eq. (35) and the relative frequency distribution of the states of 8×10^4 RSSA trajectories (computed using multiple runs of the adaptive explicit resampling algorithm with a nominal trajectory pool size of 400) at a time $\tau = 3$ prior to two different initial conditions: $n_{Xf} = 0$ and $n_{Xf} = 1$, respectively. These are the posterior probability distributions for the state of the system at this time, given the model and a uniform prior distribution over the states at the same time. As the birth rate at any instant depends on the size of the population at that time, a non-zero n_X implies that n_X must have been non-zero at all earlier times. This leads to the observed qualitative difference between the retrodictive probability distributions due to the two starting conditions as Fig. 4(a), with $n_{Xf} = 0$, shows that $R(n_X, \tau = 3|n_{X_f} = 0)$ has its mode at the extinct state while Fig. 4(b), shows that $R(n_X, \tau = 3|n_{X_f} = 1)$ maintains a zero at $n_X = 0$.

Similarly, also for individuals of type X, we consider the constant-rate birth process described by

$$0 \xrightarrow{r} X.$$

This kind of process arises naturally in situations where a parent population of constant size generates offspring which are unable to themselves reproduce – a scenario which is descriptive (over short timescales) of the production of cells by progenitors. The predictive master equation for this process is

$$\frac{\partial}{\partial t} P(n_X, t|n_{X0}, t_0) = r[(1 - \delta_{0,n_X})P(n_X - 1, t|n_{X0}, t_0) - P(n_X, t|n_{X0}, t_0)] \tag{36}$$

and therefore $T(n_X) = r(1 - \delta_{0,n_X+1})$. We thus find that the RME corresponding to the constant-rate birth process is

$$\frac{\partial}{\partial \tau} R(n_X, \tau|n_{Xf}) = r\big[R(n_X + 1, \tau|n_{Xf}) - R(n_X, \tau|n_{Xf})\big] - R(n_X, \tau|n_{Xf})\frac{\partial \Omega(\tau)}{\partial \tau}. \tag{37}$$

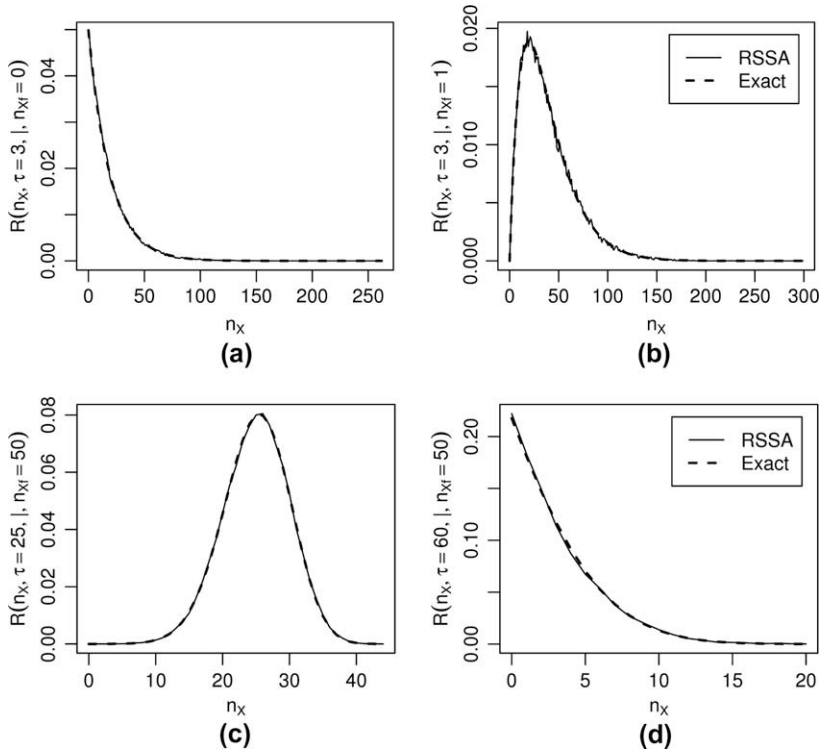

Fig. 4. Agreement between the RSSA results and those obtained through direct numerical integration of the retrodictive master equation. The upper two sub-figures display the distributions obtained at time $\tau = 3$ for a linear death/decay process with known final population sizes of (a) 0 and (b) 1, while the lower two display those obtained for a constant-rate birth process at times (c) $\tau = 25$ and (d) $\tau = 60$ given a known final population size of 50.

Fig. 4(c) and (d) illustrates the shape of the posterior probability distribution over possible values of n_X at two different times $\tau = 25$ and $\tau = 60$ before the population had a known size of $n_{Xf} = 50$, given a mean birth rate of $r = 1$ and a uniform prior probability distribution over the previous states of the system. Note that while the mode of the distribution decreases linearly with τ until it reaches zero at $\tau = 50$, there remains a significant probability of there having been more than zero individuals even at earlier times. These figures again demonstrate perfect agreement between the distribution obtained through the exact numerical integration of the RME and the relative frequency distribution of the corresponding RSSA trajectories (in this case 1.28×10^6 trajectories generated using multiple runs of the adaptive explicit resampling algorithm with a nominal trajectory pool size of 400).

4.2. The continuous-time Moran model

The Moran model [18] is a famous stochastic model of genetic drift and fixation in fixed-sized populations. Consider a population of N individuals, each possessing one of M alleles of some gene X. The state of the system can then be described by the vector $\vec{n}$, the elements of which specify the number of individuals possessing each allele X_j. The Moran model for the stochastic dynamics of the subpopulation sizes can then be summarised by the $M(M-1)$ processes of the form

$$X_i + X_j \xrightarrow{r_{ij}} 2X_i,$$

where $i \neq j$ and r_{ij} specifies the rate at which the process occurs given that a particular pair of individuals (i,j) is chosen.

The transition rate between states due to process ij is then $T_{ij}(\vec{n}) = r n_i n_j$, so the RME can be written

$$\frac{\partial}{\partial\tau}R(\vec{n},\tau|\vec{n}_f) = \sum_{i=2}^{M}\sum_{j=1}^{i-1} n_i n_j \left[r_{ij}R(\vec{n}_{+i-j},\tau|\vec{n}_f) + r_{ji}R(\vec{n}_{-i+j},\tau|\vec{n}_f) - (r_{ij}+r_{ji})R(\vec{n},\tau|\vec{n}_f)\right] - R(\vec{n},\tau|\vec{n}_f)\frac{\partial\Omega(\tau)}{\partial\tau}, \tag{38}$$

where $\vec{n}_{+i-j} = (n_1,\ldots,n_i+1,\ldots,n_j-1,\ldots,n_M)$ and $\vec{n}_{-i+j}$ is its compliment. For our purposes we will consider the neutral case were the reproductive fitness of every allele is equivalent and therefore the matrix r_{ij} can be replaced by the scalar rate r. If we further simplify the system by considering only two alleles, the RME reduces to

$$\frac{\partial}{\partial\tau}R(n,\tau|n_f) = rn(N-n)[R(n+1,\tau|n_f) + R(n-1,\tau|n_f) - 2R(n,\tau|n_f)] - R(n,\tau|n_f)\frac{\partial\Omega(\tau)}{\partial\tau}, \tag{39}$$

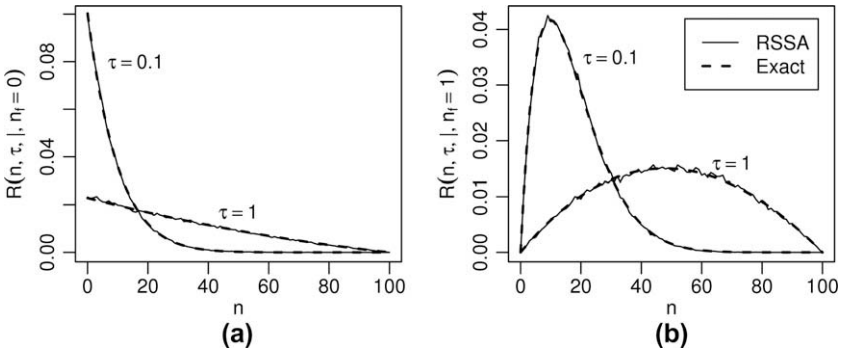

Fig. 5. Comparison between the RSSA results and those obtained via direct integration of the RME for probability distributions over states at previous times $\tau = 0.1$ and $\tau = 1.0$ for a population of two alleles, X_1 and X_2 evolving according to the continuous-time Moran model, given that the final size of the X_1 population is (a) $n = 0$ (extinct) and (b) $n = 1$.

where we have defined $n \equiv n_1$ and replaced n_2 with $N - n$, since the processes conserve the total population size and thus a single random variable is enough to fix the system's state.

Each of Fig. 5(a) and (b) displays the posterior probability distributions (given a uniform prior) over the states at two different times, $\tau = 0.1$ and $\tau = 1$, before one of two different known final states, $n_f = 0$ and $n_f = 1$. As in Fig. 4 both the results of numerically integrating the RME and those obtained stochastically are presented, again displaying perfect agreement. (The stochastic results were in this case derived from 1.28×10^6 RSSA trajectories generated via multiple runs of the adaptive explicit resampling algorithm to maintain a nominal pool size of 10^3.)

For the $n_f = 0$ final condition (Fig. 5(a)), the mode of the distribution over states at previous times remains at $(n_1 = 0,\ n_2 = N)$, while the distribution itself clearly approaches $R(n, \tau \gg 0 | n_f = 0) \propto (N - n)$. This makes sense, as one expects the probability that the system was in state n at some time well before it is known to be extinct to be proportional to the probability of extinction from state n, which for the 2 allele Moran model is known to be $(N - n)/N$.

In contrast, the distribution over previous population configurations given the $n_f = 1$ final condition (Fig. 5(b)) is strikingly different due to the fact that for this model, where extinction is a strictly one-way event, a population with a non-zero size at some time can never have been extinct at any previous time. At $\tau = 0.1$, the distribution is clearly asymmetric, with a bias toward states close to the known final state. At the earlier time of $\tau = 1$, however, the distribution is close to a steady-state where the only information it carries which is particular to the known final state is that neither allele was fixed at that time.

4.3. Genetic mutation

Finally, we consider the problem of performing retrodictive inference on a Markovian model of genetic (i.e. DNA, RNA or protein) sequence mutation. Such models are often used in phylogenetic tree reconstruction to estimate relationships between the inter-sequence Hamming distance – the number of sites at which two sequences differ – and the time necessary to accrue that many differences via mutation. (Refer to [19] for a detailed description of such methods.)

A general model of sequence mutation can be expressed in terms of reactions of the form

$$X_i \xrightarrow{\mu_{ji}} X_j,$$

where X_i represents the ith unique sequence and the mutation matrix μ_{ji} denotes a constant-rate of mutation from sequence X_i to sequence X_j. Particular models are then defined by the geometry of the sequence space and the exact structure of the mutation matrix.

In our case, we consider a simplified model in which sequences of a fixed length L contain bases chosen from the binary alphabet of "characters" 0 and 1. If we further allow only point mutations (mutations resulting in the change of a single character) and assume that all possible mutations occur at equal rates, the mutation matrix can be written

$$\mu_{ji} = \delta_{H(i,j),1} \frac{\mu}{L}, \tag{40}$$

where $H(i,j)$ is the Hamming distance between sequences i and j and $\mu = \sum_j \mu_{ji}$ is the total mutation rate of a given sequence.

This is essentially the Jukes–Cantor [20] model of DNA evolution applied to a binary rather than quaternary genetic alphabet and, despite its simplicity, it can be used to demonstrate some of the power of the methods that have been outlined in this paper.

4.3.1. Single mutating sequence

Firstly, consider the case of a single-organism possessing a single copy of a gene (or a small highly variable region of a particular gene) described by a binary sequence mutating according to the model just outlined. The state of this system

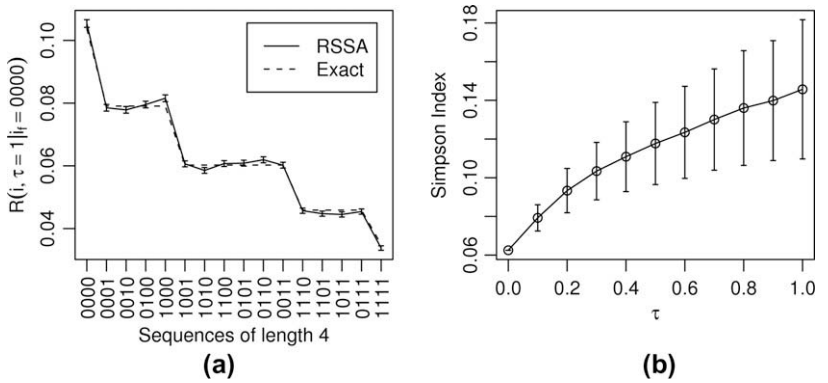

Fig. 6. (a) Probability distribution (given a uniform prior) over possible $L = 4$ binary sequences at a time $\tau = 1$ before the sequence is known to be $i_f = 0000$, again showing agreement between RSSA result (error bars represent standard error in mean) and direct integration of RME. (b) Variation of the mean Simpson Index of inferred states of a population of 80 $L = 4$ sequences with time τ before the population is known to be maximally diverse (error bars represent standard deviation).

can at any time be uniquely identified by the location i in sequence space currently occupied by the sequence. The retrodictive master equation for $R(i, \tau|i_f)$, the probability of the sequence having mutated to its final form i_f from the form i in a time τ, can then be written:

$$\frac{\partial}{\partial\tau}R(i,\tau|i_f) = \frac{\mu}{L}\sum_{j\in\mathcal{N}_i}[R(j,\tau|i_f) - R(i,\tau|i_f)], \tag{41}$$

where $\mathcal{N}_i$ is the set of sequences separated by a single point mutation from i. The normalisation term is absent because, in this case, it vanishes.

It is easily shown that the RME above is identical in form to the forward-time master equation for the single-organism system. This agrees with intuition, as an observation of the completed trajectory of an unbiased random walker cannot be used to determine the actual direction in which it was walked. We therefore expect the retrodictive probability distribution over sequence space to broaden with increasing τ.

This is exactly what we find. Fig. 6(a) illustrates the posterior probability distribution (uniform prior) over each of the 16 possible $L = 4$ sequences at time $\tau = 1$ given that the final sequence i_f is 0000 and the total mutation rate is $\mu = 4$. Shown in the figure are both the relative frequency distribution of 6.4×10^4 RSSA trajectories (computed using multiple runs of the adaptive explicit resampling algorithm and a nominal trajectory pool size of 4×10^3) and the result of directly integrating the RME, which are in good agreement with one another. (The error-bars denote the standard error in the mean of the stochastic results.) One can clearly see that the distribution has broadened substantially from its initial delta-function profile, the probability diminishing with increasing Hamming distance from i_f.

We thus find that for a single mutating sequence, inference of possible states at a particular time is dependent only on the time difference between that time and the time at which the state is known, as the distribution over likely states broadens symmetrically in both temporal directions.

4.3.2. Many mutating sequences

We now consider the more interesting situation in which we have a population of N individuals, each possessing a single copy of a gene undergoing mutation according to the model described above.

Intuitively, given that the dynamics are that of N independent random walks through sequence space, one might expect the inferred population distribution at earlier states to be broader than a current population distribution due to the of the broadening single-organism retrodictive probability distributions. However, as we will see, this intuition is incorrect.

It is easiest to express the unique states of this system in terms of state vectors $\vec{n}$ whose elements n_i describe the number of individuals possessing the sequence i. In this formalism, the RME becomes

$$\frac{\partial}{\partial\tau}R(\vec{n},\tau|\vec{n}_f) = \frac{\mu}{L}\sum_i\sum_{j\in\mathcal{N}_i} n_i[R(\vec{n}_{-i+j},\tau|\vec{n}_f) - R(\vec{n},\tau|\vec{n}_f)] - R(\vec{n},\tau|\vec{n}_f)\frac{\partial\Omega(\tau)}{\partial\tau}, \tag{42}$$

where, as in the Moran example, $\vec{n}_{-i+j} = (n_1,\ldots,n_i-1,\ldots,n_j+1,\ldots,n_{2^L})$.

Direct integration of the RME is in this case difficult even for very small systems. Indeed, for a population of $N = 80$ organisms each possessing a length 4 binary sequence, the number of distinct states $\vec{n}$ accessible to the dynamics is greater than 10^{17}.

We therefore use the RSSA to generate distributions over possible previous states of this system, at times ranging from $\tau = 0.1$ to $\tau = 1$ prior to a known uniform population distribution. In order to gauge the width of the inferred population distributions at each of these times, we employ Simpson's index of diversity [21]

$$S(\vec{n}) = \sum_i \left(\frac{n_i}{N}\right)^2 \tag{43}$$

which, for $L = 4$, ranges from between $1/16 = 0.0625$ for a population in which every possible sequence is equally represented and unity for a genetically homogeneous population.

Fig. 6(b) illustrates, at each of the various retrodictive times, the mean and standard deviation of the diversity index for the states of 6.4×10^4 trajectories generated by the RSSA. Again, the adaptive explicit resampling approach was used to maintain a nominal trajectory pool size of 4×10^3. (The results were compared with those generated using a trajectory pool size of 8×10^3 in order to check for systematic errors due to the resampling process, but none were found.) These results clearly demonstrate that, while the uncertainty in the state of the system becomes larger at earlier times, the average genetic diversity of the inferred populations is a monotonically *decreasing* function of τ.

This outcome, although perhaps surprising, can be easily reconciled with the broadening of the single-organism distribution by considering the importance of the difference in notation between Eqs. (41) and (42). In order to express $R(\vec{n}, \tau|\vec{n}_f)$ in terms of products of the probability distributions over single-organism states $R(i, \tau|i_f)$, one must symmetrise those products with respect to the labelling on individual organisms. This is due to the fact that following single-organisms of the population via $R(i, \tau|i_f)$ is equivalent to treating each as individually labelled, so that in order to obtain $R(\vec{n}, \tau|\vec{n}_f)$ one must explicitly discard these labels. The result of this symmetrization is that, while the location in sequence space of individually labelled organisms within the population may become more uncertain with increasing τ, groups of organisms are more likely to have originated from genetically similar populations simply due to the larger number of equivalent configurations that exist for such population distributions. (This is very similar to the quantum optical concept of Bose enhancement which is responsible for the macroscopic occupation of laser modes by photons.)

We therefore find that, for purely statistical reasons that have nothing to do with any kind of interaction between organisms, inferred previous states of *populations* of organisms possessing sequences evolving independently of one another tend to be less genetically diverse than the states of later populations. This is in contrast to the fact that our ability to infer the actual sequence of any given member of the population becomes progressively poorer for earlier times.

5. Conclusions

In this paper, we have demonstrated that classical master equations which govern the evolution of systems subject to continuous-time birth/death processes can be cast into a retrodictive form, the solutions of which allow one to directly calculate the probability that the known final state of a system arose from a particular initial state at some point in the past.

More importantly however, we have shown that solutions to these retrodictive master equations (RMEs) can be generated stochastically via a retrodictive stochastic simulation algorithm (RSSA), which is similar to the forward-time algorithm commonly used to solve predictive master equations. We have demonstrated that this algorithm can be used to infer the character of previous states of a variety of systems, including a genetic system operating in a state space volume exceeding 10^{17} states, where direct numerical integration of the corresponding RME is practically impossible.

We therefore conclude that the RSSA represents a useful means of systematically inferring the initial states of systems governed by arbitrary continuous-time birth/death Markov processes.

Acknowledgments

This work was supported by the Australian Research Council through a Discovery Project Grant (DP0773445).

References

[1] C.W. Gardiner, Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences, third ed., Springer-Verlag, Berlin, Heidelberg, New York, 2004.
[2] N.G.V. Kampen, Stochastic Processes in Physics and Chemistry, third ed., Elsevier, Amsterdam, Boston, Heidelberg, 2007.
[3] R.A. Blythe, A.J. McKane, Stochastic models of evolution in genetics, ecology and linguistics, J. Stat. Mech.: Theory Exp. 7 (2007) P07018.
[4] D.T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, J. Comput. Phys. 22 (1976) 403.
[5] D.T. Gillespie, Stochastic simulation of coupled chemical reactions, J. Phys. Chem. 81 (1977) 2340.
[6] B.F. Keele et al, Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection, Proc. Natl. Acad. Sci. (USA) 105 (21) (2008) 7552–7557.
[7] S.M. Barnett, D.T. Pegg, J. Jeffers, O. Jedrkiewicz, Atomic retrodiction, J. Phys. B: Atom. Mol. Opt. Phys. 33 (2000) 3047.
[8] D.T. Pegg, S.M. Barnett, J. Jeffers, Quantum retrodiction in open systems, Phys. Rev. A 66 (2002) 022106.
[9] H. Jeffreys, Theory of Probability, third ed., Oxford University Press, 1998.
[10] R.T. Cox, Probability, frequency and reasonable expectation, Am. J. Phys. 14 (1946) 1.
[11] E.T. Jaynes, Probability Theory: The Logic of Science, Cambridge University Press, 2003.
[12] A.N. Tikhonov, V.Y. Arsenin, Solutions of Ill-Posed Problems, Winston & Sons, Washington, DC, 1977.
[13] E.T. Jaynes, Prior information and ambiguity in inverse problems, SIAM-AMS Proc. 14 (1984) 151.
[14] S.X. Sun, Path summation formulation of the master equation, Phys. Rev. Lett. 96 (2006) 210602.
[15] M.A. Gibson, J. Bruck, Efficient exact stochastic simulation of chemical systems with many species and many channels, J. Phys. Chem. A 104 (2000) 1876–1889.
[16] A. Slepoy, A.P. Thompson, S.J. Plimpton, A constant-time kinetic Monte Carlo algorithm for simulation of large biochemical reaction networks, J. Chem. Phys. 128 (20) (2008) 205101.

[17] D.M. Ceperley, M.H. Kalos, Quantum many-body problems, in: K. Binder (Ed.), Monte Carlo Methods in Statistical Physics, Springer-Verlag, Heidelberg, 1979.
[18] P.A.P. Moran, Random processes in genetics, Proc. Camb. Philos. Soc. 54 (1958) 60.
[19] J. Felsenstein, Inferring Phylogenies, Sinauer Associates, Massachusetts, 2003.
[20] T.H. Jukes, C.R. Cantor, Evolution of protein molecules, in: M.N. Munro (Ed.), Mammalian Protein Metabolism, vol. 3, Academic Press, New York, 1969, p. 21.
[21] E.H. Simpson, Measurement of diversity, Nature 163 (1949) 688.