BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 6 2011, pages 837–843
doi:10.1093/bioinformatics/btr025

Genetics and population analysis

Advance Access publication January 18, 2011

Efficient simulation under a population genetics model of carcinogenesis

Tianqi Zhu[1], Yucheng Hu[1], Zhi-Ming Ma[2], De-Xing Zhang[3], Tiejun Li[1] and Ziheng Yang[3,4,5,*]

[1]School of Mathematical Sciences, Peking University, Beijing 100871, [2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, [3]Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, [4]College of Life Sciences, Peking University, Beijing 100871, China and [5]Department of Biology, University College London, London WC1E 6BT, UK

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Cancer is well known to be the end result of somatic mutations that disrupt normal cell division. The number of such mutations that have to be accumulated in a cell before cancer develops depends on the type of cancer. The waiting time T_m until the appearance of m mutations in a cell is thus an important quantity in population genetics models of carcinogenesis. Such models are often difficult to analyze theoretically because of the complex interactions of mutation, drift and selection. They are also computationally expensive to simulate because of the large number of cells and the low mutation rate.

Results: We develop an efficient algorithm for simulating the waiting time T_m until m mutations under a population genetics model of cancer development. We use an exact algorithm to simulate evolution of small cell populations and coarse-grained τ-leaping approximation to handle large populations. We compared our hybrid simulation algorithm with the exact algorithm in small populations and with available asymptotic results for large populations. The comparison suggested that our algorithm is accurate and computationally efficient. We used the algorithm to study the waiting time for up to 20 mutations under a Moran model with variable population sizes. Our new algorithm may be useful for studying realistic models of carcinogenesis, which incorporates variable mutation rates and fitness effects.

Contact: z.yang@ucl.ac.uk

Received on October 6, 2010; revised on December 20, 2010; accepted on January 12, 2011

1 INTRODUCTION

Cancer is a genetic disease caused by mutations in cancer susceptibility genes, including oncogenes, tumor suppressor genes and genetic instability genes (see e.g. Michor *et al*., 2004 for a review). Mutations that activate oncogenes can confer a selective advantage to the cell when one of the two alleles is mutated, but typically involve specific amino acid changes at specific residues of the protein. Mutations that inactivate both alleles of a tumor suppressor gene also lead to fitness benefits to the cell, and such mutations do not have to be specific as many mutations in a gene can disrupt its normal function (Forbes *et al*., 2010). Genetic instability refers to a diversity of changes in the genome on the nucleotide or chromosomal levels. The former is caused by mutations in the DNA repair pathways and involves substitutions, insertions or deletions of one or a few nucleotides. The latter includes gains or losses of whole chromosomes as well as inversions, deletions, duplications and translocations of large chromosomal segments. Mutations in genetic instability genes may cause increased mutation rates in the cell, leading to the so-called 'mutator phenotype' (Lengauer *et al*., 1998). For example, p53 mutations may impair the detection of and response to DNA damage. Many cancers have a mutator phenotype that results from mutations in genes involved in DNA mismatch repair.

It is now generally accepted that the rate-limiting step in the process of carcinogenesis is the accumulation of new mutations. Both fitness advantages and increased mutation rates can significantly shorten the waiting time until the required number of mutations. The relative importance of selection followed by clonal expansion versus raised mutation rates is somewhat controversial. Tomlinson *et al*. (1996) argue based on a computer simulation that advantageous mutations and clonal expansion were sufficient to initiate tumor without increased mutation rates. However, if more than two mutations in tumor suppressor genes are needed for carcinogenesis, increased mutation rates appear to be an important factor.

The number of mutations required for tumorigenesis depends on the type of cancer. Armitage and Doll (1954) used age-specific cancer incidence data to estimate the number of mutations necessary to develop colon cancer to be about 6. Knudson (1971) inferred from an analysis of the incidence of retinoblastoma in children that the cancer is caused by mutations to the two alleles of a gene (which was later identified to be RB1, the first tumor suppressor gene), one of which may be in the germline and inherited. Calabrese *et al*. (2005) analyzed data from 1022 colorectal cancers from nine hospitals in Finland and estimated that about five to six oncogenic mutations are required for hereditary cancers or seven or eight mutations for sporadic cancers. In contrast to those early studies which suggested that a handful of mutations were sufficient to cause cancer, recent efforts using next-generation sequencing technologies to identify mutations in protein-coding genes in common tumor types have in general identified far more genes (Sjöblom *et al*., 2006, Wood *et al*., 2007, Jones *et al*., 2008, Parsons *et al*., 2008). For example, Sjöblom

*To whom correspondence should be addressed.

et al. (2006) analyzed >13000 protein-coding genes in breast and colorectal tumors, and found that individual tumors accumulate an average of ∼90 mutations. Only a small subset of those mutations may be 'drivers', which offer the cell a selective advantage and are responsible for carcinogenesis, while the others are 'passengers', which provide no fitness benefits to the cell but hitchhike to fixation by chance. Sjöblom *et al*. (2006) estimated that as many as 14 mutations may be required to initiate colon cancer and as many as 20 may be involved in breast cancer (see Wood *et al*., 2007 for an updated analysis).

The early work of Armitage and Doll (1954) and Knudson (1971) has prompted a large body of theoretical work, which uses population genetics models to describe mutation, drift and selection and to study the waiting time until the accumulation of m mutations. Iwasa *et al*. (2004) studied a two-stage model for a population of cells evolving according to the Moran model. By assuming that the first mutation is either neutral or deleterious while the second mutation is advantageous, they studied an interesting scenario called stochastic tunneling, in which a population fixed for cells of no mutation can reach fixation for the second mutation without ever reaching fixation for the first mutation. Schweinsberg (2008) and Durrett *et al*. (2009) derived the asymptotic distributions of the waiting time until m mutations. Due to the complexity of the problem, they assumed selective neutrality and a constant mutation rate. Numerical simulations were conducted to confirm analytical results, but only for small populations ($10^3 \sim 10^4$). Beerenwinkel *et al*. (2007) studied the waiting time until m mutations under the Wright–Fisher model with large population sizes ($10^6 \sim 10^9$). With numerical simulation, they were able to show how varying population size, mutation rate and selection affect the waiting time.

As pointed out by Beerenwinkel *et al*. (2007), the Moran model is more natural than the Wright–Fisher model for describing the evolution of somatic cells. However, simulation under the Moran model is expensive, even more expensive than under the Wright–Fisher model. The large population sizes and small mutation rates mean that one has to simulate many transitions, which change the compositions of the population only slightly. In this article, we develop an approximate algorithm to simulate the waiting time until m mutations under the Moran model. We explore the similarity of the model of cancer initiation to a chemical reaction system and develop simulation algorithms that have been studied in that context.

The evolutionary process of mutation, drift and selection under the Moran model is described by a variable rate Markov chain, as is a chemical reaction system. The process can be simulated using standard algorithms by generating exponential waiting times until the next event and then deciding what event has happened. This is the so-called Gillespie algorithm (Gillespie, 1977, 2007). However, with a large population size, simulation using this exact algorithm is prohibitively slow. The τ-leaping algorithm (Gillespie, 2001; Li, 2007) was proposed to speed up the simulation by leaping over a time interval τ in which changes to the state and rates of the chain are expected to be small. However, this coarse-grained approximation does not work well for our problem, because of the existence of cell types whose population sizes may be small.

Here we develop a hybrid algorithm that can efficiently simulate the evolutionary processes of large populations. We use the exact algorithm to simulate transitions involving small populations, which may introduce substantial changes to the rates, and coarse-grained τ-leaping algorithm to simulate transitions within each large population, where transitions are unlikely to cause large changes to the state and the event rates. We test our new hybrid algorithm against the exact simulation in small populations and against asymptotic results in large populations. The results suggest that the hybrid algorithm is reliable and computationally efficient. The fast simulation algorithm may be useful for studying realistic carcinogenesis models that allow different cell types to have different finesses and different mutation rates.

2 METHODS

We consider a population of N cells, evolving according to the Moran model (Moran, 1958). Each cell's life span is an independent exponentially distributed random variable with parameter 1. When a cell dies, it is replaced by a new cell whose parent is chosen at random from the N cells in the population (including the one being replaced). We distinguish cells by the number of mutations that have accumulated, so that a type j cell has j mutations, and a new mutation in a type j cell changes the cell to type $j+1$. Initially all cells are of the wild type, i.e. type 0. The mutation rate depends on the cell type, and is μ_j for cells of type j. Thus models of mutator phenotypes can be accommodated using the model by assuming large u_js for large j. Genes are clonally inherited, so that the daughter cell inherits the mutations of the mother cell. This process continues until a cell has accumulated m mutations, when cancer develops. We are interested in the waiting time T_m until the appearance of the first type m cell.

Note that cells of different types may have different fitness. In particular, new mutations may increase the fitness, so that a type j cell may be fitter than a type $(j-1)$ cell. Each type j cell is characterized by its mutation rate μ_j and fitness $1+s_j$.

The evolutionary process of the cells is described by a Markov chain. The state of the chain at time t is $\boldsymbol{X}(t)=(x_0(t),x_1(t),\ldots,x_m(t))$, where $x_j(t)$ is the number of type j cells at time t. There are in total $m+1$ cell types. The initial state is $\boldsymbol{X}(0)=(N,0,\ldots,0)$, and the process stops when $x_m(t)>0$.

Suppose there are K possible events that can change the state of the chain. Note that some events do not change the system state (such as a type j cell being replaced by another type j cell), so we do not need to simulate them. Let the rates of the events be $a_j(\boldsymbol{X}(t))$ for $j=1,2,...,K$, and they cause the state of the chain to change by $\boldsymbol{\nu}_j$. In the language of chemical reaction (Gillespie, 2001), a_j is the reaction rate and $\boldsymbol{\nu}_j$ the state-change vector corresponding to reaction j. With this characterization, the Markov chain is described by the Kolmogorov forward equation (or master equation),

$$\frac{\partial}{\partial t}p(\mathbf{x},t|\mathbf{x}_0,t_0)=\sum_{j=1}^{K}a_j(\mathbf{x}-\boldsymbol{\nu}_j)p(\mathbf{x}-\boldsymbol{\nu}_j,t|\mathbf{x}_0,t_0)-\sum_{j=1}^{K}a_j(\mathbf{x})p(\mathbf{x},t|\mathbf{x}_0,t_0),$$

where the transition probability $p(\mathbf{x},t|\mathbf{x}_0,t_0)$ is the probability that $\boldsymbol{X}(t)=\mathbf{x}$ given that $\boldsymbol{X}(t_0)=\mathbf{x}_0$.

Under the Moran model, two kinds of events are possible in each generation:

(1) A type j cell is replaced by a type j' cell, with $j,j'=0,\ldots,m-1, j\neq j'$. This event occurs with rate $a_{jj'}=x_j(1+s_{j'})x_{j'}/\sum_{l=0}^{m-1}(1+s_l)x_l$. Here x_j cells of type j die at rate x_j, and the new born cell will be of type j' with probability $(1+s_{j'})x_{j'}/\sum_{l=0}^{m-1}(1+s_l)x_l$. The state-change vector is: $\boldsymbol{\nu}_{jj'}=\boldsymbol{\xi}_{j'}-\boldsymbol{\xi}_j$, where $\boldsymbol{\xi}_j$ is a vector with the j-th element being one and others zero.

(2) A type j cell mutates into type $(j+1)$, with $j=0,\ldots,m-1$. This occurs at rate $a_j=\mu_j x_j$, and the state-change vector is $\boldsymbol{\nu}_j=\boldsymbol{\xi}_{j+1}-\boldsymbol{\xi}_j$.

We would like to simulate the trajectories of $\boldsymbol{X}(t)$ until appearance of the first cell of type m. There are a total of $m(m-1)+m=m^2$ possible events, and we index them as $k=1,2,\ldots,m^2$. Below we review two algorithms developed for simulating such a system: the exact simulation algorithm (Gillespie, 1977, 2007) and the approximate τ-leaping algorithm (Gillespie, 2001), before describing our new hybrid algorithm.

2.1 Exact simulation algorithm

One can generate the exponential waiting time τ until the next event using the total event rate, and then sample the event among all possible events in proportion to their rates.

ALGORITHM 1. Exact algorithm
Initialization: set $t=0$ and initial value $\boldsymbol{X}(0)=(N,0,\ldots,0)$. Store the state-change vector $\boldsymbol{\nu}_k, k=1,\ldots,m^2$.

(1) Compute the rates $a_k=a_k(\boldsymbol{X}(t)), k=1,\ldots,m^2$.

(2) Generate the waiting time τ until the next event, which is an exponentially distributed random variable with parameter $a_{\text{total}}=\sum_{k=1}^{m^2} a_k$.

(3) Generate a random number $u\sim U(0,1)$. Find j such that

$$\sum_{k=1}^{j-1} a_k \le u a_{\text{total}} < \sum_{k=1}^{j} a_k.$$

(4) Update time as $t'=t+\tau$ and state as $\boldsymbol{X}(t')=\boldsymbol{X}(t)+\boldsymbol{\nu}_j$. If $x_m(t')>0$, set $T_m=t'$ and stop. Otherwise go back to step 1 with $t=t'$.

This algorithm is called the 'direct method' by Gillespie (1976). Some variations are possible, and furthermore, smart bookkeeping can lead to considerable improvement in its computational efficiency (Gibson and Bruck, 2000). However, all those exact algorithms are prohibitively slow when the cell population sizes are large (in the order of 10^6–10^9, say), because a_{total} is of order N and the expectation of the waiting time between two successive events is of order $1/N$.

2.2 The τ-leaping algorithm

To illustrate the idea of τ-leaping, consider the Moran model with only two cell types. The state vector is $\boldsymbol{X}(t)=(x_0,x_1)$, with $N=x_0+x_1\gg 1$. Suppose there are no mutations ($\mu_j=0$) and no selection ($s_j=0$), so that the model is a pure drift model. Two kinds of events drive the evolution of the population. (i) A type 0 cell is replaced by a type 1 cell with rate $a_1=x_0x_1/N$; (ii) A type 1 cell is replaced by a type 0 cell with rate $a_2=x_1x_0/N$. If $x_0=x_1=N/2\gg 1$, then $a_1+a_2=N/2\gg 1$, so that the average step size in the exact algorithm (the average waiting time until the next event) is $2/N\ll 1$. However, note that each event changes $\boldsymbol{X}$, and hence a_1 and a_2, only by a tiny proportion. The τ-leaping approximation is to consider the rates $a_1(\boldsymbol{X}(t_0))$ and $a_2(\boldsymbol{X}(t_0))$ to be constant in a time interval $t\in[t_0,t_0+\tau)$, so the numbers of occurrences of the two kinds of events over the time interval are approximately Poisson distributed with parameters $a_1(\boldsymbol{X}(t_0))\tau$ and $a_2(\boldsymbol{X}(t_0))\tau$. The *leaping time* τ can be chosen reasonably large so that many events can be updated in one step, and yet small enough so that the changes to the rates over the time interval are small to achieve good accuracy.

A widely used leaping condition (Cao *et al*., 2006) is

$$\mathbb{E}\Big[|X_i(t+\tau)-X_i(t)|\Big|X_i(t)=x_i(t)\Big]\le \epsilon x_i(t), \tag{1}$$

for all i and a small ϵ, say, in the range $0.01\sim 0.1$.

The τ-leaping algorithm may not be reliable if any x_i is small. In our model of cancer development, it is common to have some $x_i\sim\mathcal{O}(1)$ (e.g. the occurrence of mutation creating a new cell type), and thus the τ-leaping approximation cannot be applied directly.

2.3 A novel hybrid algorithm

The exact method is very inefficient as the step size is of $O(1/N)$. However, the τ-leaping algorithm is not directly usable as there may exist cell types with very small population sizes. Naive use of the τ-leaping algorithm may even lead to negative population sizes. We thus implement a hybrid algorithm.

We partition the cell types into two subgroups depending on the population size. We then use the exact algorithm to simulate events that involve any small population, and τ-leaping approximation to simulate events within each large population. Similar ideas have been used to speed up the simulation of multiscale chemical reaction systems (Haseltine and Rawlings, 2002).

A cell type i belongs to the major set Σ if $x_i(t)>N_c$, where N_c is a threshold value, or to the minor set σ otherwise. We set $N_c=10$ in our implementation. We further partition the possible events into the non-critical set Ω, which changes only the major cell types, and the critical set Λ, which changes at least one minor cell type.

The algorithm is as follows:

ALGORITHM 2. The hybrid algorithm
Initialization: Set $t=0$ and initial state $\boldsymbol{X}(0)=(N,0,\ldots,0)$. Store the state-change vector $\boldsymbol{\nu}_k, k=1,\ldots,m^2$.

(1) Compute rates $a_k=a_k(\boldsymbol{X}(t)), k=1,\ldots,m^2$.

(2) Partition the cells into the major and minor sets Σ and σ.

(3) Determine the τ-leaping step length τ. See below for details.

(4) Determine the waiting time for the next critical event: $e\sim\exp(a_\Lambda)$, where $a_\Lambda=\sum_{k\in\Lambda} a_k$.

(5) If $e<\tau$, simulate a critical event. Generate a random number $u\sim U(0,1)$ and find $j\in\Lambda$ such that

$$\sum_{k=1,k\in\Lambda}^{j-1} a_k \le u a_\Lambda < \sum_{k=1,k\in\Lambda}^{j} a_k.$$

Let $\boldsymbol{X}^*=\boldsymbol{X}+\boldsymbol{\nu}_j$.
Otherwise (if $e\ge\tau$) no critical event occurs, and $\boldsymbol{X}^*=\boldsymbol{X}$.

(6) Let $h=\min(\tau,e)$. Simulate non-critical events in Ω by τ-leaping over time h:

(a) Generate Poisson random variables $r_k\sim\mathcal{P}(a_k h), k\in\Omega$.

(b) Let $\boldsymbol{X}'=\boldsymbol{X}^*+\sum_{k\in\Omega} r_k\boldsymbol{\nu}_k$.

(7) Update time $t'=t+h$. If $x_m(t')>0$, set $T_m=t'$ and stop. Otherwise go back to step 1 with $t=t'$.

We determine the τ-leaping step length τ by applying the τ-leaping condition [Equation (1)] to the non-critical events. The rate at which x_i decreases by one is $x_i(1-x_i/N)+x_i\mu_i$. [x_i cells of type i are dying with rate x_i, and the probability that the new born cell is of the same type is x_i/N. Meanwhile x_i cells of type i mutate to type $(i+1)$ cells at rate $x_i\mu_i$.] The rate that x_i increases by one is $x_i(1-x_i/N)+x_{i-1}\mu_{i-1}$. [$N-x_i$ cells whose types are not i are dying with rate $N-x_i$, and the probability that the dying cell is replaced by a type i cell is x_i/N. Meanwhile x_{i-1} cells of type $(i-1)$ mutate to type i cells at rate $x_{i-1}\mu_{i-1}$.] Because mutation events usually occur slowly ($\mu_i\ll 1$), we ignore them here. For a similar reason, we ignore selection here. Then x_i increases or decreases by one with rate $(N-x_i)x_i/N$. This rate may be assumed to be approximately constant in the time interval $[t,t+\tau]$ when τ is small. Let

$$X_i^+(\tau):=\sharp\{s\in[t,t+\tau): X_i(s)-X_i(s-)=1\}$$

and

$$X_i^-(\tau):=\sharp\{s\in[t,t+\tau): X_i(s)-X_i(s-)=-1\}$$

be the numbers of events that increase and decrease type i cells in the time period $[t,t+\tau)$, respectively. Then we have

$$\begin{aligned}
&\mathbb{E}\Big[|X_i(t+\tau)-X_i(t)|\Big|X_i(t)=x_i(t)\Big]\\
=\;&\mathbb{E}\Big[|X_i^+(\tau)-X_i^-(\tau)|\Big|X_i(t)=x_i(t)\Big]\\
\le\;&\mathbb{E}\Big[X_i^+(\tau)\Big|X_i(t)=x_i(t)\Big]+\mathbb{E}\Big[X_i^-(\tau)\Big|X_i(t)=x_i(t)\Big]\\
=\;&\mathbb{E}\Big[\mathcal{P}(\tau(N-x_i)x_i/N)]+\mathbb{E}\Big[\mathcal{P}(\tau(N-x_i)x_i/N)\Big]\\
\le\;&2\tau x_i.
\end{aligned}$$

Thus, Equation (1) is satisfied as long as $\tau\le\epsilon/2$. In all of our numerical examples, we choose $\epsilon=0.04$ so that the leaping step length is $\tau=0.02$.

Table 1. Parameters used in Figure 1

	(a)	(b)	(c)
Population size (N)	10^3	10^4	10^6
Mutations to wait (m)	20	3	3
Mutation rate (μ)	$\mu_i = 10^{-3}$	$\mu_i = 10^{-4}$	$\mu_i = 10^{-6}$
Fitness ($1+s$)	$1+s_i = (1+0.01)^i$	1	1

3 RESULTS AND DISCUSSION

3.1 The performance of the hybrid algorithm

3.1.1 Accuracy In the first experiment, we examine the accuracy of our new hybrid method by comparison with the exact method for small N, and with an asymptotic result for large N. The parameters are shown in Table 1. For cases (a) and (b) of Table 1 with $N = 10^3$ or 10^4, the estimated probability density functions using the hybrid method are indistinguishable from those produced using the exact method (Fig. 1a and b). For case (c) with $N = 10^6$, the exact method is too slow to produce many samples, so we compare the hybrid method with an asymptotic result given in Schweinsberg (2008), which states that for a constant mutation rate and neutral evolution, if $\lim_{N\to\infty} \mu N = A$, then

$$P(\mu^{3/4} T_3 > t) \to \exp\left(-A \int_0^t \frac{1-e^{-2s}}{1+e^{-2s}} ds\right)$$

(Schweinsberg, 2008: Theorem 3 or case $\mu \sim AN^{-1}$ on p. 1451). Figure 1c shows that the hybrid method and the asymptotic approximation produced nearly identical results. $N = 10^4$ is apparently too small for the asymptotic approximation to be reliable, as the approximation produced quite different (much shorter) waiting time compared with the exact and hybrid algorithms in Figure 1b.

3.1.2 Computational efficiency We compared the computational efficiency of the hybrid and exact methods, by simulating 100 samples using both methods and measuring their computational time and average step sizes. The range of population size N is from 10^3 to 10^6. Other parameters are fixed at $\mu_i = \mu = 1/N$, $m = 4$ and $1+s_i = (1+0.01)^i$. The results are shown in Table 2. As N increases, the time taken by the exact method increases much faster than by the hybrid method . For the exact method, the average step size is of order $\mathcal{O}(N^{-1})$, while for the hybrid method it remained nearly constant.

3.2 A variable population size Moran model

To model the growth of a benign tumor (adenoma), we develop a Moran model of variable population size, with the total population size growing deterministically. The evolutionary process is described by a birth–death process. Three kinds of events are possible in the system:

(1) A type i cell mutates into a type $(i+1)$ cell, with rate $a_i^m = \mu_i x_i$ and state-change vector $\boldsymbol{\nu}_i = \boldsymbol{\xi}_{i+1} - \boldsymbol{\xi}_i$.

(2) A type i cell dies with rate $a_i^d = x_i$ and state-change vector $\boldsymbol{\nu}_i = -\boldsymbol{\xi}_i$.

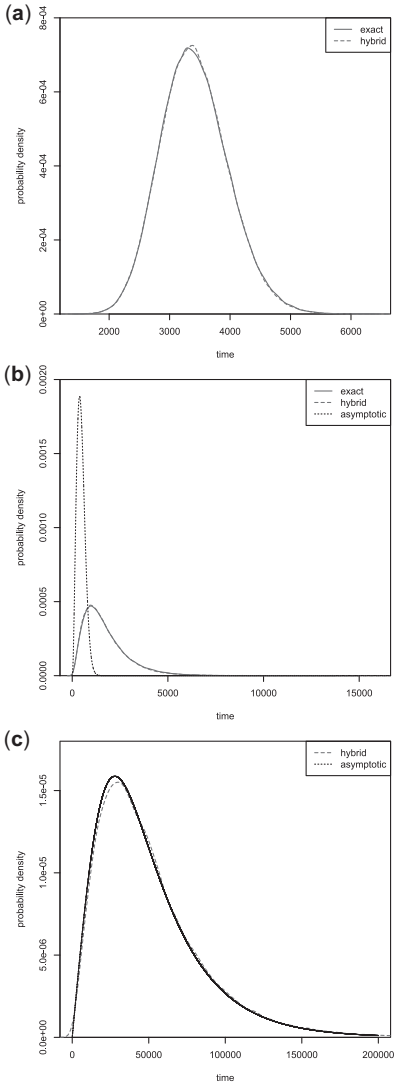

Fig. 1. Estimated probability density function of the waiting time T_m generated by the hybrid method in comparison with the exact simulation (**a** and **b**) and asymptotic result (**b** and **c**). The density functions are estimated using kernel density smoothing in R, with $2\times10^5, 2\times10^5$ and 10^5 samples for (a), (b) and (c), respectively. The parameters used are listed in Table 1.

Table 2. Time (in seconds) taken and average step sizes of the exact and the hybrid methods for simulating 100 trajectories

		$N = 10^3$	$N = 10^4$	$N = 10^5$	$N = 10^6$
Time	Exact	6	111	1595	19 414
	Hybrid	5	19	46	83
Step size	Exact	3.13e-003	3.67e-004	4.46e-005	5.45e-006
	Hybrid	0.0108	0.0104	0.0102	0.0101
Average T_m		412.9	827.6	1326.9	2056.7

The mutation rate $\mu_i = \mu$ is chosen to satisfy $\mu N = 1$. The number of mutations to cause cancer is $m = 4$ and the fitness is $1+s_i = (1+0.01)^i$.

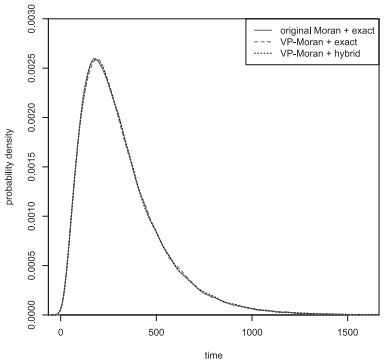

Fig. 2. Probability density function of waiting time T_3 under the original Moran model sampled by the exact method and under the variable population size Moran model sampled using both the exact and hybrid methods. The population size is $\tilde{N}(t)=N=1000$. The mutation rate $\mu_i=10^{-3}$ and there is no selection ($s_i=0$). The densities are estimated using kernel smoothing with 10^5 samples.

(3) A type i cell is born with rate $a_i^b=\tilde{N}(t)(1+s_i)x_i/\sum_k(1+s_k)x_k$ and state-change vector $\boldsymbol{\nu}_i=\boldsymbol{\xi}_i$.

Here $\tilde{N}(t)$ is not the real population size but the real population size $N(t)$ fluctuates closely around $\tilde{N}(t)$. To see this, note that when $N(t)<\tilde{N}(t)$, the total birth rate [which equals $\tilde{N}(t)$] will be larger than the total death rate [which equals $N(t)$], so that $N(t)$ will increase. Otherwise, $N(t)$ will decrease. Note that even if $\tilde{N}(t)=N$ is a constant, the new model still differs from the original Moran model. In the latter, the real population size is strictly constant, where in the new model it fluctuates around $\tilde{N}(t)$. There are $3m$ event types in the new model compared with m^2 in the original model. When m is large, the new model is faster to simulate than the original Moran model as determining the type of event requires fewer comparisons.

We note that a slightly different variable population size Moran model was described by Durrett and Mayberry (2009), in which new individuals are added into the population at a certain rate.

Two numerical examples are presented, to show (i) that the variable population size Moran model approximates the original model well when $\tilde{N}(t)=N$ and (ii) the real population size can be controlled effectively by $\tilde{N}(t)$.

In the first example, we let $\tilde{N}(t)=N$ and use the exact algorithm to simulate under the original Moran model, and both the exact and the hybrid methods to simulate under the variable population size Moran model. The estimated densities for the waiting time T_3 are indistinguishable among the three methods (Fig. 2). The variable population size Moran model approximates the original model very well.

In the second example, we demonstrate that $\tilde{N}(t)$ can effectively control the population size. We let $\tilde{N}(t)$ grow logistically from 10^4 to 10^6 with $\tilde{N}(t)=KN_0e^{rt}/(K+(N_0e^{rt}-1))$ with $N_0=10^4$, $K=10^7$ and $r=0.001$. Figure 3a gives one sampled trajectory, sampled every 100 days. We can see that $\tilde{N}(t)$ and the real population size $N(t)$ are indistinguishable. Figure 3b plots the step sizes for the exact and hybrid algorithm. For the exact method, the step size is of order $1/N$, so that the step size gets smaller as the population size increases. For the hybrid method, the step size remains at the leaping step size $\tau=0.02$.

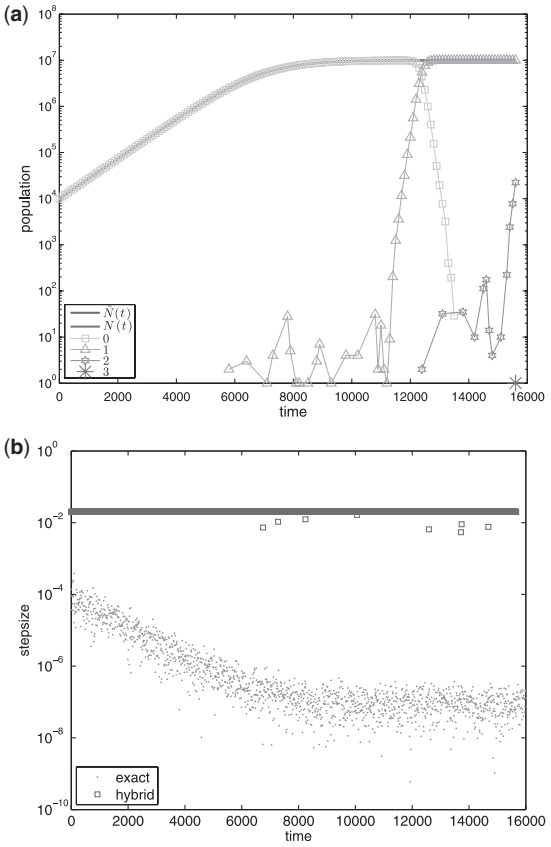

Fig. 3. (**a**) A trajectory simulated under the variable population size Moran model with $\tilde{N}(t)$ changing according to a Logistic growth curve. Samples are taken once every 100 days. $\tilde{N}(t)$ and the real population size $N(t)$ are indistinguishable on the plot. (**b**) The step lengths for the exact and hybrid methods for the simulation of (a).

3.3 Traveling waves of genetic adaptation

Beerenwinkel *et al*. (2007) conducted computer simulations to study the waiting time to $m=20$ mutations in a Wright–Fisher model of growing population size. The model mimics a colonic adenoma composed of 10^6 cells ($\sim$1 mm^3) initially that grows exponentially to reach a size of 10^9 cells ($\sim$1 cm^3) in about 12 years. Based on the data of Sjöblom *et al*. (2006), it is assumed that cancer will develop as soon as a cell accumulates $m=20$ mutations in any of the 100 susceptible genes. Mutations occur in unmutated genes at the rate of $\mu_g=10^{-7}$ per gene, with no back mutation. If any i of the 100 genes experience mutations, the cell will be of type i. The fitness of a type i cell is $1+s_i=(1+\gamma)^i$, where $\gamma=0.01$. Beerenwinkel *et al*. (2007) used the Wright–Fisher model as it allowed fast simulation even for large populations although the Moran model seemed more natural for cancer progression. Here, we use the new hybrid algorithm to simulate a similar process under the variable population Moran model.

We use $\tilde{N}(t)=N_0\exp(\alpha\bar{\omega}t)$, where $\alpha=0.0015$ controls the growth rate and $\bar{\omega}=\sum_i(1+s_i)x_i/\sum_i x_i$ is the average fitness of the population. A simulated trajectory is shown in Figure 4, which shows a pattern of traveling waves of clonal expansion. This is because a type i cell created by a new mutation in a population of type $(i-1)$ cells has greater fitness and is thus driven to fixation by natural

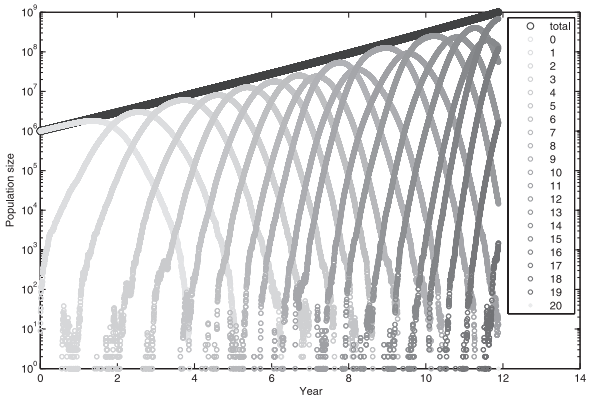

Fig. 4. A sampled trajectory of the evolutionary process of cancer initiation. The total population size is growing exponentially from 10^6 to 10^9 in about 12 years. Cell types with many mutations and thus high fitness are taking over the population successively in a pattern of traveling waves. The mutation rate is $\mu_g = 10^{-7}$ per gene and there are 100 susceptible genes, so that a type i cell mutates into a type $(i+1)$ cell at rate $\mu_i = (100-i)\mu_g$. The fitness for type i cells is $1+s_i = (1+\gamma)^i$, with $\gamma = 0.01$. The parameters used here are from Beerenwinkel *et al*. (2007).

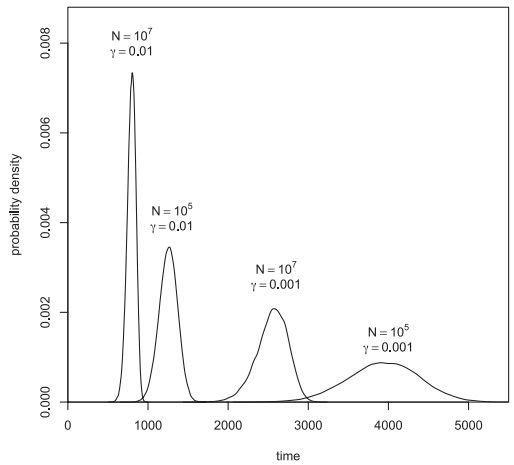

Fig. 5. Probability density function of T_{20} for different population sizes ($\tilde{N}$) and selection coefficients (s_i), estimated using kernel smoothing from 10^5 samples. The parameters used are as follows: $\mu_i = (100-i)\mu_g$ with $\mu_g = 10^{-5}$ and $1+s_i = (1+\gamma)^i$ with $\gamma = 0.01$.

selection, until the next mutation creates an even fitter type $(i+1)$ cell. This interesting phenomena is described earlier by Rouzine *et al*. (2003) in a model of asexual evolution and has recently been investigated analytically by Durrett and Mayberry (2009).

Figure 5 shows the probability density functions of T_{20}, the waiting time for 20 mutations. The model used is similar to that in Figure 4, but here we fix $\tilde{N}$ in the variable population size Moran model at 10^5 or 10^7. The selective benefit of each new mutation is $\gamma = 0.01$ or 0.001, and the mutation rate for each of the 100 susceptible genes is $\mu_g = 10^{-5}$. While the waiting time is much shorter for larger $\tilde{N}$, selection has much greater impact on the waiting time.

4 CONCLUSIONS AND PERSPECTIVES

We have developed a new hybrid algorithm for simulating the Moran model of carcinogenesis through accumulation of mutations. We use the algorithm to study the waiting time until m mutations for a range of population sizes. For small population sizes, our algorithm produces results nearly identical to those from the exact algorithm. For large populations, the exact method is too slow but the hybrid method still works accurately and efficiently. We implement a new Moran model of variable population size, which can effectively control deterministically changing population sizes. Our simulation model is general, allowing for arbitrary mutation rates and selective coefficients for different cell types. Those features of the algorithm indicate that it may be useful for studying complex and realistic evolutionary models of cancer initiation, which may be analytically intractable.

Our new hybrid simulation algorithm may be extended to accommodate more complex models of carcinogenesis. In particular, the evolutionary model we have implemented assumes a well-mixed tissue compartment without any spatial structure, in which all cells are in direct reproductive competition with each other. For certain types of cancer, it may be more realistic to model cellular differentiation and spatial structure explicitly, with stem cells generating somatic cells that undergo apoptosis in different compartments.

ACKNOWLEDGEMENT

We thank for anonymous referees for many constructive comments.

Funding: National Science Foundation of China (Grant No. 10871010); National Basic Research Program (Grant No. 2005CB321704) to Y.H. and T.L.; MOST (Grant No. 2006CB805901 to D.Z.) in part K.C. Wong Education Foundation, Hong Kong (to Z.Y.).

Conflict of Interest: none declared.

REFERENCES

Armitage,P. and Doll,R. (1954) The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br. J. Cancer*, **11**, 161–169.

Beerenwinkel,N. *et al.* (2007) Genetic progression and the waiting time to cancer. *PLoS Comput. Biol*, **3**, 2239–2246.

Cao,Y. *et al.* (2006) Efficient step size selection for the tau-leaping simulation method. *J. Chem. Phys.*, **124**, 44109–44119.

Calabrese,P. *et al.* (2005) Numbers of mutations to different types of colorectal cancer. *BMC Cancer*, **5**, 126–132.

Durrett,R. and Mayberry,J. (2009) Traveling waves of selective sweeps. *Arxiv preprint arXiv:0910.5730*.

Durrett,R. *et al.* (2009) A waiting time problem arising from the study of multi-stage carcinogenesis. *Ann. Appl. Probab.*, **19**, 676–718.

Forbes,S.A. *et al.* (2010) COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res.*, **38**, D652–D657.

Gibson,M.A. and Bruck,J. (2000) Efficient exact stochastic simulation of chemical systems with many species and many channels. *J. Phys. Chem. A*, **104**, 1876–1889.

Gillespie,D.T. (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, **22**, 403–434.

Gillespie,D.T. (1977) Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, **81**, 2340–2361.

Gillespie,D.T. (2001) Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.*, **115**, 1716–1733.

Gillespie,D.T. (2007) Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, **58**, 35–55.

Haseltine,E.L. and Rawlings,J.B. (2002) Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *J. Chem. Phys.*, **117**, 6959–6969.

Iwasa,Y. *et al.* (2004) Stochastic tunnels in evolutionary dynamics. *Genetics*, **166**, 1571–1579.

Iwasa,Y. *et al.* (2005) Population genetics of tumor suppressor genes. *J. Theor. Biol.*, **233**, 15–23.

Jones,S. *et al.* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, **321**, 1801–1806.

Knudson,A.G. (1971) Genetic instabilities in human cancers. *Proc. Natl Acad. Sci. USA*, **68**, 820–823.

Lengauer,C. *et al.* (1998) Genetic instabilities in human cancers. *Nature*, **396**, 643–649.

Li,T. (2007) Analysis of explicit tau-leaping schemes for simulating chemically reacting systems. *Multiscale Model. Simul.*, **6**, 417–436.

Michor,F. *et al.* (2004) Dynamics of cancer progression. *Nat. Rev. Cancer*, **4**,197–205.

Moran,P. (1958) Random processes in genetics. *Math. Proc. Camb. Phi. Soc.*, **54**, 60–71.

Parsons,D.W. *et al.* (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science*, **321**, 1807–1812.

Rouzine,I.M. *et al.* (2003) The solitary wave of asexual evolution. *Proc. Natl Acad. Sci. USA*, **100**, 587–592.

Schweinsberg,J. (2008) The waiting time for m mutations. *Electron. J. Probab.*, **13**, 1442–1478.

Sjöblom,T. *et al* (2006) The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**, 268–274.

Tomlinson,I. *et al.* (1996) The mutation rate and cancer. *Proc. Natl Acad. Sci. USA*, **93**, 14800–14803.

Wood,L.D. *et al.* (2007) The Genomic landscapes of human breast and colorectal cancers. *Science*, **318**, 1108–1113.