

An Algorithm for the Stochastic Simulation of Gene Expression and Heterogeneous Population Dynamics

Daniel A. Charlebois^{a,b,1}, Jukka Intosalmi^{c,d}, Dawn Fraser^{a,b}, Mads Kærn^{a,b,e,1}

^a*Department of Physics, University of Ottawa, 150 Louis Pasteur, Ottawa, Ontario, K1N 6N5, Canada.*

^b*Ottawa Institute of Systems Biology, University of Ottawa, 451 Smyth Road, Ottawa, Ontario, K1H 8M5, Canada.*

^c*Department of Mathematics, Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland.*

^d*Department of Signal Processing, Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland.*

^e*Department of Cellular and Molecular Medicine, University of Ottawa, 451 Smyth Road, Ottawa, Ontario, K1H 8M5, Canada.*

Abstract

We present an algorithm for the stochastic simulation of gene expression and heterogeneous population dynamics. The algorithm combines an exact method to simulate molecular-level fluctuations in single cells and a constant-number Monte Carlo method to simulate time-dependent statistical characteristics of growing cell populations. To benchmark performance, we compare simulation results with steady-state and time-dependent analytical solutions for several scenarios, including steady-state and time-dependent gene expression, and the effects on population heterogeneity of cell growth, division, and DNA replication. This comparison demonstrates that the algorithm provides an efficient and accurate approach to simulate how complex biological features influence gene expression. We also use the algorithm to model gene expression dynamics within ‘bet-hedging’ cell populations during their adaption to environmental stress. These simulations indicate that the algorithm provides a framework suitable for simulating and analyzing realistic models of heterogeneous population dynamics combining molecular-level stochastic reaction kinetics, relevant physiological details and phenotypic variability.

Keywords: Constant-number Monte Carlo, Stochastic simulation algorithm, Gene expression, Heterogeneous population dynamics

PACS: 87.10.Mn, 87.10.Rt, 87.16.Yc, 87.17.Ee

1. Introduction

Stochastic mechanisms play key roles in biological systems since the underlying biochemical reactions are subject to molecular-level fluctuations (see e.g. [11, 28]). Chemical reactions are discrete events occurring between randomly moving molecules. Consequently, the timing of individual reactions is nondeterministic and the evolution of the number of molecules is inherently noisy. One example of particular importance is the stochastic expression of gene products (mRNA and protein) [11, 12, 20, 23, 28]. Here, molecular-level fluctuations may cause genetically identical cells in the same environment to display significant variation in phenotypes,

¹Corresponding authors. Tel.: +1 613 562 5800 (Ext. 8691); Fax: (+1) 613 562 5636.

E-mail addresses: daniel.charlebois@uottawa.ca (Daniel Charlebois); mkaern@uottawa.ca (Mads Kærn).

loosely defined as any observable biochemical or physical attribute. While such noise is generally viewed as detrimental due to reduced precision of signal transduction and coordination, several scenarios exist where noise in gene expression may provide a fitness advantage (see Fraser and Kærn [6] for a review). For example, it has been proposed that a cell population may enhance its ability to reproduce (fitness) by allowing stochastic transitions between phenotypes to increase the likelihood that some cells are better positioned to endure unexpected environmental fluctuations [1].

Due to the importance of noise in many biological systems, models involving stochastic formulations of chemical kinetics are increasingly being used to simulate and analyze cellular control systems [9]. In many cases, obtaining analytical solutions for these models are not feasible due to the intractability of the corresponding system of nonlinear equations. Thus, a Monte Carlo (MC) simulation procedure for numerically calculating the time evolution of a spatially homogeneous mixture of molecules is commonly employed [7, 8]. Among these procedures, the Gillespie stochastic simulation algorithm (SSA) is the *de-facto* standard for simulating biochemical systems in situations where a deterministic formulation may be inadequate [7]. The SSA tracks the molecular number of each species in the system as opposed to the variation in concentrations in the deterministic framework. Hence, high network complexity, large separation of time-scales and high molecule numbers can result in computationally intensive executions. Another challenge is the need for simulating cell populations. In many cases, gene expression is measured for 10-100 thousand individuals sampled from an exponentially growing culture of continuously dividing cells. While the dynamics of individual cells can be appropriately simulated by disregarding daughter cells, repeating such simulations for a fixed number of cells may not capture population variability arising from asymmetric division, for example, or age-dependent effects. The alternative, tracking and simulating all cells within the population, is intractable beyond a few divisions due to an exponential increase in CPU demands as a function of time [22].

Here, we present a flexible algorithm to enable simulations of heterogeneous cell population dynamics at single-cell resolution. Deterministic and Langevin approaches to account for changes in intracellular content and the constant-number MC method [18, 31] were previously been combined to simulate and analyze gene expression across cell populations [21, 22]. In these studies, extrinsic heterogeneity associated with stochastic division and partitioning mechanisms, and intrinsic heterogeneity associated with molecular reaction kinetics were considered. Our algorithm, which combines the exact SSA for single-cell molecular-level modeling and a constant-number MC method for population-level modeling, is designed to incorporate user-defined biologically relevant features, such as gene duplication and cell division, as well as single cell, lineage and population dynamics at specified sampling intervals. Additionally, the SSA, which can be replaced by approximate methods if desired, is implemented within a shared-memory CPU parallelization framework to reduce simulation run-times. The emphasis of our study is to validate the accuracy of the method by directly comparing simulated results to the analytical solutions of models describing increasingly realistic biological features. Our results indicate that combining the SSA and the constant-number MC provides an efficient and accurate approach to simulate heterogeneous population dynamics, and a reliable tool for the study of population-based models of gene expression incorporating physiological detail and phenotypic variability.

This paper is organized as follows: Sections 2 and 3 briefly introduce the SSA and the constant-number MC method, respectively. The developed algorithm is described in Section 4. Section 5 provides the results of the benchmarking against analytical results. Finally, in Section 6, we demonstrate the applicability of the algorithm to more complex contexts by demon-

ing that it can quantitatively reproduce experimental measurements of gene expression dynamics within ‘bet-hedging’ cell populations during their adaption to environmental stress. The work is summarized in Section 7.

2. Stochastic Simulation Algorithm

The physical basis of the stochastic formulation of chemical kinetics is a consequence of the fact that collisions in a system of molecules in thermal equilibrium is essentially a random process [8]. This stochasticity is correctly accounted for by the Gillespie SSA, a MC procedure to numerically simulate the time evolution of chemical and biochemical reaction systems. While based on an assumption of intracellular homogeneity and mass-action kinetics, it is the *de-facto* standard for simulations of gene expression. In the Direct Method Gillespie SSA, M chemical reactions R_1, \dots, R_M with rate constants c_1, \dots, c_M among N chemical species X_1, \dots, X_N , are simulated one reaction event at a time. The next reaction to occur (index μ) and its timing (τ) are determined by calculating M reaction propensities a_1, \dots, a_M , given the current number of molecules of each of the N chemical species, to obtain an appropriately weighted probability for each reaction. It can be implemented via the following pseudocode [7, 8]:

```

1: if  $t < t_{end}$  and  $\alpha_M = \sum_{v=1}^M a_v \neq 0$  then
2:   for  $i = 1, M$  do
3:     Calculate  $a_i$  and  $\alpha_i = \sum_{v=1}^i a_v$ 
4:   end for
5:   Generate uniformly distributed random numbers  $(r_1, r_2)$ 
6:   Determine when  $(\tau = \ln(1/r_1)/\alpha_M)$  and which  $(\min\{\mu \mid \alpha_\mu \geq r_2 \alpha_M\})$  reaction will occur
7:   Set  $t = t + \tau$ 
8:   Update  $\{X_i\}$ 
9: end if
```

The SSA can be augmented to incorporate biologically relevant features, such as changes in the volume of the cell during growth, the partitioning of cell volume and content at division and DNA replication (see e.g. [2, 19, 25]). Changes in cell volume may have significant effects on reaction kinetics. First order reactions have deterministic rate constants (w_M) and stochastic rate constants (c_M) that are equal and independent of volume [14]. However, for higher order reactions, it is necessary to incorporate cell volume $V(t)$ into the reaction propensities in order to perform an exact simulation. For example, the stochastic rate constant for a bimolecular second order reaction R_μ at time t is given by

$$c_\mu = \frac{w_\mu}{N_A V_k(t)}, \quad (1)$$

where N_A is Avogadro’s number. Therefore, in the SSA, the rates of higher-order reactions must be scaled appropriately by the current cell volume before calculating propensities. This procedure has previously been demonstrated to provide a satisfying approximation as long as the kinetic time-scale is short compared with the cellular growth rate [19]. Typically, the volume of each cell k is modeled using an exponential growth law

$$V_k(t_{div}) = V_0 \exp \left[\ln(2) \left(\frac{t_{div}}{\tau_0} \right) \right], \quad (2)$$

where V_0 is the cell volume at the time of its birth, t_{div} is the time and τ_0 is the interval between volume doublings. This functional form allows for the description of dilution as a first- order decay process within a deterministic model of intracellular concentrations.

Once the SSA incorporates a continuously increasing cell volume, it is necessary also to specify rules that govern cell division. One option is ‘sloppy cell-size control’ [34] where the cell division is treated as a discrete random event that take place with a volume-dependent probability. Another simpler option is to assume that division occurs once the cell has exceeded a critical size V_{div} corresponding to one doubling of its initial volume, $V_{div} = 2V_0$. The volume doubling time τ_0 then becomes cell division time and t_{div} becomes the time since the last division. When cell division is triggered, i.e. when $V_k(t_{div}) \geq V_{div}$, additional rules must be specified to model the partitioning of cellular content between mother and daughter cells. For example, asymmetric cell division can be modeled by setting $V_{daughter} < V_{mother}$. The molecules of the cell can then be partitioned probabilistically between the two volumes [14, 27, 30, 33].

The SSA can accommodate additional discrete events. For example, the G2/M cell cycle checkpoint which ensure proper duplication of the cell’s chromosomes before division, can be modeled by defining a variable representing the completion of DNA replication such that cell division is delayed until the DNA content of the cell has doubled. The replication of individual genes, which doubles the maximum rate of gene transcription by doubling the number of corresponding DNA templates, can be modeled as a discrete event that occurs at a fixed time t_{rep} after cell division, i.e. when $t_{div} \geq t_{rep}$, or as a random event that occurs with some variable probability. In both cases, the DNA-replication event can be placed in a cell-specific stack of future events that is compared against t_{div} (or t in the above pseudocode) following each SSA step. Events in the stack scheduled to occur before this time are then executed and removed from the stack. This can be incorporate into the above pseudocode by inserting the following two lines:

- 8a: **if** $length(t_{event}) \geq 0$ **then** (there are scheduled events)
- 8b: **if** $t > t_{event}(i)$ **then** execute $t_{event}(i)$ and delete $t_{event}(i)$ from stack

This approach also provides a convenient basis for simulating the effects of time-delays [25, 26].

We note that the exact SSA can be extremely computationally intensive since the step size τ becomes very small when the total number of molecules is high or the fastest reaction occurs on a time-scale that is much shorter than the time-scale of interest. It therefore useful to develop techniques that can be used to speed up the simulation. This can be done, for example, using approximate methods such as the tau-leaping procedure in which each time step τ advances the system through possibly many reaction events [10]. Additionally, since many independent runs are required to compute population statistics, parallel computing can be used to further optimize simulation run-times.

3. Constant-Number Monte Carlo

Implementations of the modified SSA that track only one of the two cells formed during cell division may introduce artifacts in the calculation of population characteristics in the presence of significant phenotypic variability among cells. For example, gene expression capacity and division time may depend on chronological age; old cells may express genes at a reduced rate, and daughter cells may need to mature before they can reproduce. In addition, reproductive rates may be influenced by the accumulation of genetic mutations within a specific cell lineage or by the current levels of gene expression within individual cells. To simulate stochastic models of

gene expression incorporating such features, it is necessary to couple the SSA with simulation techniques used in studies of population dynamics.

The population balance equation (PBE) is a mathematical statement of continuity that accounts for all the processes that generate and remove particles from a system of interest [24], including individual members of a population [31]. In a general molecular-dynamics framework, the PBE contains terms due to nucleation, coagulation and fragmentation, and so forth, and is mathematically represented by an integro-differential equation that typically must be solved numerically to obtain particle size distribution and densities as a function of time [31]. Due to the integro-differential nature of the problem, discretization of the size distribution is required. This is problematic because features of the distribution are not known ahead of time and may change during growth [15, 31]. To resolve discretization problems that hinder the direct integration of the PBE, one can use MC methods to sample a finite subset of a system in order to infer its properties and study finite-size effects, spatial correlations, and local fluctuations not captured by a mean field approximation [10, 18, 24, 31]. Furthermore, a MC method is appropriate as its discrete nature adapts itself naturally to growth processes involving discrete events, and can simulate growth over arbitrary long times with finite numbers of simulation particles while maintaining constant statistical accuracy [18].

In order to construct a reliable and efficient algorithm to simulate biological cell populations, a constant-number MC method is adopted to simulate the birth-death processes that take place within such populations [18, 21, 22, 31]. This approach permits modeling of growing populations using a fixed number of cells while avoiding the alternative (i.e. an infinitely growing population) by sampling N particles representing the population as a whole. It essentially amounts to contracting the physical volume represented by the simulation to continuously maintain a constant number of cells [18]. The constant-number MC approach has been successfully applied to a variety of non-biological particulate processes [16, 18, 31] as well as cell population dynamics [21, 22].

In our implementation of the constant-number MC, we keep track of individual mother and daughter cells in two separate arrays. Each time a cell divides, the daughter cell is placed in the daughter array and the time of birth recorded. Then, at specified intervals, cells within the mother array are replaced one at a time, with the oldest daughter cells being inserted first. Because every mother cell is equally likely to be replaced during the sample update, the size distribution of the population remains intact for sufficiently large populations [31]. In our case, the size distribution corresponds to the distribution of cell ages (or volumes) across the population.

The constant-number MC method can be represented by the following pseudocode:

```

1: if  $t > t_{restore}$  and  $NC_{daughter} \geq 1$  then
2:   for all  $NC_{daughter}$  do
3:     Randomly select mother cell
4:     Replace mother cell with oldest available daughter cell
5:   end for
6: end if
```

Here, $t_{restore}$ is the interval between population updates and $NC_{daughter}$ the number of daughter cells born since the last update. To avoid simulating the daughters of daughter cells, $t_{restore}$ is chosen such that mother cells divide at most once, and daughter cells not at all, during a particular $t_{restore}$ interval.

4. Algorithm

Simulations are carried out using an initial population distribution, where gene expression in each cell is described by a user defined set of equations, and population statistics are obtained at a specified sampling interval. Here, stochastic simulation is carried out using the Gillespie direct method [7, 8], however any stochastic simulation method can be implemented. Parallelism is implemented across the simulation (see Fig. 1 and pseudocode in this section), as a large number of independent simulations need to be performed when simulating the dynamics of a cell population, in a shared memory multiprocessor environment.

The algorithm can be expressed by the flow diagram (Fig. 1) and the following pseudocode:

```

1: while  $t < t_{end}$  do
2:   begin parallel region
3:     for all  $NC_{population}$  such that  $t < t_{sample}$  do
4:       Gillespie SSA (see pseudocode in Section 2)
5:       Update  $V_k$ 
6:       Execute events in stack with  $t_{event} < t_{div}$ 
7:       if  $V_k(t_{div}) \geq V_{div}$  then
8:         Execute cell division
9:         Increment  $NC_{daughter}$ 
10:      end if
11:    end for
12:    Update  $t_{sample}$ 
13:  end parallel region
14:  Execute constant-number MC (see pseudocode in Section 3)
15:  Compute statistics
16: end while

```

Here, $NC_{population}$ is the total number of cells in the population, V_k the volume of cell k , and t_{sample} the user defined population sampling interval.

The algorithm can execute simulations of considerable size in reasonable times. For example, an IBM with 2 quad-core processors (1.86GHz cores) and 2.0GB of RAM completed a $10^5 s$ simulation of the network presented in Section 5.1 for 8000 cells in 81s when $v_0 = 0.3 s^{-1}$, $v_1 = 0.05 s^{-1}$, $d_0 = 0.05 s^{-1}$, $d_1 = 5 \times 10^{-5} s^{-1}$, $t_{div} = 3600 s$, and $t_{restore} = 3300 s$.

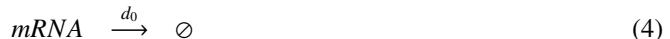
5. Numerical Results

In order to evaluate the accuracy of the present algorithm, we compare simulation results to steady-state and time-dependent analytical solutions of constitutive gene expression models. In this section, models describing increasingly realistic biological features are considered and presented along with the derivations of the corresponding analytical solutions. We have included these details to emphasize the significant complexity associated with the derivation of even simple kinetic models. Part of our motivation for developing the algorithm is the anticipation that finding analytical solutions to models incorporating complex biochemical reaction network and cellular physiology will be intractable. We begin in Subsection 5.1 by considering time-dependent gene expression, i.e., the transcription of RNA and translation of RNA into protein, and benchmark this scenario against the corresponding time-dependent analytical

distributions. In Subsection 5.2 we consider both time-dependent and time-independent gene expression using a model that incorporates the effects of gene duplication and cell division on gene expression dynamics in individual cells using the constant-number MC method. All simulations statistics were obtained from populations consisting of 8000 cells.

5.1. Time-Dependent Population Distributions

Population-based simulation algorithms have the advantage of yielding time-dependent population-distributions as the output. To evaluate the accuracy of our approach in this respect, validation against a time-dependent distribution is of interest. For this purpose, we simulate a two-stage gene expression model consisting of the following biochemical reactions:



where Eq. (3) describes transcription at a rate v_0 , Eq. (4) the degradation of the mRNA at a rate d_0 , Eq. (5) translation at a rate v_1 , and Eq. (6) the protein degradation at a rate d_1 . Here, all rates are given in probability per unit time and it is assumed that the promoter T is always active and thus the model has two stochastic variables, the number of mRNAs and the number of proteins P .

Shahrezaei and Swain [30] studied the system described by Eqs. (3)-(6) and derived an approximative protein distribution as a function of time. The approximation is based on the assumption that the degradation of mRNA is fast compared to the degradation of proteins (i.e. $d_0/d_1 \gg 1$). Consequently, the dynamics of mRNA is at the steady-state for the most of a protein's lifetime. The essential steps of the derivation are as follows (see supplementary materials in [30] for the complete derivation):

The chemical master equation (CME) describing the probability of having m mRNAs and n proteins for the system in Eqs. (3-6) at time t is given by

$$\begin{aligned} \frac{\partial P_{m,n}}{\partial t} = & v_0(P_{m-1,n} - P_{m,n}) + v_1 m (P_{m,n-1} - P_{m,n}) \\ & + d_0[(m+1)P_{m+1,n} - mP_{m,n}] \\ & + d_1[(n+1)P_{m,n+1} - nP_{m,n}]. \end{aligned}$$

If we let $u = z' - 1$ and $v = z - 1$, the corresponding generating function $F(z', z)$, defined in [30] as $\sum_{m,n}(z')^m z^n P_{m,n}$, is given by

$$\frac{1}{v} \frac{\partial F}{\partial \tau} + \frac{\partial F}{\partial v} - \gamma \left[b(1+u) - \frac{u}{v} \right] \frac{\partial F}{\partial u} = a \frac{u}{v} F, \quad (7)$$

where $a = v_0/d_1$, $b = v_1/d_0$, $\gamma = d_0/d_1$, and $\tau = d_1 t$. If r measures the distance along a characteristic, which starts at $\tau = 0$ with $u = u_0$ and $v = v_0$ for some constants u_0 and v_0 , then from Eq. 7 it is found that

$$\frac{du}{dr} = -\gamma \left[b(1+u) - \frac{u}{v} \right] \quad (8)$$

using the method of characteristics. Consequently direct integration implies that $v = r$ and Eq. 8 has the solution

$$u(v) = e^{-\gamma bv} v^\gamma \left[C - b\gamma \int^v dv' \frac{e^{\gamma bv'}}{v'^\gamma} \right] \quad (9)$$

for a constant C . By Taylor expansion of $e^{\gamma bv}$ such that $e^{\gamma bv} = \sum_n (\gamma bv)^n / n!$ the integral in Eq. 9 can be evaluated, and, if Stirling's approximation is subsequently applied, $u(v)$ is found for $\gamma \gg 1$ to obey

$$u(v) \cong \left(u_0 - \frac{bv_0}{1-bv_0} \right) e^{-\gamma b(v-v_0)} \left(\frac{v}{v_0} \right)^\gamma + \frac{bv}{1-bv} \quad (10)$$

or

$$u(v) \cong \frac{bv}{1-bv} \quad (11)$$

as $v = v_0 e^\tau > v_0$ for $\tau > 0$. When $\gamma \gg 1$, u tends rapidly to a fixed function of v and the generating function describing the distribution of proteins can be obtained from Eq. 7

$$\frac{dF}{dv} \cong \frac{ab}{1-bv} F. \quad (12)$$

Integrating Eq. 12 yields the probability distribution for protein number as a function of time

$$F(z, \tau) = \left[\frac{1 - b(z-1)e^{-\tau}}{1 + b - bz} \right]^a. \quad (13)$$

By definition of a generating function, expanding $F(z)$ in z yields

$$P_n(\tau) = \frac{\Gamma(a+n)}{\Gamma(n+1)\Gamma(a)} \left[\frac{b}{1+b} \right]^n \left[\frac{1+be^{-\tau}}{1+b} \right]^a \times {}_2F_1 \left[-n, -a, 1-a-n; \frac{1+b}{e^\tau+b} \right], \quad (14)$$

where ${}_2F_1$ and Γ are the hypergeometric and the gamma function, respectively. The initial number of proteins n is set to zero. In this case, the mean, variance, and protein noise of the process are described respectively by

$$\mu_P(\tau) = ab(1 - e^{-\tau}), \quad (15)$$

$$\sigma_P^2(\tau) = \mu_P(1 + b + be^{-\tau}), \quad (16)$$

$$\eta_P(\tau) = \sigma_P/\mu_P = \left[\frac{1 + b + be^{-\tau}}{ab(1 - e^{-\tau})} \right]^{1/2}. \quad (17)$$

To benchmark the ability of the algorithm to accurately generate time-dependent population distributions, we simulated Eqs. (3)-(6) under conditions where the assumptions of Eq. (14) are satisfied, and compared the resulting distributions with corresponding time-dependent analytical distributions. Fig. 2 shows the simulated and analytical distributions at two different values of dimensionless time τ . The population statistics, specifically μ_P and η_P , as a function of τ are shown in Fig. 3. In both cases, the simulated protein distributions and statistics are in excellent agreement with the analytical results (Eqs. (14)-(17)).

5.2. Gene Duplication, Cell Division, and Time-Dependent Validation

To explore the accuracy of the algorithm when simulating models incorporating cell growth, division, and DNA replication, we implemented the simplified reaction network presented in Swain *et al.* [33]. The reduced reaction network was obtained from a model of gene expression consisting of 8 molecular species and 11 chemical reactions. For this simplified network, it is possible to derive time-dependent analytical results for the mean protein number and coefficient of variation in protein number. Importantly, by making the appropriate approximations, the effects of gene replication and cell division can be included in the analytical solutions. The reduced model have two components - one described by the reactions in Eqs. (3)-(6) (note that the reaction rates v_1 and d_0 can be directly related to v'_1 and d'_0 in the original model [33]), and another describing pre-transcription kinetics. This component captures the reversible binding of RNAP to the promoter (rate constants b_0 and f_0), and the formation of an open promoter complex (rate constant k_0). These steps are described by the reactions



where D , C and T represent the promoter with polymerase unbound, the promoter with polymerase bound and the open promoter complex, respectively. Since the total number n of DNA molecules is conserved before and after replication, D and C can be constrained by

$$n_0 + n_1 = n, \quad (20)$$

where n_0 and n_1 are the number of promoter copies in state D and C respectively.

To derive an analytical solution, the authors invoked the assumption that the distributions of C , T , and mRNA can be approximated by their steady state distributions. While this assumption thus ignores the transient dynamics of these species, it is expected to introduce a minimal error since the protein degradation rate d_1 is much smaller compared to the other reaction rates. As a consequence, the mean and coefficient of variation protein P are time-dependent while the moments of the distributions of the other species are constant. Even with this approximation, the derivation of the analytical solutions for the mean and coefficient of variation is rather arduous. In the following, we highlight the only the main points (the complete derivation can be found in the supplementary material of Swain *et al.* [33]). It consists of three separate stages - the derivation of time-dependent expression for the population mean and noise, the incorporation of gene replication and the addition of cell division.

The first stage is analogous to the derivation of time-dependent moments in Section 5.1, that is, cell cycle effects are neglected and the probability distributions for the species C , T , mRNA, and P is described using the CME. In this case, the variables n_1 , n_2 , n_3 , and n_4 are used to describe the numbers of C , T , mRNA, and P , respectively, and $p(n_1, n_2, n_3, n_4, t)$ denotes the probability density function of the time-dependent state. The CME can be correspondingly be written in the form

$$\begin{aligned} \frac{\partial p(n_1, n_2, n_3, n_4, t)}{\partial t} &= f_0[(n - n_1 + 1)p(n_1 - 1, n_2, n_3, n_4, t) \\ &\quad - (n - n_1)p(n_1, n_2, n_3, n_4, t)] + \dots, \end{aligned} \quad (21)$$

where dots denote similar terms, one for each rate constant. The CME is then used to derive an expression for the time-dependent probability-generating function. The probability-generating

function is defined by

$$F(z_1, z_2, z_3, z_4, t) = \sum_{n_1, n_2, n_3, n_4} z_1^{n_1} z_2^{n_2} z_3^{n_3} z_4^{n_4} p(n_1, n_2, n_3, n_4, t). \quad (22)$$

It can easily be seen that differentiating F with respect to z_i and setting all z_i to unity, gives μ_{n_i} and similarly the second derivative gives $\mu_{n_i(n_i-1)}$. Applying the transformation given by Eq. (22) to the CME (Eq. (21)), an expression for the probability-generating function can be obtained. This expression has the form of the partial differential equation

$$\begin{aligned} \frac{\partial F}{\partial t} &= f_0 nwF - [f_0 w(1+w) + b_0 w - k_0(x-w)] \frac{\partial F}{\partial w} + v_0(y-x) \frac{\partial F}{\partial x} \\ &\quad + [v'_1 z(1+y) - d'_0 y] \frac{\partial F}{\partial y} - d_1 z \frac{\partial F}{\partial z}, \end{aligned} \quad (23)$$

where $w = z_1 - 1$, $x = z_2 - 1$, $y = z_3 - 1$, and $z = z_4 - 1$. This equation, just like the CME, is practically impossible to solve. However, the equation can be combined with a second order Taylor expansion of Eq. (22) which can be written in the form

$$\begin{aligned} F(w, x, y, z, t) &\simeq 1 + wX_1 + xX_2 + yX_3 + zX_4(t) + \frac{1}{2} [X_{11}w^2 + X_{22}x^2 \\ &\quad + X_{33}y^2 + X_{44}(t)z^2 + 2X_{12}wx + 2X_{13}wy + 2X_{23}xy \\ &\quad + 2X_{14}(t)wz + 2X_{24}(t) + 2X_{34}(t)yz], \end{aligned} \quad (24)$$

where the expansion is taken around $w = 0$, $x = 0$, $y = 0$, $z = 0$ so that the following holds: $X_i = \mu_{n_i}$, $X_{ii} = \mu_{n_i^2} - \mu_{n_i}$, and $X_{ij} = \mu_{n_i n_j}$, $i \neq j$. Here it is important to note that only the processes involving protein molecules are time-dependent according to the previous assumptions. The Eq. (24) is then substituted to Eq. (23), the coefficients are compared and solvable expressions for the expected values, variances, and covariances of the considered process are obtained. This gives equations governing the variables $X_4 = \mu_P$ and $X_{44} = \mu_{P(P-1)}$

$$\frac{dX_4(t)}{dt} = v'_1 X_3 - d_1 X_4(t), \quad (25)$$

$$\frac{dX_{44}(t)}{dt} = 2v'_1 X_{34}(t) - 2d_1 X_{44}(t). \quad (26)$$

Assuming that $\mu_P(0) = m$, Eqs. (25) and (26) can be solved using expressions for the other X_{ij} variables. The expressions are rather complex and the interested reader should refer to [33]. Solving Eqs. (25) and (26) yields the following expressions for the protein mean and variance

$$\mu_P(t) = \frac{v'_1 X_3}{d_1} (1 - e^{-d_1 t}) + m e^{-d_1 t}, \quad (27)$$

$$\sigma_P^2(t) = (1 - e^{-d_1 t}) (m e^{-d_1 t} + \lambda [1 + \lambda \Omega (1 + e^{-d_1 t})]), \quad (28)$$

where

$$\lambda = \frac{v'_1 f_0 k_0 n}{d'_0 d_1 l}, \quad (29)$$

and

$$\Omega = \frac{d_1}{d'_0 + d_1} \left[\eta_{33}^2 + \frac{d'_0}{d_1 + v_0} \left(\eta_{23}^2 + \frac{v_0}{d_1 + l} \eta_{13}^2 \right) \right]. \quad (30)$$

Note that Ω is a measure of the mRNA fluctuations, $l = f_0 + b_0 + k_0$, and that η_{ij}^2 is given by

$$\eta_{ij}^2 = \frac{\mu_{n_i n_j} - \mu_{n_i} \mu_{n_j}}{\mu_{n_i} \mu_{n_j}}. \quad (31)$$

The effects of gene replication are incorporated in the second stage of the derivation. The number of proteins at the beginning of each cell cycle is determined by the time evolution of the system during the cycle of a parent cell. To assess the time evolution of protein molecules during the cell cycle, the probability $q_{n|m}(t)$ of having n proteins at time t , given that there were m proteins at time $t = 0$ is defined and the probability-generating function $Q_m(z, t)$ for this distribution is constructed. By definition, the generating function has the form

$$Q_m(z, t) = \sum_n q_{n|m}(t) z^n. \quad (32)$$

The equation can be expanded around $z = 1$ which yields

$$Q_m(z, t) \cong 1 + (z - 1)\mu_P + \frac{1}{2}(z - 1)^2[\mu_{P^2} - \mu_P] + \dots \quad (33)$$

This function can be determined up to the necessary level by means of equations $\mu_P(t)$ and $\sigma_P^2(t)$. Using Eq. 33, it is obtained that

$$Q_m(z, t) = Q_0(z, t) [1 - e^{-d_1 t} + z e^{-d_1 t}]^m. \quad (34)$$

Because the gene replication occurs at time $t = t_d$, two different forms of $Q_m(z, t)$ have to be considered: $Q_m^{(1)}(z, t)$ which is valid when the gene number is n , and $Q_m^{(2)}(z, t)$ which is valid when the gene number is $2n$. Thus

$$Q_m^{(i)}(z, t) = Q_0^{(i)}(z, t) [Y + z(1 - Y)]^m, \quad (35)$$

where $Y = 1 - e^{-d_1 t}$. Now it is possible to proceed to the third stage of the derivation where cell division is included.

The third stage incorporates cell division. Cell division is in the model assumed to occur at fixed intervals given by the division time T_d . When $t = T_d$ it is assumed that each protein has a 50 % probability of being kept in this cell (symmetric division) and the probability of having n proteins immediately after the division is the binomial

$$\binom{m}{n} 2^{-m} \quad (36)$$

given that there are m proteins just before cell division. The transfer probability from one cell cycle to another can be constructed by combining the binomial distribution with the protein distribution derived earlier (Eq. 24). After many divisions, the protein number tends to a limit cycle and expressions for the mRNA and protein mean and coefficient of variation can be obtained in the limit $d_1/d'_0 \ll 1$. Through a fairly complicated set of steps, it can be shown [33] that the mean mRNA number before gene duplication ($t < t_d$), and the mRNA coefficient of variation are given by

$$\mu_{mRNA} = \frac{f_0 k_0 n}{d'_0 l} \quad (37)$$

$$\eta_{mRNA}^2 = \frac{1}{\mu_{mRNA}} - \frac{d'_0 v_0 (d'_0 + l + v_0)}{n(d'_0 + l)(l + v_0)(d'_0 + v_0)}. \quad (38)$$

The mean protein number and coefficient of variation in protein number as functions of time can be derived as

$$\mu_P(t) = \frac{v'_1}{d_1} \mu_{mRNA} \phi_0(t) \quad (39)$$

$$\eta_P^2(t) = \frac{1}{\mu_P(t)} + \frac{1}{\mu_{mRNA}} \left[1 - \frac{f_0 k_0}{l^2} \right] \frac{d_1}{d'_0} \phi_1(t), \quad (40)$$

where

$$\phi_0(t) = \begin{cases} 1 - \frac{e^{-d_1(T-t_d+t)}}{2-e^{-d_1T}}, & \text{for } 0 \leq t \leq t_d \\ 2 \left[1 - \frac{e^{-d_1(t-t_d)}}{2-e^{-d_1T}} \right], & \text{for } t_d \leq t \leq T \end{cases} \quad (41)$$

and

$$\phi_1(t) = \frac{2 - e^{-d_1 T}}{2 + e^{-d_1 T}} \times \begin{cases} \frac{4 - e^{-2d_1 T} - 2e^{-2d_1 t} - e^{-2d_1(T+t-t_d)}}{(2 - e^{-d_1 T} - e^{-d_1(T+t-t_d)})^2}, & \text{for } 0 \leq t \leq t_d \\ \frac{4 - e^{-2d_1 T} - e^{-2d_1 t} - 2e^{-2d_1(t-t_d)}}{2(2 - e^{-d_1 T} - e^{-d_1(t-t_d)})^2}, & \text{for } t_d \leq t \leq T. \end{cases} \quad (42)$$

In Eqs. (41) and (42), t_d and T denote the gene replication time and the cell division time, respectively.

It is noted that Eqs. (37) and (38) are time independent and that the value of the mean is twice this result after gene replication occurs (i.e. when $t > t_d$). The time independence follows from the assumption that the RNA is in a quasi-steady state proportional to the gene copy number n , and that all other time dependencies are absorbed into the protein distribution.

Our simulation results are compared to the corresponding steady-state and time-dependent analytical solutions (Figs. 4-6). In these simulations, we use the same assumptions as in [33]; the cell volume increases linearly up to time of cell division T , gene replication occurs at $t_{rep} = 0.4T$ and cell division is symmetric with binomial partitioning of molecules. Simulated protein number and concentration, as well as mRNA number dynamics, for single cells (Fig. 4) are comparable with the simulation results obtained by Swain *et al.* [33]. Figures 5 and 6 further compare population characteristics estimated from simulations to those predicted by the corresponding steady-state analytical solutions. Both RNA ($\mu_{mRNA}(n)$ and $\eta_{mRNA}^2(n)$, Fig. 5) and protein ($\mu_P(t)$ and η_P^2 , Fig. 6) characteristics are in good agreement with the analytical results (Eqs. (37)-(42)).

6. Simulating complex population dynamics

6.1. Asymmetric Cell Division

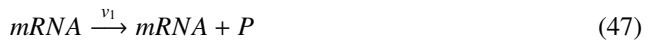
To investigate sources of external variability in eukaryotic gene expression, Volkson *et al.* [35] combined computational modelling with fluorescence data. As part of this study, the authors simulated the distribution of cell sizes within a population of *Saccharomyces cerevisiae* (budding yeast). In these simulations, cells grew exponentially until they reached a critical volume V_c where they divide. The volume at division was drawn from a normal distribution with a mean specified as a function of genealogical age and coefficient of variation 0.15. Following division, the mother cell retained 70 % of the volume ($V_0 = 0.7V_c$) while daughter cells were correspondingly smaller ($V_0 = 0.3V_c$). The resulting distribution of cell sizes obtained from an initial population of 1000 cells allowed to grow to 100000 cells was found to be in agreement with experimental and analytical results [35].

The model by Volkson *et al.* [35] is ideally suited for benchmarking the constant-number MC method. As in Volkson *et al.* [35], we first simulated the growth of a population initially consisting of 1000 cells and obtained the steady-state size distribution once the population grew to 100000 cells (Fig. 7a). Next, we repeated the simulations using the constant-number MC method to estimate the size distribution from a representative sample (8000 cells) of this cell population (Fig. 7b). A plot of the probabilities for the sample population against the probabilities of the ‘true’ population shows that the difference between these variables is minimal (Fig. 7c). These results compliment previous studies [16, 18, 21, 22, 31] demonstrating the ability of the constant-number MC method to capture complex population dynamics.

6.2. Bet-Hedging Cell Populations

One of the most interesting potential applications of the simulation algorithm described in Section 4 is investigations of interactions between environmental changes, population dynamics and gene expression in individual cells. For example, it can be used to study the optimization of fitness in fluctuating environments, which is a classic problem in evolutionary and population biology [4, 17, 29, 32]. Acar *et al.* [1] experimentally investigated how stochastic switching between phenotypes in changing environments affected growth rates in fast and slow-switching populations by using the galactose utilization network in *Saccharomyces cerevisiae*. Specifically, a strain was engineered to randomly transition between two phenotypes (*ON* and *OFF*) characterized by high or low expression of a gene encoding the Ura3 enzyme necessary for uracil biosynthesis. Each phenotype was designed to have a growth advantage over the other in one of two environments. In the first environment (E_1) which lacks uracil, cells in the *ON* phenotype have an advantage. In the second environment (E_2), cells in the *OFF* phenotype have an advantage due to the presence of a drug (5-FOA) which is converted into a toxin by the Ura3 enzyme. In this environment, which also contains uracil, cells expressing Ura3 will have low viability while cells not expression Ura3 will grow normally.

Models of gene expression often describe the promoter T as being in one of two states: a repressed state T_R (basal level of gene expression) or an active state T_A (upregulated level of gene expression) corresponding respectively to *OFF* and *ON* phenotypes. This can be described by the following biochemical reaction scheme [11]:



where Eq. (43) describes the transitions to the T_A and T_R promoter states at rates k_1 and k_2 respectively, Eqs. (44) and (45) the mRNA production from the T_A (at a rate $v_{0,A}$) and T_R (at a rate $v_{0,R}$) states respectively, Eq. (47) the protein production from mRNA at a rate v_1 , and Eqs. (46) and (48) respectively the mRNA (at a rate d_0) and protein (at a rate d_1) degradation.

We first follow the approach that was used in Acar *et al.* [1] to describe the dynamics of phenotype switching, where cells are in either the *ON* or the *OFF* state:



In this scenario, cells randomly switch between the high and low expressing states at rates k_1 and k_2 (see [1] for parameter values corresponding to slow and fast-switching cells). The growth rate (Eq. (2)) of fit cells was set higher than the corresponding growth rate for unfit cells in the same environment. In order to avoid synchronization in the population level dynamics, we set $V_{div} = 2V_0 + \xi$, where ξ is a small random number drawn from a normal distribution with zero mean and 0.2 variance.

Figure 8 shows the growth rates obtained from simulations of slow and fast-switching cell populations, where cells were transferred from E2 to E1, and vice versa, at $t = 0$. Growth rates show a transition period and a steady-state region. In agreement with experiments (see Acar *et al.* [1]), fast-switching cells were found to recover from the effect of environment change faster than slow-switching cells but have a lower steady-state growth rate.

Next we implemented the full model of gene expression described by Eqs. (43)-(48). The fitness w_k of each cell k , which is here defined as a function of the environment and cellular protein concentration $[P]$, was described by a Hill function

$$w_k(E, P) = \begin{cases} \frac{[P]^n}{[P]^n + K^n}, & \text{if } E = E1 \\ \frac{K^n}{K^n + [P]^n}, & \text{if } E = E2. \end{cases} \quad (50)$$

This equation describes partitioning of cells into fit ($w_k(E, P) > 0.5$) and unfit ($w_k(E, P) < 0.5$) phenotypes corresponding to whether or not their $[P]$ in a particular environment is above or below a particular value given by the Hill coefficient K . The volume of each cell was described by Eq. (2), except here $\tau_0 = \tau_\phi/w$, where τ_ϕ is the cell division time in absence of selection. To incorporate the effect of fitness on gene expression, the value of transcription rate parameter v_0 depended on whether or not a cell was fit in either $E1$ or $E2$ (see Fig. 9 for parameters).

The population distributions obtained for this model are shown in Figure 9. Specifically, we first obtained the steady-state protein concentration distributions for cells in $E1$ and $E2$ (Fig. 9a and 9b respectively). Here, the majority of cells either fell within a distribution centered at higher value characterizing the *ON* cells, or a distribution centered at a lower value of P characterizing the *OFF* cells, in $E1$ or $E2$ respectively. The rest of the cells fell within the distribution capturing the unfit subpopulation in both environments. These results were found experimentally in [1] and are expected, as higher levels of the uracil enzyme are either favorable or unfavorable with respect to the fitness of the cells depending on the environment. Next, the time-dependent population distributions after the transition to $E1$ from $E2$, and vice versa, were obtained (Fig. 9a and 9b respectively). Here, the dynamics of the two distinct subpopulations of cells in transition between the steady-states are visible. As time progresses after the environmental transition, less and less of the cells are in the unfit state (*ON* in Fig. 9a and *OFF* in Fig. 9b), as the cells in the more fit state (*OFF* in Fig. 9a and *ON* in Fig. 9b) grow and divide at a faster rate and therefore come to dominate the population in terms of absolute numbers.

7. Conclusions

We have presented a framework for the stochastic simulation of heterogeneous population dynamics. The accuracy of the method was verified by comparing simulation results of stochastic gene expression and population dynamics with corresponding steady-state and time-dependent analytical solutions and experimental results. Parallel execution of the algorithm was found to significantly decrease run-times in comparison to simulations run on a single processor, and did not introduce errors in numerical results.

The algorithm was also shown to be capable of simulating and capturing the dynamics of a cell population in a fluctuating environment, where phenotypic variability strongly influences gene expression dynamics. Agreement between this framework and the experimental and theoretical results obtained using a deterministic reaction-rate method in Acar *et al.* [1], serves as a further benchmark for the proposed method. Furthermore, the algorithm's ability to capture the steady-state and time-independent phenotypic distributions in this system exemplifies the utility of this approach, as these distributions cannot be obtained using a deterministic framework.

Current cellular population simulation methods, including the present algorithm, treat the extracellular environment as homogeneous (e.g. the spatial-temporal concentration profile of a nutrient required for growth is held constant). This prohibits, for example, the inclusion of competition for a limiting resource in the present implementation. However, it is possible to model feedback between cells and their environment. The simplest approach would be to assume that the environment is constant over short time intervals. The change in total population cell volume at the end of each interval could then be used to calculate how much nutrients have been consumed and the parameters describing the environment adjusted accordingly. Since the time intervals would have to be sufficiently short so that the change in concentration of the nutrient during any particular interval is negligible, the computational workload would increase substantially. The focus of future work will be on developing and benchmarking accurate and efficient augmentations that permit population simulators to handle these and other more complex scenarios.

Acknowledgments

This work was supported financially by the National Science and Engineering Research Council of Canada (NSERC), the Canadian Institutes of Health Research (CIHR), the Academy of Finland (application number 129657, Finnish Programme for Centres of Excellence in Research 2006-2011, and 124615), and the Tampere Graduate School in Information Science and Engineering (TISE).

Author Contributions

D.C., D.F., and M.K. developed the serial, and D.C. the parallel, versions of the algorithm; D.C. performed the stochastic simulations; D.C. and J.I. benchmarked the algorithm; D.C., M.K., and J.I. wrote the paper; M.K. supervised the study.

References

- [1] M. Acar, J.T. Mettetal, A. van Oudenaarden, Stochastic switching as a survival strategy in fluctuating environments, *Nat. Genet.* 40 (2008) 471-475.
- [2] D. Adalsteinsson, D. McMillen, T.C. Elston, Biochemical Network Stochastic Simulator (BioNetS): software for stochastic modeling of biochemical networks, *BMC Bioinfo.* 5 (2004) 24.
- [3] B.J. Brewer, E. Chlebowicz-Sledziewska, W.L. Fangman, Cell Cycle Phases in the Unequal Mother/Daughter Cell Cycles of *Saccharomyces cerevisiae*, *Mol. Cell. Biol.* 4 (1984) 2529-2531.
- [4] D. Cohen, Optimizing reproduction in a randomly varying environment, *J. Theor. Biol.* 12 (1966) 119-129.
- [5] D.L. Eager, J. Zahorjan, E.D. Lazowska, Speedup Versus Efficiency in Parallel Systems, *IEEE Trans. Comput.* 38 (1989) 408-423.
- [6] D. Fraser, M. Kaern, A chance at survival: gene expression noise and phenotypic diversification strategies, *Molec. Microbiol.* 71 (2009) 1333-1340.
- [7] D.T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, *J. Comput. Phys.* 22 (1976) 403-434.
- [8] D.T. Gillespie, Exact stochastic simulation of coupled chemical reactions, *J. Phys. Chem.* 81 (1977) 2340-2361.
- [9] D.T. Gillespie, Stochastic Simulation of Chemical Kinetics, *Annu. Rev. Phys. Chem.* 58 (2007) 35-55.
- [10] D.T. Gillespie, Approximate accelerated stochastic simulation of chemically reacting systems, *J. Chem. Phys.* 115 (2001) 1716-1733.
- [11] M. Kaern, T.C. Elston, W.J. Blake, J.J. Collins, Stochasticity in gene expression, *Nat. Rev. Genet.* 6 (2005) 451-464.
- [12] B.B. Kaufmann, A. van Oudenaarden, Stochastic gene expression: from single molecules to the proteome, *Curr. Opin. Genet. Dev.* 17 (2007) 107-112.
- [13] T.B. Kepler, T.C. Elston, Stochasticity in Transcriptional Regulation, *Biophys. J.* 81 (2001) 3116-3136.
- [14] A.M. Kierzek, STOCKS: STOChastic Kinetic Simulations of biochemical systems with Gillespie algorithm, *Bioinf.* 18 (2002) 470-481.
- [15] M. Kostoglou, A.J. Karabelas, Evaluation of zero-order methods for simulation particle coagulation, *J. Colloid Interface Sci.* 163 (1994) 420-431.
- [16] K. Lee, T. Matsoukas, Simultaneous coagulation and break-up using constant-N Monte Carlo, *Powder Technol.* 110 (2000) 82-89.
- [17] R. Levin, *Evolution in Changing Environments: some Theoretical Explorations*, Princeton University Press, Princeton, 1968.
- [18] Y. Lin, K. Lee, T. Matsoukas, Solution of the population balance equation using constant-number Monte Carlo, *Chem. Eng. Sci.* 57 (2002) 2241-2252.
- [19] T. Lu, D. Volfson, L. Tsimring, J. Hasty, Cellular growth and division in the Gillespie algorithm, *Syst. Biol.* 1 (2004) 121-128.
- [20] N. Maheshri, E.K. O'Shea, Living with noisy genes: how cells function reliably with inherent variability in gene expression, *Annu. Rev. Biophys. Biomol. Struct.* 36 (2007) 413-434.
- [21] N.V. Mantzaris, Stochastic and deterministic simulations of heterogeneous cell population dynamics, *J. Theor. Biol.* 241 (2006) 690-706.
- [22] N.V. Mantzaris, From Single-Cell Genetic Architecture to Cell Population Dynamics: Quantitatively Decomposing the Effects of Different Population Heterogeneity Sources for a Genetic Network with Positive Feedback Architecture, *Biophys. J.* 92 (2007) 4271-4288.
- [23] J. Paulsson, Summing up the noise in gene networks, *Nature* 427 (2004) 415-418.
- [24] D. Ramkrishna, The status of population balances, *Rev. Chem. Engng.* 3 (1985) 49-95.
- [25] A.S. Ribeiro, D.A. Charlebois, J. Lloyd-Price, *CellLine*, a stochastic cell lineage simulator, *Bioinf.* 23 (2007) 3409-3411.
- [26] M. Roussel, R. Zhu, Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression, *Phys. Biol.* 3 (2006) 274-284.
- [27] N. Rosenfeld, T.J. Perkins, U. Alon, M.B. Elowitz, P.S. Swain, A Fluctuation Method to Quantify In Vivo Fluorescence Data, *Biophys. J.* 91 (2006) 759-766.
- [28] M.S. Samoilov, G. Price, A.P. Arkin, From fluctuations to Phenotypes: The Physiology of Noise, *Sci. STKE* 366 (2006) re17.
- [29] W.M. Schaffer, Optimal efforts in fluctuating environments, *Am. Nat.* 108 (1974) 783-790.
- [30] V. Shahrezaei, P.S. Swain, Analytical distributions for stochastic gene expression, *PNAS* 105 (2008) 17256-17261.
- [31] M. Smith, T. Matsoukas, Constant-number Monte Carlo simulation of population balances, *Chem. Eng. Sci.* 53 (1998) 1777-1786.
- [32] S.C. Stearns, Life-history tactics: a review of the ideas, *Q. Rev. Biol.* 51 (1976) 3-47.
- [33] P.S. Swain, M.B. Elowitz, E.D. Siggia, Intrinsic and extrinsic contributions to stochasticity in gene expression, *PNAS* 99 (2002) 12795-12800.

- [34] J.J. Tyson, O.J. Diekmann, Sloppy size control of the cell division cycle, *Theor. Biol.* 118 (1986) 405-426.
- [35] D. Wolfson, J. Marciniak, W.J. Blake, N. Ostroff, L.S. Tsimring, J. Hasty, Origins of extrinsic variability in eukaryotic gene expression, *Nature* 439 (2006) 861-864.

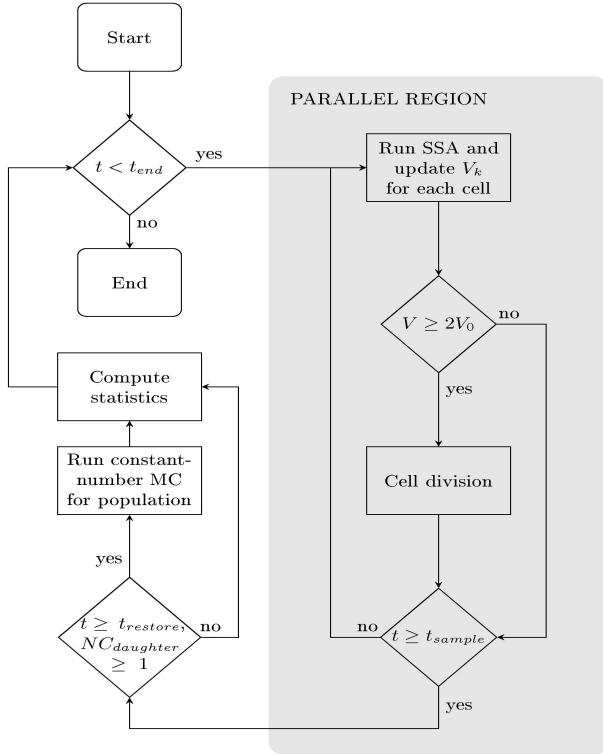


Figure 1: Flow diagram of the present algorithm for the parallel stochastic simulation of gene expression and heterogeneous population dynamics.

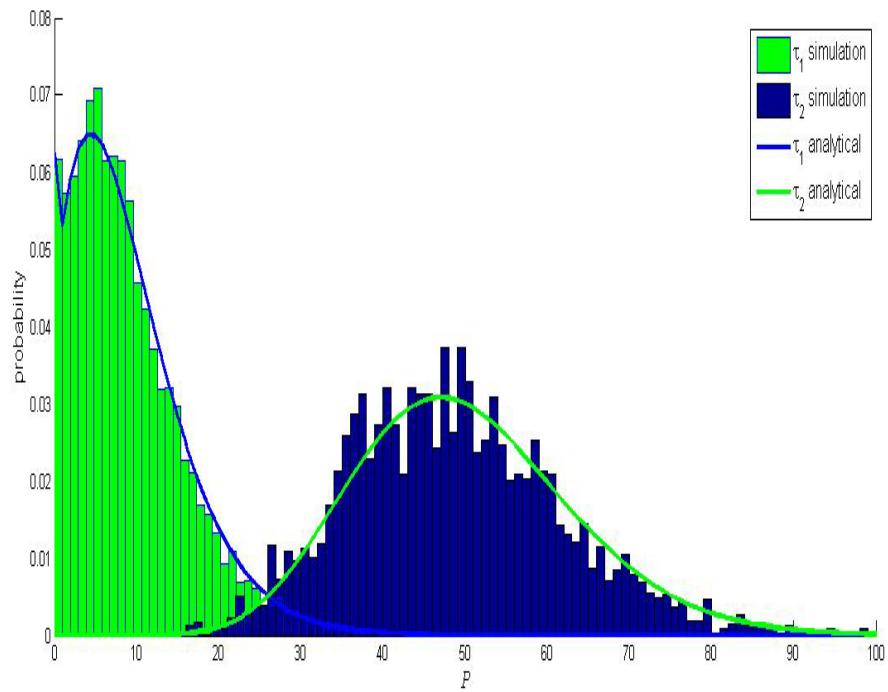


Figure 2: Simulation results and time-dependent analytical solutions of a two-stage model of gene expression [30]. The distribution of protein numbers for a population of cells at two different dimensionless times, $\tau = 0.2$ and $\tau = 10$, is shown.

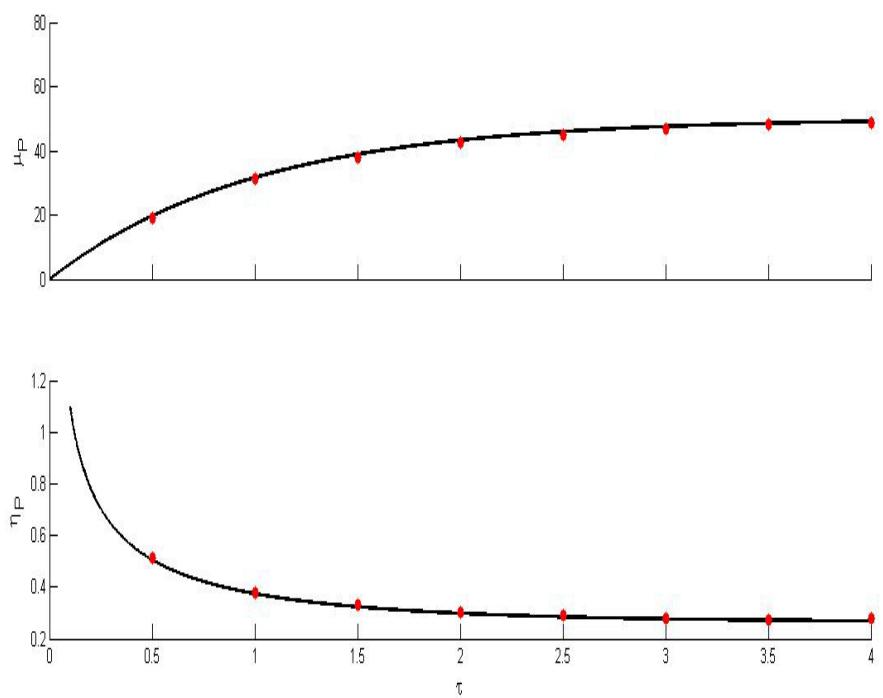


Figure 3: Simulation results and time-dependent analytical solutions of a two-stage model of gene expression [30]. Mean protein μ_P (top) and noise η_P (bottom) are plotted as a function of dimensionless time τ . Red dots indicate simulation results and black curves analytical solutions [30].

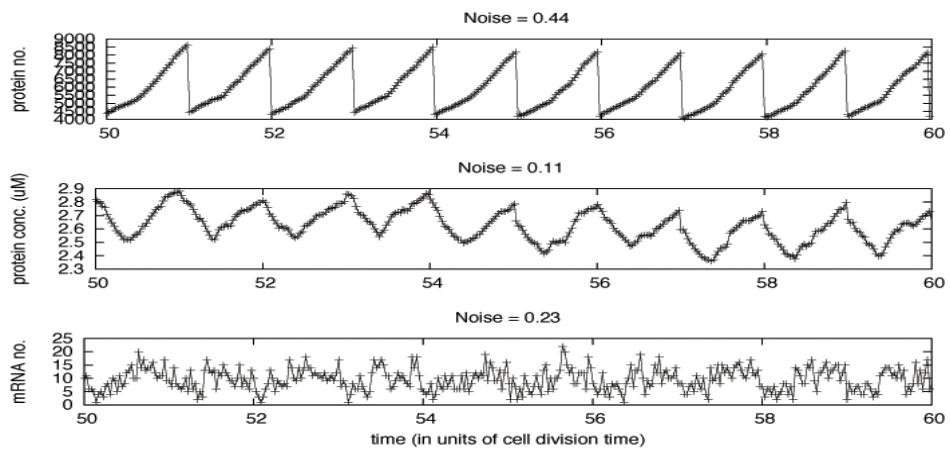


Figure 4: Time series of a single cell within a growing and dividing population. Protein number (top) and concentration (middle), and mRNA number (bottom), were obtained and found to be in agreement with a model of translation provided in [33]. Gene duplication occurs every $t_d = 0.4T$ into the cell cycle and results in an increased rate of protein production until the next cell division event where the number of genes prior to duplication is restored.

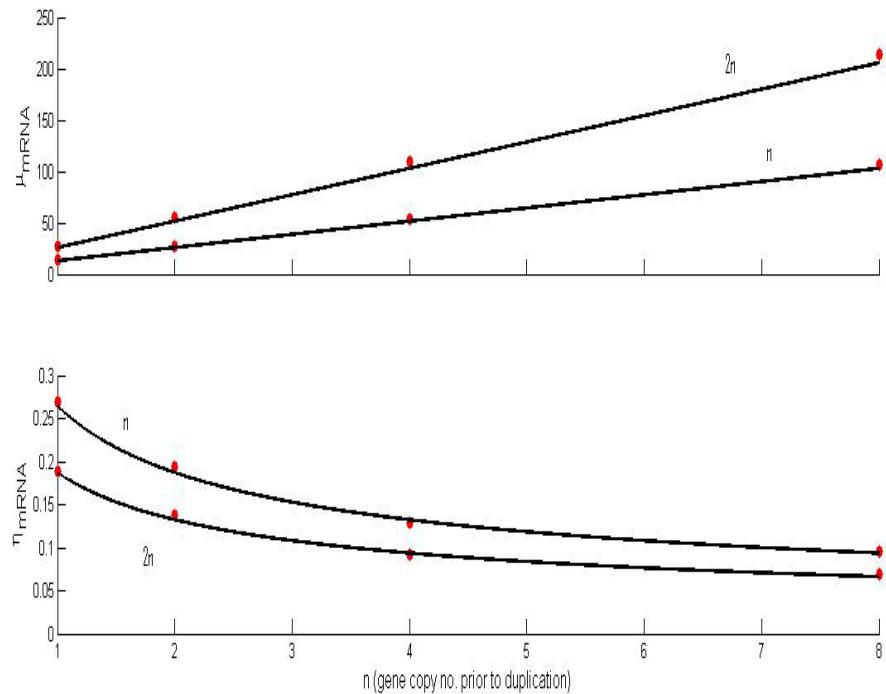


Figure 5: Comparison of simulation results and analytic solutions. Mean mRNA values are plotted as a function of gene copy number n (top). The noise in mRNA number is also plotted as a function of n (bottom). Note that mean mRNA values increase and the noise decreases after gene duplication as expected. Black curves indicate analytical values [33] and red dots simulation results.

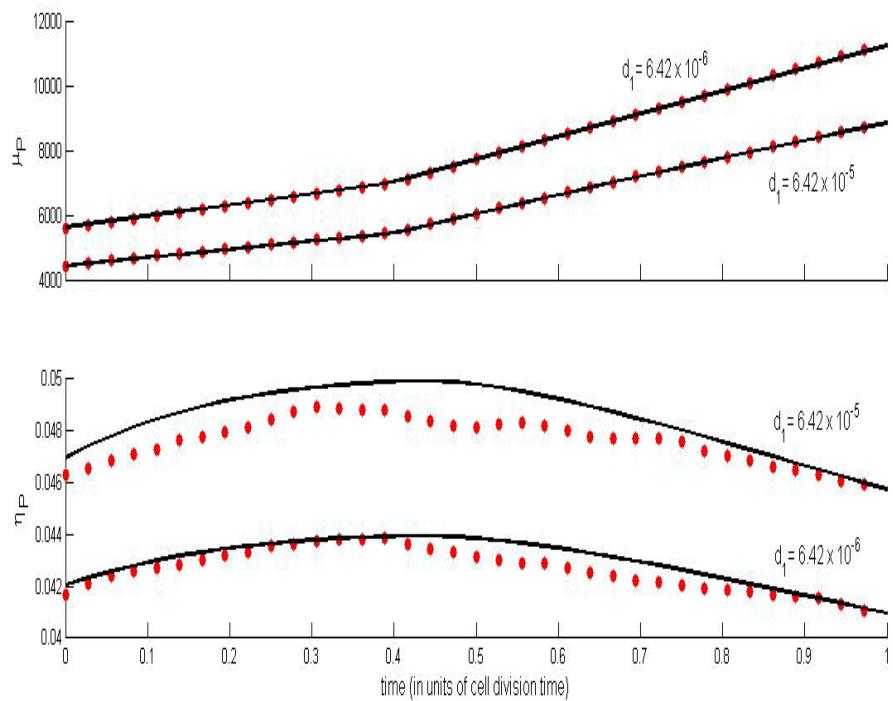


Figure 6: Comparison of simulation results and analytic solutions. Mean protein number (top) and noise (bottom) as a function of time t for two different values of the protein degradation parameter d_1 . Note the increase in protein production rate and decrease in noise levels that occurs after gene duplication at $t = 0.4$. Red dots indicate simulation results and black curves analytical values [33].

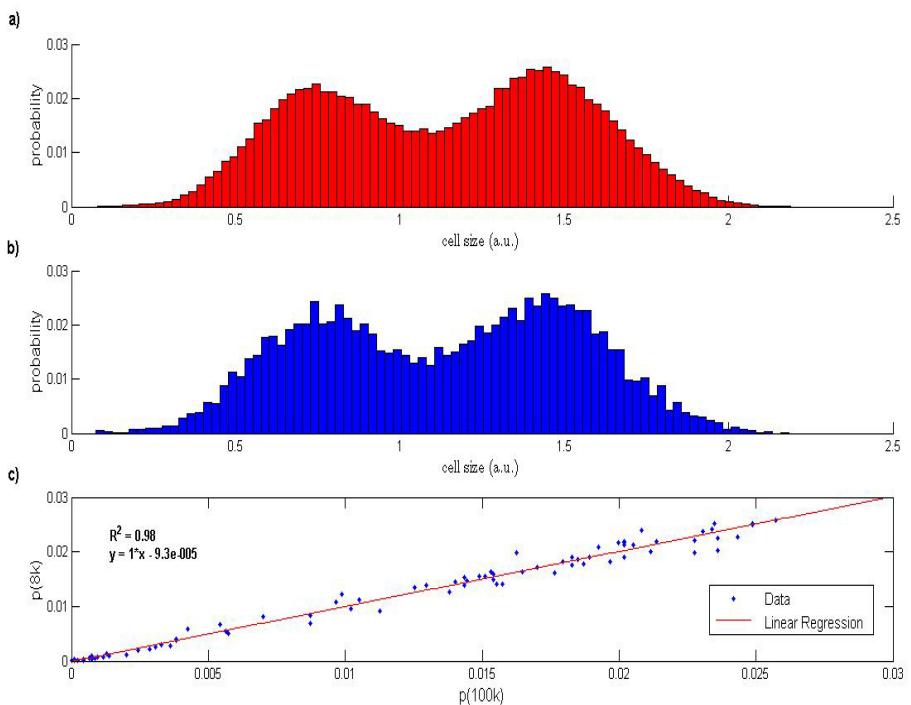


Figure 7: Simulation of a stochastic population dynamics model [35] of a *Saccharomyces cerevisiae* population undergoing stochastic (size at division) and asymmetric (partitioning of cell volume) division. (a) Steady-state distribution of cell sizes for a population of 100000 cells. (b) Steady-state size distribution of a representative sample (8000 cells) obtained using the constant-number Monte Carlo method [18, 31] of the ‘true’ population shown in (a). (c) Plot of the probabilities population shown in (b) against the probabilities of the population shown in (a) along with linear regression.

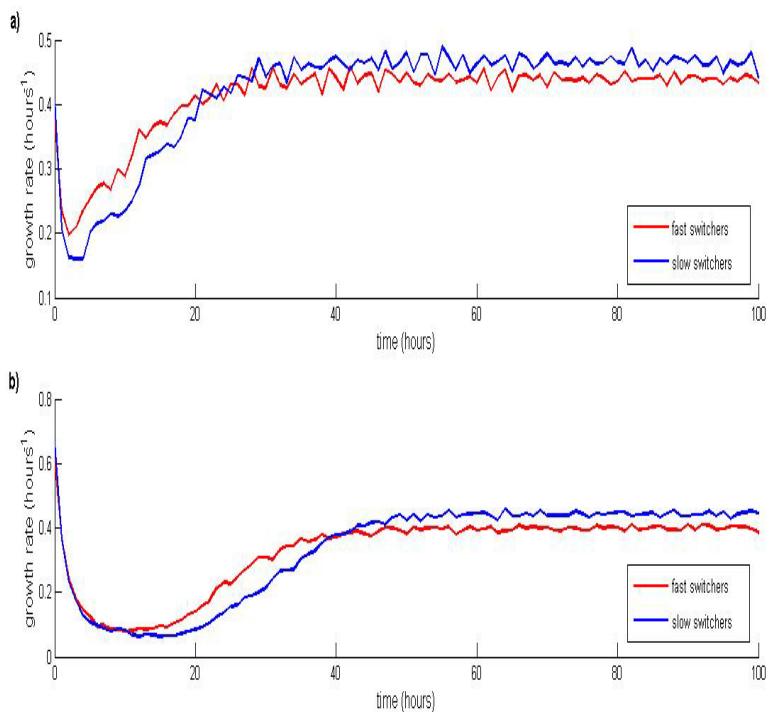


Figure 8: Simulations of populations of slow and fast-switching cells. (a) Growth rates of cells transferred from an environment containing uracil and 5-FOA (E2) to one containing no uracil (E1) at $t = 0$. (b) Growth rates of cells transferred from E1 to E2 at $t = 0$. Note that the transient before the steady-state region is shorter in (a) than in (b), and that fast-switching cells recover faster from the environment change but slow-switching cells have a higher steady-state growth, in agreement with experimental results found in [1].

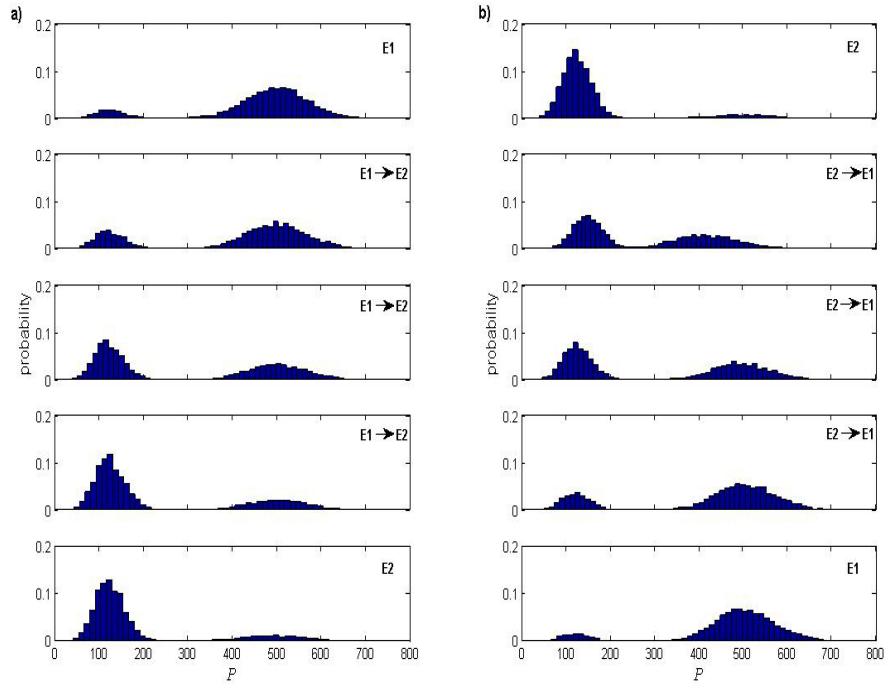


Figure 9: Simulations of environmental effects on phenotypic distribution. (a) Steady-state (top and bottom figures) and time-dependent (middle figures) protein distributions of cells resulting from an environment change from E1 to E2. (b) Steady-state (top and bottom figures) and time-dependent (middle figures) protein distributions of cells resulting from an environment change from E2 to E1. Note that when a sufficient amount of time has elapsed after the environmental transition from either E1 to E2 or vice versa, cells with either the OFF or ON phenotype proliferate, respectively, in agreement with experimental results found in [1]. The following parameters were used (units s^{-1}): $d_0 = 0.005$, $v_1 = 0.1$, $d_1 = 0.008$, $K = 200$, $n = 10$. In E1 $v_{0,A} = 0.2$ for fit cells and $v_{0,R} = 0.05$ for unfit cells - vice versa in E2. Additionally τ_ϕ was set to the mean doubling time (MDT) of 1.5 hours for *Saccharomyces cerevisiae* [3].