

Improved Inference of Mutation Rates

II. Generalization of the Luria–Delbrück Distribution for Realistic Cell-Cycle Time Distributions

Mihaela Oprea

Department of Computer Science, University of New Mexico, Albuquerque, New Mexico 87545

E-mail: mihaela@cs.unm.edu

and

Thomas B. Kepler

Biomathematics Graduate Program, Department of Statistics, North Carolina State University,

Raleigh, North Carolina 27695-8203; and Santa Fe Institute, Santa Fe, New Mexico 07501

E-mail: kepler@santafe.edu

Received December 20, 1999

In the first paper of this series (Kepler and Oprea, *Theor. Popul. Biol.* 2001) we found a continuum approximation of the Luria–Delbrück distribution in terms of a scaled variable related to the proportion of mutants in the culture. Here we show that the Luria–Delbrück distribution is inaccurate when realistic division processes are being considered due to the non-Markovian character of the cell cycle. We derive the expectation of the proportion of mutants in the culture for arbitrary cell-cycle time distributions. We then introduce a two-parameter generalization of the continuum Luria–Delbrück distribution for two of the more commonly used cell-cycle time distributions: gamma and shifted exponential. We obtain the generalized distribution by defining a map from the actual parameters to “effective” parameters. The effective mutation rate is obtained analytically, while the effective population size is obtained by fitting simulation data. Our simulations show that the second parameter depend mostly on the coefficient of variation of the cell-cycle time distribution. © 2001 Academic Press

Fluctuation analysis was introduced by Luria and Delbrück (1943) to address the question of whether the mutations that arise in bacterial cultures placed under strong selective conditions appear as a consequence of the selective agent or are only revealed by it. The theoretical prediction of the number of mutants, M , in a culture of size n , grown under nonselective conditions became known as the Luria–Delbrück distribution. Numerous studies have been devoted to the mathematical treatment of this

distribution (Luria and Delbrück, 1943; Kendall, 1948; Lea and Coulson, 1949; Bartlett, 1978; Stewart *et al.*, 1990; Sarkar *et al.*, 1992; Jones *et al.*, 1994). In the companion to this paper (Kepler and Oprea, 2001), we derived an integral representation of this distribution.

A key assumption to the mathematical tractability of the Luria–Delbrück distribution is that, at each time step, all cells have an equal probability of dividing. This amounts to assuming that the cell-cycle time is exponentially

distributed. The impact of the cell-cycle time distribution on the distribution of mutants and the fact that the Luria–Delbrück distribution is incorrect for realistic cell-cycle time models was recognized by Kendall (1952). His general treatment, however, led to intractable coupled nonlinear integral equations and was not pursued toward applications.

Fluctuation analysis has been extremely useful in the analysis of mutational processes. The scope of its applications now encompasses eukaryotic systems as well, such as mutagenesis leading to cancers (see for a review Kendal and Frost (1988)). As the precision of the experimental methods increases, biases in the statistical estimators based on uncorrected Luria–Delbrück distribution can become substantial. The aim of this study is to generalize the Luria–Delbrück distribution for more realistic cell-cycle time distributions and to derive improved methods for mutation rate estimation in these situations. The basic culture dynamics is otherwise assumed to be the same as for the Luria–Delbrück distribution: mutations are acquired at cell division, they are irreversible, and there is no cell death.

1. LURIA–DELBRÜCK DISTRIBUTION, TREES, AND TREE SHAPES

We first describe the processes underlying the Luria–Delbrück (Luria and Delbrück, 1943) distribution of mutants. A culture is seeded with n_0 (generally 1) cells. The culture then grows due to cell division up to n cells. Within a small time interval, δt , each cell has a constant, and small, probability of replicating, identical for all cells currently present in the culture. There is no cell death; cells are only “lost” when they divide to form two daughter cells. When a wild-type cell divides, each of its two daughter cells has a probability μ of undergoing a mutation that changes its phenotype. Mutants do not revert to the wild-type phenotype, and thus all progeny of a mutant cell will be mutants as well. The culture, or, more accurately, a sample of it, is then placed under selective conditions to reveal the mutants. This means that, while the culture is growing, mutant cells do not have a selective advantage and grow at the same rate as wild-type cells. The number of mutants in a final culture of size n is denoted by M .

The assumption that all cells have a constant probability of dividing per unit time, which is used for deriving the Luria–Delbrück distribution, is equivalent to assuming an exponentially distributed interdivision time. This gives a strictly decreasing probability of replication as a

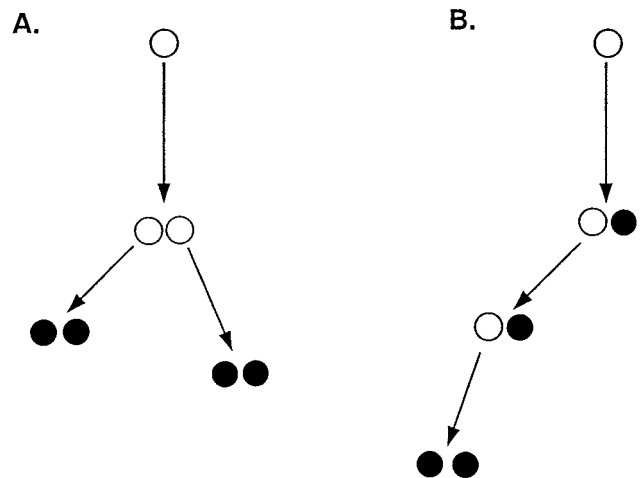


FIG. 1. Genealogical trees that can be realized in a four-cell culture: the first tree is symmetrical (A), the second asymmetrical (B). The cells that are present in the final culture are represented by filled circles. Open circles represent cells that have been present at some point in the growth of the culture, but have since undergone division. The edges denote life times of individual cells.

function of the age of the cell. However, any cell that has just divided will have a very small probability of dividing again soon. That this actually makes a difference in the distribution of mutants in a population of a given size is seen clearly when considering the case of a population of four cells arising from a single ancestor.

For such a culture, there are only two topologically distinct genealogical trees (Fig. 1). In the balanced tree (Fig. 1A), each of the four cells has two divisions in its history and therefore has the same probability of being a mutant: 2μ for μ small. In the second tree, which is skewed (Fig. 1B), the four cells experienced one, two, three, and three division events, respectively, with probability μ of mutation at each division, for a mean mutant frequency of $9\mu/4$. The variance in the mutant frequency is also higher in the case of the skewed tree. To determine the average proportion of mutants in a culture of a given size n , we have to consider the ensemble of tree topologies which could underlie the cell culture. Where the cell-cycle time distribution comes into play is in the probability of realizing different tree topologies. For example, when the cell-cycle time distribution is exponential, the skewed tree is realized twice as often as the balanced tree. At the other extreme, if cell replications are synchronous,¹ then only the balanced tree can be realized.

¹ The prime example of synchronous replication is polymerase chain reaction (PCR), although the replicating entities are strands of nucleic acids, rather than cells.

2. MEAN PROPORTION OF MUTANTS AS A FUNCTION OF THE CULTURE SIZE n

When the cell-cycle time has an arbitrary distribution, we cannot hope to obtain an analytical form for the distribution of mutants. We can, however, calculate the expected value of the number (or proportion) of mutants in the culture. The basic approach is to determine the mean number of divisions that a cell in the final culture experienced and then calculate the probability that mutation occurred at (at least) one of these divisions. The mean number of divisions that took place during the period of time, t , in which the culture grows from n_0 to n cells is given by the ratio between t and the average age of a cell at division (the cell-cycle time). We therefore set to calculate these quantities.

Let us then denote the age at which any individual cell divides by a random variable, A . The age of a cell at division is essentially the interdivision time, or the cell-cycle time. For a cell randomly chosen at birth, the age at division is described by the cumulative distribution function Ψ defined by $\text{Prob}(A < a) = \Psi(a)$, and, equivalently, by its density function $\psi(a) = d\Psi(a)/da$.

At division, the parent is lost and two cells of age zero are created. Consider a population of such cells. If the number of cells of age a in the population is denoted by $y(a, t)$ where t is the absolute time, then the equation for loss of the parent cells by division is

$$(\partial_t + \partial_a) y(a, t) = -h(a) y(a, t). \quad (1)$$

The hazard function, $h(a)$, given by

$$h(a) = \frac{\psi(a)}{1 - \Psi(a)}, \quad (2)$$

denotes the probability that a cell divides at age a , given that it has not divided yet. Note that we can write

$$h(a) = -\frac{d}{da} \log(1 - \Psi(a)). \quad (3)$$

The rate of loss of parent cells is derived by integrating Eq. (1) over a , to obtain

$$\frac{d}{dt} \int_0^\infty da y(a, t) = y(0, t) - \int_0^\infty da h(a) y(a, t). \quad (4)$$

As each division results in the loss of the parent cell and the gain of two daughter cells of age 0, the total rate of production can be written as

$$\frac{d}{dt} \int_0^\infty da y(a, t) = \int_0^\infty da h(a) y(a, t). \quad (5)$$

This results in the production equation

$$y(0, t) = 2 \int_0^\infty da h(a) y(a, t). \quad (6)$$

We seek separable solutions to Eq. (1) of the form $y(a, t) = \omega(a) n(t)$. $\omega(a)$ is the stationary age distribution, and $n(t)$ is the total population size as a function of time. Substituting this form into Eq. (1), we obtain

$$\frac{dn(t)}{dt} \frac{1}{n(t)} = -h(a) - \frac{d\omega(a)}{da} \frac{1}{\omega(a)}. \quad (7)$$

Note that the right-hand side depends only on a , while the left-hand side depends only on t . Therefore, if this equation is to hold for all values of both t and a , then either side must be constant. If we call this constant γ , we have

$$\frac{dn(t)}{dt} = \gamma n(t) \quad (8)$$

and

$$\frac{d\omega(a)}{da} = [-h(a) - \gamma] \omega(a). \quad (9)$$

The solution for n is simply

$$n(t) = n(0) e^{\gamma t}, \quad (10)$$

while for ω we have

$$\omega(a) = \omega(0)(1 - \Psi(a)) e^{-\gamma a}. \quad (11)$$

Note that γ is the growth rate of the culture. The condition on γ is obtained by substituting Eq. (11) into Eq. (6):

$$1 = 2 \int_0^\infty da h(a)(1 - \Psi(a)) e^{-\gamma a}. \quad (12)$$

Now substituting Eq. (2) yields

$$\frac{1}{2} = \int_0^{\infty} da \psi(a) e^{-\gamma a}. \quad (13)$$

This is the eigenvalue equation for γ that we seek. Since $\psi(a)$ is the density function for cell-cycle time, we can write this last result as

$$E_{\psi}(e^{-\gamma a}) = \frac{1}{2}, \quad (14)$$

where E_{ψ} denotes expectation with respect to ψ .

Solving for γ , we can then obtain the stationary age distribution of cells in the culture. We now proceed to calculate the mean age of a cell at division using the density function for age, $\psi(a)$, but weighted by the proportion of cells of age a in the culture.

The proportion of cells of age a that divide at any given time is given by the hazard function, $h(a)$, weighted by the proportion of cells of age a that are present in the culture. We then have to normalize by the number of cells that divide, regardless of their age. Therefore, the average age of a cell at division is given by

$$E(a) = \frac{\int_0^{\infty} da a h(a) \omega(a)}{\int_0^{\infty} da h(a) \omega(a)}. \quad (15)$$

If we now substitute Eqs. (2) and (11) in the above expression, we obtain

$$E(a) = \frac{E_{\psi}(ae^{-\gamma a})}{E_{\psi}(e^{-\gamma a})}, \quad (16)$$

with E denoting the average, and E_{ψ} denoting the expectation with respect to ψ . In light of the definition of γ (Eq. (14)),

$$E(a) = 2E_{\psi}(ae^{-\gamma a}). \quad (17)$$

Knowing the growth rate of the culture we find the time required to reach size n when starting from n_0 cells:

$$t = \frac{\log(n/n_0)}{\gamma}. \quad (18)$$

Thus, the mean number of divisions, g , experienced by an individual cell when the culture size is n , is given by:

$$E(g; n) = \frac{t}{E(a)} = \frac{\log(n/n_0)}{\gamma E(a)}, \quad (19)$$

where $E(\cdot; n)$ denotes the expectation over an ensemble of genealogical trees with n leaves. We may now calculate the mean number of mutants by assuming that at each cell division, each of the daughters has a probability μ of becoming mutant (and therefore all of her daughters as well). The probability that no mutation occurs in g divisions is $(1 - \mu)^g$, so the probability that at least one mutation occurs is $1 - (1 - \mu)^g$. Then the probability of a cell being a mutant is $\sum_g (1 - (1 - \mu)^g) p(g)$, where $p(g)$ is the probability that the cell underwent g divisions. For small mutation rates, such that $\mu g \ll 1$, this expression can be approximated by $\mu E(g; n)$, that is, the probability that an individual cell is mutant is $\mu E(g; n)$. The mean number of mutants in the culture is thus

$$E(M; n) = \mu n E(g; n) = \mu n \frac{\log(n/n_0)}{\gamma E(a)}. \quad (20)$$

Defining the proportion of mutants in the culture as

$$X \equiv \frac{M}{n} \quad (21)$$

and writing

$$b \equiv \frac{1}{\gamma E(a)}, \quad (22)$$

the expectation of the proportion of mutants in a culture of size n is

$$E(X; n) = b \mu \log(n/n_0). \quad (23)$$

Thus, the expected proportion of mutants in the population depends on the initial (n_0) and final (n) culture size, the probability that an individual daughter cell mutates (μ), and a factor which is only a function of the cell-cycle time distribution (b). Observe that b represents the average number of divisions that take place in the genealogy of a cell when the culture grows by a factor of e , and Eq. (23) is only written in terms of the mutation rate per daughter cell per division and the average number of divisions in the genealogy of the cell.

We now calculate the parameter b for two of the more commonly used distributions for cell-cycle time. The first is the gamma distribution, which was used to fit bacterial cell-cycle data (Kelly and Rahn, 1932). Note that the exponential distribution of cell-cycle times, which is assumed by the Luria–Delbrück distribution of mutants, is just a special case of a gamma distribution. The second type of cell-cycle time distribution that we consider

(which we call “shifted exponential”) is based on a two-phase model, introduced by Smith and Martin (1973). According to this model, cells in G_1 phase of the cell cycle are viewed as being in a state A, from which they have a constant rate per unit time, λ , of transition to phase B. Phase B corresponds to the replication phase of the cell cycle and is assumed to take a constant time, T_B . This too, contains the exponential model as a limiting case.

Let us assume that the cell-cycle time obeys a gamma distribution of shape parameter q and scale parameter θ :

$$\psi(a) = \frac{\theta^q a^{q-1} e^{-\theta a}}{\Gamma(q)}. \quad (24)$$

The mean cell-cycle time is q/θ , the variance q/θ^2 , and the coefficient of variation (the ratio of standard deviation to mean) $1/\sqrt{q}$. The growth rate in such a culture is obtained by solving Eq. (13), with the above definition of ψ , to yield

$$\gamma = \theta(2^{1/q} - 1). \quad (25)$$

The expected age at division is given by Eq. (17). Substituting ψ from Eq. (24), we obtain

$$E(a) = \frac{q}{\theta} 2^{-1/q}. \quad (26)$$

Substituting $E(a)$ in the definition of b (Eq. (22)), we find, for a gamma distribution of cell-cycle times,

$$b = \frac{1}{q(1 - 2^{-1/q})}. \quad (27)$$

For $q=1$, the case of exponential distribution, $b=2$, while for synchronous replication, i.e., $q \rightarrow \infty$, $b=1/\log(2)$. Observing that b is monotonically decreasing with q , it becomes clear that by assuming $b=2$, as the Luria–Delbrück distribution does, we overestimate the number of mutants as a function of μ and thereby underestimate the mutation rate in the culture. For example, if we take $q=20$ (which corresponds to a coefficient of variation of the cell-cycle time of 22.4%), as in the study of Kelly and Rahn (1932), the bias is 26.6%. For $q=100$, corresponding to a coefficient of variation of 10% (Van Zoelen *et al.*, 1981), the bias is 27.6%, while the maximum bias, for $q \rightarrow \infty$, amounts to 27.9%. As an example, we estimated the mutation rate for 10^4 parallel simulated cultures of 10^4 cells using the above method and the method of Luria and Delbrück. The real mutation rate was 0.001 per daughter cell per replication. As we expect,

the estimates agree when the cell-cycle time is exponentially distributed: the mean proportion of mutants was 0.0188, and the estimated mutation rate was 0.00102. When the cell-cycle time has a gamma distribution of shape parameter $q=20$, the mean proportion of mutants in 10^4 simulated cultures was 0.0135. Using our method, which takes into account the cell-cycle time distribution, we estimate that the mutation rate is 0.001. Using the Luria–Delbrück method, which assumes an exponentially distributed cell-cycle time, we obtain a smaller mutation rate, 0.000735.

A similar derivation gives the correction factor for the mean for the two-phase model of cell-cycle time. In this model, the cell cycle time has a constant component, T_B , and a component which obeys an exponential distribution of parameter λ . Thus, the distribution of the cell-cycle time is

$$\phi(a) = \begin{cases} 0 & \text{if } a < T_B \\ \lambda e^{-\lambda(a-T_B)} & \text{otherwise.} \end{cases} \quad (28)$$

The mean of the cell-cycle time is $T_B + 1/\lambda$, the variance $1/\lambda^2$, and the coefficient of variation

$$r = \frac{1/\lambda}{T_B + 1/\lambda} = \frac{1}{\lambda T_B + 1}.$$

Again, assuming that the culture is in stationary growth, with growth rate γ , the number of cells in the culture as a function of time is given by Eq. (10). By filling in the value of $E(e^{-\gamma a})$ for this cell-cycle time distribution, and requiring that γ satisfies Eq. (14), we obtain the solution for γ in terms of Lambert’s W function (Weisstein, 1998)

$$\gamma = -\lambda + \frac{W(2\lambda T_B e^{\lambda T_B})}{T_B}. \quad (29)$$

To calculate b , we need to determine the expectation of the age at division, which satisfies Eq. (17). It follows that

$$\begin{aligned} E(ae^{-\gamma a}) &= \int_{T_B}^{\infty} a \lambda e^{-\lambda(a-T_B)} e^{-\gamma a} da \\ &= \lambda e^{\lambda T_B} \left[\frac{e^{-(\lambda+\gamma)T_B}}{\lambda+\gamma} \left(T_B + \frac{1}{\lambda+\gamma} \right) \right]. \end{aligned}$$

From Eq. (14), we find that

$$E(e^{-\gamma a}) = \frac{\lambda e^{-\gamma T_B}}{\lambda + \gamma} = \frac{1}{2};$$

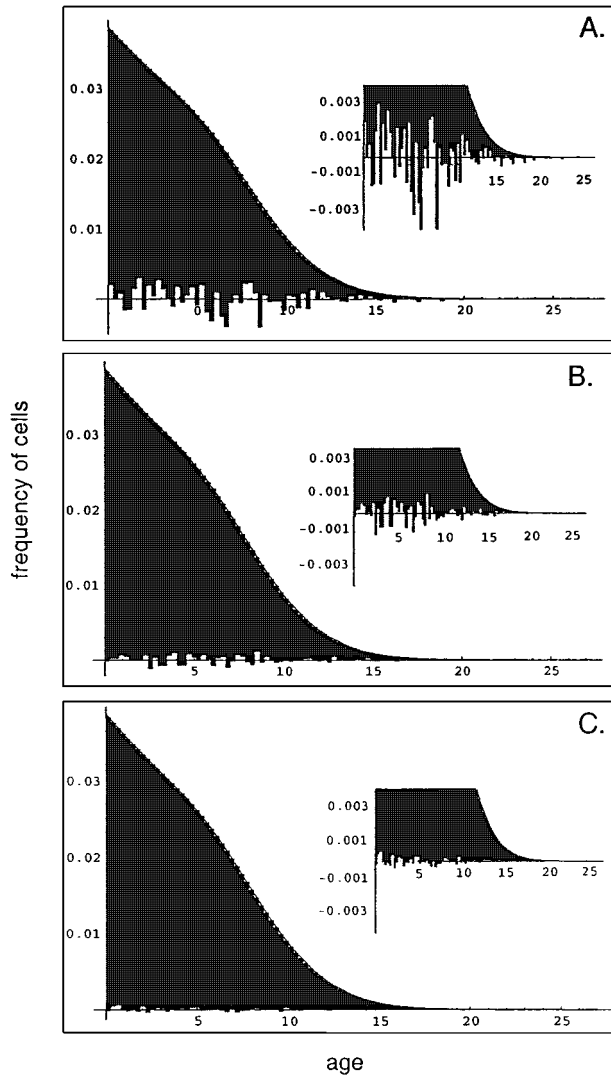


FIG. 2. Hanging histogram of the age distribution in a culture that grew from 1 to 10^6 cells. The age distribution was determined when the culture size was (A) 10^4 , (B) 10^5 , and (C) 10^6 cells, respectively. The insets show a magnification of the region around $x=0$, which corresponds to a perfect fit of the experimental and the predicted distribution.

thus

$$E(ae^{-\gamma a}) = \frac{1}{2} \left[T_B + \frac{\lambda}{\lambda + \gamma} \right].$$

The correction factor, b , is then given by

$$b = \frac{\lambda + \gamma}{\gamma [T_B(\lambda + \gamma) + 1]}. \quad (30)$$

For $T_B=0$, the case of exponential cell-cycle time distribution, we retrieve $b=2$, as we expect. Also, b decreases strictly with T_B , so neglecting the constant replication time results in underestimating μ .

Note that the above analytical expressions for the average proportion of mutants are derived assuming that the culture has a stationary age distribution. If we are to use these expressions to obtain point estimates of the mutation rate from culture data, or to estimate parameters for the mutant distributions, one might wonder whether the stationary age distribution is attained for the cell cultures sizes that we worked with. Figure 2 shows how the stationary age distribution is approached as the culture grows, for a realistic shape parameter of the gamma distribution, $q=10$ (corresponding to a coefficient of variation of 31.6%). We plotted the predicted age distribution together with the empirical age distribution obtained when the culture contained 10^4 , 10^5 , and 10^6 cells. Once the cell culture reaches 10^5 cells, the age distribution is quite close to the stationary age distribution.

3. APPROXIMATING A GENERALIZED CONTINUUM LURIA-DELBRÜCK DISTRIBUTION

We found that, due to the non-Markovian character of the cell cycle, the estimator based on the Luria-Delbrück distribution is biased. Here we attempt to improve the estimation procedure by taking into account the cell-cycle time distribution. In our previous paper, we introduced a readily computable, continuum approximation of the Luria-Delbrück distribution, which we called cLD. The integral representation for this distribution has the following form (Kepler and Oprea, 2001):

$$f_{\zeta}^0(\zeta | \beta) = \frac{1}{\pi} \int_{\varepsilon}^{\beta} dw \exp \left\{ -w[\log(w - \varepsilon) - \log(\beta - w)] \right. \\ \left. + \beta(\zeta + \log \beta) \log \left(1 - \frac{w}{\beta} \right) \right\} \sin(\pi w). \quad (31)$$

ζ is a scaled variable related to the proportion of mutants in the culture:

$$\zeta = \frac{X}{2\mu} - \frac{1}{\beta} - \log(\beta), \quad (32)$$

$\beta = 2\mu n$, and $\varepsilon = 2\mu n_0$. Empirically we found that the distributions of the proportion of mutants that we obtain with nonexponential cell-cycle time distributions are very similar to the cLD, although scaled and shifted. We therefore attempt to generalize cLD for nonexponential cell-cycle time distributions as follows. For the random variable X giving the proportion of mutant cells in a

culture defined by the parameters n , n_0 , and μ , as well as the cell-cycle time distribution ψ , we define a new set of effective parameters depending on ψ and assume that X is distributed approximately like a cLD random variable with these effective parameters. We assume that the parameters (n_0 , n , and μ) are simply rescaled by a constant and that the expected value of X is given by Eq. (23). This gives the unique rescaling

$$n \rightarrow cn \quad (33)$$

$$n_0 \rightarrow cn_0 \quad (34)$$

$$\mu \rightarrow \frac{b}{2}\mu, \quad (35)$$

where b is given in Eq. (22) and c is a free parameter that we will fit empirically using simulation data. This family of distributions will be called the generalized continuum Luria–Delbrück distribution (gcLD). We found empirically that the parameter c is most sensitive to the coefficient of variation of the cell-cycle time and does not depend strongly on n and μ .

4. COMPUTATIONAL MODEL OF A GROWING CULTURE OF CELLS

To determine the empirical distribution of the number of mutants for various cell-cycle time distributions, we need to represent individually each cell in the culture, to follow its progression through the cell cycle, and the outcome of its exposure to mutation. We implemented the cell culture as a *priority queue of cell-objects* (Sedgewick, 1988), each object encoding the pertinent information for a single cell in a culture. In our simplified view of a cell, this means its genotype (wild-type or mutant), and the time at which it will replicate. The time of replication is calculated as the sum between the time at which the cell was born and the length of its cycle. This in turn is determined by drawing a random deviate from the assumed distribution of cell-cycle times. Random deviates from the gamma (and, as a special case, the exponential) distribution are generated using the algorithm described in *Numerical Recipes* (Press *et al.*, 1988). Random deviates from the uniform distribution over the interval $[0, 1]$, which are used in the *Numerical Recipes* functions, are generated using an algorithm adapted from Knuth (1973).

The cell culture simulation proceeds as follows. We seed the system with one wild-type cell of age 0. We then perform a series of $n - 1$ replication events, to reach a culture size of n cells. A replication event always involves

the cell-object with the lowest value of the replication time, which is found at the root of the priority queue. Therefore, the retrieval operation is very short. We remove this cell-object, i.e., the parent cell, from the queue, and create two new cell-objects, i.e., the daughter cells. For each daughter cell we determine a cell-cycle length by drawing a random deviate from the cell-cycle time distribution. We add the cell-cycle length to the current time to determine the replication time of each new cell. With probability μ , each of the daughter cells of a wild-type parent may change its phenotype to become a mutant. We determine the change in phenotype by drawing a random deviate from the uniform distribution over the interval $[0, 1]$. If the value is smaller than μ , we mark the cell as mutant. We then insert the two new objects in the priority queue and rearrange the queue, such that the cell-object with the lowest time of replication is found at the root. The computation time required for these operations is logarithmic in the number of objects in the queue, i.e., the culture size. This allows us to simulate cultures of as many as 10^6 cells on an Ultra2 Sun computer in a reasonable time. At the end of the simulation, we count the number of mutants among the n cell-objects. A large number (10^4) of replicates of this experiment was used for generating the distribution of the proportion of mutants for each set of parameter values.

For the cell-cycle time distribution, we considered both the family of gamma distributions and the shifted exponentials. The process is invariant under changes in time scale; thus the choice of the scale parameter of the gamma distribution and the mean cell-cycle time for the shifted exponential are arbitrary.

5. FITTING THE FREE PARAMETER OF GCLD

For a given set of cultures, specified by the culture size, mutation rate, and cell-cycle time parameters, we determined the value of the parameter c as follows:

1. We generated the empirical distribution from simulation data.
2. We fit the parametrized cumulative density function to the cumulative distribution obtained from the simulation data. The penalty function that we use is

$$\begin{aligned} \Delta(F_{pred}, F_{obs}) = & \int dx \frac{(F_{pred}(x) - F_{obs}(x))^2}{F_{pred}(x)(1 - F_{pred}(x))} \\ & \times \Theta(F_{pred}(x) - \alpha_1) \Theta(\alpha_2 - F_{pred}(x)), \end{aligned}$$

TABLE 1

Fit of the c Parameter for Cell-Cycle Times Distributed as Shifted Exponentials

N	μ	CV	c	$\Delta(F_{pred}, F_{obs})$
10^4	10^{-3}	1	1.98	2.63×10^{-4}
10^5	10^{-4}	1	2.00	4.27×10^{-4}
10^5	10^{-3}	1	1.97	9.11×10^{-4}
10^4	3×10^{-4}	0.5	2.70	7.51×10^{-3}
10^4	10^{-3}	0.5	2.77	3.97×10^{-3}
10^4	3×10^{-3}	0.5	2.75	9.05×10^{-4}
10^5	10^{-4}	0.5	2.82	1.36×10^{-3}
10^5	3×10^{-4}	0.5	2.82	1.12×10^{-3}
10^5	10^{-3}	0.5	2.77	$3.35 \times 10^{-3\dagger}$
10^4	3×10^{-4}	0.25	2.89	7.02×10^{-3}
10^4	10^{-3}	0.25	2.98	6.17×10^{-3}
10^4	3×10^{-3}	0.25	2.96	1.59×10^{-3}
10^5	10^{-4}	0.25	3.04	2.72×10^{-3}
10^5	3×10^{-4}	0.25	3.08	5.14×10^{-4}
10^5	10^{-3}	0.25	3.02	$2.02 \times 10^{-3\dagger}$
10^4	3×10^{-4}	0.1	2.95	7.43×10^{-3}
10^4	10^{-3}	0.1	3.06	5.65×10^{-3}
10^4	3×10^{-3}	0.1	2.99	9.91×10^{-3}
10^5	10^{-4}	0.1	3.02	3.79×10^{-3}
10^5	3×10^{-4}	0.1	3.16	5.76×10^{-3}
10^5	10^{-3}	0.1	3.08	$5.87 \times 10^{-3\dagger}$

Note. Right tails truncated at 0.99, unless otherwise specified (0.98 marked by †, 0.97 by ‡).

where F_{pred} is the cumulative density function of gcLD, F_{obs} is the cumulative distribution computed from the simulation data, and

$$\Theta(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

We chose this penalty function, which puts extra weight on the tails of the distribution to provide good fits for confidence interval estimation. We truncated F at $\alpha_1 = 0.01$ and $\alpha_2 = 0.99$ (or, in some specific cases, noted below, to 0.97 or 0.98). The probability mass in these regions is negligible, while the computation of the integral becomes difficult. For minimization, we used a golden search implementation in the IMSL (copyright Visual Numerics, Inc.) subroutine UVMGS.

We found that the penalty function is a concave function of the parameter c of the distribution. More importantly, for a given set of cell-cycle time parameters, the penalty function is relatively insensitive to the value of c . Furthermore, the optimum value of c is relatively insensitive to the culture size and mutation rate and

TABLE 2

Fit of the c Parameter for Gamma-Distributed Cell-Cycle Times

N	μ	CV	c	$\Delta(F_{pred}, F_{obs})$
10^4	3×10^{-4}	1	2.00	2.71×10^{-4}
10^4	10^{-3}	1	2.00	9.95×10^{-4}
10^4	3×10^{-3}	1	1.99	1.72×10^{-2}
10^5	10^{-4}	1	2.00	4.27×10^{-4}
10^5	3×10^{-4}	1	2.00	2.12×10^{-4}
10^5	10^{-3}	1	1.97	$1.02 \times 10^{-3\dagger}$
10^4	3×10^{-4}	0.32	2.82	6.21×10^{-3}
10^4	10^{-3}	0.32	2.93	1.93×10^{-4}
10^4	3×10^{-3}	0.32	2.87	2.3×10^{-3}
10^5	10^{-4}	0.32	2.95	9.61×10^{-4}
10^5	3×10^{-4}	0.32	2.95	1.47×10^{-3}
10^5	10^{-3}	0.32	2.97	$6.17 \times 10^{-4\dagger}$
10^6	10^{-5}	0.32	3.09	2.06×10^{-2}
10^4	3×10^{-4}	0.1	2.92	9.38×10^{-3}
10^4	10^{-3}	0.1	3.08	$3.95 \times 10^{-2\dagger}$
10^4	3×10^{-3}	0.1	3.06	9.02×10^{-3}
10^5	10^{-4}	0.1	3.05	2.09×10^{-3}
10^5	3×10^{-4}	0.1	3.14	3.23×10^{-3}
10^5	10^{-3}	0.1	3.15	4.54×10^{-3}
10^6	10^{-5}	0.1	3.01	6.20×10^{-3}

Note. Right tails truncated at 0.99, unless otherwise specified (0.98 marked by †, 0.97 by ‡). Initial culture size was $n_0 = 1$ cell.

depends largely on the coefficient of variation of the cell-cycle time (see Tables 1 and 2). This allows us to design a general procedure for constructing confidence intervals for the mutation rate, knowing the parameters of the cell-cycle time distribution. To obtain a unique parameter c for all cultures with the same cycle time distribution, we fitted simultaneously all the corresponding data sets to a unique gcLD. Once we have the fitted parameter c we have the theoretical distribution of the proportion of mutants. For the purpose of illustration, in Fig. 3 we show the fit of our predicted distribution contrasted to the fit of the Luria–Delbrück distribution to simulation data. All data sets are generated using a shifted exponential form of the cell-cycle time distribution, the coefficient of variation (CV) being 25%. The upper panels corresponds to culture size $n = 10^4$, while the lower panels correspond to culture size $n = 10^5$. The mutation rates are (per daughter cell per division) 0.0003 (A), 0.001 (B), 0.003 (C), 0.0001 (D), 0.0003 (E), and 0.001 (F). The horizontal line drawn at 0 corresponds to the perfect fit. The curves in thin line give the fit of the Luria–Delbrück distribution, while the curves is heavy lines give the fit of our predicted distribution. Clearly, our predictions are considerably better than LD.

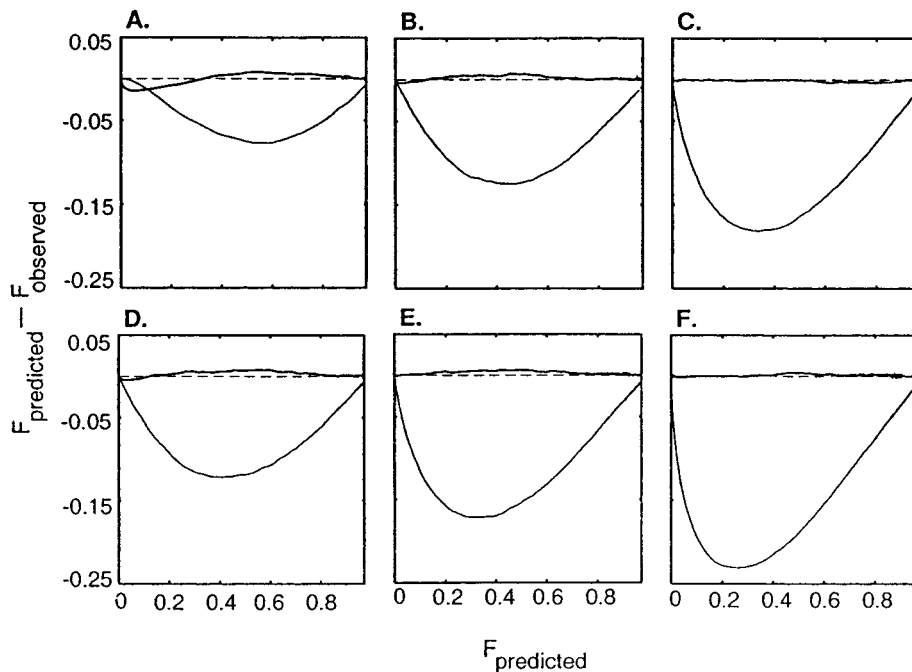


FIG. 3. Assessment of the goodness of fit of gcLD compared to LD for cultures in which the cell-cycle time, distributed as a shifted exponential, has a coefficient of variation of 25%. Top: culture size 10^4 ; bottom: culture size 10^5 . Mutation rates (per daughter cell per division) 0.003 (A), 0.001 (B), 0.003 (C), 0.0001 (D), 0.0003 (E), and 0.001 (F).

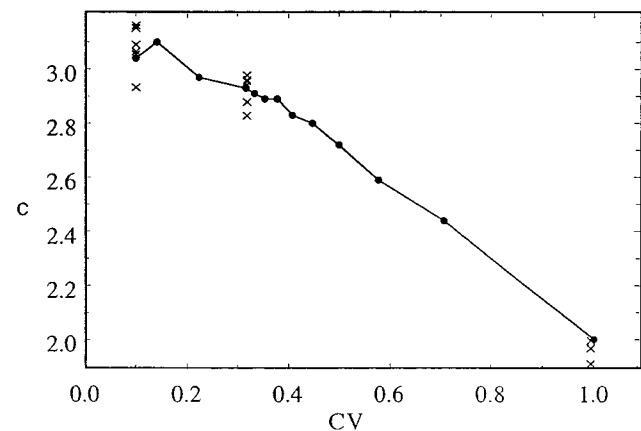


FIG. 4. Estimated parameter c as a function of the coefficient of variation of the gamma distribution of cell-cycle times. Filled circles represent the estimated value of c for simulations in which the initial population size, $n_0=1$, final population size, $n=10^4$, and mutation rate, $\mu=10^{-3}$, were kept constant, while the order parameter of the gamma distribution took values 1–10, 20, 50, and 100. Cross symbols represent estimated values of c for the parameter settings given in Table 2.

A similar fitting procedure, applied individually to the results of the simulations in which the cell cycle is distributed either as a shifted exponential or as a gamma random variable, leads to the c values given in Tables 1 and 2. We also explored the relationship between the fitted parameter c and the coefficient of variation of the cell-cycle time, the results (for gamma-distributed cell-cycle times) are presented in Fig. 4.

Actual experiments are most frequently done with large culture sizes and smaller mutation rates. Due to limitations in computational resources, such culture sizes are difficult to simulate. In fact, even culture sizes of $n=10^6$ tax our available resources so significantly that we have done fewer parameter combinations at that culture size. Our hypothesis, however, born out empirically, is that the free parameter c does not depend sensitively on culture size or mutation rate. We estimated a standard deviation for the estimator for c using individual bootstrap resampling on a pair of simulated datasets ($n=10^6$, $\mu=10^{-5}$, $CV=0.32$; $n=10^5$, $\mu=10^{-4}$, $CV=0.32$). These standard deviations (0.075 and 0.096, respectively) are roughly the same size as the dispersion of estimates from simulated datasets using different values of n and μ (Fig. 4).

6. PROCEDURE FOR ESTIMATING THE MUTATION RATE

A point estimate of the mutation rate can be obtained from the analytical value of the mean proportion of mutants in the culture. We can also construct confidence intervals for the mutation rate, using the procedure outlined in Kepler and Oprea (2001). The modification that we need to make is to use the effective parameters b and c . As an example, let us assume that the data consists of cultures which were grown from 1 to 10^4 cells, that the coefficient of variation of the cell-cycle time is 10%, and that we want to estimate the 90% confidence interval for the mutation rate given that the measured proportion of mutants was 0.01. The limits of the confidence interval correspond to the mutation rates for which the observed proportion is the 95th and 5th quantile of the proportion of mutants in the culture. We denote this interval by $[\mu_{95\%}, \mu_{5\%}]$. Table 3 gives the confidence interval for three possible scenarios: the measurement was done on a single culture; the measured proportion is the mean proportion of mutants in 10 parallel cultures of 10^4 cells; the measured proportion is the mean over 100 parallel cultures of 10^4 cells. We checked the accuracy of our estimation procedure as follows. Once we obtained the limits of the confidence interval, we simulated another set of cultures with these mutation rates. A total of 10^4 replicates of a culture of 10^4 cells were used for the empirical quantile in the single culture case, and 10^3 replicates for the 10- and 100- parallel culture situations. Note that in the last two cases, a replicate is defined as the mean proportion of mutants in 10 and 100 parallel cultures of 10^4 cells, respectively. From these cultures, we determined the quantiles corresponding to a proportion

TABLE 3

Illustration of the Estimation of the 90% Confidence Interval for the Mutation Rate

Number of parallel cultures	$\mu_{5\%}(q_e)$	$\mu_{95\%}(q_e)$	Point estimate
1	0.00245(0.045)	0.00029(0.948)	0.00075
10	0.001498(0.055)	0.000289(0.946)	0.00075
100	0.001073(0.04)	0.00042(0.947)	0.00075

Note. We assumed that the measured proportion of mutant cells is 0.01, in each of 1, 10, or 100 parallel cultures of 10^4 cells. We then searched for the mutation rates which give 0.01 as the 0.95 ($\mu_{5\%}$) and 0.05 ($\mu_{95\%}$) quantiles, to check the accuracy of these estimates, we simulated cultures with the estimated mutation rates. The quantile that a proportion of 0.01 mutants represents in these cultures is reported. They should be 0.05 and 0.95.

of mutants of 0.01. If our confidence interval is correctly estimated, these empirical quantiles (q_e) should be 0.05 and 0.95. Table 3 summarizes these values together with the point estimate that we get from the expected proportion of mutants. The empirical quantiles are very close to the desired values; thus our estimation procedure behaves adequately.

7. CONCLUSION

In the above study, we introduced a two-parameter generalization of the continuum approximation of the Luria–Delbrück distribution to correct the bias due to the non-Markovian character of the cell cycle. One of the parameters is determined analytically, from the cell-cycle time distribution. The second parameter is determined empirically, using simulation data. We find that its value is largely determined by the coefficient of variation of the cell-cycle time distribution, and it is relatively insensitive to the mutation rate and culture size. Using the empirically determined parameter, we design a procedure for constructing confidence intervals for the mutation rate given the parameters of the cell-cycle time distribution, initial and final number of cells in the culture, and the mean proportion of mutants in a number of parallel cultures. Tested against simulation data, we find that the confidence intervals that we constructed are accurate.

ACKNOWLEDGMENTS

Mihaela Oprea was supported by Office of Naval Research Grant N00014-95-1-0364 to Stephanie Forrest and Thomas B. Kepler by NSF Grant MCB 9357637. We gratefully acknowledge the support of the Santa Fe Institute, where we carried out the computer simulations.

REFERENCES

- Bartlett, M. S. 1978. "An Introduction to Stochastic Processes," 3rd ed., Cambridge Univ. Press, Cambridge.
- Jones, M. E., Thomas, S. M., and Rogers, A. 1994. Luria–Delbrück fluctuation experiments: Design and analysis, *Genetics* **136**, 1209–1216.
- Kelly, C. D., and Rahn, O. 1932. The growth of individual bacterial cells, *J. Bacteriol.* **23**, 147–153.
- Kendal, S., and Frost, P. 1988. Pitfalls and practice of Luria–Delbrück fluctuation analysis: A review, *Cancer Res.* **48**, 1060–1065.
- Kendall, D. G. 1948. On the role of variable generation time in the development of a stochastic birth process, *Biometrika* **35**, 316–330.

- Kendall, D. G. 1952. Les processus stochastiques de croissance in biologie, *Ann. Inst. H. Poincaré* **13**, 43–108.
- Kepler, T. B., and Oprea, M. 2001. Improved inference of mutation rates: I. An integral representation for the Luria–Delbrück distribution, *Theor. Popul. Biol.*
- Knuth, D. E. 1973. “The Art of Computer Programming,” Vol. 2, Addison–Wesley, Reading, MA.
- Lea, D. E., and Coulson, C. A. 1949. The distribution of the number of mutants in bacterial populations, *J. Genet.* **49**, 264–284.
- Luria, S. E., and Delbrück, M. 1943. Mutations of bacteria from virus sensitivity to virus resistance, *Genetics* **28**, 491–511.
- Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. 1988. “Numerical Recipes in C,” Cambridge Univ. Press, Cambridge.
- Sarkar, S., Ma, W. T., and Sandri, G. v. H. 1992. On fluctuation analysis: A new, simple and efficient method for computing the expected number of mutants, *Genetica* **85**, 173–179.
- Sedgewick, R. 1988. “Algorithms,” Addison–Wesley, Reading, MA.
- Smith, J. A., and Martin, L. 1973. Do cells cycle? *Proc. Natl. Acad. Sci. USA* **70**, 1263–1267.
- Stewart, F. M., Gordon, D. M., and Levin, B. R. 1990. Fluctuation analysis: The probability distribution of the number of mutants under different conditions, *Genetics* **124**, 175–185.
- Van Zoelen, E. J. J., Van Der Saag, P. T., and De Laat, S. W. 1981. Family tree analysis of a transformed cell line and the transition probability model for the cell cycle, *Exp. Cell Res.* **131**, 395–406.
- Weisstein, E. W. 1998. “The CRC Concise Encyclopedia of Mathematics,” CRC Press, Boca Raton, FL.