DOI: 10.1534/genetics.104.030338

The Amplification Model for Adaptive Mutation: Simulations and Analysis

Mats E. Pettersson,* Dan I. Andersson,† John R. Roth‡ and Otto G. Berg*,[1]

**Department of Molecular Evolution, Uppsala University EBC, SE-75236 Uppsala, Sweden, †Microbiology and Tumour Biology Center, Karolinska Institute and Department of Bacteriology, Swedish Institute for Infectious Disease Control, SE-171 82, Solna, Sweden and ‡Microbiology Section/DBS, University of California, Davis, California 95616*

Manuscript received April 22, 2004
Accepted for publication October 25, 2004

ABSTRACT

It has been proposed that the *lac* revertants arising under selective conditions in the Cairns experiment do not arise by stress-induced mutagenesis of stationary phase cells as has been previously assumed. Instead, these revertants may arise within growing clones initiated by cells with a preexisting duplication of the weakly functional *lac* allele used in this experiment. It is proposed that spontaneous stepwise increases in *lac* copy number (amplification) allow a progressive improvement in growth. Reversion is made more likely primarily by the resultant increase in the number of mutational targets—more cells with more *lac* copies. The gene amplification model requires no stress-induced variation in the rate or target specificity of mutation and thus does not violate neo-Darwinian theory. However, it does require that a multistep process of amplification, reversion, and amplification segregation be completed within ~20 generations of growth. This work examines the proposed amplification model from a theoretical point of view, formalizing it into a mathematical framework and using this to determine what would be required for the process to occur within the specified period. The analysis assumes no stress-induced change in mutation rate and describes only the growth improvement occurring during the process of amplification and subsequent elimination of excess mutant *lac* copies. The dynamics of the system are described using Monte Carlo simulations and numerical integration of the deterministic equations governing the system. The results imply that the amplification model can account for the behavior of the system using biologically reasonable parameter values and thus can, in principle, explain Cairnsian adaptive mutation.

In the neo-Darwinistic paradigm (Mayr 1982), mutations are thought to occur at random without regard to selective consequence and their formation is not thought to be purposefully regulated. Thus great interest was generated by systems whose behavior suggested that selective stress might either direct mutation to sites that improve growth (Cairns *et al.* 1988; Cairns and Foster 1991) or increase the undirected general mutation rate (Torkelson *et al.* 1997). The system generally used (Cairns and Foster 1991) employs *Escherichia coli* cells with the *lac* operon deleted from the chromosome or *Salmonella enterica* cells, whose genome does not include a *lac* region (Galitski and Roth 1995; Andersson *et al.* 1998; Hendrickson *et al.* 2002; Slechta *et al.* 2003). These cells carry, on plasmid F'_{128}, a copy of the *lac* operon with a +1 frameshift mutation that leaves cells phenotypically Lac^-. When spread on plates with only lactose as a carbon source, these cells yield Lac^+ revertant colonies at a rate several orders of magnitude faster than expected. The lawn of parental cells (between visible revertant colonies) is neither dying nor growing sufficiently to account for the observed revertants.

Three main models have been suggested to explain these results. The first two, directed mutation and hypermutable states (described below), assume that revertants arise as single-step mutations in this nongrowing lawn of parental cells. The directed mutation model proposes (Cairns *et al.* 1988; Cairns and Foster 1991; Foster and Cairns 1992) that mutations are somehow directed to sites that improve growth or to the F′-plasmid that carries the *lac* mutation. This model became less attractive after it was found that *lac* revertants (but not the parent lawn) had experienced general (undirected) mutagenesis under selective conditions (Torkelson *et al.* 1997). Another model, the hypermutable state model (Hall 1990), suggests that a small fraction of cells under stress enter a hypermutable state that is characterized by a high random mutation rate. This model was supported superficially by the observation that at least some *lac* revertants had an average 20- to 50-fold increase in the probability of carrying associated unselected mutations (Torkelson *et al.* 1997; Rosche and Foster 1999; Slechta *et al.* 2002b). The nonrevertant parent population showed very little increase in the frequency of associated mutations (Bull *et al.* 2001). Recent calculations (Roth *et al.* 2003), however, show that the hypermutable state

[1]*Corresponding author:* Department of Molecular Evolution, Uppsala University EBC, Norbyvägen 18C, SE-75236 Uppsala, Sweden. E-mail: otto.berg@ebc.uu.se

model cannot explain the number of revertants arising in Cairns' experiment without an increase in mutation rate ($\sim 10^5$-fold), much greater than that observed (20- to 50-fold). The hypermutable state model became even more problematic when it was found that mutagenesis was uneven and might affect as few as 10% of the revertants; that is, 90% seem to arise with little or no general mutagenesis.

A third model (amplification mutagenesis) assumes that reversion does not occur in the non-growing parent population, but rather within growing clones, initiated by cells with a duplication of the mutant *lac* region, which retains some weak functionality (Andersson *et al.* 1998; Hendrickson *et al.* 2002; Slechta *et al.* 2003). In this model, growth improves as unequal recombination adds more *lac* copies and, concomitantly, the probability of a reversion (−1 frameshift) event increases due to an increase in *lac* copy number within developing clones. Importantly, this model requires no change in either specificity or rate of mutagenesis but relies on standard genetic events. The phenomenon requires that the particular *lac* mutation is slightly leaky (Andersson *et al.* 1998) and that lactose be present in the selection medium (Foster and Cairns 1992). Evidence that amplification and growth precede reversion in the Cairns experiment has been presented elsewhere (Hendrickson *et al.* 2002; Slechta *et al.* 2002a,b).

It should be noted that the amplification model described here shares features with two previous theoretical explanations for directed mutation (Lenski *et al.* 1989): (1) Counterselection of revertants prior to plating is included in the sense that duplications are unstable and therefore are present in low frequency in the absence of selection and (2) existence of intermediate states prior to actual reversion is included in the form of cells carrying an amplified *lac* region. The model described here adds to those suggestions a succession of increasingly fit intermediates, which appear along the way to full reversion.

Genome-wide general mutagenesis, which occurs in the course of reversion, was interpreted as support for the hypermutable state model. However, this associated mutagenesis makes a small (4-fold) contribution to *lac* revertant number—a minor factor compared to growth and amplification (a 10^4-fold effect). This mutagenesis occurs only in clones (10 or 20% of the total) whose *lac* amplification includes the *dinB* gene, which encodes an error-prone DNA polymerase (Rosche and Foster 1999; Slechta *et al.* 2003). In the modeling presented here, this mutagenesis is largely ignored, since it is neither necessary nor sufficient to explain the behavior of the system. That is, the basic amplification model is tested.

In the amplification model, cells with a duplication of the frameshift mutant allele allow very slow growth on lactose. Because growth rate will be essentially proportional to the number of *lac* operons present in the cell, selection will strongly favor cells with increasing copy number and such cells will accumulate in microcolonies and provide an increasing target for reversion (added cells, each with multiple *lac* copies), explaining the high number of revertants. The model presumes that while amplification of *lac* improves growth, other genes that may be included are deleterious when amplified. Therefore, once a lac^+ reversion event has occurred in one *lac* copy, selection will favor retention of that copy and loss of nonrevertant copies.

The two aspects of the amplification model that are hardest to visualize intuitively are (1) the origin of the *lac* duplications in the unselected preselection population and (2) how the complicated process of amplification, reversion, and segregation can be completed within as few as 20 cell generations. It has been proposed that the process is accelerated because the *lac* operon used in the Cairns system is located on an F′ plasmid whose conjugational replication origin produces DNA ends that stimulate recombination and thereby make the underlying events more frequent (Slechta *et al.* 2002a).

MODELS

We have formalized the amplification model into a mathematical framework. Our models cover the process of forming a colony from a single cell. This allows quantitative assessment of the amplification model and shows which requirements must be fulfilled under the conditions present in the Cairns system. The models used are described in detail below.

Preexisting mutations: Two parameters affect the frequency of neutral duplications in a population, the rate at which new duplications occur and the rate at which they are lost due to recombination. The formation of the first duplication is based on the exchange between short repeats and is a relatively low probability event. Further increases in copy numbers as well as loss occur much more readily by recombination between extensive perfect sequence repeats. In an unselected growing population, the frequency of neutral duplications is expected to come to quasi-equilibrium, where formation and loss due to recombination balance out. Thus, the duplications will not be subject to Luria-Delbruck fluctuation, which assumes no loss of mutations. Fixation of a duplication is all but impossible because the loss rate significantly exceeds the formation rate.

Previous estimates suggest that the steady-state duplication frequency for a typical particular chromosomal locus is $\sim 10^{-3}$ but frequencies vary widely from one site to another (Anderson and Roth 1981). The amplification model predicts that if no new duplications arise in the nongrowing parent population, the frequency in the plated population will set an upper limit on the number of revertant colonies that can appear under selection. However, since the haploid population often shows very slow growth (Cairns and Foster 1991; Fos-

ter and Cairns 1992; Foster 1994), it is possible that a few new duplications could be added after plating. Furthermore, it is currently uncertain whether all duplications are suitable for amplification; it appears likely that the process will be affected by which surrounding genes are included in the duplication.

Adaptive amplification: The formalized model includes two types of genetic events occurring in growing cells within developing colonies: increase or decrease in the number of *lac* copies by recombination and correction of the mutant *lac* allele (*i.e.*, reversion to *lac*$^+$) by a frameshift mutation. It should be stressed that the copy number actually refers to the number of *lac* copies per plasmid in each cell, and thus the fundamental event in the model is replication of one plasmid. In most cases, it makes no difference whether there is more than one plasmid in each cell, as they replicate separately. However, plasmid growth is determined by the growth rate of the cell, which is, in turn, determined by the total number of *lac* copies present in the cell. Thus, the contribution to cell growth from each *lac* copy should be multiplied by the F′ copy number. Also, the reversion probability calculated per cell, rather than per plasmid, requires the reversion rate to be scaled in the same way.

Recombination is assumed to be equally probable between all repeated units and to occur at a constant intrinsic recombination rate. The probability of a reversion event (frameshift mutation) is, similarly, proportional to the copy number of the mutant *lac* allele per plasmid; this rate constant is the probability of mutation per *lac* copy per cell division.

This is put in a framework of an exponentially growing population, starting with a single duplication-bearing cell. This starting condition follows from the way bacterial colonies form on plates. Since the initial culture is spread out and immobilized on a selective plate, every colony is derived from a single ancestor cell. Because initial formation of a duplication is very slow, and mechanistically different, the relevant cells are those carrying a duplication when put on the plate. The fitness of any given cell is a function of its number of the (partially functional) mutant *lac* copies. Also noteworthy, death is not explicitly modeled in the system. This simplification is justified by the low likelihood of cell death in an exponential population and the typically large number of cells present (Foster 1994). When its rate is low, death would have a major influence only if the original cell, or the first revertant, died before giving rise to any descendants, either of which is very unlikely.

The recombination rate per cell division, v_{rec}, is a function of the copy number, n, and the intrinsic probability of recombination between two copies, p_{rec}. The form of this function (Equation 1) is chosen so that the rate of recombination never exceeds one per cell division. Similarly, the rate of reversion per cell division, v_{rev}, is proportional to the number of *lac* copies per plasmid, n, and the rate of mutation (*lac*$^-$ to *lac*$^+$) per *lac* copy, k_{rev}:

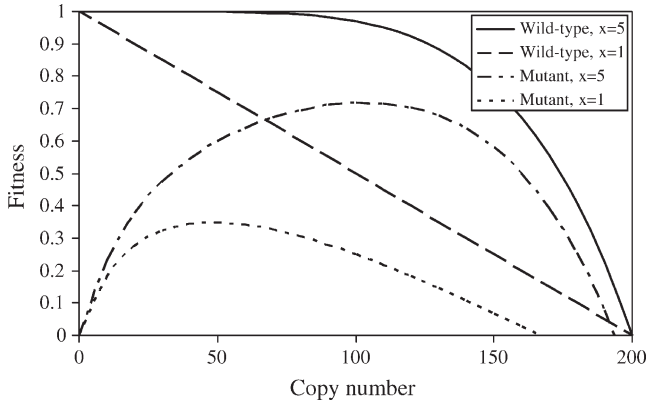

Figure 1.—The fitness of cells plotted against copy number, under different load models. The first pair of curves uses a linear load ($x = 1$) and the second pair a high-order load ($x = 5$). All curves use $y = 0.03$ and $l = 0.005$.

$$v_{rec}(n) = \frac{p_{rec} \cdot (n - 1)}{1 + p_{rec} \cdot (n - 1)} \quad (1)$$

$$v_{rev}(n) = k_{rev} \cdot n. \quad (2)$$

The fitness, w, of a mutant cell depends on its *lac* copy number, a selection parameter, y, and two load parameters, l and x, which combine to reflect the cost of carrying excess copies. The y term reflects the amount of β-galactosidase produced by a mutant copy, multiplied by the copy number of the F′-plasmid. This parameter does not affect reverted cells, which are given a fitness of 1 minus the load from excess mutant copies, should they be present. Thus, the fitness for mutant and revertant cells is, respectively,

$$w(n) = \frac{y \cdot n}{1 + y \cdot n} - (l \cdot n)^x$$

$$w_{rev}(n) = 1 - (l \cdot n)^x. \quad (3)$$

The shape of these fitness curves as a function of copy number is shown in Figure 1.

Recombination events will often, although not always, yield daughter cells with copy number different from that of the mother. In the present model, the copy number of one daughter cell remains unchanged. If recombination is treated as reciprocal, every recombination event causing a change will yield one cell with higher copy number than the original (which will improve growth in the early stages) and one with lower copy number (which will grow less well). However, simulations show that this is essentially the same as using a higher p_{rec}. Thus, for simplicity, we consider the mother cell to remain constant after recombination in all models used. Given that the original copy number is m, the resulting copy number, n, after recombination is a sample from the distribution given by

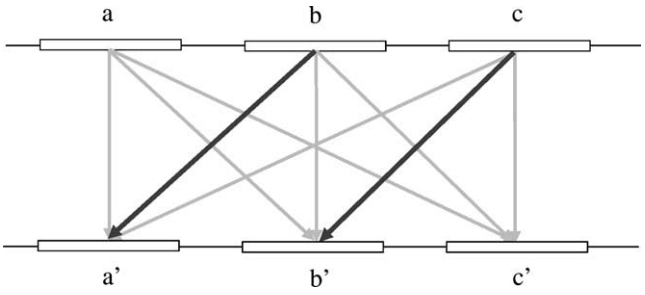

FIGURE 2.—Possible recombination events for a three-copy cell. The solid arrows show the possible events that result in a three- to four-copy transition. The shaded arrows show the remaining seven possible events. Together, they cover all nine possibilities. From the solid arrows it can be seen that the sequence of the four-copy cell will be either a-b/a′-b′-c′ or a-b-c/b′-c′.

$$p(n|m) = \frac{(2m - n)}{m^2} \quad \text{if } n \geq m \tag{4}$$

$$p(n|m) = \frac{n}{m^2} \quad \text{if } n < m. \tag{5}$$

These equations represent the number of paths of length n that can be obtained when combining two strings of length m with a single link placed at any step on either string, divided by the total number of obtainable paths. As an example, consider a transition from three to four copies (Figure 2). Obviously, two paths are possible (solid arrows), and it is readily realized that there is one path for a three- to five-copy recombination, there are two paths for a three-to-two transition, etc. It is also obvious that there are nine possible events, as there are three targets for each of the three starting positions (shaded arrows). A generalization of this example will yield the Equations 4 and 5 above. Using this, the total *rate of change* in the number, $N_n(t)$, of cells that carry n *lac* copies due to recombination events in all cells can be calculated as

$$r_{\text{rec}}(n, t) = \sum_{m=(n/2) \text{ or } m=(n/2)+(1/2)}^{m=n} N_m(t) \cdot \frac{(2m - n) \cdot v_{\text{rec}}(m)}{m^2} + \sum_{m=n+1}^{m=\infty} N_m(t) \cdot \frac{n \cdot v_{\text{rec}}(m)}{m^2}. \tag{6}$$

$r_{\text{rec}}(n, t)$ is the weighted sum of the contributions from all cells. Cells with copy number lower than or equal to n ($m \leq n$) must have more than $n/2$ copies, or they cannot contribute a daughter cell with n copies. Thus the summation starts at the first productive state. The summation to infinity may look problematic, but, because of the load constraints put upon the system (Figure 1), all terms above a limit given by the constants will be zero, eliminating the problem.

Monte Carlo simulations: Monte Carlo methods in general are based on using random numbers to explore the possible behaviors of a system. In this case the *state of the system* at any particular time can be specified by the numbers of cells, N_m, that carry m ($m = 1, 2, 3 \ldots$) *lac* copies. The simulation is a random walk through state space, where the direction of movement is governed by the relative rates away from the current state. These rates are calculated from the model formulas described above. The walk proceeds for a given number of steps, and the states visited along the way provide the output of the method. Each simulated walk gives one possible realization of the process. The expected behavior of the system can be determined by studying a number of such realizations. For each step in the walk, one new cell is added and the kind of event that occurs is determined as follows:

1. One cell from the population is chosen for replication. This is achieved by calculating the fraction of the total population fitness possessed by each cell, which is then equal to the probability that that cell will be the replicant. The probability that the replicating cell has n copies is
$$p_{\text{rep}}(n) = \frac{N_n \cdot w(n)}{\sum_{i=1}^{i=n_{\max}} N_i \cdot w(i)}, \tag{7}$$
where N_n is the current number of cells with n copies and $n_{\max}$ is the highest copy number present.
2. Then the features of the newly created daughter cell are calculated. The probability of recombination is equal to $v_{\text{rec}}(n)$ from Equation 1, n being the copy number of the mother cell. If recombination occurs, the resulting copy number is sampled from the distribution given by Equations 4 and 5.
3. Finally it is determined whether or not a reversion has occurred. The probability for this event is proportional to the number, n, of *lac* copies and given by Equation 2.

The process above is then repeated for as long as the simulation runs, adding a new cell each round.

Numerical integration: Numerical calculations can be performed in several ways. Here we use a simple Euler scheme, calculating the population state in the next time step as the previous state plus the time step times the derivative, or rate of change. That is, the number of cells in class n, *i.e.*, with n *lac* copies, at time $t + \Delta t$ will be

$$N_n(t + \Delta t) = N_n(t) + \frac{dN_n(t)}{dt} \cdot \Delta t. \tag{8}$$

The time derivative will include a term for growth and another term for recombination. The growth term is based on the fitness formula (Equation 3) used in the Monte Carlo simulations. The recombination term depends on the intrinsic recombination rate, the number of cells in other classes, and the probabilities that recombination moves a cell to class n. Like the Monte Carlo simulation, the probability that a cell with m copies yields

a daughter cell with n copies after a recombination event follows Equations 4 and 5. Finally there is a mutational flow from the mutant to the wild type, slowly seeding the revertant population.

The fully deterministic approach has the drawback of producing fractional cells, in particular fractional revertant cells, which can seriously disrupt the calculations. Because of its rapid growth, such a fractional revertant cell could quickly overtake the population even though the probability that a reversion really has occurred may be very small. This can be alleviated in a quasi-deterministic description by adding a simulation element that uses the rate of production of revertants as the probability to stochastically yield a "real" revertant as well as imposing some restrictions to allow growth only in states containing at least one cell. Thus, one can use the numerical analysis as a way to calculate the probability of reversion as a function of time. Furthermore, this can be used to obtain the probability that reversion has occurred at a given time point.

Since the probability that reversion occurs during division of a cell with a given copy number is constant (per copy) throughout the simulation, it is straightforward to calculate the total reversion probability over each time step; all that needs to be determined is the number of cells that replicated during the step, for each copy number. The reversion probability, over a given step, for the class with n copies, where ΔN_n cells replicate, is

$$q_{\text{rev}}(n) = \Delta N_n \cdot k_{\text{rev}} \cdot n; \quad \Delta N_n \cdot k_{\text{rev}} \cdot n \ll 1. \tag{9}$$

The total reversion probability during that time step is 1 minus the probability that no reversion has occurred in any of the classes n,

$$p_{\text{rev}}(T) = 1 - \prod_{n=0}^{n=\infty} (1 - q_{\text{rev}}(n)), \tag{10}$$

where T is the time at the end of the time step in question. Thus, the probability that the first reversion occurs at or before this time is

$$P_{\text{rev}}(T) = 1 - \prod_{t=0}^{t=T} (1 - p_{\text{rev}}(t)). \tag{11}$$

And, finally, the probability that the first reversion occurs during the time step ending at T is

$$Q_{\text{rev}}(T) = p_{\text{rev}}(T) \cdot \prod_{t=0}^{t=T-1} (1 - p_{\text{rev}}(t)). \tag{12}$$

The resulting distribution of reversion events can then be compared to the Monte Carlo simulations, which directly represent the possible outcomes with no approximations other than those inherent in the formulation of the underlying model.

Timescale: The calculations themselves proceed in meta-time, which is not directly related to the evolution of the system; it is simply a succession of growth events. Thus, there is a need to associate the sequence of events to a relevant time measure, to understand the behavior of the model. Two types of timescales have been used:

Generation timescale: The most natural timescale is to measure the number of cells present in the population, as the number of divisions that have occurred is central to both recombination and reversion events. However, the raw number of cells is not very intuitive and it quickly becomes unwieldy. Instead the 2-logarithm of the number of cells is used, as that is the same as the number of generations the colony would have passed through, had all cells grown equally fast. This is not the case, of course, but it is a representation of the average age of the cells.

Real timescale: While it is very convenient to allow the number of cells in the colony to constitute the timescale, it has drawbacks when the model needs to be correlated with experimental data. One can transform the generation timescale into a real timescale by introducing a wild-type growth rate, μ_0 (generations per day), and using this to scale the fitness formula, Equation 3. Thus, the growth rate of a cell with n copies can be expressed as

$$\mu(n) = \mu_0 \cdot w(n). \tag{13}$$

On the other hand, it would also be possible to do the reverse. By slightly changing the experimental setup, one could measure the number of cells in revertant colonies and use such data to correlate to the model. This would be preferable, as it eliminates some unnecessary assumptions and parameters that are difficult to estimate.

Parameter choice: It is not easy to obtain an accurate estimate of the intrinsic recombination rate, p_{rec}, but a value in the range of a few percent seems reasonable since this has been experimentally estimated for tandem duplications on the F′ plasmid (E. Kugelberg, J. Roth and D. I. Andersson, unpublished data) and slightly lower rates have been observed in the chromosome (Anderson and Roth 1977; Reams and Neidle 2003; R. Dawson and J. Roth, unpublished data). The rate of reversion (back-mutation) of the *lac* frameshift mutation used is $k_{\text{rev}} = 10^{-8}$, a rate estimated from unselected growth. Strictly, this should be scaled by the F′ copy number. However, this scaling is left out since the system is very insensitive to the value of k_{rev} and 10^{-8} is already a rough estimate. In some cases, though, the average mutation rate, and consequently the reversion rate, is further increased ~35-fold by an induction of the SOS response (Rosche and Foster 1999; Hendrickson *et al.* 2002).

The growth rate contribution from each copy is essentially equivalent to the β-galactosidase activity yielded by one copy of the frameshift mutant, which has been experimentally shown to be ~1% of the wild type (Andersson *et al.* 1998). Thus $\gamma = 0.01$, in the standard case. However, since the *lac* operon is on a plasmid, the contribu-

tion from each copy should be multiplied by the F′ plasmid copy number, which is between 2 and 3. The real-time calculations show that this is necessary to complete the process within the duration of the experiment.

The load parameters, l and x, reflect the cost of maintaining and expressing many copies of *lacZ* and other genes included in the duplication, as well as the increasing risk of deletion events where a part of the array is looped out and lost. These are more difficult to estimate from experimental data; fortunately the exact choice of load parameters is not crucial for the behavior of the system. As long as the expression follows the assumption of the model, which is that fitness increases with increasing copy number up to some sufficiently large (>50) optimum value, the system is robust to changes in these parameters. The main function of the load, in combination with the selection expression, is to determine which copy number gives maximal fitness and therefore is optimal with respect to growth. Furthermore, the load expression determines the maximum copy number that can actually grow, as well as the general shape of the fitness *vs.* copy number curve (Figure 1). The simplest possible (*i.e.*, linear with $x = 1$) load expression in Equation 3 gives a broad fitness curve, where the optimum is far below the maximum. By using a higher-order expression ($x > 1$), the effect of the load sets in much more sharply near the maximum copy number, thus allowing the positive contribution of the high copy numbers to make its presence felt without having to include a large number of nonproductive states by increasing the maximum. If the same maximum copy number is used, there will also be a distinct effect on the fitness relative to revertant cells. This is of less importance, since the main part of the dynamics takes place in a population without revertants present. It will, however, affect the rate with which the duplications are lost after a reversion has occurred. It should be noted that even a neutral duplication will tend to segregate. This is because copy numbers can change rapidly, also down to $n = 1$, due to frequent recombination. However, the process is asymmetric, as discussed above; once excess copies are lost, they will be recreated only slowly. In principle, the load affects the absolute growth rate and would be an experimentally accessible quantity.

For real-time calculations, the system is more sensitive to parameter values. In this case, the selection parameter and the fitness expression become critical to the magnitude of the timescale. However, there are also qualitative effects that are insensitive to this and are thus of genuine interest. The general effect of the real-time transformation is to extend the early generations of the colony, as mean fitness is low at that stage and, naturally, growth is slow. Late generations are correspondingly compressed, due to their high mean fitness. Thus, by switching to real time, the formulation of the fitness expression, especially the choice of fitness parameters, becomes much more important for the shape of the process.

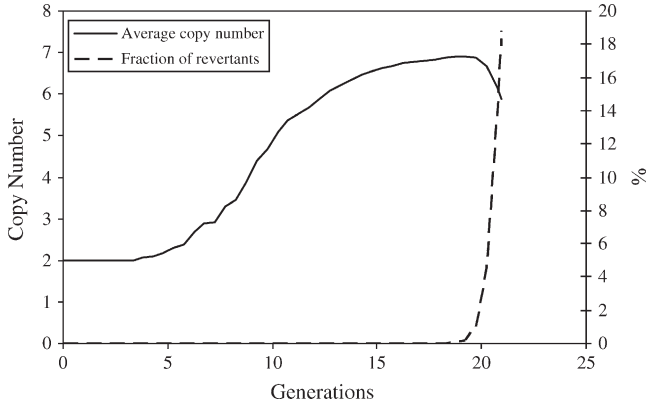

FIGURE 3.—Monte Carlo simulation of colony formation. A run with elevated rates, $p_{rec} = 0.2$, $k_{rev} = 10^{-6}$, $l = 0.005$, $x = 5$, and $y = 0.05$, is shown to illustrate the rise of revertant cells and the corresponding drop in average copy number. This will typically occur three to four apparent generations after the original revertant was born.

RESULTS

Two variants of the model have been analyzed: a complete Monte Carlo simulation and the quasi-deterministic description, which is a stochastic-deterministic hybrid. The pros and cons of each approach are discussed below.

Monte Carlo simulations: This is perhaps the best approach, as it completely obeys the proposed rules, without requiring any extra assumptions or adaptations. However, as each cell is modeled individually, it becomes unmanageable for large population sizes. Due to its closeness to the thought model, it can be, and has been, used as a control of the other variants.

The simulations show that after a period of slow growth of cells with low copy number, the average copy number takes off quickly. This is expected, as variation must be generated before selection can take effect. Once variation in copy number is present, the ability of recombination to quickly distribute the population over many copy-number states provides an excellent template for selection. In this aspect all simulation runs are more or less the same. The key event is the appearance of the first revertant. In the revertant population the duplication or amplification loses its selective advantage and is even counterselected, due to the load penalty. As expected, this subpopulation quickly loses the extra copies, and stable (single-copy) revertants appear a few generations later and start rising rapidly in the population (Figure 3).

The Monte Carlo simulations are to some extent limited by the exponential population growth, making it difficult to follow the colony to the end, in the range of 10^8–10^9 cells. Therefore, to get a reversion event within a

reasonable calculation time, the simulations displayed in Figure 3 were run with an elevated reversion rate. With a more realistic rate, the first reversion is likely to happen in the last or second to last generation simulated, too late to follow the development.

Quasi-deterministic description: This approach has the advantage of implicitly modeling the cells, by calculating the total change in each copy-number class. This means that the number of calculations required does not scale with the number of cells present in the population. On the other hand, the nature of cells as discrete entities is partially lost, which forces some artificial constraints to be put on the system. The deterministic nature of these calculations is only a minor drawback, mainly because there is no death in the system, which removes one important source of random variation. Furthermore, by including the first appearance of a revertant as a stochastic event, according to Equation 9, the main contribution to random variation is accounted for (jackpots).

The full development of the system is shown in Figure 4. Figure 4, a and b, shows how the copy numbers of the mutant *lac* allele quickly increase, driven by selection, until a stationary distribution is reached. The shape of this distribution will be governed by the fitness equation (Equation 3). Thereafter, this distribution for the nonrevertants retains essentially the same shape throughout, although the relative contribution will decrease once revertants have appeared (Figure 4, c and d). Figure 4b shows that the first revertant is likely to appear in some high-copy-number class, as expected from the model. Once their growth starts, revertant cells will swiftly lose copies, and low-copy-number classes will increase in frequency (Figure 4c) and eventually dominate the colony (Figure 4d). The rate at which revertants take over the population is dependent on the relative fitness of the revertants and the best of the nonrevertant cells; this relation, in turn, depends on the fitness equation (Equation 3).

In fact, since the timing of the first reversion event is so crucial to the history of the clone and its probability can be analytically calculated, it is possible to answer many questions without actually performing the stochastic step at all, but rather by just following the development of the reversion probability, as shown in Figure 5. This figure shows the complete development of the system, until reversion is all but inevitable. Since it is not certain that all cells initially carrying a duplication will give rise to a revertant colony within the time of the experiment, it is of interest to examine also the region where reversion is a low probability event (Figure 5, inset).

Fortunately, the added resolution of the analytic probability calculation makes it possible to examine the dependence on the recombination rate, which allows estimation of the development of the reversion probability in this key region. These distributions are shown in

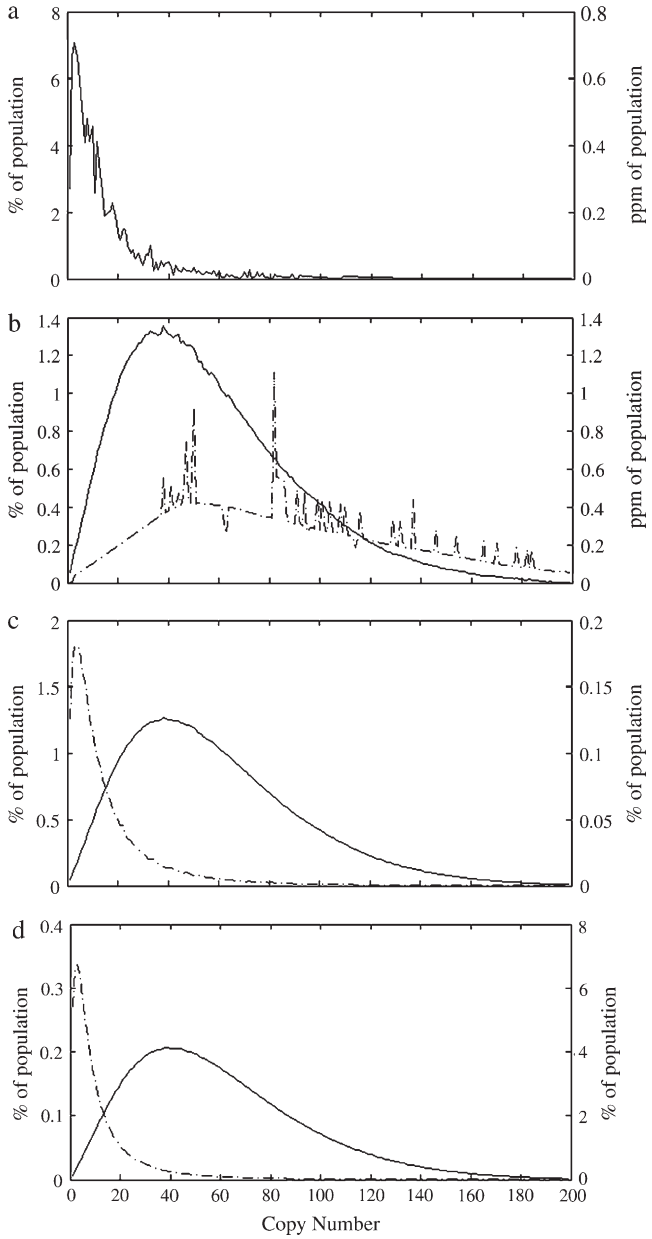

Figure 4.—The copy-number distribution at chosen time points. Mutants are plotted on the left *y*-axis, revertants on the right. Note the differences in scales as the colony develops. (a) Ten generations. The population has started to spread out across the available copy numbers, reversion has not occurred. (b) Twenty generations. The mutant population has reached the shape dictated by the fitness curve and will remain essentially unchanged throughout the experiment. The first revertants have appeared, evident as spikes in the curve. Surrounding these peaks is some numerical noise, due to the very low frequencies present at this stage, which have negligible impact on system behavior. (c) Twenty-two generations. Revertants are growing, and losing excess copies, rapidly. The peak of the distribution is at two to three copies, with also single-copy cells being abundant. (d) Twenty-six generations. The shape of the curves is largely unchanged, but the scale shifts and the single-copy cells have increased further in frequency. Revertants are now dominant. Parameters used are $p_{rec} = 0.05$; $k_{rev} = 35 \times 10^{-8}$, $r = 0.01$, $l = 0.005$, $x = 5$.

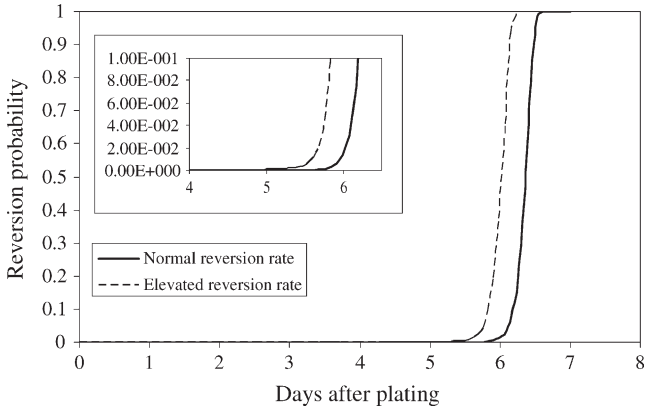

FIGURE 5.—Analytic curves for the cumulative reversion probability over time for $p_{rec} = 0.05$, $y = 0.03$, $\mu_0 = 30$ generations/day, and other parameters as in Figure 4.The solid line is the probability of reversion having occurred at or before a given time, with $k_{rev} = 10^{-8}$ (Equation 11). The dashed line follows the same equation, but with $k_{rev} = 35 \times 10^{-8}$.

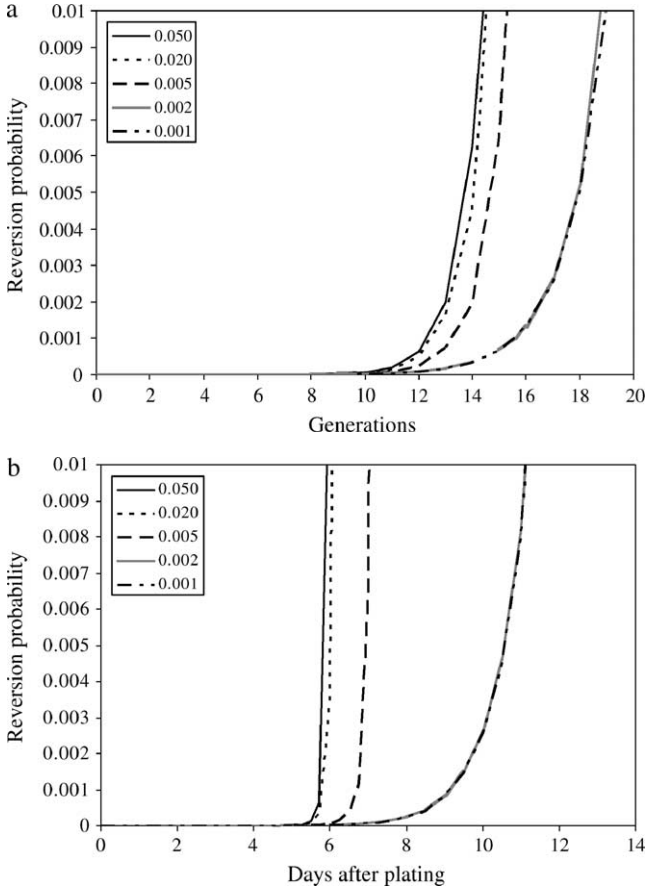

FIGURE 6.—The probability that reversion has occurred, according to Equation 11, plotted as a function of time. (a) The generation timescale, (b) the real timescale. The different lines are calculated with different recombination rates (p_{rec}), which are shown in the insets. Other parameters are the same as in Figure 5.

Figure 6. Here, it is interesting to note that there is a qualitative difference between the real-time and generation-scale behavior. When measured in generations, the recombination rate has very little impact, until it becomes so small that there is essentially no recombination occurring during the entire process. If the measure is real time, however, there is a much more noticeable effect of semi-low recombination rates. This is due to lower fitness in early generations, which accounts for a large part of the total time of colony evolution.

Figures 6 and 7 show the relation between reversion and recombination rate. There is a threshold value where recombination is too slow to have any real impact on the system. In this case, reversion occurs only at a colony size of a few times 10^7 cells, which is what one would expect if the average copy number of the colony remains at two. At the other extreme, when recombination is as frequent as one per division (the maximum), reversion occurs at a frequency that is not very different from a recombination rate of a few percent. This is not unexpected as the probability of reversion depends on the total number of *lac* copies in the colony. Thus even if the optimal copy number is reached quickly, the population must be large before reversion is likely.

Alternative model variants: The calculations above are all based on the same principal model as described. However, some aspects could be treated equally well in slightly different ways. The following is an exploration of some alternative choices and variants of the model.

Different initial conditions: While all colonies are initiated by a single cell, it is not necessary that this cell has exactly two *lac* copies. Distributions from calculations on neutral duplications (Figure 8) indicate that, while most cells with multiple gene copies have only two copies, there is a thin but long tail of cells with higher copy numbers. Starting at a higher state may of course increase the likelihood of success, perhaps mainly because cells that grow faster in the beginning are more likely to produce a revertant colony in real time. In terms of the generation in which the first revertant appears, the effect is limited. This is because the population usually reaches optimal copy numbers before reaching a size where reversion is likely even with just two initial copies. As an example, shifting the initial copy number from two to eight moves the average reversion time less than half a generation (data not shown). However, due to the increased growth rate of the initial cells, the real timescale can be substantially shortened by such a shift.

Loss of revertants: Not included in the model is the possible loss of the actual reverted copy through recombination events between flanking nonrevertant copies. This might slow down the process. Given the assumption that all recombination events are equally likely, the probability of this event is

$$p_{loss} = \frac{z}{n} \cdot \frac{(n - z)}{n}, \qquad (14)$$

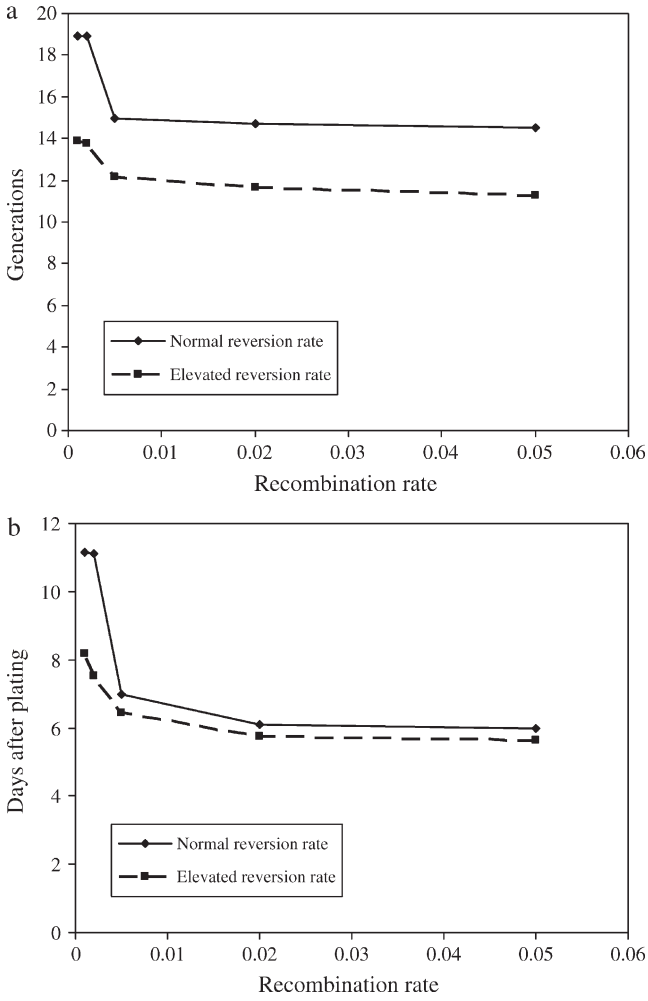

FIGURE 7.—The time where reversion has occurred with probability 0.01 plotted as a function of the recombination rate (p_{rec}). The solid line is with the normal reversion rate 10^{-8}, and the dashed line is with the elevated rate 35×10^{-8}. (a) Timescale in generations, (b) real timescale (days). Constants other than p_{rec} are the same as in Figure 5.

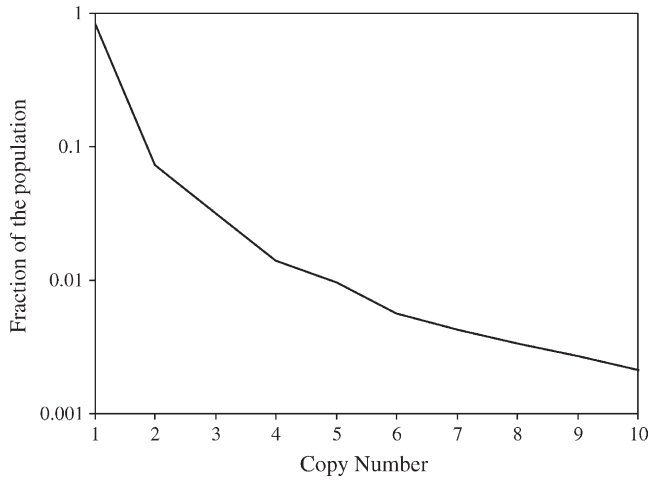

FIGURE 8.—The stationary distribution of a neutral duplication where the rate of initial duplication formation is $1/100$ of the recombination rate per copy.

where n is the number of copies, and z is the position of the reverted copy. It is obvious that this probability is maximal when the reversion is in the middle copy ($z = n/2$) and that this maximum is 25%. Thus, even in a worst-case scenario, 75% of all deletions are allowed, and the process is unlikely to be very much affected. Calculations done for this situation (25% of all deletions yield nonrevertant cells) support this expectation.

Reversion rate variations: This analysis has modeled the reversion process without using any increase in mutation rate (per gene copy). This puts the maximum burden on the amplification model to explain the observed revertants. In fact, the general mutation rate does go up somewhat in the course of this experiment, but there is a strange discrepancy between the intensity of that mutagenesis and the number of observed revertants. The associated general mutagenesis increases the mutation rate 20- to 50-fold; if this is prevented (by preventing SOS induction or eliminating the required *dinB* gene) the revertant number drops only by a factor of 4. It has been suggested that mutagenesis allows reversion to occur earlier in the history of each clone. That is, without mutagenesis, reversion still occurs but only after longer growth or higher amplification to compensate (with target number increases) for the lower mutation rate.

However, the situation is complicated by the fact that not all cells may experience the same reversion rate. Recent experiments (SLECHTA *et al.* 2003) show that detectable general mutagenesis occurs only if the amplified unit includes the *dinB* gene, which is situated near *lac* on the strains used in the experiment. About 10% of detected revertants arise in colonies whose *lac* amplification includes *dinB*, but this may not represent the fraction of initial clones with a *dinB* duplication. That is because mutagenesis allows reversion to occur earlier in clones with amplified *dinB* so that they will be overrepresented among the revertants that appear during any particular time period. The probability of reversion for microcolonies with and without an elevated mutation rate is shown in Figure 5. This figure shows that it is likely that a much larger fraction of duplications including *dinB* will result in revertant colonies, compared to when *dinB* is not included. The final result will be a weighted superposition of the two curves.

DISCUSSION

General points: The calculations and simulations described here show that the amplification model can quantitatively explain the behavior of the Cairns selection system. However, this requires that several variables be appropriately set: Recombination rates cannot be too low; if the parameter p_{rec} drops below 0.5% per cell/division, the gene amplification process will not lead to

reversion within the timescale of the experiment. Equally important is that to the genetic system can accommodate copy numbers on the order of 10^2, or the process will not be quick enough. Minimal estimates of copy number show experimentally that this can be attained in the Cairns system (Slechta *et al.* 2003). These requirements might well explain why this phenomenon is observed only on a plasmid.

Of course, the exact nature of these restrictions will depend on some particulars of the model. If, for example, the reversion rate is higher than the standard unselected rate of 10^{-8}/copy/generation, the required recombination rate and achievable copy number drop. In fact, the average rate is 30- to 50-fold higher due to induction of the SOS response (Hendrickson *et al.* 2002). However, this is achieved by a 10-fold higher increase in ~10% of the population, the ones with *dinB* included in the duplication. Also other duplications may have some effect on the mutation rate, which could explain why the appearance of revertant colonies is spread out over several days rather than in one distinct burst. Also of some importance is the fitness equation (Equation 3 and Figure 1), since it will dictate what copy number is optimal with respect to growth. This will be roughly equivalent to the average of the population, since cells with copy numbers close to the optimum grow best. However, given the costs of maintaining and translating genes, some of which may have toxic effects when overexpressed, it seems unlikely that the optimum copy number should be much larger than the number needed to achieve approximately wild-type levels of β-galactosidase activity, while the fundamental need for lactose metabolism inherent in the system will ensure that it is not much smaller. This implies that the probable range of optimal copy numbers is not very broad, which limits the impact of which expression is used to calculate the fitness of cells.

Model behavior: According to the model, development of a revertant colony can be divided into three steps. The first is a start up phase where average *lac* copy number per cell increases, slowly at first but constantly accelerating, until the average copy number approaches an optimum (limited by the loss rate and the general growth cost of the amplification). The length of this phase will be weakly dependent on the recombination rate. This is because even with recombination at every cell division, the maximum, accumulation will still be limited by the small initial step size and low number of cells early in development. In the case of low recombination rates, the exponential growth of the colony will compress events, since what really matters is the recombination rate times the number of cells. Also, the race is not to infinity, but rather to the optimal copy number. The probability of reversion will remain small until the population has increased, making it less crucial to reach the optimum quickly. Reversion is unlikely before generation 13 (~10^4 cells), even if the population starts at optimal copy numbers.

Second, as the optimal copy number is reached, a temporary stable state appears. At this point a fairly uniform population of cells is growing, with their copy numbers distributed around the optimum (*cf.* Figure 4), in some sense waiting for the appearance of the first revertant. Once a revertant is formed, the colony enters the final stage. In this stage, fitter cells will constantly be created by loss of excess gene copies, until a stable single-copy revertant appears and takes over the population.

Relation to experiments: Current estimates suggest that only a small fraction (~10^{-2}) of duplication-bearing cells succeed in generating a revertant clone on selective medium by day 6. This appears surprising, as there is an inherent inevitability in the amplification model; once the process starts, reversion is the only possible outcome. However, other experimental evidence (E. Kugelberg and D. I. Andersson, unpublished data) offers an explanation. Experiments using only cells that are known to carry a duplication show that virtually all cells give rise to colonies within 10 days, with the first ones appearing by day 4 or 5. Between days 4 and 10, the number of revertant colonies increases nearly linearly, until the number of plated cells is reached. This would indicate that the original duplication is modified, leading to a different efficiency of amplification. If it remained exactly the same, a more coherent appearance would be expected, with the majority of the colonies showing up during 1 day. Around day 5, ~1/100 cells have given rise to a colony. These are ones that either picked up a reversion very early or gained a duplication that was amplified with very high efficiency. It is only these colonies that will appear in the Cairns system, in which experiments cannot be extended to 10 days, due to the conditions used.

One more factor that could affect the number of revertants is that two copies might not be enough to sustain growth, implying that only cells with even higher copy numbers can enter the amplification process. This will have a limited effect, however, as the ratio between cells with two and, say, four copies will initially not be larger than about fourfold (Figure 8).

The main reason that prevents extension of the Cairns system experiments in time is interference from surrounding cells, both scavengers and more successful amplifying clones. It is not unreasonable to propose that this external competition will prevent growth of would-be revertant colonies that were not among the lucky few to pick up a reversion, or improved amplification efficiency, at an early stage in growth. Relating to Figure 5, this would mean that only the leftmost tail will actually contribute, and the rest of Figure 5 is relevant only as an illustration of the model or in experiments with duplication-carrying cells only.

The adaptive duplication mechanism could contribute to gene innovation: Most genomes seem to be in a

constant flux of incorporating and losing duplications. Most of these duplications are probably neutral or deleterious and are lost fairly quickly. The classical model for the evolution of new gene function relies on gene duplication followed by adaptation of one the copies. The problem with this idea is that the redundant gene copy is much more likely to be removed or inactivated rather than adapted. Duplication patterns and age distribution of paralogous genes in microbial genomes suggest that duplications that survive are under selection from the start (Hooper and Berg 2003). This could be the case for a duplicated sequence that carries some weak or ancillary function that is beneficial. A weak function would be one that is not optimized but more accidental; perhaps an enzyme that is specific for a certain substrate can also deal with another similar substrate, though not very efficiently. If the new substrate becomes abundant, further duplications of this sequence could be driven by a gene dosage effect, which in turn will produce a larger target for adaptive mutations. Thus, a gene that carries some ancillary function that can be selectively amplified stands a much larger chance of being retained and adapted (Hendrickson *et al.* 2002; Hooper and Berg 2003).

Conclusion: The combined evidence of the analyses used shows that the adaptive duplication mechanism is quantitatively sound and can explain the high reversion rate phenomenon observed without proposing the existence of any new and unproven molecular processes.

This work was supported by the National Graduate School in Scientific Computing (M.P.) and by the Swedish Research Council (O.G.B., D.I.A.).

LITERATURE CITED

Anderson, R. P., and J. R. Roth, 1977 Tandem genetic duplications in phage and bacteria. Annu. Rev. Microbiol. **31:** 473–505.

Anderson, R. P., and J. R. Roth, 1981 Spontaneous tandem genetic duplications in *Salmonella typhimurium* arise by unequal recombination between ribosomal RNA (rrn) cistrons. Proc. Natl. Acad. Sci. USA **78:** 3113–3117.

Andersson, D. I., E. S. Slechta and J. R. Roth, 1998 Evidence that gene amplification underlies adaptive mutability of the bacterial *lac* operon. Science **282:** 1133–1135.

Bull, H., M.-J. Lombardo and S. Rosenberg, 2001 Stationary-phase mutation in the bacterial chromosome: recombination protein and DNA polymerase IV dependence. Proc. Natl. Acad. Sci. USA **98:** 8334–8341.

Cairns, J., and P. L. Foster, 1991 Adaptive reversion of a frameshift mutation in *Escherichia coli*. Genetics **128:** 695–701.

Cairns, J., J. Overbaugh and S. Miller, 1988 The origin of mutants. Nature **335:** 142–145.

Foster, P. L., 1994 Population dynamics of a Lac$^-$ strain of *Escherichia coli* during selection for lactose utilization. Genetics **138:** 253–261.

Foster, P. L., and J. Cairns, 1992 Mechanisms of directed mutation. Genetics **131:** 783–789.

Galitski, T., and J. R. Roth, 1995 Evidence that F plasmid transfer replication underlies apparent adaptive mutation. Science **268:** 421–423.

Hall, B. G., 1990 Spontaneous point mutations that occur more often when advantageous than when neutral. Genetics **126:** 5–16.

Hendrickson, H., E. S. Slechta, U. Bergthorsson, D. I. Andersson and J. R. Roth, 2002 Amplification-mutagenesis: evidence that "directed" adaptive mutation and general hypermutability result from growth with a selected gene amplification. Proc. Natl. Acad. Sci. USA **99:** 2164–2169.

Hooper, S. D., and O. G. Berg, 2003 On the nature of gene innovation: duplication patterns in microbial genomes. Mol. Biol. Evol. **20:** 945–954.

Lenski, R. E., M. Slatkin and F. J. Ayala, 1989 Mutation and selection in bacterial populations: alternatives to the hypothesis of directed mutation. Proc. Natl. Acad. Sci. USA **86:** 2775–2778.

Mayr, E., 1982 *The Growth of Biological Thought: Diversity, Evolution and Inheritance.* Harvard University Press, Cambridge, MA.

Reams, A. B., and E. L. Neidle, 2003 Genome plasticity in Acinetobacter: new degradative capabilities acquired by the spontaneous amplification of large chromosomal segments. Mol. Microbiol. **47:** 1291–1304.

Rosche, W. A., and P. L. Foster, 1999 The role of transient hypermutators in adaptive mutation in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **96:** 6862–6867.

Roth, J. R., E. Kofoid, F. P. Roth, O. G. Berg, J. Seger *et al.*, 2003 Regulating general mutation rates. Examination of the hypermutable state model for Cairnsian adaptive mutation. Genetics **163:** 1483–1496.

Slechta, E. S., J. Harold, D. I. Andersson and J. R. Roth, 2002a The effect of genomic position on reversion of a *lac* frameshift mutation (*lacIZ33*) during non-lethal selection (adaptive mutation). Mol. Microbiol. **44:** 1017–1032.

Slechta, E. S., J. Liu, D. I. Andersson and J. R. Roth, 2002b Evidence that selected amplification of a bacterial *lac* frameshift allele stimulates Lac$^+$ reversion (adaptive mutation) with or without general hypermutability. Genetics **161:** 945–956.

Slechta, E. S., K. L. Bunny, E. Kugelberg, E. Kofoid, D. I. Andersson *et al.*, 2003 Adaptive mutation: general mutagenesis is not a programmed response to stress, but results from rare co-amplification of *dinB* with *lac*. Proc. Natl. Acad. Sci. USA **100:** 12847–12852.

Torkelson, J., R. S. Harris, M.-J. Lombardo, J. Nagendran, C. Thulin *et al.*, 1997 Genome-wide hypermutation in a subpopulation of stationary-phase cells underlies recombination-dependent adaptive mutation. EMBO J. **16:** 3303–3311.

Communicating editor: M. Feldman