Transposable Elements and Fitness of Bacteria

Jean-Louis Martiel
Techniques de l’Imagerie, de la Modélisation et de la Cognition, CNRS UMR5525, F-38706 La Tronche, France
E-mail: jean-louis.martiel@imag.fr

and

Michel Blot
Plasticité et Expression des Gènes Microbiens, CNRS FRE2383, Université Joseph Fourier, F-38041 Grenoble, France

Received February 14, 2002

A stochastic model was designed to describe the evolution of bacterial cultures during 10,000 generations. It is based on a decreasing law for the generation of beneficial mutations as they become fixed in the genomes. Seven beneficial mutations on average were necessary to improve the relative fitness from 1.0 to 1.43 and the model was consistent with the population biology and the genetic data of 12 experimental lines. In one bacterial line, comparison between the model and the data suggests that pivotal mutations mediated by insertion sequences account for a large part of bacterial adaptation. In a more detailed analysis of one simulation, it was shown that only 0.01% of the mutations generated by a population over 10,000 generations can go to fixation as a consequence of their improved fitness. However in the model, the probability of being better fit than its parent should be set initially at ca. 10% to promote an evolution similar to the observed data.

INTRODUCTION

Variation in bacterial genomes is the product of different mutations, including DNA rearrangements mediated by transposable elements. Insertion sequences (ISs) can insert at many sites in the genome (Campbell et al., 1977) and their multiple copies can also be involved in site-specific recombination, leading to a variety of genomic rearrangements (Chalmers and Blot, 1999). The IS elements give a major contribution to the overall mutagenesis (Rodriguez et al., 1992) and they are assumed a role in genomic adaptation and the completion of horizontal transfer. The first discovered and most obvious consequence of transposition is deleterious; for instance, when the insertions are within essential genes. This raises the question of their maintenance in bacterial genomes through evolution (see Blot (1994) for discussion). If IS elements were mostly detrimental (in the laboratory, the bacterial geneticist experiences easily unexpected gene disruptions!), they would be eliminated from bacterial genomes in the long term, except under very frequent horizontal transfers promoted by other accessory elements (plasmids and viruses). Although horizontal transfer is essential for importing IS elements (and vice versa), this is not believed enough to account for the current distribution of IS elements (Condit, 1990). The occurrence of rare beneficial IS-mediated mutations is assumed to balance the many neutral or deleterious mutations and thus account for their long-term maintenance (Blot, 1994). In the recent years, the impact of IS elements could be analyzed in the test tube because bacteria grow fast enough to give several thousands of generations in relatively short periods. Mutants with a restriction fragment length polymorphism based on IS
elements (RFLP-IS) were numerous in a long-term experimental evolution with the bacterium Escherichia coli and IS elements were shown good markers to monitor bacterial evolution (Papadopoulos et al., 1999). Some of these IS mutations became fixed very early and then conserved, while the relative fitness of the population increased from 1 to ca. 1.45 after 10,000 generations (Lenski and Travisano, 1994). These “pivotal” mutations were identified as IS insertions, IS-mediated deletions and inversions (Schneider et al., 2000; Cooper et al., 2001). Replacement of the ancestral allele at six loci into the genomic background of the evolved clones showed that these IS-mediated mutations were indeed beneficial (Cooper et al., 2001; Schneider and coll., unpublished results, Rozen and coll., unpublished results). Other similar studies have identified beneficial IS-mediated mutations (Treves et al., 1998; Faure, D., Portier, P. and Blot, M., unpublished data), showing that IS elements can be useful accessory elements (see Chalmers and Blot (1999) for discussion). The long-term maintenance of IS elements can now be viewed as a trade between many neutral or deleterious mutations slowly crippling the bacterial genomes and rare beneficial mutations hitchhiking for the presence of IS elements in populations. The conditions for an equilibrium can be searched with a mathematical model based on the observed features of IS elements in Escherichia coli and data on the evolution of fitness during a 10,000 generation experiment. In this paper, we developed a stochastic model for the population dynamics to examine the effect of individual mutations on population fitness. We thus bridge the gap between population dynamics and genome plasticity by studying the contribution of IS elements to evolution. This study accounts for a variety of experimental results and suggests that IS-mediated rearrangements play a significant role in bacterial adaptation.

THE MODEL FOR BACTERIAL EVOLUTION

We consider a population made of one isogenic cell type, in which mutant clones can arise. Depending on its fitness and chance, each mutant clone will survive and eventually expand. Each day, the culture is allowed to duplicate 100 times, and thus 99% are eliminated at the dilution time (Fig. 1). This results in 6.6 generations of bacteria and a population size varying from 5e−6 to 5e−8 individuals. The mutation rate is derived from estimations at the bgl locus in our lab (6.4e−8 mutation per cell division), and Hall (1999) finds a similar value at the ebg locus (6.3e−8 mutation per cell division). Both estimations comprise a broad spectrum of mutations, in which IS elements give the major contribution. In the ancestor of the 12 lines, nucleotide-substitution rates were found ca.1e−9 for ara+ reversion, ca. 6e−11 for nalidixic-acid resistance and ca. 6e−8 for phage-T5 resistance (Sniegowski and Lenski, 1997). The mathematical characterization of the population growth and the generation of mutants is given in Appendix A and the simulation algorithm is presented in Appendix B.

In the course of the population increase, the mutants generated may be allocated a different duplication rate, thus a different fitness. To generate the values for this parameter, we use the following probability density function:

$$f(x) = \begin{cases} \frac{1 - p}{a}, & 0 \leq x \leq a, \\ \frac{1 - p}{a} \exp \left( - \frac{(x - a)(1 - p)}{ap} \right), & a < x, \end{cases} \quad (1)$$

where $a$ is the duplication rate of the parent clone and $p$ is the probability that a new clone duplicates faster than its parent. The exact form of the density probability to get a duplication rate smaller than $a$ is of little

![FIG. 1. Representation of the algorithm used in simulations of population dynamics. Each day the culture is diluted 1:100 (6.6 generations per day); after a lag phase lasting initially 1 h, the bacteria grow and reach stationary phase. During that time they can generate deleterious, neutral or favorable mutations. The next day, 99% of the cells are randomly eliminated at the time of dilution. The simulation is made over 10,000 generations (1515 days).](image-url)
importance; this is accounted for by using the uniform density over the interval $[0,a]$. However, to prevent unbounded duplication rate values, we assume that the density function decays exponentially fast on the interval $[a,\infty)$ (Crow and Kimura, 1970). It should be noted that the expression for $f$ (Eq. (1)) is very simple and depends only on the parameter $p$ and the exponential behavior for large $x$. According to Eq. (1), the population fitness would increase in an unbounded manner in the course of time, a fact which is not supported by the experimental results. To prevent this, we assume that the parameter $p$ decreases as the fitness increases and the actual parameter $p$ used in the simulations is

$$p(t) = p_0 \exp(-\beta(F(t) - 1)), \quad (2)$$

where $p_0$ is the initial value of $p$, and $F(t)$ is the population fitness at time $t$. Our choice was motivated by the following property:

$$p(t + s) = p(t) \exp(-\beta(F(t + s) - F(t))), \quad t > 0, \quad s > 0. \quad (3)$$

If conditions (experimental or simulations) were changed at time $t$, the subsequent evolution of $p$ would start afresh from $p(t)$ and would not depend on the previous history of the population. In addition, from Eq. (2), it follows that

$$\beta = -\frac{1}{p(t)} \frac{dp(t)}{dF(t)}. \quad (4)$$

Therefore, the parameter $\beta$ may be seen as a sensibility index of the population to its new culture conditions. When sensitivity is low ($\beta \geq 0$), the probability of best-fit mutants is constant. With higher values of $\beta$, $p$ decreases when fitness increases. All together the model agrees that in these populations “the fixation of a given beneficial mutation is a decreasing function of both population size and mutation rate” (Gerrish and Lenski, 2000), but also that it is more difficult to generate a best-fit mutation while fitness increases.

**RESULTS**

**Proportion of Beneficial Mutations**

In this model, we explore first the proportion of beneficial mutations depending on time. In Lenski and Travisano (1994), the first step-increase in fitness equals 0.13; from Eq. (1), we derived that the average step-increase in fitness is $(E(x|x > a)/a) - 1 = p/(1 - p)$ and the initial probability of best-fit clones $p_0^* = 0.11$. In other words, this suggests that 11% of the mutants are expected to have a better fitness than their parent until a beneficial mutant would take over. To set the sensitivity index (a constant imposing that $p$ decreases when fitness increases), simulations were generated with $p_0^*$ and $\beta$ in [0–4] (Fig. 2). For $\beta = 0$, the relative fitness increased in an unbounded manner up to 4 at 10,000 generations. For $\beta = 4$, fitness did not increase enough to match with the experimental data because the probability of being better fit than the parent became too low in the long term. The best estimate was $\beta = 1.4$, leading to a final fitness of 1.45 after 10,000 generations.

The second set of simulations was performed to display the evolution of $p$ with time. Obviously, $\beta$ and $p_0$ are linked factors but we tested other values of $p_0$ in [0.05–0.2] to estimate the model behavior around $p_0^*$ (Fig. 3). Interestingly, all plots converged to 0.04 at 10,000 generations indicating that the probability of making best-fit mutants is lower-limited when fitness has substantially improved. Below that value, the chance that a beneficial mutation arises and settles in a population was low because the cultures are diluted 1:100 each day. Moreover, the initial value ($p_0$) was not critical because the plots decreased rapidly with fitness increase. The average fitness increment after 10,000

![FIG. 2. Evolution of fitness during 10,000 generations as a function of the adaptability index $\beta$ in [0–4]. Each plot is the average of 50 simulations with $p_0^* = 0.1$. Bold curves are compatible with the experimental data (Lenski and Travisano, 1994), and $\beta = 1.4$ provided the best estimate (not shown).](image-url)
generations observed by Lenski and Travisano (1994) was between 1.4 and 1.52 depending on the lines, and their standard deviation was ca. 0.05. Simulations with $p_0 = 0.1$ provided the best estimate to their data for fitness evolution over 10,000 generations.

**Evolution of Fitness and Mutations**

One simulation among 50 at $p_0 = 0.1$ was arbitrarily chosen to understand the link between population fitness and the number of beneficial mutations (Fig. 4). There were seven steps in this simulation, which increased fitness successively by 0.08, 0.08, 0.05, 0.04, 0.07, 0.05 and 0.05. This simulation led to a final fitness of 1.42, thus at the lower limit observed with the experimental lines (Lenski and Travisano, 1994). Thus, as recorded previously with the experimental data (Lenski and Travisano, 1994), a small number of beneficial mutations could settle in the populations and thus explain the evolution of fitness during 10,000 generations.

Using the same simulation, the number of mutants and their characteristics were detailed (Table I). There were 46,979 mutants generated over 10,000 simulated generations. Among these mutants seven were fixed, leading to a frequency of $1.4 \times 10^4$ for the beneficial and successful mutations. Amongst the mutants generated, 2126 (4.5%) experienced a fitness improvement but 2119 were not retained in the simulation. This was due to: (i) drift (they were eliminated at the first 1:100 dilutions) or (ii) the encounter of an even fitter mutant. Thus, most mutations produced in the simulations were purged by periodic selection, irrespective of their fitness.

**TABLE 1**

<table>
<thead>
<tr>
<th>Generation</th>
<th>Mutants generated</th>
<th>Beneficial mutants generated (%)</th>
<th>Beneficial mutants fixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>2308</td>
<td>224 (9.7)</td>
<td>1</td>
</tr>
<tr>
<td>1000</td>
<td>4675</td>
<td>401 (8.5)</td>
<td>2</td>
</tr>
<tr>
<td>1500</td>
<td>7065</td>
<td>525 (7.4)</td>
<td>3</td>
</tr>
<tr>
<td>2000</td>
<td>9346</td>
<td>649 (6.9)</td>
<td>3</td>
</tr>
<tr>
<td>2500</td>
<td>11,710</td>
<td>784 (6.7)</td>
<td>4</td>
</tr>
<tr>
<td>3000</td>
<td>14,057</td>
<td>907 (6.4)</td>
<td>5</td>
</tr>
<tr>
<td>4000</td>
<td>18,790</td>
<td>1084 (5.7)</td>
<td>5</td>
</tr>
<tr>
<td>5000</td>
<td>23,449</td>
<td>1284 (5.5)</td>
<td>6</td>
</tr>
<tr>
<td>8000</td>
<td>37,489</td>
<td>1770 (4.7)</td>
<td>7</td>
</tr>
<tr>
<td>10,000</td>
<td>46,979</td>
<td>2126 (4.5)</td>
<td>7</td>
</tr>
</tbody>
</table>

*Percent of beneficial to total mutants.*
Evolution of Population Structure

The same simulation was used to monitor population structure (Fig. 5). Here, not only the seven mutants improving fitness were seen, but also 14 other mutants, which had a transient success in the population and were replaced by even fitter mutants. The seven clones were fixed rapidly and they dominated the populations for 990, 1403, 2000, 1689, 2772, 2013 generations, respectively (the last one was alive at generation 10,000, thus after 5077 generations). The average lifetime of a mutant clone was 1810 generations (min 990; max 2772) and only one clone was alive at generation 10,000. The pedigree of this simulation could be reconstructed to display the mutant features (Fig. 6). As expected in an asexual haploid organism, polymorphism was only transient and it corresponded to the replacement of a clone by another. With fitness steps lower than 0.1 and a population size varying from \(5 \cdot 10^6\) to \(5 \cdot 10^8\), the replacement took less than 800 generations. Due to a large fitness step, the fastest to be eliminated was the ancestor at generation 613, thus 567 generations after the birth of the first successful mutant.

DISCUSSION

For most geneticists, mutations are often thought as functional alterations of the wild type in their experiments, and the perception of mutations as a potential for evolution is thus low. Since also phylogenies are built with the assumption that nucleotide substitutions accumulate neutrally, the impact of beneficial mutations on fitness in evolutionary biology is not well documented with experimental data. In the recent years, microbial studies have contributed useful data to estimate the role of beneficial mutations on evolution. Experimental data on the fitness evolution of 12 lines of the bacterium \(E. coli\) have shown that bacteria can adapt their life history traits to a new environment (Lenski and Travisano, 1994) with successive families of siblings, each fixing new genomic changes (Papadopoulos et al., 1999). To date, only IS-mediated mutations (transposition and IS-borne deletions and inversions) have been genetically characterized in Lenski’s collections (Cooper et al., 2000; Schneider et al., 2001; Riehle et al., 2001). This is certainly because a point mutation is difficult to identify within the 4.6 Mbp of the \(E. coli\) sequence, and systematic sequencing at a few arbitrary loci was not fruitful (M. Travisano and R. Lenski, pers. comm.).

A deterministic model on the evolution of these lines based on clonal interference (Gerrish and Lenski, 2000) derived a fixation probability for beneficial mutations to account for the observed rate of fitness increase (Lenski and Travisano, 1994) and studied the polymorphism in these populations. In particular, the sequential substitution of a genotype by another was well documented. However, since this model ignored deleterious

---

**FIG. 5.** Evolution of population structure. Seven favorable mutations were observed in the same simulation as in Fig. 4. Fourteen beneficial mutants reached significant frequencies but did not succeed to colonize the population and the polymorphism was transient.
mutations, it cannot provide correct estimates for the proportion of beneficial to total mutations (see Orr (2000) for discussion). In the present stochastic model, we have minimized the assumptions to estimate the contribution of beneficial mutations by simulations. Here, deleterious, neutral and beneficial mutations are generated, but the probability of a mutant for being more beneficial than its parent decreases. Then, by chance (being retained at the dilution time) and depending on their randomly attributed fitness, mutants can invade the population and replace their ancestor. After many trials, we generated a probability law for the beneficial mutants, which depends on the initial proportion of beneficial mutations and on a constant imposing the decrease. We found a non-critical value suggesting that 11% of the mutants should be beneficial before the first beneficial mutation was fixed, and this value decreased to 4% at generation 10,000. Because of the experimental design (daily dilution and small population size) most mutants got lost, including the majority of best-fit mutants. A posteriori, the proportion of beneficial mutants reaching fixation was seen as $1.4e^{-C0}$, and one may calculate the efficient mutation rate for beneficial mutations as $1.4e^{-C0} 	imes 6.4e^{-8} = 9e^{-12}$ per cell division. This value is in agreement with the idea that beneficial mutations are rare but they can reshuffle the population structure if the culture is long or/and large.

FIG. 6. Pedigree. In the same simulation as in Figs. 4 and 5, 46,979 mutants were generated, but only seven (0.014%) contributed to the population structure. Each mutant is labeled by (1) its identity number; (2) its birth and death date (generation); (3) its maximum frequency and its production of secondary mutants; and (4) its relative fitness to the ancestor.
We have generated plots for the fitness increase that match the experimental data and indicate that seven mutations per bacterial population contribute a fitness increase of 1.43 after 10,000 generations. In one simulation, the population structure was detailed: the first four mutations provided a relative fitness of 1.08, 1.17, 1.22, 1.27, successively (Fig. 4). It was tempting to compare these results to experimental data obtained in one line (ara+1) for which pivotal IS-mediated mutations had been identified (Papadopoulos et al., 1999). Mutations were within the following loci: *nadR*, *hok-sok* and *phpA-rodA* (Schneider et al., 2000) and in the *rbs* operon (Cooper et al., 2001). Recent data on reconstructed genotypes in the ara+1 line show that *nadR*, *hok-sok* and *phpA-rodA* mutations contribute a fitness increment in this range (Schneider, Woods, Lenski and Blot, unpublished results), while the *rbs* mutation has a smaller fitness improvement (0.02) compared to the ancestor (Cooper et al., 2001). More data are certainly necessary to fully describe the evolution of a lineage over 10,000 generations, but the current data document that most fitness changes in the ara+1 population can now be explained, and they are IS-related mutations.

This is maybe not surprising since three kinds of fittest mutations can be expected: (i) those inventing new functions; (ii) those creating new gene regulations or new molecules fluxes; (iii) those making significant genome savings. Obviously, IS elements have a major effect on gene regulation (Chalmers and Blot, 1999) and most mutations identified in the different lines (Schneider et al., 2000) and in other studies (Trevis et al., 1998; Faure, D., Portier, P. and Blot, M., unpublished data) change gene expression. The model thus suggests that a large part of the ara+1 genome optimization as an adaptation to its new environment was due to IS-mediated mutations.

Our model is not optimistic for the continuation of the experimental evolution. Due to a decreasing probability of making best-fit mutants and the daily dilution, the chance to pick a new beneficial mutant at generation 30,000 is extremely low and the mutant should be washed out at the first dilutions. If, however, that mutant would survive, its genetic analysis would be of a great value.

There are some discrepancies between our model and the data. First, polymorphism is larger in real than in the model. Second, the model only documents beneficial mutations, not neutral or deleterious mutations that could maintain. These two points make one if polymorphism is neutral. Indeed, the genetic distance calculated in Papadopoulos et al. (1999) is larger. Part of the explanation is because RFLP-IS often reports two changes for most IS-mediated mutations (conservative transposition, inversion, deletion). Moreover, the stochastic model being dedicated to individuals able to survive, we have only recorded the mutations that reshuffled the population structure. Thus, we do not document hitchhiking—i.e., secondary mutations that are not beneficial but are retained on the bacterial chromosome together with a beneficial mutation. Since several hundred generations are required before fixation of a beneficial mutation, hitchhiking is likely, especially after the fitness steps level off. The phenomenon is illustrated by the accumulation of IS150 copies (5–22 copies at generation 10,000) associated with a large genomic polymorphism (Papadopoulos et al., 1999).

Most of these mutations are seen after the fitness has gained its maximum and they thus should not contribute to a better fitness.

In the present model, one mutation rate summing all types of mutations was used. The combination of different mutation rates and fitness distributions for each mutational sources might promote a pattern in which neutral mutations can invade large clonal populations. More sophisticated stochastic models, together with more genetic data are required to address this question.

**APPENDIX A**

**Population Growth and Generation of Mutants**

1. **Clone-size characterization:** Let \((X(t), Y(t))\) be a couple of random variables for the clone size at time \(t\) and the number of mutations generated during the clone growth in the interval \([0, t]\); initially, one has: \(X(0) = N\) and \(Y(0) = 0\). Assuming that time, \(t\), is continuous (generations are asynchronous), we first characterize the jump probabilities for \((X(t), Y(t))\). We assume that bacteria of the same clone behave independently. To facilitate the model derivation, we remark that the random variable \(X(t)\) has the following Markov property: \(X(t|t_1)\) and \(X(t_0)\) with \(0 < t_0 < t_1 < t\) are independent. In addition, the distribution of \(Y(t)\) over \([t_0, t]\) is determined by \(X(t_0)\) only and is independent of the number of mutations generated in \([0, t_0]\).

To save notation complexity, we did not mention explicitly that all the probabilities listed below (A1)–(A3) are conditioned by

\[X(t) = n_X, \quad Y(t) = n_Y,\]
where \( n_X \) and \( n_Y \) are two positive integers. In Eqs. (A1)–(A3), \( k_X, k_Y \) are positive integers, \( h \) is a positive real number, \( \mu \) is the duplication rate, \( q \) is the probability that a cell division gives at least one bacterium with a mutation, \( o(x) \) refers to any continuous real function verifying

\[
\lim_{x \to 0} \frac{o(x)}{x} = 0.
\]

Because cell death is ruled out during the exponential growth phase and because the mutations accumulate in time, one has

\[
\Pr(X(t + h) - X(t) < 0; Y(t + h) - Y(t) = k_Y) = 0,
\]
\[
\Pr(X(t + h) - X(t) = k_X; Y(t + h) - Y(t) < 0) = 0. \tag{A1}
\]

The probability for transitions larger than 1 are vanishing functions of the time step \( h \):

\[
\Pr(X(t + h) - X(t) \geq 2; Y(t + h) - Y(t) = k_Y) = o(h),
\]
\[
\Pr(X(t + h) - X(t) = k_X; Y(t + h) - Y(t) \geq 2) = o(h). \tag{A2}
\]

Other transitions are:

\[
\Pr(X(t + h) - X(t) = 1; Y(t + h) - Y(t) = 0) = n_X \mu(1 - q)h + o(h),
\]
\[
\Pr(X(t + h) - X(t) = 1; Y(t + h) - Y(t) = 1) = 0,
\]
\[
\Pr(X(t + h) - X(t) = 0; Y(t + h) - Y(t) = 0) = 1 - n_X \mu h + o(h),
\]
\[
\Pr(X(t + h) - X(t) = 0; Y(t + h) - Y(t) = 1) = n_X \mu q h + o(h). \tag{A3}
\]

Let \( P_{n_X,n_Y}(t) \) be \( \Pr(X(t) = n_X, Y(t) = n_Y | X(0) = N, Y(0) = 0) \). The marginal probability distribution for \( X(t) \), denoted \( Q \), is \( Q_{n_X}(t) = \Pr(X(t) = n_X | X(0) = N) = \sum_{n_Y} P_{n_X,n_Y}(t) \). From the transition probabilities (A1)–(A3), we conclude that \( P \) obeys an infinite set of ordinary differential equations, namely,

\[
\frac{d}{dt} P_{n_X,n_Y}(t) = (n_X - 1) \mu(1 - q)P_{n_X-1,n_Y} + n_X \mu q P_{n_X,n_Y-1} - n_X \mu P_{n_X,n_Y},
\]

\[
P_{n_X,n_Y}(0) = \begin{cases} 
1 & \text{if } n_X = N \text{ and } n_Y = 0, \\
0 & \text{otherwise.} 
\end{cases} \tag{A4}
\]

To solve this differential equation system, we introduce the generating function

\[
g(s, u, t) = E s^{X(t)} u^{Y(t)},
\]

which verifies

\[
\frac{\partial g}{\partial t} + s \mu (1 - (1 - q) u - q u) \frac{\partial g}{\partial s} = 0,
\]

\[
g(s, u, 0) = s^N.
\]

The solution is

\[
g(s, u, t) = \left( \frac{s(1 - qu)}{(1 - q) s + (1 - (1 - q) s - qu) \exp(\mu t (1 - qu))} \right)^N. \tag{A5}
\]

The generating function for the marginal probability distribution \( Q \) is obtained by setting \( u = 1 \) into (A5):

\[
g_X(s, t) = \left( \frac{s \exp(-\mu(1-q)t)}{1 - s(1 - \exp(-\mu(1 - q)t))} \right)^N \tag{A6}
\]

to which corresponds the elementary probability distribution

\[
Q_{N+1}(t) = \binom{N + k - 1}{k} \exp(-N\mu(1-q)t)(1 - \exp(-\mu(1-q)t))^k, 
\]

\[ k \geq 0, \quad N \geq 1. \]

Expectation and variance for \( X(t) \) are:

\[
E X(t) = N \exp(\mu t (1 - q)),
\]
\[
\text{Var } X(t) = N \exp(\mu t (1 - q))(\exp(\mu t (1 - q)) - 1).
\]

This stochastic process, known as the Yule-Furry process of parameter \( \mu(1 - q) \), accounts for the daily exponential growth phase in Fig. 1. To obtain the generating function of the marginal distribution for \( Y(t) \) (with \( X(0) = N \) and \( Y(0) = 0 \)), we set \( s = 1 \) into (A5):

\[
g_Y(u, t) = \left( \frac{1 - qu}{1 - q + qu(1 - u) \exp(\mu t (1 - qu))} \right)^N. \tag{A7}
\]

The expectation and the variance of the number of mutations in the interval \([0, t]\) are:

\[
E Y(t) = N \frac{q}{(1 - q)}(\exp(\mu t (1 - q)) - 1),
\]
\[
\text{Var } Y(t) = N \frac{q}{(1 - q)^2}q(\exp(2\mu t (1 - q))
\]
\[
+ \exp(\mu t (1 - q))(1 - q - 2\mu t (1 - q)) - 1).
\]
APPENDIX B
Numerical Simulations

(1) Simulation of the growth phase: Growth is simulated according to the elementary probability jumps listed in Appendix I. Let us denote by \( X(t) \) the clone size at time \( t \). Using the jump probabilities (Eqs. (A1)–(A3)) and the Markov property for the couple of random variables \( (X(t), Y(t)) \), one may split the simulation of the clone growth into \( X(t) \) parallel elementary steps, each of them corresponding to one bacterium. The population net change \((AX,AY)\) resulting from one of these steps is generated according to

\[
\begin{aligned}
AX &= 1, & \text{with probability } \mu(1-q)h, \\
AY &= 0, \\
AX &= 0, & \text{with probability } \mu q h, \\
AY &= 1, \\
AX &= 0, & \text{with probability } 1-\mu h.
\end{aligned}
\]

Because of the independence between the bacteria, one gets

\[
X(t+h) = X(t) + \sum_{i=1}^{X(t)} AX_i,
\]

\[
Y(t+h) = \sum_{i=1}^{X(t)} AY_i,
\]

where we set \( Y(t) = 0 \). The value of \( Y(t+h) \) is used to generate new clones of size 1; these clones will be taken into account in the next simulation step \((t + h, t + 2h)\). The duplication rates associated with these new-born clones are randomly generated according to the probability distribution presented in the text (Eq. (1)).

(2) Simulation of the selection phase: In the experimental conditions, one-hundredth of the total population, chosen by random sampling, is inoculated in the next day culture. In the present model, we generate a random sample according to the hypergeometric distribution

\[
p_{k_1, \ldots, k_n} = \binom{f_1 N}{k_1} \cdots \binom{f_n N}{k_n} \frac{N}{dN},
\]

\[
\sum_{i=1}^{n} f_i = 1, \quad \sum_{i=1}^{n} k_i = dN,
\]

where \( N \) is the total population size before sampling, \( n \) is the number of clones in the population, \( f_i \) is the frequency of the \( i \)th clone in the population, \( d \) is the dilution factor, and \( p_{k_1, \ldots, k_n} \) is the probability that the sample is made of \( k_1 + \cdots + k_n \) cells, where \( k_i \) is the number of cells from the \( i \)th clone surviving the dilution. When \( N \) is large (5e8) and \( f_i \) close to one, the \( i \)th clone is represented by \( dN f_i \) cells in the sample. Conversely, with a large \( N \) and a low-frequency \( f_i \) (ca. \( N^{-1} \)), we use the classical Poisson limit of the hypergeometric distribution.

(3) Simulation for large clones: The algorithm presented in Sections 1 and 2 can be used, whatever the size of the clone. However, calculation time was saved by using the limit distribution for a large number of cells. In the limit of a large clone, we derive from distribution (A6) that \( X(t) \) is given by

\[
X(t+h) = \mu X(t,h) + \sigma X(t,h) N(0,1),
\]

where

\[
\mu X(t,h) = X(t)e^{p(1-q)h},
\]

\[
\sigma^2 X(t,h) = X(t)e^{p(1-q)h}(e^{p(1-q)h} - 1)
\]

and \( N(0,1) \) is the Laplace–Gauss random variable of expectation 0 and variance 1.

Similarly, most of the mutations are generated by the larger clones in the population and it would be useful to obtain the corresponding limit for the distribution probability of the mutations. We introduce

\[
q = \frac{\rho}{\rho + N},
\]

where \( \rho \) is real and positive. The characteristic function for \( Y(t) \) becomes (cf. relation (A7))

\[
g_Y(u,t) = \left( 1 + \frac{1}{\rho/N}(1 - e^{iu}) \right)^N.
\]

For large \( N \) and finite \( \rho \), one has

\[
\lim_{N \to \infty} g_Y(u,t) = \hat{g}_Y(u,t) = \exp((e^{iu} - 1)\rho(e^{iu} - 1)).
\]

Since \( \hat{g}_Y(\mu,t) \) is continuous at \( u = 0 \), the Levy theorem ensures that the mutation distribution converges to a Poisson distribution of parameter

\[
\rho(e^{iu} - 1).
\]

The number of mutation generated by a large clone \( X \times (t) > 10e4 \) in the interval \( (t, t+h) \) is simulated according to a Poisson distribution of parameter

\[
\frac{X(t)q(e^{iu} - 1)}{1-q}.
\]
In a preliminary study, we numerically tested these limit distributions and their effective use in the simulations. For the threshold values indicated in this appendix, no difference can be detected between numerical simulations using Eq. (B1) or the limit distributions (B2) and (B3) (χ² test, results not shown).

ACKNOWLEDGMENTS

This work was supported by the CNRS (GDR2157) and by the CNRS, INSERM INRIA (Bioinformatics Program). We thank D. Schneider and R.E. Lenski for fruitful discussions and R. Woods for sharing unpublished data.

REFERENCES


