Mathematical Biology

Paul Joyce · Zaid Abdo · José M. Ponciano · Leen De Gelder · Larry J. Forney ·
Eva M. Top

Modeling the impact of periodic bottlenecks, unidirectional mutation, and observational error in experimental evolution

Abstract. Antibiotic resistant bacteria are a constant threat in the battle against infectious diseases. One strategy for reducing their effect is to temporarily discontinue the use of certain antibiotics in the hope that in the absence of the antibiotic the resistant strains will be replaced by the sensitive strains. An experiment where this strategy is employed in vitro [5] produces data which showed a slow accumulation of sensitive mutants. Here we propose a mathematical model and statistical analysis to explain this data.

The stochastic model elucidates the trend and error structure of the data. It provides a guide for developing future sampling strategies, and provides a framework for long term predictions of the effects of discontinuing specific antibiotics on the dynamics of resistant bacterial populations.

1. Introduction

Bacteria expressing antibiotic resistant genes are of increasing concern. In the presence of an antibiotic the bacteria that carry resistant genes have a tremendous competitive advantage and quickly sweep through the population. Genes encoding

P. Joyce, Z. Abdo, J. M. Ponciano: Department of Mathematics, Division of Statistics, Initiative in Bioinformatics and Evolutionary Studies (IBEST), P.O. Box 441103, University of Idaho, Moscow, ID 83844-1103 e-mail: `Paul Joyce joyce@uidaho.edu`,

L. De Gelder, L. J. Forney, E. M. Top: Department of Biological Science, Initiative in Bioinformatics and Evolutionary Studies (IBEST), P.O. Box 443051, University of Idaho, Moscow, ID 83844-3051

resistance to antibiotics are often located on plasmids. It is often assumed that the carriage of an antibiotic resistance encoding plasmid is at a cost to the host when this host resides in a medium where the antibiotic is not present [1]. However, it is not known whether this cost gives resistant bacteria a considerable disadvantage in a neutral environment causing the antibiotic sensitive bacteria to replace the bacteria with the resistance genes, or if antibiotic resistant genes persist in such an environment. An experiment [5] designed to address these questions for a particular plasmid-encoded tetracycline resistance gene produced a very interesting pattern. At the beginning of the experiment nearly all the bacteria under study contained the resistance gene while a small fraction did not. When the bacteria were allowed to evolve in a neutral environment, the proportion of antibiotic sensitive bacteria increased. However, the increase was quite slow and erratic. After 500 generations the fraction of sensitive bacteria increased on average from 0.15 percent to 6 percent of the population. All clones examined had lost the entire tetracycline resistance operon due to deletion of the corresponding plasmid region (De Gelder et al. [5]). While there was an average trend upward, there was a considerable amount of variability. This led the investigators to consider the following explanation: There is little to no fitness advantage conferred to the nonresistant cells, and the slow accumulation of sensitive mutants was mainly due to unidirectional mutation. That is, from time to time the plasmid in the tetracycline (tet) resistant bacteria mutates, resulting in the deletion of the tet resistant gene.

A brief description of the experimental evolution that we are modelling is as follows. An initial number of bacteria, say N_0 (usually on the order of 10^7/ ml), is placed in a flask and the population doubles each generation for l generations. l is about 8. The process is then subjected to a series of bottlenecks. That is, at the end of each period of length l generations a sample of size N_0 is taken from the population that has grown to $2^l N_0$. This produces what we call a cycle of the experiment, where a cycle is defined to be a combination of a growth period and a bottleneck occurring at the end of that period. The sample is then placed in a new flask and the procedure starts anew.

Here we develop a stochastic model that captures the variability in an experiment where unidirectional mutation and no selection explains the data. Our model assumes that the rate of mutation from antibiotic resistant bacteria to sensitive bacteria is much higher than the reverse and that we can only detect the difference between the rates of mutation. It is this unidirectional mutation process that allowed the mutant to increase in frequency over time. The statistical model described here is in the spirit of those introduced in [14], and [15] but with important differences. The focus of those models is to provide a theoretical basis for understanding certain aspects of adaptive evolution in a controlled environment. This paper is motivated by data where an explanation involving unidirectional mutations and drift due to bottlenecks is explored. As a result the focus here is to understand certain patterns of variability for experimental evolution where no defined selection is evident.

In addition to developing the mathematical model, we develop a statistical procedure for parameter estimation and a goodness of fit test. We then apply our statistical methodology to our data.

2. The stochastic model

The experimental evolution process described in the second paragraph of the introduction (Section 1) has two sources of variability. The first involves the bottleneck. The percentage of mutants sampled at the beginning of the kth cycle could be higher or lower than the percentage of mutants at the end of the $(k-1)$st simply due to the error associated with the bottleneck. The genetic effects of such fluctuating environments is explained in [10]. The second source of variability involves the mutation process.

Taking these two sources of variability into account, and assuming that mutants arise at random with an average mutation rate of λ per individual per generation, gives rise to the following model.

- Let M_k be the number of mutants at the end of the kth cycle.
- Let Y_k be the number of mutants that arise during the kth cycle.
- Let X_k be the number of mutants sampled from the $(k-1)$st cycle or the number of mutants at the beginning of the kth cycle.
- Let l be the number of generations in a cycle.
- Let $A_k = M_k/(2^l N_0)$ be the fraction of mutants at the end of the kth cycle.
- Let $V_{j,k}$ be the number of mutants that arise during generation j ($j \le l$) of the kth cycle.
- Let $Y_{j,k}$ be the number of mutants at generation j that arise during the kth cycle. Note that

$$Y_{j,k} = 2Y_{j-1,k} + V_{j,k} \tag{1}$$

and $Y_k \equiv Y_{l,k}$.

The distribution of M_k is determined by the distributions of X_k and Y_k.

$$M_k = 2^l X_k + Y_k. \tag{2}$$

These distributions are determined by the previous generation $k-1$. Note that the bottleneck determines the conditional distribution of X_k given M_{k-1} to be binomial, which can be approximated by the Poisson distribution. That is

$$X_k|M_{k-1} \sim \mathrm{BIN}\left(N_0, \frac{M_{k-1}}{2^l N_0}\right) = \mathrm{BIN}\left(N_0, A_{k-1}\right) \sim \mathrm{POI}\left(\frac{M_{k-1}}{2^l}\right). \tag{3}$$

The error due to the mutation process is determined by the number of mutants that arise during that cycle, which is determined by the number of mutants that arise during each generation of the cycle, $V_{j,k}$, using the following conditional distribution

$$V_{j,k}|Y_{j-1,k}, X_k \sim \mathrm{POI}\left(\lambda\left(2^j(N_0 - X_k) - 2Y_{j-1,k}\right)\right). \tag{4}$$

Now that we have the model set up, we use it to calculate $E(M_k)$ the mean number of mutants after k cycles, $\mathrm{Var}(M_k)$, and $\mathrm{Cov}(M_k, M_j)$. We then use these calculations to form a moment estimator for λ, to calculate the error associated with the estimator and to develop a goodness of fit criteria.

2.1. Means

In this section we use the stochastic model described above to derive an expression for the mean number of mutants present at the end of the kth cycle, $E(M_k)$, and the mean fraction of mutants, $E(A_k)$.

We begin by defining $y_{j,k} = E(Y_{j,k}|X_k)$ and note that it follows from equations (1) and (4) that

$$E(Y_{j,k}|Y_{j-1,k}, X_k) = 2(1-\lambda)Y_{j-1,k} + \lambda 2^j(N_0 - X_k). \tag{5}$$

Which implies

$$y_{j,k} = 2(1-\lambda)y_{j-1,k} + \lambda 2^j(N_0 - X_k).$$

Let $\alpha_j = \dfrac{y_{j,k}}{2^j(N_0 - X_k)}$ then

$$\alpha_j = (1-\lambda)\alpha_{j-1} + \lambda.$$

By definition $y_{0,k} = 0$, so it follows by (35) that, again,

$$\alpha_j = 1 - (1-\lambda)^j$$

and

$$E(Y_k|X_k) = E(Y_{l,k}|X_k) = y_{l,k} = \alpha_l 2^l(N_0 - X_k). \tag{6}$$

Therefore, it follows from (6) and (2) that

$$E(M_k|X_k) = 2^l X_k + E(Y_k|X_k) = 2^l(1-\alpha_l)X_k + 2^l \alpha_l N_0 \tag{7}$$

and by definition $E(X_k|M_{k-1}) = M_{k-1}/2^l$, therefore

$$E(M_k|M_{k-1}) = (1-\alpha_l)M_{k-1} + 2^l \alpha_l N_0 \tag{8}$$

implying

$$E(M_k) = (1-\alpha_l)E(M_{k-1}) + 2^l \alpha_l N_0.$$

Now define $\beta_k = E(A_k) \equiv \dfrac{E(M_k)}{2^l N_0}$ to be the mean fraction of mutants after k cycles. This leads to the following recursion

$$E(A_k) = \beta_k = (1-\alpha_l)\beta_{k-1} + \alpha_l = 1 - (1-\beta_0)(1-\lambda)^{lk} \tag{9}$$

2.2. *Variance*

The purpose of this section is to calculate the variance of the number of mutants at the end of the kth cycle, $\mathrm{Var}(M_k)$, and from it will follow the variance of the fraction of mutants, $\mathrm{Var}(A_k)$. The variability associated with M_k depends on the variability associated with the number of mutants transferred from the previous cycle and the variability associated with the number of mutants that arise during the cycle. The following standard formula for variance shows how these two are related to the overall variance.

$$\mathrm{Var}(M_k) = E(\mathrm{Var}(M_k|X_k)) + \mathrm{Var}(E(M_k|X_k)).$$

If we condition on the number of mutants at the beginning of the cycle X_k then $\mathrm{Var}(M_k|X_k)$ represents the variability associated with the new mutants that arise during the cycle. It follows from (7) that

$$\mathrm{Var}(E(M_k|X_k)) = \mathrm{Var}\left(2^l(1-\alpha_l)X_k + 2^l\alpha_l N_0\right) = 2^{2l}(1-\alpha_l)^2\mathrm{Var}\,(X_k)$$

and so $\mathrm{Var}(E(M_k|X_k))$ is determined by the variance of X_k, the number of mutants at the beginning of the cycle.

Recall in equation (3) we noted that $X_k|M_{k-1}$ is Poisson distributed. Now using the conditional variance formula again we write

$$\mathrm{Var}(X_k) = E\left(\frac{M_{k-1}}{2^l}\right) + \mathrm{Var}\left(\frac{M_{k-1}}{2^l}\right)$$
$$= N_0\beta_{k-1} + \frac{\mathrm{Var}(M_{k-1})}{2^{2l}}. \tag{10}$$

Combining these results gives

$$\mathrm{Var}(M_k) = E(\mathrm{Var}(M_k|X_k)) + 2^{2l}(1-\alpha_l)^2 N_0\beta_{k-1} + (1-\alpha_l)^2\mathrm{Var}(M_{k-1}). \tag{11}$$

To complete the recursion we need to develop a recursion for $E(\mathrm{Var}(M_k|X_k))$. It follows from equation (2) that

$$\mathrm{Var}(M_k|X_k) = \mathrm{Var}((2^l X_k + Y_k)|X_k) = \mathrm{Var}(Y_k|X_k) \tag{12}$$

and so this term depends only on the variability associated with the number of mutants that arise during the kth cycle Y_k. To solve for $\mathrm{Var}(Y_k|X_k)$ we need to consider each generation during the kth cycle. Using the conditional variance formula again we get

$$\mathrm{Var}(Y_{j,k}|X_k) = E\left[\mathrm{Var}(Y_{j,k}|Y_{j-1,k}, X_k)|X_k\right] + \mathrm{Var}\left[E(Y_{j,k}|Y_{j-1,k}, X_k)|X_k\right]. \tag{13}$$

We consider each term on the right side of equation (13) separately. It follows from equations (1), (4) and (6) that

$$E\left[\mathrm{Var}(Y_{j,k}|Y_{j-1,k}, X_k)|X_k\right] = E\left[\mathrm{Var}(V_{j,k}|Y_{j-1,k}, X_k)|X_k\right]$$
$$= 2^j\lambda(N_0 - X_k)(1 - \alpha_{j-1}). \tag{14}$$

Now consider the second term on the right side of (13). It follows by equation (5) that

$$\text{Var}\left[E(Y_{j,k}|Y_{j-1,k}, X_k)|X_k\right] = \text{Var}\left[2(1-\lambda)Y_{j-1,k} + \lambda 2^j(N_0 - X_k)|X_k\right]$$
$$= 4(1-\lambda)^2 \text{Var}(Y_{j-1,k}|X_k).$$

So combining the above gives

$$\text{Var}(Y_{j,k}|X_k) = 2^j \lambda (N_0 - X_k)(1 - \alpha_{j-1}) + 4(1-\lambda)^2 \text{Var}(Y_{j-1,k}|X_k).$$

If we define

$$\gamma_j = \frac{\text{Var}(Y_{j,k}|X_k)}{(N_0 - X_k)2^j}$$

we get the following recursion

$$\gamma_j = \lambda(1 - \alpha_{j-1}) + 2(1-\lambda)^2 \gamma_{j-1} = \lambda(1-\lambda)^{j-1} + 2(1-\lambda)^2 \gamma_{j-1}. \quad (15)$$

The solution to the above recursion follows again by equation (35) given in the Appendix and can be expressed as follows

$$\gamma_j = \lambda(1-\lambda)^{j-1} \frac{(2(1-\lambda))^j - 1}{1 - 2\lambda}. \quad (16)$$

For λ small we can approximate γ_j by

$$\gamma_j \approx \frac{\lambda}{1 - 2\lambda}(2^j - 1). \quad (17)$$

We now use the solution for γ_j given in (16) to calculate the $E(\text{Var}(Y_{j,k}|X_k))$ for any generation j in the kth cycle. We are particularly interested in the last generation of the kth cycle when $j = l$. Using equation (12) we get an explicit expression for $E(\text{Var}(M_k|X_k))$. Note that $E(X_k) = N_0 \beta_{k-1}$, hence

$$E(\text{Var}(M_k|X_k)) = E(\text{Var}(Y_k|X_k)) \equiv E(\text{Var}(Y_{l,k}|X_k))$$
$$= (N_0 - E(X_k))2^l \gamma_l = (1 - \beta_{k-1}))2^l \gamma_l N_0.$$

Using the above result we can return now to equation (11) and recalling the formula for β_k given by (9), we get

$$\text{Var}(M_k) = (1 - \beta_{k-1})2^l \gamma_l N_0 + 2^{2l}(1-\alpha_l)^2 N_0 \beta_{k-1} + (1-\alpha_l)^2 \text{Var}(M_{k-1})$$
$$= (1-\lambda)^{2l} \text{Var}(M_{k-1}) + 2^l N_0 (1-\lambda)^{l(k-1)}(1 - \beta_0)$$
$$\times \left(\gamma_l - 2^l(1-\lambda)^{2l}\right) + 2^{2l}(1-\lambda)^{2l} N_0. \quad (18)$$

The above recursion (18) follows the form of the recursion (34) given in the Appendix, so it follows from equation (35) with $a = (1-\lambda)^{2l}$, $b = (1-\lambda)^l$, $c = 2^l N_0 (1-\beta_0)\left(\gamma_l - 2^l(1-\lambda)^{2l}\right)$, and $d = 2^{2l}(1-\lambda)^{2l} N_0$, that

$$\mathrm{Var}(M_k) = 2^l N_0 (1-\lambda)^{l(k-1)} \frac{(1-\lambda)^{lk}-1}{(1-\lambda)^l-1} (\gamma_l - 2^l(1-\lambda)^{2l})(1-\beta_0)$$
$$+ 2^{2l} N_0 (1-\lambda)^{2l} \frac{(1-\lambda)^{2lk}-1}{(1-\lambda)^{2l}-1} \tag{19}$$

where γ_l is given by equation (16).

We now have an explicit formula for $\mathrm{Var}(A_k)$

$$\mathrm{Var}(A_k) = \frac{1}{2^{2l} N_0^2} \mathrm{Var}(M_k). \tag{20}$$

2.3. Covariance

In this section we show that the covariance of M_k with M_{k+n} is determined by the variance of M_k. It follows from equation (7) that, for $n \geq 1$

$$\begin{aligned}
E(M_{k+n} M_k) &= E(M_k(E(M_{k+n}|M_{k+n-1}\cdots M_k))) \\
&= E\left(M_k[(1-\alpha_l)M_{k+n-1} + 2^l \alpha_l N_0]\right) \\
&= (1-\alpha_l)E(M_{k+n-1}M_k) + 2^{2l}N_0^2 \alpha_l \beta_k.
\end{aligned}$$

Recall that the fraction of mutants at the end of the kth cycle, is defined by $A_k = M_k/(2^l N_0)$. Define

$$a_n = E\left(\frac{M_{k+n}M_k}{\left(2^l N_0\right)^2}\right) = E(A_{k+n}A_k)$$

then

$$a_n = (1-\alpha_l)a_{n-1} + \alpha_l \beta_k.$$

Again, a_n satisfies the usual recursion (34), where in this case $a_0 = E(A_k^2)$. Therefore,

$$a_n = (1-\alpha_l)^n E(A_k^2) + \beta_k \alpha_{ln}. \tag{21}$$

Recall that $E(A_k) = \beta_k$ and therefore

$$\begin{aligned}
\mathrm{Cov}(A_{k+n}, A_k) &= E(A_{k+n}A_k) - E(A_{k+n})E(A_k) \\
&= (1-\alpha_l)^n \mathrm{Var}(A_k) + (1-\alpha_l)^n \beta_k^2 + \alpha_{ln}\beta_k - \beta_k \beta_{n+k} \\
&= (1-\lambda)^{ln} \mathrm{Var}(A_k). \tag{22}
\end{aligned}$$

3. Statistical methods

In Section 2 we characterized the structure of a model which assumes exponential growth, random mutations and periodic bottlenecks. In this section we develop several methods for data analysis. We begin by developing methods for fitting data to the model. That is, we propose methods for estimating λ. We then develop ways to assess the adequacy of the model. In Section 4 we apply the methods developed in this section to the data of De Gelder et al. [5].

3.1. Process error

Random mutations and periodic bottlenecks induce a certain amount of fluctuation that is inherent to the process. That is, even if one could observe every single mutant and wild type without error at the end of each cycle, there would still be variability between different runs of the experiment just due to inherent randomness of the process. We call this *process error*. To consider an estimator for λ where only process error is accounted for, we make the assumption that A_k is observable. In Section 3.2 we use the results of this section to consider sampling error and process error together. This provides the realistic estimator used in the data analysis.

Below we propose a moment estimator for λ, the mutation rate, assuming that A_k is observable. We refer to this estimator as $\hat{\lambda}_{process}$, and give an explicit expression for its variance. As mentioned above, the estimator is based on the A_k's which are the fraction of mutants present at the end of the k cycles. Note that the initial fraction of mutants $A_0 \equiv \beta_0$. Denote by $\bar{A}_k$ the average (over r replicates of the experiment) of the fraction of mutations at the end of k cycles.

Note that

$$E(\bar{A}_k) = \beta_k \approx \beta_0 + \lambda l k.$$

If the experiment is replicated r times then it is easy to see that

$$\text{Cov}(\bar{A}_k, \bar{A}_j) = \frac{1}{r}\text{Cov}(A_k, A_j).$$

If samples are taken at cycles $k_1, k_2, \cdots, k_n$ then

$$E(\bar{A}_{k_1} + \bar{A}_{k_2} + \cdots + \bar{A}_{k_n}) \approx \lambda l(k_1 + k_2 + \cdots + k_n) + n\beta_0.$$

This suggests the following estimator

$$\hat{\lambda}_{process} = \frac{\bar{A}_{k_1} + \bar{A}_{k_2} + \cdots + \bar{A}_{k_n} - n\beta_0}{l(k_1 + k_2 + \cdots + k_n)}. \tag{23}$$

If we define $u = l(k_1 + k_2 + \cdots + k_n)$ then

$$\text{Var}(\hat{\lambda}_{process}) = \frac{1}{ru^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{Cov}(A_{k_i}, A_{k_j})$$

$$= \frac{1}{ru^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(1 - \lambda)^{l|k_j - k_i|}\text{Var}(A_{\min\{k_i, k_j\}}). \tag{24}$$

Using the formula for $\text{Var}(A_j)$ derived in (19) and (20) and substituting into (22) gives an explicit expression for the $\text{Var}(\hat{\lambda}_{process})$.

3.2. Observational error

Recall that M_k is the number of mutants present at the end of the kth cycle and A_k is the fraction of mutants at the end of the kth cycle. On average A_k follows a curve given by

$$E(A_k) = \beta_k = 1 - (1 - \beta_0)(1 - \lambda)^{lk}.$$

Note that the $\{A_k\}$ process is subject to two sources of variability, one is due to periodic bottlenecks and the other is due to random mutation. Recall that we denote these two sources of variability as process error. However, the A_k process cannot be directly observed. At the end of each cycle a sample is removed and used to start the next cycle. The rest is diluted and from this dilution the mutants are observed. Inherent in this procedure is another source of error, which we refer to as *observational error*.

Recall that at the end of the kth cycle a sample of size N_0 is taken and placed into a new flask. We referred to the number of mutants in this sample by X_{k+1}. The fraction of mutants remaining is A_k and the dilution process is equivalent to taking a sample of size D_k. Let S_k be the number of mutants observed in the sample of size D_k at the end of the kth cycle, then

$$S_k | A_k \sim \text{Bin}\,(D_k,\ A_k) \sim \text{POI}(D_k A_k)$$

and $S_1, S_2, \dots, S_k$ are conditionally independent given $A_1, A_2, \dots, A_k$. The variability associated with the $\{A_k\}$ process is called *process error* and the conditional distribution of S_k given A_k describes the *observational error*.

The purpose of this section is to model the observational error on top of the process error to derive the mean, variance and covariances for the observable random variables S_k and then to derive an estimator based on $\{S_k\}$.

We assume that at the beginning of the experiment, $k = 0$, the fraction of mutants in the population is given by $A_0 \equiv \beta_0$. We then take r samples each of size N_0, which form the r replicates of the experiment. From the definition of X_k, given in Section 2, X_1 represents the initial number of mutants at the beginning of the experiment for a particular run of the process. We assume that X_1 is Poisson distributed with mean $N_0 \beta_0$. To estimate the fraction of mutants in the initial population we further take more samples, r_0 say, each of size D_0 and analyze those samples. Let S_0 be the number of mutants observed at the beginning of the experiment for a particular sample of size D_0. Note that S_0 is distributed Poisson with mean $\beta_0 D_0$ and that S_0 is independent of X_1. For the data that follow (Section 4) $r = 6$ and $r_0 = 2$. That is, 6 replicates of the experiment are run, but only two samples are taken at the beginning.

Note that

$$E(S_k) = E(E(S_k | A_k)) = D_k E(A_k) = D_k \beta_k$$

and

$$\begin{aligned}
\text{Var}(S_k) &= E(\text{Var}(S_k | A_k)) + \text{Var}(E(S_k | A_k)) \\
&= D_k \beta_k + D_k^2 \text{Var}\, A_k.
\end{aligned} \tag{25}$$

Finally the covariance can be calculated for $j, k \geq 1$ by

$$\text{Cov}(S_k, S_j) = D_k D_j \text{Cov}(A_k, A_j). \tag{26}$$

So the fraction of observed mutants in the sample at cycles k and j have the same covariance as the fraction of mutants in the population. That is,

$$\text{Cov}\left(\frac{S_k}{D_k}, \frac{S_j}{D_j}\right) = \text{Cov}(A_k, A_j) \tag{27}$$

Again sampling at the end of cycles $k_1, k_2, \ldots, k_n$ and performing r runs of the experiment and noting that $E\left(\frac{\bar{S}_k}{D_k} \middle| \bar{A}_k\right) = \bar{A}_k$, we get

$$E\left(\frac{\bar{S}_{k_1}}{D_{k_1}} + \frac{\bar{S}_{k_2}}{D_{k_2}} + \cdots + \frac{\bar{S}_{k_n}}{D_{k_n}}\right) \approx n\beta_0 + \lambda l(k_1 + k_2 + \cdots + k_n).$$

Recall that S_0 is the number of initially observed mutants in a particular run of the experiment and $\bar{S}_0$ the average number of observed initial mutants averaged over r_0 samples and $E(\bar{S}_0/D_0) = \beta_0$. This suggests that in order to account for observational error, we use the following estimator for λ

$$\hat{\lambda} = \frac{\bar{S}_{k_1}/D_{k_1} + \bar{S}_{k_2}/D_{k_2} + \cdots + \bar{S}_{k_n}/D_{k_n} - n\bar{S}_0/D_0}{l(k_1 + k_2 + \cdots + k_n)}. \tag{28}$$

Note that $\bar{S}_0$ is independent of $\bar{X}_1$ which implies that $\bar{S}_{k_i}$ is independent of $\bar{S}_0$. Thus again let $u = l(k_1 + k_2 + \cdots + k_n)$

$$\text{Var}(\hat{\lambda}) = \frac{1}{ru^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(S_{k_i}/D_{k_i}, \ S_{k_j}/D_{k_j}) + \frac{n^2}{r_0 u^2} \text{Var}(S_0/D_0)$$

$$= \text{Var}(\hat{\lambda}_{process}) + \frac{1}{ru^2} \sum_{i=1}^{n} \beta_{k_i}/D_{k_i} + \frac{n^2}{r_0 u^2} \beta_0/D_0. \tag{29}$$

where the process error $\text{Var}(\hat{\lambda}_{process})$ is given by equation (24) which uses equations (19) and (20). We denote the observational error by

$$\text{Var}(\hat{\lambda}_{obs}) = \frac{1}{r_0 u^2} n^2 (\beta_0/D_0) + \frac{1}{ru^2} \sum_{i=1}^{n} \beta_{k_i}/D_{k_i}. \tag{30}$$

3.3. Process error versus observational error

The fraction of mutants (per replicate) observed as a result of sampling the process at the end of the kth cycle is given by S_k/D_k. While the fraction of mutants (per replicate) in the entire population at the end of the kth cycle is given by A_k. It follows from equation (9) that

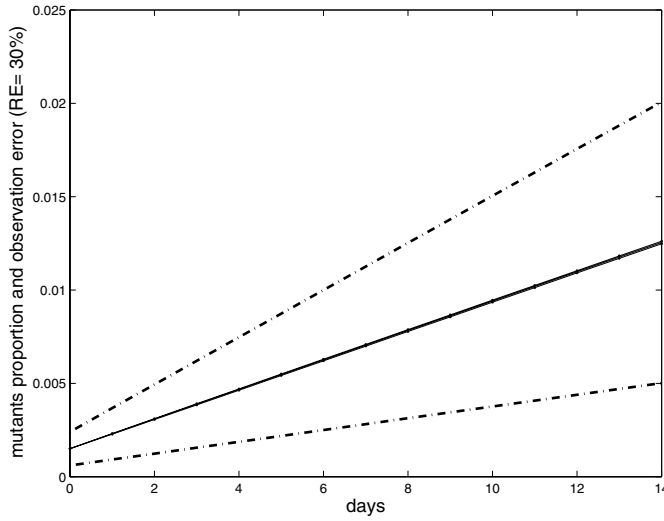$$E(S_k/D_k) = E(A_k) = 1 - (1 - \beta_0)(1 - \lambda)^{lk}.$$

Fig. 1. A plot of the number of days k versus the expected fraction of mutants $\beta_k = E(S_k/D_k) = E(A_k)$ (Equation (9)) plotted over 14 days. $\lambda = 9.93 \times 10^{-5}$ and $\beta_0 = 0.0015$. The dark trend line is actually three lines, where the solid lines above and below are $E(A_k) \pm 2\sqrt{\text{Var}(A_k)}$ demonstrating that the variability of the A_k process is quite small. The dashed lines (- - -) are $E(S_k/D_k) \pm 2\sqrt{\text{Var}(S_k/D_k)}$ where D_k was chosen so as to keep the relative error at 30%, $\text{RE} = \sqrt{\text{Var}(S_k/D_k)}/\beta_k = 0.3$.

It follows from equation (25) that

$$\text{Var}(S_k/D_k) = \beta_k/D_k + \text{Var}(A_k).$$

While we have not developed a full likelihood description of A_k we have derived exact solutions for the mean, variance and covariances for the $\{A_k\}$ process. We discovered that for a typical population size on the order of $N_0 = 10^7$, cycle length $l = 8$ and mutation rate $10^{-8} \leq \lambda \leq 10^{-4}$ the $\{A_k\}$ process was nearly deterministic and the error of the process, described by the variance and covariances was actually quite small. The graph in Figure 1 illustrates this point. The solid trend line is actually three lines. The middle line being $\beta_k = E(A_k)$ using equation (9) with $\lambda = 9.93 \times 10^{-5}$ and $\beta_0 = 0.0015$. The solid line $(-)$ above is $E(A_k) + 2\sqrt{\text{Var}(A_k)}$ and the line below is $E(A_k) - 2\sqrt{\text{Var}(A_k)}$. The dashed lines $(--)$ represent the error associated with the observed fraction of mutants S_k/D_k. The error for the observed fraction depends on the size of the sample D_k. A reasonable approach to deciding the sample size, D_k, is to consider the relative error. The relative error (RE), denoted here by η, is defined to be the standard deviation divided by the mean. In this case

$$\text{RE} \equiv \eta = \sqrt{\text{Var}(S_k/D_k)}/\beta_k.$$

To illustrate the difference between process error and sampling error we choose D_k so that the relative error is 0.3, which turns out to represent a sampling effort that is attainable for these types of experiments. (See the data analysis that follows

Table 1. The observed number of mutants for each of 6 replicates of the experiment and two initial samples. $s_{j,k}$ is the number of mutants sampled during the jth replicate of the kth cylce.

day k		samples						total
	$s_{1,0}$		$s_{2,0}$				D_k	
0	1		2				1000	3
	$s_{1,k}$	$s_{2,k}$	$s_{3,k}$	$s_{4,k}$	$s_{5,k}$	$s_{6,k}$		
14	0	1	1	1	0	0	24	3
28	1	1	0	2	0	2	52	6
42	2	0	1	2	0	1	52	6
49	4	0	0	5	2	2	52	13
56	0	5	1	0	2	4	52	12
63	1	3	4	4	3	4	52	19

in Section 4). Our main point in this section is to demonstrate that by modelling both the population variability (process error) and the sampling variability (observational error) we see that the observational error is several orders of magnitude larger than the process error.

4. Data

Consider the data given in Table 1. Samples were taken 7 days apart starting with day 14. The sample size taken on day 14 was 24, but realizing that this was too small, the sample size was increased to 52 for subsequent days. (The investigators did not have a prior view of how many sensitive mutants to expect and so the initial sampling plan was somewhat of a guess.) So the original data consisted of the analysis of days 14, 21, 28, 35, 42, 49, 56 and 63. Errors occurred on both days 21 and 35 and the data for those days had to be discarded. Two initial samples each of size 1000 were taken on day zero and a small fraction (0.0015) of mutants were observed.

The experiment was replicated 6 times. Each replicate consisted of an initial population of size $N_0 = 10^7$ individual cells. The population doubled for $l = 8$ generations per day to reach a population size of $2^8 \cdot 10^7$ followed by transferring 10^7 cells into fresh medium each day. At the end of cycle k listed in Table 1 a sample of size D_k is taken and the number of mutants observed is recorded. We first fit the data to the model by estimating λ using equation (28) and calculate the variance using (29). The results are summarized in Table 2. Table 3 gives the observed fraction of mutants compared to the expected fraction.

4.1. Assessing the goodness of fit

As indicated in the discussion in Section 3.3, we can neglect process error for the purposes of fitting data to the model and assume that the fraction of mutants in the

Table 2. An estimate of the mutation rate $\hat{\lambda}$ using equation (29) applied to the data in table 1. The variance of $\hat{\lambda}$ is given by equation (29) using equation (24) for the process error and (30) for the observational error.

Estimated rate	$\hat{\lambda}$	9.93×10^{-5}
Process error	$\text{Var}_{process}(\hat{\lambda})$	5.64×10^{-15}
Observational error	$\text{Var}_{obs}(\hat{\lambda})$	1.53×10^{-10}
Total Variance	$\text{Var}(\hat{\lambda}) = \text{Var}_{process}(\hat{\lambda}) + \text{Var}_{obs}(\hat{\lambda})$	1.53×10^{-10}
Standard error	$SE(\hat{\lambda}) = \sqrt{\text{Var}(\hat{\lambda})}$	1.24×10^{-5}

Table 3. The observed fraction of mutants calculated from table 1 versus the predicted fraction calculated using equation (9)

Cycle k	Observed fractions $\bar{S}_k/D_k$	Expected fraction $\hat{\beta}_k$
0	0.0015	0.0015
14	0.021	0 .013
28	0.019	0.024
42	0.019	0.035
49	0.042	0.040
56	0.038	0.046
63	0.061	0.052

population follows a deterministic growth pattern defined by β_k. Finally, since the data appear in an array, we denote by $S_{j,k}$ the number of mutants observed during the jth replicate at the kth cycle. With process error ignored, $S_{j,k}$ will follow the Poisson distribution with mean $D_k \beta_k$, and all of the observations are independent.

To test for model adequacy we consider the hypothesis that the model is correct, that is, the data are Poisson distributed with $E(S_{j,k}) = D_k \beta_k$ against the alternative that each $S_{j,k}$ is Poisson with a different mean $\mu_{j,k}$. We use the Poisson dispersion test statistic [11] given by

$$\chi^2 = \sum_{k \in K} \sum_{j=1}^{r_k} \left(s_{j,k} - D_k \hat{\beta}_k \right)^2 / D_k \hat{\beta}_k. \tag{31}$$

This is the familiar observed frequencies minus expected frequencies squared over the expected frequencies formula. The distribution of the test statistic given in (31) is asymptotically distributed chi-squared with $\sum_{k=1}^{n} r_k - 2$ degrees of freedom. Under the Poisson assumption the mean and variance are the same. In theory, if the data are overly dispersed, then the sample variances will be much larger than the means producing significant χ^2 values. According to the Poisson dispersion test displayed in Table 4, the data adequately fit the model. Care must be taken when interpreting the scatter about the trend line in Figure 2. Note that the residual

Table 4. The Chi-squared goodness of fit test using the data from Table 1 with $\hat{\beta}_k$ coming from table 3

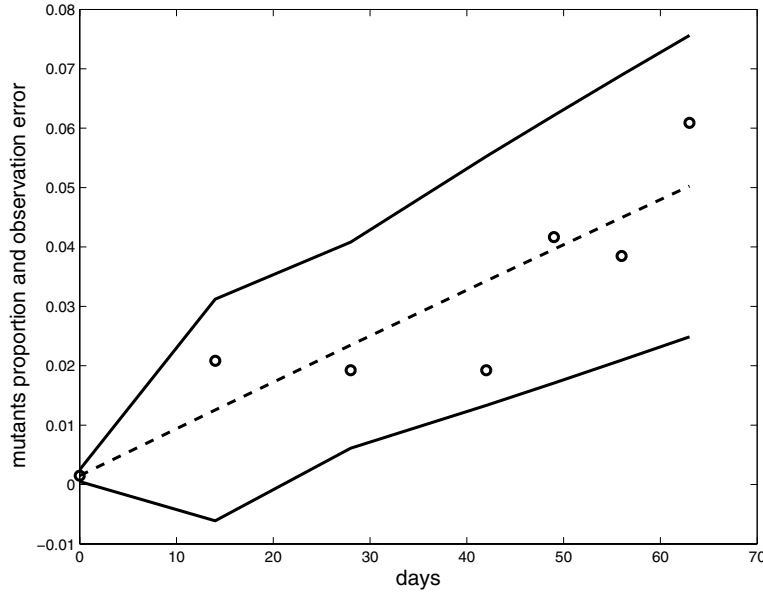|	
$\chi^2 = \sum_{k \in K} \sum_{j=1}^{r_k} \left(s_{j,k} - D_k \hat{\beta}_k \right)^2 / D_k \hat{\beta}_k$	
Degrees of Freedom	36
Value	36.5
P value	0.44

Fig. 2. Predicted (trend line) and observed (plotted points) of the fraction of mutants plotted over time (in days). The jagged lines represent two standard errors above and below the predicted values.

error is not normally distributed but is instead Poisson distributed. This means that the error depends on the cycle k, the sample size D_k and the mutation rate λ. The jagged lines in Figure 2 represent two standard errors above and below the trend line.

Note that with the exception of day 14 a constant sample size of 52 was maintained throughout. However, because of the Poisson nature of the observational error, the relative error = standard deviation/expected value of the predicted values decreases over time. Therefore, in order to keep the relative error constant, we recommend that in future studies larger samples be taken for the early days than those taken in the later days. That is $\sqrt{\mathrm{Var}(\hat{\beta}_k)} \approx \sqrt{lk\lambda/(rD_k)}$ and $E(\hat{\beta}_k) \approx lk\lambda$ and so the relative error η is

$$\eta = \frac{\sqrt{lk\lambda/(rD_k)}}{lk\lambda} = \frac{1}{\sqrt{lk\lambda r D_k}}.$$

Hence, an appropriate sampling strategy is to fix the relative error at η and choose D_k such that

$$D_k = \frac{1}{\eta^2 \lambda r l k}.$$

The sampling strategy described above was implemented in De Gelder et.al [5] which made the pattern of a slow increase in mutants much more apparent in that study.

5. Discussion

The mathematical model presented here can be used to predict the time needed to eliminate antibiotic resistance genes in an environment where the antibiotic is not present, thereby giving a preliminary assessment of the long term effects of eliminating the use of certain antibiotics. This is a considerable advantage over a purely descriptive approach to the data. For example, if we consider the mutation rate $\lambda = 9.93 \times 10^{-5}$ and $\beta_0 = 0.0015$ which were estimated from our data, then we can use equation (9) to predict the average time $t = lk$ needed to eliminate p percent of the antibiotic resistant genes in the population. That is

$$1 - (1 - \beta_0)(1 - \lambda)^t = p. \tag{32}$$

implying

$$t = \frac{\ln((1 - p)/(1 - \beta_0))}{\ln(1 - \lambda)}. \tag{33}$$

If $p = 0.95$ say, then this implies that $t = 32{,}794$ generations. With 8 generations a day this comes to 11.23 years. This means that if there is no selective advantage to the sensitive mutant, even if the antibiotic is not present, resistance genes will persist in the population for a long time (see Figure 3). If the mutant has even a slight selective advantage, this amount of time will decrease dramatically. So we can view our results using equation (33) as an upper bound. (See [5] for details.) The proposed model should be thought of as a null hypothesis against which other models can be tested.

Splitting the variance of the estimate $\hat{\lambda}$ into the two terms called process error and observational error is an idea borrowed from theoretical ecology (see [6]).
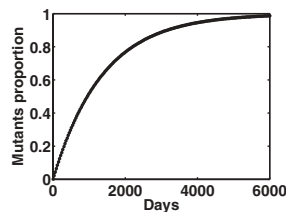

Fig. 3. Number of days k versus predicted fraction of mutants β_k.

However, this idea proves to be quite useful in experimental evolution. Because microbial systems have the unique feature that both ecological and genetic processes play an important role *on the same time scale*, exciting research opportunities exists on the boundary between the two disciplines; evolutionary biology and theoretical ecology, both of which have rich history in mathematical modelling. Several recent papers in theoretical ecology focus intensely on the issues surrounding *observational error* and *process error*. So it is important to understand how our results fit into this framework. Population dynamics theory further subdivides the process error into "demographic" and "environmental" stochasticity. *Demographic stochasticity* represents the variability due to random contributions of births, deaths and migrations of individuals in the population ([4]). *Environmental noise* represents the effect of external factors on the individuals of the population. In ecological terms, this noise represents environmentally driven fluctuations in the per capita population growth rate ([8]). These two types of error scale differently: demographic noise models predict that the variability in population size will decrease towards zero for large population sizes, whereas under environmental noise, the population fluctuates at large as well as at small densities [9].

The work of De Valpine and Hastings [6] and [7] focus mainly on environmental noise as the main source of process error. Here, we were interested in the random contributions of mutation and bottlenecks that affect population growth in a unique way that may be characterized as a special type of demographic noise. To our knowledge, compounding demographic stochasticity with observation error is still a fertile field, both in ecology and genetics.

By modelling both the process error and observational error we see that the pattern of variability in our data is explained almost entirely by the observational error. In fact, based on numerical calculations, we were able to demonstrate that the only situation where process error and observational error are comparable is when the mutation rate is extremely high (on the order of $\lambda = 10^{-3}$) and the bottleneck is extremely small ($N_0 = 1000$). So for nearly all experiments of this type the process error can be neglected. This greatly simplifies the data analysis. In fact, since we were able to demonstrate that process error can be ignored, then in the context of the observational error only model, our moment estimate for λ is indeed also the maximum likelihood estimate. These conclusions would not be possible without the mathematical model. The conclusions are facilitated by the fact that we have developed explicit formulae for the contributions of both observational error and processes error in the estimates (29). A maximum likelihood approach that relies on asymptotics to assess error can actually be less reliable than moment estimators when the asymptotic theory does not hold. (In fact, in the area of statistical genetics, where complex likelihood approaches are common, there is a growing body of literature [13],[2], that suggests that estimates based on simple summaries work as well as a full maximum likelihood approach, yet are more readily implementable in complex situations where full maximum likelihood is not computationally feasible.)

The models also suggest sensible and efficient experimental designs, which can aid the researcher in deciding how much data to collect and at what time periods should the process be observed. Finally, our approach could serve as a template for

future collaboration between experimentalists and theoreticians in designing and evaluating evolution experiments in a variety of settings.

References

1. Andersson, D.L., Levin, B.R.: The biological cost of antibiotic resistence. Current Opinion in Microbiology **2**, 489–493 (1999)
2. Beaumont, M.A., Zhang, W., Balding, D.D.: Approximate bayesian computation in population gentiecs. Genetics **162**, 2025–2035 (2002)
3. Brauer, F., Castillo-Chavez, C.: Mathematical models in population biology and epidemiology. (Texts in applied mathematics, Springer Verlag, New York, 2001), **40**, pp. 52
4. Cushing, J.M., Costantino, R.F., Dennis, B., Desharnais, R.A. Henson, S.M.: Chaos in Ecology. Amsterdam: Academic Press, 2002
5. De Gelder, L., Ponciano, J., Abdo, Z., Joyce, P., Forney, L., Top, E.: Combining mathematical models and statistical methods to understand and predict the dynamics of antibiotic sensitive mutants in a population of resistant bacteria during experimental evolution. Genetics **168**(3) in press, 2004
6. De Valpine, P., Hastings, A.: Fitting populations models incorporating process noise and observation error. Ecological Monographs **72**(1), 57–76 (2002)
7. Valpine, D.: Better inferences from population-dynamics experiments using Monte Carlo state-space likelihood methods. Ecology **84**, 3064–3077 (2003)
8. Ives, A.R., Dennis, B., Cottingham, K.L., Carpenter, S.R.: Estimating community stability and ecological interactions from time-series data. Ecological monographs **73**, 301–330 (2003)
9. Dennis, B., Munholland, P., Scott, M.: Estimation of growth and extinction parameters for endangered species. Ecological monographs **61**(2), 115–143 (1991)
10. Ewens, W.J.: The probability of survival of a new mutant in a fluctuating environment. Heredity **43**, 438–443 (1967)
11. Rice, J.A.: Mathematical statistics and data analysis. (Second edition, Duxbury Press, 1995), pp. 316
12. Ross, K.A.: Elementary analysis: The theory of calculus. (Undergraduate texts in mathematics, Springer Verlag, New York, 1991), pp. 69
13. Marjoram, P., Molitor, J., Plagnol, V., Tavar, S.: Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA **100**, 15324–15328 (2003)
14. Wahl, L.M.: Experimental evolution: analytical approaches and the need for a specific integrated theory. Commun. Theor. Biol. 189–204 (2001)
15. Wahl, L.M., Gerrish, P.J., Saika-Voivod, I.: Evaluating the impact of population bottlenecks in experimental evolution. Genetics **162**, 961–972 (2002)

6. Appendix 1- recursion

Throughout the paper we develop recursion equations that relate the process at time t to the process one unit of time into the past $t - 1$. The general equation is as follows.

$$z_t = a z_{t-1} + c b^{t-1} + d. \tag{34}$$

It follows by induction that

$$z_1 = az_0 + c + d$$
$$z_2 = az_1 + cb + d = a^2 z_0 + (a+b)c + (a+1)d$$
$$z_3 = az_2 + cb^2 + d = a^3 z_0 + (a^2 + ab + b^2)c + (a^2 + a + 1)d$$
$$z_4 = az_3 + cb^3 + d = a^4 z_0 + (a^3 + a^2 b + ab^2 + b^3)c + (a^3 + a^2 + a + 1)d$$

implying that

$$z_t = a^t z_0 + ca^{t-1} \sum_{i=0}^{t-1} \left(\frac{b}{a}\right)^i + d \sum_{i=0}^{t-1} a^i.$$

We now use a geometric series identity (See [12] page 69 equation 1) to get:

$$z_t = a^t z_0 + c \frac{a^t - b^t}{a - b} + d \frac{1 - a^t}{1 - a}. \tag{35}$$