The population biology of bacterial plasmids: a hidden
Markov model approach

José M. Ponciano\(^3\) Leen De Gelder\(^2\), Eva M. Top\(^2\) and Paul Joyce\(^1\)

\(^1\)Department of Mathematics (PO BOX 443051), and \(^2\) Department of Biological Sciences (PO Box 441103), University of Idaho, Moscow, 83844, USA, and \(^3\) Department of Ecology, (PO Box 173460) Montana State University, Bozeman, Montana 59717-3460
Running Head: A hidden Markov model of Plasmid stability and horizontal transfer

Keywords: Hidden-Markov models, plasmid stability, horizontal gene transfer

Corresponding author:
Dr. Paul Joyce
Department of Mathematics
University of Idaho
413 Brink Hall
Moscow, ID 83844-3051
USA
Phone or fax 1-208 885-6338 Email: joyce@uidaho.edu
Abstract

Horizontal plasmid transfer plays a key role in bacterial adaptation. In harsh environments, bacterial populations adapt by sampling genetic material from a horizontal gene pool through self-transmissible plasmids, and that allows persistence of these mobile genetic elements. In the absence of selection for plasmid-encoded traits it is not well understood if and how plasmids persist in bacterial communities. Here we present three models of the dynamics of plasmid persistence in the absence of selection. The models consider plasmid loss (segregation), plasmid cost, conjugative plasmid transfer and observation error. Also, we present a stochastic model in which the relative fitness of the plasmid free cells was modeled as a random variable affected by an environmental process using a hidden Markov model (HHM). Extensive simulations showed that the estimates from the proposed model are nearly unbiased. Likelihood Ratio Tests showed that the dynamics of plasmid persistence are strongly dependent on the host type. Accounting for stochasticity was necessary to explain 4 out of 7 time series data sets, thus confirming that plasmid persistence needs to be understood as a stochastic process. This work can be viewed as a conceptual starting point under which new plasmid persistence hypotheses can be tested.
INTRODUCTION

Comparative molecular phylogenies (GOGARTEN and TOWNSEND, 2005; SØRENSEN et al., 2005) and prospective, mathematical models coupled with experimental data sets have shown that Horizontal Gene Transfer (HGT), and in particular conjugative plasmid transfer (STEWART and LEVIN, 1977; LEVIN, 1980; SIMONSEN, 1991) is an important mechanism for bacterial adaptation. The search for adaptive traits within a large horizontal gene pool is often facilitated by plasmids, since these mobile genetic elements often carry genes that are advantageous to their hosts (e.g. genes required to exploit new carbon sources, antibiotic resistance genes, etc.). As these genetic functional units allow their host to occupy new ecological niches, then the persistence of plasmids in bacterial populations under local selective pressures can be understood (GOGARTEN and TOWNSEND, 2005; SØRENSEN et al., 2005). Perhaps most difficult to understand is the persistence of plasmids under non-selective conditions, that is, when the plasmid’s genetic material does not confer any advantage to its hosts. This scenario is the focus of our research. Understanding this scenario has many important applications. For example, the loss or persistence of plasmids carrying antibiotic resistance genes, when the selective pressure of the antibiotic is removed from the population, has major human health implications.

In the absence of selection, a plasmid may be maintained if a certain balance exists between three key factors (STEWART and LEVIN, 1977; SIMONSEN, 1991; FRETER et al., 1983; LENSKI and BOUMA, 1994). These factors are (i) plasmid loss by segregation during bacterial replication, (ii) the burden or fitness cost associated with carrying and/or expressing the extra piece of genetic material and (iii) plasmid transmission via conjugation. In other words, for a plasmid to persist, horizontal transmission must compensate for segregational loss and fitness cost of the plasmid.
The framework under which most of this knowledge about plasmids persistence has been built is deterministic differential equation modeling. Yet, the main biological mechanisms and principles under which evolution and adaptation are theoretically understood are essentially stochastic (Novozhilov et al., 2005).

Adequately connecting deterministic and stochastic population models to real time series data via statistical time series methods is an important yet difficult task (Cushing et al., 2002; Dennis et al., 2006). The statistical framework under which these analyses are done while considering both, process and observation uncertainty is well formalized and known as State-Space Modeling (SSM) (Carlin et al., 1992; Meyer and Millar, 1999; Dennis et al., 2006). One important class of SSM’s is the Hidden Markov Models. However, much work remains to be done in order to assess the reliability and accuracy of parameter estimates and the inferences made from them. Finally, we expect that a careful time series analysis might lead to a better understanding of plasmid persistence in bacterial populations.

The objective of our current work is two-fold: first, we formulate, fit and later compare deterministic and stochastic models to time series data on plasmid persistence in seven bacterial strains. In doing so we consider taking into account both process and observation uncertainty using analytical methods for SSM. Second, we show via extensive simulations that the statistical procedures implemented here provide the means to reliably make biological inferences from plasmid instability time series data (De Gelder et al., 2006). We briefly explain the stability experiment methods used to obtain time series data on plasmid instability and refer to De Gelder et al. (2006) for technical details on the experimental procedures. We also present a mathematical modeling section in which we state and develop each one of the deterministic and stochastic models that include segregation, selection and horizontal transfer processes used throughout the paper. In an online supplemental data file, the statistical methodology used to confront the models with the time series
data and evaluate their performance is explained in detail. Finally, we discuss the implications, significance and weaknesses of our findings in light of the current studies in the area.

THEORETICAL BACKGROUND

Mathematical modeling

Segregation and Selection (SS) model: Our simplest dynamic model summarizing the growth dynamics of the fraction of plasmid-free cells in the experiments described below (see Materials and Methods) is a simple system of difference equations where it is assumed that at any generation, the abundance of the plasmid-free cells \( m \) increases due to 1) plasmid segregation from the wild type cells \( n \) at a frequency \( \lambda \) and 2) growth of segregants at a rate \( 2^{1+\sigma} \), where \( \sigma \) represents the selection coefficient:

\[
\begin{align*}
    n_t &= 2 (1 - \lambda) n_{t-1}, \\
    m_t &= 2^{1+\sigma} m_{t-1} + 2\lambda n_{t-1}
\end{align*}
\]

and the fraction of plasmid-free cells \( x_t \) is given by

\[
x_t = \frac{m_t}{m_t + n_t}. \tag{2}
\]

This deterministic model was developed by De Gelder et al. (2004) and assumes that there is no conjugational transfer from plasmid carrying cells to segregants.

Throughout this paper, the Segregation and Selection SS model serves as our null hypothesis against which more complex models and growth behaviours were tested. The solution to the SS
model is presented in DE GELDER et al. (2004) (Appendix 1). This fraction of plasmid free cells grows logistically starting very close to 0 and approaching 1 as $t \to \infty$. We note that JOYCE et al. (2005) also showed that it can be assumed that the deterministic growth of plasmid-free cells is basically unaffected by the daily bottlenecks described below (Materials and Methods).

**Horizontal Transfer (HT) model:** A horizontal transfer (HT) model can be generated from eq. 1 by incorporating a term that accounts for the fraction of plasmid free cells that re-acquire the plasmid through conjugative transfer. The typical approach to model conjugation (LEVIN, 1980; SIMONSEN, 1991; STEWART and LEVIN, 1977) is to use the mass-action principle, where the rate at which conjugation occurs depends linearly on the concentration of plasmid-free and plasmid-carrying cells. Using the mass-action principle, the horizontal transfer model where $\gamma$ represents a constant conjugative transfer frequency would be written as follows:

$$
\begin{align*}
  n_t &= 2 \left( 1 - \lambda \right) n_{t-1} + 2^{1+\sigma} \gamma m_{t-1} n_{t-1}, \\
  m_t &= 2^{1+\sigma} \left( 1 - \gamma n_{t-1} \right) m_{t-1} + 2\lambda n_{t-1}.
\end{align*}
$$

Here we model conjugation by relaxing the mass-action principle with the following system of equations:

$$
\begin{align*}
  n_t &= 2 \left( 1 - \lambda \right) n_{t-1} + 2^{1+\sigma} m_{t-1} \gamma \frac{1 - x_{t-1}}{\theta + (1 - x_{t-1})}, \\
  m_t &= \left( 1 - \gamma \frac{1 - x_{t-1}}{\theta + (1 - x_{t-1})} \right) 2^{1+\sigma} m_{t-1} + 2\lambda n_{t-1},
\end{align*}
$$

where $x_t$ is the fraction of segregants at time $t$ (eq.2), $\gamma$ is an asymptotic maximum conjugation frequency during a time interval and $\theta$ represents the fraction of the plasmid carrying cells at which the frequency of conjugations is half its maximum. The second term in eq. (4) assumes that the transfer process works as in an enzymatic reaction, where enzyme and substrate are the plasmid
carrying and plasmid free cells respectively (ANDRUP and ANDERSEN, 1999).

The system of equations (4 and 5) can be readily reduced to a single model equation for the fraction \( x_t \) of plasmid free cells at time \( t \):

\[
x_t = \frac{\left(1 - \frac{(1-x_{t-1}) \gamma}{\theta + (1-x_{t-1})}\right)}{2^{1+\sigma} x_{t-1} + 2 (1 - x_{t-1})} 2^{1+\sigma} x_{t-1} + 2 (1 - x_{t-1}).
\]

(6)

A local stability analysis (KOT, 2001) shows that this model has 3 equilibrium solutions, the simplest of which is \( x^*_1 = 1 \). The other two solutions \( x^*_2 \) and \( x^*_3 \) are \( \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \), where \( A = (2^\sigma - 1) \), \( B = ((\theta + 1)(2^\sigma - 1) + \lambda + 2^\sigma \gamma) \) and \( C = -\lambda (\theta + 1) \). The equilibrium solution \( x^* = 1 \) is stable as long as the plasmid burden \( \sigma \) is big enough to satisfy

\[
\frac{\gamma}{\theta} \geq 1 - \frac{1 - \lambda}{2^\sigma},
\]

(7)

provided \( \theta > \gamma \). As it is illustrated in Fig. 1, when the solution \( x^* = 1 \) is stable, the plasmid-carrying cells are guaranteed to be lost from the population. When \( \sigma \) is too small so that inequality (7) is not satisfied, the equilibrium solution \( x^* = 1 \) becomes unstable and one of the other 2 solutions of the model, the one inside the interval \((0, 1)\), \((x^*_2)\), becomes stable. This new equilibrium solution basically predicts that plasmids will never be lost from the population.

Note that inequality (7) is readily interpretable: Since the fraction \( \gamma/\theta \) is a measure of the intensity of the transfer frequency, this formula basically states that a high loss of plasmids due to segregation must be balanced by a high transfer frequency for the plasmids to persist in the population. Likewise, a high cost would decrease the fraction \( (1 - \lambda)/2^\sigma \) and hence increase the size of the transfer frequency threshold needed to guarantee the persistence of plasmids in the population. This property of the HT model is analogous to the results of STEWART and LEVIN (1977), who found that in a chemostat, plasmids can be maintained only when the cell density and conjuga-
tional transfer rate constant are large enough for the transmission of the plasmid to overcome its loss through segregation and the selection against plasmid-carrying bacteria. However, we note that the HT model also shows that, when the frequency-dependence in transfer is strong, i.e., at higher values of $\theta$, the persistence threshold $\gamma/\theta$ becomes smaller and the loss of plasmids through segregation and selection can be more easily overcome. If the loss by segregation and the frequency of frequency-dependent transfer are kept fixed, a reduction in the size of the cost $\sigma$ down to a critical value (see inequality 14) allows the invasion of plasmids in the population. This behaviour is visualized by plotting both, solution trajectories and the growth rate of the fraction of plasmid-free cells at different values of $\sigma$ (see Fig. 1). As the plasmid burden $\sigma$ decreases, the growth rate ceases to be parabolic in shape (as in a typical logistic growth curve), and adopts a cubic-like form with a root inside the interval $(0, 1)$ which is a stable equilibrium. That is, it is the point where the long-term fraction of plasmid-free bacteria stagnates, thus predicting a long-run coexistence between plasmid-free and plasmid-carrying bacteria.

The Variable Selection (VS) model: The dynamic equations explained so far assume that during an entire plasmid stability experiment the growth of the fraction of plasmid free cells follows essentially a deterministic pattern. That is, all the deviations from the deterministic smooth growth equations 1, 2 and 4, 5 that appear in the data are assumed to be pure random sampling error. However, theory and experiments (De Visser and Rozen, 2005) suggest that during a 600 generations experiment, the occurrence of compensatory mutations and/or a variable host-dependent plasmid burden would dramatically alter the plasmid loss dynamics. Periods of overall heavy plasmid loss would be followed by periods where the relative frequency of segregants remains almost unchanged. Therefore, as an alternative hypothesis, we propose a stochastic formulation of the segregants growth dynamics that assumes that at each time step, the burden is a value drawn from a continuous probability distribution. By doing so, the fraction of plasmid-free cells grow stochasti-
cally. This variable selection (VS) model is then recognized as a model with environmental stochasticity (Lewontin and Cohen, 1969; Keiding, 1975; Cushing et al., 2002). Hence, to specify our variable selection model we let the selection coefficient be drawn at each time step from a Normal distribution (Lewontin and Cohen, 1969; Keiding, 1975) $S_t$ with mean $\mu$ and variance $\tau^2$. Then, the VS model can be written as:

$$N_t = 2(1 - \lambda)N_{t-1},$$  

(8)

$$M_t = 2^{1+S_t}M_{t-1} + 2\lambda N_{t-1},$$  

(9)

$$X_t = \frac{M_t}{M_t + N_t} = \frac{X_{t-1}2^{1+S_t} + 2\lambda (1 - X_{t-1})}{X_{t-1}2^{1+S_t} + 2(1 - X_{t-1})},$$  

(10)

where capital letters denote random variables and lower case letters from here on, will be used to denote realizations of the random variables involved. Then, $X_t$ becomes a Markov process whose transition probability density function (pdf) is found to be (Appendix):

$$f(x_t|x_{t-1})(x_t) = \frac{(1 - \lambda)}{\ln2 (x_t - \lambda) (1 - x_t) \sqrt{2\pi\tau^2}} \exp\left\{\frac{-(h_t - \mu)^2}{2\tau^2}\right\},$$  

(11)

where

$$h_t = \frac{\ln[(1 - x_{t-1})(x_t - \lambda)] - \ln[x_{t-1}(1 - x_t)]}{\ln2}$$

and $x_t > \lambda$. The transition probability density function (eq. 11) provides us the means to characterize the behaviour of $X_t$ via analytical and simulation results. Also, this pdf provides the proper link between parameter estimation and the hypothesized biological process.
MATERIALS AND METHODS

Stability Experiments and Time Series Data

To investigate whether the ability of a broad-host-range plasmid to be stably maintained in a bacterial population varies between different hosts, plasmid stability experiments were performed with different bacterial strains carrying the same plasmid pB10, as described by DE GELDER et al. (2006, 2005). The experimental approach is briefly summarized below. The 64.5 Kb plasmid pB10 (SCHLÜTER et al., 2003), isolated from the bacterial community of a wastewater treatment plant (DRÖGE et al., 2000), is a self-transmissible, BHR IncP-1/β plasmid that mediates resistance to the antibiotics tetracycline (Tc), streptomycin (Sm), amoxicillin, and sulfonamide, and to HgCl₂. For each strain, stability experiments were performed in triplicate, starting from three separate colonies, which were each inoculated in 5 mL LB with the appropriate concentrations of Tc and Sm to select for pB10. After incubation for 24 hrs, these cultures were washed to remove the antibiotics by spinning down 1 mL culture and re-suspending the pellet in 1 mL saline. From these cell suspensions, 4.88 µL was transferred to 5 mL of LB such that the cells went through 10 generations of growth during each 24 hrs growth cycle. These freshly inoculated cultures constituted time point zero. After they were plated on LB plates and an aliquot was archived at −80°C, they were incubated on a rotary shaker for 24hrs. As of then, 4.88 µL of the full grown cultures was transferred each 24 hrs to fresh 5 mL LB. These were the daily bottlenecks mentioned in the Mathematical modeling section. At various time points, the cultures were diluted and plated on LB plates. Determining the fraction of plasmid-free cells in the population was done by replica picking 50 colonies per culture at random from the LB plates onto LB-Tc, LB-Sm and LB plates and scoring Tc−Sm− colonies. Random Tc−Sm− isolates of each strain were confirmed as true segregants through comparison of their genomic fingerprints (BOX PCR) (RADEMAKER et al., 1997) with those of the original
strains from which they were derived, and through gel electrophoresis of plasmid extracts (Kado and Liu, 1981; Top et al., 1990). The time series data thus obtained were analyzed using three population dynamics models.

Statistical Analysis

Deterministic modeling (sampling error with no environmental noise): A sample of size $d_{tj}$ colonies were taken at random from a replicated culture $j$, $j = 1, 2, \ldots, r$ at day $t$. Each individual has a probability $x_t$ of being a segregant and $(1 - x_t)$ of being a wild type, where $x_t$ is the model predicted fraction of segregants at generation $t$. This defines a binomial sampling process with $d_{tj}$ trials and the fraction of segregants $x_t$ changes deterministically in time according to the dynamic equations 1, 2 and 4-5. Then, the number of plasmid-free cells observed in culture $j$ at day $t$, denoted $Y_{tj}$, is a binomial random variable, and

$$P(Y_{tj} = y_{tj}) = \binom{d_{tj}}{y_{tj}} x_t^{y_{tj}} (1 - x_t)^{d_{tj} - y_{tj}}.$$  \hfill (12)

The two deterministic models, SS (eqs. 1-2) and HT (eqs. 4 and 5), together with the sampling process defined by eq. 12, account for the deviations of the observations from the predicted growth pattern. Therefore, the next step in the model building process was to rigorously connect the data with the model using the above binomial sampling process. This was done using the method of maximum likelihood (Rice, 1995) in the case of the SS and HT models, as in De Gelder et al. (2004). For these two models, a Likelihood Ratio Test (LRT) was carried out, where under the null hypothesis the data were binomially distributed according to eq. 12, and under the alternative hypothesis the data were still binomially distributed but with a different mean $p_{tj}$ unrelated to the model. Under the null hypothesis, the estimated mean trend is $\hat{E}[Y_{tj}] = d_{tj}\hat{x}_t$, where $\hat{x}_t$ is the
model predicted fraction of segregants at generation $t$ using the ML estimates for the parameters $\sigma$, $\lambda$, $\gamma$, $\theta$ and $x_0$. Under the alternative hypothesis, the estimated mean trend is $\hat{E}[Y_{tj}] = d_{tj}\hat{p}_{tj}$, where $\hat{p}_{tj}$ is just the empirical estimate of the segregants’ proportion at replicate $j$ and generation $t$. After taking the natural logarithm and multiplying by $-2$, the LRT $\Lambda$ reduces to

$$-2\ln \Lambda = \sum_{t=1}^{q} \sum_{j=1}^{r} y_{tj} \left[ \ln (\hat{x}_t) - \ln (\hat{p}_{tj}) \right] + (d_{tj} - y_{tj}) \left[ \ln (1 - \hat{x}_{tj}) - \ln (1 - \hat{p}_{tj}) \right].$$  \hspace{1cm} (13)

To approximate the distribution of $-2\ln \Lambda$ we used Parametric Bootstrap Likelihood Ratio Tests (EFRON and TIBSHIRANI, 1993) and proceeded as in DE GELDER et al. (2004). While the Chi-squared distribution is often used to approximate the distribution of $-2\ln \Lambda$, it is only valid if the sample sizes are large enough. The asymptotic theory is known to be unreliable for small sample sizes, which is exactly the case in our data during the early time periods. Our approach has the advantage that it does not rely on asymptotic theory and the accuracy of the approximation is determined by the number of simulations, which we can completely control. A fuller discussion of this issue appears on page 1137 of (DE GELDER et al., 2004).

**Stochastic modeling (sampling error plus environmental noise):** To carry out parameter estimation for the stochastic model in eqs.8-10, we reformulate it as a SSM or HMM. For each replicated time series of the process, we denote its realizations as $X_{t,j}$, $t = 1, 2, \ldots, q$ and $j = 1, 2, 3$. The stochastic growth equation that governs each of the unobserved $X_{t,j}$ random realizations is therefore given by (see eq. 10):

$$X_{t,j} = \frac{X_{t-1,j}2^{1+S_t} + 2\lambda (1 - X_{t-1,j})}{X_{t-1,j}2^{1+S_t,j} + 2 (1 - X_{t-1,j})}.$$  \hspace{1cm} (14)

Given an (unobservable, or “hidden”) replicated random path $X_j = [X_{1,j}, X_{2,j}, \ldots, X_{q,j}]'$ that
starts from a fixed (unknown) proportion of plasmid free cells \( x_{0,j} \), each observation in the vector of recorded plasmid-free colonies counts \( Y_j = [Y_{0,j}, Y_{1,j}, Y_{2,j}, \ldots, Y_{q,j}]' \) is assumed to be drawn from a binomial probability distribution with samples sizes \( d_j = [d_{0,j}, d_{1,j}, \ldots, d_{q,j}]' \) and probabilities vector \( \begin{bmatrix} x_{0,j} \\ X_j \end{bmatrix} \). That is,

\[
(Y_j | X_j) \sim \text{Binom} \left( d_j, \begin{bmatrix} x_{0,j} \\ X_j \end{bmatrix} \right).
\]

Notice that eq. 15 differs from eq. 12 in an important way: in eq. 15, except for \( x_{0,j} \), the probabilities used to evaluate the binomial sampling distribution are themselves random variables, and not fixed quantities as in eq. 12. Eq. 14 is called the “state equation” and eq. 15 is called the “observation equation”. Together, equations 14 and 15 constitute the State-Space Model formulation (DENNIS et al., 2006) of the segregation-selection problem.

We used a Monte Carlo technique to retrieve the maximized likelihood scores and carry out model selection. The maximized likelihood scores were computed by evaluating the likelihood function of the observed time series at the ML estimates. Let \( \varphi = [\lambda, \mu, \tau^2, x_{0,1}, x_{0,2}, \ldots, x_{0,r}]' \) be the model parameters of interest for \( r \) replicated time series of the process. Then, the likelihood function of the observed time series for these \( r \) replicates, denoted by \( L(\varphi) \) is:

\[
L(\varphi) = P( [Y_1, Y_2, \ldots, Y_r] | \varphi) = \prod_{j=1}^{r} P(Y_j | \varphi)
\]

\[
= \prod_{j=1}^{r} \int P(Y_j | \varphi, X_j) P(X_j | \varphi) dX_j.
\]

The integral in eq 16 cannot be computed directly and was approximated using Importance Sampling as in GEORGE and THOMPSON (2002). Before doing so, the ML estimates of the model parameters were computed using Gibbs sampling CARLIN et al. (1992). However, we note that the
methods used here generalize their approach, as CARLIN et al. (1992) only treat the case where the state and observation equations have additive state and observation errors into their non-linear and non-Gaussian models. CARLIN et al. (1992) formulated their methods using the Bayesian paradigm of statistical estimation. In DE GELDER et al. (2006) however, we adopted the frequentist perspective to find the SS and HT model parameter estimates. So, in order to make the results of the SS model comparable to those of DE GELDER et al. (2006, 2004), we adopted the strategy of GEORGE and THOMPSON (2002) and used the Bayesian methodology of CARLIN et al. (1992) just as a numerical device to compute the ML estimates of the VS model parameters: by adopting uniform priors for all the parameters, the posterior modes of the parameters of interest are equivalent to the ML estimates.

In an online supplemental data file, we present first the Gibbs Sampling Algorithm for a single time series of plasmid stability, with no replicas. Then we extend this procedure to the case were a number $r$ of replicated time series are recorded. We also present there the details of an extensive simulation experiment done to evaluate the performance of the parameter estimation method, using the concept of Bootstrap (EFRON and TIBSHIRANI, 1993). Also, in the online supplemental data we explain in detail the calculation of the likelihood equation 16. For a very good description of the Gibbs sampling algorithm, and why it works, we refer the reader to CASELLA and GEORGE (1992).

**RESULTS**

**Results of fitting the models to the plasmid stability data via LRTs:** The different mathematical models were fitted to time course data that represent the stability of plasmid pB10 in seven
different hosts (2). In the top panels of Fig. 2 we plotted the model predicted fraction of segre-
gants’ growth along with the replicated data, for strains H2, R28 under the SS model and for the
strain P21 under the HT model. The results of the Segregation-Selection (SS) model fitting to the
plasmid stability time series, (Tables 1 , upper panel of Fig. 2) showed that for the strains Pseu-
domonas putida H2 and Pseudomonas koreensis R28 only two simple factors, plasmid cost and the
segregation frequency, were necessary to explain the segregant fraction time series and most of the
variation in the data, as confirmed by the absolute goodness of fit p-values. For Stenotrophomonas
malthophiilia P21, fitting the SS model was not sufficient and it was necessary (Table 1, Fig. 2) to
include frequency-dependent plasmid transfer in the deterministic model equations. For the strains
Pseudomonas plecoglossicida P18, Pseudomonas veronii S34, Ochrobactrum tritici S55 and Ens-
sifer adhaerens S96 both deterministic models failed to fit the data as they showed more variability
than what could have arisen from simple random sampling off the deterministic trajectories (Fig. 2,
eqs. 1, 4, 5). For these strains the Variable-Selection (VS) model provided an adequate explanation
of the data.

Note that if two models are nested, the ML score of the more complex model has to be the
biggest. In Table 1, after rounding, the ML score of both the H2 and R28 models is the same. Note
that, in the HT model, when $\gamma = 0$, the HT model is identical to the SS model. In De Gelder et
al, 2006, we show that the ML estimates of $\gamma$ for the strains H2, R28 are $2.11E - 09, 1.99E - 30$
respectively. Thus, the ML estimate under the HT model for these strains basically says that $\gamma \approx 0$,
and therefore the HT model converges to the SS model, and hence, the ML score under both
models has to be very close. For the S55 strain the estimate of $\gamma$ is $1.348896e - 05$, and, although
this number is also close to 0, we note that neither the SS nor the HT model fit the data, and that
these models were rejected according to the Absolute Goodness of fit test.
Results of predicted growth patterns: The SS and HT models predict a smooth trajectory and the deviations of the data from those trajectories were assumed to be due to sampling noise (see Fig 2. The SS and HT model have the advantage of providing a very simple explanation of the data, yet they only explain a small number of growth patterns. In the bottom panels of Fig. 2 we plotted the recorded data for each replicated stability experiment of the strain P18, along with the posterior mode of the estimated trajectory under the VS model. Following the HMM assumptions, the process of growth is itself stochastic, so that the recorded data set would not deviate from a smooth unknown trajectory, but instead from a variable growth pattern. Thus, the bottom panels of Fig. 2 should not be interpreted as a typical “observed vs. predicted” plot. A major advantage of the VS model is that it can explain much more complex growth patterns. Its one disadvantage is that it does not provide a precise explanation as to what might cause the variation in selection over time. How well the VS model approximates a more mechanistic model is a topic for further research.

Results of the stability analysis under the Horizontal Transfer (HT) model: As explained before, the plots in Fig. 1 reveal that there exist particular combinations of parameter values for which a long-term coexistence of plasmid-free and plasmid-carrying bacteria is predicted. Fig. 3 depicts explicitly the regions of the parameter space for which this coexistence occurs. Take for instance the first subplot in Fig. 3, where the plasmid burden $\sigma$ is located in the abscissa and the segregation frequency $\lambda$ is in the ordinate. The picture inside these axes was produced as follows: First, note that solving for $\sigma$ in inequality (7), it follows that for fixed (biologically meaningful) values of $\theta$, $\lambda$, and $\gamma$, $x^*_1 = 1$ is stable (i.e. the plasmid will always go extinct) whenever

$$\sigma > \frac{\ln(\theta(1 - \lambda)) - \ln(\theta - \gamma)}{\ln 2}.$$ 

Then, for each combination of $\sigma$ and $\lambda$ the right hand side of the above inequality was evaluated
using the ML estimates of $\theta$ and $\gamma$. Then, a dot was simply plotted if the above inequality was satisfied. By repeating the same procedure for many combinations of $\sigma$ and $\lambda$ in the quadrant, a shaded area enclosing the parameter region where the point $x^* = 1$ is stable appeared. The white area denotes the set of parameter values for which $x^* = 1$ is an unstable equilibrium. As mentioned before, when this occurs, a new stable equilibrium $x^*$ appears and the fraction of plasmid free bacteria never converges to 1, but to a point in $(0, 1)$. In other words, the plasmid-carrying bacteria remain in the population at a certain fraction.

The ML estimates for the data set of the strain P21 were superimposed on these stability boundaries plots, and fell inside the region where no long-run coexistence between plasmid-free bacteria and plasmid-carrying bacteria is predicted. To account for sampling uncertainty we added the approximate 95% parametric bootstrap confidence cloud around the ML estimates (DE GELDER et al., 2004; DENNIS et al., 1995; HILBORN and MANGEL, 1997). The joint confidence interval can be interpreted as an inverted likelihood ratio test (RICE, 1995): the points inside that interval denote the plausible parameter values under which the data at hand could have arisen, given the specified model. That (blue) cloud of points in Fig. 3 lies within both the shaded and white areas. Therefore, due to this observation it cannot be concluded that the plasmids will always be completely lost in the population in repeated experiments, using the strain P21.

**Assessing the reliability and accuracy of the Variable-Selection (VS) model via simulations:**

In order to assess the performance of the HMM formulation of the VS model, an extensive simulation study was performed. Data was simulated under the assumptions of the VS model using the parameter values $x_0 = 0.0066$, $\mu = 0.01166667$ per generation, $\tau^2 = 0.1583333$ per generation, and $\lambda = 1.533333 \times 10^{-05}$ per generation. Figures 4, 5, and 6 were all based on simulated data sets.

Assessing how well the HMM formulation of the VS model predicted the underlying frequency
dynamics was done as follows: A single simulated trajectory of the growth of the frequency of segregants over time under the VS model was produced and plotted in Fig.4 with triangles. Then, a random sample under the binomial model eqs. 14 and 15 with 50 trials and probability equal to the size of the segregants fraction at each time step was taken. Those samples were then treated as an observed data set and plotted with open circles in Fig.4. The samples were then analyzed using the Gibbs sampling algorithm (see the online supplemental data file). The results were plotted with solid circles in Fig. 4. As Fig. 4 suggests, the VS model does well at predicting the underlying frequency dynamics.

The parameter estimation method (see the online supplemental data file) for the HMM formulation of the segregants’ population dynamics provides reliable estimates of the plasmid cost and plasmid segregation frequency. 1000 data sets containing each three replicated time series of length 22 were simulated using the VS model. For each replicated realization of the process within each one of the “true” 1000 data sets, one random sample using the binomial observation error model was generated, thus obtaining 3000 simulated “observed” time series. The simulated “observed” time series were used to estimate 1) the posterior distributions of the model parameters and 2) the posterior distributions of the “true” simulated trajectories. Thus, 1000 posterior distributions were obtained for each of the model parameters $\mu$, $\tau^2$ and $\lambda$.

The posterior mode of the model parameters posterior distributions were retrieved each time and were plotted in a histogram and compared to the true values used to generate the simulations in Fig.5. Because uniform priors were used throughout, the posterior modes of the model parameters are identical to the ML estimates of the model. Therefore, the variability around the mode of the posterior modes histograms is an estimate of the variance of the ML estimates. In these histograms, the 3 modes of the ML estimates of the model parameters $\mu$, $\tau^2$ and $\lambda$ lied very close to the true values, thus showing that the HMM parameter estimation framework provides nearly
unbiased estimates of the model parameters. Also, the posterior distributions of the estimates of each point in the simulated “true” trajectories were computed (thus obtaining a total of $3000 \times 22$ posterior distributions). The posterior mode of each of those $3000 \times 22$ posterior distributions was computed and the difference between those modes and the “true” unknown trajectory point values were computed and plotted as histograms in Fig.6. The mode of those histograms lies in general right above 0, implying that the mode of the ML estimates is usually an unbiased estimate. The mode of the ML estimate of $x_0$ was around 0.050 and thus conveyed a biased estimate of the true original value $x_0 = 0.0066$. A closer inspection of the Monte Carlo Markov Chains generated using Gibbs sampling and simulations not presented here showed that the Gibbs algorithm had trouble converging for $x_0$.

**DISCUSSION**

The patterns generated by the dynamics of plasmid loss in seven bacterial strains were effectively explained by our deterministic and stochastic population dynamics models. The Segregation Selection and Horizontal Transfer models with added observation noise together explained 3 out of 7 data sets. Non-negligible plasmid transfer was a necessary component in our models to explain the data set for the P21 strain. The need of including process noise was evidenced by the rejection of the pure deterministic observation error models in 4 of the 7 data sets in favor of the Variable Selection (VS) model. Hence, our analysis shows clearly that different strains presented different plasmid loss dynamics. Moreover, the parameter estimates helped to explain which of the three underlying mechanisms was most responsible for the observed rapid plasmid loss. In the case of hosts H2 and P21, segregation rate was estimated to be respectively very low to low, but the plas-
mid cost was high to very high, suggesting that few segregants were formed but swept through the population once they appeared. In contrast, the very high segregation frequency estimate for host R28 together with a low cost suggest that the major cause of rapid plasmid loss is a high segregational loss rate, and not so much the growth advantage of plasmid-free segregants. Importantly, in the case of $\sigma$, the plasmid cost, which was also measured empirically, the model-based estimates were very similar to the experimental estimates obtained by De Gelder et al. (2006) for the same strains.

Although the HT model predicts that there exist certain parameter combinations for which the plasmid-carrying and plasmid-free bacteria coexist in the long-term, the ML parameter estimates for the strain P21 data indicated that the amount of transfer was not high enough to compensate for the loss via segregation and the selection against the plasmid-carrying cells. However, sampling variability can blur the certainty that over time, the plasmids will disappear from a bacterial population. Even if the ML parameter estimates predict the eventual loss of the plasmids, the joint confidence interval of the model parameters may encompass values that are both consistent with the loss of plasmid carrying cells and their long-term presence, as shown by our results (Fig.3).

The VS model is a major departure from other modeling approaches (Cooper et al., 1987; Gasunov and Brilkov, 2002; Bergstrom et al., 2000; Levin et al., 1979; Levin and Stewart, 1980; Levin, 1980; Simonsen, 1991; Tolker-Nielsen and Boe, 1994; Proctor, 1994; Stewart and Levin, 1977; Boe, 1996; Boe and Rasmussen, 1996; Novozhilov et al., 2005; Freter et al., 1983; Seneta and Tavaré, 1982, 1983; Tanaka et al., 2003; Wahl et al., 2002). The process noise included in the VS model represents the variability due to the environment, where the environment may be understood as the host itself and the host’s growth environment (De Gelder et al., 2006). Our model suggests that the plasmid cost changes from one day to the other. The growth rate of fraction of plasmid-free bacteria, a function of the plasmid cost then
becomes a random variable. Modeling plasmid cost as random effect substantially improves our ability to adequately explain much of the data. These stochastic changes in the plasmid cost may be due to the effect of beneficial or deleterious mutations in the host chromosome and/or the plasmid. Plasmid-carrying bacteria can acquire for instance compensatory mutations that allow them to grow as fast as or faster than the plasmid-free bacteria. These processes are not however explicitly modeled in the VS model equations and further work is required to unravel the underlying biological mechanisms behind this variable plasmid cost. How well the VS approximates a more mechanistic model is a topic for further research.

The fact that the VS model provided a substantially better fit to most of the data sets, makes the use of SSM models an important tool for understanding plasmid growth dynamics. However, careful analysis, followed by extensive simulations are required when using SSM models. A recent study in theoretical population dynamics (Dennis et al., 2006) has shown that in the simple case of a Linear and Gaussian SSM, the likelihood function is highly multi-modal and ML estimation is not a trivial task. In fact, in some cases, the globally highest peak in the likelihood function produces parameter estimates which are quite biased compared with estimates based on other modes. However, (Jensen and Petersen, 1999) showed that the maximum likelihood estimates for SSM’s are indeed asymptotically normal which implies consistency. This does not contradict the simulation results of (Dennis et al., 2006), since the multi-modality will eventually disappear and in the limit as the samples size goes to infinity and the maximum likelihood estimate will converge to the true parameter. Yet, for finite amount of data, the maximum likelihood estimate may not be the best. Stochastic methods such as the Gibb sampler used in this paper can often fail to sample all of the modes. For this reason, if multi-modality of the likelihood surface is suspected then extensive validation of the statistical methodology is required. Fortunately, the SSM models presented in this paper did not exhibit any of the exotic behavior found in (Dennis et al., 2006).
Stochastic population dynamics modeling in microbial systems is a subject that is still in its infancy and we’d like to stress that much remains to be done in the experimental, mathematical and statistical areas. One of the final comments of Novozhilov et al. (2005) was that “unfortunately, quantitative estimates (of the different process rates) are lacking, which precludes us from supplementing the mathematical analysis of the model with empirical estimates...” Our study is an attempt to answer the plea of Novozhilov et al. (2005).

Acknowledgements

This work was supported by grants from the National Institutes of Health (P20 RR16448, NIH-R01 GM076040-01, NIH 1 R01 GM73821-02) and the National Science Foundation (NSF-DEB-0515738). We wish to thank Larry J. Forney and Zaid Abdo for their comments and help during the preparation of the manuscript and research.

References


SCHLÜTER, A., H. HEUER, R. SZCZEPANOWSKI, L. J. FORNEY, C. M. THOMAS et al., 2003 The 64,508 bp IncP-1β antibiotic multiresistance plasmid pB10 isolated from a wastewater treatment
plant provides evidence for recombination between members of different branches of the IncP-1β group. Microbiology 149: 3139–3153.


APPENDIX

Derivation of the VS model transition pdf

Recall that in the RHS of eq. 8, that $x_{t-1}$ is the realized value of the process $X_t$ at time $t - 1$. Also, $S_t$ comes from a Normal distribution with mean $\mu$ and variance $\tau^2$. Provided that $0 \leq \lambda < X_t|X_{t-1} = x_{t-1} < 1$, it follows from eq. 10 that

\[
P( X_t \leq x_t | X_{t-1} = x_{t-1} ) = P \left( \frac{x_{t-1} 2^{1 + S_t} + 2 \lambda (1 - x_{t-1})}{x_{t-1} 2^{1 + S_t} + 2(1 - x_{t-1})} \leq x_t \right)
\]

\[
= P \left( S_t \leq \frac{\ln[(1 - x_{t-1})(x_t - \lambda)] - \ln[(1 - x_t)x_{t-1}]}{\ln 2} \right).
\]

After differentiation and setting $h_t = (1/\ln 2) / (\ln[(1 - x_{t-1})(x_t - \lambda)] - \ln[(1 - x_t)x_{t-1}])$ we get:

\[
f_{X_t|X_{t-1}}(x_t) = \frac{(1 - \lambda)}{\ln2 (x_t - \lambda) (1 - x_t) \sqrt{2\pi\tau^2}} \exp \left\{ - \frac{(h_t - \mu)^2}{2\tau^2} \right\},
\]

which is the transition pdf shown in eq. 11.
Tables
Table 1: Likelihood ratio tests and model selection results

<table>
<thead>
<tr>
<th>Strain</th>
<th>$-\ln \hat{L}$ SS (pvalue abs. g.o.f.)</th>
<th>$-\ln \hat{L}$ HT (pvalue abs. g.o.f.)</th>
<th>$-\ln \hat{L}$ VS (pvalue abs. g.o.f.)</th>
<th>P-values for likelihood ratio test of:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SS vs. HT</td>
</tr>
<tr>
<td>H2</td>
<td>44.913 (0.2498)</td>
<td>44.913 (0.1625)</td>
<td>N.A.</td>
<td>1</td>
</tr>
<tr>
<td>R28</td>
<td>49.77485 (0.1304)</td>
<td>49.77485 (0.2457)</td>
<td>N.A.</td>
<td>1</td>
</tr>
<tr>
<td>P21</td>
<td>45.738 (0.00052)</td>
<td>30.24335 (0.5313)</td>
<td>N.A.</td>
<td>1.86534E-07</td>
</tr>
<tr>
<td>P18</td>
<td>246.0254 (0)</td>
<td>245.2713 (0)</td>
<td>104.4029</td>
<td>N.A.</td>
</tr>
<tr>
<td>S34</td>
<td>189.7365 (0)</td>
<td>175.8767 (0)</td>
<td>64.91768</td>
<td>N.A.</td>
</tr>
<tr>
<td>S96</td>
<td>644.6472 (0)</td>
<td>520.86 (0)</td>
<td>59.69714</td>
<td>N.A.</td>
</tr>
<tr>
<td>S55</td>
<td>227.3623 (0)</td>
<td>227.3623 (0)</td>
<td>83.82895</td>
<td>N.A.</td>
</tr>
</tbody>
</table>

The estimated negative log-likelihood score $-\ln \hat{L}$ for each model (SS-segregation selection, HT-horizontal transfer, VS-variable selection) and strain combination is given under the first three columns. The better the model fit, the lowest its computed $-\ln \hat{L}$ value.

For the SS and HT model columns, the absolute goodness of fit p-values in parentheses indicate whether the model describes adequately (p-value > 0.05) or not (p-value<0.05) the data at hand. Finally, the last three columns present the p-values for the Likelihood Ratio Tests to evaluate which model was the best at describing the data.
Figure legends

Figure 1: Sample deterministic trajectories (left) and growth rates (right) of the Horizontal Transfer (HT) model. These plots illustrate that for certain parameter values, the HT model predicts that a long-term coexistence of plasmid-free and plasmid-carrying cells will occur. When coexistence is predicted, the growth rate of the segregants fraction as a function of the fraction of plasmid-free cells adopts a cubic-like form reminiscent of an Allee effect model. Here, however, the interior equilibrium is stable and not unstable as in typical Allee effect models. The different curves correspond to different plasmid cost $\sigma$ values. The other parameters remained fixed. The parameter values used were close to the ML estimates for the strain P21: $\lambda = 6.851044 \times 10^{-05}$, $\theta = 0.25$, $\gamma = 2.443239 \times 10^{-02}$.

Figure 2: For each of four bacterial strains (H2, R28, P21 and P18) the data (black dots) and the estimated model trajectories (lines) are plotted. The top panel figures corresponds to a Maximum Likelihood observation error fit of the dynamic equations 1 for H2 and R28 and 4 and 5 for P21. There, it is assumed that the underlying trajectories obey a deterministic dynamic equation (continuous line) and that any deviation from that trajectory seen in the data (black circles) is attributed to sampling error. The plots in the bottom panel correspond to a process noise plus observation error fit for each of the three replicates of strain P18. The model parameters where estimated using the three replicates simultaneously and are presented separately with a different vertical scale for clarity. In each case, the empty circles joined by a continuous line show the observations and the black circles represent the location of the underlying estimated trajectory (the $X_t$ process) from which the observations were assumed to be drawn.

Figure 3: Stability boundaries for the Horizontal Transfer (HT) Model, location of the estimated
model parameters for the P21 strain and their joint approx. 95% confidence region. In each subplot, the HT model stability boundaries are plotted as a function of each parameter combination. The grey area corresponds to the set of parameter values for which the plasmid-carrying cells eventually disappear (i.e., where the point $x = 1$ is stable) and the white area denote the parameter space for which the point $x = 1$ ceases to be a stable equilibrium and a new stable equilibrium $x^* \in (0, 1)$ appears. In this white area, a long-run coexistence between plasmid-free cells and plasmid carrying cells is predicted. In each case, the black dot locates the Maximum likelihood estimated parameter values and the cloud of blue points around them is their approximate joint 95% confidence cloud.

Figure 4: Example of a simulated trajectory under the Variable Selection (VS) model (triangles), sampled observations from the trajectory according to a binomial sampling process (empty circles) and the HMM estimates of the true simulated trajectory (solid circles) using only the observations (empty circles). The parameter values used for the simulations are $x_0 = 0.0066$, $\mu = 0.01166667$ per generation, $\tau^2 = 0.1583333$ per generation and $\lambda = 1.533333 \times 10^{-05}$ per generation.

Figure 5: Evaluation of the performance of the HMM parameter estimation methods. 1000 data sets containing 3 replicated time series of plasmid loss each were simulated using the Variable Selection (VS) model plus binomial sampling error. For each data set, the posterior modes of the model parameters were found and the thus obtained set of 1000 posterior modes for each parameter was plotted in a histogram and compared against the true values used to simulate (vertical lines).

Figure 6: Evaluation of the performance of the HMM parameter estimation methods. 1000 data sets containing 3 replicated time series of plasmid loss each where simulated using the Vari-
able Selection (VS) model plus binomial sampling error. For each data set, the posterior modes of the underlying (unseen) trajectory were estimated and their difference against the true simulated trajectory was computed. The 1000 sets of differences between the posterior mode of the $X$ process trajectories and the true simulated value of the $X$ process are shown as histograms, from day 0 to day 21.
Figure 1:

Fraction of plasmid free cells at time $t$, $x_t$

Time, in days (10 gen/day)

Growth rate of the fraction of plasmid free bacteria

$\sigma = 0.41$

$\sigma = 0.15$

$\sigma = 0.09$

$\sigma = 0.065$

$\sigma = 0.055$

$\sigma = 0.045$

$\sigma = 0.040$

$\sigma = 0.035$

$\sigma = 0.01$

Fraction of plasmid free bacteria, $x$

Time, in days (10 gen/day)

Fraction of plasmid free cells at time $t$, $x_t$
Figure 2:

- **H2**
- **R28**
- **P21**
- **P18 rep. 1**
- **P18 rep. 2**
- **P18 rep. 3**
Figure 3:
Figure 4:
Figure 5:
Figure 6: