The population biology of bacterial plasmids: a hidden Markov model approach
Supplementary information for the Gibbs Sampler

Jose M. Ponciano, Leen De Gelder, Eva.M. Top and Paul Joyce

To use the Gibbs Sampling algorithm for parameter estimation of the SSM eqs. (15) and (16) in the main text according to the method of Carlin *et al.* (1992) we need first some definitions and establish notation conventions. Let

- $\boldsymbol{\theta} = \left[X_0, X_1, X_2, \ldots, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\tau^2},\right]'$ be the vector of unobservable random variables, where we use greek bold letters to distinguish random variables form point values of the model parameters (eq. 15). Note that $\boldsymbol{\theta}$ is not to be confused here with the parameter θ of the HT model.
- Let $\mathbf{Y}$ be a short-hand notation for the vector of observation from one replicate of the stochastic process X_t (see eq. 12), denoted $\mathbf{Y_j} = [Y_{0,j}, Y_{1,j}, Y_{2,j}, \ldots, Y_{q,j}]'$ in the main text.
- Let $\mathbf{X}$ be a short-hand notation for the vector $[X_0, X_1, X_2, \ldots, X_q]'$. Under the Bayesian paradigm, x_0 is no longer viewed as a point value but as a a random variable. Under both, the frequentist and Bayesian paradigms, $X_1, X_2, \ldots, X_q$ are all random variables defined by the Markov process X_t (see eq. 12).
- Let $p_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{\theta})$ be the sampling density function of the observations given $\boldsymbol{\theta}$.
- Let $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ be the joint (multivariate) prior distribution of the unobservables.
- Let $p_{\mathbf{Y},\boldsymbol{\theta}}(\mathbf{y}, \boldsymbol{\theta})$ be the joint density of the unobservables and observations.
- Finally, we introduce the following short hand notation for three important conditional distributions: recalling that time is indexed from 0 to q, we set

  1. $X_i|X_{t\neq i} \equiv X_i|\left(X_0, X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_q\right)$,
  2. $X_i|X_{t<i} \equiv X_i|\left(X_0, X_1, X_2, \ldots, X_{i-1}\right)$,
  3. $X_i|X_{t>i} \equiv X_i|\left(X_{i+1}, X_{i+2}, X_{i+3}, \ldots, X_q\right)$,

  where $0 \leq t \leq q$ and $0 \leq i \leq q$.

It follows that the joint posterior density function of the process $p_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{\theta})$ is:

$$
\begin{aligned}
p_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y}) &= \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\theta})p_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{\theta})}{p_{\mathbf{Y}}(\mathbf{y})} \\
&\propto p_{\boldsymbol{\theta}}(\boldsymbol{\theta})p_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{\theta}). \qquad \text{(S1)}
\end{aligned}
$$

The conditional posterior distribution for the element θ_i of the vector of unobservables random variables $\boldsymbol{\theta}$ is simply the product of the terms in $p_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y}) \propto p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) p_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{\theta})$ that involve θ_i. The joint $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ prior is simply assumed to be

$$p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = p_{\tau^2}\left(\tau^2\right) p_{\mu}\left(\mu\right) p_{\lambda}\left(\lambda\right) p_{X_0}\left(x_0\right) \prod_{i=1}^{q} p_{X_i|X_{i-1},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\tau^2}}\left(x_i|x_{i-1}, \lambda, \mu, \tau^2\right), \tag{S2}$$

and the sampling density $p_{\mathbf{Y}|\boldsymbol{\theta}}\left(\mathbf{y}|\boldsymbol{\theta}\right)$ is

$$p_{\mathbf{Y}|\boldsymbol{\theta}}\left(\mathbf{y}|\boldsymbol{\theta}\right) = \prod_{t=0}^{q} p_{Y_t|X_t,\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\tau^2}}\left(y_t|x_t, \lambda, \mu, \tau^2\right). \tag{S3}$$

The objective of the Bayesian analysis of Carlin *et al.* (1992) is to marginalize the joint posterior eq. S1 over the components of $\boldsymbol{\theta}$ to obtain posterior means, medians and modes using Gibbs sampling.

The Gibbs algorithm for our state-space model formulation eqs. 15 and 16 is specified below. We denote the m^{th} value of the Gibbs sequence using superscripts in parentheses:

0. Set the random variables: $X_0^{(0)}, X_1^{(0)}, \ldots, X_q^{(0)}, \boldsymbol{\lambda}^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\tau^2}^{(0)}$ at the arbitrary values $x_0^{(0)}, x_1^{(0)}, \ldots, x_q^{(0)}, \lambda^{(0)}, \mu^{(0)}, \tau^{2(0)}$.

1. Generate a sample $\lambda^{(1)}$ from the conditional posterior distribution $p_{\boldsymbol{\lambda}|\mathbf{x},\boldsymbol{\mu},\boldsymbol{\tau^2}}\left(\lambda|\mathbf{x}^{(0)}, \mu^{(0)}, \tau^{2(0)}\right)$, where $\mathbf{x}^{(0)} = [x_0^{(0)}, x_1^{(0)}, \ldots, x_q^{(0)}]$.

2. Generate a sample $\tau^{2(1)}$ from the conditional posterior $p_{\boldsymbol{\tau^2}|\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda}}\left(\tau^2|\mathbf{x}^{(0)}, \mu^{(0)}, \lambda^{(1)}\right)$.

3. Generate a sample $\mu^{(1)}$ from the conditional posterior $p_{\boldsymbol{\mu}|\mathbf{x},\boldsymbol{\tau^2},\boldsymbol{\lambda}}\left(\mu|\mathbf{x}^{(0)}, \tau^{2(1)}, \lambda^{(1)}\right)$.

4. Generate $X_0^{(1)} = x_0^{(1)}$ from the conditional posterior $p_{X_0|Y_0,d_0,X_1}\left(x_0|y_0, d_0, x_1^{(0)}\right)$, where y_0, d_0 denote respectively the observed number of segregant colonies at time 0 and the total number of colonies screened at time 0.

5. Iteratively generate the samples $X_i^{(1)} = x_i^{(1)}$, $i = 1, 2, \ldots, q$ from the conditional posterior distributions $p_{X_i|X_{t\neq i},\boldsymbol{\lambda},\boldsymbol{\tau^2},\boldsymbol{\mu},\mathbf{y},\mathbf{d}}\left(x_i|x_{t<i}^{(1)}, x_{t>i}^{(0)}, \lambda^{(1)}, \tau^{2(1)}, \mu^{(1)}, \mathbf{y}, \mathbf{d}\right)$, where $\mathbf{d} = [d_0, d_1, \ldots, d_q]'$

6. Set $X_0^{(1)} = x_0^{(1)}, X_1^{(1)} = x_1^{(1)}, \ldots, X_q^{(1)} = x_q^{(1)}, \boldsymbol{\lambda}^{(1)} = \lambda^{(1)}, \boldsymbol{\mu}^{(1)} = \mu^{(1)}, \boldsymbol{\tau^2}^{(1)} = \tau^{2(1)}$ and repeat steps 1 to 6 m times. Here we used $m = 31001$.

The above procedure is repeated so as to obtain B samples $\boldsymbol{\theta}^{(b)}, b = 1, \ldots, B$. Following Casella and George's argument above, the $\boldsymbol{\theta}^{(\boldsymbol{b})}$ are then samples of the multivariate joint posterior eq. S1. Then, a sample from the marginal posterior distribution of the i^{th} element of the vector $\boldsymbol{\theta}$, (say $\theta_i = \lambda$) is simply given by $\left\{\theta_i^{(b)}, b = 1, 2, \ldots, B\right\}$, and no high-dimensional integration is necessary (Meyer and Millar 1999). We also note that, alternatively, instead of repeating the procedure B times, a single very long chain can be generated and B samples from it be taken at lags multiple of ℓ, provided that at lag ℓ, the serial autocorrelation in the chain has basically disappeared.

The conditional posterior densities involved in the Gibbs algorithm were:

$$\begin{aligned}
p_{\boldsymbol{\lambda}|\mathbf{x},\boldsymbol{\mu},\boldsymbol{\tau^2}}\left(\lambda|\mathbf{x}^{(0)},\mu^{(0)},\tau^{2(0)}\right) &\propto p_{\boldsymbol{\lambda}}\left(\lambda\right)\textstyle\prod_{t=1}^{q}p_{X_t|X_{t-1},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\tau^2}}\left(x_t|x_{t-1},\lambda,\mu,\tau^2\right),\\
p_{\boldsymbol{\tau^2}|\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda}}\left(\tau^2|\mathbf{x}^{(0)},\mu^{(0)},\lambda^{(1)}\right) &\propto p_{\boldsymbol{\tau^2}}\left(\tau^2\right)\textstyle\prod_{t=1}^{q}p_{X_t|X_{t-1},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\tau^2}}\left(x_t|x_{t-1},\lambda,\mu,\tau^2\right),\\
p_{\boldsymbol{\mu}|\mathbf{x},\boldsymbol{\tau^2},\boldsymbol{\lambda}}\left(\mu|\mathbf{x}^{(0)},\tau^{2(1)},\lambda^{(1)}\right) &\propto p_{\boldsymbol{\mu}}\left(\mu\right)\textstyle\prod_{t=1}^{q}p_{X_t|X_{t-1},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\tau^2}}\left(x_t|x_{t-1},\lambda,\mu,\tau^2\right),\\
p_{X_0|X_{t>0},\mathbf{Y},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\tau^2}}\left(x_0|x_{t>0},\mathbf{y},\lambda,\mu,\tau^2\right) &\propto p_{Y_0|X_0}\left(y_0|x_0\right)p_{X_1|X_0}\left(x_1|x_0\right)p_{X_0}\left(x_0\right),\\
p_{X_t|X_{j\neq t},\mathbf{y},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\tau^2}}\left(\lambda,\mu,\tau^2\right) &\propto p_{Y_t|X_t}\left(y_t|x_t\right)p_{X_{t+1}|X_t}\left(x_{t+1}|x_t\right)p_{x_t|x_{t-1}}\left(x_t|x_{t-1}\right),
\end{aligned}\tag{S4}$$

where $p_{Y_t|X_t}\left(y_t|x_t\right)$ and $p_{x_t|x_{t-1}}\left(x_t|x_{t-1}\right)$ are given by the binomial likelihhood and the Markov process transition density function eq. (12) respectively. To obtain random samples from those conditional posterior pdf's, the Sampling Importance Resampling (SIR) method (Liu 2001) was used. As Carlin *et al* 1992 mention, the SIR method can be inefficient and difficult to use when the cover distributions are difficult to find. However, here all of the unknown trajectories locations have to be a number between 0 and 1. Thus, we used uniform cover distributions and uniform priors for all of the unknowns, and this enormously simplifies sampling according to the SIR algorithm. Assuming uniform priors for our unknowns gave more confidence in the resulting analysis and avoided having to rely in the specification of conjugate priors, as Carlin *et al* 1992 and Meyer and Millar (1999) do, therefore avoiding having to investigate the effect of the priors on our posterior distribution estimates. When $r = 3$ different realizations of the stochastic process are sampled at each time step, the algorithm is basically the same as in the unidimensional case explained above, except that the conditional posterior distributions of the X process (see eq. 12 in the main text) is of multivariate dimension 3.

To evaluate the validity of the Bayesian procedures, the frequentist concept of bootstrap (Efron and Tibshirani 1993) was used and 1000 data sets with 3 replicated sample paths each was simulated and each time the parameter estimates and their posterior distributions were found. Recall that the 1000 data sets containing each three replicated time series of length 22 were simulated using the VS model $x_0 = 0.0066$, $\mu = 0.35$ per 30 generations, $\tau^2 = 4.75$ per 30 generations and $\lambda = 0.00046$ per 30 generations. The prior distributions used where $\mathbf{X}_t \sim \text{Unif}(0,1)$, $t = 1,\ldots 21$, $\boldsymbol{\mu} \sim \text{Unif}(0.01,10)$ and $\boldsymbol{\tau^2} \sim \text{Unif}(0.01,60)$. For $\boldsymbol{\lambda}$ a uniform distribution was also used. From the term h_t in eq. 12 it is readily seen that the value of λ is restricted to be less than x_t, for $t = 1,\ldots,q$, and that is an assumption that derives from the model itself. So strictly, the prior for λ, $p_{\boldsymbol{\lambda}}(\lambda)$ is not independent from x_t and rather, it should be written as $\boldsymbol{\lambda}|X_t = x_t \sim \text{Unif}(0,x_t)$. As it is shown in the results, this fact does not seem to have affected the outcome of the MCMC computations.

The quality of the model fitting to the data in Fig. 3 was assessed and compared between the deterministic and the stochastic models. In previous papers, De Gelder *et al* (2004, 2006), we used Likelihood Ratio Tests (LRTs) to compare among the different deterministic models involved. Similarly we use LRTs to determine if the stochastic model is a significant better explanation of the data relative to the deterministic model eqs. (1,2,4,5). To do so, we defined first $\varphi = [\lambda, \mu, \tau^2, x_{0,1}, x_{0,2}, \ldots, x_{0,r}]'$ to be the parameters of interest. The reader should be aware that this notation now conforms the frequentist paradigm and assumes that the parameters are unknown point values and not random variables. Then, the likelihood function for the observations, denoted $L(\varphi)$ is (see eq. 17 in the main text):

$$\begin{aligned} L(\varphi) &= \textstyle\prod_{j=1}^{r} P(\mathbf{Y_j}|\varphi) \\ &= \textstyle\prod_{j=1}^{r} \int P(\mathbf{Y_j}|\varphi, \mathbf{X_i}) P(\mathbf{X_j}|\varphi) d\mathbf{X_j} \\ &= \textstyle\prod_{j=1}^{r} \int \frac{P(\mathbf{Y_j}|\varphi,\mathbf{X_i})P(\mathbf{X_j}|\varphi)}{P(\mathbf{X_j}|\mathbf{Y_j},\varphi)} P(\mathbf{X_j}|\mathbf{Y_j}, \varphi) d\mathbf{X_j}. \end{aligned} \tag{S5}$$

To calculate the Likelihood score for a particular data set, we evaluated the likelihood at the ML estimates of φ for that data set, *i.e,*,

$$\begin{aligned} L(\hat{\varphi}) &= \textstyle\prod_{j=1}^{r} \mathbb{E}_{(\mathbf{X_j}|\mathbf{Y_j},\hat{\varphi})} \left\{ \frac{P(\mathbf{Y_j}|\hat{\varphi},\mathbf{X_i})P(\mathbf{X_j}|\hat{\varphi})}{P(\mathbf{X_j}|\mathbf{Y_j},\hat{\varphi})} \right\} \\ &\approx \textstyle\prod_{j=1}^{r} \frac{1}{m} \sum_{i=1}^{m} \frac{P(\mathbf{Y_j}|\hat{\varphi},\mathbf{X_i}^{(m)})P(\mathbf{X_j}^{(m)}|\hat{\varphi})}{P(\mathbf{X_j}^{(m)}|\mathbf{Y_j},\hat{\varphi})}, \end{aligned} \tag{C6}$$

where the $(m)^{th}$ sample ($m = 2000$) of the vector X_j was drawn at random using SIR from the conditional posterior density $\mathbf{X_j}|\mathbf{Y_j}, \hat{\varphi})$. This posterior density is the Importance Sampling distribution and is calculated as the product from time 1 to q of the last conditional posterior density shown in eq. S4 for the single replicate case. $P(\mathbf{X_j}|\hat{\varphi})$ is the joint probability distribution of a particular sample path of the X_t process and is computed as $\prod_{t=1}^{q} p_{X_t|X_{t-1},\hat{\varphi}}(x_t|x_{t-1}, \hat{\varphi})$. Finally, $P(\mathbf{Y_j}|\hat{\varphi}, \mathbf{X_i})$ is just the binomial density evaluated at a particular data set and at the ML estimates for that data set. All the calculations were done in the freely available software R (`http://www.r-project.org`) and a Beowulf cluster with 132 nodes.

`http://styx.ibest.uidaho.edu/help/servers/server_info.html`

References

[1] Carlin, B.P., N.G. Polson and D.S. Stoffer, 1992 A Monte-Carlo approach to nonnormal and nonlinear state-space modeling. JASA **87** : 493-500.

[2] De Gelder, L., J.M. Ponciano, Z. Abdo, P. Joyce, L.J. Forney and E.M Top, 2004 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** : 1131-1144.

[3] De Gelder, L., L.F. Vandecasteele, C. Brown, L.J. Forney and E.M. Top, 2005, Plasmid donor affects host range of the promiscuous Inc-1β plasmid pB10 in a sewage sludge microbial community. Appl. Environ. Microbiol. **71** : 5309-5317.

[4] Efron, B. and R. Tibshirani, 1993 *An introduction to the bootstrap.* Chapman & Hall, New York, N.Y.

[5] Liu, J.S., 2001 *Monte Carlo strategies in scientific computing.* Springer Verlag, New York, N.Y.

[6] Meyer, R. and R. B. Millar, 1999 Bayesan stock assessment using a state-space implementation of the delay difference model. Can. J. Fish. Aquat. Sci **56** : 37-52.