Mutation rates: estimating phase variation rates when fitness differences are present and their impact on population structure

Nigel J. Saunders,† E. Richard Moxon1 and Mike B. Gravenor2

1Molecular Infectious Diseases Group, Institute of Molecular Medicine, University of Oxford, Headington, Oxford OX3 9DS, UK
2Institute for Animal Health, Compton, Berkshire RG20 7NN, UK

Phase variation is a mechanism of ON–OFF switching that is widely utilized by bacterial pathogens. There is currently no standardization to how the rate of phase variation is determined experimentally, and traditional methods of mutation rate estimation may not be appropriate to this process. Here, the history of mutation rate estimation is reviewed, describing the existing methods available. A new mathematical model that can be applied to this problem is also presented. This model specifically includes the confounding factors of back-mutation and the influence of fitness differences between the alternate phenotypes. These are central features of phase variation but are rarely addressed, with the result that some previously estimated phase variation rates may have been significantly overestimated. It is shown that, conversely, the model can also be used to investigate fitness differences if mutation rates are approximately known. In addition, stochastic simulations of the model are used to explore the impact of ‘jackpot cultures’ on the mutation rate estimation. Using the model, the impact of realistic rates and selection on population structure is investigated. In the absence of fitness differences it is predicted that there will be phenotypic stability over many generations. The rate of phenotypic change within a population is likely, therefore, to be principally determined by selection. A greater insight into the population dynamics of mutation rate processes can be gained if populations are monitored over successive time points.

INTRODUCTION

Phase variation describes a process of reversible, high-frequency phenotypic switching that is mediated by DNA mutations, reorganization or modification. Phase variation is used by several bacterial species to generate population diversity that increases bacterial fitness and is important in niche adaptation including immune evasion (Saunders, 2003; Salaün et al., 2003). Being able to determine the rate at which these processes occur and the nature of any factors that influence them is integral to understanding the impact of these processes on the evolution and dynamics of the population as a whole and on the host–bacterium interaction. To do this, tools with which to reliably determine and compare phase variation rates within and between experiments and bacterial populations are needed. The estimation of mutation rates in bacteria has a long history. The methods in use, however, are not general across all systems and it is important that the assumptions behind the methods are recognized. Here, we present a new mathematical model that can be used to estimate phase variation rates in the presence of fitness differences, and explore the impact of proposed rates on population structure over time. It is also timely to compare and contrast some of the many different methods and terminology in this field. To put our approach in context, we begin with a brief review of the previous work in this field.

Background to mutation rate estimation

Luria & Delbrück, Lea & Coulson and Stocker. The estimation of mutation rates predominantly uses methods derived from a classic paper by Luria & Delbrück (1943). In these studies, a number of cultures are grown under identical conditions, starting with an inoculum of cells of the same genotype. As they divide some give rise to clones of mutants. At the end of the experiment the cells are plated out and the number that have the mutant phenotype is determined. Because mutation can occur at any time, the number of mutant colonies at the end of the experiment represents the number of mutations plus the accumulation of mutants from the replication of those that arose prior to the final round of division. By chance,
in some cultures mutation will occur earlier than in others, and a much larger proportion of mutants will be present at the end of the experiment. The occurrence of so-called ‘jackpot cultures’ due to very early mutations is a feature of the stochastic nature of this process. It was this insight that led Luria & Delbrück to conclude that mutations (to bacteriophage resistance) were occurring prior to exposure to the selective pressure for which they were adaptive. They showed that the variance between replica experiments was much greater than (instead of equal to) the mean, and the distribution of the number of mutants was characterized by a long tail of rare cases with high numbers of mutant bacteria. Their analysis demonstrated that the mutations were spontaneous and their approach was used by others to show similar spontaneous generation of mutants resistant to antibiotics (Demerec, 1945) and ultraviolet radiation (Withkin, 1946).

Luria & Delbrück described two methods to determine the mutation rate. Their first method used the proportion of cultures in which no mutants could be detected ($P_0$). At low mutation rates, the distribution of mutations will be approximately Poisson, with mean $m$. Accordingly a proportion $P_0 = e^{-m}$ cultures will have no mutant colonies, and $m$ can be estimated by $-\ln(P_0)$, and the mutation rate by $m/(N_t - N_0)$, where $N_t$ is the population size at time $t$. Although it is efficient and simple (Li & Chu, 1987), this method has drawbacks for phase variation experiments, in particular that all cells in a culture are plated and all mutants detected. The efficiency declines greatly if investigation of the whole culture is impractical (Kendal & Frost, 1988; Jones et al., 1994).

The second method is based on the number of mutants in a final culture ($\rho$). In its simplest form, a rate can be obtained by dividing the proportion of mutants ($\rho/N$) by the number of generations that have elapsed ($g$). This is the method most frequently used for estimating phase variation rates (e.g. Eisenstein, 1981). However, this equation was not used by Luria & Delbrück. They recognized that the mean proportion of mutants from several cultures may be unsuitable, due to the possibility of outliers caused by ‘jackpot cultures’, which lead to an overestimate of the mutation rate. To improve accuracy, they calculated the ‘likely’ number of mutants ($r$) that would be observed in a culture given a certain mutation rate. For a mutation rate $a$, and $C$ similar cultures, the likely number of mutants is $r = \ln(NaC)aN_c$. Given experimental observations on the number of mutants, the mutation rate is obtained by numerical methods or from a plot of $r$ against $aN_c$ made for the relevant value of $C$.

Lea & Coulson (1949) extended the Luria & Delbrück model considerably by calculating a precise distribution for the number of mutants. They also determined several methods for generating more accurate mutation rate estimates. Each method was based on initially calculating the number of mutations from the distribution of mutants. The median and maximum likelihood methods have received most attention. However, the solutions are not commonly used by experimentalists as they are computationally complex.

The first paper to specifically consider rates of phase variation was a study of Salmonella flagella by Stocker (1949). He uses a discrete model based on the arguments of Luria & Delbrück, and arrives at the most simple formula, $(\rho/N)/g$. His clear description of this method and its lack of mathematical complexity probably led to its popularity.

**Standard assumptions underlying the estimation of mutation rates.** The basic methods in general use tend to make the following assumptions. (1) The experiment is seeded with cells of a single phenotype/genotype. (2) The growth rates (or fitness) of the different types are equal. (3) All cells have an equal probability of mutating. (4) The mutant population is always small and the total number of wild-type cells approximates the total number of cells in the culture. (5) There is no back-mutation. (6) There is no ‘phenotypic lag’ (delay in the expression of detectable mutant characteristics, Armitage, 1952).

Each of these assumptions can provide difficulties for specific experimental systems. With regard to phase variation, problems may arise if fitness differences are excluded and, as observed by both Bunting (1940) and Stocker (1949), high mutation rates can increase the importance of back-mutation. It is, therefore, important that care is taken in recognizing the underlying assumptions of the models that are applied, some of which are particular issues in studies of phase variation. The first assumption listed above describes starting conditions that are difficult to achieve under the experimental constraints imposed when studying populations with high mutation rates. To ensure an experiment is initiated with cells of a single phenotype, this frequently has to be a single cell. This increases the number of generations prior to the population achieving a size in which a mutation is likely to occur. In addition, the number of divisions may be constrained to those that can be generated upon solid media between colony-forming unit to colony. In many phase variation experiments, therefore, the number of mutants will remain at zero for a significant number of divisions, and the final proportion of mutants will be reduced. In general, a mutation rate will be more accurately estimated if the proportion of mutants in a (non-jackpot) culture is high. This problem is compounded by a frequent need to address small sample populations because it is not possible to select for mutants within the whole population as can be done when the phenotype of interest is directly selectable (such as an antimicrobial resistance). For example, when detecting phenotypic variants by immunoblotting this reduces the sample size of the population and the number of detectable mutants to those that can be addressed on a limited number of plates.

Methods addressing the distribution of mutant accumulation in parallel cultures have been developed to study phenotypes with differing fitness (e.g. Koch, 1982). Stewart et al. (1990) introduced formulae for the prediction of
mutant distributions that could be used when the growth rate of the mutant and the wild-type differ. These studies are important because they illustrate clearly that when the assumptions inherent to the Luria & Delbrück method are not concordant with the experimental conditions (which is often the case) then divergence from their predicted distribution is to be expected, and does not necessarily indicate the action of ‘directed’ mutation (Cairns et al., 1988; Mittler & Lenski, 1992). Subsequent work has resolved many issues related to the prediction of the expected Luria & Delbrück distribution using predetermined mutation rates (Sarker et al., 1992). However, these studies did not directly provide a computationally simple and widely used solution to the determination of mutation rates from experimental data.

**Methods in use for phase variation rate determination.** Despite this extensive background work, the study of phase variation has not consistently applied the same methods. The use of different methods, which have different underlying assumptions, has made it difficult to compare rates between experiments and between phase variable systems. This is compounded when the method or the primary data from which the rates were determined are not stated.

Some papers avoid the issue of calculating a rate, and simply describe the proportion of mutants without reference to the number of generations elapsed (e.g. Roche et al., 1994; Belland et al., 1997). These analyses cannot be easily compared with the results of other studies. This approach has also been used in cases where phenotypes are stated to have different growth rates (Weiser, 1993). This is a widespread and important source of error that may lead to high phase variation rates being described. Other papers, even ones cited for their methods of rate determination (e.g. Hammerschmidt et al., 1996, cited by Bucci et al., 1999), are not explicit about essential steps in the estimation process. These examples serve to illustrate common problems with this type of study. If mutation rates are expressed as the frequency of detectable mutations in a culture of $10^6$ or $10^7$ cells, this is not an appropriate use of the term ‘rate’, unless it is related to the appropriate number of generations. Such an approach can suggest that the rate is higher than its true value. In addition, it is also often not stated whether cultures have been subjected to subculture or if the experiments are initiated with single colonies. Finally, the number of repeated cultures is not often stated. ‘Rates’ determined in this way have been directly compared with those determined on a per cell per generation basis (Bucci et al., 1999). Such a comparison is difficult and might exaggerate differences between the two processes.

**The problem of ‘jackpots’.** An important problem, particularly for the simple rate estimators, is how to summarize the data for the rate formula. The mean number of mutants in replicate cultures is often a poor indicator of mutation rates because of the influence of ‘jackpot cultures’. It is, therefore, more common to use the median. One comparison of the methods indicated that performance decreases from maximum likelihood, the median, the upper quartile, to the mean method (Li & Chu, 1987). Each method was used to derive the parameters of the Lea & Coulson distribution from an experiment, then this distribution was compared to the actual data. It is unclear how the methods compare in systems that depart from this theoretical distribution, when backmutation and fitness differences are present, for instance.

In a subsequent study of simulated experiments, the estimators were compared using predetermined mutation rates (Stewart, 1994). The maximum likelihood measurement was the best predictor of mutation rate. The other methods, particularly the median estimator of Lea & Coulson, performed well and did not normally introduce errors of greater than 10%. However, a few of the estimations gave a mutation rate an order of magnitude too high due to the inclusion of ‘jackpot cultures’. Bayesian procedures have now been presented to analyse fluctuation studies, but these are computationally intense. Each of these methods still requires a relatively large dataset. The Bayesian models were used to investigate the effect of ‘jackpot cut-offs’. This showed that a cut-off of three- to four-fold the median number of mutants did not significantly affect the accuracy of the estimates (Asteris & Sarker, 1996).

There are no general criteria by which one might eliminate ‘jackpots’ from the analysis. However, mutants in a final population that are descended from events that occurred during the initial divisions can be identified as lying outside the primary aim of the experiment. This underscores the methods of adjusting the number of generations used to determine the mutation rate (Drake, 1991), or adjusting the likely number of mutants that will arise (Luria & Delbrück, 1943). By the same argument, an adjustment can be made by excluding ‘jackpots’, after which the statistical objections to the use of the mean are reduced. If enough data are available and a full model for the distribution under study is unavailable, it is pragmatic to exclude unrepresentative data that are understandable from a biological perspective. Excluding values of three times the mean could do this, for example, but specific rules must depend on the estimator used and system under study. Computer simulation experiments using known mutation rates and a model framework appropriate to the experimental system offer a good method of testing the efficiency of excluding ‘jackpots’, and this approach is followed in the sections below.

**A new model for phase variation rate estimation**

Here, we present a mathematical model of the mutation process, with the aim of expanding on the limiting assumptions made in most phase variation studies.

\[
A_{n+1} = 2(1 - \beta)A_n + \alpha A_n + \beta B_n
\]

\[
B_{n+1} = 2(1 - \beta)B_n + \beta B_n + \alpha A_n
\]

**Model 1**
This model describes synchronous growth of a population of bacteria. $A_n$ and $B_n$ represent the numbers of original phenotype and mutant phenotype at generation $n$. Variation can occur from $A \rightarrow B$ and $B \rightarrow A$, potentially at different rates. The probability of a bacterium of type $B$ arising from division of a type $A$ bacterium is $\alpha$; hence, the discrete variation rate is $\alpha$ per cell per generation. The 'backward-variation' rate is $\beta$ per cell per generation. At division, if variation from $A$ does not occur (probability $1-\alpha$), two progeny of type $A$ arise. If phase variation occurs, the progeny consist of one $A$ and one $B$. This represents a departure from the assumptions of Stocker, since in his model both progeny have the mutant phenotype. This was appropriate when considering DNA reorganization prior to division. However, slippage-like processes in repetitive DNA sequences commonly mediate phase variation. In this case, variation is most likely to occur independently on both replicated chromosomes (Fig. 1). Consequently, use of the original formula leads to a two-fold underestimation of the mutation rate. In addition, the inversion that mediates fimbrial phase variation investigated by Stocker can occur at any time in the life-cycle of the bacterium, which contrasts with slippage events which are likely to occur at division. The latter mechanism is more amenable to analysis with discrete (rather than continuous) models, as used here.

We extend Model 1 to include fitness differences, expressed as the probability of surviving to division. In Model 2, proportions $d_A$ and $d_B$ of types $A$ and $B$, respectively, survive each generation.

$$A_{n+1} = 2d_A(1-\alpha)A_n + d_A\alpha A_n + d_B\beta B_n$$

$$B_{n+1} = 2d_B(1-\beta)B_n + d_B\beta B_n + d_A\alpha A_n \quad \text{Model 2}$$

Starting from any initial number and ratio of normal and mutant types ($A_0$, $B_0$), these models can easily be iterated on a spreadsheet (available from the authors) to determine the numbers of each phenotype expected from given variation rates and other parameters.

Analytical solutions for $A$ and $B$ after $n$ generations are given in the Appendix. These provide expressions for the phase variation rates in terms of the proportion of mutants at $n$ generations, which we denote by $p_n$. In the absence of fitness differences (Model 1), starting with a culture consisting solely of type $A$ ($B_0 = 0$), the expected proportion of mutants is:

$$p_n = \frac{\alpha}{2} \left[ 2^n - (2 - \alpha - \beta)^n \right]$$

(1)

Using a summary observation of $p_n$ (e.g. mean or median from experimental replicates), both phase variation rates cannot be calculated separately. For equal forward- and backward-variation, the expression can be rearranged to give:

$$\alpha = 1 - \sqrt{1 - 2p_n}$$

(2)

The analytical solution for Model 2 is cumbersome (an easy way of exploring the model behaviour is to obtain an exact solution by iterating the equations on a spreadsheet). Under the conditions of approximately equal switching rates and initially a population of all phenotype $A$, the following formula expresses the phase variation rate in terms of $p_n$ and the fitness difference $f$ (expressed as the ratio of the individual phenotype growth rates $d_A/d_B$).

$$\alpha = \frac{2p_n f^{n-1} (f-1)}{f^n - 1 + p_n (1 - f^{n-1})}$$

(3)

Performance of the models. Using the notation of the new model, the most common estimator is Stocker’s $\alpha = p_d/n$. For comparison with our model we correct this to allow for a phase variation to produce only one mutant at division (rather than two), giving $\alpha = 2p_d/n$. How does...
this standard formula perform in the presence of back-switching? Unless phase variation rates are very high \((10^{-1} \text{ to } 10^{-2} \text{ per generation})\), the adjusted Stocker formula is useful even if back-switching occurs. If experiments are run over a large number of generations, there will be a tendency to underestimate the rate. However, this will be insignificant (in the presence of other likely variability) even if back-switching occurs at a relatively high rate.

In contrast, the inclusion of fitness differences has a marked effect on the estimators. For example, in an experiment with a starting culture all type \(A\), given underlying switching rates of \(10^{-3}\) and no fitness difference, the expected proportion of phenotype \(B\) after 25 generations is 1.2%. Using Model 2 we can show that a fitness advantage of only 10% for phenotype \(B\) leads to 5.5% \(B\) at 25 generations, on average. If the fitness differences are ignored, an almost five-fold overestimate of the switching rate results (using Equation 2 or Stocker’s). If the fitness difference increases to over 20%, phase variation rates may be incorrectly estimated at several orders of magnitude higher than the true rate.

Equation 3 can be used to derive the correct phase variation rate from the proportion of mutants if fitness differences are known. A more likely use of this estimator is to provide a range of phase variation rates that are consistent with an observed proportion of mutants and range of assumed fitness differences. This is illustrated in Fig. 2. As an example, if after 25 generations the proportion of mutants is 0.03, the phase variation rate can be estimated at between 0.0005 and 0.0025 for an assumed fitness difference of between 10 and 0%.

Might it be possible to estimate fitness differences and phase variation rates simultaneously? If an experiment with a large number of replicates is carried out, an estimate of all parameters in Model 2 is possible using the full distribution of mutant phenotype proportions between cultures. A second method is available if multiple samples of a single culture can be taken to generate a time series of \(p\) against \(n\). This provides considerably more information than the single snapshots of cultures at the end of an experiment. A fit of the Model 2 simulation to such data could possibly distinguish between a case of high phase variation rate/small fitness advantage and low phase variation rate/large fitness advantages (see below). Another approach to this problem is to initially estimate phase variation rates under experimental conditions that minimize the fitness difference. Once the rates have been identified (using Equation 2), subsequent experiments are carried out with the aim of identifying fitness difference under specific conditions (using Equation 3). Equation 3 cannot be solved explicitly for \(f\). However, Fig. 2 can be rearranged and a plot made to show how \(f\) is dependent on a known phase-variation rate and the observed proportion of mutants at the end of an experiment. In Fig. 3, the fitness advantage can be estimated from the proportion of mutants (here, for example, at 25 generations) for an assumed phase variation rate of between \(10^{-5}\) and \(10^{-2}\).

**The mean or median mutant proportion? A stochastic model.** The above exercises explore the models needed when estimating phase variation rates and fitness differences if an appropriate value for the proportion of mutants \((p_n)\) is available from an experiment. Methods of correcting the mean number of mutants observed such as excluding ‘jackpot cultures’, using the median proportion of mutants or altering the time period of the experiment (to adjust for the likely first appearance of a mutant) all address the same issue. Here, we provide some simple suggestions for choosing \(p_n\) based on a fully stochastic simulation of Model 1.

Model 1 is deterministic; it provides the expected value or mean of each population at each generation. This is
because at each generation exactly a proportion $x$ vary from $A$ to $B$ (effectively generating fractions of bacteria). In a stochastic formulation a bacterium either varies or it does not, and at each step an integer number of bacteria undergo phase variation that is drawn at random from a binomial distribution with (for switching $A$ to $B$) mean $A_0x$. Due to the random component, a different outcome occurs every time the model is run, mimicking the experimental process. Fig. 4(a) shows results from 1000 stochastic simulations using parameters $x = \beta = 0.005$. The proportion of mutant phenotype $B$ at 25 generations shows the characteristic skewed distribution explained by Luria & Delbrück (1943). The skew occurs due to the rare occurrence of phase variation in the very first few generations when the culture is exclusively a small number of phenotype $A$. In three runs of the simulation, phase variation occurred in the very first division, and an approximately 50% mutant culture resulted.

The mean of the distribution in Fig. 4(a) equals the output ($p_{n}$) given by Model 1 (5.9% after 25 generations). This might imply the mean of several experiments should be used as the value of $p_{n}$ in Equation 2. If enough replicates are performed, this is indeed the case and use of the median will underestimate the phase variation rate. However, with practical numbers of replicates, estimates based on the median have the attraction of being less variable. The mean provides an unbiased estimate (the expected estimate is the true value of 0.005), but is less precise (more variable). We simulated 100 ‘experiments’, each with 10 replicate cultures (runs of the model). The range of $x$ estimated with Equation 2 varied between ‘experiments’ from 0.0039 to 0.011 if the mean was used, and from 0.0038 to 0.0052 for the median. (Note that these estimates have small absolute errors because fitness difference and the ratio forward/ backward phase variation rates are assumed to be known.) A practical approach when replicates are few is, therefore, to use the median. The distribution of estimates for each method is shown in Fig. 4(b). We also investigated exclusion of ‘jackpot cultures’. Cultures with mutant proportions greater than three times the mean were excluded, and the phase variation rate was calculated from a new mean. This gave results intermediate of the mean and median approach, which were less biased than the median and less variable than the mean.

**Impact of phase variation on population structure.**

An advantage of representing the mutation rate process by a dynamical model (the coupled equations of Models 1 or 2) is that the effect of mutations on population structure can easily be explored over time. We generated a series of simulations of Models 1 and 2 using different values for the switching rate and fitness parameters. Rates of phase variation in the order of $10^{-3}$ to $10^{-2}$ per generation were used in the simulations. Unless stated, initial conditions for the simulations were that all bacteria were phenotype $A$ and none were of type $B$ ($A_0 = 1, B_0 = 0$).

When the phenotypes have equal fitness, phase varying cultures will approach an equilibrium determined by their mutation rates (Bunting, 1940; Stocker, 1949). The proportion of variant bacteria ($B$) will tend to a stable value, $x/(x + \beta)$; hence, if the forward- and backward-switching rates are equal, a 50:50 composition will result. In some instances, phase variation is mediated by repeats located within open reading frames such that alterations in the repeat tract length affect expression by moving the 3’ reading frame in or out of frame with the 5’ initiation codon. In this situation, given the three possible reading frames, the observed phenotypic variation would be expected to be modelled by $x = 2\beta$ and the population approaches a 1/3:2/3 composition. One consequence of the approach to equilibrium is that mutation rates cannot easily be estimated if the culture is grown close to this stage, because the change in population structure slows down and it is unlikely to be determined accurately enough to relate it to the number of generations elapsed.

**Fig. 4.** (a) Distribution of mutant proportion in 1000 simulations of stochastic version of Model 1 (phase variation rates $x = \beta = 0.005$). Mean $p_{25} = 0.059$ (at 25 generations, cultures seeded with type $A$ only). Right-skew indicates jackpots. (b) Use of mean or median proportion of mutants to estimate phase variation (PV) rate. Distribution of estimates obtained from simulation experiments using stochastic model. One-hundred experiments were performed each consisting of 10 runs of the model with a known phase variation rate. In each case, an estimate of $x$ was obtained using the mean (○) and median (line) proportion of mutants ($p_{r}$) from the 10 simulations. The true phase variation rate was 0.005 per generation (vertical line). Mean estimates were 0.005 using the mean, and 0.004 for the median, with a lower variance for the median estimator. Distributions were smoothed by kernel density estimation.
In the absence of fitness differences, and for equal forward- and backward-switching rates, the time taken before a stable population structure is reached is approximately inversely proportional to the mutation rate. Starting with a population of type A bacteria, the number of generations \( n \) before a specified proportion \( x \) of variants is reached is given by:

\[
\ln \left[ \frac{1 - \frac{x(a + b)}{b}}{\ln \left( \frac{1 - x}{b} \right)} \right]
\]

For \( a = b = 0.001 \), the number of generations before a new variant reaches 10% is 223, and for phase variation rates in the order of \( 10^{-4} \) it is over 2000. Thus, considering the time frame and bacterial population sizes that exist in colonization, these rates are not expected to lead to large changes in the proportions of the original and variant phenotypes in the absence of fitness differences (or ‘jackpot’ events) (Fig. 5). Likewise, in the absence of selection, the equilibrium population structure is unlikely to be reached except following very prolonged colonization.

However, supposing mutation rates of \( 10^{-2} \), a large proportion of variants might arise, even in the absence of selection (\( =10\% \) in 22 generations). It could be argued that such high rates are difficult to reconcile with the process of immune evasion because the host would be quickly exposed, and have opportunity to respond, to the full antigenic repertoire of the colonizing population. Furthermore, genome analyses reveal that a single strain may have many phase variable characteristics, as in *Haemophilus influenzae* (Hood et al., 1996), *Helicobacter pylori* (Saunders et al., 1998), *Neisseria meningitidis* and *Neisseria gonorrhoeae* (Saunders et al., 2000; Snyder et al., 2001), *Treponema pallidum* (Saunders, 1999) and *Campylobacter jejuni* (Parkhill et al., 2000). If each of these were phase variable at a rate in the region of \( 10^{-2} \) then there would be insufficient stability for any clone to become adapted to a specific environmental niche without the accumulation of a large proportion of variant, and potentially less-fit, phenotypes. If it occurs, a possible role for phase variation at a very high rate may occur under circumstances associated with small populations, particularly perhaps during colonization. Under these conditions a small inoculum may benefit from rapid diversification in order to generate a clone with advantages in the initial stage of infection, although this has yet to be observed experimentally.

The situation is different when there are fitness differences between the phenotypes. Even at low mutation rates, a more-fit mutant population will almost completely replace the starting phenotype after a time period determined by the difference in fitness. Fig. 6(a, b, c) illustrates the influence of 1, 10 and 50% reductions in the fitness of the starting phenotype relative to the variant phenotype. A 1% fitness advantage results in a significant change in the predominant phenotype, but only over a very large number of generations. In contrast, fitness differences of 10 and 50% (estimates that reflect, for example, the selective pressure of a specific immune response) result in almost complete replacement of the starting population phenotype within 90 and 25 generations, respectively. This is sufficiently rapid to facilitate changes in the phenotype of colonizing and infecting bacterial populations.

Since phase variation continually regenerates the original phenotype, even with a large fitness difference it cannot be completely eliminated. The equilibrium proportion of the less-fit phenotype is proportional to, and at large fitness differences (25% or more) approximately equal to, the...
variation rate. A significant proportion can only coexist at equilibrium, therefore, if the variation rate is very high. Fig. 7(a, b, c) shows that at low mutation rates, the selective difference is the predominant determinant of population composition. Equilibrium populations are similar, and phase variation induces only small differences in the lag before the mutant population dominates, though this lag might still be biologically important to an initial colonizing population.

These findings have implications for experimental design and the interpretation of results. An analysis of the study of Bunting (1940) demonstrates that our model allows interpretation of data in which the starting populations are mixed, and the dynamics of population change are monitored over time. In this study of colour variants in Serratia marcescens grown in vitro in liquid cultures, it was recognized that phase varying populations approached an equilibrium state determined by the relative mutation rates. Since a stable population of 97% of one variant was reached, the author concluded there was an approximately 32:1 difference between the mutation rates mediating the colour changes in each direction. This was assuming that the phenotypes were equally fit in the culture conditions, and the equilibrium of \( f/(\alpha + \beta) \) was reached. The populations were observed over a period of several days in serial subcultures such that the prevalence of one variant was observed to increase from 22% to over 80% in 12 days. If the starting population proportions and the approximate number of generations in Bunting's experiment (generation time of 65 min) are used in our model, then the fitness difference required to alter the population structure as described is only around 1% \( (d_A/d_B=0.99) \). This fitness difference was stated not to be within the limits of detection of her experiment. It is noteworthy that similar conclusions are drawn in our model using variation rates between \( 10^{-3} \) and \( 10^{-5} \) showing that, under these conditions and this degree of selection, the rate of switching within this range is relatively unimportant.

If changes in a single culture can be monitored on several occasions over time, alternative explanations for the changes in population structure can sometimes be distinguished with the model. As an example, Fig. 8(a) shows model simulations whereby population change is determined by either fitness differences (10%) or a biased (1:28) forward-/backward-switching rate. Both mechanisms lead to 50% phenotype \( B \) at 50 generations and are, therefore, indistinguishable given a single snapshot of the process at this time. However, when caused by fitness differences, the population approaches this composition in a sigmoidal manner (Fig. 8a, dotted line). In contrast, population changes caused by biased switching rates approach the same point at a continuously decreasing rate (Fig. 8a, solid line). Such dynamics of population change were monitored by Bunting (1940) indicating, in combination with our model, that the population structure is indeed likely to be influenced by a bias in forward-/backward-switching rate. However, unless very small differences in growth rate can be discounted, the ratio of forward- to backward-switching could be much less than reported. Fig. 8(b) indicates the data can equally well be described by the model with a 1% fitness difference and a 1:3 ratio of forward- to backward-switching. This approach also allows an estimate of the absolute switching rates, in this case in the order of \( 3 \times 10^{-3} \) to \( 9 \times 10^{-3} \) per cell per generation (higher estimates are obtained if equal fitness is assumed). This exercise illustrates the benefits, to both the analysis of the population structure and determination of mutation rates, of considering the dynamics of the population changes rather than snapshots made at one point in time.

As a second example, Weiser et al. (1998) investigated the effects of phase variation of phosphorylcholine (ChoP) in the LPS of Haemophilus influenzae in an infant rat

Fig. 6. Influence of selection on population structure. Proportion of variant phenotype \( B \) shown over 600 generations. (a) \( d_A/d_B=0.99 \); (b) \( d_A/d_B=0.9 \); (c) \( d_A/d_B=0.5 \). Phase variation rates of \( \alpha = \beta = 0.0005 \), and initial proportions \( A_0=1 \), \( B_0=0 \) were used in each simulation.

Fig. 7. Interaction between phase variation and selection. For a 10% difference in fitness \( (d_A/d_B=0.9) \), the effect of phase variation rates on the proportion of variant \( B \) is simulated. (a) \( \alpha = \beta = 0.00005 \); (b) \( \alpha = \beta = 0.0005 \); (c) \( \alpha = \beta = 0.005 \). \( A_0=1 \), \( B_0=0 \). Initial proportion of phenotype \( A \) was 1. Residual percentage of \( A \) at equilibrium is 0.025% (a), 0.25% (b) and 2.5% (c).
Fig. 8. Interaction of fitness differences and biased switching rates. (a) Comparison of the effects of fitness differences and biased forward/backward phase variation rates. Dotted line, population structure determined by fitness difference (dA/DB = 0.9, \( \alpha = \beta = 0.001 \)); solid line, population structure determined by biased forward/backward-switching rates (dA/DB = 1, \( \alpha = 28\beta, \beta = 0.001 \)). Initial proportion phenotype \( A \) = 1. Both simulations generate 50% phenotype \( B \) at 50 generations. (b) Population structure of Serratia marcescens, data from Bunting (1940) table 1. Triangles, relative increase in phenotype \( B \) ('dark-red' variant) in serial culture; circles, relative decrease in phenotype \( A \) ('bright-pink' variant). Initial mixed inoculum of 23% type \( B \). Solid lines show fit of model with estimated parameter values \( dA/DB = 0.99, \alpha = 0.009, \beta = 0.003 \).

Estimating mutation rates

nasopharyngeal colonization model. The data indicate starting proportions of \( A = 0.98 \) \( (\text{ChoP}^+) \) and \( B = 0.02 \) \( (\text{ChoP}^-) \), and after 16 days (approx. 384 generations at 1 h\(^{-1}\)) 73% of the population was of phenotype \( B \). These dramatic changes need not represent a high mutation rate and can be modelled using \( \alpha = \beta = 0.001 \), with a fitness difference of only 1%. The importance of accurately comparing the relative growth rates of the alternate phenotypes. Where this is not possible the assumptions must be clearly stated and the alternative interpretations of the data considered.

DISCUSSION

Any estimate of a mutation rate must be based on an underlying population model. Many such models have been constructed so it is important that their underlying assumptions are clearly presented and understood. The model presented here includes a number of assumptions covered by other authors, but includes back-mutation and fitness differences. The aim is to present an overall model of the population dynamics of phase variation that can be summarized in a small number of coupled equations. Previous models have been presented as a series of arguments, where different assumptions and approximations are introduced at each step. Here, we integrate all assumptions about the underlying process into two simple discrete equations. We hope our approach clarifies the simplifying assumptions that must be made in any mathematical model, and also allows the behaviour of the model system to be easily explored on a spreadsheet. Model simulations are useful to gain a feel for the causes and dynamics of the population changes under study, and a guide to the problem inherent in estimating the parameters. Our model is most appropriate for a synchronous system. If the conditions of the experiment lead to large variation in bacterial generation times, a continuous formulation of the model can be made.

Wherever possible, analyses should include an assessment of the relative fitness of the phenotypes being investigated. If fitness differences are present and unknown, mutation rates cannot easily be accurately estimated. We develop a basic formula so that mutation rates can be estimated in the presence of fitness differences if they are known. A contrasting approach is possible, i.e. to initially estimate mutation rates when fitness differences can be excluded, and then use this rate in Equation 2 to estimate fitness differences when they are present.

Defining a model that characterizes the population changes is only one stage of the process of estimating phase variation rates. A long-standing problem has been what summary value of the proportion of mutants (from replicate experiments) to use in the rate formula. We have used simulations of a stochastic model to compare the mean and median approach using known mutation rates. This gave conflicting results. On average, the mean returned the correct value and the median an underestimate. But the variance and range was much higher using the mean. Bearing in mind all other factors that lead to errors in estimates, the median is likely to be the most acceptable approach. Surprisingly, an estimate based on excluding 'jackpot cultures' (three times greater than the mean), followed by calculation of a new mean, often appeared better than either the mean or median. It outperformed the median and mean in the frequency that an estimate lay within 10–50% of the true value. Unfortunately, this \( \text{ad hoc} \) approach is not likely to apply to all systems. Further statistical work is needed for this question and for the calculation of confidence intervals for estimates. Ideally, the full distribution of mutants numbers, rather than a summary value, should be used and maximum likelihood (or Bayesian) techniques can be pursued based on our model framework.
In addition to providing the mutation rate, the model can be used to explore the phenotypic dynamics of populations over time. We find that in the absence of selection, phase variable populations will tend towards an equilibrium state, after a time period approximately inversely proportional to the variation rate. Phase variation, at the rates observed in vitro, will result in a relatively stable phenotypic population composition in the absence of selection over biologically relevant time periods. When a fitness difference is present, the rate of change in the population composition is largely determined by the relative fitness of the alternate phenotypes. Phase variation also ensures that the less-fit phenotype is not eliminated, and the proportion of the ‘residual’ phenotype at equilibrium is proportional to the phase variation rate.

These predictions are consistent with observations from studies of phase variable systems. For example, the bovine pathogen *Haemophilus somnus* has a phase variable LPS phenotype that is stable over weeks of daily subculture. However, during infection in the natural host there are rapid changes in the LPS of serial isolates. The appearance of variant phenotypes is associated with the generation of specific immune responses, and the phenotypes occur sequentially as the animals are exposed and respond to each (Inzana et al., 1992).

So far the influence of variation of a single gene has been addressed. It should be noted that the situation in vivo is frequently more complex. The presence of multiple phenotypes increases the complexity of analyses of population dynamics, especially if fitness differences occur. In addition, it cannot be assumed that the expression of one phase variable gene does not affect the increased or reduced fitness associated with the expression of others (Blake et al., 1995). It is possible for independent stochastic switching processes to become effectively co-ordinated, and this complex process represents a fruitful avenue for future research.

The model proposed here is designed to encourage investigators to be more explicit about the assumptions underlying their mutation-rate calculations, and to show that additional insights into population change can be gained by considering the dynamics of the process over time. The experimental systems are much more variable than allowed by most models in use. Given a relative fitness difference and assumed relative rates of mutation for the populations under study, a simple formula expresses the per generation rate of mutation. It also facilitates the study of phase variation and population structure under selective conditions once either the variation rate has been determined under non-selective conditions or when the relative fitness of the phenotypes in the experimental conditions is known.

**ACKNOWLEDGEMENTS**

Thanks to Assad Jalali, Rowland Kao and Dan Haydon for helpful comments. N. J. S. was supported by a Wellcome Advanced Research Fellowship in Medical Microbiology, and currently by a Wellcome Advanced Research Fellowship.

**REFERENCES**


Estimating mutation rates

APPENDIX

Analytical solution for Model 1 (no fitness differences)

\[
A_n = \frac{1}{x + \beta} \left[ 2^n \beta (A_0 + B_0) + (2 - x - \beta)^n (xA_0 - \beta B_0) \right]
\]

\[
B_n = \frac{1}{x + \beta} \left[ 2^n x (A_0 + B_0) + (2 - x - \beta)^n (-xA_0 + \beta B_0) \right]
\]

Ignoring fitness differences, if it is assumed that the back-variation rate is \(x\) times the forward rate (\(\beta = x\)), then:

\[
x = \frac{2}{1 + x} \left[ 1 - \sqrt{1 - (1 + x)p_n} \right]
\]

For the special case (described in the text) where back-switching is half as frequent as forward-switching (\(\beta = 0.5x\)):

\[
x = \frac{4}{3} \left( 1 - \sqrt{1 - \frac{3}{2}p_n} \right)
\]

When back-switching is twice as frequent as forward-switching (\(\beta = 2x\)):

\[
x = \frac{2}{3} \left( 1 - \sqrt{1 - 3p_n} \right)
\]

Analytical solution for Model 2 (fitness differences)

Although a full solution is available, the following provides a very good approximation (specifically it is assumed that the quantity \(4d_A d_B \beta = 0\)):

\[
A_n = kA_0 \left[ (2 - x) \right]^n (2 - x) d_A - ((2 - x) d_A)^n (2 - \beta) d_B \]
\[
+ kB_0 \left[ (2 - x) \right]^n \beta d_B - ((2 - \beta) d_B)^n \beta d_B \]
\]

\[
B_n = kB_0 \left[ (2 - \beta) d_B \right]^n (2 - x) d_A - ((2 - \beta) d_B)^n (2 - \beta) d_B \]
\[
+ kA_0 \left[ (2 - x) d_A \right]^n \alpha d_A - ((2 - \beta) d_B)^n \alpha d_A \]
\]

with:

\[
k = \frac{1}{(2 - x) d_A - (2 - \beta) d_B}
\]