

Real time forecasting of near-future evolution

Philip J. Gerrish and Paul D. Sniegowski

J. R. Soc. Interface published online 18 April 2012
doi: 10.1098/rsif.2012.0119

Supplementary data

["Data Supplement"](#)

<http://rsif.royalsocietypublishing.org/content/suppl/2012/04/18/rsif.2012.0119.DC1.html>

References

[This article cites 57 articles, 21 of which can be accessed free](#)

<http://rsif.royalsocietypublishing.org/content/early/2012/04/18/rsif.2012.0119.full.html#ref-list-1>

P<P

Published online 18 April 2012 in advance of the print journal.

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click [here](#)

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

Real time forecasting of near-future evolution

Philip J. Gerrish^{1,2,*} and Paul D. Sniegowski³

¹*Center for Evolutionary and Theoretical Immunology, Department of Biology, University of New Mexico, 230 Castetter Hall, MSC03-2020, 1 University of New Mexico, Albuquerque, NM 87131, USA*

²*Centro de Matemática e Aplicações Fundamentais, Faculdade de Ciências, Universidade de Lisboa, Avenida Professor Gama Pinto, 2, 1649-003 Lisbon, Portugal*

³*Department of Biology, 213 Leidy Laboratories, University of Pennsylvania, Philadelphia, PA 19104, USA*

A metaphor for adaptation that informs much evolutionary thinking today is that of mountain climbing, where horizontal displacement represents change in genotype, and vertical displacement represents change in fitness. If it were known *a priori* what the ‘fitness landscape’ looked like, that is, how the myriad possible genotypes mapped onto fitness, then the possible paths up the fitness mountain could each be assigned a probability, thus providing a dynamical theory with long-term predictive power. Such detailed genotype–fitness data, however, are rarely available and are subject to change with each change in the organism or in the environment. Here, we take a very different approach that depends only on fitness or phenotype–fitness data obtained in real time and requires no *a priori* information about the fitness landscape. Our general statistical model of adaptive evolution builds on classical theory and gives reasonable predictions of fitness and phenotype evolution many generations into the future.

Keywords: distribution of mutational effects; statistical forecasting; fundamental theorem of natural selection; Price equation; cumulant expansion

1. INTRODUCTION

An intuitive association exists between the notions of quantitative science and predictive science. Biology’s accelerating shift into the quantitative realm thus brings with it attendant expectations of predictive powers. Yet evolutionary biology, the discipline built on biology’s most distinguishing process, has been famously silent on the matter of prediction [1,2], and this silence persists despite a well-established quantitative foundation.

Some observations may help us to explain this silence: (i) mutation, the ultimate source of heritable genetic novelty that feeds evolution, is the result of random replication infidelity that is unpredictable, (ii) fitnesses of newly arising mutations can be highly contingent on genetic background and are thus thought to be very unpredictable, (iii) environments, both abiotic and biotic, can change in unpredictable ways, changing fitnesses of existing variants as well as newly arising mutations, (iv) the trajectories and fates of newly arising mutant lineages are strongly affected by random sampling (genetic drift) and (v) recombination

crossovers, despite having hotspots on the genome, can occur at random positions.

These considerations, and others, have understandably given rise to the established (and sometimes celebrated [2]) position that evolution is inherently unpredictable [1]. Evidence in support of this position has been claimed in palaeontology [3,4] (but see [5]), molecular evolution [6], long-term evolutionary studies [7], experimental evolution [8,9] and even *in silico* evolution [10,11]; such evidence, however, typically takes unpredictability to be the null hypothesis and merely reports a failure to reject this null, leaving ample room for type II error. Other evidence is less clear, showing mixed results and reaching ambiguous conclusions [12–14]. In contrast, some work finds clear evidence of parallel evolution in experimental [13,15–17] and natural [18] populations; i.e. *in retrospect*, the evolution of these populations had a degree of predictability. Finally, there are a handful of papers that entertain the notion of predicting evolution in a *prospective* way and speculate how this might be achieved [12,19–21], especially in the context of the influenza virus [22–25].

Predicting the future course of evolution requires a model. We focus on evolutionary models and thus do not consider purely statistical time-series models. Evolutionary models that may be applied to the problem of predicting evolution can be divided into two categories:

*Author for correspondence (pgerrish@unm.edu).

Electronic supplementary material is available at <http://dx.doi.org/10.1098/rsif.2012.0119> or via <http://rsif.royalsocietypublishing.org>.

(i) ‘fitness landscape’ or ‘adaptive topography’ models [26–31] that focus on the mutational input to adaptation and largely ignore, or statistically summarize, the net effects of [32–36], the population dynamics of adaptation and (ii) ‘adaptive process’ models [37–40] that focus on the population dynamics of adaptation; most of these either ignore or make simplistic assumptions about the mutational input to adaptation. (We note that there are some models that resist this simple categorization, cf. [41–43].)

Adaptive topography approaches to predicting evolution are well-suited to long-term prediction, because they map out the mutations underlying the often slow ‘adaptive walk’ of populations; however, there are two significant difficulties with this approach: (i) adaptive topographies are usually contingent on many factors and the assumption of their long-term stasis, implicit in most such models (but see [44,45]), is thus questionable at best, and (ii) mapping out potentially adaptive mutations requires either a vast amount of genotype–fitness data or it requires the experimental evolution of a test population [20]. Acquiring an exhaustive genotype–fitness map is so time-consuming and labour-intensive that this approach cannot rightly be said to offer real time prediction. Furthermore, identifying adaptive mutations through such maps or through *a priori* experimental evolution offers predictions that are not truly prospective in nature.

In light of the significant difficulties with adaptive topography methods, we would argue that a first step towards predicting evolution should implement an adaptive process approach. In what follows, we re-derive a standard evolutionary model that describes the adaptive process and even allows for mutational input in a very general way. We then apply standard transforms to this model resulting in a cumulant expansion (a hierarchy of statistical relations), which can be used to predict evolution in real time, with no *a priori* information about the organism in question or its environment.

Predicting the near-future course of evolution based on these statistical relations is then achieved in three steps: first, a number of fitness or phenotype–fitness measurements are taken from the population in question; these are taken from a number of individuals sampled from the population at one or more time points. Second, these measurements are inserted into the statistical relations to infer the mutational input that feeds ongoing evolution, i.e. mutation rate and distribution of mutational effects. Third, a discretization of the statistical relations gives rise to a telescoping algorithm for projecting the statistical properties of fitness or a fitness-related phenotype into the near future. Using this method, we have succeeded in predicting the future course of *in silico* evolution (fully stochastic simulations) roughly 40 generations into the future.

Our approach differs from many previous discussions of predicting evolution in that we start at a very elemental level: (i) we ask first how the evolution of fitness itself might be forecast, (ii) we then ask how the evolution of a fitness-related phenotype might be forecast and (iii) we soberly aspire to predict evolution only tens of generations into the future. While the

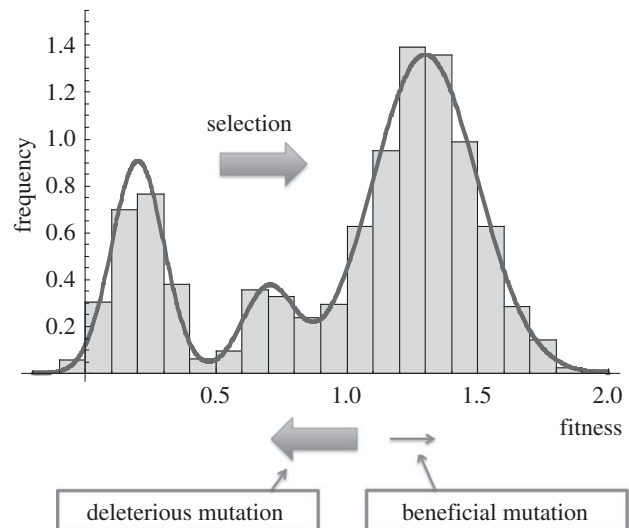


Figure 1. An illustration of the different evolutionary forces that shape the fitness distribution of a population. The histogram represents a finite population and the solid curve represents the corresponding infinite population approximation.

developments presented here offer only a rudimentary beginning, they offer the novel prospect of a theoretical framework onto which more comprehensive prediction algorithms might be built.

2. METHODS

2.1. Assumptions and general model description

We present and analyse a standard model of fitness evolution. We focus on the most general form of this model that makes no asymptotic assumptions and no assumptions about the shapes of the fitness or mutational distributions. In what follows, the word ‘fitness’ will mean ‘haploid fitness’ or ‘additive genic fitness’. Extension to diploid and polyploid organisms is discussed later. The continuous function $u(x,t)$ is defined such that, at time t , the fraction of individuals in the population whose fitness lies in the interval $(x, x+dx)$ becomes identically equal to $u(x,t)dx$ in the limit as dx approaches zero. The deterministic convergence to $u(x,t)dx$ reflects underlying assumptions of infinite population size and infinitely divisible fitness. The model is conservative and thus ensures that the fitness distribution at any given time is a probability density, i.e. $\int_{-\infty}^{\infty} u(x,t)dx = 1$, for all t . In what follows, we refer to the group of individuals whose fitness lies in the interval $(x, x+dx)$ as ‘fitness class x ’.

An example fitness distribution might look something like that illustrated in figure 1, where the histogram represents frequencies of individuals in different fitness bins in a finite population, and the solid line represents the continuous approximation of that distribution. The arrows indicate the different evolutionary ‘forces’ acting on this distribution. The distribution in this illustrative figure is randomly multimodal, and perhaps even a bit improbable, in order to emphasize the point that no assumptions restrict the shape, smoothness or previous dynamics of the distribution. The analysis begins by

making the continuous approximation (solid line), but a comparison with individual-based stochastic simulations demonstrates that the results obtained using the continuous approximation can apply quite well to finite populations (histogram).

2.2. Model of fitness evolution

To model selection, we assume that the *per capita* growth rate of individuals with fitness x is equal to the relative fitness of those individuals, $x - \bar{x}$, where $\bar{x}$ is mean fitness, i.e. $\bar{x} = \bar{x}(t) = \int_{-\infty}^{+\infty} xu(x, t)dx$. To model mutational influx, fitness class x grows in weighted proportion to the sizes of neighbouring fitness classes. The weights depend on the overall mutation rate and on the distance from a given point to x ; they are prescribed by the distribution of fitness-effects of mutations, $g(\gamma, t)$, where γ is the horizontal distance from x , and $g(\gamma, t)$ is a probability density, i.e. $\int_{-\infty}^{+\infty} g(\gamma, t)d\gamma = 1$ for all t . Mutational influx is thus modelled by the convolution $\mu \int_{-\infty}^{+\infty} u(x - \gamma, t)g(\gamma, t)d\gamma$, where μ is genomic mutation rate. Mutational outflux is modelled as a decrease in fitness class x in direct proportion to the size of that fitness class: $-\mu \int_{-\infty}^{+\infty} u(x, t)g(\gamma, t)d\gamma = -\mu u(x, t)$. Summing the effects of selection and mutation gives rise to a general equation for fitness dynamics:

$$\frac{\partial}{\partial t} u(x, t) = (x - \bar{x})u(x, t) + \mu \left(\int_{-\infty}^{+\infty} u(x - \gamma, t)g(\gamma, t)d\gamma - u(x, t) \right). \quad (2.1)$$

This evolutionary model is not new. The equation expresses, in continuous mathematical form, an evolutionary model that has been expressed previously in continuous mathematical, discrete mathematical and verbal form. The model, in the exact form taken in equation (2.1), has been studied previously [41]. In discrete form, it has been dubbed the ‘replicator equation’ [46] and used to describe and study viral ‘quasi-species’ [37, 47–49]. In verbal form, it has been conjectured numerous times and may be as old as evolutionary theory itself.

2.3. Statistical properties of fitness evolution

Our analysis of equation (2.1) uses slight variations on standard transforms. We first transform equation (2.1) to central-moment-generating function (*cmgf*) form; the *cmgf* is defined as $\Pi(\theta, t) = \int_{-\infty}^{+\infty} e^{\theta(x-\bar{x})} u(x, t)dx$ (essentially equivalent to a Laplace transform). Next, the equation is transformed to central-cumulant-generating function (*ccgf*) form, defined as $\psi(\theta, t) = \ln \Pi(\theta, t)$. We define $M(\theta, t)$ to be the moment-generating function (*mgf*) for the distribution of mutational effects, $g(\gamma, t)$. Then the transformed equation for evolutionary dynamics is

$$\psi_t(\theta, t) = \psi_\theta(\theta, t) - \dot{\bar{x}}\theta + \mu(M(\theta, t) - 1), \quad (2.2)$$

where subscripts denote partial derivatives with respect to the subscripted variable, and $\dot{\bar{x}} = (\partial/\partial t)\bar{x}(t)$, the rate of fitness increase or adaptation rate. From this transformed

equation, the following unclosed hierarchy of equations is immediate:

$$\left. \begin{aligned} \dot{\bar{x}} &= \kappa_2 + \mu m_1, \\ \dot{\kappa}_2 &= \kappa_3 + \mu m_2 \\ &\vdots \\ \dot{\kappa}_j &= \kappa_{j+1} + \mu m_j \end{aligned} \right\}. \quad (2.3)$$

where $\kappa_j = \kappa_j(t)$ denotes the j th central cumulant of the fitness distribution at time t (central cumulants are functions of central moments), $\dot{\kappa}_j = (\partial/\partial t)\kappa_j(t)$, the rate of change of the j th central cumulant, and $m_j = m_j(t)$ denotes the j th raw moment of the underlying fitness-effect distribution of newly arising mutations. Fitness measurements from a population can give estimates for the κ_i , which can then be inserted into these equations. In a more general context, equation (2.3) was derived previously by Bürger [41].

2.4. Model of phenotype evolution

Here, we model the evolution of a fitness-related phenotype. The continuous function $u(z, x, t)$ is defined such that, at time t , the fraction of individuals in the population whose fitness lies in the interval $(x, x + dx)$ and whose phenotype lies in the interval $(z, z + dz)$ becomes identically equal to $u(z, x, t)dx dz$ in the limit as dx and dz approach zero. Again, the model is conservative and thus ensures that the fitness–phenotype distribution at any given time is a probability density.

Again, *per capita* growth rate of individuals with fitness x is equal to the relative fitness of those individuals, $x - \bar{x}$, where $\bar{x}$ is the mean fitness, i.e. $\bar{x} = \bar{x}(t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xu(z, x, t)dx dz$. The mutational influx is given by the convolution $\mu \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} u(z - \gamma_z, x - \gamma_x, t)g(\gamma_z, \gamma_x, t)d\gamma_z d\gamma_x$, where μ is the genomic mutation rate. The mutational outflux is $-\mu u(z, x, t)$. Summing the effects of selection and mutation gives rise to an equation for phenotype dynamics:

$$\begin{aligned} \frac{\partial}{\partial t} u(z, x, t) &= (x - \bar{x})u(z, x, t) \\ &+ \mu \left(\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} u(z - \gamma_z, x - \gamma_x, t) \right. \\ &\times g(\gamma_z, \gamma_x, t)d\gamma_z d\gamma_x - u(z, x, t) \Big). \end{aligned} \quad (2.4)$$

2.5. Statistical properties of phenotype evolution

We analyse equation (2.4) by first transforming it to *cmgf* form; the *cmgf* is defined as $\Pi(\phi, \theta, t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{\phi(z-\bar{z}) + \theta(x-\bar{x})} u(z, x, t)dx dz$. Next, the equation is transformed to *ccgf* form, defined as $\psi(\phi, \theta, t) = \ln \Pi(\phi, \theta, t)$. The transformed evolutionary equation is

$$\begin{aligned} \psi_t(\phi, \theta, t) &= \psi_\phi(\phi, \theta, t) - \dot{\bar{z}}\phi - \dot{\bar{x}}\theta \\ &+ \mu(M(\phi, \theta, t) - 1), \end{aligned} \quad (2.5)$$

where $\dot{\bar{z}} = (\partial/\partial t)\bar{z}(t)$, the rate of increase in mean phenotype, and again $\dot{\bar{x}} = (\partial/\partial t)\bar{x}(t)$, the rate of fitness increase. In a way similar to previous

derivations, but now in two spatial dimensions, the dynamic fitness–phenotype distribution is characterized by this set of equations:

$$\left. \begin{aligned} \dot{\kappa}_{i,0} &= \kappa_{i,1} + \mu m_{i,0}, \\ \dot{\kappa}_{i,1} &= \kappa_{i,2} + \mu m_{i,1} \\ &\vdots \\ \dot{\kappa}_{i,j} &= \kappa_{i,j+1} + \mu m_{i,j} \end{aligned} \right\} \quad (2.6)$$

where $\kappa_{i,j} = \kappa_{i,j}(t)$ denotes the (i,j) th central cumulant of the fitness distribution at time t , $\dot{\kappa}_{i,j} = (\partial/\partial t)\kappa_{i,j}(t)$, the rate of change of the (i,j) th central cumulant, and $m_{i,j} = m_{i,j}(t)$ denotes the (i,j) th raw moment of the underlying fitness-effect distribution of newly arising mutations. Fitness–phenotype measurements from a population can give estimates for the $\kappa_{i,j}$, which can then be inserted into these equations. We note that setting $i=1$ gives the dynamics of mean phenotype ($\kappa_{1,0} = \bar{z}$), a fact we will use later under §3.1.

2.6. Refinements

2.6.1. Ploidy and recombination

For organisms whose ploidy is $n > 1$, equations (2.3) and (2.6) may be approximated as follows: (i) account for masking effects by replacing the κ_{j+1} or $\kappa_{i,j+1}$ on the right-hand side by $s\kappa_{j+1}$ or $s\kappa_{i,j+1}$, where $s \in (0,1)$ is a strength-of-selection factor (the first of these replacements was also noted previously by Burger [41]), and (ii) replace μm_j by $\mu m_j \bar{D}_j$ or $\mu m_{i,j}$ by $\mu m_{i,j} \bar{D}_j$, respectively, where $\bar{D}_j = \sum_{i=1}^n (1 + \lambda_i)^j$, the *de novo* contribution of dominance to the j th moment in an n -ploid organism, where λ_i is the gradient of mean dominance of the i th ploid (or gene segment copy, in the case of segmented viruses) with respect to fitness (see the electronic supplementary material). The foregoing analyses apply when the unit of selection is either the entire genome, if it is haploid, or the ‘gene’ if ploidy is $n > 1$. It may be the case that recombination occurs within the unit of selection, especially if the unit of selection is the entire genome. When such intra-unit recombination occurs at a high rate with no net effect of dominance, its effects may be approximated by adding the term $2^{-(j+1)}r\pi_j$ and $2^{-(j+1)}r\pi_{i,j}$ to the right-hand side of all but the first relation in equations (2.3) and (2.6), respectively, where r is the recombination rate and π_j is the j th central moment of the fitness distribution and $\pi_{i,j}$ is the (i,j) th central moment of the phenotype–fitness distribution.

2.6.2. Epistasis

The foregoing analyses allow mutational input to be time-dependent. While this approach should capture most mutational dynamics reasonably well, it may miss certain aspects attributable to the contingency of mutational effects on genetic background, or epistasis. We account for epistasis by assuming that the fitness effects of newly arising mutations are a function, not only of time but also of the fitness of the genetic background onto which the mutations arise. We define δ_j to be the gradient of the j th moment of the mutational effects distribution

with respect to fitness. Then, the moments m_j and $m_{i,j}$ in equations (2.3) and (2.6) are simply replaced by

$$\tilde{m}_j = m_j + \delta_{j+1} + \sum_{k=0}^{j-1} \binom{j}{k} \delta_k \kappa_{j-k+1}$$

and

$$\tilde{m}_{i,j} = m_{i,j} + \delta_{j+1} + \sum_{k=0}^{j-1} \binom{j}{k} \delta_k \kappa_{i,j-k+1},$$

respectively, where the m_j and $m_{i,j}$ now denote ‘baseline’ moments (see the electronic supplementary material). These baseline moments may be functions of time, reflecting a distribution of mutational effects that is both time-dependent and contingent on genetic background.

3. RESULTS

3.1. Predicting evolution

The foregoing statistical relations may be used to predict the near-future course of evolution, following the three steps outlined in §1. To avoid unnecessary repetition, in much of what follows, we treat the cases of fitness and phenotype evolution in parallel. Results pertaining to fitness evolution derive from equation (2.3), while results pertaining to phenotype evolution derive from equation (2.6). While we give the procedure to estimate phenotype–fitness mutational input quite generally, i.e. for all the two-dimensional moments ($\mu m_{i,j}$ for all (i,j)), we focus only on predicting the evolution of mean phenotype after that; this is achieved by setting $i=1$ and limiting the equations to $\mu m_{1,j}$ and $\kappa_{1,j}$. The Discussion outlines the generalization that projects the full distribution.

3.1.1. The necessary measurements

Predicting the future course of fitness evolution requires fitness measurements taken from several individuals in the population. As a minimal requirement, several such measurements must be taken at a single time point, namely, the present. To make predictions based on such minimal data, it requires two assumptions: (i) that the mutational input to evolution does not change over time, and (ii) that the fitness distribution has converged to a constant shape that simply propagates towards increased fitness. These two ‘dynamic equilibrium’ assumptions, while somewhat questionable, may provide a reasonable first approximation; indeed, adaptive evolution under this regime, called the ‘solitary wave’ of adaptation, has been studied appreciably [50–54]. To relax these assumptions, fitness measurements must be taken at the present time and at one or more time points in the past; now, predictions do not rely on the ‘dynamic equilibrium’ assumptions described earlier but on the less-restrictive single assumption that mutational input follows unknown but generally characterizable trends.

Predicting the future evolutionary course of a fitness-related phenotype requires measurements of both the fitness and the phenotype of several individuals in the population. As with the fitness-only predictions described

earlier, and for the same reasons outlined there, it is preferable to have measurements taken at the present time and at one or more time points in the past.

3.1.2. Inferring mutational input

Data taken from a single time point. As indicated already, evolutionary predictions can be made based on measurements from only a single sample taken at the present time; however, this relies on the two ‘dynamic equilibrium’ assumptions described earlier. The second of these two assumptions translates to the conditions $\dot{\bar{x}} = c_x$ and $\dot{\kappa}_j = 0$ in equation (2.3) (c_x is constant) for the case of fitness evolution; and it translates to the conditions $\dot{\bar{z}} = c_z$ and $\dot{\kappa}_{i,j} = 0$ in equation (2.6) (c_z is constant) for the case of phenotype evolution. Applying these conditions to equations (2.3) and (2.6) directly gives the quantities μm_j or $\mu m_{i,j}$ that characterize the fitness or phenotype–fitness mutational input, respectively:

$$\begin{aligned} \text{fitness:} \quad & \begin{cases} \mu m_1 = c_x - \kappa_2 \\ \mu m_2 = -\kappa_3 \\ \vdots \\ \mu m_j = -\kappa_{j+1} \end{cases} \\ \text{phenotype:} \quad & \begin{cases} \mu m_{i,0} = \tilde{c}_z - \kappa_{i,1} \\ \mu m_{i,1} = -\kappa_{i,2} \\ \vdots \\ \mu m_{i,j} = -\kappa_{i,j+1}, \end{cases} \end{aligned} \quad (3.1)$$

where

$$\tilde{c}_z = \begin{cases} c_z, & i = 1 \\ 0, & i > 1 \end{cases}.$$

Data taken from more than one time point. Using data taken from more than one time point, we can relax the questionable ‘dynamic equilibrium’ assumptions described earlier. This is achieved simply by a coarse discretization of equations (2.3) and (2.6). The quantities μm_j or $\mu m_{i,j}$ (now dynamic) are thus estimated as follows:

$$\begin{aligned} \text{fitness:} \quad & \begin{cases} \mu m_1(t) = (\bar{x}(t) - \bar{x}(t - \Delta t)) / \Delta t - \kappa_2(t) \\ \mu m_2(t) = (\kappa_2(t) - \kappa_2(t - \Delta t)) / \Delta t - \kappa_3(t) \\ \vdots \\ \mu m_j(t) = (\kappa_j(t) - \kappa_j(t - \Delta t)) / \Delta t - \kappa_{j+1}(t) \end{cases} \\ \text{phenotype:} \quad & \begin{cases} \mu m_{i,0}(t) = (\kappa_{i,0}(t) - \kappa_{i,0}(t - \Delta t)) / \Delta t - \kappa_{i,1}(t) \\ \mu m_{i,1}(t) = (\kappa_{i,1}(t) - \kappa_{i,1}(t - \Delta t)) / \Delta t - \kappa_{i,2}(t) \\ \vdots \\ \mu m_{i,j}(t) = (\kappa_{i,j}(t) - \kappa_{i,j}(t - \Delta t)) / \Delta t - \kappa_{i,j+1}(t), \end{cases} \end{aligned} \quad (3.2)$$

where Δt is the sampling interval, and everything else on the right-hand side of the equations is obtainable from the fitness or phenotype–fitness measurements. No asymptotic assumptions are required here, and the dynamics of the distribution of mutational effects can thus be studied by estimating $\mu m_j = \mu m_j(t)$ or $\mu m_{i,j} = \mu m_{i,j}(t)$ at different time points during the evolution of a population.

We note that mutation rate cannot be estimated separately in the earlier-mentioned non-parametric scheme (only the products $\mu m_j(t)$ and $\mu m_{i,j}(t)$ can be estimated). Furthermore, if recombination is present and r is to be estimated too, then there is one too many parameters, i.e. the system of equations is under-determined. To remedy this, it may be assumed that the underlying distribution, $g(\gamma, t)$ or $g(\gamma_z, \gamma_x, t)$, has a certain form (making the problem parametric); this can result in a well-determined or over-determined system that can, in principle, be used to estimate μ , r and the parameters of the now-parametric underlying distribution, a method we have tested with some success.

Comparing estimates of μm_i and $\mu m_{i,j}$ to their known values in simulations. As a first means to evaluate the theory developed here, we employed the foregoing estimation procedures on virtual data from fully stochastic simulations of populations undergoing adaptive evolution and compared these estimates with the known values that were put into the simulations. Figure 2 plots the results of one such comparison for fitness evolution and one for phenotype evolution.

3.1.3. Projecting the near-future course of evolution

With the estimates of μm_i or $\mu m_{i,j}$ in hand, we have the information necessary to project near-future evolution. Here again, we employ a discretization of equations (2.3) and (2.6), but this discretization is much less coarse than that used to determine mutational input. The first equation becomes

$$\begin{aligned} \text{fitness:} \quad & \bar{x}(t+1) = \bar{x}(t) + \kappa_2(t) + \mu m_1 \\ \text{phenotype:} \quad & \bar{z}(t+1) = \bar{z}(t) + \kappa_{1,1}(t) + \mu m_{1,0}. \end{aligned} \quad (3.3)$$

This gives mean fitness only one generation from now (where ‘now’ is time t). The formula for mean fitness (or mean phenotype) two generations from now is

$$\begin{aligned} \text{fitness:} \quad & \bar{x}(t+2) = \bar{x}(t+1) + \kappa_2(t+1) + \mu m_1 \\ \text{phenotype:} \quad & \bar{z}(t+2) = \bar{z}(t+1) + \kappa_{1,1}(t+1) + \mu m_{1,0}, \end{aligned} \quad (3.4)$$

where $\bar{x}(t+1)$ (or $\bar{z}(t+1)$) have already been calculated in equation (3.3). To complete this calculation, however, we need a predicted value for $\kappa_2(t+1)$ (or $\kappa_{1,1}(t+1)$); this is obtained from discretizing the second expression in equations (2.3) and (2.6), which gives:

$$\begin{aligned} \text{fitness:} \quad & \kappa_2(t+1) = \kappa_2(t) + \kappa_3(t) + \mu m_2 \\ \text{phenotype:} \quad & \kappa_{1,1}(t+1) = \kappa_{1,1}(t) + \kappa_{1,2}(t) + \mu m_{1,1}. \end{aligned} \quad (3.5)$$

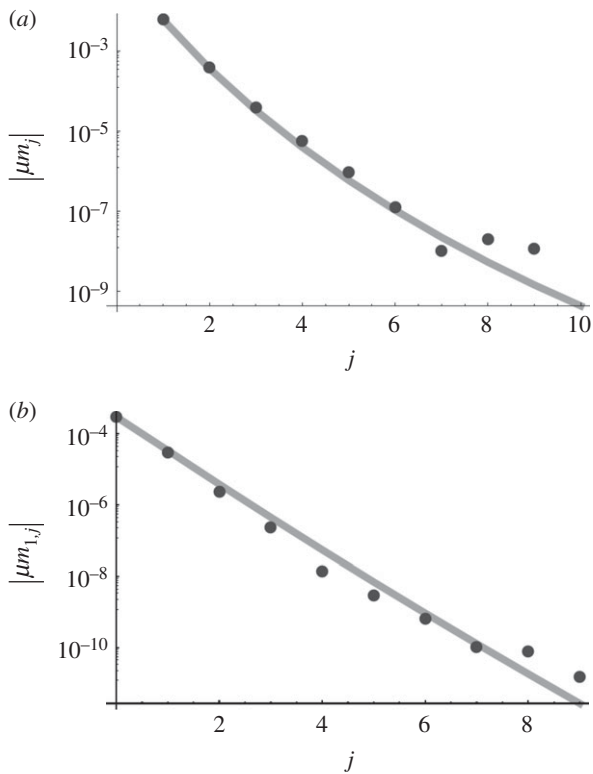


Figure 2. Assessing inference of mutational input. Absolute values of the μm_j and $\mu m_{1,j}$. Known values (solid lines), calculated from the parameters of the simulations, follow the curve. Circles show the averages of values estimated from equation (3.2). In the individual-based stochastic simulations, deleterious mutation rate was 0.0001, effects of mutations were drawn from an exponential distribution with mean 0.03 and population size was 10 000. (a) Fitness evolution and (b) phenotype evolution.

To predict mean fitness (or mean phenotype) three generations from now, we use

$$\left. \begin{aligned} \text{fitness: } \bar{x}(t+3) &= \bar{x}(t+2) + \kappa_2(t+2) + \mu m_1 \\ \text{phenotype: } \bar{z}(t+3) &= \bar{z}(t+2) + \kappa_{1,1}(t+2) + \mu m_{1,0}, \end{aligned} \right\} \quad (3.6)$$

where $\bar{x}(t+2)$ and $\bar{z}(t+2)$ are obtained from equation (3.4), and $\kappa_2(t+2)$ and $\kappa_{1,1}(t+2)$ are obtained from

$$\left. \begin{aligned} \text{fitness: } \kappa_2(t+2) &= \kappa_2(t+1) + \kappa_3(t+1) + \mu m_2 \\ \text{phenotype: } \kappa_{1,1}(t+2) &= \kappa_{1,1}(t+1) + \kappa_{1,2}(t+1) + \mu m_{1,1}, \end{aligned} \right\} \quad (3.7)$$

where

$$\left. \begin{aligned} \text{fitness: } \kappa_3(t+1) &= \kappa_3(t) + \kappa_4(t) + \mu m_3 \\ \text{phenotype: } \kappa_{1,2}(t+1) &= \kappa_{1,2}(t) + \kappa_{1,3}(t) + \mu m_{1,2}. \end{aligned} \right\} \quad (3.8)$$

And so on. In general, this iterative ‘telescoping’

procedure can be written compactly as

fitness:

$$\kappa_j(\tau) = \kappa_j(\tau-1) + \kappa_{j+1}(\tau-1) + \mu m_j(\tau-1)$$

phenotype:

$$\kappa_{1,j}(\tau) = \kappa_{1,j}(\tau-1) + \kappa_{1,j+1}(\tau-1) + \mu m_{1,j}(\tau-1) \quad (3.9)$$

for $\tau = t+1, t+2, \dots, t+R$, where R is the predictive reach of the algorithm. In practice, the hierarchy of equations cannot be infinite, and there must be some $j_{\max}$ at which the system of equations is closed (equivalent to the problem of moment closure). To close the system of equations, we suppose that $\kappa_{j_{\max}+1}(\tau-1) = 0$ and $\kappa_{1,j_{\max}+1}(\tau-1) = 0$, thereby closing the system of equations with this last equation:

$$\left. \begin{aligned} \text{fitness: } \kappa_{j_{\max}}(\tau) &= \kappa_{j_{\max}}(\tau-1) + \mu m_{j_{\max}}(\tau-1) \\ \text{phenotype: } \kappa_{1,j_{\max}}(\tau) &= \kappa_{1,j_{\max}}(\tau-1) + \mu m_{1,j_{\max}}(\tau-1). \end{aligned} \right\} \quad (3.10)$$

3.2. Testing predictions against simulations

As a first step towards assessing the accuracy of our methods for predicting evolution, we wrote simulations to mimic evolving populations. These simulations kept track of every individual and every replication event in the population. Offspring could differ from parents in fitness. At the outset, every individual in the population had a fitness of one and the wild-type genomic mutation rate set by the user.

3.2.1. Mechanics of the simulations

At each time step, every individual produced a number of offspring, X , drawn at random from a Poisson distribution whose mean was $w_i/\bar{w}$, where w_i is the fitness of the i th individual in the population and $\bar{w} = (1/N) \sum_{j=1}^N w_j$ is the mean fitness of the population. Each offspring acquired: (i) a number X_D of new deleterious mutations that each decreased fitness by a factor $(1 - M_D)$, (ii) a number X_B of new beneficial mutations that each increased fitness by a factor $(1 + M_B)$. X_D and X_B are Poisson random variables with means μf_D , μf_B , respectively, where μ is the base genomic mutation rate, and the f s are fractions of deleterious and beneficial mutations, respectively—parameters set by the user. To simulate mutational effects, $M_D \geq 0$ and $M_B \geq 0$ were continuous random variables with means m_D , and m_B , respectively. These methods were employed for figure 2. Parameter values for most simulations were chosen to reflect the adaptive regime of some experimental evolution populations [55,56]: population size and beneficial mutation rate were well into the ‘clonal interference’ regime of adaptive evolution, in which several competing beneficial mutations are likely to co-exist in the population at any given time.

To mimic a static environment in which the degree of adaptedness changes over time, and to assess predictive accuracy under these non-asymptotic conditions, some simulations implemented a limited supply of beneficial mutations (figures 3–6). This was achieved by considering all beneficial mutations to reside on a finite, indeed

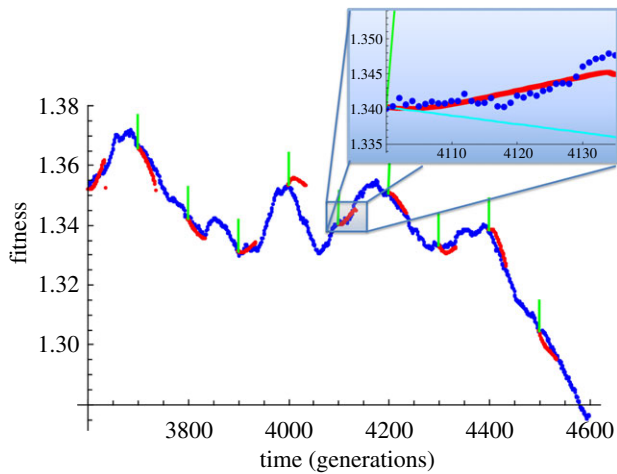


Figure 3. Predictions for fitness evolution. Blue line plots the fitness trajectory of the simulated population; red lines plot predicted fitness trajectories based on our telescoping method. The data used for the telescoping method are: a sample of 100 fitness measurements taken from the population at the start of each red line, and second sample of 100 fitness measurements taken 100 generations prior to that (to estimate the $\mu m_j(t)$). Green lines plot the extrapolated predictions of Fisher's theorem [39], and the light blue line in the inset plots the extrapolated prediction of Kimura's related equation [57] (the F/K equation; see §4). Parameters for the simulations are the same as those given in figure 2, except here the supply of beneficial mutations is limited (there are 64 available at the outset) and is thus used up as the population evolves.

small, 'beneficial genome', modelled as a string of ones and zeros, i.e. a 'bit-string' genome. Figures 3 and 4 plot example simulations (blue curves) and predictions based on 'data' taken from those simulations for the near-future course of evolution (red curves).

3.2.2. Assessing predictive accuracy of our methods

Taking a lead from meteorology, we employed normalized root-mean-square error (r.m.s.e.) as our indicator of predictive accuracy, given by the formula:

$$\frac{1}{\bar{x}_{\max} - \bar{x}_{\min}} \sqrt{\frac{1}{n} \sum_{i=1}^n (\bar{x}_{p,i} - \bar{x}_{o,i})^2},$$

where $\bar{x}_{p,i}$ is the i th predicted value for mean fitness, $\bar{x}_{o,i}$ is the i th observed value for mean fitness, $\bar{x}_{\max}$ and $\bar{x}_{\min}$ are maximum and minimum observed mean fitness. Figure 5 shows how this indicator of predictive accuracy is affected by population size. As expected, our methods perform much better when selection is the 'limiting step' to evolution (empty symbols, figure 5), i.e. in the 'clonal interference' regime [60]; they do not perform as well when mutation is the limiting step (filled circles, figure 5), i.e. in the 'periodic selection' regime.

The difficulty with a straight comparison between observed data and predicted values, as plotted in figure 5, is that there is no obvious way to formulate a null hypothesis representing 'inaccurate prediction', whose rejection would thereby support 'accurate prediction'. To quantitatively assess predictive accuracy, therefore, we compared our predictions with those of

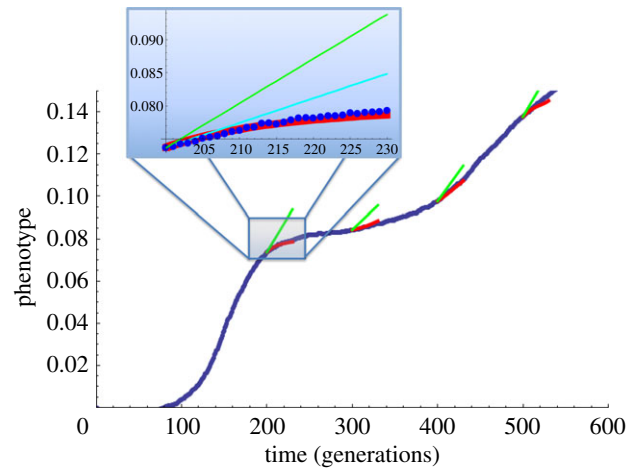


Figure 4. Predictions for phenotype evolution. Blue line plots the fitness trajectory of the simulated population; red lines plot predicted phenotype trajectories based on our telescoping method. The predictions made, using our method (red lines), are based on one set of 100 phenotype–fitness measurements taken from the population at the start of each red line, and a second set of 100 phenotype–fitness measurements taken 100 generations prior to that (to get current estimates of the $\mu m_{1,j}(t)$). Green lines plot the extrapolated predictions of the Price equation, and the light blue line in the inset plots the extrapolated prediction of the full Price equation [40,58,59]. Parameters for the simulations are the same as those given in figure 2, except for the distribution of mutational effects, which is a two-dimensional Gaussian distribution in fitness and phenotype with means $(-0.03, -0.08)$ and covariance matrix $\begin{pmatrix} 0.0020 & 0.001 \\ 0.0010 & 0.001 \end{pmatrix}$.

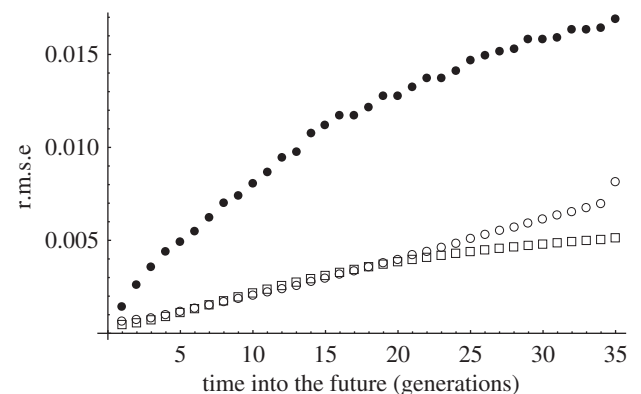


Figure 5. Effect of population size on predictive accuracy. Normalized r.m.s.e. of predictions (see text) is plotted as a function of time into the future that predictions are made. Predictions are compared against individual-based simulations in which deleterious mutation rate was $\mu f_D = 0.05$, beneficial mutation rate was $\mu f_B = 5 \times 10^{-5}$, and fitness effects of mutations were drawn from an exponential distribution with mean 0.03. Different symbols indicate different population sizes and different corresponding quantities $F = 2N\mu f_B \log(Ns/2)$ indicating the adaptive regime [60]: filled circles indicate a population of size $n = 1000$ that is in the periodic selection regime ($F = 0.27 < 1$); empty circles and empty rectangles indicate populations of size $n = 10000$ and $n = 100000$ that are in the clonal interference regime ($F = 5.01 > 1$ and $F = 73.13 \gg 1$, respectively).

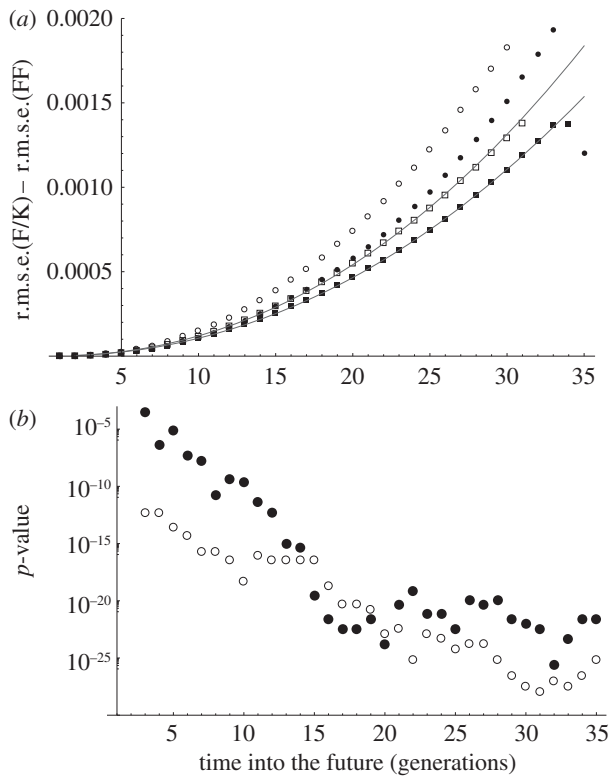


Figure 6. Improvement over F/K projections. We employ simulations to assess the predictive accuracy of our methods by comparing it to the accuracy of linear extrapolations of F/K predictions (see §4). Simulations were fully stochastic and individual-based; deleterious mutation rate was 0.05, beneficial mutation rate was 0.0001, effects of mutations were drawn from an exponential distribution with mean 0.03. Circles indicate that a population size of 10 000 was used; squares indicate that a population size of 50 000 was used. Filled indicates that the entire population was sampled to compute the cumulants; empty indicates that cumulants were calculated from a sample of size 100 taken from the population. Values plotted are computed from 400 non-overlapping prediction-assessment intervals of 50 generations each, taken from two simulated evolving populations, each with a non-renewable supply of 96 beneficial mutations that is used up as the population adapts. (a) The r.m.s.e. of F/K projections (r.m.s.e.(F/K)) minus the r.m.s.e. of our fitness forecasting method (r.m.s.e.(FF)), quantifying the improvement in predictive accuracy of our methods over F/K projections. Curiously, these curves follow a power-law (solid curves) with exponents 1.82 (whole population) and 1.86 (sample size 100). (b) Probabilities supporting the null hypothesis (p -values) that our methods perform no better than F/K projections. These values were calculated using the non-parametric sign test as suggested by Diebold & Mariano [61], which if anything should err on the conservative side.

the ‘next best’ predictor, which is linear extrapolation of the Fisher/Kimura (F/K) equation (figure 6), given by the equation $\bar{x}(t) = \bar{x}(0) + \kappa_2 t + \mu m_1$, where $\bar{x}(t)$ is mean fitness projected t generations into the future, $\kappa_2 = \sigma_x^2$ is variance in additive fitness measured at the present time, $t = 0$ (the ‘Fisher term’), and μm_1 is the genomic mutation rate times the mean effect of mutations on additive fitness based on present and perhaps past data (the ‘Kimura term’). Figure 6 shows that improvement of our methods over F/K projections can increase with decreasing population size. This is

because, while our methods are less accurate in smaller populations, the F/K projections are also less accurate; the amount by which the accuracy of our methods is reduced in a smaller population may be less than the amount by which F/K projections are reduced, so the improvement of our methods over F/K projections can be larger in a small population. Figure 6b shows that these improvements are highly significant and that their significance increases with time.

4. DISCUSSION

We note that when there is no mutation, i.e. when $\mu = 0$, the fitness equations (equation (2.3)) reduce to the set of relations: $\dot{\bar{x}} = \kappa_2$, $\dot{\kappa}_2 = \kappa_3$, ..., $\dot{\kappa}_j = \kappa_{j+1}$, the first of which is Fisher’s fundamental theorem of natural selection, $\dot{\bar{x}} = \kappa_2 = \sigma_x^2$ (adhering to our definition of x as additive genic fitness). This celebrated relation states that the instantaneous rate of increase of mean fitness equals the fitness variance; it is an exact but dynamically insufficient result [62]. If it were known how the variance were changing over time at the same instant, this could slightly extend our knowledge of the future trajectory of mean fitness, thereby incrementally increasing the dynamical sufficiency of Fisher’s theorem. The rate of change in fitness variance is precisely what is given by the second relation, $\dot{\kappa}_2 = \kappa_3$; this rate in turn depends on the third cumulant, whose rate of change is given by the next equation, $\dot{\kappa}_3 = \kappa_4$, and so on. In this way, the effects of the higher cumulants percolate down through the equations, ultimately extending the projected dynamics of mean fitness in a telescoping fashion. This observation suggests that predictions for the trajectory of mean fitness may be extended beyond that predicted by Fisher’s theorem, and it might be viewed as adding a degree of dynamical sufficiency to Fisher’s theorem. This is the logic behind our method for forecasting the near-future course of fitness evolution. When mutation is added, the full system is that given by equation (2.3), the first of which is analogous to Kimura’s more general equation [57]. And just as with Fisher’s theorem, equation (2.3) might be viewed as adding a degree of dynamical sufficiency to the Kimura equation.

In a similar fashion, when there is no mutation and setting $i = 1$, the phenotype equations (equation (2.6)) reduce to the set of relations: $\dot{\bar{z}} = \kappa_{1,1}$, $\dot{\kappa}_{1,1} = \kappa_{1,2}$, ..., $\dot{\kappa}_{1,j} = \kappa_{1,j+1}$, the first of which is the Price equation [40], $\dot{\bar{z}} = \kappa_{1,1} = \text{cov}(z, x)$ (adhering again to our definition of x as additive fitness): the instantaneous rate of change of a fitness-related phenotype equals the covariance between fitness and phenotype. Like Fisher’s theorem, it is an exact but dynamically insufficient result [58]. And just as the equations for the higher (one-dimensional) cumulants allow the future projection of mean fitness beyond that predicted by Fisher’s theorem, the equations for the higher two-dimensional cumulants allow the future projection of mean phenotype beyond that predicted by the Price equation. This is the logic behind our method for forecasting the near-future course of phenotype evolution. When mutation is added, the full system is that given by equation (2.6), the first of which is analogous to the

‘full’ Price equation [40,58,59]. And just as with the Price equation, the hierarchy of expressions in equation (2.6) might be viewed as adding a degree of dynamical sufficiency to the full Price equation.

Figures 3 and 4 plot two typical examples of evolutionary trajectories from fully stochastic, individual-based simulations, compared with the predicted trajectories calculated using the telescoping method described earlier. As points of comparison to classical theory, figure 3 also plots the predictions of Fisher’s ‘fundamental theorem’, as well as the related Kimura equation (extrapolated beyond their one-generation reach for illustration); figure 4 also plots the predictions of the Price equation, as well as the ‘full’ Price equation (also extrapolated beyond their one-generation reach for illustration).

In principle, the telescoping method we describe can be used to predict not only the future trajectory of mean fitness or mean phenotype but also the future dynamics of the full fitness or phenotype distribution: equation (3.9) gives not only mean fitness or mean phenotype at time τ , but it also gives the higher cumulants at time τ . We have not exploited this fact in the current paper. It may be desirable, for example, to predict not only the future adaptedness of a population (indicated by its mean fitness) but also its future ‘adaptive capacity’ (or adaptation rate); this latter quantity is given by $\kappa_2(\tau)$ in equation (3.9). To project the future course of evolution for the entire phenotype marginal distribution, the general form of equation (3.9) may be written compactly as $\kappa_{i,j}(\tau) = \kappa_{i,j}(\tau-1) + \kappa_{i,j+1}(\tau-1) + \mu m_{i,j}(\tau-1)$ for all $i \geq 1$ and $j \geq 0$. This allows projection of the $\kappa_{i,0}$, for all $i \geq 1$, which uniquely projects the future dynamics of the full marginal distribution of the phenotype.

Our analyses and statistical models implicitly assume infinite population size, a fact that is reflected by the absence of a parameter N for population size. We suspect that this assumption is quite weak for our purposes, because of the timescales involved. This suspicion is corroborated by the observations (i) that predictions based on the continuous (infinite N) approximation appear to work quite well with simulated data from finite N populations, and (ii) the improvement over F/K projections is actually slightly greater for a smaller population size (figure 6). The most significant error introduced by the infinite N assumption comes about because, in an infinite population, all possible fitness variants—even variants with very high fitness—are present at all times, even if at miniscule frequency. When a model based on this assumption is applied to finite populations, it allows for the growth of minute fractions of individuals, which is of course a biological absurdity. The most significant consequence of this absurdity is that high fitness variants can grow, under the infinite population assumption, even when they do not yet exist in a finite population. (Consequently, it has been shown [51] that the infinite population assumption leads to unboundedly accelerating fitness increase, whereas a finite population with the same parameters shows a linear fitness increase.) However, the error introduced by the infinite population assumption typically takes a considerable amount of time to affect predictions.

To illustrate, we imagine a population of 10^5 individuals whose fitnesses follow a normal distribution with mean one and standard deviation 0.05. A fitness variant with selective advantage 4 pre-exists in the infinite population model, but it exists at such a low frequency that it would require 797 generations for that variant to reach a frequency of 10^{-5} , which would correspond to a single individual in our finite population of 10^5 . This number of generations is much greater than the current predictive reach of our model. We are nevertheless concerned that the error introduced by the infinite population assumption may limit the predictive reach of our methods, and we are currently exploring application of previous work that remedies these limitations: (i) employing a right-truncated fitness distribution with a ‘stochastic edge’, as outlined by Rouzine *et al.* [50,53,54], and (ii) employing a tuning technique for evolutionary models introduced by Hallatschek [52] that absorbs the stochastic effects of finite N .

In addition to accounting for the effects of finite N , future work will address the related issues of how to determine an appropriate time interval, given the real data. Furthermore, we plan to perform a more rigorous evaluation of the predictability provided by our methods as well as a comprehensive sensitivity analysis. Finally, we are planning to test this theory with evolving *in vitro* populations of *Escherichia coli*. We have already completed pilot experiments (K. Sprouffske, C. Gentile, P. Sniegowski, & P. Gerrish 2012, unpublished data), which indicate that a strong signature of predictability can be obtained using the methods outlined here with a modest amount of experimental work.

As mentioned in the Introduction, the present study offers a rudimentary beginning to the problem of predicting evolution. While the focus of this study has been the evolution of fitness or a fitness-related phenotype, most applications of this work will be more complex and will seek ways to predict the evolution of multiple phenotypes or genotypic features. The cumulant expansion presented here is easily extended to n phenotypic or genotypic dimensions. Such an extension could potentially predict the trajectories and perhaps fates of n existing polymorphisms; however, its applicability to de novo phenotypes or genotypic features is questionable and a subject for further study.

Notably, our results carry no asymptotic requirements and apply equally well to recently perturbed populations that are far from dynamic equilibrium. As time progresses, our model also allows for changes in the mutational landscape through the dynamic mutational term; this term can accommodate epistatic contingencies as long as they admit of statistical characterization. Eventually, however, the dynamics of mutational effects and contingencies will surely resist further prospective characterization, a fact that will limit the timescale of applicability.

We thank Warren Ewens, Michael Lässig, Nico Stollenwerk, Jorge Carneiro and Isabel Gordo for helpful discussions at various stages, and three anonymous reviewers for helpful comments. Much of this research was developed, thanks to fertile environments provided by two institutes: the Kavli Institute for Theoretical Physics in Santa Barbara, CA,

USA (2011 Microbial and Viral Evolution workshop), and the Instituto Gulbenkian de Ciências in Oeiras, Portugal. We acknowledge support from the US National Institutes of Health grants R01 GM079843-01 (P.J.G./P.D.S.) and ARRA R01GM079483-02S1 (P.J.G./P.D.S.), and from European Commission grant FP7 231807 (P.J.G.).

REFERENCES

- Mayr, E. 1988 *Toward a new philosophy of biology: observations of an evolutionist*. Cambridge, MA: Harvard University Press.
- Scriven, M. 1959 Explanation and prediction in evolutionary theory. *Science* **130**, 477–482. (doi:10.1126/science.130.3374.477)
- Gould, S. J. 1989 *Wonderful life: the Burgess Shale and the nature of history*, 1st edn. New York, NY: W.W. Norton.
- Simpson, G. G. 1950 Evolutionary determinism and the fossil record. *Sci. Mon.* **71**, 262–267.
- Conway Morris, S. 1998 *The crucible of creation: the Burgess Shale and the rise of animals*. Oxford, UK: University Press.
- Monod, J. & Wainhouse, A. 1972 *Chance and necessity: an essay on the natural philosophy of modern biology*. London, UK: Collins.
- Grant, P. R. & Grant, B. R. 2002 Unpredictable evolution in a 30-year study of Darwin's finches. *Science* **296**, 707–711. (doi:10.1126/science.1070315)
- Blount, Z. D., Borland, C. Z. & Lenski, R. E. 2008 Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **105**, 7899–7906. (doi:10.1073/pnas.0803151105)
- Woods, R. J., Barrick, J. E., Cooper, T. F., Shrestha, U., Kauth, M. R. & Lenski, R. E. 2011 Second-order selection for evolvability in a large *Escherichia coli* population. *Science* **331**, 1433–1436. (doi:10.1126/science.1198914)
- Yedid, G. & Bell, G. 2002 Macroevolution simulated with autonomously replicating computer programs. *Nature* **420**, 810–812. (doi:10.1038/nature01151)
- Gerrish, P. 2002 Computation biology: evolution plays dice. *Nature* **420**, 756–757. (doi:10.1038/420756a)
- Bull, J. J. & Molineux, I. J. 2008 Predicting evolution from genomics: experimental evolution of bacteriophage T7. *Heredity* **100**, 453–463. (doi:10.1038/sj.hdy.6801087)
- Nakatsu, C. H., Korona, R., Lenski, R. E., de Bruijn, F. J., Marsh, T. L. & Forney, L. J. 1998 Parallel and divergent genotypic evolution in experimental populations of *Ralstonia* sp. *J. Bacteriol.* **180**, 4325–4331.
- Wichman, H. A., Badgett, M. R., Scott, L. A., Boulianne, C. M. & Bull, J. J. 1999 Different trajectories of parallel evolution during viral adaptation. *Science* **285**, 422–424. (doi:10.1126/science.285.5426.422)
- Cunningham, C. W., Jeng, K., Husti, J., Badgett, M., Molineux, I. J., Hillis, D. M. & Bull, J. J. 1997 Parallel molecular evolution of deletions and nonsense mutations in bacteriophage T7. *Mol. Biol. Evol.* **14**, 113–116.
- Bull, J. J., Badgett, M. R., Wichman, H. A., Huelsenbeck, J. P., Hillis, D. M., Gulati, A., Ho, C. & Molineux, I. J. 1997 Exceptional convergent evolution in a virus. *Genetics* **147**, 1497–1507.
- Wood, T. E., Burke, J. M. & Rieseberg, L. H. 2005 Parallel genotypic adaptation: when evolution repeats itself. *Genetica* **123**, 157–170. (doi:10.1007/s10709-003-2738-9)
- Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A. & Cresko, W. A. 2010 Population genomics of parallel adaptation in threespine stickleback using sequenced RAD Tags. *PLoS Genet.* **6**, e1000862. (doi:10.1371/journal.pgen.1000862.t003)
- Bairoch, A. & Murzin, A. 1997 Sequences and topology: predicting evolution. *Curr. Opin. Struct. Biol.* **7**, 367–368. (doi:10.1016/S0959-440X(97)80053-X)
- Hall, B. G. 2002 Predicting evolution by *in vitro* evolution requires determining evolutionary pathways. *Antimicrob. Agents Chemother.* **46**, 3035–3038. (doi:10.1128/AAC.46.9.3035-3038.2002)
- Reeve, J. P. 2000 Predicting long-term response to selection. *Genet. Res.* **75**, 83–94. (doi:10.1017/S0016672399004140)
- Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J. & Fitch, W. M. 1999 Predicting the evolution of human influenza A. *Science* **286**, 1921–1925. (doi:10.1126/science.286.5446.1921)
- Plotkin, J. B., Dushoff, J. & Levin, S. A. 2002 Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. *Proc. Natl Acad. Sci. USA* **99**, 6263–6268. (doi:10.1073/pnas.082110799)
- Huang, J.-W. & Yang, J.-M. 2011 Changed epitopes drive the antigenic drift for influenza A (H3N2) viruses. *BMC Bioinform.* **1**(Suppl. 12), S31. (doi:10.1186/1471-2105-12-S1-S31)
- Liao, Y. C., Lee, M. S., Ko, C. Y. & Hsiung, C. A. 2008 Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus. *Bioinformatics* **24**, 505–512. (doi:10.1093/bioinformatics/btm638)
- Mustonen, V., Kinney, J., Callan Jr, C. G. & Lässig, M. 2008 Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proc. Natl Acad. Sci. USA* **105**, 12 376–12 381. (doi:10.1073/pnas.0805909105)
- Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. & Tans, S. J. 2007 Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445**, 383–386. (doi:10.1038/nature05451)
- Rokyta, D. R., Joyce, P., Caudle, S. B. & Wichman, H. A. 2005 An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. *Nat. Genet.* **37**, 441–444. (doi:10.1038/ng1535)
- Macken, C. A. & Perelson, A. S. 1989 Protein evolution on rugged landscapes. *Proc. Natl Acad. Sci. USA* **86**, 6191–6195. (doi:10.1073/pnas.86.16.6191)
- Perelson, A. S. & Macken, C. A. 1995 Protein evolution on partially correlated landscapes. *Proc. Natl Acad. Sci. USA* **92**, 9657–9661. (doi:10.1073/pnas.92.21.9657)
- Beardmore, R. E., Gudelj, I., Lipson, D. A. & Hurst, L. D. 2011 Metabolic trade-offs and the maintenance of the fit-test and the flattest. *Nature* **472**, 342–346. (doi:10.1038/nature09905)
- Joyce, P., Rokyta, D. R., Beisel, C. J. & Orr, H. A. 2008 A general value theory model for the adaptation of DNA sequences under strong selection and weak mutation. *Genetics* **180**, 1627–1643. (doi:10.1534/genetics.108.088716)
- Orr, H. A. 1998 The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution* **52**, 935–949. (doi:10.2307/2411226)
- Rokyta, D. R., Beisel, C. J. & Joyce, P. 2006 Properties of adaptive walks on uncorrelated landscapes under strong selection and weak mutation. *J. Theor. Biol.* **243**, 114–120. (doi:10.1016/j.jtbi.2006.06.008)
- Rokyta, D. R., Beisel, C. J., Joyce, P., Ferris, M. T., Burch, C. L. & Wichman, H. A. 2008 Beneficial fitness effects are not exponential for two viruses. *J. Mol. Evol.* **67**, 368–376. (doi:10.1007/s00239-008-9153-x)
- Rozen, D. E., de Visser, J. A. G. M. & Gerrish, P. J. 2002 Fitness effects of fixed beneficial mutations in microbial

- populations. *Curr. Biol.* **12**, 1040–1045. (doi:10.1016/S0960-9822(02)00896-5)
- 37 Eigen, M. 1971 Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58**, 465–523. (doi:10.1007/BF00623322)
 - 38 Domingo, E. *et al.* 1985 The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: biological relevance: a review. *Gene* **40**, 1–8. (doi:10.1016/0378-1119(85)90017-4)
 - 39 Fisher, R. A. 1930 *The genetical theory of natural selection*. Oxford, UK: Clarendon Press.
 - 40 Price, G. R. 1970 Selection and covariance. *Nature* **227**, 520–521. (doi:10.1038/227520a0)
 - 41 Bürger, R. 1991 Moments, cumulants, and polygenic dynamics. *J. Math. Biol.* **30**, 199–213. (doi:10.1007/BF00160336)
 - 42 Jones, A. G., Arnold, S. J. & Bürger, R. 2007 The mutation matrix and the evolution of evolvability. *Evolution* **61**, 727–745. (doi:10.1111/j.1558-5646.2007.00071.x)
 - 43 Turelli, M. 1990 Dynamics of polygenic characters under selection. *Theor. Popul. Biol.* **38**, 1–57. (doi:10.1016/0040-5809(90)90002-D)
 - 44 Mustonen, V. & Lässig, M. 2009 From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends Genet.* **25**, 111–119. (doi:10.1016/j.tig.2009.01.002)
 - 45 Mustonen, V. & Lässig, M. 2010 Fitness flux and ubiquity of adaptive evolution. *Proc. Natl Acad. Sci. USA* **107**, 4248–4253. (doi:10.1073/pnas.0907953107)
 - 46 Page, K. & Nowak, M. 2002 Unifying evolutionary dynamics. *J. Theor. Biol.* **219**, 93–98.
 - 47 Briones, C. & Domingo, E. 2008 Minority report: hidden memory genomes in HIV-1 quasispecies and possible clinical implications. *AIDS Rev.* **10**, 93–109.
 - 48 Domingo, E., Baranowski, E., Ruiz-Jarabo, C. M., Martin-Hernandez, A. M., Saiz, J. C. & Escarmis, C. 1998 Quasispecies structure and persistence of RNA viruses. *Emerg. Infect. Dis.* **4**, 521–527. (doi:10.3201/eid0404.980402)
 - 49 Mateu, M. G., Martinez, M. A., Rocha, E., Andreu, D., Parejo, J., Giralt, E., Sobrino, F. & Domingo, E. 1989 Implications of a quasispecies genome structure: effect of frequent, naturally occurring amino acid substitutions on the antigenicity of foot-and-mouth disease virus. *Proc. Natl Acad. Sci. USA* **86**, 5883–5887. (doi:10.1073/pnas.86.15.5883)
 - 50 Rouzine, I. M., Wakeley, J. & Coffin, J. M. 2003 The solitary wave of asexual evolution. *Proc. Natl Acad. Sci. USA* **100**, 587–592. (doi:10.1073/pnas.242719299)
 - 51 Tsimring, L. S., Levine, H. & Kessler, D. A. 1996 RNA virus evolution via a fitness-space model. *Phys. Rev. Lett.* **76**, 4440–4443. (doi:10.1103/PhysRevLett.76.4440)
 - 52 Hallatschek, O. 2011 The noisy edge of traveling waves. *Proc. Natl Acad. Sci. USA* **108**, 1783–1787. (doi:10.1073/pnas.1013529108)
 - 53 Brunet, E., Rouzine, I. M. & Wilke, C. O. 2008 The stochastic edge in adaptive evolution. *Genetics* **179**, 603–620. (doi:10.1534/genetics.107.079319)
 - 54 Rouzine, I. M., Brunet, E. & Wilke, C. O. 2008 The traveling-wave approach to asexual evolution: Muller's ratchet and speed of adaptation. *Theor. Popul. Biol.* **73**, 24–46. (doi:10.1016/j.tpb.2007.10.004)
 - 55 Gerrish, P. J. & Lenski, R. E. 1998 The fate of competing beneficial mutations in an asexual population. *Genetica* **102/103**, 127–144. (doi:10.1023/A:1017067816551)
 - 56 Perfeito, L., Fernandes, L., Mota, C. & Gordo, I. 2007 Adaptive mutations in bacteria: high rate and small effects. *Science* **317**, 813–815. (doi:10.1126/science.1142284)
 - 57 Kimura, M. 1958 On the change of population fitness by natural selection. *Heredity* **12**, 145–167. (doi:10.1038/hdy.1958.21)
 - 58 Gardner, A. 2008 The Price equation. *Curr. Biol.* **18**, R198–202. (doi:10.1016/j.cub.2008.01.005)
 - 59 Frank, S. A. 1995 George Price's contributions to evolutionary genetics. *J. Theor. Biol.* **175**, 373–388. (doi:10.1006/jtbi.1995.0148)
 - 60 Sniegowski, P. D. & Gerrish, P. J. 2010 Beneficial mutations and the dynamics of adaptation in asexual populations. *Phil. Trans. R. Soc. B* **365**, 1255–1263. (doi:10.1098/rstb.2009.0290)
 - 61 Diebold, F. X. & Mariano, R. S. 1995 Comparing predictive accuracy. *J. Bus. Econ. Stat.* **13**, 253–263. (doi:10.2307/1392185)
 - 62 Frank, S. A. & Slatkin, M. 1992 Fisher's fundamental theorem of natural selection. *Trends Ecol. Evol.* **7**, 92–95. (doi:10.1016/0169-5347(92)90248-a)