nature genetics

An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus

Darin R Rokyta[1–3], Paul Joyce[2–5], S Brian Caudle[1,6] & Holly A Wichman[1–3]

The primary impediment to formulating a general theory for adaptive evolution has been the unknown distribution of fitness effects for new beneficial mutations[1]. By applying extreme value theory[2], Gillespie circumvented this issue in his mutational landscape model for the adaptation of DNA sequences[3–5], and Orr recently extended Gillespie's model[1,6], generating testable predictions regarding the course of adaptive evolution. Here we provide the first empirical examination of this model, using a single-stranded DNA bacteriophage related to ϕX174, and find that our data are consistent with Orr's predictions, provided that the model is adjusted to incorporate mutation bias. Orr's work suggests that there may be generalities in adaptive molecular evolution that transcend the biological details of a system, but we show that for the model to be useful as a predictive or inferential tool, some adjustments for the biology of the system will be necessary.

Evolution by natural selection is one of the major generalizations in biology, yet a framework for a quantitative, rigorous study of adaptation has remained elusive. Fisher's geometric model[7] has been useful for deriving qualitative predictions[8–10], but its use requires arbitrary selection of a fitness function and mutation definition, and it assumes a continuous, unlimited phenotypic space. Gillespie's mutational landscape model[3–5] seems to be a better approximation to biological reality. It uses the discrete nature of DNA sequence space and requires only modest assumptions about the population under study. Here we test Orr's extensions to this model[6]. The results that we address concern a single step in adaptive molecular evolution but may serve as a solid foundation on which to construct a general theoretical treatment of adaptive molecular evolution.

The mutational landscape model[3–5] considers the adaptation of a large population of DNA sequences under strong selection and weak mutation[5]. Under these conditions, a single sequence will dominate the population with occasional, rapid fixations of new beneficial mutations, and the only accessible sequences will be those that differ from the current sequence by a single mutation[4]. For notation, assume that the current sequence has fitness rank i and the fittest allele has fitness rank 1. If most mutations are deleterious, then the $i - 1$ beneficial mutations will be drawn from the tail of the fitness distribution. If this distribution has an exponential tail[2], extreme value theory predicts that the fitness differences between adjacently ranked alleles will be asymptotically independent exponential random variables (**Fig. 1**). The mean of the fitness difference between the fittest allele and second fittest allele is some constant, C; all other fitness spacings are scaled versions of the first. The mean fitness difference between the second fittest and third fittest allele is $C/2$, between the third and forth is $C/3$, and so on. These spacings allow calculation of fitness effects and thus selection coefficients for mutations. By assuming that the probability of fixation is given by Haldane's $2s$ approximation[11], where s is the allele's selection coefficient, Orr used this regularity property to calculate the mean transition probability as

$$E[P_{ij}] = \frac{1}{i-1}\sum_{k=j}^{i-1}\frac{1}{k},$$

which gives the probability that the population jumps to the j^{th} fittest allele averaged over the possible assignments of selection coefficients to the $i - 1$ beneficial alleles. On average, the course of adaptive evolution for a single step depends only on the number of beneficial mutations available. From Orr's equation, the mean rank of the next allele fixed can be shown to be

$$E[j] = \frac{i+2}{4}.$$

As Orr pointed out[6], this is midway between the expected rank under perfect or 'gradient'[12] adaptation, where the fittest allele (rank 1) is always fixed, and a random choice from among the beneficial alleles, where, on average, rank $i/2$ is fixed.

To test Orr's mean transition probabilities and to assess whether the average behavior of adaptive evolution predicts the outcome for a single realization, we carried out 20 single-step adaptations from a single ancestral genotype, selecting for rapid phage replication under standard batch culturing conditions. We used an icosahedral, single-stranded DNA bacteriophage, ID11, which is related to ϕX174. Replicate populations were allowed to fix a single beneficial mutation under strong selection and weak mutation. We determined the

[1]Department of Biological Sciences, [2]Program in Bioinformatics and Computational Biology, [3]Initiative for Bioinformatics and Evolutionary Studies, [4]Department of Mathematics and [5]Department of Statistics, University of Idaho, Moscow, Idaho 83844, USA. [6]Present address: Department of Biological Sciences, Section of Integrative Biology, University of Texas, Austin, Texas 78712, USA. Correspondence should be addressed to H.A.W. (hwichman@uidaho.edu).

Published online 20 March 2005; doi:10.1038/ng1535

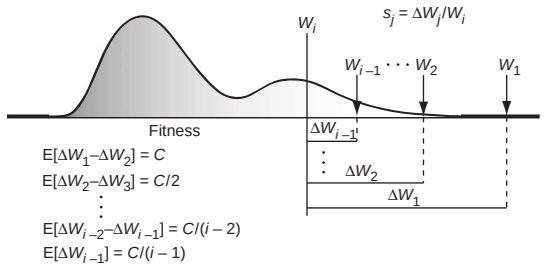

Figure 1 A schematic depiction of the extreme value theory predictions for a single step. W_j is the absolute fitness of the j^{th} fittest allele, ΔW_j is its fitness effect and s_j is its selection coefficient. The current wild-type is the i^{th} fittest. Extreme value theory predicts that the differences in fitness effects of adjacent alleles are independent exponential random variables.

identity of the substitution by whole-genome sequencing of each final population. To estimate the ranks of alleles, we determined the fitness of each unique mutation through standard fitness assays[13]. All ten observed substitutions were nonsynonymous (**Table 1**). Two different nucleotide substitutions at the same site generated the same amino acid replacement and were treated as a single allele; hence, we considered nine different beneficial mutations.

Orr's model predicts the proportion of times the population should fix its available beneficial alleles, provided the number of beneficial alleles is known. It is unlikely that we have observed all the beneficial mutations, but under Orr's model, the observed number is the maximum likelihood estimate for the true value. To assess the goodness of fit for Orr's predictions, we used a multinomial likelihood ratio goodness-of-fit test, which suggests that Orr's model does not adequately explain our data ($P = 0.10$, $-2\mathrm{Ln}\Lambda = 13.50$, degrees of freedom (d.f.) = 8; parametric bootstrapping $P = 0.10$; **Fig. 2a**).

Under Orr's model, the fittest mutation should be the most frequently substituted, the second fittest mutation should be the second most frequently substituted, and so on. In our data, the two most frequent mutations were both C→T transitions, whereas the fittest genotype arose through a G→T transversion. The most salient discrepancy between Orr's predictions and our data set is the number of times the fittest allele was fixed (**Fig. 2a**). It should have been fixed in ~6 of the 20 replicates, but it fixed only once. Transitions are known to occur at a higher rate than transversions; therefore, mutation bias is an obvious explanation for this discrepancy. As Orr pointed out[1], his predictions deal with the average behavior of adaptation. If there is no correlation between the fitness effect of a mutation and the rate at which it arises, then mutation bias will not alter the predictions of the model, because mutation rates are essentially averaged out. For specific cases, it may be acceptable to consider the average behavior of fitness effects, but perhaps differences in mutation rates cannot be ignored. We therefore adjusted Orr's model to allow the beneficial alleles to have different mutation rates:

$$E[P_{ij}] = \frac{\mu_j}{\sum_{k=1}^{i-1} \mu_k} \sum_{k=j}^{i-1} \frac{1}{k},$$

where

$$\bar{\mu}_k = \frac{1}{k} \sum_{i=1}^{k} \mu_i.$$

The mutation rate for allele j is denoted μ_j. This adjustment works for both absolute mutation rates and relative rates, because any scaling factor cancels. Using 13 additional unpublished phage sequences, we estimated the relative mutation rates for a general time-reversible model of DNA sequence evolution. The third fittest allele arose through two different single-nucleotide changes and could have arisen through an additional, unobserved mutation; therefore, it was assigned the sum of these rates. The fittest allele also had an alternate single mutation pathway, and its rate was adjusted accordingly. All other amino acid changes could only be achieved through a single-nucleotide substitution. The model performed well ($P = 0.49$, $-2\mathrm{Ln}\Lambda = 7.42$, d.f. = 8; parametric bootstrapping $P = 0.49$; **Fig. 2b**). Our data are 21 times more likely under the mutation-adjusted version of Orr's model than under the original model. One caveat, however, is that we are no longer guaranteed that the observed number of mutations is the maximum likelihood estimate.

Even the mutation-adjusted version of Orr's model incorporates few of the biological and experimental details of our system. Therefore, it seems germane to assess the explanatory power lost in this simplification. Extending work by Wahl *et al.*[14,15], we derived fixation probabilities explicitly for our experimental system and protocol, incorporating population bottlenecks, growth rates and selection coefficients. Assuming that the fixation time for an allele is exponentially distributed with rate $\mu_j\Pi(s_j)$, where μ_j is the mutation rate for allele j and $\Pi(s_j)$ is its fixation probability, the probability that an allele is the next one fixed is the ratio of its rate to the sum of the rates for all alleles[5]. Using the estimated selection coefficients (**Table 1**) and our phylogenetic estimates for mutation rates, this model provided a better fit than the original Orr model ($P = 0.67$, $-2\mathrm{Ln}\Lambda = 5.76$, d.f. = 8; parametric bootstrapping $P = 0.78$; **Fig. 2c**). The mutation-adjusted Wahl model and the mutation-adjusted Orr model both incorporate the effects of selection and mutation bias, and both adequately explain our data. Although Orr's model is based on Haldane's $2s$ fixation probability, which is a small s approximation,

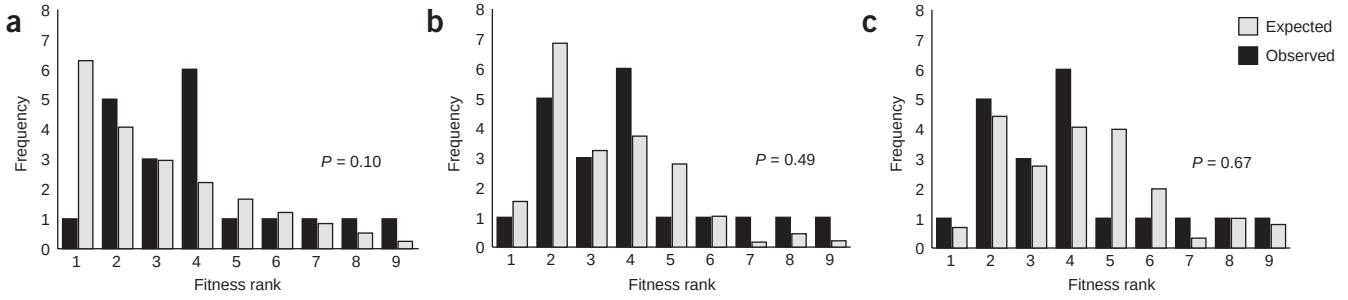

Figure 2 A comparison of the observed data with the expectations under Orr's model and the mutation-adjusted models. (**a**) Orr's model. (**b**) Mutation-adjusted Orr model using a six rate mutation model. (**c**) Mutation-adjusted Wahl model incorporating selection coefficients, the experimental protocol and differences in mutation rates. The P values indicate goodness of fit.

it requires only that fixation probabilities be proportional to *s*. The mutation-adjusted Wahl model is a better fit to the data, most likely because it incorporates the defining features of our system: severe population bottlenecks and large selection coefficients.

Orr's approach to modeling adaptation has resulted in general predictions for adaptive evolution that are relatively parameter-free; they depend only on the number of beneficial mutations. This theory represents a first attempt to propose general properties of adaptive walks. It may allow predictions about the magnitudes of fitness gains that will occur during adaptive evolution, something that could be applied in many contexts. Additionally, the model could potentially serve as a tool for statistical inference. The single parameter in Orr's mean transition probability formula, the fitness rank of the wild type, is of central importance in the study of molecular evolution and adaptive walks but has never been measured. Our work shows that his model holds for empirical data, except that differences in mutation rates for alleles must be incorporated to be applicable to a single realization of the adaptive process.

Many real-time outcomes of natural selection, such as the evolution of drug-resistant bacteria and viruses and the evolution of pest resistance to insecticides and herbicides, have been to our detriment. Understanding adaptive evolution offers potential solutions. For example, understanding the biological details of evolution has led to protocols for treating human immunodeficiency virus that monitor evolution and circumvent drug resistance, and to mandated refugia of non–genetically modified crops aimed at delaying the evolution of insect resistance to Bt toxin genes in transgenic plants. Understanding and mimicking processes of evolution has led to 'directed evolution' in which the biotechnology industry uses enhanced natural selection to produce proteins and nucleic acids with specific functions. Although the biological details will always be important, we need to keep moving beyond details toward generalities, if such generalities exist. Orr's work suggests that they may indeed exist, and we have now taken the first step towards experimental verification of his theoretical results.

Table 1 Substitutions observed in 20 single-step adaptations

Genome position	Nucleotide substitution	Amino acid position	Amino acid substitution	Number	Fitness	s.d.	*s*
Ancestor					14.61	0.18	
2534	G→T	J20	V→L	1	20.31	0.40	0.39
3665	C→T	F355	P→S	5	20.05	0.49	0.37
3850	G→A/T	F416	M→I	3	19.45	0.59	0.33
2520	C→T	J15	A→V	6	19.29	0.66	0.32
3543	C→T	F314	A→V	1	19.13	0.61	0.31
3857	A→G	F419	T→A	1	19.04	0.52	0.30
2609	G→T	F3	V→F	1	17.56	0.43	0.20
3567	A→G	F322	N→S	1	16.74	0.32	0.15
3864	A→G	F421	D→G	1	16.22	0.70	0.11

Fitness is measured as the $\log_2$ increase in phage number per h. The amino acid position gives the gene name followed by residue number: gene J is a single-stranded DNA binding protein, and gene F is the major capsid protein. Genome positions correspond to the published G4 sequence.

METHODS

Strains and culture conditions. The phage used in this work, ID11, was isolated from the University of Idaho barnyards. Its single-stranded DNA genome is 5,577 nt long and encodes 11 genes. Like ϕX174, it is a member of the family Microviridae, differs from the bacteriophage G4 at ~3% of its sites and has the same genome size and map. The host for all experiments was *Escherichia coli* C. We carried out all experiments at 37 °C in Luria-Bertani broth supplemented with 2 mM $CaCl_2$.

Adaptations. Our protocol has been described elsewhere[13]. We grew hosts to a density of $\sim 10^8$ cells per ml, using a 10-ml volume in 125-ml flasks, shaking at 200 r.p.m. We added $\sim 10^4$ phage and grew them for 40 min to a population size of $\sim 10^8$. We terminated growth with chloroform and then sampled the population to begin the next flask culture. We determined phage titers by plating for each flask and used them to monitor phage fitness, which is the $\log_2$ increase in phage numbers per h. Each adaptation was initiated from an independent isolate of the same ancestral genotype. We terminated adaptations after a discernible increase in fitness (8–16 flask transfers).

Sequencing. We determined the genome sequences of the final populations for all 20 adaptations by standard methods. If the final population contained two substitutions (5 of the 20 adaptations), we sequenced earlier populations and the initial isolate to determine the timing of the substitutions. Three initial isolates contained silent substitutions; these were assumed to have no effect on fitness or the allele fixed and were not present in the isolates used for determining fitness. Two final populations had fixed two nonsynonymous substitutions. In both cases, only one was present at the midpoint population and was selected as the first substitution. We sequenced isolates for fitness assays as described for populations.

Fitness assays. Our fitness assay protocol has been described previously[13]. We measured fitness, the $\log_2$ increase in the phage population per h, under our adaptation conditions. For each unique mutation, we obtained a sequence-confirmed genetic isolate by plating, picking a plaque and sequencing it to assure that only the mutation of interest was obtained. We measured fitness ten times for each mutation.

Statistics. We obtained fitnesses by averaging the results of ten experiments for each of the nine mutations. Because the variation was similar across replicates and alleles, we pooled the variances to produce an overall standard error of 0.20. All fitnesses were more than one standard error apart, indicating that the observed fitness ranking is fairly accurate. But we can account for potential error in the fitness rankings by considering all possible rankings consistent with the data; the results are similar.

To test goodness of fit, we used a multinomial likelihood ratio test. We tested a multinomial model with the proportions given by Orr's mean transition probability formula, assuming that the number of beneficial alleles is nine, against a model with class proportions estimated from the data. The test assumes that twice the difference in log likelihoods follows a χ^2 with d.f. given by 1 minus the number of classes. We verified *P* values through parametric bootstrapping with 10,000 replicates.

Although the true number of beneficial alleles is probably greater than the nine observed alleles, the use of the χ^2 with 8 d.f. is conservative. If we knew the true number of beneficial alleles and incorporated this into our analysis, then both the mean and variance of the log likelihood ratio statistic would increase under the null hypothesis, which would increase our *P* value.

To estimate the mutation rates for our phage, we estimated, by maximum likelihood, the phylogeny of our phage and 13 unpublished natural isolate sequences within 10% uncorrected sequence distance from our ancestor using PAUP* 4.0 (ref. 16). We aligned whole genomes using ClustalW[17] and excluded gaps. We estimated initial parameters on a neighbor-joining phylogeny. We used the GTR+I+G model of sequence evolution, as selected by DT_ModSel[18]. Final parameter values were estimated from the ML tree. Relative rates estimates were as follows: A↔C, 0.78; A↔G, 4.31; A↔T, 1.01; C↔G, 0.23; C↔T, 8.50; G↔T, 1.0.

Derivation of fixation probabilities. Details of the derivation procedure are given in **Supplementary Note** online. We assumed exponential growth followed by a bottleneck to the original population size. Following Wahl *et al.*[15], we defined *r* as the growth rate, *D* as the dilution factor or the proportion of the

population surviving the bottleneck, s as the selection coefficient and τ as the time between bottlenecks. Wahl *et al.*[14] considered the time at which a mutant arises as a random variable and calculated a small s fixation probability. Below is the formula without assuming a small s. First solve for y in $1-y=e^{-Dye^{r(1+s)\tau}}$, and then substitute y into

$$\prod(s)=(Dy)^{\frac{1}{s+1}}\Gamma\left(\frac{s}{s+1}\right)P\left(\frac{Dy}{\alpha}\le Y\le 1\right)+e^{-\alpha-r\tau}\left[1+\alpha\frac{s+1}{s}\right]+1-e^{-Dy}\left[1+\frac{s+1}{s}Dy\right],$$

where $\alpha=Dye^{r(1+s)\tau}$ and Y is a gamma-distributed random variable with scale parameter given by α and shape parameter

$$\frac{s}{s+1}+1.$$

This means that

$$P\left(\frac{Dy}{\alpha}\le Y\le 1\right)=\frac{1}{\Gamma(s/(s+1)+1)}\int_{Dy/\alpha}^{1}\alpha(\alpha x)^{\frac{s}{s+1}}e^{-\alpha x}dx$$

and $\Gamma(t)=\int_0^\infty x^{t-1}e^{-x}dx$.

GenBank accession numbers. Phage ID11, AY751298; bacteriophage G4, AF454431.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank J.J. Bull, Z. Abdo, H.A. Orr and L. Wahl for discussions and comments. This work was supported by a grant from the US National Institutes of Health.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 27 October 2004; accepted 8 February 2005
Published online at http://www.nature.com/naturegenetics/

1. Orr, H.A. The distribution of fitness effects among beneficial mutations. *Genetics* **163**, 1519–1526 (2003).
2. Gumbel, E.J. *Statistics of Extremes* (Columbia University Press, New York, 1958).
3. Gillespie, J.H. A simple stochastic gene substitution model. *Theor. Popul. Biol.* **23**, 202–215 (1983).
4. Gillespie, J.H. Molecular evolution over the mutational landscape. *Evolution* **38**, 1116–1129 (1984).
5. Gillespie, J.H. *The Causes of Molecular Evolution* (Oxford University Press, New York, 1991).
6. Orr, H.A. The population genetics of adaptation: the adaptation of DNA sequences. *Evolution* **56**, 1317–1330 (2002).
7. Fisher, R.A. *The Genetical Theory of Natural Selection* (Oxford University Press, Oxford, UK, 1930).
8. Orr, H.A. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution* **52**, 935–949 (1998).
9. Orr, H.A. Adaptation and the cost of complexity. *Evolution* **54**, 13–20 (2000).
10. Hartl, D.L. & Taubes, C.H. Towards a theory of evolutionary adaptation. *Genetica* **102/103**, 525–533 (1998).
11. Haldane, J.B.S. A mathematical theory of natural and artificial selection. V. selection and mutation. *Proc. Camb. Philos. Soc.* **28**, 838–844 (1927).
12. Orr, H.A. A minimum on the mean number of steps taken in adaptive walks. *J. Theor. Biol.* **220**, 241–247 (2003).
13. Rokyta, D., Badgett, M.R., Molineux, I.J. & Bull, J.J. Experimental genomic evolution: extensive compensation for loss of DNA ligase activity in a virus. *Mol. Biol. Evol.* **19**, 230–238 (2002).
14. Wahl, L.M., Gerrish, P.J. & Saika-Voivod, I. Evaluating the impact of population bottlenecks in experimental evolution. *Genetics* **162**, 961–971 (2002).
15. Wahl, L.M. & Gerrish, P.J. The probability that beneficial mutations are lost in populations with periodic bottlenecks. *Evolution* **55**, 2606–2610 (2001).
16. Swofford, D.L. *Phylogenetic Analysis using Parsimony* (PAUP*)* ver. 4.0. (Sinauer Associates, Sunderland, Massachusetts, 1998).
17. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
18. Minin, V., Abdo, Z., Joyce, P. & Sullivan, J. Performance-based selection of likelihood models for phylogeny estimation. *Syst. Biol.* **52**, 1–10 (2003).