

# Mutation-Biased Adaptation in a Protein NK Model

Arlin Stoltzfus

Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute; and Biochemical Science Division, National Institute of Standards and Technology, Gaithersburg, Maryland

Evolutionary trends responsible for systematic differences in genome and proteome composition have been attributed to GC:AT mutation bias in the context of neutral evolution or to selection acting on genome composition. A possibility that has been ignored, presumably because it is part of neither the Modern Synthesis nor the Neutral Theory, is that mutation may impose a directional bias on adaptation. This possibility is explored here with simulations of the effect of a GC:AT bias on amino acid composition during adaptive walks on an abstract protein fitness landscape called an “NK” model. The results indicate that adaptation does not preclude mutation-biased evolution. In the complete absence of neutral evolution, a modest GC:AT bias of realistic magnitude can displace the trajectory of adaptation in a mutationally favored direction, to such a degree that amino acid composition is biased substantially and persistently. Thus, mutational explanations for evolved patterns need not presuppose neutral evolution.

## Introduction

Although many early evolutionists (e.g., Mivart 1871) supposed that “internal” mutational-developmental tendencies influence the course of evolution, in the mid-20th century, advocates of the neo-Darwinian “Modern Synthesis” theory argued for a strict externalist position in which selection is “the only direction-giving factor in evolution” (Mayr 1980, p. 3; for discussion, see Stoltzfus 2006). The population-genetical rationale for this position, originally given by Fisher (1930) and Haldane (1932), was that mutation rates are too small to overcome the opposing pressure of selection (for discussion, see Yampolsky and Stoltzfus 2001). This “opposing pressures” argument was interpreted very broadly to exclude variation as a factor shaping the course or direction of evolution, and thus to eliminate all rival theories, leaving neo-Darwinism as “the sole surviving theory” (Fisher 1930, p. 21).

Yet, starting with the early work of Sueoka (1962) and Freese (1962), molecular evolutionists have identified mutation bias as a direction-giving factor, invoking the effects of deletion:insertion bias (Petrov and Hartl 1998), strand-specific nucleotide biases (Beletskii and Bhagwat 1996), CpG bias (Fryxell and Zuckerkandl 2000), and GC:AT bias (Lobry 1997; Singer and Hickey 2000). The hypothesis of mutation-biased molecular evolution is often treated as a case of “neutral” evolution (Sueoka 1988; Knight et al. 2001) or associated with changes at the so-called “unconstrained” or “neutral” sites (Wolfe 1991; Lafay et al. 1999), which makes the hypothesis superficially consistent with the “opposing pressures” argument, in the sense that mutation has free rein when the opposing pressure of selection is effectively absent (e.g., as in the rationale given in Maynard Smith et al. 1985, p. 282).

Yet, to apply this reasoning at all is problematic. The “opposing pressures” argument is essentially an argument against mutation as a mass-action pressure capable of driving alleles to fixation, within a theory that defines “evolution” as a more or less deterministic process of mass-action shifts in frequencies of preexisting alleles (i.e., “shifting

gene frequencies”). By contrast, in a stochastic theory in which preexisting variation is not taken for granted, the order of occurrence of individual mutational events may determine the outcome of evolution, without any involvement of mass-action mutational shifts in frequencies (Mani and Clarke 1990). When molecular evolutionists refer to evolutionary change as a “mutation-driven” process (Li 1997), presumably they mean to invoke mutation in its role as a novelty-introducing process (not as a mass-action pressure on allele frequencies), within a conception of evolution as a 2-step process of the origin of new alleles by mutation and their subsequent fixation (or loss) by drift or selection. In typical formal treatments (Kimura 1983; Bulmer 1991; Ohta 1992), this origin-fixation process occurs under “mutation-limited” conditions, often in an infinite genotypic space, in which case evolutionary change will reach a steady-state rate equal to the rate of introduction of new alleles,  $n\mu$  (where  $\mu$  is the relevant mutation rate and  $n$  is the population size; for diploids, replace  $n$  with  $2n$ ), multiplied by  $p$ , the probability of fixation faced by these new alleles.

Within this conception of evolution as an origin-fixation process, it would seem that a mutational bias in the introduction of new alleles is an immanent directional or orienting factor regardless of whether the mutations are beneficial or neutral. The bias in rates of evolution via fixation of 2 types of mutations with rates  $\mu_1$  and  $\mu_2$  would be  $\mu_1 p_1 / (\mu_2 p_2) = (\mu_1 / \mu_2)(p_1 / p_2)$ , and this shows that the mutation bias term  $\mu_1 / \mu_2$  is a factor, along with  $p_1 / p_2$ . In the case of beneficial changes,  $p \approx 2s$  (where  $s$  is a selection coefficient: Haldane 1927); thus,  $p_1 / p_2 \approx s_1 / s_2$ . Of course there is a special case in which  $\mu_1 / \mu_2$  is the only factor because  $p_1$  and  $p_2$  are identical—as in the strictly neutral case, or in the case of equivalently beneficial changes—but  $\mu_1 / \mu_2$  is a factor regardless of whether it is the only factor, and because it depends on relative rates (not the absolute magnitude of mutation “pressure”), the effect does not diminish if mutation rates are small. Thus, mutation-biased adaptation should be possible as well as mutation-biased neutral evolution.

This possibility is not merely speculative but has a clear (albeit narrow) theoretical basis and some experimental support. Yampolsky and Stoltzfus (2001) used population simulations of a 1-step model of adaptation to show that mutation bias strongly influences the choice of alternative adaptive steps and that—contrary to the “opposing pressures” argument—this occurs when mutation rate are small

Key words: mutation bias, GC content, NK model, simulation, adaptive evolution, adaptive walk.

E-mail: arlin.stoltzfus@nist.gov.

*Mol. Biol. Evol.* 23(10):1852–1862. 2006

doi:10.1093/molbev/msl064

Advance Access publication July 20, 2006

and selection coefficients are large, and when mutational and selective preferences differ in direction, for example, when  $\mu_1/\mu_2 < 1$  whereas  $s_1/s_2 > 1$ . Recently, the influence of mutation bias on adaptation was observed in the experiments of Rokytá et al. (2005), who carried out 1-step adaptive walks with a laboratory population of bacteriophages, finding that the likelihood of the observed results given an origin-fixation model of adaptive steps is increased 21-fold (relative to the model of Orr 2002, which ignores mutational effects) by taking into account mutational effects, which include an approximately 12-fold transition:transversion bias as well as a maximum 3-fold difference in the multiplicity of mutational paths to alternative amino acid states. This provides a specific example of discarding a model (Orr 2002) precisely because it ignores mutation-biased adaptation.

Thus, it is of interest to consider the possibility (discussed at length by Stoltzfus 2006) that biases or nonrandomness in the rate of origin of new variants by mutation (and more generally, by mutation and altered development) are a general cause of nonrandomness in evolution, a possibility that cuts across traditional scientific disputes over selection versus drift, the Modern Synthesis versus the Neutral Theory, and morphological versus molecular evolution.

In pursuit of a better theoretical understanding of this possibility, the influence of GC:AT-biased adaptation on amino acid composition of proteins is chosen here as a case amenable to modeling. Briefly, Sueoka (1962) and Freese (1962) first invoked systematic GC:AT biases in nucleotide mutation to account for systematic differences in DNA and protein composition. This conjecture gained credibility when Cox and Yanofsky (1967) isolated a laboratory strain of *Escherichia coli* with GC-biased mutation and when Kimura (1968) argued that, for random fixations of neutral alleles, the rate of evolution is proportional to the mutation rate. Critics argued that mutation bias cannot be the cause of protein composition changes because these changes would have systematic structural consequences (D'Onofrio et al. 1999), that is, assuming that mutation-biased evolution requires neutrality (and that neutral changes cannot have systematic structural consequences). However, genomic differences in GC content account largely for differences in genomic codon and amino acid usage, whereas the converse is not true (Singer and Hickey 2000; Knight et al. 2001). Genomic nucleotide composition, in turn, has been attributed to mutation-biased neutral evolution or to selection for an optimal nucleotide composition (Gillespie 1991, p. 85; Duret 2002), though the latter hypothesis lost much of its appeal when an anticipated relationship between the genomic GC content and thermal habitat failed to materialize (Galtier and Lobry 1997).

Using available theory, the influence of GC:AT bias can be addressed in terms of 2 extreme special cases, infinite neutral evolution and 1-step adaptation. For a model of purely neutral change in an infinite space, the equilibrium nucleotide composition is analogous to the chemical equilibrium for a simple isomerization, that is, the GC:AT ratio reaches an equilibrium value of  $1/\beta$  where  $\beta = \mu_{GC \rightarrow AT} / \mu_{AT \rightarrow GC}$ , as described by Sueoka (1988). From this, one may use the genetic code to determine the equilibrium composition of amino acids, including the ratio of amino acids with AT-rich codons (phenylalanine, tyrosine, methionine,

isoleucine, asparagine, and lysine) to those with GC-rich codons (glycine, alanine, arginine, and proline) or more simply (using the single-letter code for amino acids) the “FYMINK:GARP” ratio. In regard to anticipating the effects of such a bias during adaptation, the Yampolsky–Stoltzfus model represents a short-term case in which mutation bias may influence the choice between two 1-step adaptive paths that are mutually exclusive. In this model, the mutation bias can be interpreted as a GC:AT bias in mutation, for example, a bias between a C  $\rightarrow$  A mutation with rate  $\mu_1$  changing an ACT codon (threonine) to AAT (asparagine), and an A  $\rightarrow$  G mutation with rate  $\mu_2$  changing the same codon to GCT (alanine). The resulting bias in choice of alternative adaptive steps is roughly  $(\mu_1/\mu_2)(s_1/s_2)$ , as explained above (Yampolsky and Stoltzfus 2001).

An obvious case of interest, then, is the influence of GC:AT bias on amino acid composition during adaptive evolution in a large but finite space. To study adaptation in phenotypic dimensions requires a fitness function  $w(\phi)$  defined for every phenotype  $\phi$ , here a protein sequence. Such a function would tell us, for instance, what is the fitness of the sequence KYETLISTH, what is the selection coefficient of an L  $\rightarrow$  S change in the fifth position, and how this selection coefficient would be affected by changing the final H to R. Such a function is not currently available. Physics-based models have achieved only limited success in predicting a folded protein structure from a sequence and do not address the link from structure to activity or to fitness. An experimental approach to measuring the contours of a fitness “landscape” would be prohibitively expensive, even for a tiny local part of sequence space (e.g., for a 100-residue sequence, there are  $6.6 \times 10^9$  neighbors that differ by 3 or fewer amino acid changes) and has not been done (for a limited example, see Gregoret and Sauer 1993). In the absence of a fitness model based on physical principles or laboratory results, here I use a random field model of the “NK” type popularized by Kauffman and Levin (1987), implemented so that each of  $N$  sites in a protein interacts with  $K$  other sites.

Combining this fitness model with development, inheritance, mutation, and population genetics gives a model of evolutionary change. Here the genetic code is the developmental model linking a genotype to a protein phenotype, and population genetics is reduced to the mutation-limited case of a 2-step origin-fixation process that can be restricted to include only beneficial changes. This model is used to explore the effect of GC:AT bias on amino acid composition during adaptive walks. The results indicate that adaptation does not preclude mutation-biased evolution. In the complete absence of neutral evolution, a GC:AT bias in mutation of realistic magnitude can displace the trajectory of adaptation in a mutationally favored direction, to such a degree that amino acid composition (the FYMINK:GARP ratio) is biased substantially and persistently.

## Methods

### Model

In the model used here, an adaptive walk proceeds by a series of changes from an initial sequence (here, a random sequence) until a local optimum is reached, that is, until

there is no single-mutant neighbor with a higher fitness. The individual changes occur by a 2-step procedure in which a single mutation (e.g., A → C at position 31) is proposed, then accepted or rejected. With exceptions noted below, only mutations that increase fitness are accepted, with a probability proportional to the selection coefficient  $s = w_m/w - 1$ , where  $w_m$  is the mutant fitness and  $w$  is the parental fitness. This rule is chosen here, as in Orr’s models of 1-step adaptive walks (Orr 2002), to reflect the theoretical result of Haldane (1927) that the probability of fixation is approximately  $2s$  for a newly introduced beneficial allele.

The fitness of a protein is a function of its amino acid sequence and is defined by an “NK” model in which each of  $N$  amino acid sites interacts with  $K$  neighboring sites, so that  $K$  modulates what Kauffman and others refer to as the “ruggedness” or “roughness” of the fitness “landscape” (Kauffman 1993; Altenberg 1997). Relevant features of this type of model are illustrated with a simplified example in figure 1. In the NK model implemented here, the fitness of the protein is the sum of  $N$  fitness components. The initial sequence is drawn at random from the set of all amino acid–encoding codons; the complete set of component values, which remains fixed during an adaptive walk, is drawn from a uniform distribution bounded at 0 and 1. The fitness effect of a mutation at site  $i$  is computed by replacing the component values at site  $i$  and at  $K$  interacting sites with the appropriate values for the mutant sequence (see fig. 1 for an example).

When  $K = 0$ , the  $N$  sites are independent and the model is analogous to a nonepistatic house-of-cards model for  $N$  loci (for a comparative explanation, see Welch and Waxman 2005). As  $K$  increases, the landscape becomes more rugged, with an increasing number of local peaks that have decreasing mean fitness (Kauffman 1993). The inability of an optimization procedure to reach a global optimum due to such interactions is called “frustration.”

### Implementation

The above model is implemented in a simulation software called “PNK,” which consists of about 2,500 lines of object-oriented C++ code in 23 files, maintained in a Concurrent Versioning System repository and available from the author on request. The random number generator is the Mersenne Twister algorithm MT19937 (Matsumoto and Nishimura 1998), as implemented by Shawn Cokus. This code compiles and runs identically on Macintosh systems running OS X and on Linux systems. On an Apple PowerBook with a 1.5-GHz CPU and with 1 GB of memory, 100 walks with  $N = 100$ ,  $K = 2$ , and  $S = 20$  typically take less than a minute. Although PNK does not have hard-coded limits, physical memory limitations are important given that the size of the fitness component table (see example in fig. 1) is  $NS^{K+1}$  floating point values. The practical limit for the coding sequence (CDS) model with  $N = 100$  is  $K = 4$ .

### Analysis

The simulation software produces a simple line-based output, in which results are reported each time a mutation is accepted (e.g., the fitness, the current number of more-fit

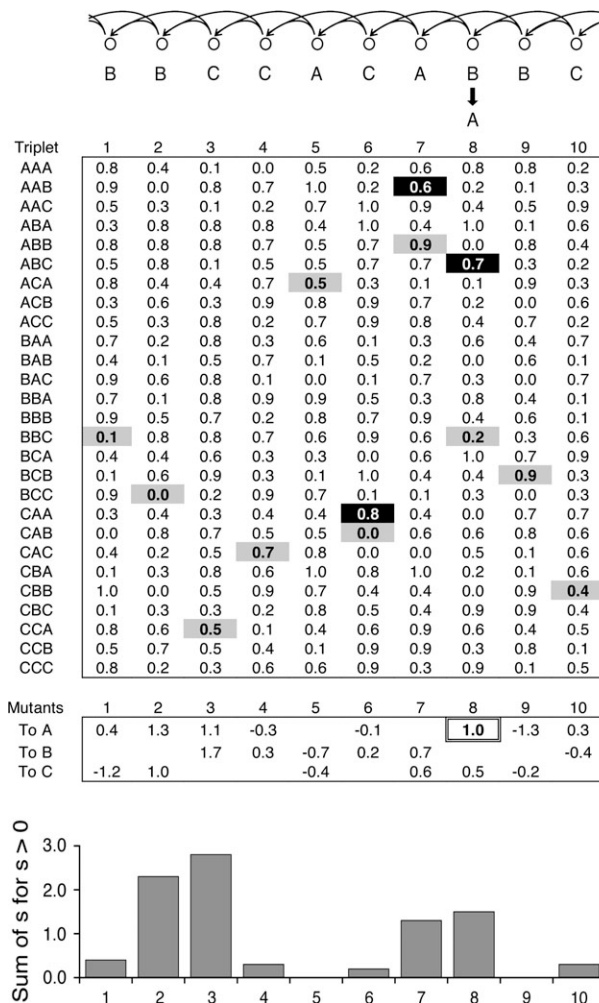


FIG. 1.—An illustration of how fitness is computed in an NK model. In this example where  $N = 10$ ,  $K = 2$ , and  $S = 3$  possible residues per site (or alleles per locus, in a genetic interpretation), we compute the fitness for the sequence BBCCACABBC, illustrate the effect of a B-to-A mutation at site 8, and consider the sum of beneficial effects of mutations at all 10 sites. Because  $K = 2$ , the fitness contribution at each site is affected by the residues at the next 2 sites, as shown by the arrows in the top figure. Thus, to specify the fitness landscape fully requires a table of fitness component values with  $N = 10$  columns and  $S^{K+1} = 27$  rows, that is, each value is indexed by a site (column) and a triplet of states (row), as shown in the upper table, which is filled with random values from 0 to 1 (a “random field” model). Given this table, the fitness for the type BBCCACABBC is 4.2, the sum of the 10 shaded values. The effect of a B-to-A mutant at site 8 is to replace 3 of these shaded values (in columns 6, 7, and 8) summing to 1.1, with 3 mutant values (values in white text on black background) summing to 2.1. Thus, the B-to-A mutation at site 8 increases fitness by 1.0 from 4.2 to 5.2. The effect of each of the 20 possible single-site mutants (i.e., 10 sites multiplied by 2 alternatives per site) is shown in the lower table, with the value for the B-to-A mutant at site 8 shown boxed. Overall, 8 mutants are deleterious and 12 are beneficial. Some sites (5 and 9) have 2 deleterious mutants (negative change in fitness), whereas others (2, 3, 7, and 8) have 2 beneficial mutants. Thus, the initial expected rate of change is nonuniform across sites, as indicated by the histogram (lower panel), which shows the sum of selection coefficients for beneficial mutants (this is one source of heterogeneity among sites; mutational effects, not shown, are another).

neighbors). In typical usage, the program is called many times, and the results parsed, using ad hoc Perl scripts that produce tabular data, which then are loaded into Excel for plotting and further analysis.

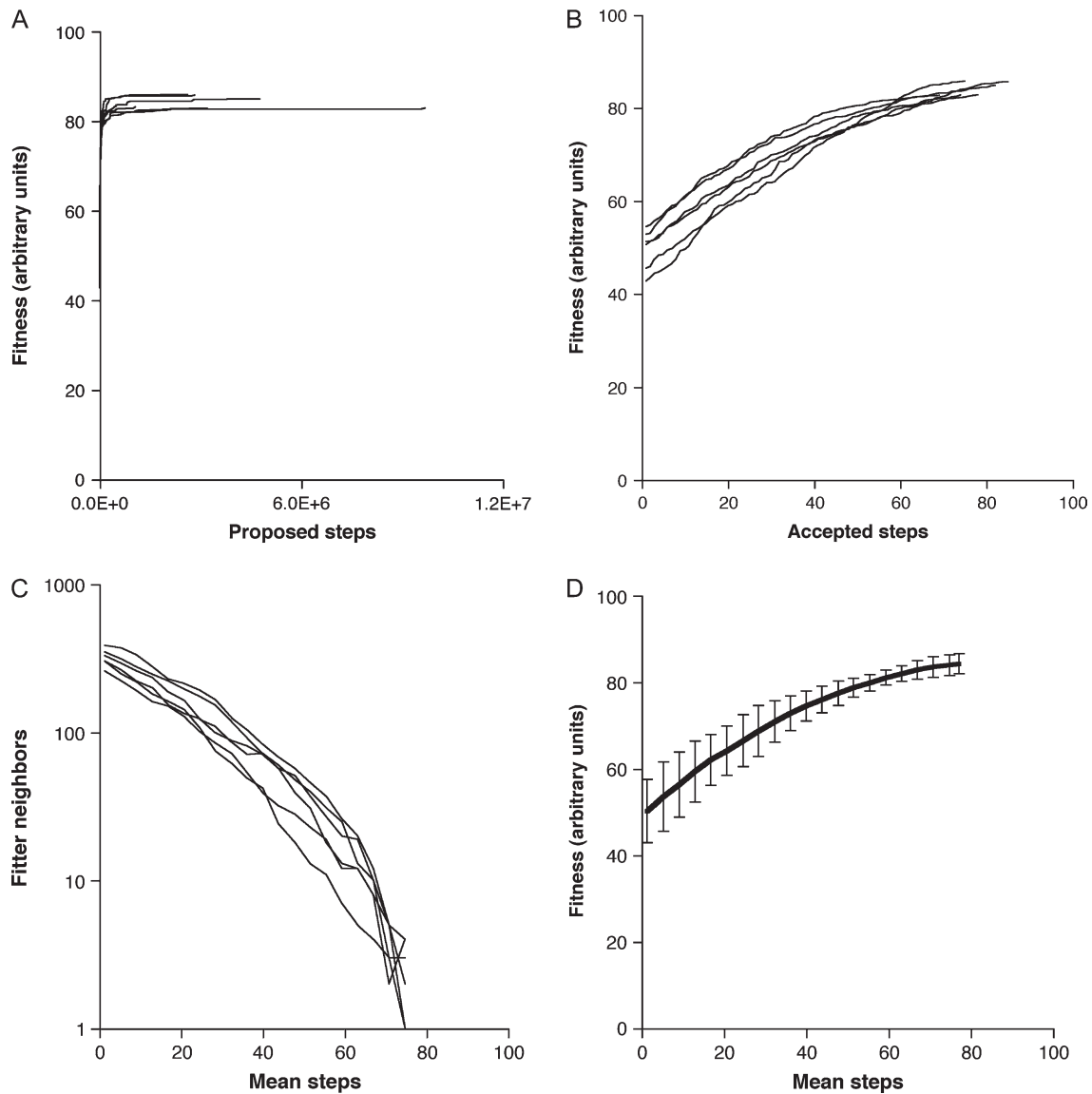


FIG. 2.—Sample behavior of 6 adaptive walks for the CDS model ( $N = 100$ ,  $K = 2$ ). In (A), (B), and (C), each line represents the progress of a single adaptive walk. Each walk begins with a different random sequence on a different random landscape; thus, any consistency in the set of walks is not due to a trivial recurrence of the same path. In (A) and (B), the increase in fitness is shown, scaled by the number of mutations proposed (A), which is proportional to time, or the number of accepted steps (B). The final fitness values are 83.2, 85.9, 82.9, 84.9, 82.8, and 85.8 (in the model used here, fitness ranges from 0 to 100, with a mean of 50). Because walks typically differ in starting point, length, and ending point, in order to characterize the average behavior for a set of walks, it is necessary to impose some kind of registration. Here the walks are registered by start and end and scaled linearly in between, in 5% intervals. Thus, the average behavior of a set of walks is characterized by a series of 21 average values: the average starting value, the average values at 19 internal intervals, and the average end value. These averages can be plotted against the interval number or, more meaningfully, the average number of steps for that interval. This method is used in (C), which shows the number of fitter neighbors for each interval of the individual walks, as a function of the mean number of steps. In (D), the same method is applied to the data on fitness, except that only the mean value is shown at each interval, with error bars representing the 95% confidence interval.

Figure 2 shows sample results for the CDS model, which represents a gene of 300 nt encoding a protein of  $N = 100$  residues via the canonical genetic code. Large numbers of replicate walks (typically 1,000) are used to characterize the average behavior of the model for a set of parameters. Because walks typically differ in starting point, length, and ending point, in order to characterize the average behavior for a set of walks, it is necessary to impose some kind of registration on each time series. Here the walks are registered by start and end and scaled linearly in between in 5% intervals. That is, each walk is divided

into 20 intervals of width  $L/20$ , where  $L$  is the number of steps for that walk, and the behavior of a set of walks is characterized by a series of 21 average values: the average starting value, the average values at 19 internal intervals, and the average end value. This method of registering results is explained and illustrated in figure 2.

## Results

Because most readers will be unfamiliar with models of adaptive walks (with or without mutation bias), it is

helpful to begin with a simplified model to demonstrate generic principles from limiting cases, including some limiting cases that cannot be implemented in the more complex model of a protein-coding sequence introduced below.

In interpreting these results, it is important to bear in mind 2 things. First, except in the special case of deterministic adaptation (below), every adaptive walk is unique. Because the number of local fitness optima is enormous, the chance is negligible (in a typical set of 1,000 walks) that 2 adaptive walks will end at the same peak, except in the special cases noted specifically below. Likewise, because the fitness components have many digits of precision, the chance is negligible that 2 sequences (e.g., a sequence and a 1-mutant neighbor, or the sequences representing 2 different peaks) have identical fitness. Second, because the issue here is in regard to global effects on composition, the results from many different walks are averaged, so that the uniqueness of adaptive walks is not apparent. For instance, although each walk ends in a different place (fig. 2*A* and *B*), the final fitness for a set of walks is represented by a mean value (fig. 2*D*); in a typical set of 1,000 walks, the confidence interval on this mean value is so narrow that it cannot be displayed meaningfully (figs. 3 and 4).

#### Simplified 2-Allele Model without Coding Effects

As a simplified model useful for illustrating basic effects, I consider a sequence of length  $N$  consisting of the symbols “F” and “G” that is both genotype and phenotype (it does not encode another sequence). A mutation bias is allowed such that mutations from “G” to “F” are  $\beta = \mu_G \rightarrow_F / \mu_F \rightarrow_G$  times more likely than the reverse mutations. Of interest is the effect of this bias on the dynamics of evolutionary walks and on sequence composition, measured as the F:G ratio.

The first special case to consider is a pure neutral model, in which every mutation, whether  $F \rightarrow G$  or  $G \rightarrow F$ , is accepted, regardless of fitness effects (fig. 3*A*). In this case, sequence composition simply approaches a steady state that directly reflects  $\beta$ , the bias in mutation.

In the second special case, only the best of all possible mutations is accepted at each step. In models of adaptive walks, this is called “greedy,” “perfect,” or “gradient” adaptation (Kauffman 1993; Orr 2002), though the precise meaning of this rule might be evoked more reliably by referring to Voltaire’s fictional character, Dr. Pangloss, with his theodicy of “the best of all possible worlds.” This rule suggests the Modern Synthesis view that selection chooses the best possible variant from the abundance of the “gene pool.” Under Panglossian adaptation, if all fitness values are unique, adaptive walks proceed deterministically, so that the same starting sequence will always take the same series of steps. Thus, mutation bias should have no effect, an expectation that is confirmed by the results shown in figure 3*B*.

The third special case, shown in figure 3*C*, is to use a more naturalistic acceptance rule based on population genetics, namely, Haldane’s  $2s$ , but to consider the unnatural case of  $K = 0$  (no interactions), so that every sequence must evolve to the same global optimum. Whereas in the Panglossian case, the entire path of each walk is determined by effects of fitness alone, in the  $K = 0$  case, the final out-

come is determined by fitness effects alone, but the path taken by each walk is subject to stochastic effects and to the kinetic bias imposed by mutation.

What accounts for the recurring behavior seen in figure 3*C*, in which the rate of change in composition per accepted step begins at an extreme, slows, and then reverses? The explanation is straightforward given the simplicity of the “FG” model (it becomes more complex for the CDS model below). A given starting sequence differs from the optimal sequence at  $x$  sites; thus—given only 2 possible residues, with no fitness interactions—the adaptive walk will have  $x$  steps, including some F-to-G changes and some G-to-F changes. Because  $N = 100$ , on average  $x = 50$ . After  $x$  steps, the sequence will arrive at the global optimum, which has a mean composition of  $F/G = 1$ . However, because the order of steps is not determined (the number of possible orders is  $x! \approx 10^{64}$ ) and the chance of taking a given step is proportional to  $\mu s$ , the mutationally preferred changes (and changes with larger fitness benefits) tend to occur early in the adaptive walk. Thus, when the mutational bias is toward “F,” the changes from “G” to “F” tend to take place earlier in the walk, leading to a composition bias; but this bias is later reversed when the remaining (slower) changes from “F” to “G” take place. In metaphorical terms, the  $K = 0$  landscape is like a smooth cone: the trajectory of adaptation may spiral upwards to the right under the influence of a rightwards mutation bias, or upwards to the left under the influence of a leftward mutation bias, but ultimately the same end point will be reached. This effect (reversal of the initial mutation-biased trajectory) should tend to diminish as  $K$  increases, because the rougher the landscape is, the less likely that mutations passed over early in the walk will remain beneficial.

The behavior of adaptive walks for  $K = 4$  and  $K = 16$  are shown in figure 2*D* and *E*, respectively. For the most extreme biases ( $\beta = 10$  or  $\beta = 1/10$ ), the rate of change in composition is 3% to 4% per accepted step in the first few steps, the same as in the neutral case or the case of  $K = 0$ . Because the adaptive walks are shorter as  $K$  increases, only 62% as long when  $K = 4$  and 29% as long when  $K = 16$ , the maximum change in composition is not as great as when  $K = 0$ . This effect is offset by the fact that the trajectory does not recurve as far for larger values of  $K$ .

As suggested above, the effect of  $K$  on this recurring behavior can be understood by noting that  $K$ , by specifying how many sites of context are necessary to assign the fitness component for a site, modulates how long a possible mutation will remain beneficial in the face of changes that may alter its context. When  $K = 0$ , mutant effects are context-independent, and slow beneficial steps that are passed over early in the adaptive walk simply take place later. As  $K$  increases, beneficial mutations with a kinetic disadvantage are just as likely to be passed over early in the walk, but they are less likely to occur late in the walk because they are less likely to remain beneficial. In the limit of  $K = N - 1$ , the fitness component at each site depends on the entire sequence, and the effect of an accepted change at one site is to randomize the effects of all possible mutations; thus, the recurring behavior should disappear. This expectation is borne out by results (not shown).

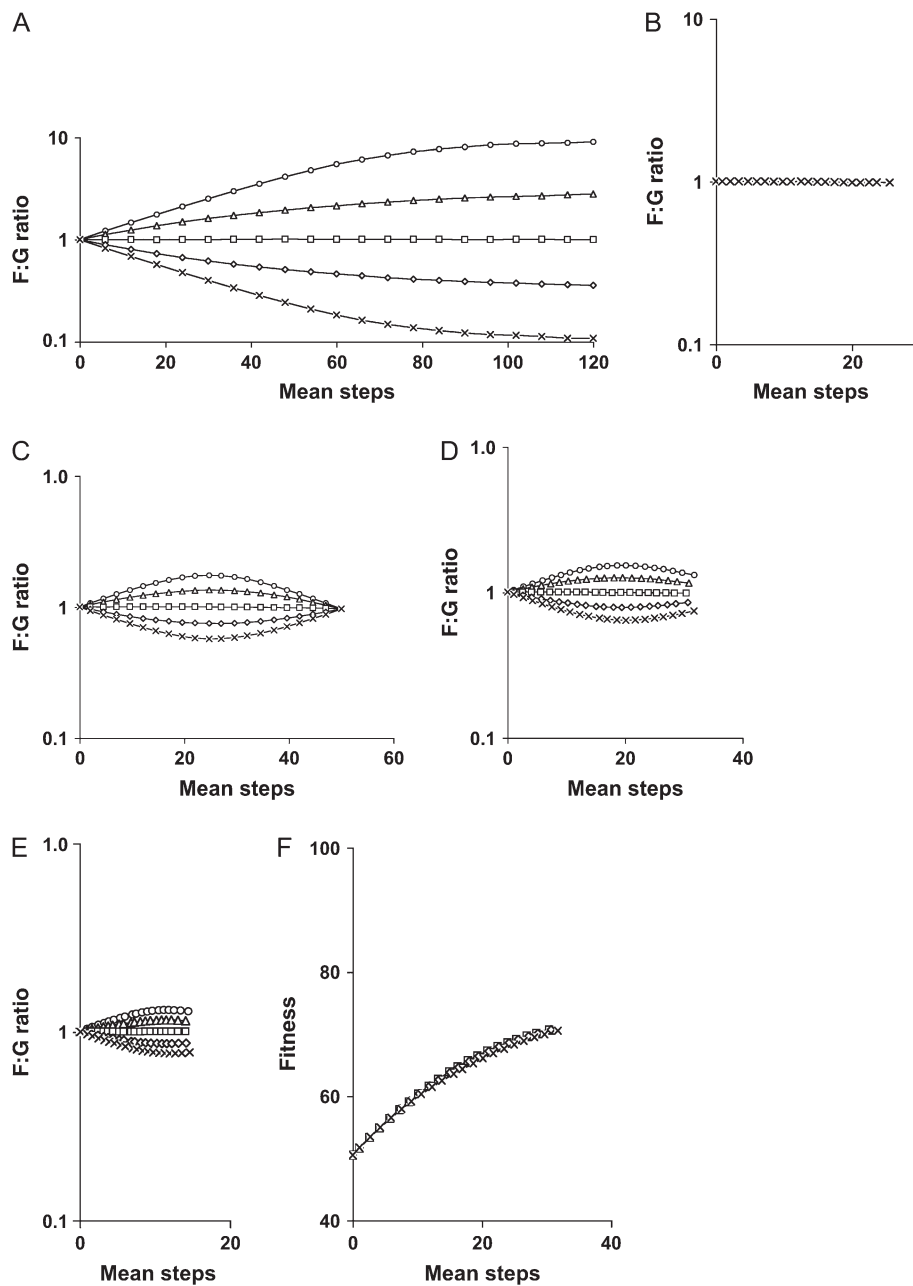
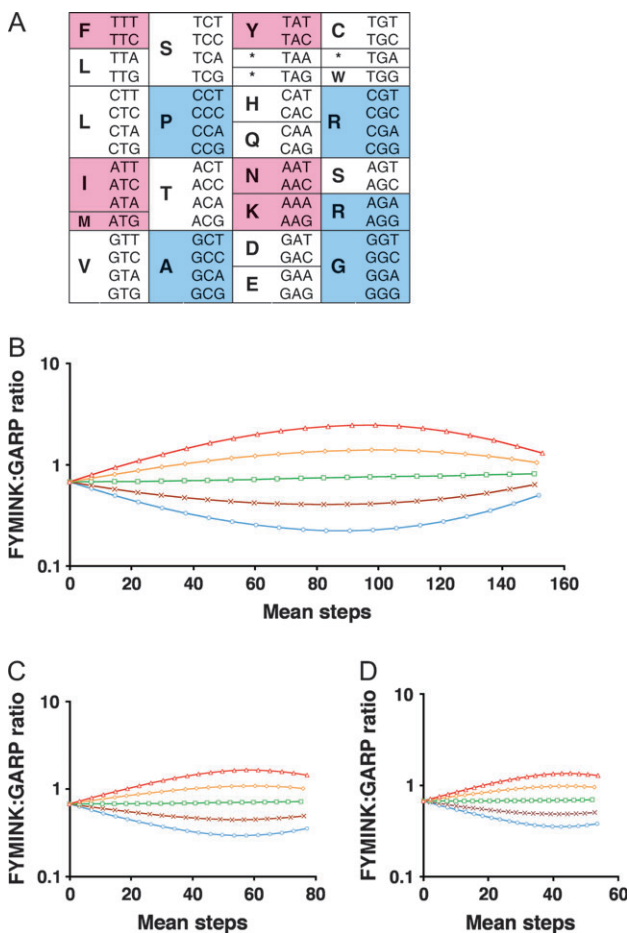


FIG. 3.—Limiting cases of mutational effects in a simple NK model. Each plot shows the mean behavior of adaptive walks for  $N = 100$  and subject to 5 different values of mutation bias  $\beta$  of 1/10 (crosses), 1/3 (diamonds), 1 (squares), 3 (triangles), and 10 (circles). The lines connect the data points for each series, which represents the average behavior of 1,000 walks (20 walks on each of 50 landscapes) that differ only in the mutation bias (95% CI are too small to display meaningfully). Panel (A) shows a pure neutral model, in which every mutation is accepted up to a limit of 300 steps. The mean final fitness is  $50.13 \pm 0.25$  (95% confidence interval) on a scale from 0 to 100. In (B), a proposed step is only accepted if it is the best possible step, with  $K = 4$ . Because this condition is deterministic, each set of 1,000 walks yields the same set of 1,000 outcomes (which have a mean walk length of  $25.49 \pm 0.25$  and a mean final fitness of  $71.03 \pm 0.10$ ). In (C), beneficial steps are accepted with a probability proportional to their selection coefficient, but the landscape is smooth, so that replicate walks reach the same end points regardless of mutation bias (mean walk length,  $50.05 \pm 0.32$ ; mean final fitness,  $67.09 \pm 0.16$ ). In (D) and (E), conditions are as in (C), but with  $K = 4$  and  $K = 16$ , respectively. The resulting mean walk lengths are  $31.10 \pm 0.54$  and  $14.30 \pm 0.22$  steps for  $K = 4$  and  $K = 16$ , respectively, and the mean final fitness values are  $70.61 \pm 0.05$  and  $66.04 \pm 0.06$ , respectively. The degree of mutation bias does not affect the level of fitness achieved, as shown in (F) using the data for adaptive walks when  $K = 4$ .

### A Coding-Sequence Model

Now it is possible to return to the CDS model, which can be applied more directly to the expected effect of mutation-biased adaptation on protein composition biases. In this model, a protein-coding gene evolves under a GC:AT mutation bias of  $\beta = \mu_{GC \rightarrow AT} / \mu_{AT \rightarrow GC}$ , where mutations

that do not change GC content have the intermediate rate  $\beta^{0.5} \mu_{AT \rightarrow GC}$ . Patterns of nucleotide composition in different genomes may be used to define a realistic range of parameter values. For instance, GC content in the third position of codons in diverse genomes ranges from about 11% to 90% (Singer and Hickey 2000; Knight et al. 2001).



**FIG. 4.**—Mutation-biased adaptation in the CDS model. Amino acid composition may be measured in terms of an FYMINK:GARP ratio following the canonical genetic code (A). The effect of mutation bias on amino acid composition is shown for adaptive walks for the CDS model,  $N = 100$ , and with  $K = 0$  (B),  $K = 2$  (C), and  $K = 4$  (D). The 5 data series represent 5 values for the GC:AT mutation bias parameter  $\beta$ , namely, 10 (triangles, red), 3 (diamonds, orange), 1 (squares, green),  $1/3$  (crosses, brown), and  $1/10$  (circles, blue). The lines connect the data points for a series, which represents the average behavior of 1,000 walks (20 walks each on 50 landscapes; 95% confidence intervals [CIs] are too small to be displayed meaningfully). For the 5 sets of conditions, the means of walk length and final fitness, respectively, are  $151.40 \pm 0.95$  (95% CI) and  $89.56 \pm 0.09$  for  $K = 0$ ;  $76.33 \pm 0.72$  and  $84.58 \pm 0.02$  for  $K = 2$ ; and  $52.89 \pm 0.62$  and  $80.73 \pm 0.05$  for  $K = 4$ . The narrow CIs on the fitness averages, which are averages for just 5 values of mutation bias that cover a 100-fold range, indicate that mutation bias has no substantial effect on the final fitness achieved.

If these are interpreted as steady-state frequencies reflecting only GC:AT bias, then  $\beta$  would range from about 9 to  $1/10$ ; thus, values of  $\beta$  are chosen from  $\{1/10, 1/3, 1, 3, 10\}$ . As noted previously, the bias in amino acid composition of a protein can be defined, following Singer and Hickey (2000), by distinguishing 2 classes of amino acids: the FYMINK class including amino acids with AT-rich codons and the “GARP” class with GC-rich codons (fig. 4A).

The results of simulations for the CDS model, shown in figure 4, reveal a substantial effect of GC:AT bias on amino acid composition. This effect, it should be noted, is far less than that expected at equilibrium under a pure

neutral model, which is a composition bias of 0.0074 for  $\beta = 1/10$ , 0.080 for  $\beta = 1/3$ , 0.67 for  $\beta = 1$ , 4.9 for  $\beta = 3$ , and 33 for  $\beta = 10$  (these expectations follow from the equilibrium GC:AT ratio, which is  $1/\beta$ , and the genetic code).

The CDS model differs from the simplified FG model presented earlier in that the effect of a mutation bias is not transient when  $K = 0$  (fig. 4B). The reason for this is somewhat counterintuitive given that, whereas it is conventional to attribute a property of “smoothness” or “roughness” to a “fitness landscape,” the amount of frustration (inability to reach the global optimum within some large period of time) depends on how the landscape is explored. For instance, if any mutation, no matter how complex, is allowed (e.g., changing the entire sequence by one mutation), there is no frustration in the sense that there is always an upward path to the global optimum. In the CDS model, the fitness function is defined on the space of protein sequences, but it is explored via single-nucleotide changes. Frustration occurs because an encoded amino acid cannot change to all 19 alternative amino acids by a single-nucleotide mutation but typically can change to each of only 6 or 7 other amino acids (e.g., an alanine codon cannot change into a leucine codon with 1-nucleotide mutation). In the CDS model, then, there are effectively many local peaks even when  $K = 0$ ; as  $K$  increases, the effects on walk length and composition are qualitatively similar to those seen above in the simplified model.

In the above simulations, only beneficial changes that alter the amino acid sequence are allowed, in order to make clear that mutation can impose a systematic bias on “strictly adaptive” walks, with no neutral or deleterious changes confounding the interpretation of results. Having demonstrated this point, I briefly consider the relation of amino acid composition to nucleotide composition in the more natural case in which GC content of the gene is influenced, not just by adaptive amino acid changes but by neutral synonymous changes. Simulations were performed as for the results in figure 4, except that synonymous changes were accepted with a probability of  $10^{-4}$  and that simulations continued for 400 steps to allow some equilibration of nucleotide composition. The resulting amino acid and nucleotide compositions are plotted in figure 5, along with data from real genomes (Singer and Hickey 2000). Under these conditions, the sequence typically ends at, or very near, a peak with properties similar to those for the strict adaptation model in figure 4. The vast majority of adaptive steps take place early in the walk. This is because early in the walk there are many beneficial mutations; their probability of fixation is twice the selection coefficient, which typically begins in the range of  $10^{-2}$  for the first step, then decreases (e.g., for the walks in fig. 2, the average selection coefficients for the first and last steps are  $1.0 \times 10^{-2}$  and  $1.9 \times 10^{-3}$ , respectively). Interestingly, the number of adaptive steps is slightly higher when neutral synonymous changes are included because synonymous changes effectively increase the number of paths for nonsynonymous changes (e.g., the number of accepted beneficial steps is  $85.66 \pm 0.03$  for  $K = 2$  when synonymous changes are allowed, as compared with  $84.58 \pm 0.02$  in the strict adaptation model in fig. 4B).

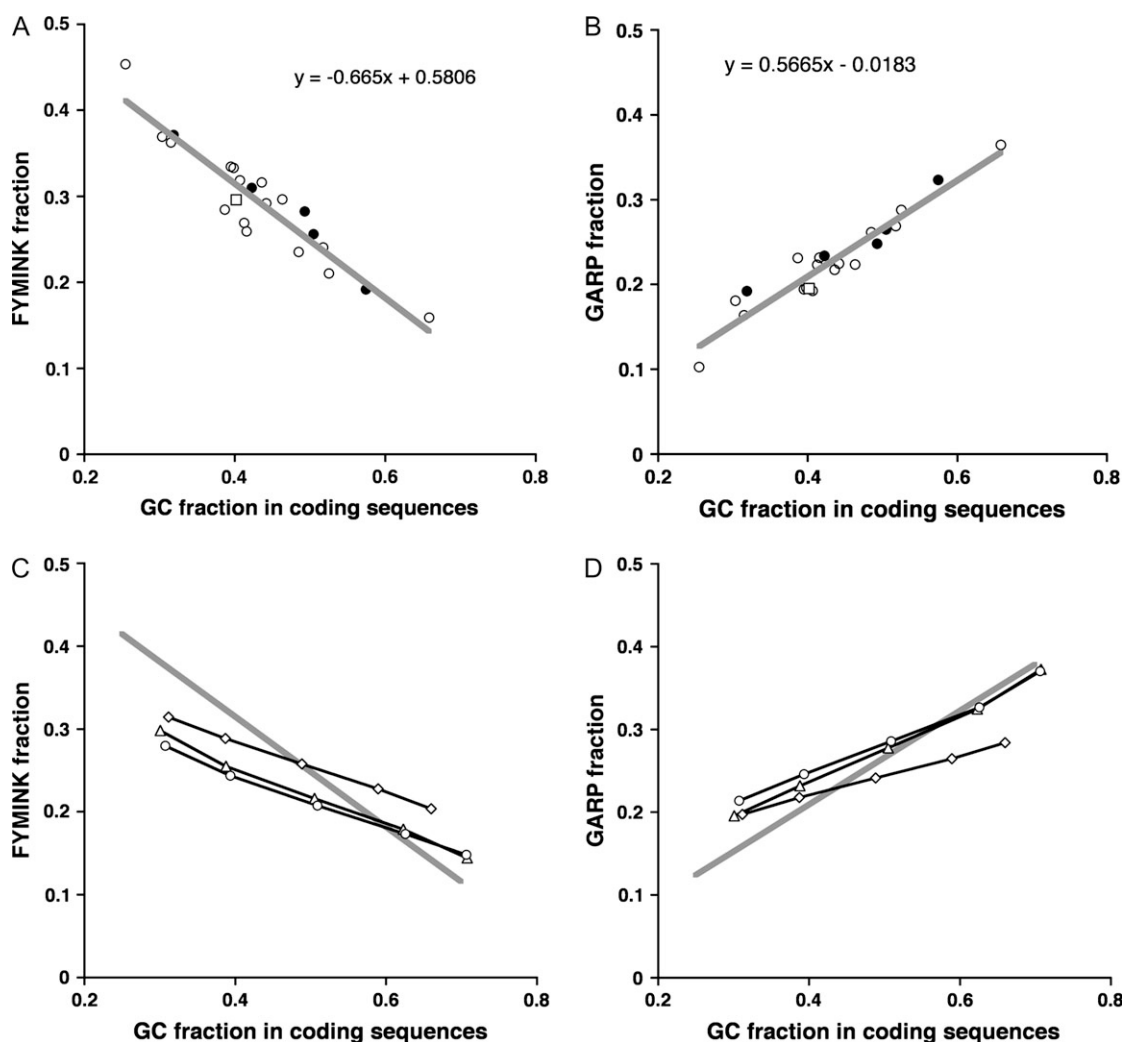


FIG. 5.—Composition effects in observed and simulated data. FYMINK and GARP frequencies are plotted as a function of GC content. Panels (A) and (B) show data from actual genomes of eubacteria (open circles), archaeobacteria (filled circles), and eukaryotes (squares), along with linear regressions (gray lines), and correspond to Figs. 1A and 1B of Singer and Hickey (2000), using data kindly provided by Greg Singer. Plots (C) and (D) show the regression (gray line) from the foregoing analysis of actual genomes, along with compositions of 1,000 simulated sequences for  $K = 0$  (diamonds),  $K = 2$  (triangles), and  $K = 4$  (circles; the black lines connect the points in each series). These simulations were done under the same conditions as for figure 4, except that synonymous changes were accepted with probability of  $10^{-4}$  and simulations ended at 400 steps whether or not a peak is reached. Thus, the 5 points in each simulated series represent the 5 values of the mutation bias parameter  $\beta$  used earlier.

## Discussion

Computer simulations of adaptive walks using an NK fitness model demonstrate that a modest mutation bias can impose a substantial composition bias, with the degree of displacement in composition, and its ultimate persistence, dependent on the degree of mutation bias and the roughness of the fitness landscape. This kind of behavior is common both to the simplistic FG model and the more complex CDS model. For the CDS model, significant biases in composition arise from modest GC:AT biases of a few fold (within the range of realistic biases), and modest degrees of fitness interaction (i.e., modest values of  $K$ ).

Whereas these results demonstrate that mutation bias is a possible cause of systematic trends or patterns in evolution, they do not prove that it is a plausible or likely cause. In particular, whereas the influence of GC:AT bias in mutation evidently has contributed to massive systematic

differences in protein composition in different genomes (Singer and Hickey 2000; Knight et al. 2001), this effect is not necessarily mediated by adaptive changes but may be due largely to neutral ones. The effect of including neutral (or nearly neutral) amino acid changes in the model used to generate figure 5 would be that the more the neutrality allowed, the greater the response of FYMINK:GARP ratio to GC:AT mutation bias. Presumably, there is some degree of neutrality that would maximize the fit with observed data. In the extreme case in which all changes are neutral, the equilibrium amino acid composition would follow from the equilibrium GC composition, as described in Introduction. For the present purposes, there is no reason to make an assumption about the amount of neutral evolution: the fraction of adaptive changes may be 50% or 10% or only 2%, but whatever is the fraction of adaptive changes, that is the fraction addressed by the model used here.

With respect to this fraction of adaptive changes, the model used here may seem highly arbitrary, and thus some discussion is warranted in regard to whether its treatment of evolutionary genetics and protein fitness are realistic or justifiable. The model uses a naturalistic approach in the sense that evolutionary steps are chosen by rules based on population genetics, e.g., the probability of acceptance is  $2s$ , whereas in the models of Kauffman (1993), all beneficial changes have an equal chance of being chosen. The model is defined for the space of nucleotide sequences that encode, by way of the canonical genetic code, sequences of the 20 natural amino acids whereas, for instance, the NK model of Ohta (1997) assumes only 9 amino acid states not encoded by nucleotide codons. The use of the genetic code is crucial because the primary statistical pattern of protein evolution is not something having to do with structure or amino acid properties, but with mutation: evolutionary change in proteins occurs overwhelmingly via “singlet” exchanges, that is, involving the 75 (out of 190) pairs of amino acids that can be exchanged via single-nucleotide mutations (Fitch 1966).

The model also is crudely naturalistic in that amino acids are assumed to interact strongly with other residues but only with a small number of them. Studies of proteins engineered with combinations of mutations indicate that most mutant effects are additive, not interactive (Gregoret and Sauer 1993). However, strong interactions exist (e.g., Chen et al. 1997; Kondrashov et al. 2002), and these context-specific effects dominate the spectrum of effects of amino acid exchanges at a given site. This principle is the operative basis of various successful methods in bioinformatics including profile-based homology searches and the Sorting Intolerant From Tolerant (SIFT) method for predicting effects of single-nucleotide polymorphisms (Ng and Henikoff 2003). In the NK model used here, each site interacts with a small number of other sites, but most mutant effects are additive. This is crucial because, as noted here and by Kauffman (1993), allowing just a small number of interactions per site (small  $K$ ) has an enormous effect on the “ruggedness” of the fitness landscape and thus on the dynamics of adaptation.

Recent studies (Tang et al. 2004; Yampolsky and Stoltzfus 2005b) also have revealed systematic differences in the mean exchangeability of amino acids that are not included in the protein NK model used here, but this is a quantitatively minor effect compared with the effect of the genetic code or of amino acid interactions. Considering singlet exchanges only, the mean probabilities of fixation vary only 10-fold (Tang et al. 2004), much smaller than the “mutational distance” effect of the genetic code and comparable to other mutational effects such as transition:transversion bias (Yampolsky and Stoltzfus 2005a). More importantly, the mean effects of amino acid exchanges account for only 4% of the variance in effects of specific amino acid exchanges (Yampolsky and Stoltzfus 2005b, Table 4), the remainder being due to context, that is, interactions as discussed above. The NK model used here ignores the (quantitatively less important) mean differences in exchangeability but addresses the (quantitatively more important) context-specific effects. Here such interactions are treated arbitrarily, but in future, it might be possible to develop a model of fitness interactions using an empirical contact-energy function (Berrera et al. 2003).

Finally, the model of an adaptive walk used here begins with a random sequence and ends when a limit of adaptation is reached and there is no further change. The choice to start with a random sequence is somewhat arbitrary but is intended to address the full course of adaptation, as opposed to the Darwinian assumption that adaptation can be treated as a matter of minor adjustments to a highly adapted state (e.g., Orr 2002). It might be more realistic to begin with a sequence that has higher-than-random fitness because, presumably, the evolution of new “functions” begins with some gratuitous excess capacity (Hall 2001). The choice to end simulations when a “peak” is reached is problematic because real proteins do not stop evolving. This is a problem for any model of strictly adaptive walks on a fixed landscape. The possible remedies include allowing neutral changes (Huynen 1996); allowing the fitness landscape to change, that is, to consider adaptation to a varying environment (Gillespie 1993); and allowing rare jumps beyond the immediate mutational neighborhood (Kauffman 1993). In a model that includes all of these possibilities, evolutionary changes would include neutral wandering, local climbing in regions of the landscape that have remained fixed, local climbing in regions that have not remained fixed, and rare nonlocal jumps.

At present, it is difficult to foresee how these conditions would influence the impact of mutation bias. However, the purpose of this study is not to provide a hyper-realistic model in which every possible contingency can be evaluated but to illustrate an unfamiliar principle that is expected to operate in evolving systems: biases in mutation, or more properly, biases in the origin of new variants by mutation, are expected to be an evolutionary cause of orientation or direction. This causal principle may be invoked in regard to various cases in which mutational non-uniformities appear to shape evolutionary change (Golding 1987; Gutierrez et al. 1994; Hancock 1995; Macey et al. 1997; Zhivotovsky et al. 1997; Petrov and Hartl 1998; Beletskii et al. 2000; Rokyta et al. 2005). Its effects are not limited to neutral evolution but apply to adaptive change, both in the short term (Yampolsky and Stoltzfus 2001) and, as shown here, in the longer term. Furthermore, the effects of this kind of causation are not limited to the biases in mutation per se that arise from molecular processes of replication, damage, and repair, but also would include phenotypic biases that arise from asymmetries in developmental systems (Yampolsky and Stoltzfus 2001; Stoltzfus 2006). In this regard, it may be noted that the genetic code is a developmental model—a set of rules that maps nucleotide genotypes to amino acid phenotypes, encapsulating a complex, regulated, self-organizing process involving hundreds of interacting components—and that, in the results presented here, it mediates the effect of a GC:AT bias in mutation.

### Acknowledgments

The author thanks Eric Nawrocki for his development of the initial version of the simulation software and Alex Mont for help in data analysis. This work was supported by the Center for Advanced Research in Biotechnology (a research institute jointly supported by the National Institute

of Standards and Technology and the University of Maryland Biotechnology Institute) and by NIH grant R01-LM007218 from the National Library of Medicine's Computational Biology Program. The identification of specific commercial software products in this paper is for the purpose of specifying a protocol and does not imply a recommendation or endorsement by the National Institute of Standards and Technology.

## Literature Cited

- Altenberg L. 1997. NK fitness landscapes. In: Back T, Fogel D, Michalewicz Z, editors. *The handbook of evolutionary computation*. Bristol, UK: IOP Publishing. p B2.7:5–10.
- Beletskii A, Bhagwat AS. 1996. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. *Proc Natl Acad Sci USA* 93:13919–24.
- Beletskii A, Grigoriev A, Joyce S, Bhagwat AS. 2000. Mutations induced by bacteriophage T7 RNA polymerase and their effects on the composition of the T7 genome. *J Mol Biol* 300:1059–67.
- Berrera M, Molinari H, Fogolari F. 2003. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. *BMC Bioinformatics* 4:8.
- Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897–907.
- Chen R, Greer AF, Dean AM. 1997. Structural constraints in protein engineering—the coenzyme specificity of *Escherichia coli* isocitrate dehydrogenase. *Eur J Biochem* 250:578–82.
- Cox EC, Yanofsky C. 1967. Altered base ratios in the DNA of an *Escherichia coli* mutator strain. *Proc Natl Acad Sci USA* 58:1895–902.
- D'Onofrio G, Jabbari K, Musto H, Bernardi G. 1999. The correlation of protein hydropathy with the base composition of coding sequences. *Gene* 238:3–14.
- Duret L. 2002. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev* 12:640–9.
- Fisher RA. 1930. *The genetical theory of natural selection*. London: Oxford University Press.
- Fitch WM. 1966. An improved method of testing for evolutionary homology. *J Mol Biol* 16:9–16.
- Freese E. 1962. On the evolution of the base composition of DNA. *J Theor Biol* 3:82–101.
- Fryxell KJ, Zuckerkandl E. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol Biol Evol* 17:1371–83.
- Galtier N, Lobry JR. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J Mol Evol* 44:632–6.
- Gillespie JH. 1991. *The causes of molecular evolution*. New York: Oxford University Press.
- Gillespie JH. 1993. Substitution processes in molecular evolution. I. Uniform and clustered substitutions in a haploid model. *Genetics* 134:971–81.
- Golding GB. 1987. Nonrandom patterns of mutation are reflected in evolutionary divergence and may cause some of the unusual patterns observed in sequences. In: Loeschcke V, editor. *Genetic constraints on adaptive evolution*. Berlin, Germany: Springer-Verlag. p 151–72.
- Gregoret LM, Sauer RT. 1993. Additivity of mutant effects assessed by binomial mutagenesis. *Proc Natl Acad Sci USA* 90:4246–50.
- Guiérrez G, Casadesus J, Oliver JL, Marin A. 1994. Compositional heterogeneity of the *Escherichia coli* genome: a role for VSP repair? *J Mol Evol* 39:340–6.
- Haldane JBS. 1927. A mathematical theory of natural and artificial selection. V. Selection and mutation. *Proc Camb Phil Soc* 26:220–30.
- Haldane JBS. 1932. *The causes of evolution*. New York: Longmans, Green and Co.
- Hall BG. 2001. Predicting evolutionary potential. I. Predicting the evolution of a lactose-PTS system in *Escherichia coli*. *Mol Biol Evol* 18:1389–400.
- Hancock JM. 1995. The contribution of slippage-like processes to genome evolution. *J Mol Evol* 41:1038–47.
- Huynen MA. 1996. Exploring phenotype space through neutral evolution. *J Mol Evol* 43:165–9.
- Kauffman S, Levin S. 1987. Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol* 128:11–45.
- Kauffman SA. 1993. *The origins of order: self-organization and evolution*. New York: Oxford University Press.
- Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217:624–6.
- Kimura M. 1983. *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press.
- Knight RD, Freeland SJ, Landweber LF. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol* 2: research0010.1-0010.13.
- Kondrashov AS, Sunyaev S, Kondrashov FA. 2002. Dobzhansky-Muller incompatibilities in protein evolution. *Proc Natl Acad Sci USA* 99:14878–83.
- Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, Wolfe KH. 1999. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res* 27:1642–9.
- Li W-H. 1997. *Molecular evolution*. Sunderland, MA: Sinauer.
- Lobry JR. 1997. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene* 205:309–16.
- Macey JR, Larson A, Ananjeva NB, Papenfuss TJ. 1997. Replication slippage may cause parallel evolution in the secondary structures of mitochondrial transfer RNAs. *Mol Biol Evol* 14:30–9.
- Mani GS, Clarke BC. 1990. Mutational order: a major stochastic process in evolution. *Proc R Soc Lond B Biol Sci* 240:29–37.
- Matsumoto M, Nishimura T. 1998. Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Trans Model Comput Simul* 8:3–30.
- Maynard Smith J, Burian R, Kauffman S, Alberch P, Campbell J, Goodwin B, Lande R, Raup D, Wolpert L. 1985. Developmental constraints and evolution. *Q Rev Biol* 60:265–87.
- Mayr E. 1980. Some thoughts on the history of the evolutionary synthesis. In: Mayr E, Provine W, editors. *The evolutionary synthesis*. Cambridge, MA: Harvard University Press. p 1–48.
- Mivart SG. 1871. *On the genesis of species*. London: R. Clay, Sons and Taylor.
- Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31:3812–4.
- Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst* 23:263–86.
- Ohta T. 1997. Role of random genetic drift in the evolution of interactive systems. *J Mol Evol* 44:S9–14.
- Orr HA. 2002. The population genetics of adaptation: the adaptation of DNA sequences. *Evol Int J Org Evol* 56:1317–30.
- Petrov DA, Hartl DL. 1998. High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. *Mol Biol Evol* 15:293–302.
- Rokyta DR, Joyce P, Caudle SB, Wichman HA. 2005. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. *Nat Genet* 37:441–4.

- Singer GA, Hickey DA. 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol Biol Evol* 17:1581–8.
- Stoltzfus A. 2006. Mutationism and the dual causation of evolutionary change. *Evol Dev* 8:304–17.
- Sueoka N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci USA* 48:582–92.
- Sueoka N. 1988. Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci USA* 85:2653–7.
- Tang H, Wyckoff GJ, Lu J, Wu CI. 2004. A universal evolutionary index for amino acid changes. *Mol Biol Evol* 21:1548–56.
- Welch JJ, Waxman D. 2005. The nk model and population genetics. *J Theor Biol* 234:329–40.
- Wolfe KH. 1991. Mammalian DNA replication: mutation biases and the mutation rate. *J Theor Biol* 149:441–51.
- Yampolsky LY, Stoltzfus A. 2001. Bias in the introduction of variation as an orienting factor in evolution. *Evol Dev* 3:73–83.
- Yampolsky LY, Stoltzfus A. 2005a. Untangling the effects of codon mutation and amino acid exchangeability. *Pac Symp Biocomput*:433–44.
- Yampolsky LY, Stoltzfus A. 2005b. The exchangeability of amino acids in proteins. *Genetics* 170:1459–72.
- Zhivotovsky LA, Feldman MW, Grishechkin SA. 1997. Biased mutations and microsatellite variation. *Mol Biol Evol* 14:926–33.

Kenneth Wolfe, Associate Editor

Accepted June 9, 2006