The Rate of Compensatory Evolution

Wolfgang Stephan

Department of Zoology, University of Maryland, College Park, Maryland 20742-4415

Manuscript received February 2, 1996
Accepted for publication June 13, 1996

ABSTRACT

A two-locus model is presented to analyze the evolution of compensatory mutations occurring in stems of RNA secondary structures. Single mutations are assumed to be deleterious but harmless (neutral) in appropriate combinations. In proceeding under mutation pressure, natural selection and genetic drift from one fitness peak to another one, a population must therefore pass through a valley of intermediate deleterious states of *individual* fitness. The expected time for this transition is calculated using diffusion theory. The rate of compensatory evolution, k_c, is then defined as the inverse of the expected transition time. When selection against deleterious single mutations is strong, k_c depends on the recombination fraction r between the two loci. Recombination generally reduces the rate of compensatory evolution because it breaks up favorable combinations of double mutants. For complete linkage, k_c is given by the rate at which favorable combinations of double mutants are produced by compensatory mutation. For $r > 0$, k_c decreases exponentially with r. In contrast, k_c becomes independent of r for weak selection. We discuss the dynamics of evolutionary substitutions of compensatory mutants in relation to WRIGHT's shifting balance theory of evolution and use our results to analyze the substitution process in helices of mRNA secondary structures.

ONE of the most important problems in population genetics is to understand the significance of epistatic selection in the evolutionary process. The analysis of epistatic fitness interactions has played an important role in the history of population genetics since it was introduced by HALDANE (1931) and WRIGHT (1931). Historically, epistatic interactions are defined as interactions between genes. Epistatic selection is expected to lead to nonrandom associations between polymorphisms at different loci within populations. However, nonrandom associations between loci have been rarely detected in natural populations. Most notably, extensive studies of linkage disequilibrium based on allozyme variation in natural populations of Drosophila have failed to lend support to the importance of the epistasis concept (LANGLEY 1977; HEDRICK *et al.* 1978; LEWONTIN 1985). In contrast, several population surveys based on RFLP analysis and DNA sequencing have revealed strong linkage disequilibria between polymorphisms that have been attributed to epistatic selection (MIYASHITA and LANGLEY 1988; MIYASHITA *et al.* 1993; SCHAEFFER and MILLER 1993). This work and other lines of research [for instance, LAURIE's experimental approaches to understanding the effect of polymorphisms within the *Adh* gene region on alcohol dehydrogenase expression (*e.g.*, LAURIE and STAM 1994)] suggest that it is appropriate to extend the classical epistasis concept such that interactions between polymorphisms within genes are also included. This generalized definition will be used in this paper.

Author e-mail: stephan@zool.umd.edu

One molecular mechanism that might be subject to epistatic selection is the maintenance of RNA secondary structure (KIRBY *et al.* 1995). These structures comprise single-stranded (loop) and double-stranded (stem) regions, where the stems are formed by Watson-Crick pairing of complementary bases. If these structures played an important functional role, then epistatic selection would act to conserve the form of the stems of these structures. Thus, individual mutations occurring at nucleotides that form Watson-Crick pairs within stems are expected to be deleterious if they break up the pairing of an intact structure. But fitness can be restored when a second ("compensatory") mutation occurs at the appropriate position on the opposite strand of the stem and reestablishes the pairing.

In this paper, a two-locus, two-allele model is described to analyze the evolution of compensatory mutations. Our main goal is to model the mode of evolution of compensatory mutations associated with stems of RNA secondary structures, but other processes (for instance the evolution of amino acids subject to protein folding) may also be described. Emphasis is placed on evolutionary rather than populational properties. Our fitness scheme follows essentially KIMURA's (1985a) compensatory neutral model: individual mutations are assumed to be deleterious, but harmless (neutral) in appropriate combinations (*i.e.*, complementary Watson-Crick pairs); furthermore, we neglect dominance effects at each locus, so that we can adopt the scheme of genic selection. However, we allow for differences in the fitness effects of individual mutations. This seems to be important for RNA secondary structures because

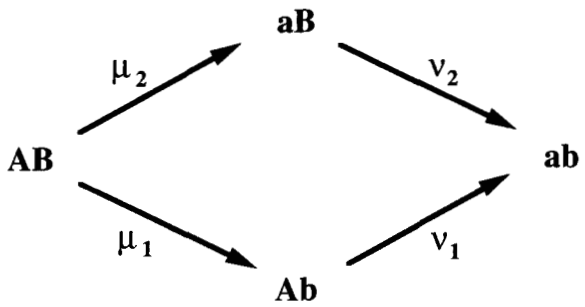

FIGURE 1.—Model of sequential mutation. The rate of the first mutation step is μ_i, that of the second "compensatory" mutation is ν_i ($i = 1, 2$).

noncanonical pairs (*e.g.*, A/C) appear to be under stronger selective constraints than wobble pairs (ROUSSET *et al.* 1991). Using diffusion theory, we are able to derive analytical approximations for the expected time to fixation of double mutants (*e.g.*, A-U) under tight linkage, assuming the population was originally fixed for a favorable combination (*e.g.*, G-C). Based on this result, we define the rate of compensatory evolution as the inverse of the expected time to fixation and apply our formulas to relevant data on substitution rates in mRNA secondary structures.

THEORY

Model: We consider a randomly mating, diploid population of effective size N. Furthermore, we assume two linked loci, with alleles A and a at locus 1 and alleles B and b at locus 2. Allele A is allowed to mutate irreversibly to a and likewise B to b. Back mutation is neglected because we assume that multiple hits at nucleotide sites are infrequent. The entire mutation scheme (including mutation rates) is depicted in Figure 1. The transition from haplotype AB to ab may occur through two different pathways with the intermediates Ab and aB. These pathways are denoted by the indices 1 and 2. The rate of the first mutation step is μ_i, that of the second "compensatory" mutation is ν_i ($i = 1, 2$). For selection, we assume a haploid (or genic) mechanism such that the alleles a and b are individually deleterious; *i.e.*, the fitnesses of the intermediate haplotypes Ab and aB are $1 - s_1$ and $1 - s_2$ ($0 < s_i < 1$), respectively, whereas the combinations AB and ab have fitness 1. The recombination fraction between the two loci is r.

These assumptions provide sufficient flexibility to model the evolution of compensatory mutations associated with RNA secondary structure. For instance, identify allele A with guanine (G), B with cytosine (C), a with adenine (A), and b with uracil (U). Then, the mutation scheme of Figure 1 shows a transition between the Watson-Crick pairs G-C and A-U via the intermediates GU and A/C, where the first one refers to a wobble pair and the other to a noncanonical pair. Assuming that the stability of pairing regions is an important determinant of the functional integrity of RNA molecules, it is reasonable to assign the highest fitness values to Watson-Crick pairings in stems of RNA secondary structures, whereas GU wobble pairs (which are less stable) and noncanonical pairs (such as A/C combinations) have smaller values.

The deterministic recurrence equations for this two-locus model with selection, recombination and mutation can be derived by standard methods (EWENS 1979; Chapt. 2). Let x_1, x_2, x_3, and x_4 be the frequencies of the haplotypes AB, Ab, aB, and ab, respectively. Then,

$$
\begin{aligned}
x_1' &= \frac{1}{\bar{w}}\{(x_1 w_1 - rD)(1 - \mu_1 - \mu_2)\} \\
x_2' &= \frac{1}{\bar{w}}\{(x_1 w_1 - rD)\mu_1 + (x_2 w_2 + rD)(1 - \nu_1)\} \\
x_3' &= \frac{1}{\bar{w}}\{(x_1 w_1 - rD)\mu_2 + (x_3 w_3 + rD)(1 - \nu_2)\} \\
x_4' &= \frac{1}{\bar{w}}\{(x_2 w_2 + rD)\nu_1 + (x_3 w_3 + rD)\nu_2 + (x_4 w_4 - rD)\},
\end{aligned}
\qquad (1)
$$

where w_i is the marginal fitness of the haplotype with index i, and $\bar{w}$ the mean fitness of the population. $D = x_1 x_4 - x_2 x_3$ is the coefficient of linkage disequilibrium.

This scheme is more general than KIMURA's (1985a) model of compensatory evolution because all mutation steps are characterized by individual parameters and because the selective values for the intermediate haplotypes may be different. This flexibility is necessary for modeling the evolution of compensatory mutations associated with RNA secondary structure.

Diffusion approximation for small recombination fractions: Our goal is to determine the expected time until fixation of the double mutant ab, assuming the population was initially fixed for the wild-type AB. Following KIMURA (1985a), we apply the diffusion equation method. Because of $\Sigma_{i=1}^{4} x_i = 1$, we must consider a three-dimensional diffusion process. However, analytical solutions of multidimensional diffusion equations are hard to obtain. We therefore make a series of assumptions that will lead to analytical approximations that are still useful for describing the evolution of compensatory mutations associated with RNA secondary structure. These assumptions are:

- The intermediate haplotypes Ab and aB are in an approximate mutation-selection balance. Necessary conditions for this balance to hold are that selection is much stronger than mutation ($s_i \gg \mu_i, \nu_i$; $i = 1, 2$) and also much stronger than genetic drift ($2Ns_i \gg 1$).
- Recombination rate is small relative to the selection coefficients; *i.e.*, $r < s_i$ ($i = 1, 2$). Since RNA secondary structures are formed within genes, this assumption does not appear to be too restrictive.

From the first assumption, it follows that, on the time scale of the change of x_4,

$$\Delta x_i = x_i' - x_i \approx 0, \quad i = 2, 3. \tag{2}$$

With and $x_1 \approx 1 - x_4$, we then obtain from (1) and (2)

$$x_2 \approx (1 - x_4)\left(\frac{\mu_1}{s_1} + \frac{r}{s_1}x_4\right) \quad \text{and}$$

$$x_3 \approx (1 - x_4)\left(\frac{\mu_2}{s_2} + \frac{r}{s_4}x_4\right). \tag{3}$$

Furthermore, the first assumption implies that the second-order moments of the changes in x_2 and x_3 can be neglected. Thus, the three-dimensional diffusion process can be reduced to one dimension. Introducing $x = x_4$ as the sole diffusion variable and measuring time in units of $2N$ generations, the infinitesimal drift and diffusion coefficients become (EWENS 1979; Chapt. 4)

$$a(x) \approx \gamma(1 - x) + \frac{1}{2}x(1 - x) \times [\theta_1 + \theta_2 - R(1 - 2x)] \quad \text{and}$$

$$b(x) = x(1 - x), \tag{4}$$

respectively. Here we have used the following abbreviations:

$$\theta_i = 4N\mu_i, \quad i = 1, 2$$

$$R = 4Nr, \quad \text{and}$$

$$\gamma = \frac{1}{2}\left(\frac{\theta_1}{s_1}\nu_1 + \frac{\theta_2}{s_2}\nu_2 + R\frac{\mu_1\mu_2}{s_1 s_2}\right). \tag{5}$$

To interpret the results, it is convenient to divide the time T until fixation of the double mutant *ab* into the time T_1 spent in the frequency interval $[0, 1/(2N)]$ and into T_2, the time spent in $[1/(2N), 1]$. In other words, T_1 is the time until a double mutant *ab* is formed. Then, the expected times (in generations) are (EWENS 1979; Chapt. 4),

$$\mathrm{E}(T) = \mathrm{E}(T_1) + E(T_2),$$

where

$$\mathrm{E}(T_1) \approx \frac{2N}{\gamma}\int_{1/2N}^{1} \psi(x)\,dx,$$

$$\mathrm{E}(T_2) \approx 4N\int_{1/2N}^{1} [b(y)\psi(y)]^{-1}\,dy\int_{y}^{1} \psi(x)\,dx, \quad \text{and}$$

$$\psi(x) = (2Nx)^{-2\gamma}\exp[-(\theta_1 + \theta_2 - R)x - Rx^2]. \tag{6}$$

To obtain the expression for $\mathrm{E}(T_1)$, EWENS' formula (4.42) is used. The integral in this formula is decomposed into an integral from x to $1/(2N)$ and an integral from $1/(2N)$ to 1. The contribution stemming from the first integral can be neglected. This leads immediately to the above expression. For $\gamma \ll 1$, which holds because of our strong selection-weak mutation assumption, we find by numerical analysis that $\mathrm{E}(T_1) \gg \mathrm{E}(T_2)$; that is, most of the time is spent until a double mutant is formed. Then, it follows from (5) and (6) that

$$\mathrm{E}(T) \approx \frac{1}{\frac{\mu_1}{s_1}\nu_1 + \frac{\mu_2}{s_2}\nu_2 + r\frac{\mu_1\mu_2}{s_1 s_2}}\int_{1/2N}^{1} \psi(x)\,dx. \tag{7}$$

This result has an intuitive interpretation because the term before the integral is the inverse of the rate at which double mutants *ab* are produced from the intermediate haplotypes *Ab* and *aB* by compensatory mutation or recombination. The integral accounts largely for the effect of recombination on the survival of newly formed double mutants. For weak mutation and recombination, this term is close to 1, but increases exponentially with the recombination fraction when $R > \theta_1 + \theta_2$. Therefore, because the main effect of recombination is to reduce the frequency of the double mutant *ab* by crossing-over with the wild-type *AB*, the expected time to fixation of a double mutant may be very long.

The rate of compensatory evolution: The inverse of the expected time to fixation of *ab*, $\mathrm{E}(T)^{-1}$, can be interpreted as the rate of compensatory evolution, k_c. This follows from the definition of a rate used in the physical sciences. In the molecular evolution literature, the rate of molecular evolution, k, is defined differently as the product of the per-generation mutation rate times the probability of ultimate fixation of a mutant. However, in the limit of small nucleotide mutation rates that are relevant here these two defintions are equivalent (STEPHAN and KIRBY 1993). Hence, we find for the rate of compensatory evolution under strong selection-weak mutation

$$k_c \approx \left(\frac{\mu_1}{s_1}\nu_1 + \frac{\mu_2}{s_2}\nu_2 + r\frac{\mu_1\mu_2}{s_1 s_2}\right)\frac{1}{\int_{1/2N}^{1} \psi(x)\,dx}. \tag{8a}$$

It is instructive to express the rate of compensatory evolution as a function of r for weak recombination and weak mutation (which is appropriate for intragenic nucleotide data). For complete linkage ($r = 0$), the rate of compensatory evolution becomes

$$k_c(0) \approx \frac{\mu_1}{s_1}\nu_1 + \frac{\mu_2}{s_2}\nu_2. \tag{8b}$$

This result corresponds to the well-known formula of the neutral theory, which says that the rate of molecular evolution is equal to the mutation rate. We may relate Equation 8a to an even broader body of theory by Taylor-expanding $\ln[\mathrm{E}(T)]$ with regard to r. Considering terms up to the first order shows that (see SIMULATION; Equation 16)

$$k_c(r) \approx \left(\frac{\mu_1}{s_1}\nu_1 + \frac{\mu_2}{s_2}\nu_2\right)e^{-rA}, \tag{8c}$$

where A is a positive function that depends, in a rather complicated way, on all parameters of the model (ex-

cept r). Equation 8c has the form of the classical Arrhenius law of chemical reaction kinetics (GARDINER 1990; Chapt. 5). This expresses the rate of a chemical reaction as a product of the frequency at which reactants try to overcome an activation barrier times a negative exponential term containing the height of the barrier. Clearly, the analogy with the present model is very close, since the first factor of the right-hand side of (8c) gives the frequency at which favorable double mutants are formed by mutation, whereas the exponential term describes the retarding effect of the recombination barrier.

Diffusion approximation for free recombination between loci: Although free recombination ($r = 0.5$) may not be applicable to RNA secondary structures, it is desirable for certain limiting cases (see below) to derive an explicit solution for the expected time until fixation of *ab* double mutants under continued mutation pressure. This problem is technically very difficult; therefore we consider here only the special case of equal mutation rates and equal selection coefficients; *i.e.*, $\mu \equiv \mu_i = \nu_i$, and $s \equiv s_i (i = 1, 2)$. This model has been treated by KIMURA (1985b) numerically and by simulation, but not analytically. We start with KIMURA's formulation. Let $E(T) = \bar{T}(p, q)$ be the expected time (in units of $2N$ generations) to fixation of *ab* gametes, given that the initial frequencies of *a* and *b* are p and q, respectively. Then, $\bar{T}(p, q)$ satisfies the partial differential equation

$$\frac{1}{2}p(1-p)\frac{\partial^2 \bar{T}}{\partial p^2} + \frac{1}{2}q(1-q)\frac{\partial^2 \bar{T}}{\partial q^2} + a(p, q)\frac{\partial \bar{T}}{\partial p} + a(q, p)\frac{\partial \bar{T}}{\partial q} + 1 = 0, \quad (9)$$

where $a(p, q)$ is the drift coefficient

$$a(p, q) = -\alpha p(1-p)(1-2q) + \frac{1}{2}\theta(1-p), \quad (10)$$

with $\alpha = 2Ns$ and $\theta = 4N\mu$. I was not able to find a general solution of Equation 9 under the appropriate boundary conditions (see KIMURA 1985b). However, the qualitative behavior of the corresponding deterministic differential equations

$$\frac{dx(t)}{dt} = -s\,x(t)[1-x(t)] \times [1-2y(t)] + \mu[1-x(t)],$$

$$\frac{dy(t)}{dt} = -s\,y(t)[1-y(t)] \times [1-2x(t)] + \mu[1-y(t)] \quad (11)$$

suggests the following heuristic argument for finding an approximative solution in certain parameter ranges. The dynamical system (11) has three equilibrium points. Besides (1, 1), these are

$$(x_1, y_1) \approx \left(\frac{\mu}{s}, \frac{\mu}{s}\right) \text{ and}$$

$$(x_2, y_2) \approx \left(\frac{1}{2} - \frac{\mu}{s}, \frac{1}{2} - \frac{\mu}{s}\right). \quad (12)$$

Another important observation is that a stochastic process, which starts at (0, 0), will get absorbed at (1, 1) with high probability due to selection as soon as this process reaches (x_2, y_2) [note that the first term in Equation 10 becomes positive at (x_2, y_2)]. In the following, we calculate the time T_1 until the first mutant reaches frequency $1/2 - \mu/s$. Since alleles *a* and *b* arrive first with equal probability, we model the process until the first mutant reaches frequency $1/2 - \mu/s$ as a one-dimensional diffusion on the interval $[-(1/2 - \mu/s), +(1/2 - \mu/s)]$, with p^* being the diffusion variable. The process starts at $p^* = 0$. For $p^* > 0$, the drift coefficient of this diffusion is $a(p^*, q^*)$ (defined by Equation 10), where $q^*(p^*)$ is the solution of

$$a(q^*, p^*) = 0. \quad (13)$$

The rationale here is that the selective pressure acting on the first mutant is much greater than that on the second one (see Equation 11). For $p^* < 0$, the drift coefficient is defined symmetrically. The expected time (in generations) until the first mutant reaches frequency $1/2 - \mu/s$ is then given by formulae 4.23 and 4.24 in EWENS (1979) as

$$E(T_1) \approx 4N\int_0^{p^{**}} \exp(-2\alpha x)\left(\frac{x}{1-2x}\right)^{\theta}\frac{1}{x(1-x)}dx \times \int_x^{p^{**}} \exp(2\alpha y)\left(\frac{1-2y}{y}\right)^{\theta}dy, \quad (14)$$

where $p^{**} = (1/2) - (\mu/s)$.

If selection against deleterious intermediates is very strong, we may have the situation that occasionally a mutant allele runs away to reach frequency $(1/2) - (\mu/s) \approx 1/2$ due to genetic drift and mutation (overpowering selection), while the frequency of the other mutant allele stays close to zero due to selection. Alternatively, the second allele may be in an intermediate frequency range when the first reaches $(1/2) - (\mu/s)$. However, this scenario is highly unlikely when strong selection acts against it.

Thus, under very strong selection against intermediates, the stochastic process reaches fixation at (1, 1) only if new *b* mutants are produced by mutation on *aB* chromosomes at times $t \geq T_1$ and if these new *ab* haplotypes survive recombination. The probability of this joint event is given by the product of the probabilities for the two single events, which can be calculated as follows. The probability for the creation of an *ab* type by mutation is approximately $1/2$ because about one half of the chromosomes are *aB* types at that time. Given an *ab* type has been produced by mutation in a certain

generation, it survives recombination due to strong selection if crossing-over with *AB* types does not occur in the same generation. The probability that a new *ab* type survives recombination is therefore the fraction of *ab* types, $N\mu/(N\mu + 1/2)$, which is produced more quickly than crossing-over between *ab* and *AB* chromosomes can occur. Thus, the expected time to fixation of *ab* double mutants, starting from $(p, q) = (0, 0)$, is

$$\mathrm{E}(T) \approx \frac{2N\mu + 1}{N\mu}\mathrm{E}(T_1). \qquad (15)$$

SIMULATION

Our analytical approximations assume weak mutation and strong selection. To examine the validity of these approximations and to explore a larger parameter space, simulations were performed as follows. In each generation, the effects of mutation, recombination and selection on the frequencies of the gametes *AB*, *Ab*, *aB*, and *ab* were computed using the recurrence Equation 1. The resulting distribution of gamete frequencies was then subject to multinomial sampling (described in STEPHAN *et al.* 1993). The simulation routine was tested against exact results (KIMURA 1985b) and against simulation results obtained from faithful sampling of gametes (KIMURA 1985a,b). The values of the time to fixation reported in the following are averages over 1000 runs.

Small recombination fractions: In Figure 2a, the simulation results are compared with the analytical approximations for small recombination fractions. The *x*-axis is scaled as $r/\max(s_1, s_2)$; the time to fixation on the y-axis is given in generations (on a log scale). For very small recombination rates, the agreement between the simulations and Equation 6 is good for both parameter sets:

Example 1: $\mu_i = \nu_i = 0.0001$, $s_i = 0.01$, $N = 500$;

Example 2: $\mu_i = \nu_i = 0.0005$, $s_1 = 0.025$, $s_2 = 0.05$, $N = 250$.

The simulations suggest that the time to fixation increases with recombination rate in a nearly exponential fashion; however, the theoretical curves in Figure 2a increase more than exponentially in r, thus causing huge differences between approximation and simulation. This suggests the following improvement of the analytical solution. First, a better agreement with the simulations is found when Equation 7 is used instead of Equation 6 because the analytical approximations generally overestimate the time to fixation. Second, an expansion of the function $\ln[\mathrm{E}(T)]$ with respect to r [where $\mathrm{E}(T)$ is given by the right-hand side of Equation 7] shows that the terms of the Taylor series are positive. This suggests that an approximation which considers only the zeroth- and first-order terms of this expansion is most useful. Hence,

$$\mathrm{E}(T) \approx \mathrm{E}(T)|_{r=0} \exp\left(r\frac{d}{dr}\ln(\mathrm{E}(T))|_{r=0}\right), \qquad (16)$$

where $\mathrm{E}(T)$ is given by the right-hand side of Equation 7. Equation 16 immediately leads to Equation 8c.

This approximation is compared with the simulations in Figure 2b. Indeed, it reproduces the simulation results much better than Equation 6. For very weak recombination such that the scaled recombination fraction is 0.1 or smaller, the difference between the analytical results and the simulations is within one standard error. It is also important to note that large values of $\theta_i = 4N\mu_i$ are not required for our analytical approximations to hold as one might expect for a mutation-selection balance model with drift (EWENS 1979; Chapt. 5.6). In both examples, the mutation parameter assumes values smaller than 1, which appears to be realistic for a wide range of species (including those with large population sizes). However, huge discrepancies arise for both sets of parameter values as the scaled recombination fraction approaches 0.5.

Free recombination: Figure 3 shows the simulation and analytical results for $N = 100$ and $\theta = 0.1$, 0.5, 1, and 2. For $\theta = 0.5$, 1, and 2, we find good to excellent agreement of the analytical approximations with the simulations when selection is sufficiently strong. For weak selection, however, large discrepancies may arise, in particular when θ is small. For $\theta = 0.1$, we found huge differences when $2Ns \leq 6$; for this case the analytical approximation is not shown in Figure 3. Asymptotically, for large values of $2Ns$, the fixation time increases exponentially with the strength of selection, as expected. For weak selection and large θ values, the curves go through a shallow minimum. This phenomenon was also noted by KIMURA (1985b).

Relaxing the assumptions: We have carried out simulations to examine the impact of the first assumption on the time to fixation. Increasing the effects of drift by reducing population size leads to marked discrepancies even for very small recombination fractions. For example, consider the parameter set of Example 1 with $N = 250$ instead of 500. Then the simulation yields an average fixation time of 342,266.3 ± 11,494.4 generations for $r = 0$, whereas Equation 6 produces 453,398.9 and the above linearized equation gives 447,516.0. Thus, the relative errors are much larger here than one standard error.

It is also interesting to relax the strong selection-weak mutation assumption because this encompasses the nearly neutral case of very weakly selected mutations (*i.e.*, $2Ns_i \leq 1$). For example, consider again the parameter set of Example 1 with the selection coefficients changed such that $2Ns_i = 0.5$. In this case, the simulations yield an average fixation time of 16, 058.8 ± 313.7 generations for $r = 0$, whereas Equation 6 gives 22,629.0.

Most interestingly, the time to fixation does not increase with recombination as we found in the case of

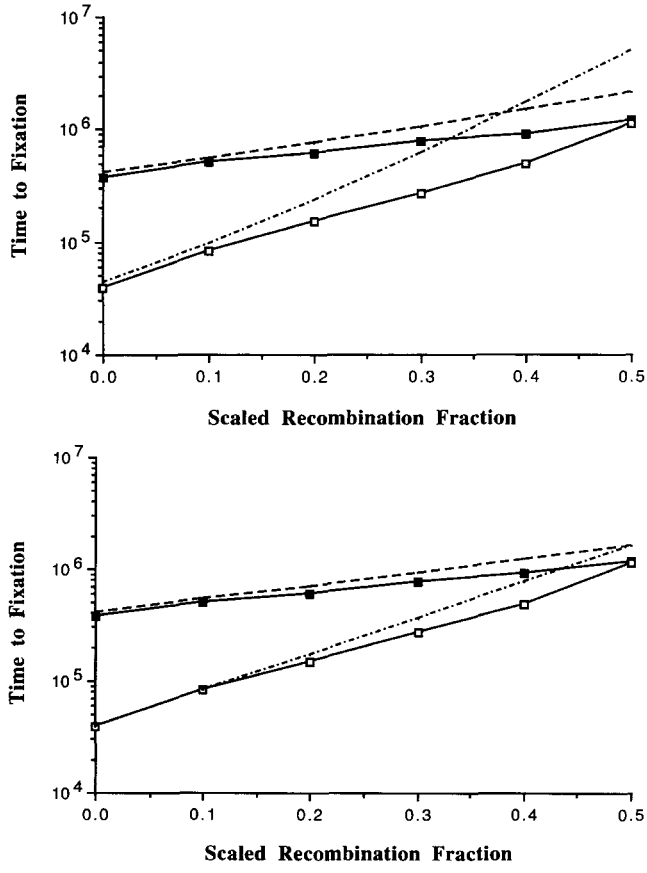

FIGURE 2.—Expected time to fixation as a function of recombination. Time is measured in generations and plotted on a log scale; recombination fraction is scaled as $r/\max(s_1, s_2)$. Simulations (denoted by squares) and theoretical results (denoted by dashed lines) are compared for two parameter sets (see text). Parameter set 1 (Example 1; ■) and Example 2 (□) simulation results are shown as are theoretical results for Example 1 (– – –) and Example 2 (·–·). (a) The theoretical values are calculated based on Equation 6; (b) the theoretical values are obtained from Equations 7 and 16. The latter approximations reproduce the simulation results much better. For $r/\max(s_1, s_2) \leq 0.1$, the theoretical results are within one standard error of the simulations.

strong selection-weak mutation. We have simulated the model for $r = 0$, 0.0001, 0.001, 0.01, 0.1, and 0.5 and found not the slightest indication of an effect of recombination. This behavior is not due to the choice of the selective values such that $2Ns_i \leq 1$; for increasing population size from $N = 500$ to $N = 2000$ leads to the same behavior. If mutation rates and selection coefficients are comparable, the population is near a quasilinkage equilibrium, so that the gain of *ab* double mutants produced by crossing-over between the *Ab* and *aB* intermediates is approximately equal to their loss due to crossing-over with *AB* gametes. The free-recombination approximation developed above (Equation 15) does not reproduce this case of weak selection-weak mutation well, as our diffusion results work only for strong selection. For example, for the case of $N = 2000$ the simulations yielded an average fixation time of ~20,000 generations (independent of recombination), whereas the free-recombination diffusion result is 26,376.0 generations. For the $N = 500$ example (in which case $2Ns_i = 0.5$), the discrepancy is even greater.

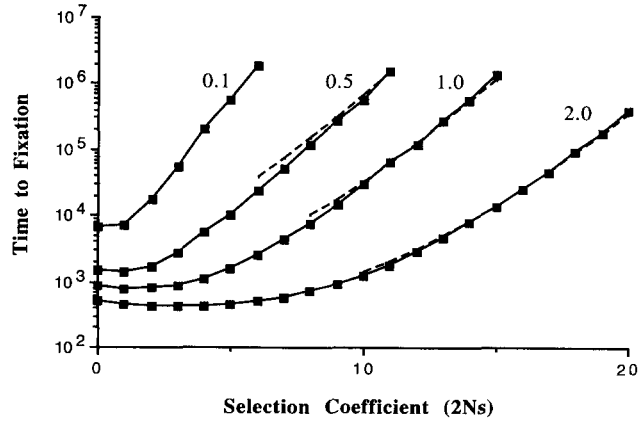

FIGURE 3.—Expected time to fixation (in generations) as a function of selection under free recombination. Simulations (■) are presented for $N = 100$ and four different θ values, which are written by the curves. Analytical approximations (– – –) are shown for $\theta = 0.5$, 1.0 and 2.0. The theoretical values are based on Equation 15. The analytical approximations agree with the simulations within one standard error when $2Ns$ is sufficiently large.

DISCUSSION

Using diffusion theory, we have obtained analytical approximations for the expected time to fixation of pairs of mutations (at different loci) under continued mutation pressure. We considered a two-locus model with two alleles (nucleotides) at each locus (site). Mutations at these sites were assumed to be individually deleterious but neutral in appropriate combinations. We found diffusion approximations for strong selection-weak mutation and tight linkage. Furthermore, we were able to treat analytically the case of free recombination between loci. The analytical results were examined by computer simulations. Simulations were also used to explore relevant parameter ranges that were not amenable to mathematical treatment (for instance

the case of weak selection-weak mutation). The diffusion process of our two-allele, two-locus model is three-dimensional. Progress could be made by reducing the number of diffusion variables using the assumptions that $\mu_i \ll s_i$ (strong selection-weak mutation) and $r < s_i$ (tight linkage). Under these assumptions, an explicit formula for the rate of compensatory evolution, k_c, was obtained.

Our most important finding is the relationship between recombination rate and the rate of compensatory evolution. For strong selection-weak mutation, we found that the rate of compensatory evolution decreases in an exponential fashion with the recombination fraction r, when r is small (Equation 8c). For complete linkage ($r = 0$), the rate of compensatory evolution is approximately equal to the rate at which double mutants *ab* are produced from the intermediate, detrimental haplotypes *Ab* or *aB* by compensatory mutation (Equation 8b). This result is analogous to the classical Arrhenius formula of chemical reaction theory. The effect of recombination on the rate of compensatory evolution can be explained as follows. Under strong selection-weak mutation, individual deleterious mutations remain in low frequencies. When a double mutant is formed by mutation, recombination will then most likely occur between the double mutant and the wild type. This will reduce the frequency of the double mutant and thus retard its fixation probability. In contrast, when selection (relative to mutation) is weaker, individual mutations may not remain at low frequencies. Then the elimination of double mutants *ab* by recombination with wild-types *AB* may be counterbalanced by the production of double mutants due to crossing-over between the intermediates *Ab* and *aB*; in other words, the population is in an approximate linkage equilibrium. As a consequence, the effect of recombination on the rate of compensatory evolution should disappear for weak selection. This was indeed found in the simulation of the weak selection-weak mutation case.

KIMURA (1985a, 1990, 1991) has likened the substitution process of compensatory mutations to WRIGHT's (1931, 1932) shifting balance theory. There is certainly some resemblance in that both processes describe shifts from one adapted peak to another one. However, there are also important differences between these two models. WRIGHT's theory is primarily concerned on how a species moves from one fitness peak through a (deep) valley of population *mean* fitness to another peak. Our analysis suggests that compensatory evolution does not necessarily occur through a deep adaptive valley of population mean fitness, even if selection against deleterious intermediates is strong (*i.e.*, *individual* fitness of the intermediate haplotypes is markedly lower than 1). For the biologically important case of strong selection and tight linkage (which is presumably relevant for RNA secondary structure; see below), we found that the population moves along a ridge of selective values from one fitness peak to the other one without passing through a deep valley of population mean fitness. This is because the deleterious intermediates are at very low frequency while waiting for a compensatory mutation to occur. On the other hand, if selection is weak, the population may pass through an adaptive valley of mean fitness, which in this case is not deep. For loose linkage and strong selection, the resemblance between compensatory evolution and the shifting balance process is closer. Our free-recombination analysis indicates that the deleterious intermediates may reach high frequency before a compensatory mutation occurs and may thus reduce population mean fitness. However, this process is extremely slow and may therefore be very rare in nature.

Next we will relate our results to data on the evolution of compensatory mutations associated with stems of RNA secondary structures. STEPHAN and KIRBY (1993) used the sequences of 10 Drosophila species to infer *Adh* mRNA secondary structures. They obtained for each inferred helix an average number of compensatory substitutions ("covariations") per pair and the physical distance (calculated between the nucleotides of the bottom pair of the stem). Their main observation was that the number of covariations decreased with physical distance. There was a much stronger distance effect for short-range pairings (with physical distances <40 nucleotides) than for longer-range pairing regions. Our theoretical findings may suggest the following explanation for this behavior. The strongly negative correlation between the number of covariations and distance for short-range pairings indicates that selection pressure against individual deleterious mutations within stems is relatively high. On the other hand, the weak distance effect for longer-range pairings is consistent with selection on single mutations within stems being weak. At first, the explanation for the long-range pairings seems implausible since long-range pairings that order several short-range helices into discrete structural units appear to be functionally more important and therefore under stronger selective constraints. However, all that is required for our explanation to work is that selection pressure per site (not per pairing region) differs between short- and longer-range helices. STEPHAN and KIRBY (1993) found a highly significant correlation between stem length and physical distance. This seems to suggest that selective constraints on (entire) longer-range helices can be larger than on short-range pairings, even if individual mutations in longer-range helices are subject to weaker selection pressure.

Of great interest for any RNA higher-order structure is a comparison between the rate of evolution in paired *vs.* unpaired regions of the structure. From such comparisons, we may be able to estimate some of the parameters of our model. Unfortunately, the currently available data permit only a crude analysis. Using Equation 16, the rate of compensatory evolution $k_c(r)$ from Equation 8c can be fitted to the data on the number of

covariations and physical distances discussed above. In this fit, we make the (reasonable) assumption that recombination rate scales with physical distance in a linear manner. This fit reveals that on average 0.77 covariations per pair would have occurred in the history of the 10 species if linkage were complete ($r = 0$). This value is proportional to $k_c(0)$. For comparison, one needs to know the number of substitutions in the unpaired regions of the *Adh* mRNA secondary structure for the same 10 species, which is proportional to the rate of molecular evolution k. MORIYAMA and GOJOBORI (1992) estimated the average number of synonymous substitutions for the *Adh* gene without accounting for secondary structure. Their results may therefore not be exactly applicable; on the other hand, a saturated mRNA secondary structure is not available at present. Based on their data, we estimated that on average 3.17 synonymous substitutions per site have occurred in the history of the 10 species. This suggests that

$$\frac{k_c(0)}{k} \approx \frac{0.77}{3.17} = 0.24.$$

Thus, the rate of compensatory evolution is considerably lower than that of molecular evolution (which assumes that the nucleotide sites in *Adh* evolve independently). Based on Equation 8b, one would have expected an even lower value for $k_c(0)$. If k were comparable with the mutation rates ν_1 and ν_2, $k_c(0)$ should be by a factor $(\mu_1/s_1) + (\mu_2/s_2)$ smaller than k. There may be several reasons for the relatively high value of $k_c(0)$. First, a more detailed analysis of *Adh* mRNA secondary structure may shift the estimates for the rates of evolution in paired and unpaired regions. Such an analysis requires more sophisticated methods for the inference of secondary structure (see KIRBY *et al.* 1995; MUSE 1995) than those used in STEPHAN and KIRBY (1993). Second, a more realistic model may take into account multiple hits at pairing sites and hence more pathways between stable pairs within a helix. This would increase the rate at which double mutants are formed (relative to the rate given by the right-hand side of Equation 8b). Third, the mutation rates ν_1 and ν_2 in paired regions may be larger than k due to templated mutation (GOLDING 1987). In this process, which acts over short distances, the frequency of nucleotide substitutions is thought to be elevated by the palindromic structure of DNA.

I thank the Population Genetics Discussion Group of the University of California at Davis and two reviewers for their comments on an earlier version of this paper. This research was supported in part by National Science Foundation grant DEB-9407226 and a Semester Research Award from the University of Maryland.

LITERATURE CITED

EWENS, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, Berlin.

GARDINER, C. W., 1990 *Handbook of Stochastic Processes*. Springer-Verlag, Berlin.

GOLDING, G. B., 1987 Nonrandom patterns of mutation are reflected in evolutionary divergence and may cause some of the unusual patterns observed in sequences, pp.151–172 in *Genetic Constraints on Adaptive Evolution*, edited by V. LOESCHCKE. Springer-Verlag, Berlin.

HALDANE, J. B. S., 1931 A mathematical theory of natural selection. VIII. Stable metapopulations. Proc. Cambridge Philos. Soc. **27:** 137–142.

HEDRICK, P. W., S. JAIN and L. HOLDEN, 1978 Multilocus systems in evolution. Evol. Biol. **11:** 101–182.

KIMURA, M., 1985a The role of compensatory neutral mutations in molecular evolution. J. Genet. **64:** 7–19.

KIMURA, M., 1985b Diffusion models in population genetics with special reference to fixation time of molecular mutants under mutational pressure, pp. 19–39 in *Population Genetics and Molecular Evolution*, edited by T. OHTA and K. AOKI. Jpn. Sci. Soc. Press, Tokyo.

KIMURA, M., 1990 Some models of neutral evolution, compensatory evolution, and the shifting balance process. Theor. Popul. Biol. **37:** 150–158.

KIMURA, M., 1991 Recent developments of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. Proc. Natl. Acad. Sci. USA **88:** 5969–5973.

KIRBY, D. A., S. V. MUSE and W. STEPHAN, 1995 Maintenance of pre-mRNA secondary structure by epistatic selection. Proc. Natl. Acad. Sci. USA **92:** 9047–9051.

LANGLEY, C.H., 1977 Nonrandom associations between allozymes in natural populations of *Drosophila melanogaster*, pp. 265–273 in *Measuring Selection in Natural Populations*, edited by F. B. CHRISTIANSEN and T. M. FENCHEL. Springer-Verlag, Berlin.

LAURIE, C. C., and L. F. STAM, 1994 The effect of an intronic polymorphism on alcohol dehydrogenase expression in *Drosophila melanogaster*. Genetics **138:** 379–385.

LEWONTIN, R. C., 1985 Population genetics. Annu. Rev. Genet. **19:** 81–102.

MIYASHITA, N., and C. H. LANGLEY, 1988 Molecular and phenotypic variation of the *white* locus region in *Drosophila melanogaster*. Genetics **120:** 199–212.

MIYASHITA, N. T., M. AGUADÉ and C. H. LANGLEY, 1993 Linkage disequilibrium in the *white* locus region of *Drosophila melanogaster*. Genet. Res. **62:** 101–109.

MORIYAMA, E. N., and T. GOJOBORI, 1992 Rates of synonymous substitution and base composition of nuclear genes in Drosophila. Genetics **130:** 855–864.

MUSE, S. V., 1995 Evolutionary analyses of DNA sequences subject to constraints on secondary structure. Genetics **139:** 1429–1439.

ROUSSET, F., M. PÉLANDAKIS and M. SOLIGNAC, 1991 Evolution of compensatory substitutions through a GU intermediate state in *Drosophila* rRNA. Proc. Natl. Acad. Sci. USA **88:** 10032–10036.

SCHAEFFER, S. W., and E. L. MILLER, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Genetics **135:** 541–552.

STEPHAN, W., and D. A. KIRBY, 1993 RNA folding in Drosophila shows a distance effect for compensatory fitness interactions. Genetics **135:** 97–103.

STEPHAN, W., L. CHAO and J. G. SMALE, 1993 The advance of Muller's ratchet in a haploid asexual population—approximate solutions based on diffusion theory. Genet. Res. **61:** 225–231.

WRIGHT, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

WRIGHT, S., 1932 The roles of mutation, inbreeding, crossbreeding, and selection in evolution. Proc. Int. Congr. Genet. **1:** 356–366.

Communicating editor: G. B. GOLDING