Selection Intensity Against Deleterious Mutations in RNA Secondary Structures and Rate of Compensatory Nucleotide Substitutions

Hideki Innan* and Wolfgang Stephan†

*Department of Biological Sciences, University of Southern California, Los Angeles, California 90089-1340 and †Department of Evolutionary Biology, University of Munich, 80333 Munich, Germany

Manuscript received February 17, 2001
Accepted for publication June 4, 2001

ABSTRACT

A two-locus model of reversible mutations with compensatory fitness interactions is presented; single mutations are assumed to be deleterious but neutral in appropriate combinations. The expectation of the time of compensatory nucleotide substitutions is calculated analytically for the case of tight linkage between sites. It is shown that selection increases the substitution time dramatically when selection intensity $Ns > 1$, where N is the diploid population size and s the selection coefficient. Computer simulations demonstrate that recombination increases the substitution time, but the effect of recombination is small when selection is weak. The amount of linkage disequilibrium generated in the process of compensatory substitution is also investigated. It is shown that significant linkage disequilibrium is expected to be rare in natural populations. The model is applied to the mRNA secondary structure of the *bicoid* 3′ untranslated region of Drosophila. It is concluded that average selection intensity Ns against single deleterious mutations is not likely to be much larger than 1.

MODELS of compensatory evolution involve mutations from two or more loci. These mutations are assumed to be deleterious when they occur independently, but, in combination, they (at least partially) compensate each other for their deleterious effects. Kimura (1985) proposed a two-locus, two-allele model, with alleles *A* and *a* at the first locus and *B* and *b* at the second locus, as illustrated in Figure 1A. He assumed that the two intermediate haplotypes, *Ab* and *aB*, are deleterious, while the wild-type *AB* and the double mutant *ab* are neutral. He further assumed that selection intensity against the deleterious intermediates is so strong that their frequencies in a population are very low. Accordingly, he considered only unidirectional mutations from *A* to *a* and from *B* to *b* and ignored back mutations (Figure 2A).

An important example of compensatory evolution is found in RNA secondary structures. In single-stranded RNAs, Watson-Crick (WC) pairing of complementary nucleotide bases is the basic mechanism in the formation of stem-loop structures. It is believed that an individual mutation that breaks up a WC pairing is deleterious and that a second "compensatory" mutation at the complementary site can reestablish pairing and restore fitness. The relatively simple pattern of intramolecular WC base-pairing involved in RNA structures has made them a suitable model for the study of compensatory evolution (Stephan and Kirby 1993; Golding 1994; Schöniger and von Haeseler 1994; Kirby *et al.* 1995; Muse 1995; Rzhetsky 1995; Tillier and Collins 1995; Stephan 1996; Higgs 2000; Savill *et al.* 2001). Most of these authors analyzed rRNA secondary structures.

Phylogenetic analysis has revealed a number of compensatory nucleotide changes between species. On the other hand, however, mismatches, including not only GU wobble pairs but also other noncanonical pairs, are frequently observed, indicating that selection against deleterious intermediates may not be very strong (Rousset *et al.* 1991; Parsch *et al.* 2000). This suggests that Kimura's compensatory evolution model, which assumes strong selection against deleterious intermediates, may not be generally applicable to the evolution of RNA secondary structures.

In this article, a compensatory evolution model is described in which selection against deleterious single mutations is not necessarily strong but covers a broad range of selection coefficients. Under weak selection, deleterious haplotypes may increase in frequency and even fix in the population as illustrated in a simulation run shown in Figure 2B. Since back mutations play an important role when the frequencies of deleterious haplotypes become large, we use a two-locus, two-allele model in which bidirectional mutations are considered (Figure 1B). This model is different from those used by Kimura (1985), Iizuka and Takefu (1996), and Stephan (1996), who assumed that selection against deleterious intermediates is so strong that the mutation process may be considered unidirectional (see above). Our analysis is also different from that of Higgs (1998), who assumed bidirectional mutation but analyzed the model only in

Corresponding author: Wolfgang Stephan, Department of Evolutionary Biology, University of Munich, Luisenstrasse 14, 80333 Munich, Germany. E-mail: stephan@zi.biologie.uni-muenchen.de

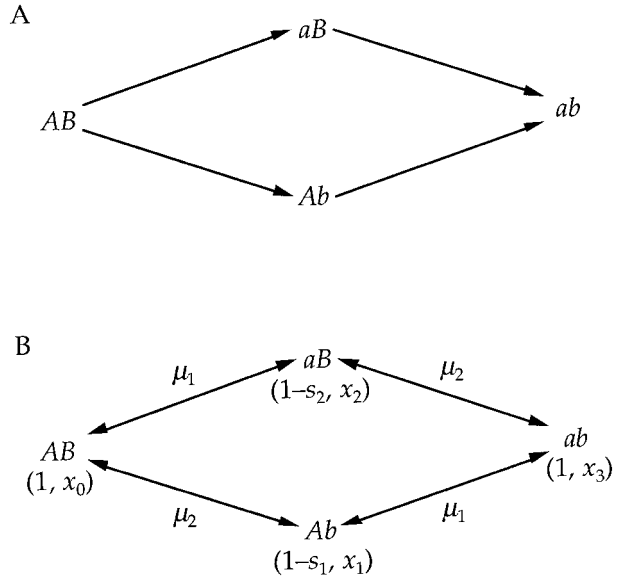

FIGURE 1.—Models of compensatory evolution. (A) KIMURA's (1985) model in which very strong selection is assumed and only unidirectional mutations are considered. (B) The model used in this article. Bidirectional mutations are considered. The fitnesses and frequencies are presented in parentheses.

the parameter range of strong selection (*i.e.*, $Ns \gg 1$, where N is the diploid population size and s the selection coefficient).

Our goal is to calculate the time to proceed from the wild-type state *AB* to the fixation of the double mutant *ab* in this bidirectional mutation model. It should be noted, however, that neither *AB* nor *ab* is an absorbing state. That means we do not calculate a "fixation" time *sensu stricto* but do calculate the time to "flip" back and forth between *AB* and *ab*. This is possible as the nucleotide mutation rate is much smaller than $1/N$ (see below). To describe the transition between *AB* and *ab*, it is therefore reasonable to use the term "fixation" (or "substitution"). We present analytical results for the rate of this compensatory substitution event when there is no recombination between the two loci and use computer simulation to obtain this time under the influence of recombination. The theoretical results are applied to DNA sequence data of the *bicoid* 3′ untranslated region (UTR) of Drosophila to estimate the selection intensity against deleterious mutations.

Another purpose of this article is to evaluate the amount of linkage disequilibrium generated by the compensatory evolution model. It is known that epistatic selection may produce significant linkage disequilibrium in natural populations (LEWONTIN 1974). SCHAEFER and MILLER (1993) reported linkage disequilibria in two clusters of DNA polymorphisms in introns of *Adh* in *Drosophila pseudoobscura*. These disequilibria are likely due to epistatic selection maintaining pre-mRNA secondary structures (KIRBY *et al.* 1995). Here we examine whether these findings are consistent with our model of compensatory evolution.

THEORY

Consider a two-locus, two-allele model for a randomly mating population with N diploids. There are alleles A and a at the first locus and B and b at the second locus. The (bidirectional) mutation rate between A and a is given by μ_1 and that between B and b is given by μ_2 (Figure 1B). The mutation rates are assumed to be much smaller than $1/(2N)$, as is the case for nucleotide substitution rates. The relative fitnesses of the haplotypes *AB*, *Ab*, *aB*, and *ab*, are given by 1, $1 - s_1$, $1 - s_2$, and 1, respectively. Effects of fitness are assumed additive within locus and therefore the diploid model is equivalent to a haploid model. The haplotype frequencies are denoted by x_0, x_1, x_2, and x_3, respectively.

In the following, we consider the process of compensatory substitution under the joint action of drift, selection, mutation, and recombination. We assume the process starts at $x_0 = 1$ at $t = 0$. First, mutations in *AB* produce *Ab* and *aB* types. Next, mutations in *Ab* and *aB* may create *ab*, and some of them fix in the population. Recombination between *Ab* and *aB* may also produce *ab*. We are interested in the expected time of compensatory substitution, defined as the time from $t = 0$ (when the system is at state $x_0 = 1$) to the time point when the double mutant *ab* is fixed ($x_3 = 1$). Because the mutation process is bidirectional, the latter state is not an absorption state. The time we are calculating is therefore a first passage time. Because of the assumption $\mu_i \ll 1/(2N)$, double mutants *ab* either get lost by drift or go to fixation; the proportion of *ab* haplotypes reverting back to *Ab* or *aB* is very small. Recombination can only retard the fixation process (KIMURA 1985; STEPHAN 1996). The fixation time can be divided into two parts, T_1 and T_2. T_1 is the waiting time for the "successful" double mutant *ab* (that will eventually get fixed) to appear in the population, and T_2 is the time from the appearance of *ab* to the fixation event (see Figure 2).

Symmetrical model: Consider a symmetrical model with $s = s_1 = s_2$ and $\mu = \mu_1 = \mu_2$. In the analytical derivations, recombination is neglected. The four haplotypes are divided into two groups: one consists of *AB* and *ab*, and the other is the group of the deleterious intermediates *Ab* and *aB*. Let X be the frequency of the group of deleterious intermediates ($X = x_1 + x_2$) and Y the frequency of the other group ($Y = x_0 + x_3$). Denote the distribution of X by $\Phi(X)$. In phase 1 when the system waits for a successful double mutant to appear, it will reach a quasi-equilibrium after a short initial period. At this quasi-equilibrium, its distribution is given approximately by

$$\Phi(X) = C \exp(-4NsX) X^{2\theta-1}(1 - X)^{2\theta-1}, \quad (1)$$

where $\theta = 4N\mu$ and C is a constant determined such

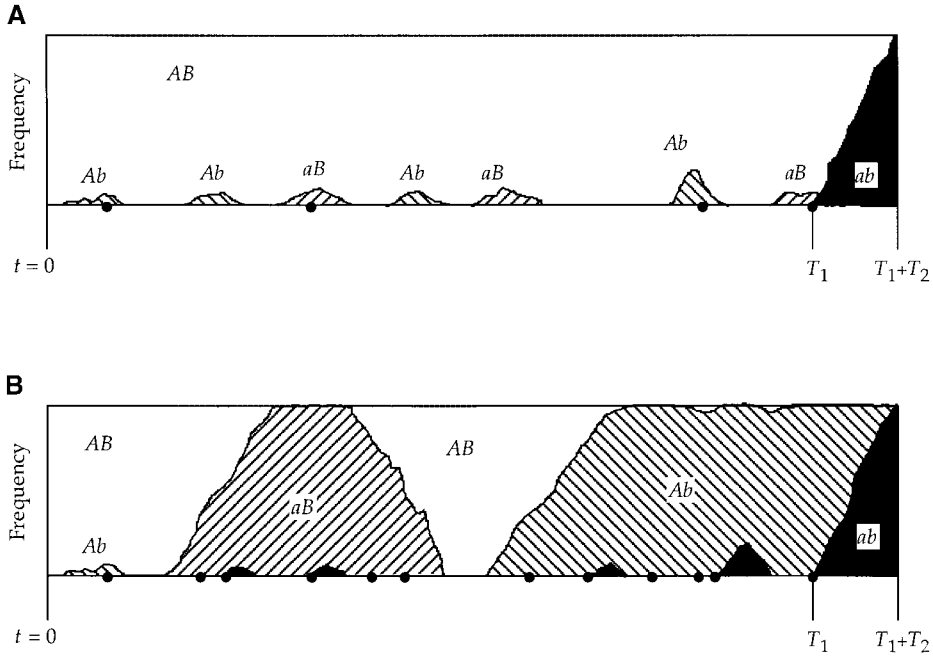

FIGURE 2.—(A) Illustration of the process of compensatory substitution when selection is very strong and the intermediates *Ab* and *aB* are maintained in low frequencies. Solid circles present births of *ab* double mutants. (B) Illustration of the process of compensatory substitution when selection is weak. *Ab* and *aB* sometimes fix in the population.

that $\int_0^1 \Phi(X)\,dX = 1$ (WRIGHT 1931, 1937). Equation 1, however, is not valid very shortly after $t = 0$ because we use the initial condition $X = 0$ (*i.e.*, $Y = 1$).

Thus, we assume that when a new *ab* appears by a mutation in *Ab* or *aB*, the distribution of X before the mutation is given by (1). After the mutation, $Y\,(= 1 - X)$ changes to $Y' = Y + 1/(2N)$. Let p be the fixation probability of the Y group. Since the mutation rate is assumed to be very small, p can be approximated by

$$p = \frac{1 - \exp(-4NsY')}{1 - \exp(-4Ns)} \tag{2}$$

(KIMURA 1962), and the probability, p', that the new *ab* mutant fixes becomes

$$p' = p/(2NY'). \tag{3}$$

Therefore, since a new *ab* appears with probability $2N\mu X$ per generation, the expected number of *ab* that will fix in the population is given by

$$\alpha = 2N\mu \int_0^1 p' X\Phi(X)\,dX, \tag{4}$$

and the waiting time for the appearance of a successful double mutant *ab* becomes

$$T_1 = 1/\alpha. \tag{5a}$$

This result suggests that the waiting time for the appearance of a successful double mutant, T_1, is approximately exponentially distributed as

$$F(T_1 = t) \approx \alpha e^{-\alpha t}. \tag{5b}$$

In the case of neutrality, since p' is $1/(2N)$, the time becomes

$$T_{1\text{neu}} = \frac{1}{\mu \int_0^1 X\Phi(X)\,dX} = \frac{2}{\mu'} \tag{6}$$

because $\int_0^1 X\Phi(X)\,dX = \frac{1}{2}$ in this symmetrical model. In a one-locus neutral model, a substitution occurs every $1/\mu$ generations on average (KIMURA 1983). Thus, $T_{1\text{neu}}$ can be considered as the expected waiting time for two independent neutral substitutions.

When selection is very strong, X is maintained in very low frequency (approximately at the deterministic mutation-selection balance). The expected frequency is then given by

$$\int_0^1 X\Phi(X)\,dX \approx 2\mu/s. \tag{7}$$

If $2\mu/s \ll 1$, p' is $\sim 1/(2N)$ because the average fitness of the population is nearly one. Therefore, T_1 becomes

$$T_1 = s/(2\mu^2). \tag{8}$$

Equation 8 agrees with Equation 8b in STEPHAN (1996), which was obtained for the expected waiting time in Kimura's model of unidirectional mutation pressure by a different method.

T_2 is the time from the appearance of a successful double mutant haplotype to the fixation event. In the case of neutrality, the expectation of T_2 is $\sim 4N$ (KIMURA and OHTA 1969) since we assumed $\mu_i \ll 1/(2N)$. When selection is very strong, T_2 may be close to $4N$ again. That is expected because x_1 and x_2 are very small and *ab* has almost no selective advantage in the population. T_2 for moderate selection intensities is expected to be smaller than $4N$ because *ab* has a selective advantage when x_1 and x_2 are not very small.

General model: Consider a general model where $\mu_1 \neq \mu_2$ and/or $s_1 \neq s_2$. Recombination is again neglected in

the analytical treatment. In this case, it is necessary to investigate the frequency distributions of *Ab* and *aB* separately. Let $\phi_1(x)$ and $\phi_2(x)$ be the frequency distributions of *Ab* and *aB*, respectively. To obtain these two distributions, we reconsider the symmetrical model, where the distribution of the sum of x_1 and x_2 is given by (1). In a population with N diploids, the probability that i haplotypes are *Ab* or *aB* is given by

$$P(i) = \int_{i/2N-1/4N}^{i/2N+1/4N} \Phi(X)\,dX \tag{9a}$$

when $0 < i < 2N$, and

$$P(0) = \int_0^{1/4N} \Phi(X)\,dX \quad \text{and} \quad P(2N) = \int_{1-1/4N}^{1} \Phi(X)\,dX. \tag{9b}$$

Let $P_1(i)$ and $P_2(i)$ be the probability distributions of the numbers of *Ab* and *aB*, respectively. If we assume that μ is so small and Ns is so large that *Ab* and *aB* do not coexist frequently, $P_1(i)$ and $P_2(i)$ are given approximately by

$$P_1(i) = P_2(i) = P(i)/2 \tag{10a}$$

for $i > 0$ and by

$$P_1(0) = P_2(0) = P(0) + \frac{1}{2}\int_{1/4N}^{1} \Phi(X)\,dX. \tag{10b}$$

These results indicate that $\phi_1(x)$ and $\phi_2(x)$ may be given approximately by

$$\phi_1(x) = \phi_2(x) = \Phi(x)/2 \tag{11}$$

for $x > 1/(4N)$.

Next we consider $\phi_1(x)$ and $\phi_2(x)$ in the general model, where $\mu_1 \neq \mu_2$ and/or $s_1 \neq s_2$. If we assume that *Ab* and *aB* do not coexist frequently, the frequencies of *Ab* and *aB* follow two independent distributions. From the arguments in Equations 9–11, it is expected that $\phi_1(x)$ and $\phi_2(x)$ are given for $x > 1/(4N)$ by

$$\phi_1(x) = \Phi_1(x)/2 \quad \text{and} \quad \phi_2(x) = \Phi_2(x)/2, \tag{12}$$

where

$$\Phi_1(x) = C_1 \exp(-4Ns_1x)\,x^{2\theta_2-1}(1-x)^{2\theta_2-1} \tag{13a}$$

and

$$\Phi_2(x) = C_2 \exp(-4Ns_2x)\,x^{2\theta_1-1}(1-x)^{2\theta_1-1}, \tag{13b}$$

where $\theta_1 = 4N\mu_1$ and $\theta_2 = 4N\mu_2$. C_1 and C_2 are constants that are determined such that $\int_0^1 \Phi_1(x)\,dx = 1$ and $\int_0^1 \Phi_2(x)\,dx = 1$ respectively.

Denote by p_1' the fixation probability of a new *ab* produced by a mutation in *Ab* given x_1. Since we assume that *Ab* and *aB* do not coexist in the population, Y is given by $1 - x_1$ before the mutation and $Y' = Y + 1/(2N)$ after the mutation. Then, p_1' is given by

$$p_1' = \frac{1 - \exp(-4Ns_1Y')}{2NY'[1 - \exp(-4Ns_1)]}. \tag{14a}$$

In the same way, the fixation probability of a new *ab* produced by a mutation in *aB* given $Y' = 1 - x_2 + 1/(2N)$ becomes

$$p_2' = \frac{1 - \exp(-4Ns_2Y')}{2NY'[1 - \exp(-4Ns_2)]}. \tag{14b}$$

Therefore, the expected number of *ab* that will fix in the population per generation is given by

$$\alpha = 2N\mu_1\int_0^1 p_1' x_1 \phi_1(x_1)\,dx_1 + 2N\mu_2\int_0^1 p_2' x_2 \phi_2(x_2)\,dx_2, \tag{15}$$

and T_1 is given by

$$T_1 = 1/\alpha. \tag{16}$$

When selection is very strong, T_1 becomes

$$T_1 = s_1s_2/[(s_1 + s_2)\mu_1\mu_2], \tag{17}$$

which agrees with Equation 8b in Stephan (1996). No simple formula for the case of neutrality can be obtained in this way because we assume that Ns_i is so large that *Ab* and *aB* do not coexist frequently.

T_2 for the general model is the same as for the symmetrical model. That is, $T_2 \approx 4N$ when selection is very strong and $T_2 < 4N$ when selection intensity is moderate.

COMPUTER SIMULATIONS

Computer simulations were carried out for the following reasons. The first one is to check the theoretical results for T_1 shown above, because we use the following assumptions in the derivation. We use approximate formulas for the distribution of the haplotype frequencies ignoring the initial condition ($X = 0$ at $t = 0$). In the general asymmetric model, we assume that two haplotypes, *Ab* and *aB*, do not coexist. The second reason is to examine the effect of recombination on T_1 under a broad range of selection coefficients. In contrast to the strong selection case (Kimura 1985; Stephan 1996), we were unable to find analytical expressions for T_1 with recombination when selection is weak. T_2 is also investigated by simulations with and without recombination. Another purpose of the simulations is to evaluate the amount of linkage disequilibrium generated in the process of compensatory substitution.

Monte Carlo simulations with mutation, selection, recombination, and random genetic drift were conducted in a constant size population of N diploids as follows. Each replication of the simulations starts from the initial condition $(x_0, x_1, x_2, x_3) = (1, 0, 0, 0)$. In every generation, the frequencies are determined by the pseudosampling method (Kimura and Takahata 1983). The recombination rate between the two loci is assumed to be r per generation. Every $4N$ generations, the frequencies (x_0, x_1, x_2, x_3) are scored to investigate their frequency distributions. The amount of linkage disequilibrium ($D = x_0x_3 - x_1x_2$) is also calculated. Each replication ends when $(x_0, x_1, x_2, x_3) = (0, 0, 0, 1)$ is reached for the first time, and T_1 and T_2 are recorded. When $0 \leq t' \leq T_2$

(with $t' = t - T_1$), (x_0, x_1, x_2, x_3) is recorded every $N/10$ generation to calculate D.

Computer simulations with no recombination were conducted assuming $N = 100$, $s = s_1 = s_2$, and $\mu = \mu_1 = \mu_2$, and the results for T_1 and T_2 are summarized in Table 1. The theoretical results for T_1 in the symmetrical model were compared with the simulation results with no recombination (Figure 3). In Figure 3A, the theoretical expectation calculated by (5a) is shown with the simulation results of $\theta = 0.01$. Ns was changed from 0 to 5. The theory is in very good agreement with the results of the simulations. Figure 3B shows the results of theory and simulations for a relatively high mutation rate ($\theta = 0.1$) although our model assumes very low mutation rates. It is shown that Equation 5a gives a quite good approximation even for $\theta = 0.1$, unless $Ns = 0$. The underestimation for $Ns = 0$ may occur because frequent recurrent mutations reduced the fixation probability of *ab*. The degree of reduction is expected to be relatively small when T_1 is large. For all parameter sets examined, the standard deviation of T_1 is similar to the mean (Table 1), supporting the exponential distribution of T_1 as suggested by (5b). Similar results were obtained from computer simulations with $N = 1000$ (data not shown).

The effect of recombination on T_1 was also investigated by computer simulations with $\theta = 0.01$ and 0.1 (Table 1). The effect of recombination is very small when selection is weak. The effect, however, can be seen for selection intensity $Ns > 1$. In Figure 4, a clear positive correlation is observed between $4Nr$ and T_1 when $Ns = 2$ and 5. These results are consistent with those of Kimura (1985) and Stephan (1996). Similar relationships between $4Nr$ and T_1 were observed in other two-locus models with epistatic interactions (Michalakis and Slatkin 1996; Christiansen *et al.* 1998).

The results for T_2 are also shown in Table 1. T_2 is much smaller than T_1 in all the parameter sets investigated. First, we consider T_2 with no recombination. In the case of neutrality, T_2 obtained from simulations for $\theta = 0.01$ is close to 400 as expected from the theory, although T_2 for $\theta = 0.1$ is a little $>4N$. When selection is very strong ($Ns \geq 5$), T_2 is close to $4N$ again. T_2 for moderate selection intensity is $<4N$.

T_2 may be negatively correlated with recombination rate (Table 1). The degree of reduction in T_2 is larger when selection is stronger. T_2 for $4Nr = 0$ is similar to that for $4Nr = 10$ when $Ns \leq 1$, while T_2 for $4Nr = 10$ is much smaller than that for no recombination when $Ns \geq 2$. The result for the T_2 phase may be understood as follows. Since recombination usually occurs between *AB* and *ab* in T_2, it reduces the fixation probability of *ab* and increases T_1. As the fixation probability is reduced, only *ab*, which increases its frequency quickly, can successfully fix in the population.

Theoretical results for T_1 in the general model were compared with the results of computer simulations, and the results for $\theta_1 = 0.02$ and $\theta_2 = 0.01$ are shown in Figure 5 with the theoretical expectations calculated from (16). It is shown that the theoretical expectations are in good agreement with the results of simulations when $Ns_1 \geq 0.5$ and $Ns_2 \geq 0.5$. If one of the selection intensities is very small, Equation 16 overestimates T_1 because the assumption used to derive (16) does not hold. In the derivation, we obtained approximate formulas for $\phi_1(x)$ and $\phi_2(x)$ under the assumption that *Ab* and *aB* do not coexist at the same time. If one of the selection intensities, say Ns_1, is very small, $\phi_1(x)$ given by (12) does not agree with the frequency distribution of *Ab* obtained by simulations. Similar results were obtained for other values of θ_1 and θ_2. Good agreement between the theory and simulations was observed when neither Ns_1 nor Ns_2 is small (data not shown).

Table 2 shows the amount of linkage disequilibrium ($D = x_0x_3 - x_1x_2$) in the process of compensatory substitution obtained by simulations under the symmetrical model with no recombination. D was calculated every $4N$ generations as long as $0 < t < T_1$. For this interval of $4N$ generations, we found almost no correlation between sampling. Mean and variance were calculated for all runs. For $\theta = 0.01$ the level of linkage disequilibrium is extremely low. Even for $\theta = 0.1$ D is very small although much larger than that for $\theta = 0.01$. This indicates that almost no linkage disequilibrium is expected during the waiting time for the appearance of a successful double mutant haplotype. Simulations with recombination showed that recombination reduces the level of linkage disequilibrium (data not shown).

On the other hand, strong positive linkage disequilibrium is generated after a successful double mutant has appeared and is on its way to fixation ($0 < t' < T_2$). The average of linkage disequilibrium for $t' \geq 0$ is plotted in Figure 6. Selection increases the level of linkage disequilibrium significantly (Figure 6, A and B). Each distribution of D has a peak near $t' = 100$. As Ns increases, the peak is getting larger and seems to saturate at $D \approx 0.2$. Note that the theoretical maximum value of D is 0.25, which is reached when $x_0 = x_3 = 0.5$ and $x_1 = x_2 = 0$. When selection is weak, the level of linkage disequilibrium is higher for $\theta = 0.1$ than for $\theta = 0.01$.

Figure 6C shows the effect of recombination on the level of linkage disequilibrium when $\theta = 0.01$ and $Ns = 2$. As the recombination rate increases, the level of linkage disequilibrium is getting weaker. Similar results were obtained for other mutation rates and selection intensities (data not shown).

DISCUSSION

Theory: We analyzed a model of reversible mutations with compensatory fitness interactions; *i.e.*, single mutations are assumed to be deleterious but harmless (neutral) in appropriate combinations. In proceeding under mutation pressure, epistatic selection, and genetic drift from one fitness peak to another, a population must pass through a valley of lower individual fitness. This

TABLE 1

Results of computer simulations for T_1 and T_2

			T_1		T_2	
θ	Ns	$4Nr$	Average	SD	Average	SD
0.01	0	0	81426.85	71271.30	409.15	219.93
0.01	0	0.1	80537.43	65183.00	398.57	213.15
0.01	0	1	78167.93	67177.99	398.07	214.43
0.01	0	10	81957.87	68238.48	388.13	210.59
0.01	0.1	0	80883.28	69894.66	399.72	225.44
0.01	0.1	0.1	81386.29	72934.91	402.71	216.01
0.01	0.1	1	84578.92	76346.41	395.08	200.90
0.01	0.1	10	82925.50	71282.31	412.50	227.51
0.01	0.5	0	140676.78	133823.65	374.22	199.43
0.01	0.5	0.1	137427.44	127141.43	375.56	200.14
0.01	0.5	1	145571.54	133570.20	358.46	202.11
0.01	0.5	10	137442.80	125953.56	388.20	208.63
0.01	1	0	493283.25	480379.41	332.75	169.16
0.01	1	0.1	482967.47	460795.95	330.53	162.58
0.01	1	1	500449.93	500982.35	332.07	167.98
0.01	1	10	501808.16	496928.55	326.84	154.32
0.01[a]	2	0	6405239.28	6181700.70	315.72	164.69
0.01[a]	2	0.1	6625633.54	6517420.66	307.46	165.18
0.01[a]	2	1	6815009.08	7214899.60	308.92	164.90
0.01[a]	2	10	7910228.05	7804179.12	277.95	122.08
0.01[b]	5	0	39703724.89	35816596.55	407.11	227.50
0.01[b]	5	0.1	37156482.42	33499185.63	357.58	172.43
0.01[b]	5	1	38959494.54	35755352.26	380.00	194.22
0.01[b]	5	10	143457987.78	133644325.59	263.44	94.02
0.1	0	0	9580.31	8416.24	428.49	232.73
0.1	0	0.1	9578.16	8839.07	419.04	240.69
0.1	0	1	9342.23	8083.69	425.07	234.17
0.1	0	10	9245.41	7741.13	436.99	251.59
0.1	0.5	0	12300.10	11647.45	411.70	226.53
0.1	0.5	0.1	12154.04	11510.89	406.76	226.92
0.1	0.5	1	12322.47	11576.16	418.43	241.57
0.1	0.5	10	12854.10	12573.64	419.30	214.68
0.1	1	0	27265.74	26622.66	387.76	215.25
0.1	1	0.1	27158.13	27295.67	386.77	210.34
0.1	1	1	27157.18	26443.99	381.92	201.11
0.1	1	10	29813.13	28505.98	373.17	180.51
0.1	2	0	109963.11	107380.55	362.89	192.97
0.1	2	0.1	106409.55	110384.68	374.45	200.37
0.1	2	1	111068.48	105695.78	357.52	189.19
0.1	2	10	173029.66	180312.76	313.34	131.74
0.1	5	0	367142.52	383788.82	393.48	214.23
0.1	5	0.1	404861.45	406154.41	389.55	210.28
0.1	5	1	440442.75	427200.82	374.25	192.38
0.1	5	10	1254455.44	1229729.63	270.56	102.06
0.1	10	0	842863.83	873207.49	399.17	222.50
0.1	10	0.1	889375.90	933199.65	384.10	195.38
0.1	10	1	1025621.37	986362.60	363.63	182.71
0.1	10	10	3752636.84	3743833.43	250.16	99.06

The averages and standard deviations of T_1 and T_2 from computer simulations with 1000 replications are shown.

[a] The number of replications is 500.

[b] The number of replications is 200.

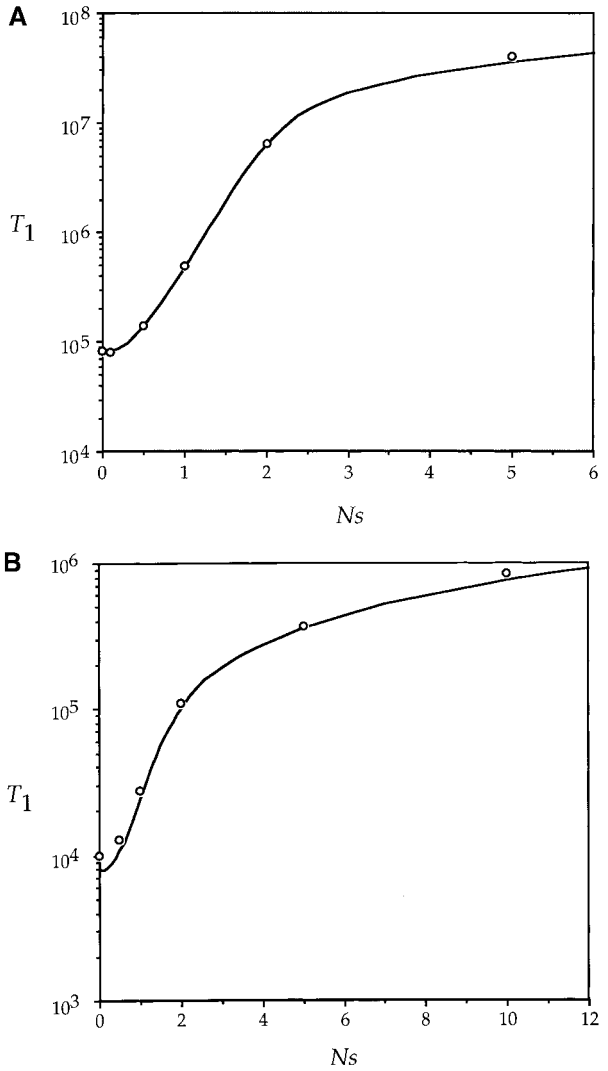

FIGURE 3.—Relationship between T_1 and Ns under the symmetrical model without recombination. The theoretical expectation is obtained from (5a) and represented by a solid curve. The simulation results are based on Table 1 and presented by open circles. (A) Results for $\theta = 0.01$. (B) Results for $\theta = 0.1$.

process of compensatory evolution is investigated by analytical approximation and computer simulation. Our model is more general than KIMURA's (1985), which assumed that mutation pressure is unidirectional and that deleterious intermediates are very strongly selected against. In contrast, our model covers a broad range of selection coefficients, including very small ones, and agrees with analyses of Kimura's model when selection is strong (KIMURA 1985; STEPHAN 1996).

As in these latter analyses, we focus on the expected time for the compensatory substitution process to go from one fitness peak to another. In addition, we study the structure of variation (linkage disequilibrium) within a population during this transition. The process of compensatory substitution is analyzed by dividing it into two phases defined by the time periods T_1 and T_2. T_1 is the waiting time for a successful double mutant

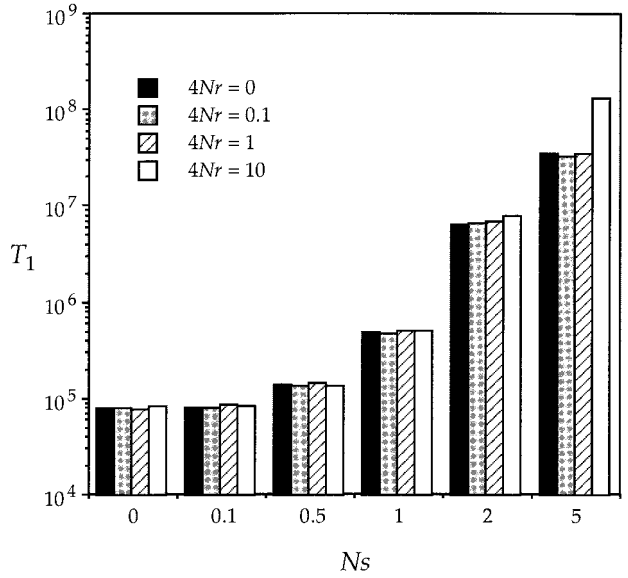

FIGURE 4.—Effect of recombination on T_1 under the symmetrical model with $\theta = 0.01$. The results of computer simulations are from Table 1.

haplotype (*ab*) that will fix in the population, and T_2 is the time from the appearance of a successful double mutant to the fixation event. The results of theory and computer simulations show that $T_1 \gg T_2$ (Table 1) as suggested by STEPHAN (1996).

The expectation of T_1 is obtained analytically for the case without recombination. It is shown that selection

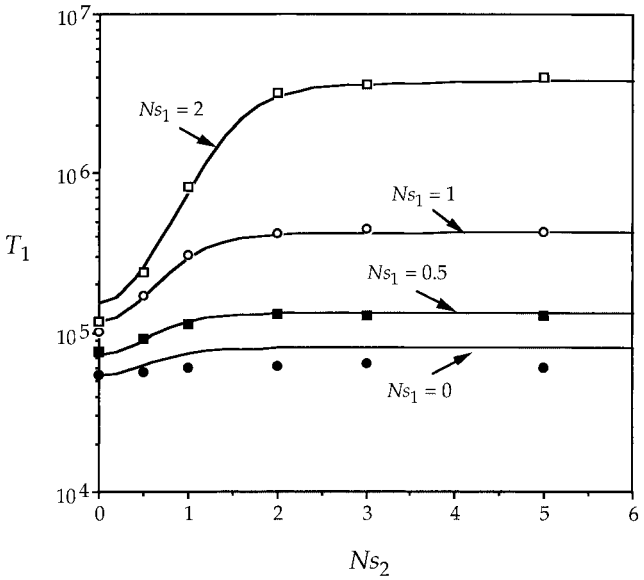

FIGURE 5.—Relationship between T_1 and selection intensities Ns_1 and Ns_2 under the general model without recombination. The theoretical expectations are obtained from (16) and represented by solid curves. $\theta_1 = 0.02$ and $\theta_2 = 0.01$ are assumed. Solid circles, solid squares, open circles, and open squares represent the results of simulations for $Ns_1 = 0$, 0.5, 1, and 2, respectively.

TABLE 2

Results of computer simulations for linkage disequilibrium

θ	*Ns*	Average ($\times 10^{-6}$)	Variance ($\times 10^{-6}$)
0.01	0	−12.10	13.73
0.01	0.1	−0.17	12.37
0.01	0.5	27.93	9.86
0.01	1	19.30	4.56
0.01	2	13.03	2.60
0.01	5	5.34	0.96
0.1	0	97.71	713.98
0.1	0.5	1325.20	591.55
0.1	1	1652.73	419.46
0.1	2	1164.52	218.09
0.1	5	475.63	82.89
0.1	10	221.47	37.91

dramatically increases T_1 (Figure 3). Although the theory is based on the assumption of very low mutation rates, it is shown that it is also applicable to very high nucleotide mutation rates relative to $1/N$ (*e.g.*, $\theta = 0.1$) unless *Ns* is very small. Thus, our analysis is useful for natural populations for which θ is generally $\ll 0.1$ (Kimura 1983; Nei 1987; Gillespie 1991). Our computer simulations demonstrate that there is almost no effect of recombination on T_1 when selection is relatively weak ($Ns \leq 1$), while T_1 increases significantly with increasing recombination rates when selection is strong (Figure 4).

Parameter estimation: An important parameter of the compensatory evolution model is the intensity of selection *Ns* against deleterious intermediates. In the following, we attempt to estimate this parameter for mRNA secondary structures. Our theoretical results predict that T_1 is much larger than T_{1neu} unless *Ns* is very small (Figure 3), suggesting that nucleotide substitutions occur very slowly in pairing regions of mRNA secondary structures. It is known that such regions are highly conserved between species in contrast to other (unpaired) regions (Muse 1995; Parsch *et al.* 2000). Thus, it may be possible to estimate *Ns* in pairing regions from DNA sequence comparisons.

We compare the rates of substitutions between species in pairing regions with those in regions that are considered selectively neutral. As an example, we analyze the *bicoid* 3′ UTR of Drosophila. It has been shown that *bicoid* mRNA has a complex secondary structure in the 3′ UTR (Macdonald 1990; Seeger and Kaufman 1990; Ferrandon *et al.* 1997; Macdonald and Kerr 1998). Based on the alignment of DNA sequences from nine Drosophila species, Parsch *et al.* (2000) identified eight highly conserved pairing regions, of which seven have been supported by mutational analysis (Ferrandon *et al.* 1997; Macdonald and Kerr 1998). We consider these eight stems as pairing regions and the remainder of the 3′ UTR as unpaired. It is also assumed that there

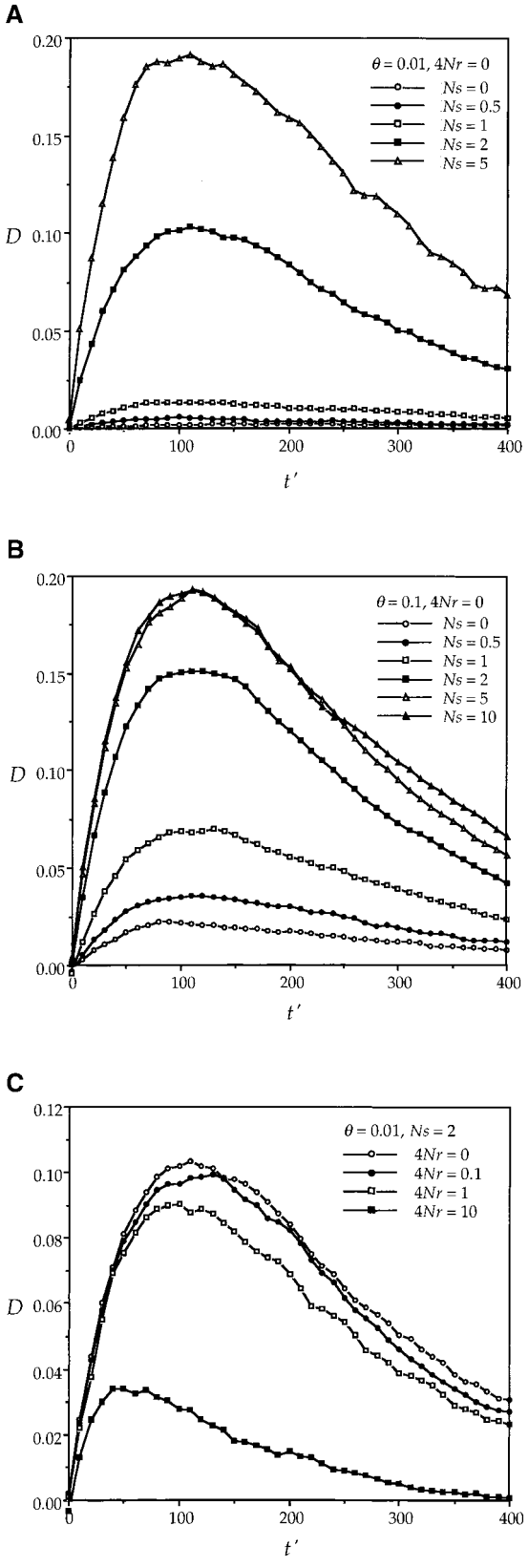

Figure 6.—Linkage disequilibrium in phase T_2. (A) Results of computer simulations without recombination for $\theta = 0.01$. The effect of selection intensity is investigated. (B) Results of computer simulations without recombination for $\theta = 0.1$. (C) Results of computer simulations for $\theta = 0.01$ and $Ns = 2$. The effect of recombination is investigated.

TABLE 3

Summary of nucleotide differences in the *bicoid* 3′ UTR of Drosophila

Species compared	*mel/sim*[a]	*mel/pse*[a]
Total		
No. of nucleotides compared	875	743
No. of nucleotide differences	34	248
No. of substitutions per site[b]	0.0399	0.4416
Pairing regions		
No. of pairs of complementary sites	71	71
No. of different pairs (WC/WC)[c]	0	7
No. of different pairs (WC/WC)[d]	3	5(1)[e]
No. of different pairs (WC/NO)[f]	0	5(1)[g]
No. of nucleotide differences	3	26
No. of substitutions per site[b]	0.0214	0.2099
Unpaired regions		
No. of sites compared	733	601
No. of differences	31	222
No. of substitutions per site[b]	0.0435	0.5087

The aligned sequence data are from Parsch *et al.* (2000).

[a] *mel*, *D. melanogaster*; *sim*, *D. simulans*; *pse*, *D. pseudoobscura*.

[b] The expected number of substitutions was calculated by Jukes and Cantor's (1969) method.

[c] The number of pairs of complementary nucleotide sites where both species have different Watson-Crick (WC) pairs. The minimum number of nucleotide changes is two.

[d] The number of pairs of complementary nucleotide sites where one species has a WC pair and the other has a GU wobble (WO) pair. The minimum number of nucleotide differences is usually one, but see below for an exception.

[e] Out of five WC/WO differences, one requires at least two nucleotide changes between the two species, where *D. melanogaster* has a GU pair and *D. pseudoobscura* has a UA Watson-Crick pair.

[f] The number of pairs of complementary nucleotide sites where one species has a WC pair and the other does not have a WC pair or a GU wobble pair (NO pair). The minimum number of nucleotide differences is usually one, but see below for an exception.

[g] Out of five WC/NO differences, one requires at least two nucleotide changes between the two species, where *D. melanogaster* has an AC pair and *D. pseudoobscura* has a UA Watson-Crick pair.

is no selection in the unpaired regions. The total length of the paired segments is 142 nucleotides, corresponding to 71 bp.

We first compare the nucleotide sequences between *D. melanogaster* and *D. simulans*, using the alignment suggested by Parsch *et al.* (2000). In 142 nucleotides of the pairing regions, 3 nucleotide differences are observed and the number of substitutions per site (d_p) is estimated to be 0.0214 by the Jukes and Cantor (1969) method (Table 3). In the remaining regions of the 3′ UTR, 31 nucleotide differences are observed and the number of substitutions per site (d_n) is estimated to be 0.0435. This suggests that the rate of nucleotide substitutions is reduced by roughly a factor of 2 in the pairing regions in comparison with the remainder of the 3′ UTR. For the pair of *D. melanogaster* and *D. pseudoobscura*, d_p and d_n are estimated to be 0.2099 and 0.5087, respectively. The ratio of d_n to d_p is ~2.4, similar to that of the comparison between *D. melanogaster* and *D. simulans*.

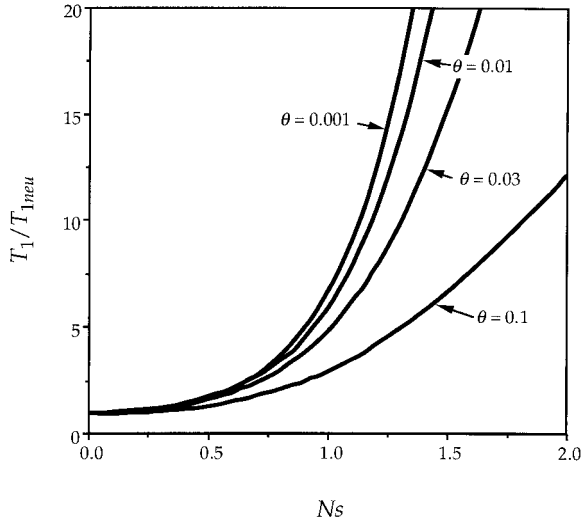

Figure 7.—Relationship between T_1/T_{1neu} and Ns for various values of θ.

Since it is assumed that the unpaired regions are selectively neutral, the ratio of d_n to d_p is comparable to T_1/T_{1neu}. In Figure 7, T_1/T_{1neu} is plotted as a function of Ns for the symmetrical model without recombination. T_1/T_{1neu} is calculated by (5a) and (6). The results show that T_1/T_{1neu} increases rapidly with increasing Ns, as soon as $Ns > 1$. This is particularly the case for $\theta \leq 0.01$. T_1/T_{1neu} depends weakly on θ when $\theta \leq 0.01$.

In the *bicoid* 3′ UTR of Drosophila (Table 3), the ratio of d_n to d_p is ~2.0–2.4. The estimate of θ in the unpaired regions for a *D. melanogaster* population from Zimbabwe is ~0.003 (J. F. Baines, Y. Chen and W. Stephan, unpublished results). Figure 7 suggests therefore that the observed ratio of d_n to d_p can be explained if Ns is ~0.6–0.7. If we consider only the number of complete compensatory substitutions (WC/WC in Table 3) for the comparison between *D. melanogaster* and *D. pseudoobscura*, the ratio of d_n to d_p becomes ~5 and the estimate of $Ns \approx 1$.

There are, however, some *caveats*.

1. Ns could be larger than this estimate because d_n is underestimated if selection is acting in the regions that we consider as unpaired. There may be some evidence for weak selection in these regions. First, Macdonald (1990) suggested the possibility of long-range pairings encompassing almost the entire 3′ UTR. However, his suggestion was not supported by a strict phylogenetic analysis (Parsch *et al.* 2000). Second, average silent divergence in the unpaired segments of the 3′ UTR between *D. melanogaster* and *D. simulans* is ~0.0435, which is about a factor of 2.5

lower than in the rest of the *bicoid* gene upstream of the 3′ UTR (J. F. Baines, Y. Chen and W. Stephan, unpublished results). On the other hand, even if weak selection is acting in the unpaired regions, the estimate of *Ns* for the pairing regions does not increase much, as the average time of compensatory substitutions becomes extremely large for $Ns > 1$ when mutation pressure is relatively weak (*i.e.*, $\theta \leq 0.01$; Figure 7). Thus, it may be concluded that the selection intensity in the pairing regions of the *bicoid* 3′ UTR is on average not much >1. This estimate is similar to the estimate of average selection intensity for codon usage in Drosophila (Akashi 1995).

2. To estimate the selection intensity, we used the average d_p of eight pairing regions. In other words, *Ns* is the average selection intensity of these eight pairing regions. Parsch *et al.* (2000) found heterogeneity for d_p among pairing regions, caused by variation in both stem length and the physical distance between base-pairing residues. One reason is that long stems are under less selective constraints than short ones (see Figure 3A of Parsch *et al.* 2000). Another factor is that short-range pairings (hairpins) experience a higher rate of evolution than long-range pairings because of the retarding effects of recombination when selection is sufficiently strong (Figure 4A of Parsch *et al.* 2000). As a consequence, the estimates of *Ns* appear to vary substantially among pairing regions.
3. Parsch *et al.* (2000) were able to distinguish the effect of stem length from that of physical distance when only pairing regions with covariations were considered. According to their Figure 4A, they found an approximately fivefold drop in the rate of compensatory evolution over a physical distance of nearly 200 bp between base-pairing residues. We have to ask whether such a large decrease of the rate of compensatory evolution is consistent with an estimate of $Ns \approx 1$. Assuming that a physical distance of 200 bp of the *bicoid* 3′ UTR corresponds to a value of $4Nr$ in the order of 10 (*i.e.*, using the standard estimates of effective population size and recombination rate for *D. melanogaster* that are in the order of 10^6 and 10^{-8}, respectively), this distance effect can be explained by our model only if *Ns* is ~5 (see Table 1). Thus, it appears that this value is not compatible with our estimate of $Ns \approx 1$ obtained without taking recombination into account. However, one has to keep in mind that for technical reasons the analysis of Parsch *et al.* (2000) is based on pairing regions *with* covariations only. A much weaker distance effect would presumably result if all pairing regions were considered, including those with no covariations (Parsch *et al.* 2000). A much larger data set and more sophisticated methods are needed to take the distance effect into account.

Linkage disequilibrium: Strong linkage disequilibrium is sometimes considered as evidence for epistatic selection (Lewontin 1974). We investigated the amount of linkage disequilibrium in the process of compensatory substitution. It is demonstrated that the level of linkage disequilibrium is very low during time period T_1. Strong positive linkage disequilibrium, however, is observed in phase T_2 if selection is strong. This suggests that significant linkage disequilibria due to compensatory interactions should be rarely observed in natural populations because T_2 is much smaller than T_1 if selection is strong. On the other hand, if selection is weak, linkage disequilibrium is not very large even in phase T_2.

Schaeffer and Miller (1993) detected two clusters of polymorphisms in the *Adh* introns of *D. pseudoobscura* that exhibit significant linkage disequilibrium. In both cases, the disequilibria are due to two highly diverged haplotypes, ha1 and ha2, that have been shown to form different pre-mRNA secondary structures (Kirby *et al.* 1995). It was also revealed that this structural polymorphism has predated the species split of *D. pseudoobscura*, *D. persimilis*, and *D. miranda* because ha1 and ha2 are similar to *D. persimilis* and *D. miranda* haplotypes, respectively (Kirby *et al.* 1995). This observation is not consistent with our results, which show that a compensatory substitution requires a long waiting time for a successful double mutant to occur and that the fixation event of *ab* follows relatively quickly. In other words, our model does not predict that the secondary-structure-forming haplotypes of *Adh* are maintained for such a long time, as observed in these species. This suggests that our model of compensatory evolution is either too simple, as it allows only two sites to undergo base changes, or that some additional form of selection (for instance, balancing selection) may be maintaining the haplotypes ha1 and ha2. While there is no evidence for the latter suggestion, the fact that in both examples the haplotypes were subject to significant rearrangement during evolutionary time (due to insertions and deletions of bases) may indicate that the underlying compensatory process is much more complicated than our two-locus model assumes. Therefore, to model such complex compensatory changes, models need to be developed that include compensatory insertions and deletions in addition to base substitutions.

We thank an anonymous reviewer for helpful suggestions that improved the manuscript. John F. Baines and Ying Chen kindly provided unpublished results. This research was supported in part by a fellowship from the Japan Society for the Promotion of Science to H.I. and by National Institutes of Health grant GM-58405 to W.S.

LITERATURE CITED

Akashi, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139:** 1067–1076.

Christiansen, F. B., S. P. Otto, A. Bergman and M. W. Feldman,

1998 Waiting with and without recombination: the time to production of a double mutant. Theor. Popul. Biol. **53:** 199–215.

Ferrandon, D., I. Koch, E. Westhof and C. Nüsslein-Volhard, 1997 RNA-RNA interaction is required for the formation of specific *bicoid* mRNA 3′ UTR-STAUFEN ribonucleoprotein particles. EMBO J. **16:** 1751–1758.

Gillespie, J. F., 1991 *The Causes of Molecular Evolution*. Oxford University Press, London.

Golding, B., 1994 Using maximum likelihood methods to infer selection from phylogenies, pp. 126–139 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

Higgs, P. G., 1998 Compensatory neutral mutations and the evolution of RNA. Genetica **102/103:** 91–101.

Higgs, P. G., 2000 RNA secondary structure: physical and computational aspects. Q. Rev. Biophys. **33:** 199–253.

Iizuka, M., and M. Takefu, 1996 Average time until fixation of mutations with compensatory fitness interaction. Genes Genet. Syst. **71:** 167–173.

Jukes, T. H., and D. R. Cantor, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Kimura, M., 1962 On the probability of fixation of mutant genes in a population. Genetics **47:** 713–719.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

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

Kimura, M., and T. Ohta, 1969 The average number of generations until fixation of a mutant gene in a finite population. Genetics **61:** 763–771.

Kimura, M., and N. Takahata, 1983 Selective constraint in protein polymorphism: study of the effectively neutral mutation model by using an improved pseudosampling method. Proc. Natl. Acad. Sci. USA **80:** 1048–1052.

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.

Lewontin, R. C., 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Macdonald, P. M., 1990 *bicoid* mRNA localization signal: phylogenetic conservation of function and RNA secondary structure. Development **110:** 161–171.

Macdonald, P. M., and K. Kerr, 1998 Mutation analysis of an RNA recognition element that mediates localization of *bicoid* mRNA. Mol. Cell. Biol. **18:** 3788–3795.

Michalakis, Y., and M. Slatkin, 1996 Interaction of selection and recombination in the fixation of negative-epistatic genes. Genet. Res. **67:** 257–269.

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

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Parsch, J., J. M. Braverman and W. Stephan, 2000 Comparative sequence analysis and patterns of covariation in RNA secondary structures. Genetics **154:** 909–921.

Rousset, F., M. Pélandakis and M. Solignac, 1991 Evolution of compensatory substitutions through a G · U intermediate state in *Drosophila* rRNA. Proc. Natl. Acad. Sci. USA **88:** 10032–10036.

Rzhetsky, A., 1995 Estimating substitution rates in ribosomal RNA genes. Genetics **141:** 771–783.

Savill, N. J., D. C. Hoyle and P. G. Higgs, 2001 RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. Genetics **157:** 399–411.

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.

Schöniger, M., and A. von Haeseler, 1994 A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phylogenet. Evol. **3:** 240–247.

Seeger, M. A., and T. C. Kaufman, 1990 Molecular analysis of the *bicoid* gene from *Drosophila pseudoobscura*: identification of conserved domains within coding and noncoding regions of the *bicoid* mRNA. EMBO J. **9:** 2977–2987.

Stephan, W., 1996 The rate of compensatory evolution. Genetics **144:** 419–426.

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

Tillier, E. R. M., and R. A. Collins, 1995 Neighbor-joining and maximum likelihood with RNA sequences: addressing interdependence of sites. Mol. Biol. Evol. **12:** 7–15.

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

Wright, S., 1937 The distribution of gene frequencies in populations. Proc. Natl. Acad. Sci. USA **23:** 307–320.

Communicating editor: G. B. Golding