Time for Spreading of Compensatory Mutations Under Gene Duplication

Tomoko Ohta

National Institute of Genetics, Mishima 411, Japan
Manuscript received February 20, 1989
Accepted for publication July 14, 1989

ABSTRACT

Evolution by compensatory mutations is accelerated by gene duplication because selective constraint is relaxed by gene redundancy. A mutation is called compensatory if it corrects the effect of an earlier deleterious mutation. Without duplication, Kimura has shown that the time for spreading of compensatory mutations is much reduced by tight linkage between the two chromosomal sites of mutations. In this report, the time for spreading with gene duplication was studied by using the diffusion equation method of Kimura, together with computer simulations. It was shown that, when $2N\nu_-$ is much less than unity, the time for spreading is greatly shortened by gene duplication as compared with the case of complete linkage between the two sites of mutations, where $2N$ is the effective population size (haploid) and $\nu_-$ is the rate of compensatory mutations. However, if $2N\nu_- > 1$, gene duplication is not effective for accelerating the evolution by such mutations.

Gene duplication is now considered to occur fairly frequently in eukaryote chromosomes. Previous studies suggest that the rate of gene duplication is $10^{-4} \sim 10^{-3}$ per ordinary gene locus (GELBART and CHOY 1979; SHAPIRA and FINNERTY 1986; MARONI et al. 1987), which is not negligibly low as compared with the nucleotide substitution rate, and gene duplication provides important material for evolution. Indeed, numerous examples of utilizing gene duplication and subsequent differentiation for organismal evolution have been reported (for a recent review, see OHTA 1988c). I have been studying the role of gene duplication in evolution from the standpoint of population genetics (OHTA 1987a,b; 1988a,b). One role is to provide relaxation of selective constraint by means of gene redundancy (OHNO 1970; KIMURA 1983). In my previous report (OHTA 1988d), it has been shown, by extensive Monte Carlo simulations, that compensatory mutant substitution is accelerated through relaxation when gene duplication is incorporated. In the present paper, a more quantitative analysis is presented by using KIMURA's diffusion equation method (KIMURA 1985a,b), together with computer simulations. It will be shown that, when the rate of compensatory mutations is low, such that the product of the mutation rate and the population size is much less than unity, the time for spreading of the mutants is greatly shortened by gene duplication.

BASIC THEORY

Compensatory mutations are individually deleterious, but become neutral or advantageous when two mutations accumulate in the same individual. In this paper, compensation is assumed to occur in the same gene. Such a mutant is assumed to occur with the constant rate, $\nu_-$, per gene per generation. In the present study, no completely neutral mutations are assumed. Initially, there is a single gene in a haploid individual, and with the constant rate, $\gamma$, per gene copy per generation, a gene is either duplicated or deleted by interchromosomal unequal crossing over. If unequal crossing over results in the total loss of the gene, the individual becomes lethal. A finite haploid population with effective size $2N$ is assumed, and no interchromosomal crossing over is considered. The above model of compensatory mutations is the same as my previous one (OHTA 1988d), but in this study, neither additional neutral mutation nor sexual recombination is introduced. Thus the model is simpler than the previous one, but is similar to the two-locus model of compensation by KIMURA (1985a,b) and may be treated analytically under certain conditions.

A gene carrying a single mutation lowers the fitness of the individual by $s_-$. If an individual has the gene, the fitness of this individual becomes as follows.

$$w_{-i} = \exp(-\gamma s_-),$$

where the subscript, $i$, denotes the $i$th individual. When genes are redundant, relaxation of selective constraint is introduced, i.e., a deleterious mutation becomes neutral so long as one or more normal genes exist on the same chromosome. This model is intended for those genes that are tissue specific, and produce small amounts of proteins, such as genes of various membrane proteins. For them, the amount of gene products would not be critical for organisms, and once genes become redundant, deleterious effects
of mutations in one of them may be covered by the activity of the remaining one or more functional genes. On the other hand, for gene families that produce large amounts, the effect of deleterious mutations would be more difficult to be covered. This is especially so for gene families with uniform members such as genes of ribosomal RNA or histone. For them, another approach that is different from the present one, is required for understanding evolution of compensatory mutations, that seem to be fairly common (Brimacombe 1984).

When a gene has two mutations, they compensate their deleterious effect and become neutral or advantageous, so that genes contain two mutations. An additional (third) mutation is deleterious but may be again compensated by the fourth mutation. Thus compensatory mutations accumulate two at a time. The model of positive selection is similar to the previous one (Ohta 1987a,b; 1988a,b,d), i.e., if the number of alleles affected by compensatory mutations in an individual is more than the population average, such an individual enjoys a selective advantage according to the fitness function,

$$w_{+,k} = \begin{cases} 1 & \text{for } k_i \geq k \\ \exp[-s_+(k - k_i)] & \text{for } k_i < k \end{cases}$$

where $k_i$ is the number of alleles with an even number of mutations in the $i$th individual, $k$ is the population average, and $s_+$ is the coefficient of positive selection. Note that the term “allele” is used here for genes belonging to any locus in the family of duplicated loci as in my previous reports. Also, a slight modification on counting the number of alleles in an individual is incorporated, i.e., the original allele without mutations is not counted as an allele here whereas it was counted in my previous study. The purpose of this modification is to enable examination of the case when there is no gene duplication, to which Kimura’s (1985a,b) theory may be directly applied.

When genes are redundant, relaxation of selective constraint is introduced, i.e., a deleterious mutation becomes neutral so long as one or more normal genes exist on the same chromosome. Under such an assumption, it is convenient to partition the process of compensatory mutations into the following three phases: 1) spreading of a duplication-bearing chromosome into the population; 2) spreading of a deleterious mutation, that occurs in one of the duplicated genes and is therefore neutral, into the population; and 3) spreading of a compensatory mutation, that occurs in one of the genes with the previous mutant, into the population. These phases may be treated separately when $2N\gamma$ and $2N\nu_-$ are much less than unity. Here it is assumed that the products of population size and unequal crossing-over rate $(2N\gamma)$, or mutation rate $(2N\nu_-)$ are much less than unity. Later, this assumption is weakened.

The first phase is the same as in the previous model (Ohta 1988b) and is analogous to the spreading of a neutral mutant under mutation pressure. Kimura’s formula gives such time in number of generations, and becomes for this phase

$$t_1 \approx \frac{4N}{V_1 - 1} \left[ 0.577 + \psi(V_1) \right]$$

where $V_1 = 2N\gamma$ and $\psi(\cdot)$ stands for the digamma function. (For details, see derivation of Equation 13 of Kimura, 1985a.)

The second phase is again the spreading of a neutral mutant, but $V_2$ should now be replaced by $V_2 = 4N\nu_-$, since there are two genes in which mutation may occur.

$$t_2 \approx \frac{4N}{V_2 - 1} \left[ 0.577 + \psi(V_2) \right].$$

If the mutants are compensatory neutral ($s_+ = 0$), the third phase is also neutral spreading, but $V_3$ should be replaced by $V_3 = 4N\nu_-$, because the second mutation must occur in the gene with the first mutant. When mutants are compensatory advantageous, the third phase becomes the spreading of an advantageous mutation under mutation pressure. In the present model of positive selection, the increase of mutant frequency becomes the same as that for a recessively advantageous mutant (Ohta 1987a), and Equations 7–9 of Kimura (1985a) may be used. Thus, the third phase becomes,

$$t_3 \approx \frac{4N}{V_3 - 1} \left[ 0.577 + \psi(V_3) \right], \text{ for } s_+ = 0 \quad (5)$$

$$t_3 \approx 4N \int_0^1 e^{-S\eta} \eta^{-V_3} d\eta \int_0^\eta \frac{e^{S\xi} \xi^{V_3-1}}{1 - \xi} d\xi, \quad \text{for } s_+ > 0 \quad (6)$$

where $S = 2N\eta_+$ and $V_3 = 4N\nu_-$. The total time for spreading of a compensatory mutation under gene duplication is the sum of the above values.

$$T_d = t_1 + t_2 + t_3. \quad (7)$$

Now, when there is no gene duplication, the theory for estimating the time is again found in Kimura (1985a,b). The present model is analogous to his case of complete linkage, but additional selective force should be incorporated for compensatory advantageous mutations. Equation 7 of Kimura (1985a) as applied to the present problem becomes,

$$T_0 \approx 4N \int_0^1 e^{-K\eta} \eta^{-V} d\eta \int_0^\eta \frac{e^{K\xi} \xi^{V-1}}{1 - \xi} d\xi, \quad (8)$$
TABLE I

<table>
<thead>
<tr>
<th>2Nα</th>
<th>2Nγ</th>
<th>2Nv-</th>
<th>Time in units of 4N generations</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>0.1</td>
<td>611.9 ± 574.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>143.0 ± 146.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>57.0 ± 52.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.4</td>
<td>34.9 ± 31.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>23.7 ± 20.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.0</td>
<td>4.9 ± 3.2</td>
</tr>
<tr>
<td>0.2</td>
<td></td>
<td>0.1</td>
<td>20.2 ± 15.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>11.6 ± 5.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>10.1 ± 7.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.4</td>
<td>8.8 ± 5.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>6.4 ± 4.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.0</td>
<td>4.9 ± 2.7</td>
</tr>
<tr>
<td>20</td>
<td>0.0</td>
<td>0.1</td>
<td>101.7 ± 77.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>25.7 ± 28.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>9.5 ± 8.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.4</td>
<td>6.3 ± 6.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>3.9 ± 3.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.0</td>
<td>1.6 ± 1.2</td>
</tr>
<tr>
<td>0.2</td>
<td></td>
<td>0.1</td>
<td>11.3 ± 7.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>5.7 ± 3.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>4.6 ± 3.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.4</td>
<td>3.2 ± 3.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>2.6 ± 1.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.0</td>
<td>1.2 ± 0.8</td>
</tr>
</tbody>
</table>

observed* | expected

* Mean ± SD of 40 replications. Other parameter: 2Nα = 10 and 2Nβ = 100.

where \( V = 2A^2/(D + 3A) \) and \( B(\xi) = 2N\alpha\xi^2 + 2AD\xi/(D + 3A) \), with \( A = 2N\alpha_\beta \) and \( D = 4N\alpha_\beta \). Results of numerical studies by using Equations 7 and 8 will be given in the next section.

NUMERICAL ANALYSES AND MONTE CARLO SIMULATIONS

In order to check the validity of the above results, Monte Carlo simulations were performed. The simulated population is made of 2N haploid individuals. At the start, each individual has a single locus free of mutation. The method of simulation was the same as described in Ohta (1987a,b; 1988a,b,d). Each generation consisted of unequal crossing over, mutation, random sampling and selection. Deleterious mutations were stored by negative integers, and compensation was done by changing two negative integers into positive ones. If a third mutation occurs in a gene with two compensatory mutations, the gene becomes deleterious again. Selection was carried out by combining the two fitness functions (1) and (2).

All experiments were carried out until all individuals in the population contained at least one gene with compensatory mutations. The number of generations needed until the above condition was attained was recorded. Forty replicate runs were performed in each set of parameters.

Table 1 gives the comparison of the observed (Monte Carlo simulation) and the expected (Equations 7 and 8) time for spreading of compensatory mutations. The figures are in units of 4N generations. Four classes examined are combinations of with and without gene duplication, and of with and without positive selection. As can be seen from the table, the agreement between the observed and the expected values are generally satisfactory. Note that the standard deviation is almost as large as the mean, but that the standard errors of the observed values become \( 1/\sqrt{40} \approx 0.063 \) of the listed values of the standard deviation. Since the distribution of the time for spreading is expected to be highly skewed like that of a neutral mutant infinite populations (Kimura 1970), it is difficult to carry out statistical tests of significance here.

A few points need caution. One is on the assumption of \( 2N\gamma \ll 1 \) and \( 2Nv- \ll 1 \). This condition is not satisfied in some sets of parameters examined. Then the three phases are not separate as described above, but overlap. Thus Equation 7 overestimates the time for spreading as may be seen from the disagreement between the observed and the theoretical time when \( \gamma > 0 \) and \( 2Nv- = 0 \) or more. This is especially evident when \( \gamma > 0 \), since the time estimated by Equation 7 is longer than that by Equation 8, which can not be true.

Another point that needs attention is the case when \( 2Nv- \ll 1 \) and \( \gamma = 0 \). Kimura's formula appears to slightly overestimate the time under such circum-
stance, presumably because the derivation is based on the assumption that the frequency of individuals with a single (deleterious) mutation is kept low by mutation-selection balance, while the second (compensatory) mutant increases its frequency. When $2Nv_\gamma \ll 1$, this balance is not stable, and may cause deviation from the theoretical prediction.

Next, let us examine how the time for spreading of compensatory mutations is influenced by various parameter values through numerical analyses of Equations 7 and 8. Figure 1 shows the time ($T$, in units of $4N$ generations) as functions of $2Nv_\gamma$, with and without gene duplication. It is very clear from the figure that the time gets longer as $2Nv_\gamma$ becomes smaller, especially when $\gamma = 0$. Indeed, KIMURA's formula tells that the time is inversely proportional to $v_\gamma$. When $\gamma > 0$, however, it is simply proportional to $1/v_\gamma$. Also note that, as mentioned earlier, when $2N\gamma$ is $0.4$ or more, the time for $2N\gamma = 0.2$ overestimates the true value.

Figure 2 shows the time as functions of $2N\gamma$, with and without positive selection. As can be seen from the figure, unequal crossing-over greatly shortens the time especially for compensatory neutral mutations ($s_+ = 0$). When $s_+ > 0$, the effect of gene duplication is not so pronounced, as the large difference of the two curves is seen for $2N\gamma = 1$.

Figure 3 shows the time for spreading as functions of $2Ns_\gamma$. When $\gamma = 0$, the time linearly increases with $2Ns_\gamma$. However when $\gamma > 0$, the time is independent of $2Ns_\gamma$, since the process of spreading is composed of three selectively neutral phases. Figure 4 represents the time as functions of $2Ns_\gamma$, with and without gene duplication. When $\gamma = 0$, the time is inversely proportional to $2Ns_\gamma$, but when $2N\gamma = 0.2$, the effect of positive selection is very small. This is because the third phase is accelerated by positive selection.

**DISCUSSION**

The present analyses show clearly that gene duplication greatly shortens the time for spreading compensatory mutations when the mutation rate is low such that $2Nv_\gamma \ll 1$. This condition would be usually satisfied as discussed below. Several examples of molecular compensatory mutations, as discussed by OHTA (1974), KIMURA (1985a,b), and OHTA (1988d), include amino acid changes of pancreatic ribonuclease (WYCKOFF 1968), tryptophan synthetase (YANOFSKY, HORN and THORPE 1964), and ferredoxin (TSUKIHARA et al. 1982), and base changes of tRNA (MUTO, YAMAO and OSAWA 1987; JUKES 1985). These examples are likely to be the results of compensatory mutant substitutions, however, some clustered amino acid substitutions in a short region may be due to gene conversion of short segments (POWERS and SMITHIES 1986; MELLOR et al. 1983). Thus one should be careful in evaluating the role of compensatory mutant substitutions.

In the above examples, the compensatory second site change corrects the specific defect of the original mutation. KLIG, OXENDER and YANOFSKY (1988) further studied the nature of molecular compensation at the tryptophan synthetase locus of *Escherichia coli*, and found that a substantial fraction of amino acid changes that resulted in a revertant phenotype were those changes previously described as “superactive.” Based on this finding, they concluded that many second-site
compensatory changes act globally at the molecular level, i.e., they do not correct the original defect specifically. Thus there appear to be two types of compensatory mutations at the molecular level; specific and global compensations.

On the other hand, it is now known that the mutation rate of nucleotide substitution is generally $10^{-9} \sim 10^{-6}$ per site per year (Kimura 1983). The effective population size of higher organisms is usually $10^4 \sim 10^6$ (Kimura 1983). If the generation time of many higher organisms on an evolutionary scale is roughly one year, the important parameter, the product of mutation rate and population size, would be $10^{-2} \sim 10^{-5}$ per nucleotide site. The number of nucleotide sites related to molecular compensation would be rather small and at most of the order of ten at any locus. Then the value of $2Nv_m$ of our model would be much less than unity, and gene duplication may greatly accelerate the evolution by compensatory mutations.

In the present model, relaxation of selective constraint is assumed, i.e., when at least one gene copy free of mutation exists on the chromosome, deleterious mutations that occur in redundant gene copies become neutral. The acceleration of the mutant spreading is caused by this assumption, and gene duplication may not be so effective as estimated here. However, if the value of $2Nv_m$ is extremely small, compensatory mutations will almost never spread without duplication. Once redundant genes have been attained, selective penalties may be lessened even if it may take some time after duplication, and compensatory mutations may increase their frequencies.

One may conclude, from the present results, that gene duplication provides favorable opportunities for evolution by compensatory mutations when $2Nv_m \ll 1$. Kimura (1985a,b) has shown that compensatory mutations can be fixed under complete linkage much more rapidly than under free recombination between the two sites of mutations. The time for spreading becomes much shorter under gene duplication even compared with the case of complete linkage, if $2Nv_m$ is much less than unity.

I thank B. S. Weir, J. B. Walsh and two anonymous referees for their many valuable suggestions and comments on the manuscript. This work is supported by a grant-in-aid from the Ministry of Education, Science and Culture of Japan. Contribution no. 1807 from the National Institute of Genetics, Mishima, 411 Japan.

REFERENCES


Communicating editor: B. S. WEIR