Simulating Evolution by Gene Duplication

Tomoko Ohta

National Institute of Genetics, Mishima, 411 Japan

Manuscript received May 23, 1986
Revised copy accepted October 11, 1986

ABSTRACT

By considering the recent finding that unequal crossing over and other molecular interactions are contributing to the evolution of multigene families, a model of the origin of repetitive genes was studied by Monte Carlo simulations. Starting from a single gene copy, how genetic systems evolve was examined under unequal crossing over, random drift and natural selection. Both beneficial and deteriorating mutations were incorporated, and the latter were assumed to occur ten times more frequently than the former. Positive natural selection favors those chromosomes with more beneficial mutations in redundant copies than others in the population, but accumulation of deteriorating mutations (pseudogenes) have no effect on fitness so long as there remains a functional gene. The results imply the following: (1) Positive natural selection is needed in order to acquire gene families with new functions. Without it, too many pseudogenes accumulate before attaining a functional gene family. (2) There is a large fluctuation in the outcome even if parameters are the same. (3) When unequal crossing over occurs more frequently, the system evolves more rapidly. It was also shown, under realistic values of parameters, that the genetic load for acquiring a new gene is not as large as J. B. S. HALDANE suggested, but not so small as in a model in which a system for selection started from already redundant genes.

GENE duplication has evidently played a major role in the evolution of complexity of higher organisms. A long time ago, BRIDGES (1935) and MULLER (1936) recognized the importance of gene duplication for evolution, but rather little attention has been paid to their theory until recently. OHNO (1970) advocated the significance of gene duplication from the standpoint of cytogenetics and biochemistry, and for more than 10 years, I have emphasized its role in the evolution of higher organisms in relation to the prevalence of multigene families in eukaryote genomes (for review, see OHTA 1980). In the model of gene duplication, it has been customarily assumed that, once redundant gene copies are available in a genome, useful mutations may accumulate in one of the copies while another copy is carrying out the original function (for review, see CHARLESWORTH 1985). In another model that emphasizes natural selection, "permanent heterozygosity" is attained by duplication when heterozygotes are advantageous (FINCHAM 1966), and SPOFFORD (1969) formulated this model as a deterministic process.

Recent developments in molecular biology, however, have revealed that evolution by gene duplication is not such a simple process as the above models imply. The prevalence of multigene families and their concerted evolution suggests that unequal crossing over, gene conversion and duplicative transposition are frequently occurring in genomes of higher organisms and are therefore providing ample opportunities for the origination of new genes (for reviews, see OHTA 1980, 1983). I have studied a model in which the more a multigene family contains genetic diversity, the more beneficial to an organism it is. I assumed an initial gene family with a few identical gene copies and then showed that the system is an extremely efficient way to accumulate genetic diversity, or genetic information in another expression (OHTA 1987). This study is useful for understanding the origin of large multigene families with functionally diverse gene members, such as those of immunoglobulin, T-cell receptor and cytochrome P450.

In this report, a more general model of evolution by gene duplication is analyzed, *i.e.*, simulations were carried out to find out how duplicated genes evolve starting from a single copy, if unequal crossing over is continuously occurring and if natural selection works in such a way that individuals with more diverse gene members than others in the population are selectively advantageous. One problem of such models is how the accumulation of deleterious mutations in repetitive gene copies is prevented. Indeed, the rate of deleterious mutation is usually much higher than that of useful mutations at the molecular level (*e.g.*, see KIMURA 1983), and pseudogenes are commonly found. Thus, in the present model, together with unequal crossing over and advantageous mutation, deleterious mutation is incorporated. It will be shown in the following sections that gene organization on the chromosome reflects more faithfully the effects of

positive and negative natural selection than does the primary structure of DNA or proteins. It is also shown that the system nevertheless provides an efficient way to increase the number of alleles in a genome, and that stochastic differentiation among the genomes is important.

MODEL AND SIMULATION PROCEDURE

In the present model, initially there is a single gene copy in each genome of the population, and all genes in the population are identical. With a constant rate, γ, per gene copy per generation, the chromosomal region of this gene undergoes unequal intrachromosomal (between sister-chromatids) crossing over, *i.e.*, at the rate, $\gamma/2$, a gene is duplicated to become two tandem genes, and at the same rate, $\gamma/2$, it is deleted from the chromosome. This rate is assumed to be constant per gene copy, and if there are n tandem genes, the rate of unequal crossing over to produce either $(n + 1)$ or $(n - 1)$ genes is $n\gamma$. Unequal crossing over is assumed to be always one gene shift, with no bias in duplication or deletion.

With constant rate, v_+, per generation, a gene mutates to a new allelic type under the infinite allele model (KIMURA and CROW 1964). With a different constant rate, v_-, per generation, a gene becomes nonfunctional, and once it deteriorates, it remains as a pseudogene.

Both positive and negative natural selection are incorporated. A specific model of positive selection is used that is similar to that of OHTA (1987), in which, if genetic diversity (the number of different alleles in a genome in this study) is lower than the population average, the gamete is disadvantageous according to the fitness function,

$$\begin{aligned} W_i &= 1 \quad \text{for} \quad k_i \geq \bar{k} \\ W_i &= 1 - s(\bar{k} - k_i) \quad \text{for} \quad k_i < \bar{k}, \end{aligned} \tag{1}$$

where the subscript, i, denotes the ith genome, k_i is the number of alleles in the ith genome, $\bar{k}$ is the population average, and s is a positive selection coefficient. This fitness function may seem rather arbitrary, but it is based on the consideration that selection coefficient should depend on genes at the loci concerned in the population, since we are studying the process of how novel genes not previously existing are acquired. As the simplest function, I have chosen (1), which is easy to handle in the simulations. Negative selection means simply that a gamete is lethal if all gene copies become nonfunctional, or if the copy number becomes zero by unequal crossing over.

The simulated population is made of $2N$ gametes, and each generation undergoes unequal crossing over, mutation, sampling and selection in this order. Actually, sampling and selection were done simultaneously. Unequal crossing over was carried out, according to the specified probability for each genome, by using random numbers. Decisions about deletion or duplication, as well as chromosomal position, were determined by random numbers. Mutation was again performed by random numbers, and allelic states were stored as integers.

In each generation, the number of different alleles in each genome was counted in order to perform selection. The selective elimination of a sampled gamete was determined by the fitness function (1), *i.e.*, a gamete is sampled randomly and it is eliminated with probability, 1—fitness. This is repeated until the number of gametes retained becomes $2N$. Thus, the simulated population is haploid, but so long as the fitness is multiplicative, the result may be extended to the diploid situation (CROW and KIMURA 1970). No interchromosomal recombination is incorporated in this study.

After every N generation, several quantities, such as the mean and variance of copy number, the actual number of different alleles, and the number of pseudogenes per genome, were counted and recorded. The number of deaths was counted every generation, and the total was recorded so that the genetic load could be examined. Each Monte Carlo experiment was continued for $50N$ generations, and 15 replicate runs were performed for each set of parameter values, except for one case of large population size. In the following section, some results will be presented which were obtained under realistic combinations of parameters, *i.e.*, the products of population size and rates of unequal crossing over and of mutation, $N\gamma$, Nv_+ and Nv_-, were chosen to be realistic.

RESULTS

Starting from a single gene, the experimental results show how copy number, pseudogene number and number of different alleles change with time under various intensities of selection. The mutation rate per locus is usually very low, *i.e.*, of the order of 10^{-6} per locus per generation. The population size of higher organisms may be often $10^4 \sim 10^5$ (*e.g.*, see KIMURA 1983). Thus, I chose the product of population size and deleterious mutation rate, $2Nv_-$ to be 0.1. The rate of useful mutation is likely to be much less than that of deleterious mutation, and a value of one-tenth of deleterious mutation rate was used, *i.e.*, $2Nv_+ = 0.01$. This value may be too high for a protein coding region, but may not be so for a noncoding region, since a large proportion of DNA sequences in the noncoding region appears to be irrelevant to transcription or translation. The rate of unequal crossing over may differ from locus to locus. Based on a recent estimate on immunoglobulin V genes (OHTA 1984), three values of unequal crossing-over rate were chosen; $2N\gamma = 0.1$, 0.25 and 0.5.

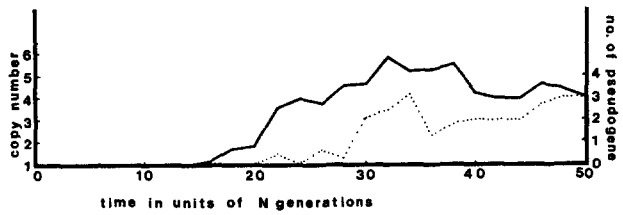

FIGURE 1.—A sample path of the simulation with no positive selection ($s = 0$). The copy number and the number of pseudogene increased, but the actual number of different alleles remained unity. The other parameters are, $2N = 100$ and $\gamma = 0.0025$.

Since very small populations of sizes, $2N = 50 \sim 200$, were simulated, but with realistic values of products, $N\gamma$ etc., as above, large selection coefficients were used: $s = 0 \sim 0.8$. Again, the product, $2Ns$, is important, and it was given the values 0, 20 and 40. When s is large, the effect of selection depends not only on $2Ns$ but also on s itself, and the following application to real populations is an approximation. At any rate, in extrapolating the results to a real population of $2N = 10^4 \sim 10^5$, the selection coefficient is very small. Figures 1–3 show some typical sample paths under various sets of parameters; a solid line shows the change of copy number; dotted line, the number of pseudogenes; broken line, the number of different alleles. Figure 1 is for the case of no positive selection ($s = 0$), with $2N\gamma = 0.25$. There is no increase of functional genes, whereas pseudogenes gradually accumulate. At the $50N$th generation, there are three copies of pseudogenes, and only one functional gene remains.

Figure 2 shows a sample path for $s = 0.4$ ($2Ns = 40$), with $2N\gamma = 0.25$. In this path, there is no increase of copy number until the $22N$th generation, and after that it increases. Between the $28N$th and the $30N$th generation, a different allele accumulates, and at the $50N$th generation, three different alleles and almost two copies of a pseudogene accumulate. In Figure 3, a sample path for the same selection intensity, but with a higher rate of unequal crossing over ($2N\gamma = 0.5$), is shown. As can be seen, the accumulation of both useful and deleterious mutations is more rapid here than in the previous case of low γ.

Because the accumulation process is stochastic, a large variance is observed even if the parameters are

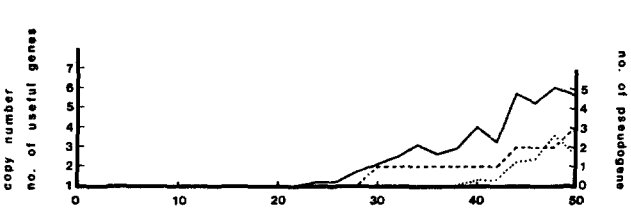

FIGURE 2.—A sample path of the simulation with positive selection ($2Ns = 40$). The copy number, the pseudogene number and the actual number of different alleles increased in this sample path. Parameters are, $2N = 100$ and $\gamma = 0.0025$.

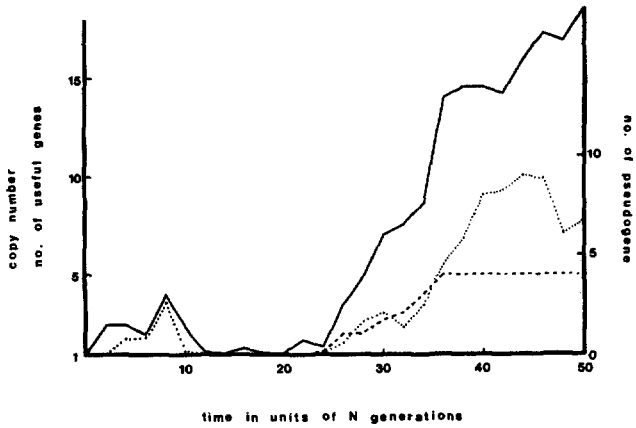

FIGURE 3.—A sample path of the simulation with $2Ns = 40$. The copy number, the pseudogene number and the actual number of different alleles increased more rapidly than in the path of Figure 2, since unequal crossing-over rate is higher here ($2N = 100$ and $\gamma = 0.005$).

the same. Except for the case of $2N = 200$, for each set of parameters, 15 replicate runs were performed, and Table 1 represents the mean and the standard deviation of 15 replications for copy number, number of different alleles, and pseudogene number at the $50N$th generation. In many cases, the standard deviation is almost as large as the mean. This means that 15 runs may differ greatly from each other, ranging from no accumulation to rapid increase of different alleles.

However, there is a clear indication that positive selection is effective for increasing useful genes. Indeed, under the present sets of parameters, positive selection is necessary for the accumulation of functional genes. When $s = 0$, the mean number of different alleles is, at most, 1.071 from Table 1. It is noted that, even if several redundant copies accumulate, all extra copies deteriorate to become pseudogenes. This is because $v_- = 10v_+$ in the present simulations, and the chance for deterioration is ten times larger than that for functional differentiation, and once a gene deteriorates, it remains as a pseudogene.

Let us examine the relative rates of useful *vs.* nonfunctional differentiation when positive selection is involved. Consider, for simplicity, the case where a single gene is present with the parameters, $2N\gamma$, $2Nv_+$ and $2Nv_-$, much less than unity. Such a case is close to the classical model of gene duplication in which a duplication becomes fixed in the population, presumably by drift, and then useful mutations accumulate by natural selection. In our case, a duplication also spreads in the population by drift, and chromosomes with two tandem genes increase. One of the two genes deteriorates at the rate, v_-, again by genetic drift, whereas the rate at which a gene acquires a useful mutation is enhanced by selection. Let x be the frequency of the chromosomes in a population with newly occurring useful mutation, *i.e.*, with two useful

TABLE 1

Results of Monte Carlo experiments for the copy number, the number of different alleles, the pseudogene number and R (observed and expected) at the $50N$th generation, and the number of genetic death per one increase of different allele

Parameters	Selection coefficient	Copy no.	No. of different alleles	No. of pseudogene	R		Average genetic death
					Observed	Expected	
$2N = 100$	0	2.111 ± 1.339	1.017 ± 0.067	0.846 ± 1.161	0.020	0.100	
$2N\gamma = 0.1$	0.2	1.977 ± 0.939	1.282 ± 0.582	0.513 ± 0.806	0.550	0.504	$9.69N$
	0.4	3.019 ± 1.860	1.813 ± 0.983	0.777 ± 0.754	1.046	0.713	
$2N = 100$	0	3.711 ± 3.134	1.005 ± 0.021	1.969 ± 2.877	0.003	0.100	
$2N\gamma = 0.25$	0.2	4.286 ± 2.559	1.667 ± 0.617	1.306 ± 1.234	0.511	0.504	$10.55N$
	0.4	6.176 ± 5.136	2.566 ± 2.051	2.533 ± 3.384	0.618	0.713	
$2N = 100$	0	3.673 ± 4.541	1.003 ± 0.010	1.570 ± 4.332	0.002	0.100	
$2N\gamma = 0.5$	0.2	8.515 ± 5.096	2.058 ± 1.210	4.212 ± 4.260	0.251	0.504	$9.55N$
	0.4	16.577 ± 12.655	5.597 ± 0.399	6.201 ± 6.323	0.741	0.713	
$2N = 50$	0	4.471 ± 4.854	1.071 ± 0.248	2.727 ± 3.187	0.026	0.100	
$2N\gamma = 0.25$	0.4	3.749 ± 2.813	1.657 ± 1.375	1.519 ± 2.335	0.433	0.504	$10.33N$
	0.8	6.687 ± 3.606	2.528 ± 1.302	2.948 ± 2.026	0.518	0.713	
$2N = 200$	0	2.771 ± 2.977	1.055 ± 0.123	1.368 ± 2.870	0.040	0.100	
$2N\gamma = 0.25$	0.1	9.320 ± 10.739	2.745 ± 2.944	4.860 ± 5.838	0.359	0.504	$9.63N$
	0.2	7.643 ± 9.267	2.808 ± 2.189	3.700 ± 5.598	0.489	0.713	

Results of 15 replications, except for the case of $2N = 200$ with five replications. Other parameters: $2Nv_+ = 0.01$, $2Nv_- = 0.1$.

genes (different alleles) as compared to the other chromosomes that have only one useful gene. Then the fitness of the chromosomes with two useful genes is unity, whereas that of those with one functional gene is, from (1),

$$w_a = 1 - sx.$$

Therefore, the change of x becomes the same as that of a recessively advantageous mutant,

$$\Delta x = \frac{sx^2(1-x)}{1 - sx(1-x)}. \tag{2}$$

The fixation probability of a recessive mutant in finite populations is approximately (KIMURA 1964),

$$u_+ \approx 1.128 \sqrt{\frac{s}{2N}} = \sqrt{2Ns}\,\frac{1.128}{2N}. \tag{3}$$

Thus, the chance of spreading useful mutants on duplicated genes is enhanced roughly by the amount, $1.128\sqrt{2Ns}$. The ratio, R, of the rate of spreading of useful mutations to that of deteriorating mutations becomes, for $s > 0$,

$$R = \frac{u_+v_+}{v_-} = \frac{1.128\sqrt{2Ns}\,v_+}{v_-}. \tag{4}$$

Comparison of observed and expected values of R is also given in Table 1. Although some of the assumptions do not hold in our simulations, data in the table show that the above prediction by (4) is roughly applicable.

When $s = 0$, mutants are no longer useful, and the expected ratio is simply $v_+/v_- = 0.1$. Because of unidirectional deterioration of redundant genes in our simulations, the observed values of R are less than 0.1 for $s = 0$. For $s > 0$, however, useful genes once fixed in the population are seldom lost because of selection. Therefore, the average ratio would take similar values for a long period of time.

The validity of (4) depends on whether or not (2) and (3) hold. When the products $N\gamma$, Nv_- and Nv_+ are much less than unity, the spreading of duplication, of deteriorating mutation and of beneficial mutation occur separately and singly, and (2) and (3) adequately describe the process. However, when these products take larger values, the spreading of duplications and mutations occurs simultaneously, and (2) and (3) would become inappropriate. Under the present sets of parameters, in which the products are slightly less than unity, our approximations seem applicable.

More generally, the ratio is expressed as follows.

$$R' = \frac{u_+v_+}{u_-v_-}, \tag{5}$$

where u_- is the fixation probability of a deteriorating mutant, and can be much less than $1/2N$ if a nonfunctional gene is disadvantageous to the organism. In the next section, I shall discuss the meaning of this ratio in relation to the observed facts.

By increasing the rate of unequal crossing over, the gene system changes more quickly, but the pattern of increase of copy number is most difficult to formulate. However, there is a tendency that, by increasing selection intensity, the total copy number increases more rapidly. This tendency was clearer in the previous model in which positive selection is similar to

that here, but the genetic system starts from a gene family with several identical copies (OHTA 1987). It was also possible, in the previous model, to estimate selection response under realistic values of parameters.

In the last column of the table, the average genetic death (load) needed for the increase of one useful gene (different allele) is given. Note that it does not include death due to negative selection, *i.e.*, lethals caused by the total loss of functional gene. Thus, the above load corresponds to the cost of natural selection (HALDANE 1957) or the substitutional load (KIMURA 1961). Genetic load per gene increase seems to be fairly uniform as a multiple of population size. We examine the implication of this and other findings for the evolution of complexity of higher organisms in the next section.

DISCUSSION

It has been shown in the previous section that the ratio of the rate of spreading of useful genes to that of deteriorating mutations, R, is crucial for understanding the functional organization of eukaryote genomes. Nonfunctional genes in our model correspond to DNA-mediated pseudogenes in many clusters of gene families, and are different from dispersed pseudogenes that arise through RNA intermediate. Thus, pseudogenes in our model should be called "junk" (OHNO 1970), rather than "selfish," since they accumulate by drift (CAVALIER-SMITH 1985); and the ratio, R, tells how much junk DNA accumulates in order to acquire useful genes.

It should be noted that the quantity R is the ratio of the initial rates of accumulation of pseudogenes and beneficial ones. As it has been pointed out in the previous section, however, R would remain unchanged on the average because useful genes once spread will be maintained by selection in the population. In other words, their elimination either by deteriorating mutation or unequal crossing over is prevented by selection for $s > 0$, provided that Ns is sufficiently large. On the other hand, pseudogenes once fixed neither increase nor decrease on the average, because they are assumed to be neutral, and their average number will unchange. Thus, the average value of R will remain stable for a long time.

In the simulations, I assumed that deteriorating mutation occurs ten times more frequently than useful mutation, and that its accumulation causes no deleterious effect to the organisms so long as one functional gene remains. We do not know exactly the real situation, but R' (equation 5) cannot be very small; otherwise, only junk would accumulate. Actual examples of multigene families, such as genes for hemoglobin α and β (JEFFREYS 1982) or human leukocyte interferon (ALLEN and FANTES 1980), indicate that one in several gene copies has deteriorated. Then it is likely that R' is considerably larger than unity or $u_+v_+ > u_-v_-$. It is possible that pseudogenes themselves are deleterious to the organism, particularly when they produce nonfunctional proteins. Then individuals with pseudogenes would be eliminated, and u_- may be much less than $1/2N$. It should also be noted that the present model is not appropriate for the cases in which the copy number is selectively regulated with respect to the amount of gene products. Several large multigene families with uniform gene members, such as histone or rRNA genes, apparently need to produce large amounts of protein or RNA, and copy number has increased because of such a requirement. The rate of unequal crossing over is usually higher in these gene families (FEDOROFF 1979) than that considered in this study.

One important point of the present study is that gene organization of relatively young gene families of proteins reflects more faithfully the action of natural selection than the primary structure of DNA or proteins. It is almost impossible to find out, or to estimate, the advantageous amino acid and nucleotide substitutions that constitute only a minor fraction of the total change by comparing primary structures (KIMURA 1983). On the other hand, the proportion of DNA-mediated pseudogenes would reflect the ratio, R', and tell us how positive selection has acted, even if not exactly. It may also be meaningful to relate the present theory with the observed acceleration of amino acid substitutions of some duplicated genes (GOODMAN 1976; LI 1985). Although most nucleotide substitutions in the long course of evolution since duplication are likely to be selectively neutral (KIMURA 1983), during the short period just after duplication, positive selection as considered here may be responsible for accelerating the evolution of duplicated genes. Particularly noteworthy is the observation that amino acid substitutions at the surface of proteins that would result in minor modification of protein function are accelerated after duplication (LI 1985). It should be noted that my interpretation is different from that of GOODMAN (1976), who claims that most amino acid substitutions occurred by positive natural selection.

As for the positive selection scheme, the present model is only one possibility. If selection is dominant or semidominant, it would be more effective, and u_+ would be larger. The significance of the present scheme is that selection coefficients are not based on absolute genotypes but on relative frequencies of genotypes in the population.

Next, let us examine the genetic load for acquiring useful genes. As I pointed out in the previous report (OHTA 1987), the important difference between my approach and the classical model of gene substitution by selection (HALDANE 1957) is that the present model

is stochastic, whereas the classical one is deterministic. Therefore, the parameters, Nv_+, Nv_-, $N\gamma$ and Ns determine the amount of genetic load for increase of one useful gene here, whereas the load depends on the initial frequency of advantageous mutants in the deterministic model. The present model is not very helpful for reducing genetic load as compared with the previous one (OHTA 1987). From Table 1, the load for the increase of one useful gene is around $10N$ under the present sets of parameter values. This is one-third of the famous estimate of HALDANE (1957). In the previous model, the load was only $2N \sim 3N$, for increase of genetic diversity that is almost equivalent to one useful gene, because the genetic system for selection contained five gene copies from the beginning and no deteriorating mutation was considered. The previous model may be helpful for understanding the origin of some large multigene families, such as genes of immunoglobulin or cytochrome P450, whereas the present model helps clarify the mechanism of evolution by gene duplication in general.

It is also interesting to compare the present model with that of LOOMIS and GILPIN (1986). These authors simulated evolution of a single chromosome by assuming that crossing over is nonhomologous. In my model, crossing over is assumed to be unequal but homologous, except for the first production of two tandem genes from a single copy, and genes always duplicate as units. But in LOOMIS and GILPIN'S model, nonhomologous crossing over occurs, and junk DNA is created by the production of incomplete genes. In the present model, junk DNA accumulates by deteriorating mutations in the redundant copies.

Another topic in evolution by gene duplication is concerned with other mechanisms recently shown to be important but not considered here, *i.e.*, gene conversion and transposition. For example, by gene conversion, various "mosaic" genes could be produced which may be useful for the organisms (GILBERT 1978, 1985; BALTIMORE 1981; MIYATA *et al.* 1980, SLIGHTOM, BLECHL and SMITHIES 1980). Also, such "reshuffling" of exons between duplicated gene members may provide an opportunity for a pseudogene to become functional again. The effects of such mechanisms on the origin and evolution of genetic systems are left for future study.

Finally, I should like to emphasize that, in this model and in the previous one (OHTA 1987), chance effects are important, in that there are many different outcomes even if similar selection is responsible for increasing useful genes. This would accord with the quite variable organization of gene families among different species, such as found in genes of hemoglobins (JEFFREYS 1982), or major uninary proteins of Myomorpha (HASTIE, HELD and TOOLE 1979; GHAZAL, CLARK and BISHOP 1985), or immunoglobulin V_H genes (HINDS and LITMAN 1986).

I thank B. S. WEIR, H. TACHIDA and an anonymous referee for their many valuable suggestions and comments for improving the presentation. This work is supported by a grant-in-aid from the Ministry of Education, Science and Culture of Japan. Contribution no. 1699 from the National Institute of Genetics, Mishima, 411 Japan.

LITERATURE CITED

ALLEN, G. and K. H. FANTES, 1980 A family of structural genes for human lymphoblastoid (leukocyte-type) interferon. Nature **287:** 408–411.

BALTIMORE, D., 1981 Gene conversion: some implications for immunoglobulin genes. Cell **24:** 592–594.

BRIDGES, C. B., 1935 Salivary chromosomes maps. J. Hered. **26:** 60–64.

CAVALIER-SMITH, T. (Editor), 1985 *The Evolution of Genome Size.* John Wiley & Sons, New York.

CHARLESWORTH, B., 1985 Recombination, genome size and chromosome number. In: *The Evolution of Genome Size*, Edited by T. CAVALIER-SMITH. pp. 489–513. John Wiley & Sons, New York.

CROW, J. F. and M. KIMURA, 1970 *An Introduction to Population Genetics Theory.* Harper & Row, New York.

FEDOROFF, N. V., 1979 On Spacers. Cell **16:** 697–710.

FINCHAM, J. R. S., 1966 *Genetic Complementation.* Benjamin, Elmsford, New York.

GHAZAL, P., A. J. CLARK and J. D. BISHOP, 1985 Evolutionary amplification of a pseudogene. Proc. Natl. Acad. Sci. USA **82:** 4182–4185.

GILBERT, W., 1978 Why genes in pieces? Nature **271:** 501.

GILBERT, W., 1985 Genes-in-pieces revisited. Science **228:** 823–824.

GOODMAN, M., 1976 Protein sequences in phylogeny. In: *Molecular Evolution*, Edited by F. J. AYALA, pp. 141–159. Sinauer Associates, Sunderland, Massachusetts.

HALDANE, J. B. S., 1957 The cost of natural selection. J. Genet. **55:** 511–524.

HASTIE, N. D., W. A. HELD and J. J. TOOLE, 1979 Multiple genes coding for the androgen-regulated major urinary proteins of mouse. Cell **17:** 449–457.

HINDS, K. R. and G. W. LITMAN, 1986 Major reorganization of immunoglobulin V_H segmental elements during vertebrate evolution. Nature **320:** 546–549.

JEFFREYS, A., 1982 Evolution of globin genes. pp. 157–176. In: *Genome Evolution*, Edited by G. A. DOVER and R. B. FLAVELL. Academic Press, New York.

KIMURA, M., 1961 Natural selection as the process of accumulating genetic information in adaptive evolution. Genet. Res. **2:** 127–140.

KIMURA, M., 1964 Diffusion models in population genetics. J. Appl. Probab. **1:** 177–232.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, London.

KIMURA, M. and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics **49:** 725–738.

LI, W.-H., 1985 Accelerated evolution following gene duplication and its implication for the neutralist-selectionist controversy. pp. 333–352. In: *Population Genetics and Molecular Evolution*, Edited by T. OHTA and K. AOKI. Japan Scientific Society Press, Tokyo/Springer-Verlag, Berlin.

LOOMIS, W. F. and M. E. GILPIN, 1986 Multigene families and vestigial sequences. Proc. Natl. Acad. Sci. USA **83:** 2143–2147.

MIYATA, T., T. YASUNAGA, Y. YAMAWAKI-KATAOKA, M. OBATA and T. HONJO, 1980 Nucleotide sequence divergence of

mouse immunoglobulin γ_1 and γ_{2b} chain genes and the hypothesis of intervening sequence-mediated domain transfer. Proc. Natl. Acad. Sci. USA **77:** 2143–2147.

MULLER, H. J., 1936 Bar duplication. Science **83:** 528–530.

OHNO, S., 1970 *Evolution by Gene Duplication.* Springer-Verlag, Berlin.

OHTA, T., 1980 *Evolution and Variation of Multigene Families,* Lecture Notes in Biomathematics, Vol. 37, Springer-Verlag, New York.

OHTA, T., 1983 On the evolution of multigene families. Theor. Pop. Biol. **23:** 216–240.

OHTA, T., 1984 Population genetics theory of concerted evolution and its application to the immunoglobulin V gene tree. J. Mol. Evol. **20:** 274–280.

OHTA, T., 1987 A model of evolution for accumulating genetic information. J. Theor. Biol. In press.

SLIGHTOM, J. L., A. E. BLECHL and O. SMITHIES, 1980 Human fetal $^{G}\gamma$- and $^{A}\gamma$-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell **21:** 627–638.

SPOFFORD, J. B., 1969 Heterosis and the evolution of duplication. Am. Nat. **103:** 407–432.

Communicating editor: B. S. WEIR