Genet. Res., Camb. (1981), **37**, *pp.* 133–149
With 5 text-figures
Printed in Great Britain

Genetic variation in small multigene families*

By TOMOKO OHTA
National Institute of Genetics, Mishima 411, Japan

(*Received* 24 *June* 1980, *and in revised form* 14 *August* 1980)

SUMMARY

In order to understand the evolution of genetic systems in which two genes are tandemly repeated (small multigene family) such as has been recently found in the haemoglobin α loci of primates, haemoglobin β loci of mouse and rabbit and other proteins, a population genetics approach was used. Special reference was made to the probability of gene identity (identity coefficient), when unequal crossing-over is continuously occurring as well as random genetic drift, inter-chromosomal recombination and mutation. Two models were studied, cycle and selection models. The former assumes that unequal crossing-over occurs in cycles of duplication and deletion, and that the equilibrium identity coefficients were obtained. The latter is based on more realistic biological phenomena, and in this model it is assumed that natural selection is responsible for eliminating chromosomes with extra or deficient gene dose. Unequal crossing-over, inter-chromosomal recombination and natural selection lead to a duplication–deletion balance, which can then be treated as though it were a cycle model. The basic parameter is the rate of duplication–deletion which is shown to be approximately equal to $2(u+2\beta)X$, where u is the unequal crossing-over rate, 2β is the inter-chromosomal recombination rate and X is the frequency of chromosomes with three genes or of that with one gene. Genetic variation of the globin gene family, of which gene organization is known in most detail, is discussed in the light of the present analyses.

1. INTRODUCTION

Remarkable progress in molecular biology has recently revealed several unexpected features in gene organization of eukaryotes. One is the fact that genes are often duplicated in the genome. It has been suggested that the two copies of haemoglobin α gene are usually present in a human genome (e.g. Dayhoff, 1972, page 77). Such a feature is now confirmed by molecular cloning technique and is extended to the genes of many other proteins; haemoglobin β of mouse (Konkel, Tilghman & Leder, 1978), haemoglobin β of rabbit (Hardison *et al.* 1979), ovalbumin of chicken (Royal *et al.* 1979), corticotropin-β-lipotropin of bovine (Nakanishi *et al.* 1979), major urinary proteins of mouse (Hastie *et al.* 1979), δ-crystallin of chicken lens (Bhat *et al.* 1980), etc.

Goossens *et al.* (1980) have performed restriction enzyme mapping studies on

* Contribution No. 1334 from the National Institute of Genetics, Mishima 411, Japan.

the human genome and found that, while the majority of chromosomes contain two tandemly repeated haemoglobin α genes, the frequency of chromosomes with three globin α genes is 0·0036 in American blacks, less than 0·004 in Sardinians and 0·05 in a Greek population. They suggest that such triple α-globin gene loci should have occurred through unequal crossing-over. Zimmer *et al.* (1980) extended the analysis to the genomes of several species of primates, and found that such a feature is characteristic to all species examined; chimpanzee, pygmy chimpanzee, gorilla, orang-utan and gibbon. These authors emphasize that, even if genes are differentiated between the species, repeated genes remain almost identical within a species, for which they suggest the phrase 'concerted evolution'. The situation is quite similar to the 'coincidental evolution' of multigene families of large size such as the histone, ribosomal RNA or immunoglobulin genes reviewed by Hood, Campbell & Elgin (1975). Unequal crossing-over is considered to be the most probable mechanism of coincidental evolution of large multigene families (Smith, 1974; Hood *et al.* 1975; Tartof, 1975), as well as of the concerted evolution of small multigene family (Zimmer *et al.* 1980; Lauer, Shen & Maniatis, 1980). The hypothesis has been verified in ribosomal RNA genes of yeast by introducing a marker in the region of this gene family (Szostak & Wu, 1980; Petes, 1980).

In order to understand the process of coincidental or concerted evolution, however, a population genetics approach is needed as in my previous study on large multigene families (Ohta, 1978, 1980; Kimura & Ohta, 1979). The purpose of the present report is to clarify the nature of gene diversity of multigene families with only a few gene copies per genome, when unequal crossing-over occurs continuously together with mutation, random genetic drift and inter-chromosomal recombination.

2. BASIC THEORY

Let us consider a finite population with effective size N_e. A small multigene family is present on a chromosome and evolving under mutation, random genetic drift, unequal intra-chromosomal (between sister-chromatids) crossing-over at somatic division of germ cell lines and inter-chromosomal recombination at meiosis. I shall investigate two models; the cycle model and the selection model. In the first one, the unequal crossing-over is assumed to occur in cycles as in my previous analyses on large multigene family (Ohta, 1976, 1978), i.e. the duplication phase of unequal crossing-over is followed by the deletion phase. In the second model, natural selection is assumed so that the gene number per family remains unchanged. The crossover products with extra genes or less genes are selected against and their frequencies in the population are in balance between crossing-over and selection. The second model is biologically real, whereas the first one is not, i.e. the second leads to a balance of duplication and deletion based on known biological phenomena, whereas the first assumes that this happens cyclically without asking how.

(i) *Cycle model*

We assume that a chromosome contains a multigene family with two gene copies, and cycles of two unequal intra-chromosomal crossing-overs of duplication and deletion as in Fig. 1 occur with the rate γ per chromosome per generation. Thus the gene number per chromosome does not change as in my previous model

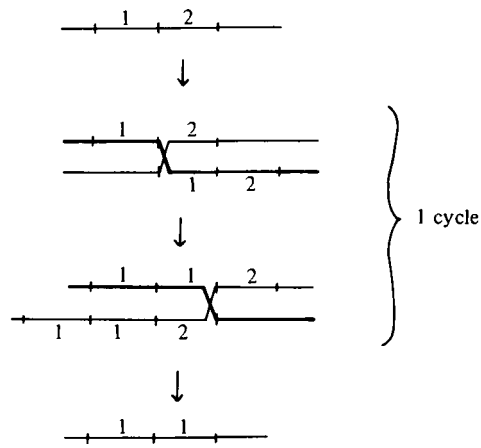

Fig. 1. Diagram showing one cycle of unequal crossing-over.

for large multigene families (Ohta, 1976). Here, for simplicity, it is assumed that the crossover takes place between the gene units and not inside the gene. Also note that the intra-chromosomal unequal crossing-over occurs at somatic division of germ cell lines. At meiosis, inter-chromosomal crossing-over takes place as in Fig. 2. Let β be the rate of such crossing-over. Here, we assume no inter-chromosomal unequal crossing-over. Let v be the mutation rate per gene per generation with all mutations assumed to be unique as in Kimura & Crow (1964).

Under this model, the gene family may differentiate by new mutations, however genes may be kept fairly homogeneous in one population through unequal crossing-over, i.e. coincidental evolution. Let us investigate a measure of gene homogeneity, the identity coefficient, which is the probability of identity of two genes of the family as in the previous studies on a large multigene family (Ohta, 1978, 1980; Kimura & Ohta, 1979). We shall define the following set of identity coefficients; c_1 is the probability of the two genes on the same chromosome being identical, f_0 is that of the two genes of the same (homologous) position of different chromosomes being identical, and f_1 is that of the two genes of the non-homologous position of different chromosomes being identical (see Fig. 3). Next we shall formulate the change of these identity coefficients by unequal crossing-over, inter-chromosomal crossing-over, random drift and mutation.

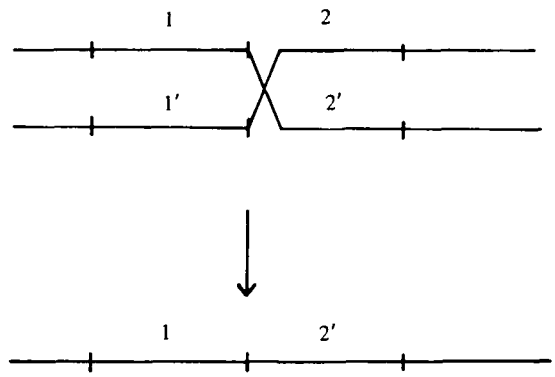

Fig. 2. Diagram showing inter-chromosomal recombination.

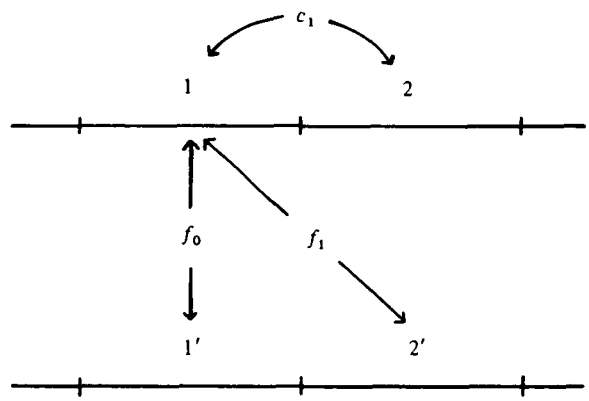

Fig. 3. Diagram illustrating the meaning of three identity coefficients of the cycle model.

By cycles of unequal crossing-over (Fig. 1) with the rate γ, c_1 changes to c_1' on the average according to the following equation,

$$c_1' = c_1 + \frac{\gamma}{2}(1 - c_1). \tag{1}$$

This is because one half of the crossover products contain duplicated gene members (Fig. 1). f_0 and f_1 change according to the following formulas, by considering that one-half of the gene comparisons for f_0 and f_1 (Fig. 3) involving the crossover chromosome are shifted.

and

$$\left.\begin{aligned} f_0' &= f_0 + \frac{\gamma}{4}(f_1 - f_0) \\ f_1' &= f_1 + \frac{\gamma}{4}(f_0 - f_1). \end{aligned}\right\} \tag{2}$$

Next, by inter-chromosomal crossing-over with the rate β (Fig. 2), c_1' changes to c_1'' as follows,

$$c_1'' = c_1' + \beta(f_1' - c_1'). \tag{3}$$

https://doi.org/10.1017/S0016672300020115 Published online by Cambridge University Press

As to f_1', only comparisons between the two crossover chromosomes would be affected, and the proportion of such comparisons is $\beta/(2N_e-1)$ (see Ohta, 1978; Kimura & Ohta, 1979), whereas f_0' does not change. Therefore,

$$\left.\begin{aligned} f_0'' &= f_0' \\ \text{and} \quad f_1'' &= f_1' + \frac{\beta}{2N_e-1}(c_1'-f_1'). \end{aligned}\right\} \tag{4}$$

We shall proceed to calculate the changes due to random drift and mutation. By sampling, c_1'' is not influenced, however f_0'' and f_1'' change as in the case of a large multigene family (Ohta, 1978). Also, all coefficients are reduced by mutation, and we have

$$\left.\begin{aligned} c_1''' &= (1-v)^2 c_1'' \\ f_0''' &= (1-v)^2\{f_0'' + \frac{1}{2N_e}(1-f_0'')\} \\ \text{and} \quad f_1''' &= (1-v)^2\{f_1'' + \frac{1}{2N_e}(c_1''-f_1'')\}. \end{aligned}\right\} \tag{5}$$

By combining the equations (1)–(5), one can generate the identity coefficients from one generation to the next.

Of particular interest are the identity coefficients when equilibrium is reached. The identity coefficients do not change with time and we have $c_1''' = c_1$, $f_0''' = f_0$ and $f_1''' = f_1$. When all parameters, γ, β, $1/(2N_e)$ and v are much less than unity, the equilibrium solution can be obtained, by letting hats ($\wedge$) denote the equilibrium values,

$$\left.\begin{aligned} \hat{f}_1 &= \frac{\gamma(6N_e v + N_e\gamma + N_e\beta + 1)}{\left(4N_e v + \frac{N_e\gamma}{2} + 1\right)^2(4v+\gamma+2\beta) - \left(\frac{N_e\gamma}{2}\right)^2(4v+\gamma+2\beta) - 2\beta\left(4N_e v + \frac{N_e\gamma}{2} + 1\right)} \\ \hat{f}_0 &= \frac{N_e\gamma\hat{f}_1 + 2}{(8N_e v + N_e\gamma + 2)} \\ \text{and} \quad \hat{c}_1 &= \frac{2\beta\hat{f}_1 + \gamma}{4v+\gamma+2\beta}. \end{aligned}\right\} \tag{6}$$

In the following, the selection model will be investigated, which may be closer to real examples like the haemoglobin α gene loci, and the results will be compared with the above equations (6).

(ii) *Selection model*

In the second model, the unequal crossing-over is not assumed to occur in cycles, but selection is responsible for keeping the average number of genes constant at two copies in a genome. Let us assume that, by unequal intra-chromosomal crossing-over, the chromosomes with one and three gene copies occur as in Fig. 4. These crossover products are selected against and their frequencies are kept low in the

population by the balance between selection and unequal crossing-over. If the rate of unequal crossing-over is u per chromosome per generation and s is the heterozygous disadvantage of chromosomes with one or three doses, the equilibrium frequencies of such chromosomes become

$$X_1 = X_3 = X = u/(2s), \tag{7}$$

where X_1 and X_3 are the equilibrium frequencies of chromosomes with one and three copies. At meiosis, these chromosomes pair mostly with the normal chromosomes with two copies, and recombination takes place between the two chromosomes. Let 2β be the rate of inter-chromosomal recombination at the region of this family in one generation, when two genes pair as in the upper diagram of Fig. 5. This rate is β when the pairing is by only one gene as in the lower diagram of Fig. 5. In the following, let us investigate the nature of gene diversity by again using the identity coefficient.

Table 1. *Classification of twelve identity coefficients investigated in the selection model*

	Identity coefficients				
	One chromosome		Two chromosomes		
No. of steps . . .	1	2	0	1	2
Group					
Three-gene	c_{31}	c_{32}	f_{30}	f_{31}	f_{32}
Two-gene	c_{21}	—	f_{20}	f_{21}	—
One-gene	—	—	f_{10}	—	—
Between-group			$\phi_{1,2}$	$\phi_{2,3}$	$\phi_{1,3}$

Let us define the following set of identity coefficients, by dividing the population into three groups, i.e. chromosomes with one, two and three genes. The identity coefficients of different groups are classified by the subscripts. The letter c denotes the identity coefficients with respect to one chromosome, and f denotes those between the chromosomes as before. The first subscript denotes the groups, and the second subscript denotes distance between the genes compared for identity (see Table 1). For example, c_{31} is the identity probability of genes one step apart on a chromosome carrying three genes and f_{31} is the probability for genes one step apart on pairs of chromosomes both carrying three genes. In addition to c and f, the identity coefficients between chromosomes of different groups are needed, and ϕ denotes these coefficients. The two subscripts with a comma denote the groups (see Table 1). For example, $\phi_{2,3}$ is the identity probability for a gene on a chromosome carrying two genes and one on a chromosome carrying three. In the following, the changes of these twelve identity coefficients by crossing-over and random genetic drift are obtained.

Let us investigate the changes of identity coefficients through intra-chromosomal unequal crossing-over which occurs with rate u per generation. As to the coefficient

c_{31}, which is the identity probability of genes with one step distance and on the same chromosome with three genes, the amount $(1-2X)u/2$ comes through unequal crossing-over from the two-gene group in one generation, as is seen by noting that one half of the crossover products will have three genes (see Fig. 4). Since the frequency of the three-gene group is X, the fraction of the crossover products in this group is $(1-2X)u/\{2X+(1-2X)u\}$. Here we assume that u is much

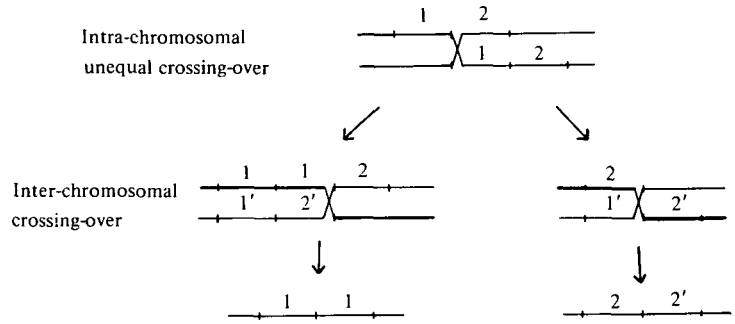

Fig. 4. Diagram showing the unequal crossing-over followed by inter-chromosomal recombination of the selection model.

less than unity. Among the chromosomes of this fraction, one half of the identity probability between two genes of one-step is unity by gene duplication, and the remaining half is c_{21}, which is the identity coefficient of different genes in chromosomes carrying two genes. Therefore, c_{31} changes to c_{31}' by unequal crossing-over according to the following equation,

$$c_{31}' = (1-g)c_{31} + \frac{g}{2}(1+c_{21}), \tag{8}$$

where $g = (1-2X)u/\{2X+(1-2X)u\}$. Similarly, c_{32} is transformed by the following formula,

$$c_{32}' = (1-g)c_{32} + gc_{21}. \tag{9}$$

The changes of identity coefficients of genes on different chromosomes of the three-gene group take a similar form, as is seen by referring to Fig. 4, and noting the heterozygous frequency of the crossover products.

$$f_{3i}' = \{1-2g(1-g)\}f_{3i} + 2g(1-g)\phi_{2,3}, \quad i = 0, 1, 2. \tag{10}$$

Note that $\phi_{2,3}$ is the identity coefficient of genes of the two-gene group and the three-gene group.

The identity coefficients of the one-gene group are similarly transformed by unequal crossing-over:

$$f_{10}' = \{1-2g(1-g)\}f_{10} + 2g(1-g)\phi_{1,2}. \tag{11}$$

https://doi.org/10.1017/S0016672300020115 Published online by Cambridge University Press

Next, the changes of the identity coefficients between the different groups may be shown to be

$$\left.\begin{aligned} \phi_{1,2}' &= (1-g)\phi_{1,2}+\frac{g}{2}(f_{20}+f_{21}), \\ \phi_{2,3}' &= (1-g)\phi_{2,3}+\frac{g}{2}(f_{20}+f_{21}), \\ \text{and} \\ \phi_{1,3}' &= \{1-2g(1-g)\}\phi_{1,3}+g(1-g)(\phi_{1,2}+\phi_{2,3}). \end{aligned}\right\} \tag{12}$$

The identity coefficients of the two-gene group are influenced through intra-chromosomal unequal crossing-over via the three-gene group, i.e. chromosomes with two genes are produced by unequal crossing-over of chromosomes carrying three genes. Let us assume that the crossing-over rate of the chromosomes with three genes is twice as large as that of the chromosomes with two genes, because the paired region at the unequal crossing-over of the former would be twice as large as that of the latter. Here we also assume that the crossover chromosomes with four genes are immediately eliminated from the population by selection and do not make any contribution to the gene pool of the population. Further, let the unequal crossing-over rate be negligibly small for the chromosomes carrying one gene. Under such assumptions, the fraction of chromosomes coming into the two-gene group from the three-gene group is

$$h = (2uX \times \tfrac{1}{2})/\{1-2X+2uX \times \tfrac{1}{2}\} = uX/(1-2X+uX).$$

Therefore we have,

$$\left.\begin{aligned} c_{21}' &= (1-h)c_{21}+hc_{31} \\ \text{and} \\ f_{2i}' &= \{1-2h(1-h)\}f_{2i}+2h(1-h)\phi_{2,3}, \quad i = 1, 2. \end{aligned}\right\} \tag{13}$$

Next, let us evaluate the changes of identity coefficients due to inter-chromosomal crossing-over (see Fig. 5). The rate of crossing-over is assumed to be 2β per family with two genes (upper diagram of Fig. 5) and β per family when the recombination is between one-gene and two-gene chromosomes (lower diagram of the figure). Then the frequencies of the crossover products are $4X(1-2X)\beta$ from the pairs between the two-gene and the three-gene chromosomes and $2X(1-2X)\beta$ from the one-gene and the two-gene chromosomes. Let us assume that the frequency of the crossover products between the one-gene and the three-gene chromosomes is negligibly small which is $2X^2\beta$. The fraction of recombinant chromosomes in the three-gene group is $2X(1-2X)\beta/\{X+2X(1-2X)\beta\} \approx 2\beta$ by assuming $X, \beta \ll 1$. The fractions in different gene groups may be similarly obtained and are shown in Fig. 5. In addition, the proportions of various recombinant types are also given at the left of the diagram. The proportion is determined by the assumption that the point of crossing-over is equally likely to occur at any point in the paired region.

Based on the frequency of recombinant chromosomes and the proportion of recombinant types, it is possible to calculate the changes of the identity coefficients.

First, c_{21}' changes to c_{21}'' by inter-chromosomal crossing-over according to the following equation,

$$c_{21}'' = (1-2\beta X)c_{21}' + \frac{\beta X}{2}(c_{31}' + 2\phi_{2,3}' + \phi_{1,2}'). \tag{14}$$

The changes of f_{20} and f_{21} take a similar form,

$$f_{2i}'' = (1-\tfrac{5}{2}\beta X)f_{2i}' + 2\beta X\phi_{2,3}' + \frac{\beta X}{2}\phi_{1,2}', \quad i = 0, 1. \tag{15}$$

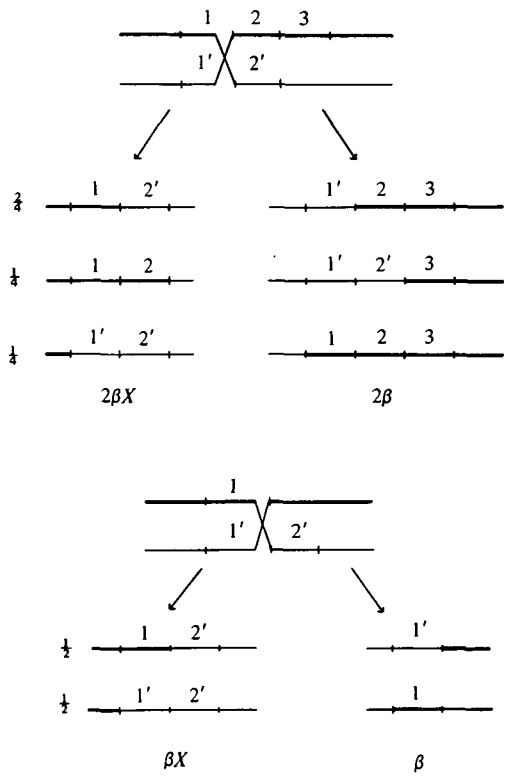

Fig. 5. Diagram illustrating the inter-chromosomal crossing-over of the selection model. The proportion of various recombinant types is shown by the figures on the left, and the frequency of recombinant chromosomes produced in one generation by the symbols below the diagram.

The coefficients of formula (15) were derived by considering the heterozygous frequencies of recombinant chromosomes and recombinant type frequencies as given in Fig. 5. The changes of the identity coefficients of the three-gene group may be similarly obtained.

$$c_{31}'' = (1-\beta)c_{31}' + \tfrac{1}{4}\beta c_{21}' + \tfrac{3}{4}\beta\phi_{2,3}', \tag{16}$$

and

$$c_{32}'' = (1-\tfrac{3}{2}\beta)c_{32}' + \tfrac{3}{2}\beta\phi_{2,3}'. \tag{17}$$

https://doi.org/10.1017/S0016672300020115 Published online by Cambridge University Press

The changes of f_{3i} ($i = 0, 1, 2$) are more complicated. As before, consider heterozygous frequency of recombinant chromosomes. As to the gene comparisons of two chromosomes, all three genes contribute equally for f_{30} (zero step), however for f_{31} (one step) the middle gene contributes twice as much as the others do, and for f_{32} there is no contribution of the middle gene. Therefore, the changes become

$$\left.\begin{aligned} f_{30}'' &= (1-\tfrac{4}{3}\beta)f_{30}' + \tfrac{4}{3}\beta\phi_{2,3}' \\ f_{31}'' &= (1-\tfrac{5}{4}\beta)f_{31}' + \tfrac{5}{4}\beta\phi_{2,3}' \\ f_{32}'' &= (1-\tfrac{3}{2}\beta)f_{32}' + \tfrac{3}{2}\beta\phi_{2,3}'. \end{aligned}\right\} \quad \text{and} \quad (18)$$

The transition of the identity coefficient of one-gene group becomes

$$f_{10}'' = (1-\beta)f_{10}' + \beta\phi_{1,2}'. \tag{19}$$

Let us proceed to obtain the changes of the identity coefficients between the groups. In order to calculate the change of $\phi_{2,3}$, let us define the following average identity coefficients:

$$\left.\begin{aligned} \bar{f}_2 &= \tfrac{1}{2}(f_{20}' + f_{21}') \\ \bar{f}_3 &= \tfrac{1}{9}(3f_{30}' + 4f_{31}' + 2f_{32}'). \end{aligned}\right\} \quad \text{and} \quad (20)$$

From the fraction of crossover products, and by considering the proportion of affected genes as before, we get for the change of $\phi_{2,3}'$

$$\phi_{2,3}'' = (1-\tfrac{5}{4}\beta X - \tfrac{2}{3}\beta)\phi_{2,3}' + \beta X\bar{f}_3 + \tfrac{1}{4}\beta X\phi_{1,3}' + \tfrac{2}{3}\beta\bar{f}_2. \tag{21}$$

Next, the transition of $\phi_{1,2}$ becomes

$$\phi_{1,2}'' = (1-\tfrac{5}{4}\beta X - \beta)\phi_{1,2}' + \beta X\phi_{1,3}' + \tfrac{1}{4}\beta X f_{10}' + \beta\bar{f}_2. \tag{22}$$

As to $\phi_{1,3}$,

$$\phi_{1,3}'' = (1-\tfrac{7}{6}\beta)\phi_{1,3}' + \tfrac{2}{3}\beta\phi_{2,3}' + \tfrac{1}{2}\beta\phi_{1,2}'. \tag{23}$$

In addition to the above transition of identity coefficients, it is necessary to incorporate the changes by inter-chromosomal *equal* crossing-over between the chromosomes of the same group. Let us assume that the term $\beta/(2N_e)$ is negligibly small, then the changes of the identity coefficients between the chromosomes are negligible and only the following transitions need to be considered, as in the case of a large multigene family (Ohta, 1978).

$$\left.\begin{aligned} c_{21}''' &= (1-\beta)c_{21}'' + \beta f_{21}'' \\ c_{31}''' &= (1-\beta)c_{31}'' + \beta f_{31}'' \\ c_{32}''' &= (1-2\beta)c_{32}'' + 2\beta f_{32}''. \end{aligned}\right\} \quad \text{and} \quad (24)$$

Here I assume that the recombination rate is proportional to the number of steps between the two genes to compare identity, i.e. it is βi when the two genes are i steps apart.

Finally, we shall proceed to obtain the effects of random genetic drift and mutation. For calculation, it is assumed that the frequency, X, is constant and random sampling occurs within each group. Then we get the following set of transition equations:

$$f_{30}''' = \left(1-\frac{1}{2N_eX}\right) f_{30}'' + \frac{1}{2N_eX}, \tag{25}$$

$$f_{3i}''' = \left(1-\frac{1}{2N_eX}\right) f_{3i}'' + \frac{1}{2N_eX}\, c_{3i}''', \quad \text{for} \quad i = 1, 2, \tag{26}$$

$$f_{10}''' = \left(1-\frac{1}{2N_eX}\right) f_{10}'' + \frac{1}{2N_eX}, \tag{27}$$

$$f_{20}''' = \left\{1-\frac{1}{2N_e(1-2X)}\right\} f_{20}'' + \frac{1}{2N_e(1-2X)}, \tag{28}$$

and

$$f_{21}''' = \left\{1-\frac{1}{2N_e(1-2X)}\right\} f_{21}'' + \frac{1}{2N_e(1-2X)}\, c_{21}'''. \tag{29}$$

Other identity coefficients are not influenced by random genetic drift. The changes of identity coefficients by mutation are rather simple, and all coefficients are reduced by a factor equal to twice the mutation rate (Kimura & Crow, 1964; Ohta, 1978),

$$\left.\begin{aligned} c_{ij}^{\mathrm{IV}} &= (1-2v)c_{ij}''', \\ f_{kl}^{\mathrm{IV}} &= (1-2v)f_{kl}''' \\ &\text{and} \\ \phi_{m,n}^{\mathrm{IV}} &= (1-2v)\phi_{m,n}''', \quad \text{for all } i, j, k, l, m \text{ and } n \text{ defined,} \end{aligned}\right\} \tag{30}$$

where v is the mutation rate per gene, which is much less than unity, and $\phi_{m,n}''' = \phi_{m,n}''$.

By using the equations (8)–(30), it is possible to generate identity coefficients from one generation to the next. The equilibrium values are particularly interesting, since they may be compared with the observed data such as those by Zimmer *et al.* (1980). Also, it would be very convenient if a correspondence between the present rather complicated system and the much simpler cycle model in the previous section is available. In the next section, comparison of the equilibrium identity coefficients between the cycle and the selection models will be given.

(iii) *Comparison of the two models and some applications*

The equilibrium values of identity coefficients of the selection model were numerically obtained starting from the homogeneous gene family (i.e. all twelve identity coefficients were unity) and by generating a large number of generations until the coefficients become unchanged, $2/v$ generations in the present case. In order to compare these values with the prediction of the cycle model (equations 6), let us examine the correspondence of parameters of the two models. Let us refer to the figures 1, 4 and 5. In the cycle model, the fraction of the new type chromo-

Table 2. *Examples of identity coefficients at equilibrium, calculated numerically by equations* (8)–(30). *The identity coefficients of the two-gene group may be compared with the prediction of the cycle model (equation 6), which are given in parentheses.*

Parameters		$\beta = 5\times10^{-3}$	$\beta = 10^{-3}$	$u = 5\times10^{-3}$	$u = 10^{-3}$	$X' = 0{\cdot}01$	$X' = 0{\cdot}05$	$X' = 0{\cdot}1$	$N_e = 100$	$N_e = 1000$
c_{31}	0·749	0·729	0·694	0·719	0·654	0·695	0·794	0·803	0·810	0·693
c_{32}	0·540	0·483	0·395	0·514	0·518	0·422	0·656	0·708	0·649	0·439
c_{21}	0·524	0·472	0·392	0·482	0·433	0·405	0·642	0·701	0·630	0·426
(c_1, equation 6)	(0·518)	(0·476)	(0·469)	(0·482)	(0·446)	(0·379)	(0·677)	(0·771)	(0·655)	(0·426)
f_{30}	0·689	0·671	0·642	0·698	0·786	0·669	0·726	0·748	0·894	0·550
f_{31}	0·653	0·632	0·599	0·640	0·623	0·608	0·703	0·728	0·797	0·527
f_{32}	0·624	0·597	0·556	0·598	0·559	0·554	0·687	0·718	0·716	0·508
f_{20}	0·753	0·761	0·769	0·760	0·765	0·769	0·748	0·752	0·937	0·606
(f_0, equation 6)	(0·766)	(0·778)	(0·795)	(0·769)	(0·773)	(0·781)	(0·757)	(0·764)	(0·944)	(0·619)
f_{21}	0·524	0·473	0·398	0·483	0·434	0·406	0·639	0·694	0·632	0·423
(f_1, equation 6)	(0·500)	(0·452)	(0·426)	(0·466)	(0·432)	(0·368)	(0·640)	(0·712)	(0·648)	(0·400)
f_{10}	0·689	0·671	0·642	0·699	0·792	0·669	0·726	0·749	0·894	0·550
$\phi_{2,3}$	0·638	0·616	0·583	0·620	0·596	0·587	0·692	0·720	0·784	0·514
$\phi_{1,2}$	0·638	0·616	0·583	0·620	0·596	0·587	0·692	0·720	0·784	0·514
$\phi_{1,3}$	0·638	0·616	0·583	0·620	0·594	0·587	0·691	0·719	0·784	0·514

Parameters are $u = 10^{-2}$, $\beta = 10^{-2}$, $N_e = 500$, $v = 10^{-4}$ and $X' = 0{\cdot}02$, otherwise indicated at the top line.

some in which gene position is shifted is $\gamma/2$. In the two-gene group of the selection model, the fraction of chromosomes which receive genes from the three-gene or one-gene group is approximately $uX + 2\beta X$. Note equations (1), (13) and (14) in which these proportions are coefficients to give the change of identity coefficient of one-step distance and on the same chromosome. Thus, it would be expected that γ (rate of cycles of unequal crossing-over) in equations (6) corresponds to $2(u + 2\beta)X$ of the selection model. Other parameters (N_e, v and β) should be equivalent in both the models. Numerical examples of the equilibrium identity coefficients of the selection model are given in Table 2. Among them, the coefficients of the two-gene group, c_{21}, f_{20} and f_{21}, may be compared with the prediction calculated from equations (6) of the cycle model, by replacing γ with $2(u + 2\beta)X$ which are also given in the table in parentheses. From the table, it can be seen that the agreement between the predictions of the two models is generally satisfactory.

From the above results, it may be concluded that the formulae (6) are applicable to the observed data by equating,

$$\gamma = 2(u + 2\beta)X, \tag{31}$$

even if the cycle model is biologically unrealistic. Now, let us examine recent observations in the light of the above analyses. From Table 1 of Zimmer *et al.* (1980), the two duplicated loci of haemoglobin α of primates are almost identical, i.e. only one species (orang-utan) out of five produces two kinds of haemoglobin α which differ by one amino acid. Then the identity coefficients approximately become $c_1 \approx f_1 \approx 0{\cdot}8$ and $f_0 \approx 1$, in terms of the amino acid identity of the total polypeptide, and one would expect from the formulae (6), that $8N_e v + N_e\gamma \ll 2$, $4v + \gamma \ll 2\beta$ and $4v < \gamma$. By estimating $v \approx 1{\cdot}5 \times 10^{-7}$ per year from the amino acid substitution rate at the haemoglobin loci, γ is predicted to be $10^{-5} \sim 10^{-6}$ per year, agreeing with Zimmer *et al.*'s estimation of the age of the duplication event. If X (frequency of chromosomes with three genes or that with one gene) is about 10^{-2} (Goossens *et al.* 1980), the above value of γ implies that $2(u + 2\beta) \approx 10^{-3} \sim 10^{-4}$ from equation (31). It would be expected that $u < \beta$, then u is probably about 10^{-4} per generation, and the selection coefficient against chromosomes with three genes or one gene would be roughly 10^{-2}. The above argument may be too simple, and in some cases the chromosomes with three genes may spread in the population, as was found in the chimpanzee (Zimmer *et al.* 1980).

So far, unequal crossing-over is considered to be the sole mechanism of concerted evolution. However, it is possible that gene conversion is also responsible for the concerted evolution of tandem genes (Lauer *et al.* 1980). If one of the two copies corrects the other, the duplication of the former and the deletion of the latter will result. Thus one gene conversion would have the same effect as one cycle of the simple model.

Haemoglobin α gene family of primates could be rather exceptional, and other well-studied cases often reveal differentiation between the tandem genes both for the primary structure and for the amount used; haemoglobin β vs. δ in primates, haemoglobin β major vs. β minor in mouse (Konkel *et al.* 1978) and in rabbit

(Hardison *et al.* 1979). Table 1 of Zimmer *et al.* (1980) states that the average difference between haemoglobin δ and β of primates is 9·3 amino acids. In such cases, the parameter γ must be much smaller than in the above haemoglobin α gene family. A probable cause is that the selection coefficient against chromosomes with unbalanced gene content is much larger when tandem genes are differentiated, in addition to less occurrence of unequal crossing-over.

Another interesting example is haemoglobin γ loci in man (Jeffreys, 1979). There are two haemoglobin γ genes in man and they are another example of a small multigene family. Jeffreys (1979) has found that there is a polymorphism with respect to a restriction enzyme map in the intervening sequence of this gene. What is remarkable is that this polymorphism exists in both of the tandem γ globin genes. Thus the identity coefficient, f_0, is not unity and may be the same as f_1 when measured by this restriction enzyme, i.e. gene identity may be the same whether or not the two genes locate at the homologous site. From Table 2, and from equations (6), it may be seen that f_0 is larger than c_1 and f_1 under the usual condition of $8N_e v + N_e\gamma < 2$. However, they merely represent average values, and individual cases may differ greatly even if the parameters take similar values. By chance, we may observe $f_0 \approx f_1$ or even $f_0 < f_1$. The situation is analogous to the enzyme loci and identity coefficients accompany large variance.

More recent studies on this gene family are directed towards utilizing the linkage phase of such polymorphisms for pre-natal diagnosis of sickle-cell anaemia (Phillips III *et al.* 1980) and thalassaemia (Little *et al.* 1980). It would be preferable to establish a theory of linkage disequilibrium under the present model.

All these examples belong to the globin gene family. It is expected that more data will be available in the near future on the fine structure of other gene families to apply the present results.

3. DISCUSSION

The theory presented here is too idealistic in many respects. In particular, the cycle model is not based on biologically known phenomena. Nevertheless, it provides a simple way to formulate the process of concerted evolution. The rate of cycles of unequal crossing-over (Fig. 1), γ, may be regarded as the rate of duplication and loss, which are the basis of concerted evolution, and are the consequence of unequal crossing-over, natural selection and inter-chromosomal recombination of the more realistic selection model. It was shown that $\gamma = 2(u + 2\beta)X$, i.e. twice the product of the frequency of chromosomes with three genes or that with one gene (X) and the sum of the rate of unequal crossing-over (u) and that of inter-chromosomal recombination (2β). As estimated in the previous section, $\gamma = 10^{-5} \sim 10^{-6}$ and $u \approx 10^{-4}$ for the haemoglobin α gene family of primates. Thus u may be considerably high and unequal crossing-over should be seriously considered as one of the driving forces of evolution.

In the present analyses, natural selection was considered only as a factor which keeps the number of duplicated genes stable, and mutations are assumed to be

selectively neutral. For multigene families, the theory of natural selection becomes very complicated and awaits future study. For some models of natural selection on large multigene families, see Ohta (1980). Even selection on the number of genes per chromosome may not be so simple as considered here. It is possible that selection is not so effectively at work in eliminating chromosomes with three or single genes. Particularly, the chromosomes with one extra copy of the gene may be almost selectively neutral. In fact, Nishioka, Leder & Leder (1980) found that an extra gene copy of α globin exists in a mouse genome, which is apparently degenerated and not used. Such non-functional genes could also have originated in the genome from unequal crossing-over, when selection is not so effective as assumed in our second model and defective mutations have accumulated before they are eliminated from the population. If the rate of occurrence of defective mutations is v_D per generation, and if a chromosome with extra-gene which is non-functional is selectively neutral, then the rate of accumulation of defective genes in the population would be $v_D \times X = v_D u/(2s)$ (see equation 7). Such an argument may generally apply to ordinary single-copy genes, although u is much smaller than that for a multigene family.

Another factor which should be considered is that the rate of unequal crossing-over is likely to vary with time. The rate may depend on the presence or absence of internal repetition at the non-coding region as well as on the identity and total length of repeating units (Fedoroff & Brown, 1978; Seidman *et al.* 1978). Thus one would need to investigate a model in which the rate of unequal crossing-over is some function of the internal structure of genes.

Recently, considerable effort has been made to investigate the population genetics of a pair of duplicated genes (Allendorf, Utter & May, 1975; Ferris & Whitt, 1977; Bailey, Poulter & Stockwell, 1978; Kimura & King, 1979; Takahata & Maruyama, 1979). These studies treat the case where gene duplication is caused by polyploidization, and unequal crossing-over was not considered. When genes are tandemly duplicated, for which data are rapidly accumulating, however, unequal crossing-over needs to be taken into account. Ohno (1970) emphasized the importance of gene duplication in evolution (based on biochemical and cytological facts). Now it is time to develop population genetics of gene duplication.

I thank Dr A. Robertson for suggesting correct formulation of equations (13) and for his many other useful comments on the manuscript.

REFERENCES

Allendorf, F. W., Utter, F. M. & May, B. P. (1975). Gene duplication within the family salmonidae. II. Detection and determination of the genetic control of duplicate loci through inheritance studies and the examination of populations. In *Isozymes*, vol. 4 (ed. C. L. Markert), pp. 415–432. New York: Academic.

Bailey, G. S., Poulter, R. T. M. & Stockwell, P. A. (1978). Gene duplication in tetraploid fish: model for gene silencing at unlinked duplicate loci. *Proceedings of the National Academy of Sciences, U.S.A.* **75**, 5575–5579.

BHAT, S. P., JONES, R. E., SULLIVAN, M. A. & PIATIGORSKY, J. (1980). Chicken lens crystallin DNA sequences show at least two δ-crystallin genes. *Nature* **284**, 234–238.

DAYHOFF, M. O. (1972). *Atlas of Protein Sequence and Structure 1972*. National Biomedical Research Foundation, Silver Spring, Maryland.

FEDOROFF, N. V. & BROWN, D. D. (1978). The nucleotide sequence of the repeating unit in the oocyte 5S ribosomal DNA of *Xenopus laevis*. *Cold Spring Harbor Symposia on Quantitative Biology* **42**, 1195–1200.

FERRIS, S. D. & WHITT, G. S. (1977). Loss of duplicate gene expression after polyploidisation. *Nature* **265**, 258–260.

GOOSSENS, M., DOZY, A. M., EMBURY, S. H., ZACHARIADES, Z., HADJIMINAS, M. G., STAMATOYANNOPOULOS, G. & KAN, Y. W. (1980). Triplicated α-globin loci in humans. *Proceedings of the National Academy of Sciences, U.S.A.* **77**, 518–521.

HARDISON, R. C., BUTLER III, E. T., LACY, E., MANIATIS, T., ROSENTHAL, N. & EFSTRATIADIS, A. (1979). The structure and transcription of four linked rabbit β-like globin genes. *Cell* **18**, 1285–1297.

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

HOOD, L., CAMPBELL, J. H. & ELGIN, S. C. R. (1975). The organization, expression, and evolution of antibody genes and other multigene families. *Annual Review of Genetics* **9**, 305–353.

JEFFREYS, A. J. (1979). DNA sequence variants in the $^{G}\gamma$-, $^{A}\gamma$-, δ- and β-globin genes of man. *Cell* **18**, 1–10.

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

KIMURA, M. & KING, J. L. (1979). Fixation of a deleterious allele at one of two 'duplicate' loci by mutation pressure and random drift. *Proceedings of the National Academy of Sciences, U.S.A.* **76** 2858-2861.

KIMURA M. & OHTA T. (1979). Population genetics of multigene family with special reference to decrease of genetic correlation with distance between gene members on a chromosome. *Proceedings of the National Academy of Sciences, U.S.A.* **76** 4001–4005.

KONKEL, D. A., TILGHMAN, S. M. & LEDER, P. (1978). The sequence of the chromosomal mouse β-globin major gene: homologies in capping, splicing and poly (A) sites. *Cell* **15**, 1125–1132.

LAUER, J., SHEN, C. J. & MANIATIS, T. (1980). The chromosomal arrangement of human α-like globin genes: sequence homology and α-globin gene deletions. *Cell* **20**, 119–130.

LITTLE, P. F. R., ANNISON, G., DARLING, S., WILLIAMSON, R., CAMBA, L. & MODELL, B. (1980). Model for antenatal diagnosis of β-thalassaemia and other monogenic disorders by molecular analysis of linked DNA polymorphisms. *Nature* **285**, 144–147.

NAKANISHI, S., INOUE, A., KITA, T., NAKAMURA, M., CHANG, A. C. Y., COHEN, S. N. & NUMA, S. (1979). Nucleotide sequence of cloned cDNA for bovine corticotropin-β-lipotropin precursor. *Nature* **278**, 423-427.

NISHIOKA, Y., LEDER, A. & LEDER, P. (1980). An unusual alpha globin-like gene that has cleanly lost both globin intervening sequences. *Proceedings of the National Academy of Sciences, U.S.A.* **77**, 2806–2809.

OHNO, S. (1970). *Evolution by Gene Duplication*. Berlin: Springer.

OHTA, T. (1976). A simple model for treating the evolution of multigene families. *Nature* **263**, 74–76.

OHTA, T. (1978). Theoretical study on genetic variation in multigene families. *Genetical Research* **31**, 13–28.

OHTA, T. (1980). *Evolution and Variation of Multigene Families*. Lecture Notes in Biomathematics, vol. 37. New York: Springer.

PETES, T. D. (1980). Unequal meiotic recombination within tandem arrays of yeast ribosomal DNA genes. *Cell* **19**, 765–774.

PHILLIPS III, J. A., PANNY, S. R., KAZAZIAN, JR., H. H., BOEHM, C. D., SCOTT, A. F. & SMITH, K. D. (1980). Prenatal diagnosis of sickle cell anemia by restriction endonuclease analysis: hind III polymorphism in γ-globin genes extend test applicability. *Proceedings of the National Academy of Sciences, U.S.A.* **77**, 2853–2856.

ROYAL, A., GARAPIN, A., CAMI, B., PERRIN, F., MANDEL, J. L., LEMEUR, M., BREGEGEGRE, F., GANNON, F., LEPENNEC, J. P., CHAMBON, P. & KOURILSKY, P. (1979). The ovalbumin gene region: common features in the organization of three genes expressed in chicken oviduct under hormonal control. *Nature* **279**, 125–132.

SEIDMAN, J. G., LEDER, A., NORMAN, M. N. B. & LEDER, P. (1978). Antibody diversity. *Science* **202**, 11–17.

SMITH, G. P. (1974). Unequal crossover and the evolution of multigene families. *Cold Spring Harbor Symposia on Quantitative Biology* **38**, 507–513.

SZOSTAK, J. W. & WU, R. (1980). Unequal crossing-over in the ribosomal DNA of *Saccharomyces cerevisiae*. *Nature* **284**, 426–430.

TAKAHATA, N. & MARUYAMA, T. (1979). Polymorphism and loss of duplicate gene expression: a theoretical study with application to tetraploid fish. *Proceedings of the National Academy of Sciences, U.S.A.* **76**, 4521- 4525.

TARTOF, K. D. (1975). Redundant genes. *Annual Review of Genetics* **9**, 355–385.

ZIMMER, E. A., MARTIN, S. L., BEVERLEY, S. M., KAN, Y. W. & WILSON, A. C. (1980). Rapid duplication and loss of genes coding for the α chains of hemoglobin. *Proceedings of the National Academy of Sciences, U.S.A.* **77**, 2158–2162.