*Genet. Res., Camb.* (1980), **36**, *pp.* 181–197
*With 4 text-figures*
*Printed in Great Britain*



# Linkage disequilibrium between amino acid sites in immunoglobulin genes and other multigene families*

By TOMOKO OHTA

*National Institute of Genetics, Mishima, 411, Japan*

(*Received* 22 *March* 1979 *and in revised form* 12 *May* 1980)

SUMMARY

Linkage disequilibrium between segregating amino acid sites in repeated genes which form a multigene family was investigated by using the population genetics theory. The degree of non-random association of amino acids is measured by the disequilibrium coefficient which is determined by the balance among various forces; unequal and equal crossing-over, mutation, random genetic drift and recombination which takes place between the two sites in question. Another measure of disequilibrium, 'standardized identity excess', represents excess probability of simultaneous identity at the two sites over that expected from random association of amino acids. Sequence data of variable region of immunoglobulins provide most interesting example of multigene family to apply the theory. Statistical analyses on identity excess for various groups and subgroups of variable region sequences of immunoglobulins suggest that a multigene family such as human $\kappa$ or mouse $\kappa$ gene family consists of several subfamilies between which recombination is limited. The analyses also indicate that the recombination may take place between any gene members in one subfamily.

## 1. INTRODUCTION

The evolution of repeated genes forming a multigene family provides a very exciting subject in molecular evolutionary studies of higher organisms. In particular, the origin of the antibody diversity is one of the most stimulating problems in recent years (e.g. Hood, Campbell & Elgin, 1975). One factor which should not be forgotten in the discussion on sequence variability of immunoglobulins is the degree of association between amino acids *within a polypeptide*. Consider, for example, two sites in a polypeptide each with two kinds of amino acids. Then under the random combination, four kinds of polypeptide are produced. However, if one amino acid of the first site completely associates with an amino acid of the second one, only two kinds of polypeptide would be produced. This sort of association is called 'linkage disequilibrium' in population genetics.

In the present paper, linkage disequilibrium *within* a gene such as those of the variable part of immunoglobulins of multigene families will be investigated. I consider amino acid sites in a polypeptide and study the nature of combination

* Contribution no. 1312 from the National Institute of Genetics, Mishima, Shizuoka-ken, 411, Japan.

0016–6723/80/2828–8210 $01.00 

of amino acids between the sites. In case of variable regions of immunoglobulins, presence of subgroups is reported in human $\kappa$, mouse $\kappa$ and other gene families. Their presence indicates non-random association of amino acids in polypeptides. The model of unequal crossing-over is used since it provides one of the most satisfactory explanations for the coincidental evolution and the contraction–expansion of the gene number of the multigene family (Smith, 1974; Black & Gibson, 1974). Here, it is assumed that the unequal crossing-over is continuously occurring in a germ cell line and is responsible for the horizontal expansion or contraction of mutant genes in a chromosome. I have shown that the theory of gene frequency in finite population may be applicable for analysing mutant dynamics in a gene family of one chromosome (Ohta, 1976). The analyses are further extended to the evolutionary studies of multigene family at the level of the population which take into account of not only unequal crossing-over and mutation but also random sampling of chromosomes and ordinary inter-chromosomal crossing-over (Ohta, 1978*a*, *b*). Non-random association of amino acids is essentially due to the expansion of some lucky combinations. The non-randomness is broken by recombination which takes place between the sites in question. In the present report, I shall evaluate the degree of non-random association under simple assumptions. The results will be applied to the analyses on sequence variability of immunoglobulins.

## 2. BASIC THEORY

A random mating population with effective size $N_e$ is assumed and a multigene family is present in a chromosome and undergoes continuously the unequal somatic crossing-over at mitotic division of germ cells, ordinary inter-chromosomal crossing-over, mutation and random genetic drift. Thus, one generation consists of the following events: mutation → random genetic drift → intra-chromosomal (sister-chromatid) unequal crossing-over → inter-chromosomal equal and unequal crossing-over. Here inter-chromosomal crossing-over may be equal or unequal. However, I shall start from the simple case in which inter-chromosomal crossing-over is always equal. Later, I shall consider more general cases. Let $\beta$ be the rate of inter-chromosomal crossing-over per family. There are $n_g$ genes in a family and each gene contains $n_s$ sites as the following diagram illustrates. Let $v$ be the mutation rate per one site per generation and all mutations are assumed to be unique (Kimura & Crow, 1964). Thus, in one generation, $n_s v$ new mutations per one gene unit, $n_g n_s v$ new mutations per one gene family and $2N_e n_g n_s v$ in a total population are expected to occur. One cycle of unequal crossovers is defined as two consecutive unequal crossovers of duplication and deletion of *one gene* and thus the gene number per family ($n_g$) stays the same (Ohta, 1976; Perelson & Bell, 1977). Let $\kappa$ be the equivalent (effective) number of cycles of unequal crossover per family per generation, when the shift at unequal crossing-over is one or more genes. By letting $m$ be the average number of gene shift, $\kappa$ is very roughly $m/2$ times the rate of unequal crossing-over. Then the process of unequal crossing-over may be described by a basic quantity just like random genetic drift, i.e. the rate of increase of the probability of two homologous units being identical (identity coefficient). This rate ($\alpha$) may be approximately

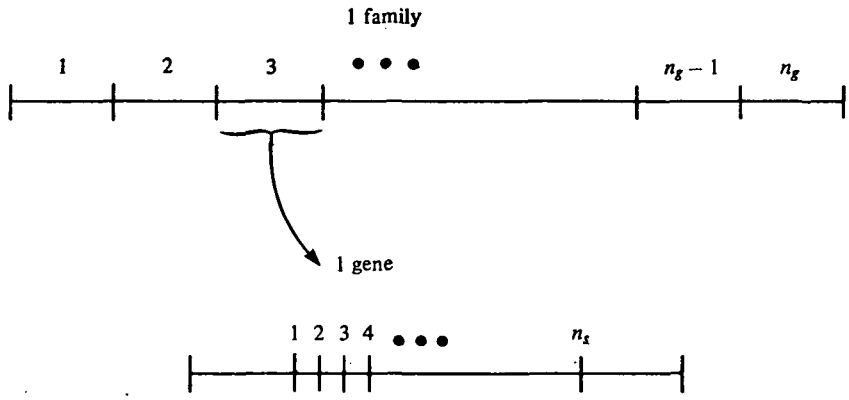


Fig. 1. Diagram illustrating the structure of a multigene family.

expressed as follows, by letting $\gamma$ be the rate of unequal crossing-over per gene family with mean shift of $m$ genes, $\alpha \approx 2\kappa/n_g^2 \approx m\gamma/n_g^2$ (Ohta, 1980).

Although, the shift at unequal crossing-over occurs by multiple of the whole gene unit, actual crossing-over may take place anywhere inside of the gene. Let us fix our attention to a particular pair of sites ($A$ and $B$) and consider recombination between these two (Fig. 2). Let $\xi$ be the recombination rate between the two sites per gene per generation.

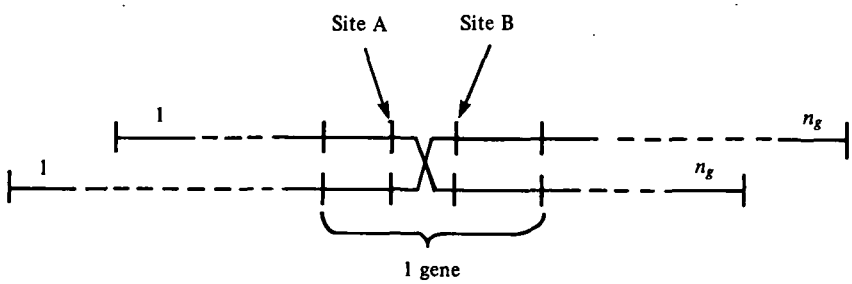


Fig. 2. Diagram showing the recombination between the two sites at unequal crossing-over.

Under the equilibrium among mutation, random drift and crossing-over, how much non-random association between mutants in the two sites of a gene of the family may be generated? This is analogous to the two-locus problem in finite populations in population genetics (Hill & Robertson, 1968; Kimura & Ohta, 1971). In our case, the non-random association is mainly created by expansion of certain lucky combination of mutants through unequal crossing-over, whereas, in the latter case, random genetic drift plays the essential role. The analogy may be extended to use the established theory in population genetics. Thus I shall treat the two sites as two linked loci.

Let us designate by $x_{i,k}$ and $y_{j,k}$ the frequency of the $i$-th amino acid of the first site ($A_i$) and that of the $j$-th amino acid ($B_j$) of the second site in the $k$-th gene family

https://doi.org/10.1017/S0016672300019790 Published online by Cambridge University Press

of the $2N_e$ families which form the population. Let $f_{ij,k}$ be the frequency of a gene with $A_iB_j$ combination in the $k$-th family. Then the degree of association between $A_i$ and $B_j$ in the $k$-th family is expressed by $D_{ij,k} = f_{ij,k} - x_{i,k}y_{j,k}$, which is called linkage disequilibrium between $A_i$ and $B_j$. In our case, the situation is more complicated because there are $2N_e$ gene families and more than one kind of gene families is present in the population. Thus, another measure of disequilibrium would be needed in the following form:

$$d_{ij} = \underset{l \neq k}{E} \{x_{i,k}y_{j,k} - x_{i,k}y_{j,l}\} \tag{1}$$

in which $\underset{l \neq k}{E}$ denotes to taking the expectation for all pairs of $k$ and $l$ in the population. This measure represents the difference of the probability of having $A_i$ and $B_j$ in the two gene members of one family and that of having $A_i$ in a gene of one family and $B_j$ in a gene of another family of the population. The two probabilities are also defined:

$$\left.\begin{aligned} &\zeta_{ij,1} = E(x_{i,k}y_{j,k}) \\ \text{and} \quad &\\ &\zeta_{ij,2} = \underset{l \neq k}{E}(x_{i,k}y_{j,l}). \end{aligned}\right\} \tag{2}$$

Next, let us consider to make summation of $x_{i,k}y_{j,k}$ or $x_{i,k}y_{j,l}$ for all $i$ and $j$ (amino acids), and then to take expectation for $k$ and $l$ in the population:

$$\left.\begin{aligned} &\zeta_{w1} = E\{\sum_i \sum_j (x_{i,k}y_{j,k})\} \\ \text{and} \quad &\\ &\zeta_{w2} = \underset{l \neq k}{E}\{\sum_i \sum_j (x_{i,k}y_{j,l})\}. \end{aligned}\right\} \tag{3}$$

By considering the similar formulation of $\zeta_{w1}$ and $\zeta_{w2}$ with the identity coefficients, $C_{w1}$ and $C_{w2}$ (see Ohta, 1978*a*) it is easy to find their rates of changes through random genetic drift and inter-chromosomal crossing-over (gene exchange between the families). By random genetic drift of one generation, $\zeta_{w1}$ does not change but $\zeta_{w2}$ changes by the amount $(\zeta_{w1} - \zeta_{w2})/(2N_e)$. Through inter-chromosomal gene exchange, $\zeta_{w1}$ changes by the amount of $\beta(\zeta_{w2} - \zeta_{w1})/3$ whereas the change of $\zeta_{w2}$ is almost negligible (Ohta, 1978*a*). Thus, when $1/(2N_e)$ and $\beta$ are much larger than the other parameters; $\alpha$ (rate of increase of identity coefficient within a family), $v$ (mutation rate per site) and $\xi$ (recombination rate between the two sites), $\zeta_{w1}$ and $\zeta_{w2}$ are expected to be almost equal. Therefore, under such a condition, the second measure of disequilibrium defined by the formula (1) would be negligible and we need to consider only the expectations of disequilibrium within a single gene family. In the present study, it is assumed that this condition is satisfied.

When the inter-chromosomal crossing-over is unequal, it is divided into gene exchange (between the families) process and contraction–expansion process (Ohta, 1979). The former is reduced to the process of equal inter-chromosomal crossing-over with a proper correction for the rate (Ohta, 1980). When $N_e$ is sufficiently small, the latter may be similarly treated as the process of intra-chromosomal unequal

crossing-over. If the rate of inter-chromosomal unequal crossing-over is $\gamma'$ and the mean gene shift at unequal pairing is $m'$, and if $m' > n_g/10$, the total rate of gene exchange between the families is very roughly $\beta' = \beta + \gamma'(1 - 3m'/n_g)$ (Ohta, 1980).

The disequilibrium coefficient, $D_{ij,k}$ may increase due to expansion of certain combinations of $A_i$ and $B_j$ through unequal crossing-over, whereas it decreases when the recombination takes place between the site $A$ and site $B$ at unequal (intra- or inter-chromosomal) crossing-over. This recombination rate between the two sites is $\xi$ per gene per generation. Under such a model, the mean of disequilibrium coefficients would be zero for any combination ($D_{ij,k} = 0$ implies random association of $A_i$ and $B_j$), yet the variance of disequilibrium is not zero which is mainly determined by the balance between unequal crossover and recombination between the two sites. Another measure of disequilibrium, $d_{ij}$, is also expected to be zero on the average. This is because the mean of $\zeta_{w1}$ and $\zeta_{w2}$ would be the same if no systematic pressure is involved. Particularly at equilibrium, $\hat{\zeta}_{w2} = \hat{\zeta}_{w1}$, since $\zeta_{w2}$ changes through random drift by the amount $(\zeta_{w1} - \zeta_{w2})/(2N_e)$, but does not change by mutation. Throughout this paper, hat ($\wedge$) on a letter denotes the equilibrium value. The variance of $d_{ij}$ is not zero again but it is expected to be small when $1/(2N_e)$ and $\beta$ are much larger than $\alpha$, $v$ and $\xi$. In the following, let us proceed bearing the above considerations in mind.

In order to obtain the variance of $D_{ij,k}$, one needs the following vector of moments (Hill, 1975).

$$\mathbf{V} = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} E\{(1 - \sum_i x_{i,k}^2)(1 - \sum_j y_{j,k}^2)\} \\ 4E(\sum_i \sum_j x_{i,k} y_{j,k} D_{ij,k}) \\ 2E(\sum_i \sum_j D_{ij,k}^2) \end{bmatrix}, \tag{4}$$

where $E$ denotes to taking the expectation and summation is for all possible amino acids ($1 \sim 20$). In the following, the theory for the change of $\mathbf{V}$ from one generation to the next will be given and then the equilibrium properties of $\mathbf{V}$ will be presented.

(i) *Change of* $\mathbf{V}$ *due to unequal crossing-over*

Since the process of unequal crossing-over may be treated by the theory of random genetic drift (Ohta, 1976, 1978*a*), the same transition matrix for the moments of the two-locus problems may be used. Here the crucial factor is the rate of increase per generation of identity coefficient, $\bar{C}_A = \sum_i x_{i,k}^2$, which is the probability of identity of two amino acids at homologous sites taken from the gene family. This rate is shown to be $\alpha = 2\kappa/n_g^2 = m\gamma/n_g^2$, provided $m > n_g/10$ ($\kappa$ = the effective number of cycles of unequal crossover, $\gamma$ = rate of unequal crossing-over and $m$ = mean number of gene shift) and is analogous to the rate of increase of homozygosity ($1/2N_e$) due to random genetic drift. $\mathbf{V}$ changes into $\mathbf{V}'$ through one generation of unequal crossing-over. Let $\alpha \ll 1$. Then the transition matrix of $\mathbf{V}$

https://doi.org/10.1017/S0016672300019790 Published online by Cambridge University Press

takes the following form ignoring the higher order terms of $\alpha$ (Hill & Robertson, 1968; Ohta & Kimura, 1969)

$$\mathbf{V}' = \mathbf{K}\mathbf{V} \tag{5}$$

with

$$\mathbf{K} = \begin{bmatrix} 1-2\alpha & \alpha & 0 \\ 0 & 1-5\alpha & 2\alpha \\ 2\alpha & 2\alpha & 1-3\alpha \end{bmatrix}.$$

When inter-chromosomal crossing-over is unequal, the treatment becomes more complicated. However, when $4N_e v$ (four times the product of the effective population size and mutation rate per site) is much less than unity, the inter-chromosomal crossing-over process is almost the same as the intra-chromosomal one in terms of the expansion–contraction process of gene units (Ohta, 1979). Here the inter-chromosomal unequal crossing-over consists of the expansion–contraction process and the gene exchange process between the chromosomes and the latter will be treated in the later section of equal crossing-over (iii).

If $\kappa'$ is the effective number of cycles of inter-chromosomal unequal crossing-overs per generation and $\gamma'$ is the rate of unequal crossing-over with the mean shift of $m'$ genes, $\alpha' = 2\kappa'/n_g^2 = m'\gamma'/n_g^2$ corresponds to the rate of increase of homozygosity due to random drift, just as the case of intra-chromosomal unequal crossing-over. In terms of this parameter, it is possible to express the changes of various moments. The basic quantities needed for the calculation are the mean and variance of the change of random variables, $x_{i,k}$, $y_{i,k}$ and $f_{ij,k}$, in unit time. These quantities in one generation may be expressed by the following formulae (Ohta, 1979), by letting $M(\cdot)$, $V(\cdot)$, and $W(\cdot)$ be respectively the mean, the variance and the covariance of the random variables in parentheses.

$$\left.\begin{aligned} M(\Delta x_{i,k}) &= \frac{n_g\alpha'}{2} \underset{l}{E}(x_{i,l}-x_{i,k}) \\ M(\Delta y_{j,k}) &= \frac{n_g\alpha'}{2} \underset{l}{E}(y_{j,l}-y_{j,k}) \\ V(\Delta x_{i,k}) &= \frac{\alpha'}{2} \underset{l}{E}\{x_{i,k}(1-x_{i,l})+x_{i,l}(1-x_{i,k})\} \\ V(\Delta y_{j,k}) &= \frac{\alpha'}{2} \underset{l}{E}\{y_{j,k}(1-y_{j,l})+y_{j,l}(1-y_{j,k})\} \end{aligned}\right\} \tag{6}$$

and

$$W(\Delta x_{i,k}\Delta x_{g,k}) = -\frac{\alpha'}{2} \underset{l}{E}\{x_{i,k}x_{g,l}+x_{g,k}x_{i,l}\}$$

in which $\underset{l}{E}$ denotes to taking expectation for all $l$ ($\neq k$) in the population.

Let us extend the above formulations to the present two-site model. Clearly we have,

https://doi.org/10.1017/S0016672300019790 Published online by Cambridge University Press

$$\left.\begin{aligned} M(\Delta f_{ij,k}) &= \frac{n_g\alpha'}{2}\underset{l}{E}(f_{ij,l}-f_{ij,k}) \\ V(\Delta f_{ij,k}) &= \frac{\alpha'}{2}\underset{l}{E}\{f_{ij,k}(1-f_{ij,l})+f_{ij,l}(1-f_{ij,k})\} \\ W(\Delta f_{ij,k}\Delta f_{gm,k}) &= -\frac{\alpha'}{2}\underset{l}{E}\{f_{ij,k}f_{gm,l}+f_{gm,k}f_{ij,l}\}. \\ & \qquad (i \neq g \quad \text{and/or} \quad j \neq m) \end{aligned}\right\} \tag{7}$$

and

The covariance of $\Delta x_{i,k}$ and $\Delta y_{j,k}$ becomes by equating

$$\Delta x_{i,k} = \sum_j \Delta f_{ij,k}, \quad \Delta y_{j,k} = \sum_i \Delta f_{ij,k}$$

and by taking the expectation,

$$W(\Delta x_{i,k}\Delta y_{j,k}) = -\frac{\alpha'}{2}\underset{l}{E}\{x_{i,k}y_{j,l}+x_{i,l}y_{j,k}-f_{ij,k}-f_{ij,l}\}. \tag{8}$$

Similarly, the other formulae may be obtained by using $D_{ij,k} = f_{ij,k} - x_{i,k}y_{j,k}$, although the exact resulting equations for $M(\Delta D_{ij,k})$, $W(\Delta D_{ij,k}\Delta x_{i,k})$ and $V(\Delta D_{ij,k})$ are rather complicated and are not given here. Nevertheless, under the simplifying assumption of large values of $1/(2N_e)$ and $\beta'$ as compared with the other parameters, the expected values of the cross products of $x_{i,k}$, $y_{j,l}$ and $f_{ij,h}$ would be almost identical whether they are taken for a single gene family ($k = l = h$) or for two or more gene families ($k \neq l \neq h$, $k = l \neq h$ or $k \neq l = h$).

As an example, let us examine a third moment,

$$\eta_2 = \underset{l \neq k}{E}\{\sum_{ij} x_{i,k}y_{j,k}f_{ij,l}\}.$$

Through one generation of random genetic drift, it changes by the amount, $(\eta_1 - \eta_2)/(2N_e)$ in which $\eta_1 = E\{\sum_{ij} x_{i,k}y_{j,k}f_{ij,k}\}$ and therefore $\eta_1$ and $\eta_2$ should approach when $1/(2N_e) \gg \alpha$, $v$ and $\xi$. The similar argument applies to the other third and the fourth moments of $x_{i,k}$, $y_{j,l}$ and $f_{ij,h}$. Actually, I have found in the numerical studies of the single site model, this principle holds for the third and the fourth moments of $x_{i,k}$. Then, the transition equation corresponding to (5) becomes much simplified and very similar to the formula (5) by replacing $\alpha$ with $\alpha'$. The change of the first element, $X = E\{(1-\sum_i x_{i,k}^2)(1-\sum_j y_{j,k}^2)\}$, is an exception due to the involvement of the second moments, as the following calculation shows.

$$\begin{aligned} E(\Delta X) = & -\sum_i \{V(\Delta x_{i,k}) + 2x_{i,k}M(\Delta x_{i,k})\}(1-\Sigma y_{j,k}^2) \\ & -\sum_j \{V(\Delta y_{j,k}) + 2y_{j,k}M(\Delta y_{j,k})\}(1-\Sigma x_{i,k}^2) \\ & +4\sum_i\sum_j x_{i,k}y_{j,k}W(\Delta x_{i,k}\Delta y_{j,k}) \\ \approx & -2\alpha' X + E\{(1-\sum_j y_{j,k}^2)n_g\alpha'(\sum_i x_{i,k}^2 - \sum_i x_{i,k}x_{i,l})\} \\ & +E\{(1-\sum_i x_{i,k}^2)n_g\alpha'(\sum_j y_{j,k}^2 - \sum_j y_{j,k}y_{j,l})\} + \alpha' Y \\ = & -2\alpha'(1+4n_gN_ev)X + \alpha' Y + 8n_gN_ev\alpha'(1-\hat{C}_{w1}) \end{aligned} \tag{9}$$

in which

$$E\{\sum_i x_{i,k}^2 - \sum_i x_{i,k}x_{i,l}\} = E\{\sum_j y_{j,k}^2 - \sum_j y_{j,k}y_{j,l}\} \approx 4N_e v\hat{C}_{w1}$$

at equilibrium when $4N_e v \ll 1$ and by denoting the equilibrium value of $E\{\Sigma x_{i,k}^2\}$ as $\hat{C}_{w1}$.

The resulting transition equation of the vector $\mathbf{V}$ due to the expansion–contraction process of inter-chromosomal unequal crossing-over becomes as follows.

$$\mathbf{V}' = \mathbf{K}'\mathbf{V} + \mathbf{A}' \tag{10}$$

where

$$\mathbf{K}' = \begin{bmatrix} 1-2\alpha'(1+4n_g N_e v) & \alpha' & 0 \\ 0 & 1-5\alpha' & 2\alpha' \\ 2\alpha' & 2\alpha' & 1-3\alpha' \end{bmatrix}$$

$$\mathbf{A}' = \begin{bmatrix} 8n_g N_e v\alpha'(1-\hat{C}_{w1}) \\ 0 \\ 0 \end{bmatrix}.$$

(ii) *Change of* $\mathbf{V}$ *due to intra-genic recombination*

Disequilibrium is reduced with the rate of intra-genic (between sites) recombination ($\xi$) each generation. $\mathbf{V}'$ changes to $\mathbf{V}''$ in one generation according to the formula (e.g. Hill & Robertson, 1968)

$$\mathbf{V}'' = \mathbf{Q}\mathbf{V}' \tag{11}$$

with

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1-\xi & 0 \\ 0 & 0 & (1-\xi)^2 \end{bmatrix}.$$

Note that the rate $\xi$ should not include the case of equal crossing-over. The intra-genic recombination effectively takes place when unequal crossover occurs between the sites within paired genes. Unequal crossing-over may be intra-chromosomal or inter-chromosomal one. It is assumed here that the arrangement of gene lineages are more or less random along the chromosome, hence the unequal pairing between the homologous genes implies that each unit pairs with a random member of the gene family.

(iii) *Change of* $\mathbf{V}$ *due to inter-chromosomal (equal) crossing-over*

By this process, only the exchange of gene members between the chromosomes is taken into account and neither unequal nor intra-genic recombination are considered. The following figure illustrates an example of inter-chromosomal crossing-over which is assumed to take place with the rate $\beta$ per family per generation. Let us assume that the gene lineages are randomly arranged along the chromosome and suppose that, by the crossing-over, the gene family is divided into two parts; $q$ and $1-q$ where $q$ is assumed to be uniformly distributed. Under these assumptions, it is possible to calculate the amount of changes of various quantities (Ohta, 1978*a*, *b*). The assumption of random arrangement of gene lineages on the

chromosome is not usually satisfied. When the mean number of gene shift at unequal crossing-over is more than 10% of the total gene number in a family, gene arrangement becomes roughly random (Kimura & Ohta, 1979; Ohta, 1980). As a parameter for the rate of gene exchange between the families, let us use

$$\beta' = \beta + \gamma'(1 - m/n_g),$$

the corrected total rate (Ohta, 1980). The last term takes account of the inter-chromosomal unequal crossing-over.

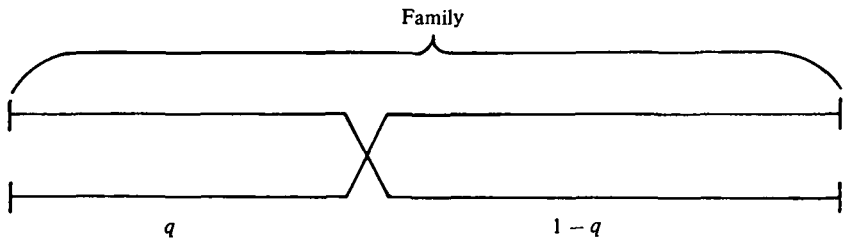


Fig. 3. Diagram of inter-chromosomal equal crossing-over.

As to the change of the first element, $X$, one needs the theory of the single site model of the multigene family; the equilibrium and the transitional properties of the identity coefficient investigated in my previous report (Ohta, 1978*a*; 1979). The first element, $X$, may be expressed, $X = E\{(1-\bar{C}_A)(1-\bar{C}_B)\}$ in which $\bar{C}_A$ and $\bar{C}_B$ designate the identity coefficient of the first and the second sites respectively. At equilibrium,

$$E(\bar{C}_A) = E(\bar{C}_B) = \hat{C}_{w1} = \frac{\alpha + \alpha'}{\alpha + \alpha' \dfrac{1 + 4n_g N_e v}{1 + 4N_e v} + 2v + \dfrac{\beta'}{3} \dfrac{4N_e v}{(1 + 4N_e v)}} \tag{12}$$

and the change of $X$ due to inter-chromosomal crossover becomes

$$\Delta X \approx -2\Delta\hat{C}_{w1}(1 - \hat{C}_{w1}).$$

It was also shown that the change of $\hat{C}_{w1}$ due to inter-chromosomal exchange is $\Delta\hat{C}_{w1} \approx \dfrac{\beta'}{3} \dfrac{4N_e v}{1 + 4N_e v} \hat{C}_{w1}$ at equilibrium (Ohta, 1978*a*). Therefore

$$\Delta X \approx \frac{2\beta'}{3} \cdot \frac{4N_e v}{(1 + 4N_e v)} \{(1 - \hat{C}_{w1}) - X\}$$

$$\approx \frac{8N_e v \beta'}{3} \{(1 - \hat{C}_{w1}) - X\} \tag{13}$$

when $4N_e v \ll 1$. To be exact, $E\{(\Delta C_{w1})^2\}$ should be added to $\Delta X$. However, this involves the difference between the fourth moments of $x_{i,k}$, $x_{i,l}$, $y_{j,k}$ and $y_{j,l}$ and is expected to be quite small when $1/(2N_e)$ and $\beta \gg \alpha$, $v$ and $\xi$.

https://doi.org/10.1017/S0016672300019790 Published online by Cambridge University Press

The change of $Y$ due to gene exchange process is approximately written as follows.

$$\Delta Y \approx 4E\Big[\sum_{ij}\{x_{i,k}y_{j,k}\Delta D_{ij,k}\}+\sum_{ij}\{D_{ij,k}\Delta(x_{i,k}y_{j,k})\} + \sum_{ij}\{\Delta D_{ij,k}\}\{\Delta(x_{i,k}y_{j,k})\}\Big].$$

Through one generation of gene exchange with rate $\beta'$, $E\{\sum_{ij} x_{i,k}y_{j,k}\} = \zeta_{w1}$ changes by the amount $\beta'(\zeta_{w2}-\zeta_{w1})/3$ as explained before. Since $D_{ij,k} = f_{ij,k} - x_{i,k}y_{j,k}$, and $f_{ij,k}$ does not change on the average, $\sum_{ij}(\Delta D_{ij,k}) = -\Delta\zeta_{w1}$. However, at equilibrium $\zeta_{w1} = \zeta_{w2}$ and only the last term of the above equation needs to be considered. Also $\Delta Z \approx 2E\Big[2\sum_{ij} D_{ij,k}(\Delta D_{ij,k}) + \sum_{ij}(\Delta D_{ij,k})^2\Big]$. Therefore, by taking the expectation of the last terms of $\Delta Y$ and $\Delta Z$, and by noting the uniform distribution of $q$ between zero and one (see Fig. 3),

$$2\Delta Z \approx -\Delta Y \approx 16\beta' E[q^2(1-q)^2\sum_{ij}(x_{i,k}y_{j,k}-x_{i,k}y_{j,l})^2] \approx 0.$$

The last approximate equality holds when $1/(2N_e)$ and $\beta'$ are large compared with the other parameters. Thus $\mathbf{V}''$ changes to $\mathbf{V}'''$ by inter-chromosomal crossing-over as follows

$$\mathbf{V}''' = \mathbf{R}\mathbf{V}'' + \mathbf{B} \tag{14}$$

in which

$$\mathbf{R} = \begin{bmatrix} 1-\dfrac{8N_e v\beta'}{3} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and

$$\mathbf{B} = \beta' \begin{bmatrix} \dfrac{8N_e v}{3}(1-\hat{C}_{w1}) \\ 0 \\ 0 \end{bmatrix}.$$

As long as the moments concerning a single gene family are almost equal to their corresponding moments concerning two or more gene families in the population as assumed in our analyses, the moments under consideration would not be influenced by random sampling of chromosomes. However note that the sampling factor ($N_e$) comes into the formula (14) through its effect on individual site.

(iv) *Change of* $\mathbf{V}$ *due to mutation*

Since all mutations are assumed to be unique, every element of $\mathbf{V}$ reduces to $(1-v)^4$ of the previous value each generation by mutation (e.g. Kimura & Crow, 1964). Not only the reduction of the moments but also their increments due to new

https://doi.org/10.1017/S0016672300019790 Published online by Cambridge University Press

mutations need to be considered (Ohta & Kimura, 1971; Hill, 1975). The increment in $X$ due to new mutations is,

$$\Delta X_m \approx E\{2v(1-\bar{C}_A)+2v(1-\bar{C}_B)\} = 4v(1-\hat{C}_{w1}).$$

As to the increment of $Y$, let us consider that a mutant appeared at the $A$ site, and let it be the $z$-th mutant. If it occurred on a gene carrying $B_j$, $D_{zj} = (1-y_{j,k})/n_g$, whereas $D_{zh} = -y_{h,k}/n_g$ for $h \neq j$. Also, $x_{z,k} = 1/n_g$. Therefore the increment of $Y$ by one mutant becomes,

$$\{y_{j,k}(1-y_{j,k}) - \sum_{h \neq j} y_{h,k}^2\}/n_g^2.$$

The probability of occurring of a mutant on a gene with $B_j$ is $y_{j,k}$, and $n_g v$ mutants are expected to occur in each site in one generation. Therefore the expected increment of $Y$ becomes,

$$\Delta Y_m \propto E[\sum_j y_{j,k}\{y_{j,k}(1-y_{j,k}) - \sum_{h \neq j} y_{h,k}^2\}]/n_g^2 = 0.$$

$\Delta Z_m$ may be similarly obtained, which becomes,

$$\Delta Z_m = 4vE(1-\hat{C}_{w1})/n_g.$$

Thus we have,

$$\mathbf{F} = \begin{bmatrix} \Delta X_m \\ \Delta Y_m \\ \Delta Z_m \end{bmatrix} = \begin{bmatrix} 4v(1-\hat{C}_{w1}) \\ 0 \\ \dfrac{4v(1-\hat{C}_{w1})}{n_g} \end{bmatrix}.$$

Therefore, $V''$ changes to $V'''$ due to mutation by the following formula,

$$\mathbf{V}''' = (1-v)^4 \mathbf{V}'' + \mathbf{F}. \tag{15}$$

The total change of $\mathbf{V}$ per generation may be obtained by multiplying the above equations (5), (10), (11), (14), and (15).

$$\mathbf{V}_{t+1} = (1-v)^4 \mathbf{K}\mathbf{K}'\mathbf{Q}\mathbf{R}\mathbf{V}_t + \mathbf{A}' + \mathbf{B} + \mathbf{F} \tag{16}$$

in which $t$ is the time measured in generations and it is assumed that all parameters $\alpha$, $n_g\alpha$, $\beta'$, $\xi$ and $4N_e v$ are much smaller than unity. At equilibrium, $\mathbf{V}_{t+1} = \mathbf{V}_t = \hat{\mathbf{V}}$ and the moments may be readily obtained by solving the following equation for $\hat{\mathbf{V}}$

$$\hat{\mathbf{V}} = (1-v)^4 \mathbf{G}\hat{\mathbf{V}} + \mathbf{A}' + \mathbf{B} + \mathbf{F}, \tag{17}$$

where

$$\mathbf{G} = \mathbf{K}\mathbf{K}'\mathbf{Q}\mathbf{R} \approx \begin{bmatrix} 1-2\alpha-2\alpha'(1+4n_g N_e v) - \dfrac{8N_e v\beta}{3} & \alpha+\alpha' & 0 \\ 0 & 1-5(\alpha+\alpha')-\xi & 2(\alpha+\alpha') \\ 2(\alpha+\alpha') & 2(\alpha+\alpha') & 1-3(\alpha+\alpha')-2\xi \end{bmatrix}.$$

https://doi.org/10.1017/S0016672300019790 Published online by Cambridge University Press

### 3. APPLICATION TO ANALYSES ON SEQUENCE VARIABILITY OF IMMUNOGLOBULINS

Immunoglobulin gene family is a most interesting example to apply the theory. Large amounts of sequence data are now available and are compiled by Kabat, Wu & Bilofsky (1979). It is known that the variable region and the constant region of light chain or heavy chain of immunoglobulins are separately encoded on the chromosome and they are joined together during differentiation (Seidman *et al.* 1978; Tonegawa *et al.* 1978). With the exception of mouse $\lambda$ chain so far studied, it is considered that a couple of hundred genes are present in each of the variable region gene families, whereas only a few genes are present in one family for constant region (Seidman *et al.* 1978). It is also known that the heavy chain, the $\kappa$ chain and the $\lambda$ chain genes are on different chromosomes and therefore they have evolved as different gene families from the common ancestor (e.g. Watson, 1976, p. 622). For a review on organization of immunoglobulin gene families, see Ohta (1980). In the

| | | | | | | |
|---|---|---|---|---|---|---|
| Asp | Ile | Val | Leu | Thr | Gln | ··· |
| Asp | Ile | Val | Leu | Thr | Gln | ··· |
| Asn | Ile | Val | Leu | Thr | Gln | ··· |
| Glu | Ile | Val | Leu | Thr | Gln | ··· |
| Asp | Val | Val | Met | Thr | Gln | ··· |
| Asp | Ile | Gln | Met | Thr | Gln | ··· |
| Asn | Ile | Val | Met | Thr | Gln | ··· |
| Asp | Val | Gln | Met | Ile | Gln | ··· |
| Asp | Ile | Val | Ile | Thr | Gln | ··· |
| Asp | Ile | Val | Met | Thr | Gln | ··· |

Fig. 4. Some examples of the amino-terminal sequences of the mouse kappa chain are shown. Ordinary three letter code is used. For this set of sequences, $C_1 = 22/45$. $C_2 = C_3 = 29/45$, $C_{12} = 12/45$ and so forth.

following, I shall apply the present theory to the statistical analyses on sequence variability of variable region of immunoglobulins.

By applying the theory to the actual data of immunoglobulins, let us consider the following comparison of amino acid sequences. If the two sequences compared have the identical amino acid at the $A$-th homologous site, we score one ($C_A = 1$). If the homologous amino acids are different, we score zero ($C_A = 0$). The second $B$-th site is scored in the same way. Next, if the two sequences have the identical amino acid at the $A$-th and the $B$-th sites, we say that the two sites are simultaneously identical and score $C_{AB} = 1$. Otherwise, $C_{AB} = 0$. All comparisons are made with available sequences and average scores are calculated for various groups of immunoglobulins. Figure 4 illustrates some examples of the amino-terminal sequences of the mouse kappa chain. As given under the Figure, $C_A$, $C_B$ and $C_{AB}$ become 22/45, 29/45 and 12/45 respectively for this set of sequences when $A = 1$ and $B = 2$.

Next, we further take averages of $C_A$, $C_B$ and $C_{AB}$ for all pairs of sites (all combinations of $A$ and $B$ over the sequence). The compilation of immunoglobulin sequences by Kabat *et al.* (1979) is used. Recently it is reported that the whole vari-

able region of the light chain actually consists of $V$ (variable) regions and $J$ (junction) segment and apparently the two regions constitute different gene families (Tonegawa *et al.* 1978; Seidman *et al.* 1978). Thus, $J$ segment should be excluded from the analyses which has only about 15 amino acid sites at the junction part with the constant region. By using Kabat *et al.*'s numbering system, the $V$ region is $0 \sim 94$th amino acid sites. The amino acid diversity patterns of homogeneous antibody molecules suggest that the variable regions of heavy chains are also encoded by $V$ and $J$ segments (Schilling *et al.* 1980), and that $V$ region is $0 \sim 95$th amino acid sites of Kabat *et al.*'s numbering system. Calculations of $C_A$, $C_B$ and $C_{AB}$ are performed by using the variable region sequences compiled by Kabat *et al.* (1979), excluding the J segments both for the light and the heavy chains. As in my previous analyses, immunoglobulin chains of which only a small part is sequenced are not used. Positions with deletions or ambiguous data (such as glutamine or glutamic acid) are omitted from the calculation.

Now, the expectation of $C_A$ or $C_B$ is the identity coefficient at each site.

$$E(C_A) = E(C_B) = \hat{C}_{w2} \approx \hat{C}_{w1}. \tag{19}$$

Here we shall define a measure of linkage disequilibrium, $\Delta$, as the excess probability of simultaneous identity over that expected from random combination of the identity at the two sites. This quantity is equivalent to the covariance of non-identity (heterozygosity) at the two loci within populations (Avery & Hill, 1979) and may be expressed by the following formula (20). Note that my previous formulation (Ohta, 1980) was erroneous as pointed out by Dr A. Robertson (personal communication). By letting $\underset{l \neq k}{E}$ denote to taking expectation for all pairs of $k$ and $l$ in the population as before,

$$\begin{aligned}
\Delta &= E(C_{AB}) - E(C_A)\,E(C_B) \\
&= \underset{l \neq k}{E} \{\sum_i \sum_j f_{ij,k} f_{ij,l}\} - \hat{C}_{w2}^2 \\
&= \underset{l \neq k}{E} \{\sum_i \sum_j (x_{i,k} y_{j,k} + D_{ij,k})(x_{i,l} y_{j,l} + D_{ij,l})\} - \hat{C}_{w2}^2 \\
&\approx E\{\sum_i x_{i,k}^2 \sum_j y_{j,k}^2 + 2\sum_i \sum_j x_{i,k} y_{j,k} D_{ij,k} + \sum_i \sum_j D_{ij,k}{}^2\} - \hat{C}_{w1}^2 \\
&= \tfrac{1}{2}(Y + Z),
\end{aligned} \tag{20}$$

under the assumption that the moments concerned take the same value whether $k$ and $l$ are different or not.

By using the theory in the previous section, it is possible to get a rough idea of the magnitude of $\Delta$. For this purpose, we use 'squared standard linkage deviation ($\sigma_d^2$)' which is defined as the ratio of $Z$ (third element of $\hat{\mathbf{V}}$) to twice of $X$ (Ohta & Kimura, 1969, 1970; Hill, 1975). From the equation (17), $\sigma_d^2$ is shown to become roughly

$$\sigma_d^2 = \frac{Z}{2X} \approx \frac{1}{3 + 4v' + 2\xi' - 4/(5 + 4v' + \xi')}, \tag{21}$$

where $v' = v/(\alpha + \alpha')$ and $\xi' = \xi/(\alpha + \alpha')$. On the other hand, the relationship

between $Z$ and $Y$ is, when the second element of the vector, $\mathbf{A}'+\mathbf{B}+\mathbf{F}$, is zero,

$$\{4v+5(\alpha+\alpha')+\xi\}Y = 2(\alpha+\alpha')Z.$$

From the above equation, we get,

$$Y = \frac{4(\alpha+\alpha')}{\{4v+5(\alpha+\alpha')+\xi\}}\sigma_d^2 X$$

since $\sigma_d^2 = Z/(2X)$. Hence $\Delta$ becomes a function of $\sigma_d^2$ of the following form,

$$\Delta = \tfrac{1}{2}(Y+Z) = \left\{1+\frac{2(\alpha+\alpha')}{4v+5(\alpha+\alpha')+\xi}\right\}\sigma_d^2 X$$

$$\approx \left\{1+\frac{2}{5+4v'+\xi'}\right\}\sigma_d^2(1-\hat{C}_{w1})^2, \tag{22}$$

where $v' = v/(\alpha+\alpha')$ and $\xi' = \xi/(\alpha+\alpha')$. When $\alpha+\alpha'$ is much larger than $v$ and $\xi$, $\sigma_d^2$ approaches $5/11$, and $\Delta$ becomes maximum, $\frac{14}{22}\times(1-\hat{C}_{w1})^2$. Roughly speaking, $\alpha+\alpha'$ is the rate of increase of homozygosity within a gene family by unequal crossing-over with $\alpha = m\gamma/n_g^2$ and $\alpha' = m'\gamma'/n_g^2$, in which intra- and inter-chromosomal unequal crossing-overs are assumed to occur with the rate $\gamma$ and $\gamma'$ and with the mean shift of $m$ and $m'$ genes, respectively. Also note that $v$ is the mutation rate per amino acid site and $\xi$ is the recombination rate between the $A$-th and the $B$-th sites by unequal crossing-over (Fig. 2).

By using a computer, I have calculated the average identity coefficients ($\bar{C}_A$ and $\bar{C}_B$) and the average simultaneous identity ($\bar{C}_{AB}$) for various groups of immunoglobulins. The results are shown in Table 1. In calculation, the first site ($A$) is moved from the first to the second last site (the $(n_s-1)$th), whereas the second ($B$) site is moved from the $(A+1)$th to the last one, therefore $\bar{C}_A$ and $\bar{C}_B$ are slightly different. The identity excess, $\Delta$, is calculated for each combination of $A$ and $B$ and average identity excess is obtained over all pairs of $A$ and $B$ of the $V$ region. The average identity excess, $\bar{\Delta}$, may be standardized by dividing with the product of non-identity probabilities at the $A$-th and the $B$-th sites, $(1-\bar{C}_A)(1-\bar{C}_B)$, and the standardized identity excess is also given in the table. From the equations (21) and (22), the standardized identity excess is approximately,

$$\Delta_s = \left\{1+\frac{2}{5+4v'+\xi'}\right\}\left\{\frac{1}{3+4v'+2\xi'-4/(5+4v'+\xi')}\right\}, \tag{23}$$

which becomes maximum, $14/22$, when $\alpha+\alpha'$ is much larger than $v$ and $\xi$. In my previous study (Ohta, 1978*c*), I have shown that the observed identity coefficients agree with the theoretical prediction of my model if $N_e = 10^4$, $n_g = 500$, $\beta = 10^{-3}$, $\alpha+\alpha' = 2\times10^{-7}$ and $v = 4\times10^{-9}$ for the framework region and $v = 8\times10^{-9}$ for the hypervariable region. Then $v' \ll 1$. If $\xi' \approx 2$ or $2(\alpha+\alpha') \approx \xi$ implying that twice the rate of increase of homozygosity within a family is about the same as the recombination rate, the standardized identity excess becomes about 0·20 which is roughly the same as the observed values for a group of immunoglobulin sequences corresponding to a gene family of a species. A more recent report (Weigert *et al.* 1978) suggests that the number of $V$ genes in one genome is a few hundred. Thus, let

us use a conservative estimate for $n_g$ and apply our theory. Let $N_e = 2{\cdot}5 \times 10^4$, $n_g = 100$, $\beta = 10^{-4}$, $v = 10^{-8}$ and $\alpha = \alpha' = 5 \times 10^{-8}$. Then $v' = 0{\cdot}1$. If $\xi' \approx 2$, the standardized identity excess becomes 0·17, which again agrees with the observed values.

On the whole, the identity excess of subgroups are smaller than that of the families, however the subgroup III of mouse heavy chain is exceptional and it should not be called a subgroup. From these results, it is expected that $\xi$ is at least

Table 1. *Observed average values of identity coefficients* ($\bar{C}_A$ *and* $\bar{C}_B$), *simultaneous identity* ($\bar{C}_{AB}$), *identity excess* ($\bar{\Delta}$) *and standardized identity excess,* $\bar{\Delta}/\{(1-\bar{C}_A)\times(1-\bar{C}_B)\}$, *are given for various groups of variable region of immunoglobulins*

| | $\bar{C}_A$ | $\bar{C}_B$ | $\bar{C}_{AB}$ | $\Delta$ | $\frac{\bar{\Delta}}{(1-\bar{C}_A)(1-\bar{C}_B)}$ | Number of sequences |
|---|---|---|---|---|---|---|
| Human $\kappa$ | 0·6585 | 0·6614 | 0·4445 | 0·0109 | 0·0943 | 31 |
| Mouse $\kappa$ | 0·6803 | 0·6609 | 0·4756 | 0·0246 | 0·2269 | 29 |
| Rabbit $\kappa$ | 0·7694 | 0·7253 | 0·5634 | 0·0053 | 0·0837 | 24 |
| Human $\lambda$ | 0·6209 | 0·6239 | 0·3966 | 0·0088 | 0·0617 | 18 |
| Human $H$ | 0·5990 | 0·5694 | 0·3603 | 0·0189 | 0·1095 | 24 |
| Mouse $H$ | 0·7272 | 0·6474 | 0·4986 | 0·0268 | 0·2786 | 19 |
| Rabbit $H$ | 0·7576 | 0·7676 | 0·5873 | 0·0047 | 0·0834 | 9 |
| Human $\kappa$ | | | | | | |
| Subgroup I | 0·8131 | 0·7476 | 0·6094 | 0·0010 | 0·0212 | 19 |
| Subgroup II | 0·8285 | 0·8272 | 0·6852 | 0·0013 | 0·0439 | 5 |
| Subgroup III | 0·8021 | 0·7932 | 0·6369 | 0·0020 | 0·0489 | 6 |
| Human $\lambda$ | | | | | | |
| Subgroup I | 0·7407 | 0·6775 | 0·5028 | 0·0008 | 0·0096 | 4 |
| Subgroup II | 0·8018 | 0·7799 | 0·6247 | 0·00004 | 0·0009 | 6 |
| Subgroup III | 0·7124 | 0·6918 | 0·4962 | 0·0045 | 0·0508 | 5 |
| Human $H$ | | | | | | |
| Subgroup II | 0·6140 | 0·5684 | 0·3606 | 0·0129 | 0·0774 | 5 |
| Subgroup III | 0·7271 | 0·6878 | 0·5022 | 0·0023 | 0·0270 | 17 |
| Mouse $H$ | | | | | | |
| Subgroup III | 0·7710 | 0·6872 | 0·5524 | 0·0215 | 0·3001 | 17 |

$\kappa$, $\lambda$ and $H$ denote the kappa, the lambda and the heavy chains respectively.

as large as $\alpha + \alpha'$ ($\approx 10^{-7}$) and that recombination takes place between genes of one subgroup. Thus, from the present result, I suggest that the subgroups compiled by Kabat *et al.* (1979) correspond to the gene family as considered here and that the whole group such as human $\kappa$ or mouse $\kappa$ consists of several subfamilies between which recombination would be limited. However, no evidence is available on the existence of many sub-subgroups as suggested by immunologists (Weigert & Riblet, 1977). In other words, recombination does not seem to be restricted to occur between the genes belonging to a very small sub-subgroup which contain only several gene members but appears to take place between any genes in one subgroup.

The analyses presented here are not exact in many respects. First, I did not take into account of the effect of sampling sequences from a large population such as

human and mouse species. Second, the difference in mutation rate among the sites was not considered in the analyses. The difference is particularly evident if the hypervariable and the framework regions are compared. As long as linkage disequilibrium is concerned, however, the difference of mutation rate would have a minor effect. Thirdly, genetic correlation with chromosomal distance within the family was not considered. At the moment, the arrangement of genes in a family is not known, however, the existence of subgroups suggests that each group is clustered on the chromosome so that genetic correlation would decrease with chromosomal distance. As Kimura & Ohta (1979) and Ohta (1980) show, it is possible to obtain theoretically the decrease of genetic correlation with chromosomal distance under various assumptions. When the number of genes of shift at unequal crossing-over is small, the correlation becomes significant, whereas when the mean number of genes of shift becomes more than 10% of the total size of the family, the correlation becomes insignificant. It is likely that each subfamily is clustered on the chromosome, however, it is still possible that, in one subfamily, genes are more or less randomly arranged on the chromosome. This is expected if subfamily size is 50 genes and if the mean gene number of shift is 5 at unequal crossing-over. It is also expected that, although rarely, the unequal crossing-over takes place between the genes of different subgroups. By chance, single subfamily may also differentiate into two subfamilies and thus the total family evolves. Important aspect is that this process is essentially random. Definition of sub-subfamilies and assignment of a fixed number of genes to each sub-subfamily in every genome of a species (e.g. Weigert & Riblet, 1977) seem to be highly arbitrary.

On the other hand unequal crossing-over and gene conversion are suggested as factors which generate antibody diversity during ontogeny (Seidman *et al.* 1978). This is because they create new association of amino acids in one sequence. They tend to decrease linkage disequilibrium during ontogeny but this effect would be quite small in quantitative analyses. If gene conversion also takes place in germ cells, it becomes essentially the same process as unequal crossing-over in a sense of expansion of certain gene copies with the expense of the others. At any rate, both unequal crossing-over and gene conversion tend to eliminate genetic diversity of the gene pool of a species if they continually occur in germ cells.

I thank Dr M. Kimura for his stimulating discussions and encouragements. I also thank Dr A. Robertson and Dr W. Hill for pointing out an error on the expectation of identity excess and for their other comments which greatly clarified the presentation.

## REFERENCES

AVERY, P. J. & HILL, W. G. (1979). Variance in quantitative traits due to linked dominant genes and variance in heterozygosity in small populations. *Genetics* **91**, 817–844.

BLACK, J. A. & GIBSON, D. (1974). Neutral evolution and immunoglobulin diversity. *Nature* **250**, 327–328.

HILL, W. G. (1975). Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theoretical Population Biology* **8**, 117–126.

HILL, W. G. & ROBERTSON, A. (1968). Linkage disequilibrium in finite populations. *Theoretical Applied Genetics* **38**, 226–231.

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.

KABAT, E. A., WU, T. T. & BILOFSKY, H. (1979). *Sequences of immunoglobulin chains*. U.S. Dept. of Health, Education, and Welfare, Public Health Service, National Institute of Health.

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

KIMURA, M. & OHTA, T. (1971). *Theoretical Aspects of Population Genetics*. Princeton University Press, Princeton.

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.

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

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

OHTA, T. (1978*b*). Theoretical population genetics of repeated genes forming a multigene family. *Genetics* **88**, 845–861.

OHTA, T. (1978*c*). Sequence variability of immunoglobulins considered from the standpoint of population genetics. *Proceedings of the National Academy of Sciences, U.S.A.* **75**, 5108–5112.

OHTA, T. (1979). An extension of a model for the evolution of multigene families by unequal crossing-over. *Genetics* **91**, 591–607.

OHTA, T. (1980). *Evolution and variation of multigene families*, Lecture Notes in Biomathematics, vol. 37, Springer, New York.

OHTA, T. & KIMURA, M. (1969). Linkage disequilibrium due to random genetic drift. *Genetical Research* **13**, 47–55.

OHTA, T. & KIMURA, M. (1971). Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics* **68**, 571–580.

PERELSON, A. S. & BELL, G. I. (1977). Mathematical models for the evolution of multigene families by unequal crossing-over. *Nature* **265**, 304–310.

SCHILLING, J., CLEVINGER, B., DAVIE, J. M. & HOOD, L. (1980). Amino acid sequence of homogeneous antibodies to dextran and DNA rearrangements in heavy chain V-region gene segments. *Nature* **283**, 35–40.

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*, vol. 38, pp. 507–513.

TONEGAWA, S., MAXAM, A. M., TIZARD, R., BERNARD, O. & GILBERT, W. (1978). Sequence of a mouse germline gene for a variable region of an immunoglobulin light chain. *Proceedings of the National Academy of Sciences, U.S.A.* **75**, 1485–1489.

WATSON, J. D. (1976). *Molecular Biology of the Gene*, third edition, W. A. Benjamin, Menlo Park, Calif., U.S.A.

WEIGERT, M., GATMAITAN, L., LOH, E., SCHILLING, J. & HOOD, L. (1978). Rearrangement of genetic information may produce immunoglobulin diversity. *Nature* **276**, 785–790.

WEIGERT, M. & RIBLET, R. (1977). Genetic control of antibody variable regions. *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 41, 'Origins of Lymphocyte Diversity', pp. 837–846.