Genet. Res., Camb. (1981), **38**, pp. 97–102
Printed in Great Britain

A mathematical study on the distribution of the number of repeated genes per chromosome

NAOYUKI TAKAHATA

National Institute of Genetics, Mishima, 411 *Japan*

(*Received* 6 *January* 1981 *and in revised form* 10 *February* 1981)

SUMMARY

I develop a mathematical model which can account for a distribution of the number of repeated genes per chromosome under the joint effects of sister chromatid exchange (SCE), inter-chromosomal crossing-over (ICC), and selection. The model can be applied not only to the cases of small gene clusters but also to multigene families. Based on this model, an appropriate mathematical formula is derived and used to obtain the equilibrium distribution. Assuming stabilizing selection and two simple schemes concerning SCE and ICC, I numerically calculate the equilibrium distribution and compare the result with observations on frequencies of single and triple α-haemoglobin genes in primates. It is also shown that if SCE and ICC occur according to the same probabilistic law, the distinction between them does not make much sense in the equilibrium distribution.

1. INTRODUCTION

Isozyme, restriction enzyme mapping and nucleotide sequencing techniques have been applied to a variety of organisms. One of the most striking findings revealed thereby may be the widespread occurrence of gene duplication, although the phenomenon has been well-known since the early report of Bridges (1919).

Recently, Harris (1980) has examined the extensive data available on human enzymes and concluded that at least 35% of the enzymes studied so far are coded by multiple loci. Gene organization of such 'multilocus enzymes' is still unclear, but it is likely that they have emerged by gene duplication. More conclusive instances of duplicated genes are the haemoglobin α and β gene clusters, located on the human 16 and 11 chromosomes, respectively (Weatherall & Clegg, 1979; Proudfoot *et al.* 1980). It is shown that at least 7 homologous genes, including pseudogenes, are involved in the β cluster (Fritsch, Lawn & Maniatis, 1980), and there are five in the α cluster (Lauer, Shen & Maniatis, 1980). Surprisingly, the genes in both clusters are arranged on the chromosome in the order of their expression during development. A similar pattern of gene organization of haemoglobin loci has been found in rabbit (Hardinson *et al.* 1979; Lacy *et al.* 1979) and mouse (Jahn *et al.* 1980; Proudfoot & Maniatis, 1980). Unlike the situation in mammals, however, adult α and β globin genes in *Xenopus laevis* are closely linked (Jeffreys *et al.* 1980).

To study a distribution of the number of repeated genes per chromosome, Krüger & Vogel (1975) have developed a model which incorporates unequal crossing-over between homologous chromosomes at meiosis. Their model is quite similar to the present one in this respect, but they have not considered the influence of sister chromatid exchange at

mitosis. In addition, they have assumed that selection operates on gametes. However, when selection acts on zygotes, their mathematical formulation to treat inter-chromosomal unequal crossing-over must be extended in terms of zygote frequencies, since the frequencies are not in Hardy-Weinberg proportion after selection. On the other hand, in order to understand the evolution of 'small multigene families' such as haemoglobin genes, Ohta (1981) has proposed two models to treat the continuous effect of unequal crossing-over on the amount of genetic variability maintained in small multigene families. Although the second model, referred to as the 'selection model' is more realistic than the 'cycle' model, the application of the former is limited to the case where only three types of chromosomes, having one, two or three repeated genes, exist in a population. Therefore, it is necessary to generalize the selection *model*. It is also an important problem in population genetics to quantitatively study how the number of genes per chromosome is regulated by selection under the continuous effect of unequal crossing-over, which presumably occurs during both mitosis and meiosis.

In this note, I present a mathematical formulation to treat this problem. A few numerical results are compared with the observations by Goossens *et al.* (1980) and Zimmer *et al.* (1980).

2. A MATHEMATICAL MODEL

Let us consider a random mating population of diploid organisms. For simplicity, I assume that the population is so large that we can ignore the effect of random genetic drift. If necessary, however, we can easily incorporate this effect into the following formulation. Let A_i be the chromosomes carrying i repeated genes, and x_i be their frequency within a population. Likewise, I denote the individuals with A_i and A_j chromosomes by A_iA_j and their frequency by y_{ij}. Thus, $x_i = \sum_{j=0}^{\infty} y_{ij}$. The number of repeated genes per chromosome increases or decreases primarily due to sister chromatid exchange (SCE) and inter-chromosomal crossing-over (ICC). Initially, I assume that SCE between A_i sister chromatids occurs at a rate β_i per generation while ICC between A_i and A_k chromosomes at meiosis occurs at a rate γ_{jk}. The subscripts of β and γ indicate that the rates may depend on the number of repeated genes on a chromosome. Once SCE and ICC occur, two new chromosomes emerge according to certain probability laws. Suppose that a new chromosome A_i is produced by SCE between A_j sister chromatids with probability $P_{i,2j}$, and by ICC between A_j and A_k chromosomes with probability $Q_{i,j+k}$. By definition, the number i cannot exceed $2j$ or $j+k$, i.e. the transition probabilities are concentrated on $[0,2j]$ or $[0,j+k]$. To show the situation explicitly, let us introduce the following function

$$H(i,n) = \begin{cases} 1 & \text{for} \quad 0 \leqslant i \leqslant n \\ 0 & \text{otherwise.} \end{cases}$$

Although I have assumed a random mating population, when selection is taken into account it is necessary to consider the zygote frequency of A_iA_j individuals at meiosis. In other words, we cannot expect that the frequency of A_iA_j is in Hardy-Weinberg proportion when ICC occurs. Let w_{ij} be the relative fitness of A_iA_j individuals. I also assume that selection acts on individuals before SCE occurs in a germ line. Then, we have

$$y'_{ij} = \frac{w_{ij}y_{ij}}{\bar{\bar{w}}} \tag{1}$$

after selection, where $\bar{\bar{w}} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} w_{ij}y_{ij}$, and $y_{ij} = x_ix_j$ holds at the stage of fertilization.

https://doi.org/10.1017/S0016672300020437 Published online by Cambridge University Press

Using the transition probability $P_{i,2j}$, SCE changes the zygote frequency y'_{ij} in (1) to

$$y''_{ij} = (1-\beta_i)(1-\beta_j)y'_{ij} + \sum_{k=0}^{\infty} P_{j,2k} H(j,2k)(1-\beta_i)\beta_k y'_{ik} + \sum_{k=0}^{\infty} P_{i,2k} H(i,2k)(1-\beta_j)\beta_k y'_{kj},$$

or approximately

$$y''_{ij} = \{1-(\beta_i+\beta_j)\}y'_{ij} + \sum_{k=0}^{\infty} \{P_{j,2k} H(j,2k)\beta_k y'_{ik} + P_{i,2k} H(i,2k)\beta_k y'_{kj}\}. \quad (2)$$

With these frequencies of A_iA_j, ICC occurs at meiosis and we get the frequency x'_i of A_i chromosomes in the next generation

$$x'_i = \sum_{j=0}^{\infty} (1-\gamma_{ij}) y''_{ij} + \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} Q_{i,j+k} H(i,j+k)\gamma_{jk} y''_{jk}. \quad (3)$$

The formulae (1) to (3) provide the basic equations to treat the present problem.

In the following discussion, I assume that selection is 'multiplicative', i.e. $w_{ij} = w_i w_j$, and that $w_i = 1-s(i-n_{op})^2$ for $|i-n_{op}| \leqslant 1/\sqrt{s}$ and otherwise 0, where s is the selection coefficient, and n_{op} is the optimum number of repeated genes on a chromosome (Krüger & Vogel, 1975). This type of stabilizing selection was first proposed by Kimura (1965) to study the maintenance of genetic variability in quantitative characters, although he assumed that $w_{ij} = 1-s(i+j-2n_{op})^2$ (see also Crow & Kimura, pages 295–296). Note that if selection is not multiplicative we must use (1) to (3).

Under the above assumptions, the change of frequency of A_i chromosomes $x_i = \sum_{j=0}^{\infty} y_{ij}$ per generation is approximately given by

$$\Delta x_i = -\frac{1}{\bar{w}}\left(\beta_i + \frac{1}{\bar{w}}\sum_{j=0}^{\infty}\gamma_{ij} w_j x_j\right) w_i x_i + \frac{1}{\bar{w}}\sum_{j=0}^{\infty} P_{i,2j} H(i,2j)\beta_j w_j x_j + \frac{1}{\bar{\bar{w}}}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} Q_{i,j+k} H(i,j+k)\gamma_{jk} w_j w_k x_j x_k + \frac{w_i-\bar{w}}{\bar{w}} x_i \quad (4)$$

neglecting the higher order terms of $\beta_i\beta_j$, $\beta_i\gamma_{ij}$ and so on. In the above equation, $\bar{\bar{w}} = \bar{w}^2$ and $\bar{w} = \sum_{i=0}^{\infty} w_i x_i$.

To demonstrate the continuous effect of SCE and ICC, I assumed here $\beta_i = \beta$ and $\gamma_{ij} = \gamma$ where β and γ are constants, and two types of transition probabilities. In the first model, the transition probabilities are

$$P_{i,2j} = \frac{1}{2j+1} \quad \text{and} \quad Q_{i,j+k} = \frac{1}{j+k+1}, \quad (5)$$

that is, that crossing-over takes place uniformly for all possible pairings between two chromosomes. In the other model, they are

$$P_{i,2j} = \frac{1}{j}\left\{1-\left|\frac{i}{j}-1\right|\right\} \quad \text{and} \quad Q_{i,j+k} = c\left\{1-\left|\frac{2i}{j+k}-1\right|\right\} \quad (6)$$

where $c = \frac{2}{j+k}$ for even $j+k$ and $\frac{2(j+k)}{(j+k)^2-1}$ for odd $j+k$. In the second model, crossing-over is most frequent at the centre of repeated gene clusters.

https://doi.org/10.1017/S0016672300020437 Published online by Cambridge University Press

3. RESULTS AND DISCUSSION

Based on the formulae (4) to (6), I numerically calculated the equilibrium distribution of the number of repeated genes per chromosome. Note that the equilibrium distribution can always exist under the present selection model because it is assumed that all individuals are lethal if the number of repeated genes on one of their two chromosomes is larger than $n_{op}+1/\sqrt{s}$, or smaller than $n_{op}-1/\sqrt{s}$. The results for small multigene families assuming $n_{op} = 2$ or 10 are presented in Table 1. The equilibrium distributions in both models are rather similar but the distribution in the first model is slightly broader than that in the second model, if the rates β and γ are the same (see the column headed

Table 1. *Equilibrium frequencies of chromosomes with repeated genes. Here it is assumed that* $\beta = 10^{-4}$ and $\gamma = 5\times10^{-4}$

(The mean and variance of the number of repeated genes per chromosome are indicated by M and V).

$n_{op}=2$	Model	$\hat{x}_0$	$\hat{x}_1$	$\hat{x}_2$	$\hat{x}_3$	$\hat{x}_4$	M	V
$S = 0{\cdot}01$	1	0·0029	0·0118	0·9709	0·0117	0·0028	2·0	0·0476
	2	0	0·0147	0·9710	0·0142	$6{\cdot}3\times10^{-5}$	2·0	0·0295
$S = 0{\cdot}1$	1	0·0002	0·0011	0·9975	0·0011	0·0002	2·0	0·0048
	2	0	0·0013	0·9973	0·0013	$3{\cdot}9\times10^{-7}$	2·0	0·0030
$n_{op}=10$	Model	$\hat{x}_7$	$\hat{x}_9$	$\hat{x}_{10}$	$\hat{x}_{11}$	$\hat{x}_{12}$	M	V
$S = 0{\cdot}001$	1	0·0071	0·0277	0·9138	0·0277	0·0071	10·0	0·5685
	2	0·0118	0·0505	0·8676	0·0501	0·0117	10·0	0·5312
$S = 0{\cdot}01$	1	0·0007	0·0028	0·9917	0·0028	0·0007	10·0	0·0543
	2	0·0012	0·0053	0·9855	0·0053	0·0012	10·0	0·0539
$S = 0{\cdot}1$	1	$4{\cdot}3\times10^{-5}$	$2{\cdot}6\times10^{-4}$	0·9994	$2{\cdot}6\times10^{-4}$	$4{\cdot}3\times10^{-5}$	10·0	0·0123
	2	$7{\cdot}2\times10^{-5}$	$4{\cdot}9\times10^{-4}$	0·9989	$4{\cdot}9\times10^{-4}$	$7{\cdot}2\times10^{-5}$	10·0	0·0072

V of Table 1). However, they are somewhat different from the Gaussian distribution with the same values of mean (M) and variance (V) of repeated genes per chromosome obtained from (4). The frequencies of chromosomes two or three steps apart from n_{op} are considerably higher than those expected from the Gaussian distribution. As noted in the previous section, the selection model used here is different from that of Kimura (1965) and Crow & Kimura (1970), but we can expect a Gaussian distribution even in the present model if we take their approach. Therefore, the discrepancy must come from the over-simplification in treating the process of crossing-over as a diffusion type, particularly when n_{op} is small.

In the case of $n_{op} = 2$, the equilibrium frequencies are given approximately by

$$\hat{x}_2 = 1-\frac{\beta+\gamma}{2s} \quad \text{and} \quad \hat{x}_1 = \hat{x}_3 = \frac{\beta+\gamma}{4s} \tag{7}$$

for both models, when $s \gg \beta+\gamma$ (see (7) in Ohta, 1981). Using (7) and the data on the frequency of triple α chromosomes in human populations observed by Goossens *et al.* (1980), I estimated the selection coefficient s as about 50 times larger than the crossing-over rate $\beta+\gamma$. If the crossing-over rate $\beta+\gamma$ is of the order of 10^{-3} or 10^{-4} (Ohta, 1981), the selection coefficient against single and triple α becomes 0·05 or 0·005. However the observations of the much higher frequency of the single α globin gene compared with the triple α gene in human populations (Goossens *et al.* 1980), and a high frequency of the triple α gene in chimpanzee (Zimmer *et al.* 1980), suggest that the selection coefficient s may be asymmetic to the dosage. Another possible explanation may be that random

genetic drift, neglected here, is responsible for high frequencies of such chromosomes in these populations.

In a recent lucid paper, Schimke (1980) has suggested that SCE in mammalian cells occurs at a considerable rate, and has shown that we can easily select cultured cells with more repeated genes when we artificially shift the optimum number of repeated genes per chromosome towards a larger value. If crossing-over spontaneously occurs at a rate as high as 10^{-4}, and the selection coefficient s against less fit chromosomes is about 10^{-2}, as predicted in the case of haemoglobin genes, the fittest chromosomes with duplicated genes must have spread in a population very rapidly during the course of evolution.

In the above analyses, I tacitly assumed that the pattern of SCE is essentially the same as that of ICC. In other words, the transition probability of SCE is assumed to be the same function as that of ICC in each model (see (5) and (6)). To evaluate relative effects of SCE and ICC on the equilibrium distribution, I calculated some other cases in which each value of β and γ is changed but the total rate $\beta+\gamma$ is kept constant. So long as $\beta+\gamma$ remains unchanged, I found no significant difference in the patterns of the equilibrium distribution. This in turn indicates that it is at least theoretically unnecessary to distinguish SCE from ICC if they do occur in the same way. However, this assumption may not be warranted in reality, and the distinction between SCE and ICC would become important. As there is no evidence for models (5) and (6), they might be unrealistic, but formula (4) will be still appropriate even when we know more precisely the mechanisms of SCE and ICC. In particular, β_i and γ_{ij} might depend heavily on the repeated number per chromosome. At any rate, the present formulation is potentially useful for studying the distribution of the number of repeated genes per chromosome under the joint effects of SCE, ICC and selection.

I would like to thank Dr T. Ohta for her stimulating discussions. I am much indebted to an anonymous referee for his many useful comments in improving the manuscript. This is Contribution no. 1361 from the National Institute of Genetics, Mishima, Japan.

REFERENCES

Bridges, C. B. (1919). Duplication. *Anatomical Record* **15**, 357–358.

Crow, J. F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York, London: Harper and Row.

Fritsch, E. F., Lawn, R. M. & Maniatis, T. (1980). Molecular cloning and characterization of the human β-like globin gène cluster. *Cell* **19**, 959–972.

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 Science* **77**, 518–521.

Hardinson, R. C., Butler, III, E. T., Lacy, E. & Maniatis, T. (1979). The structure and transcription of four linked rabbit β-like globin genes. *Cell* **18**, 1285–1297.

Harris, H. (1980). Multilocus enzymes in man. CIBA foundation Symposium 66 (new series), 187–204.

Jahn, C. L., Hutchinson, III, C. A., Phillips, S. J., Weaver, S., Haigwood, N. L., Volivia, C. F. & Edgell, M. H. (1980). DNA sequence organization of the β-globin complex in the BALB/c Mouse. *Cell* **21**, 159–168.

Jeffreys, A. J., Wilson, V., Wood, D. & Simons, J. P. (1980). Linkage of adult α- and β-globin genes in *X. laevis* and gene duplication by tetraploidization. *Cell* **21**, 555–564.

Kimura, M. (1965). A stochastic model concerning the maintenance of genetic variability in quantitative characters. *Proceedings of the National Academy of Science* **54**, 731–736.

Krüger, J. & Vogel, F. (1975). Population genetics of unequal crossing over. *Journal of Molecular Evolution* **4**, 201–247.

Lacy, E., Hardison, R. C., Quon, D. & Maniatis, T. (1979). The linkage arrangement of four rabbit β-like globin genes. *Cell* **18**, 1273–1283.

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

OHTA, T. (1981). Genetic variation in small multigene families. *Genetical Research*. (In the Press.)

PROUDFOOT, N. J. & MANIATIS, T. (1980). The nucleotide sequence of a rabbit β-globin pseudogene. *Cell* **21**, 545–553.

PROUDFOOT, N. J., SHANDER, M. H. M., MANLEY, J. L., GEFTER, M. L. & MANIATIS, T. (1980). Structure and in vitro transcription of human globin genes. *Science* **209**, 1329–1336.

SCHIMKE, R. T. (1980). Gene amplification and drug resistance. *Scientific American* **243**, 50–59.

WEATHERALL, D. J. & CLEGG, J. B. (1979). Recent developments in the molecular genetics of human hemoglobin. *Cell* **16**, 467–479.

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 Science, U.S.A.* **77**, 2158–2162