ROCKY MOUNTAIN
JOURNAL OF MATHEMATICS
Volume 9, Number 1, Winter 1979

STOCHASTIC MODELS FOR THE EVOLUTION OF MULTIGENE FAMILIES

ALAN S. PERELSON*

1. **Introduction.** Modern efforts to determine the organization and structure of the eukaryotic chromosome have shown that there are two general organizational classes of DNA: repetitive and non-repetitive. In animals the fraction of DNA in repetitive sequences typically lies between 0.25 and 0.45 [1]. For example, in man the *Y* chromosome is about 50% repetitive DNA [2]. Satellite DNA from the fruit fly *Drosophilia virilis* contains about 10^7 copies of a single heptanucleotide repeat known as satellite sequence I, as well as 10^6 copies of satellite sequences II and III, heptanucleotide repeats which could be derived from sequence I by single base substitutions [3]. In the frog *Xenopus laevis*, as well as many other species, there are units, repeated several hundred times [3], which contain the genes coding for 18S and 28S ribosomal RNA separated by spacer regions of unknown function [4–6]. An unusual feature of these units is that among different species within the genus *Xenopus* the 18S and 28S ribosomal RNA genes are identical, but the spacer regions may differ both in length and nucleotide sequence [5, 7–8].

The name *multigene family* has been proposed to describe those sets of nucleotide sequences or genes that show multiplicity, close linkage, sequence homology, and have similar or overlapping functions [3]. Besides satellite DNA and the genes coding for 18S and 28S ribosomal RNA many other genes in eukaryotes are arranged in multigene families. These include the genes coding for 5S ribosomal RNA, transfer RNA, histones, hemoglobin, and antibody [3].

The existence of multigene families raises some intriguing biological questions. Genetic mechanisms must be found which can not only generate repetitive gene sequences, but which also explain the novel evolutionary features of multigene families. That is, how can the genes within a multigene family be maintained virtually identical within a species, and yet show divergence between species, as is found with the spacer portions of the ribosomal RNA genes and the species-specific residues found in the conserved portion of the antibody V gene? This feature of multigene families, called *coincidental evolution*, is difficult to explain with classical population genetics theory. For example, in a family of

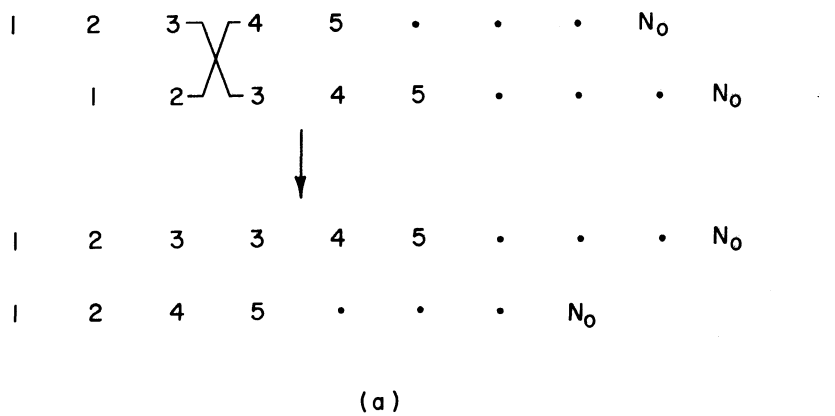
Received by the editors on December 1, 1977.

*This work was performed under the auspices of the U.S. Energy Research and Development Administration.

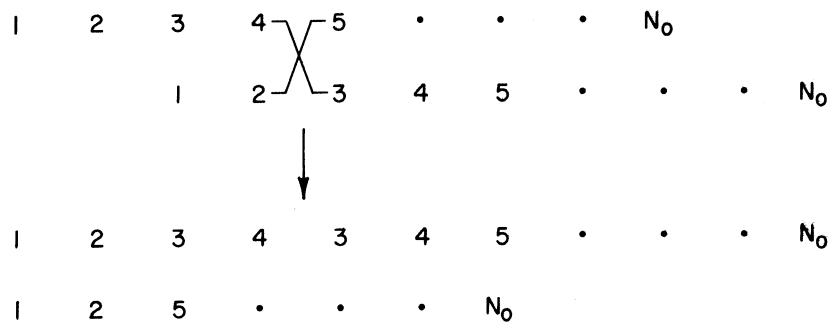
450 ribosomal RNA genes how could selection operate to eliminate a mutant nonfunctional gene when 449 functional genes were still present? One additional novel evolutionary feature of multigene families is that the family size expands and contracts rapidly with respect to evolutionary time. Different but closely related species of kangaroo rats, for example, contain multigene families which constitute from very little to 50% of their genomic DNA [3]. Also, size heterogeneity of multigene families is observed among the individual members of a population [3].

One genetic mechanism that can explain the generation and evolution of multigene families, and which has attracted both biological and mathematical attention is *homologous but unequal crossing-over* [3, 9–14]. During intrachromosomal unequal crossing-over, sister chromatids carrying closely linked homologous genes mispair, crossover, and produce two new sister chromatids, one containing a greater number and the other a lesser number of gene repeats (Fig. 1). When mispairing is by one repeat unit a single gene is duplicated in one chromatid and eliminated from the other chromatid. A mispairing by k units leads to a duplication and elimination of k tandem repeats. The effects of many unequal cross-overs on a multigene family have been studied by Smith [9, 11] and Black and Gibson [10] by Monte-Carlo simulations. Although such studies are useful they can only deal with limited numbers of genes. Here I shall present some analytical models for the evolution of multigene families developed by Ohta [12] and by me in collaboration with Bell [13]. Since unequal crossing-over requires the presence of at least two repeats, other mechanisms must be invoked to explain the initial duplication of a gene.

2. **Birth-Death Models.** I shall first analyze the effects of unequal crossing-over on a multigene family when the mispairing is by a single repeat. Assume that at $t = 0$ there are $N_0 \geqq 2$ adjacent homologous DNA sequences or "genes" which represent a multigene family. These sequences need not be identical, only similar enough to allow mispairing. For simplicity, I shall only consider the case in which all of the N_0 repeats at $t = 0$ are distinct. Pick one particular repeat and let $P_n(t)$ be the probability that there are n copies of this repeat at time t. In the time interval $[t, t + \delta t]$ one of three things can happen: The number of copies of the repeat can increase by 1, decrease by 1, or remain the same. If there are n copies of the repeat at time t, the probabilities of these events are $\lambda_n(t)\delta t + o(\delta t)$, $\mu_n(t)\delta t + o(\delta t)$, and $1 - \lambda_n(t)\delta t - \mu_n(t)\delta t + o(\delta t)$, respectively. Hence



(a)



(b)

Figure 1. Unequal crossing-over between sister chromatids in a region containing N_0 tandem homologous genes and the products of the events. a, Mispairing by one gene; b, mispairing by two genes.

$$
\begin{aligned}
P_n(t+\delta t) &= [1-\lambda_n(t)\delta t-\mu_n(t)\delta t]P_n(t) \\
&\quad +\lambda_{n-1}(t)\delta t\,P_{n-1}(t) \\
&\quad +\mu_{n+1}(t)\delta t\,P_{n+1}(t)+o(\delta t), \qquad n\geqq 1, \\
P_0(t+\delta t) &= [1-\lambda_0(t)\delta t]P_0(t) \\
&\quad +\mu_1(t)\delta t\,P_1(t)+o(\delta t).
\end{aligned} \tag{2.1}
$$

In the limit $\delta t \to 0$ one obtains the following system of differential equations:

$$\begin{aligned} \frac{dP_n}{dt} &= -[\lambda_n(t) + \mu_n(t)]P_n(t) + \lambda_{n-1}(t)P_{n-1}(t) \\ &\quad + \mu_{n+1}(t)P_{n+1}(t), \qquad n \geqq 1 \\ \frac{dP_0}{dt} &= -\lambda_0(t)P_0(t) + \mu_1(t)P_1(t), \end{aligned} \tag{2.2}$$

with initial conditions

$$P_n(0) = \begin{cases} 1 & \text{for } n = 1 \\ 0 & \text{for } n \neq 1. \end{cases} \tag{2.3}$$

This description of unequal crossing-over is a birth and death process with $\lambda_n(t)$ and $\mu_n(t)$ the birth and death coefficients, respectively. In what follows, I shall consider two separate and biologically reasonable methods of choosing these coefficients.

2.2. If there is an equal *a priori* probability of each repeat in a multigene family being influenced by crossover, i.e., duplicated or eliminated, then the probability that a gene with multiplicity n is influenced is $n/N(t)$, where $N(t)$ is the total number of genes in the family at time t. If we let p and $1 - p$ be the respective probabilities that a gene influenced by crossover is duplicated and eliminated, and if we measure time by the crossover rate so one crossover equals one unit of time, then

$$\lambda_n(t) = p \frac{n}{N(t)}, \tag{2.4}$$

$$\mu_n(t) = (1 - p) \frac{n}{N(t)}. \tag{2.5}$$

The difficulty with this approach is that $N(t)$ is a random variable. At each crossover the total family length increases by one with probability p and decreases by one with probability $1 - p$. Thus $N(t)$ varies as a one-dimensional random walk. To proceed we shall replace $N(t)$ by its mean value, $\bar{N}(t)$. Thus we are approximating the true birth and death coefficients. In section (2.6) I shall comment on a more precise method of solution employing a bivariate Markov process.

One can show quite simply that

$$\frac{d\bar{N}(t)}{dt} = (2p - 1)[1 - \pi_{0,N_0}(t)] \tag{2.6}$$

where $\pi_{0,N_0}(t)$ is the probability of the family going extinct [i.e., $N(t) = 0$] given $N(0) = N_0$. Thus for $p \geqq 1/2$, $d\bar{N}/dt \geqq 0$ whereas for $p \leqq 1/2$, $d\bar{N}/dt \leqq 0$.

Restricting our attention to the case $p = 1/2$, so that $\bar{N}(t) = N_0$, Eqs. (2.4) and (2.5) become

$$\lambda_n(t) = \mu_n(t) = \frac{n}{2N_0}. \tag{2.7}$$

For these values of the birth and death coefficients, Eq. (2.2) with initial conditions (2.3) can be solved [15] yielding

$$P_n(t) = \frac{(t/2N_0)^{n-1}}{(1 + t/2N_0)^{n+1}}, \quad n \geqq 1, \tag{2.8}$$

$$P_0(t) = \frac{t}{t + 2N_0}. \tag{2.9}$$

The mean value of n, $\bar{n}$, can be shown to be

$$\bar{n} = 1 \tag{2.10}$$

which is simply $\bar{N}(t)/N_0$, as one would expect, and the variance, σ^2, is a linearly increasing function of time given by

$$\sigma^2 = \frac{t}{N_0}. \tag{2.11}$$

2.3. If the number of copies of a gene ever reaches 0 then that gene is lost forever. From (2.9) one sees that the probability of ultimate extinction

$$\lim_{t \to \infty} P_0(t) = 1,$$

Thus if $p = 1/2$ all genes will ultimately be lost. However, since evolution has only had a finite time to operate, this asymptotic result should not be disturbing. Additionally, in reality, the loss of the whole multigene family would be highly selected against.

2.4. One can easily imagine circumstances in which initial conditions other than (2.3) might apply. If one initially has $n_0 > 1$ identical copies of the particular repeat we are following, then the probability of extinction, P_0, is obtained by raising (2.9) to the (n_0)th power and the probability of ultimate extinction is still unity. The expressions for the mean and variance are obtained by multiplying Eqs. (2.10) and (2.11) by n_0. These results follow from the fact that the n_0 repeats evolve independently.

2.5. Initially there were N_0 distinct genes in the multigene family. Let us call each of these distinct genes a gene type. Above, I examined the effects of unequal crossing-over on a single gene type and showed that any particular gene type may disappear from the family. If the total number of gene types in the family ever reaches one, then the remaining gene (or more precisely, the remaining gene type) is said to have become *fixed.* The multigene family would then be homogeneous and contain on the average $\bar{N}(t)$ copies of the fixed gene. Gene fixation is therefore a mechanism for explaining the coincidental evolution seen in multigene families. Furthermore, even after fixation occurs $N(t)$ will vary, explaining the observed expansion and contraction of multigene families.

When $p = 1/2$ gene fixation is certain to occur since all genes will eventually be lost. The mean time to fixation can be computed as follows. Let $p_{i,N_0}(t)$ be the probability of having i gene types in the multigene family at time t, given that there were N_0 gene types at $t = 0$. Then

$$p_{0,N_0} = P_0^{N_0}, \tag{2.12}$$

$$p_{1,N_0} = N_0(1 - P_0)P_0^{N_0-1}. \tag{2.13}$$

A gene becomes fixed when a multigene family with two gene types loses all the genes of one type. A family containing one gene type can become a family with zero gene types, i.e., go extinct. Since $dp_{1,N_0}/dt$ measures the net rates of these competing processes, one can easily see that the rate of gene fixation, dp_f/dt, is given by

$$\frac{dp_f}{dt} = \frac{dp_{1,N_0}}{dt} + \frac{dp_{0,N_0}}{dt} \tag{2.14}$$

and that the cumulative probability of fixation by time t, $p_f(t)$, is

$$p_f(t) = \int_0^t \frac{dp_f}{d\tau}\, d\tau = p_{1,N_0} + p_{0,N_0}. \tag{2.15}$$

Substituting (2.9) into (2.12) and (2.13) and differentiating gives

$$\frac{dp_f}{dt} = \frac{4N_0^3(N_0 - 1)t^{N_0-2}}{(t + 2N_0)^{N_0+1}} \tag{2.16}$$

and

$$p_f(t) = \frac{(2N_0^2 + t)t^{N_0-1}}{(t + 2N_0)^{N_0}}. \tag{2.17}$$

Thus the probability of ultimate fixation

$$\lim_{t\to\infty} p_f(t) = 1 \tag{2.18}$$

and the mean time to ultimate fixation determined by this birth death process, T_{bd}, is given by

$$T_{bd} = \int_0^\infty t \frac{dp_f}{d\tau} d\tau. \tag{2.19}$$

One can easily show

$$\int \frac{t^{N_0-1}}{(t + 2N_0)^{N_0+1}} d\tau = \frac{t^{N_0}}{2N_0{}^2(t + 2N_0)^{n_0}} \tag{2.20}$$

and hence

$$T_{bd} = 2N_0(N_0 - 1). \tag{2.21}$$

2.6. The results obtained so far were based upon an approximation in which the birth and death coefficients were obtained by replacing $N(t)$ in (2.4) and (2.5) by its mean valuue $\overline{N}(t)$. A precise model of unequal crossing-over can be constructed by viewing both $n(t)$ and $N(t)$ as random variables and computing the probability, $P(n, N, t)$, that at time t there are n copies of a particular repeat and a total of N repeats in the whole multigene family. Figure 2 illustrates the possible transitions from state (n, N) by unequal crossing with unit mispairing. If one assumes that there is an equal *a priori* chance of a crossover influencing each repeat in the family then the transition probabilities shown in Figure 2 have the values:

$$\text{(2.22a)} \quad \lambda_{n,N} = pn/N \quad ; \quad \lambda_{n,N}: (n, N) \to (n + 1, N + 1),$$

$$\text{(2.22b)} \quad \gamma_{n,N} = p(1 - n/N) \quad ; \quad \gamma_{n,N}: (n, N) \to (n, N + 1),$$

$$\text{(2.22c)} \quad \mu_{n,N} = (1 - p)n/N \quad ; \quad \mu_{n,N}: (n, N) \to (n - 1, N - 1),$$

$$\text{(2.22d)} \quad \omega_{n,N} = (1 - p)(1 - n/N) \quad ; \quad \omega_{n,N}: (n, N) \to (n, N - 1),$$

and

$$\lambda_{n,N} + \gamma_{n,N} + \mu_{n,N} + \omega_{n,N} = 1. \tag{2.23}$$

Although one can easily construct the forward and backward equations for this process, they do not appear to be easily soluble because of the factor N in the denominator of the transition probabilities.

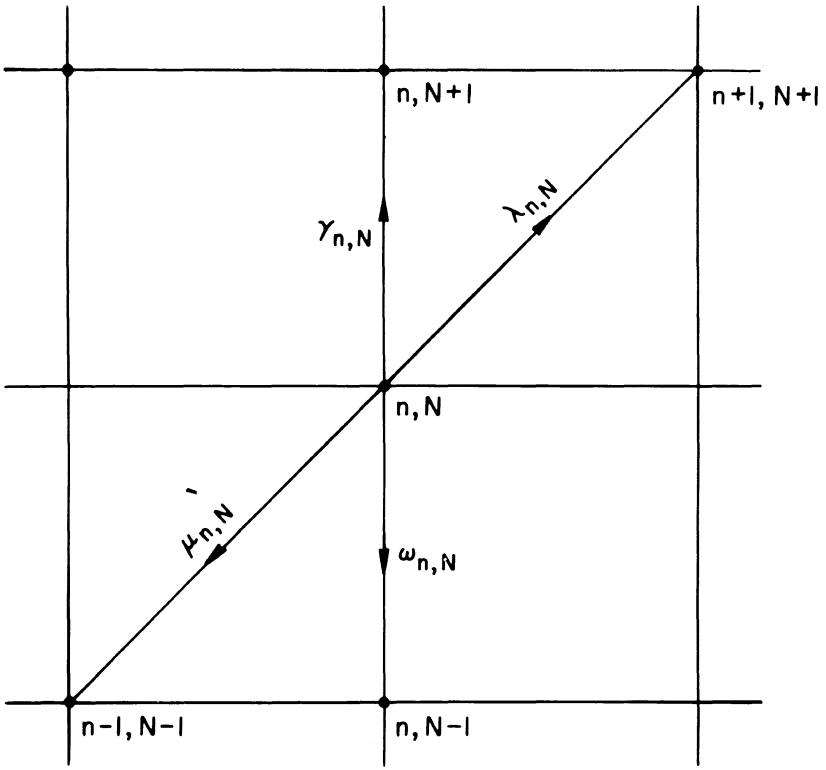


Figure 2. Possible transitions and their rates from the state (n, N) by unequal crossing-over with unit mispairing.

2.7. An alternative approach to assessing the validity of replacing $N(t)$ by its mean is to compare the results of Monte Carlo simulations with those predicted by the linear birth-death process. Black and Gibson [10] performed such a simulation for the case of unit mispairing with $p = 1/2$. In Figure 3 I have superimposed Black and Gibson's results for the percentage of original gene types that remain in the family after a given number of crossovers, with plots of $1 - P_0(t)$ as obtained from Eq. (2.9), for different values of N_0. The Monte-Carlo results consistently appear below the computed curves, with the disagreement being least at high values of N_0. The disagreement at small values of N_0 may in part be due to poor statistics in the Monte-Carlo simulations which start with few repeats.

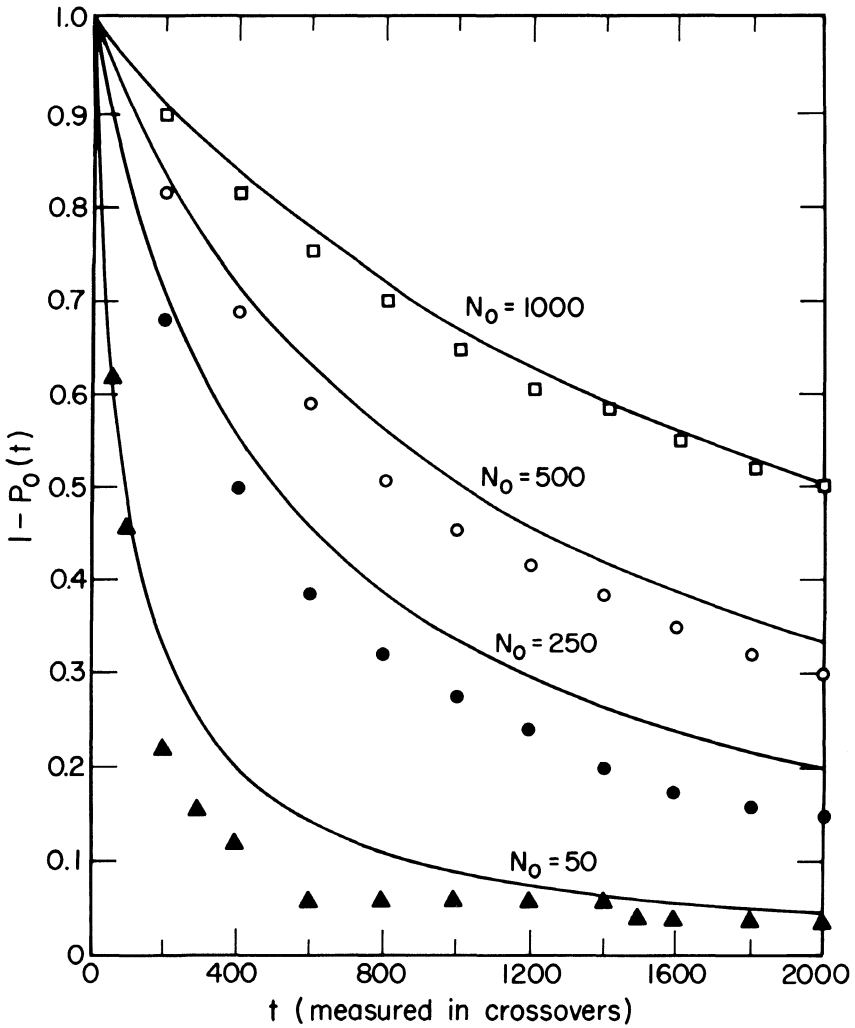


Figure 3. Comparison of the predictions of the birth-death (solid lines) with the Monte-Carlo simulations of Black and Gibson [10] (data points) for $p = 1/2$. Plotted is the expected fraction of remaining genes after various numbers of crossovers for different size families. ▲, $N_0 = 50$; ●, $N_0 = 250$; ○, $N_0 = 500$; ■, $N_0 = 1000$. Reproduced with permission from [13].

2.8. There is a second, biologically reasonably way to view the unequal crossing process which rigorously leads to linear birth and death coefficients. In the model discussed above, time was measured in units

of crossover events. If there were a constant rate of unequal crossing-over within a multigene family this approach would be appropriate. However, there is some biological evidence [16] which suggests that the rate of crossing-over is proportional to the size of the multigene family. If this is in fact the case, then it is reasonable to suppose that the number of crossovers per unit time which influence a particular gene type is proportional to $n(t)$, the number of copies of that gene at time t, so that

$$\begin{aligned} \lambda_n(t) &= kpn(t) \\ \mu_n(t) &= k(1-p)n(t) \end{aligned} \tag{2.24}$$

where k is the crossover rate per gene copy. For $p = 1/2$ this model reduces to the previous one if k^{-1} is replaced by N_0, however it has the advantage of being exactly soluble for all values of p. Using the classical results for linear birth death processes [15] one finds

$$P_0(t) = \frac{(1-p)[1-e^{k(2p-1)t}]}{1-p-pe^{k(2p-1)t}}, \tag{2.25}$$

$$P_n(t) = \frac{p^{n-1}(1-2p)^2 e^{k(2p-1)t}[1-e^{k(2p-1(t]^{n-1}}}{[1-p-pe^{k(2p-1)t}]^{n+1}}, \tag{2.26}$$

The mean and variance of n are

$$\bar{n} = e^{k(2p-1)t} \tag{2.27}$$

$$\sigma^2 = \frac{e^{k(2p-1)t}[e^{k(2p-1)t}-1]}{2p-1}, \tag{2.28}$$

For $p = 1/2$ these equations reduce to

$$P_0(t) = \frac{kt/2}{1+kt/2}, \tag{2.29}$$

$$P_n(t) = \frac{[kt/2]^{n-1}}{[1+kt/2]^{n+1}}, \tag{2.30}$$

$$\bar{n} = 1, \tag{2.31}$$

$$\sigma^2 = kt. \tag{2.32}$$

The probability of ultimate extinction is therefore

$$\lim_{t\to\infty} P_0(t) = \begin{cases} 1 & \text{if } p \leqq 1/2 \\ \dfrac{1-p}{p} & \text{if } p > 1/2. \end{cases} \tag{2.33}$$

The probability of fixation, ettermined by using (2.12), (2.13), (2.15), and (2.25), is

$$p_f(t) = \frac{(1-p)^{N_0-1}[1-e^{k(2p-1)t}]^{N_0-1}\{N_0(1-2p)e^{k(2p-1)t} + (1-p)(1-e^{k(2p-1)t}\}}{[1-p-pe^{k(2p-1)t}]^{N_0}}, \tag{2.34}$$

which simplifies when $p = 1/2$ to

$$p_f(t) = \frac{(kt/2)^{N_0-1}(N_0 + kt/2)}{(1+kt/2)^{N_0}}. \tag{2.35}$$

The probability of ultimate fixation

$$\lim_{t\to\infty} p_f(t) = \begin{cases} 1 & \text{if } p \leqq 1/2 \\ \dfrac{(1-p)^{N_0-1}[(1-p) + N_0(2p-1)]}{p^{N_0}} & \text{if } p > 1/2. \end{cases} \tag{2.36}$$

Thus when $p > 1/2$ fixation is not certain.

The mean time to ultimate fixation can be computed as in section 2.5. When $p = 1/2$ this computation is simple and leads to $2(N_0 - 1)/k$.

3. **A Gambler's Ruin Model.** If one assumes that there exists some biological mechanism for maintaining the length of a multigene family constant, say at $N(t) = N_0$, then one can interpret the unequal crossing-over problem as a classical gambler's ruin problem [15]. Again I shall follow one particular gene type, so that at $t = 0$, $n = 1$. Each time a crossover influences a gene of this type, n either increases or decreases by 1. If n reaches N_0 then all other gene types have necessarily been eliminated and n has been fixed. Thus considered as a game, one gene type plays against all others which constitute the "bank." Initially the bank has $N_0 - 1$ genes and the player 1 gene. At each unit of play, the player wins or loses 1 gene with probabilities p and $q = 1 - p$, respectively. Since a crossover can affect any gene in the family, plays do not occur at a constant rate. From well known results on the gambler's ruin problem one can conclude that the probability of gene fixation is $1/N_0$ when $p = 1/2$, and is

$$\left(\frac{q}{p} - 1\right) \Big/ \left[\left(\frac{q}{p}\right)^{N_0} - 1\right]$$

when $p \neq 1/2$ [15]. Furthermore, given that fixation occurs, the mean time of this event can be computed using results on conditional random

walks developed by Beyer and Waterman [17]. If we again measure time in units of crossover events, then one notices that the probability of a step being taken at a crossover is n/N_0. Including the non-uniform rate of play into the gambler's ruin, one finds that the mean time to gene fixation determined by a gambler's ruin process, T_{gr}, is given by

$$T_{gr} = \begin{cases} N_0(N_0 - 1) & p = 1/2 \\ \dfrac{N_0}{(p^{N_0} - q^{N_0})(p - q)} & \\ \cdot \displaystyle\sum_{i=1}^{N_0} \frac{(p^i - q^i)p^{N_0 - i} - q^{N_0 - i}}{i} & p \neq 1/2. \end{cases} \tag{3.1}$$

For $p = 1/2$ this differs from the previous estimate, T_{bd}, by a factor of 2. It is smaller because in this model the game definitely stops if n reaches N_0. In other models n could exceed N_0.

4. **Diffusion Equation Method.** Ohta [12] has used the diffusion equation method of population genetics [18] to compute the fixation time in a unit mispairing problem. Here I shall summarize her work and show how it can be generalized for the case of unequal crossing-over with multiple mispairing.

4.1 Let us assume N_0 is large so that it is appropriate to deal with continuous changes in gene frequency. Again we shall follow one particular gene type and denote its frequency by x. In order to compute the mean time for gene fixation one can utilize some classical results of Kimura and Ohta [19] based upon the use of the Kolmogorov backward equation

$$\frac{\partial p_{xi}}{\partial t} = \frac{V_{\Delta x}}{2} \frac{\partial^2 p_{xi}}{\partial i^2} + M_{\Delta x} \frac{\partial p_{xi}}{\partial i} \tag{4.1}$$

where $p_{xi}(t)$ is the probability that the gene frequency is x at time t, given the gene frequency was i at time 0. The change in the mean and variance of the gene frequency per unit time are $M_{\Delta x}$ and $V_{\Delta x}$, respectively. For gene fixation one considers the case in which $x = 1$ and determines $u_i(t) = p_{1i}$, the probability that the gene becomes fixed by time t, given that its frequency is i initially. In order to proceed Ohta assumed that mispairing occurs by one repeat, so that one gene is either duplicated or eliminated per crossover. Further, she assumed that crossovers occur at random between the N_0 repeats of the family, $p = 1/2$, and that duplication and deletion occur alternately in a process she called a "cycle." Measuring time in cycles, the family size remains constant, i.e., $N(t) = N_0$ for all t.

Ohta [12] then showed that the mean and variance of the frequency change per cycle are given by

$$\begin{aligned} M_{\Delta x} &= 0, \\ V_{\Delta x} &= 2x(1 - x)/n_0^2, \end{aligned} \tag{4.2}$$

and the mean fixation time, measured in cycles, is

$$T_{\text{dif}} = -N_0^2(N_0 - 1)\ln(1 - 1/N_0). \tag{4.3}$$

Recognizing that each cycle is two crossovers, one can compare this diffusion equation result with the fixation time estimate derived by the birth-death process. Doing so one finds that with time measured in crossovers

$$\begin{aligned} T_{\text{dif}} &= -2N_0^2(N_0 - 1)\ln(1 - 1/N_0) \\ &= -N_0 \ln(1 - 1/N_0)T_{bd} \\ &\sim \left(1 + \frac{1}{2N_0} + \frac{1}{3N_0^2} + \cdots\right) T_{bd} \end{aligned} \tag{4.4}$$

where T_{bd} is given by (2.21). The two estimates converge to the same result as N_0 becomes large. For small values of N_0, T_{bd} is probabily a more accurate estimate since it is derived without assuming that the repeat frequencies are continuous variables and that unequal crossovers occur in cycles.

4.2. Now let us consider the case of unequal crossovers with multiple and random mispairing. Recall from Figure 1 that a mispairing by k repeats leads to a duplication of k repeats on one chromatid and an elimination of k repeats from the sister chromatid. Assuming a viable offspring has equal probability of obtaining the expanded or contracted chromosome, then $p = 1/2$. Following Ohta I shall make the more stringent assumption that unequal crossovers occur in cycles composed of a duplication of m repeats followed by a deletion of m repeats, where the mispairing per cycle, m, is a random variable with mean $\overline{m}$. If one further assumes that the repeats in the multigene family are distributed randomly along the chromosome and the position of the crossover is random, then the probability that the gene frequency of a particular gene type increases from x to $x + \xi_1$, $P_{x \to x+\xi_1}$, during a duplication event in which the mispairing is by m is

$$P_{x \to x+\xi_1} = \binom{m}{\xi_1 N_0} x^{\xi_1 N_0}(1 - x)^{m - \xi_1 N_0}, \quad 0 \leqq \xi_1 N_0 \leqq m. \tag{4.5}$$

Similarly, the probability of decreasing the gene frequency by ξ_2 during the deletion stage of the cycle is

$$P_{x+\xi_1 \to x+\xi_1-\xi_2} = \binom{m}{\xi_2 N_0} x^{\xi_2 N_0}(1-x)^{m-\xi_2 N_0}, \tag{4.6}$$

$$0 \leqq \xi_2 N_0 \leqq m,$$

where I have neglected the small changes in gene frequency and repeat length caused by the preceeding duplication event. Then, for a series of cycles with random mispairings m, occurring with frequency $f(m)$, and ξ_1 and ξ_2 taking on all possible values for a given m, one finds

$$V_{\Delta x} = \sum_{m=1}^{N_0} f(m) \sum_{\xi_1 N_0=0}^{m} \sum_{\xi_2 N_0=0}^{m} (\xi_1 - \xi_2)^2 P_{x \to x+\xi_1} P_{x+\xi_1 \to x+\xi_1-\xi_2}$$

$$= \frac{2\bar{m}x(1-x)}{N_0^2} \tag{4.7}$$

and

$$M_{\Delta x} = \sum_{m=1}^{N_0} f(m) \sum_{\xi_1 N_0=0}^{m} \sum_{\xi_2 N_0=0}^{m} (\xi_1 - \xi_2) P_{x \to x+\xi_1} P_{x+\xi_1 \to x+\xi_1 \times \xi_2} = 0. \tag{4.8}$$

Using the results of Kimura and Ohta [19] one can now show that the mean time to fixation measured in crossovers is

$$T_{\text{dif}} = -\frac{2N_0^2}{\bar{m}}(N_0 - 1)\ln(1 - 1/N_0). \tag{4.9}$$

This differs from (4.4) for the unit mispairing problem by the presence of $\bar{m}$ in the denominator. Thus to the crude level of approximation implied by this model, the only effect of multiple mispairing is to speed up fixation; crossovers with average mispairing $\bar{m}$ are equivalent to $\bar{m}$ unit mispairing unequal crossovers.

4.3. Much work remains to be done on the multiple mispairing problem. The assumption of duplications and deletions occurring in cycles needs to be relaxed and the effects of selection ($p \neq 1/2$) need to be explored. Possibly this should be done by a branching process approach. Further, the importance of gene order in multiple mispaired unequal crossovers needs to be ascertained. The physical process of unequal crossover with mispairing of two or more repeats can be represented as a "necklace problem." At each time unit, a section of m tandem beads

on a necklace are removed and replaced by two identical tandem sections with probability p, and not replaced with probability $1 - p$. Given some initial ordering of beads on the necklace, what is the composition of the necklace at later times? What is the probability of certain bead types going extinct, becoming fixed, and what is the probability distribution of the times of occurrence of these processes? Ohta [23] has shown that if the maximum mispairing is 10%–15% of the total family length then a random gene order is to be expected.

5. **Conclusions.** The process of unequal crossing-over between sister chromatids can explain both the existence and the novel evolutionary features of multigene families. Before a multigene family can arise by unequal crossing-over at least two tandem homologous genes must be present. Experiments with prokaryotic organisms suggest that de novo gene duplications do occur and at relatively high frequencies [20–22]. In order for a small number of tandem repeats to generate a large multigene family by unequal crossing-over, one would require selection of the expanded chromosome over the contracted one, i.e., $p > 1/2$. With $p > 1/2$ the mean family size would increase with time. At some point one would suppose that the selective pressures for expanding the family size would cease, p would take on the value of $1/2$, and the mean family size would remain constant.

One method of generating a homogeneous family which exhibited coincidental evolution would be the fixation of one gene in the multigene family at a rate which was rapid compared to the time needed for mutations to accumulate. When $p = 1/2$ fixation of one gene in the family is certain to occur with a mean fixation time of $2N_0(N_0 - 1)$ [or $2(N_0 - 1)/k$] when mispairings of one unit occur. Here N_0 can be interpreted as the constant mean family size, and I have assumed that due to mutations all repeats in the family may be different when the constant size of N_0 is attained. If multiple mispairing occurs this fixation time can be shortened by a factor equal to the mean mispairing per crossover.

References

1. G. A. Galau, M. E. Chamberlin, B. R. Hough, R. J. Britten, and E. H. Davidson, *Evolution of repetitive and nonrepetitive DNA*, in *Molecular Evoluution*, (F. J. Ayala, ed.), Sinauer, Sunderlund, Massachusetts, (1976), 200–224.

2. H. Cooke, *Repeated sequence specific to human males*, Nature **262** (1976), 182–186.

3. L. Hood, J. H. Campbell, and S. C. R. Elgin, *The organization, expression, and evolution of antibody genes and other multigene families*, Ann. Rev. Genet. **9** (1975), 305–353.

4. D. D. Brown and C. S. Weber, *Gene linkage by RNA-DNA hybridization. I. Unique DNA sequences homologous to 4S RNS, 5S RNA and ribosomal RNA,* J. Mol. Biol. **34** (1968), 661–680.

5. D. D. Brown, P. C. Wensink, and E. Jordan, *A comparison of the ribosomal DNAs of Xenopus laevis and Xenopus mulleri: the evolution of tandem genes,* J. Mol. Biol. **63** (1972), 57–73.

6. M. L. birnstiel, M. Chipchase, and J. Spiers, *The ribosomal RNA cistrons,* Prog. Nucleic Acid Res. Molec. Biol. **11** (1971), 351–389.

7. P. K. Wellauer and I. B. Dawid, *Secondary structure maps of ribosomal RNA and DNA. I. Processing of Xenopus laevis rRNA and structure of single-stranded rDNA,* J. Mol. Biol. **89** (1974), 379–395.

8. P. K. Wellauer, R. H. Reeder, D. Carroll, D. D. Brown, A. Deutch, T. Higashinakagawa, and I. B. Dawid, *Amplified ribosomal DNA from Xenopus laevis has heterogeneous spacer lengths,* Proc. Nat. Acad. Sci. **71** (1974), 2823–2827.

9. G. P. Smith, *Unequal crossover and the evolution of multigene families,* Cold Spring Harbor Symp. Quant. Biol. **38** (1973), 507–514.

10. J. A. Black and D. Gibson, *Neutral evolution and immunoglobulin diversity,* Nature **250** (1974), 327–328.

11. G. P. Smith, *Evolution of repeated DNA sequences by unequal crossing-over,* Science **191** (1976), 528–535.

12. T. Ohta, *Simple model for treating evolution of multigene families,* Nature **263** (1976), 74–76.

13. A. S. Perelson and G. I. Bell, *Mathematical models for the evolution of multigene families by unequal crossing-over,* Nature **265** (1976), 304–310.

14. T. Ohta, *Genetic variation in multigene families,* Nature **267** (1977), 515–517.

15. W. Feller, *An Introduction to Probability Theory and its Applications* 1, 3rd ed., Wiley, New York, 1968.

16. E. M. Southern, *Long range periodicities in mouse satellite DNA,* J. Mol. Biol. **94** (1975), 51–69.

17. W. A. Beyer and M. S. Waterman, *Mean absorption time for a conditioned random walk,* Stud. Appl. Math. (in press).

18. J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory,* Harper and Row, New York, 1970.

19. M. Kimura and T. Ohta, *The average number of generations until fixation of a mutant gene in a finite population,* Genetics **61** (1969), 763–771.

20. C. W. Hill and G. Combriato, *Genetic duplications induced at very high frequency by ultraviolet irradiation in E. coli,* Molec. Gen. Genet. **127** (1973), 197–214.

21. S. W. Emmons, V. MacCosham, R. L. Baldwin, and J. O. Thomas, *Tandem genetic duplications in phage lambda. III. The frequency of duplication mutants in two derivates of phage lambda is independent of known recombination systems,* J. Mol. Biol. **91** (1975), 133–146.

22. J. Langridge, *Mutations conferring quantitative and qualitative increases in β-galactosidase activity in Escherichia coli,* Molec. Gen. Genet. **105** (1969), 74–83.

23. T. Ohta, *Theoretical study on genetic variation in multigene families,* Genet. Res. **31** (1978), 13–28.

THEORETICAL DIVISION, UNIVERSITY OF CALIFORNIA, LOS ALAMOS SCIENTIFIC LABORATORY, LOS ALAMOS, NEW MEXICO 87545

CURRENT ADRESS: DIVISION OF BIOLOGY AND MEDICINE, BROWN UNIVERISTY, PROVIDENCE, RHODE ISLAND 02912