THE EVOLUTION OF MULTIGENE FAMILIES UNDER INTRACHROMOSOMAL GENE CONVERSION

THOMAS NAGYLAKI

Department of Biophysics and Theoretical Biology, The University of Chicago, Chicago, Illinois 60637

Manuscript received July 21, 1983
Revised copy accepted November 15, 1983

ABSTRACT

A model for the evolution of the probabilities of genetic identity within and between loci of a multigene family in a finite population is formulated and investigated. Unbiased intrachromosomal gene conversion, equal crossing over between tandemly repeated genes, random genetic drift and mutation to new alleles are incorporated. Generations are discrete and nonoverlapping; the diploid, monoecious population mates at random. Formulas for the equilibrium values of the probabilities of identity and a cubic equation for the rate of convergence are deduced. Numerical examples indicate the following. The amount of homology at equilibrium generally decreases as the mutation rate, the population size and the number of repeats increase; it may increase or decrease with increasing crossover rate. The intralocus homology has an intermediate minimum, whereas the interlocus homology increases, as the rate of gene conversion increases. The intralocus homology decreases, whereas the interlocus homology increases, as the proportion of symmetric heteroduplexes increases. The characteristic convergence time can be sufficiently short to imply that intrachromosomal gene conversion may be an important mechanism for maintaining sequence homogeneity among repeated genes. The convergence time decreases as the conversion rate and the proportion of symmetric heteroduplexes increase; although exceptions occur, it generally increases as the population size and the number of repeats increase; it may increase or decrease with increasing crossover rate.

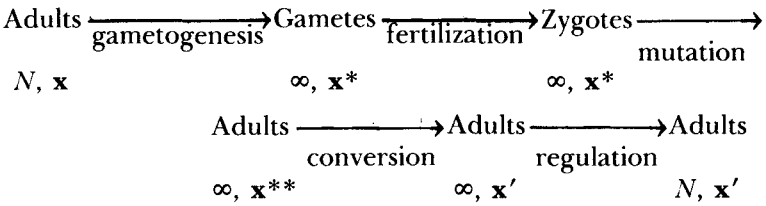
THERE has been a great deal of recent interest in the evolution of multigene families under gene conversion; DOVER (1982), NAGYLAKI and PETES (1982), OHTA (1982, 1983a, b, 1984) and OHTA and DOVER (1983) discuss the relevant data and the biological importance and background of this problem. NAGYLAKI and PETES (1982) studied the maintenance of sequence homogeneity among tandemly repeated genes under intrachromosomal gene conversion. They showed for a single chromosome lineage that a small conversional bias can have a dramatic effect on the fixation probability of a new variant and that conversion can act sufficiently rapidly to be an important mechanism for producing and conserving sequence homogeneity. Recent investigations of the influence of biased interchromosomal gene conversion on population dynamics at a single locus (NAGYLAKI 1983a, b; WALSH 1983) indicate that the extension of the analysis of NAGYLAKI and PETES (1982) to the population level is not only important but also probably nontrivial.

The data on interchromosomal gene conversion reviewed by LAMB and HELMI (1982), NAGYLAKI and PETES (1982) and NAGYLAKI (1983a) suggest that evolutionarily significant disparities are probably common in intrachromosomal gene conversion. Those data also suggest, however, that in many cases the disparities may be extremely small, possibly even zero. In this special but informative situation, the population problem can be analyzed in terms of probabilities of genetic identity (OHTA 1982, 1983a,b, 1984; OHTA and DOVER 1983). This approach, which we follow in this paper, leads directly to the equilibrium values of the intralocus and interlocus homologies and to the rate of approach to equilibrium. In the absence of mutation, the latter yields the characteristic time to sequence homogeneity.

Here, we shall study the evolution of a multigene family under the joint action of unbiased intrachromosomal gene conversion, equal crossing over between tandemly repeated genes, random genetic drift and mutation to new alleles. In FORMULATION, we shall offer a detailed formulation of our model; in particular, we shall present explicitly the assumptions and parameters that relate to gene conversion. If all evolutionary forces are weak, all heteroduplexes are asymmetric (MESELSON and RADDING 1975) and OHTA's conversion rate is suitably identified, then our recursion relations for the probabilities of identity reduce to hers (OHTA 1982; 1983a, b). In EQUILIBRIUM, we shall examine the amount and pattern of homology at equilibrium; we shall investigate the rate of convergence in CONVERGENCE. We shall consider extensions of and alternatives to our model in DISCUSSION.

FORMULATION

Generations are discrete and nonoverlapping; the diploid, monoecious population mates at random. The life cycle starts with N adults; n represents the number of repeats, which are arranged in tandem. We use three probabilities of identity to summarize the genetic structure of the population; these provide much important biological information but do not fully specify the state of the population. The term "identity" must be interpreted in accordance with the type of data available: at the most detailed level, it refers to identity of the DNA sequences of two genes; if less information is available, it can signify coincidence of restriction sites or the ability to hybridize. We assume that the n loci are exchangeable (*i.e.*, equivalent). Let f_t denote the probability that two distinct genes at the same locus, chosen at random from adults just before gametogenesis in generation t (= 0, 1, 2, . . .), are identical. Then f represents the expected homozygosity; $h = 1 - f$, the expected heterozygosity, is a measure of intralocus genetic variability in the population. Let g_t denote the probability that two distinct genes on the same chromosome, chosen at random from adults just before gametogenesis in generation t, are identical. Clearly, g is an index of homology between repeats within a chromosome. Finally, let l_t denote the probability that two genes at different loci and on different chromosomes, chosen at random from adults just before gametogenesis in generation t, are identical. Thus, l incorporates both intralocus and interlocus variation. We posit the life cycle shown below; $\mathbf{x}$ designates the vector of the probabilities of identity, and the prime signifies the next generation.



We neglect the dependence of the probabilities of identity on the positions of the genes sampled. This dependence is absent if, and only if, there are only two repeats or there is no crossing over. We shall discuss this simplifying assumption in DISCUSSION.

We suppose that at least one generation of panmixia has occurred before we examine the population. If two genes are chosen from distinct individuals, we neglect the second-order probability that gene conversion has occurred in both individuals. Then the corresponding probabilities of identity within and between individuals are equal. At the beginning of the life cycle, N adults produce infinitely many gametes, which fuse at random to form zygotes. Thus, a proportion $1/N$ of zygotes are produced by self-fertilization. If r denotes the frequency of equal reciprocal recombination between two distinct loci chosen at random, then the probabilities of identity in zygotes read

$$f^* = \theta + (1 - \theta)f, \tag{1a}$$

$$g^* = (1 - r)g + rl, \tag{1b}$$

$$l^* = \theta g + (1 - \theta)l, \tag{1c}$$

where $\theta = 1/(2N)$. Equation (1a) is standard (WRIGHT 1931; MALÉCOT 1946, 1948; KIMURA 1963); we owe (1b) and (1c) to KIMURA (1963).

The crossover probability r will generally be an increasing function of the number of repeats, n, whereas the probability of equal crossing over between two adjacent loci, β, should not depend on n. We assume that at most one crossover occurs per generation in the entire multigene family. This is a reasonable approximation if there is complete positive interference or, more likely, if $(n - 1)\beta \ll 1$. To express r in terms of β, note that the probability of first choosing locus i $(1 \leq i \leq n)$ and then locus j $(j \neq i)$ is $[n(n - 1)]^{-1}$, and that $|i - j|\beta$ gives the probability of crossing over between i and j. Therefore,

$$r = \sum_{i=1}^{n} \sum_{j=1}^{n} [n(n - 1)]^{-1} |i - j| \beta = \tfrac{1}{3}(n + 1)\beta, \tag{2}$$

as given by OHTA (1983a). Her previous definition of β is different, and hence the factor $n + 1$ is absent in OHTA (1982).

Mutation is next. We suppose that every allele mutates to a new allele at rate u $(0 \leq u \leq 1)$. This model of "infinite alleles" was proposed by MALÉCOT (1946, 1948) for identity by descent and by WRIGHT (1949) and KIMURA and CROW (1964) for identity in state. After mutation, we have

$$f^{**} = vf^*, \qquad g^{**} = vg^*, \qquad l^{**} = vl^*, \tag{3}$$

where $v = (1 - u)^2$.

To incorporate gene conversion, we posit the following. First, an interaction

between two alleles cannot produce a third allele. All previous work on this problem (NAGYLAKI and PETES 1982; OHTA 1982, 1983a, b; OHTA and DOVER 1983) is subject to this restriction, which is discussed in NAGYLAKI (1983a). Second, each interaction involves the formation of heteroduplexes between two repeated genes. The heteroduplexes may be either symmetric (HOLLIDAY 1964) or asymmetric (MESELSON and RADDING 1975). Third, interactions occur between repeats within a single tandem array (intrachromatid conversion), before chromosome duplication. We neglect sister-chromatid and interchromosomal interactions. Fourth, there is at most one interaction per individual per generation. If an interaction occurs, it does so between two randomly chosen genes. Fifth, all mismatches are repaired. Sixth, parity (FOGEL, MORTIMER and LUSNAK 1981; NAGYLAKI and PETES 1982) obtains for both the initiation of asymmetric heteroduplex formation and the repair of mismatches. Seventh, if symmetric heteroduplexes are formed, the direction of correction of one heteroduplex is independent of that of the other. Eighth, crossing over is not associated with gene conversion. Consult NAGYLAKI and PETES (1982) for discussion of assumptions 2 to 8.

We introduce now the basic parameters that describe gene conversion, and we derive some simple preliminary relations. Let μ designate the probability per individual per generation that an interaction occurs. We denote by I_b and $\not{I}_b$ the events that gene b interacts with some other gene and that it does not, respectively. Similarly, I_{bc} and $\not{I}_{bc}$ represent the events that genes b and c interact with each other and that they do not, respectively. To calculate the corresponding probabilities, observe that an interaction occurs on the chromosome b is on with probability $\mu/2$ and this interaction involves b with probability $2/n$. Therefore,

$$P(I_b) = (\mu/2)(2/n) = \mu/n. \tag{4}$$

Since there are $n(n-1)/2$ combinations of two genes on a chromosome, we have

$$P(I_{bc}) = (\mu/2)[n(n-1)/2]^{-1} = \mu[n(n-1)]^{-1}. \tag{5}$$

The probability μ should increase at least linearly with n and may increase as fast as quadratically (NAGYLAKI and PETES 1982). Hence, the parameters

$$\lambda = \tfrac{1}{4}P(I_b) = \mu/(4n), \tag{6}$$

$$\alpha = \tfrac{1}{2}P(I_{bc}) = \mu/[2n(n-1)] \tag{7}$$

will depend more weakly on n than μ does. We shall call the interaction probabilities λ and α conversion rates.

We write $b = c$ and $b \neq c$ to signify the identity and nonidentity of alleles b and c, respectively. The notations $b \to c$ and $b \not\to c$ mean, respectively, that b is converted to c and that it is not. To describe the effect of gene conversion on f, g and l, we shall require the probabilities

$$p = P(b \to c \mid b \neq c, I_{bc}), \tag{8}$$

$$q = P(b \to c, c \not\to b \mid b \neq c, I_{bc}). \tag{9}$$

Let σ ($0 \leq \sigma \leq 1$) and $1 - \sigma$ represent the proportions of symmetric and

asymmetric heteroduplexes. Since all mismatches are repaired and parity holds for mismatch repair, the probability that a single bc heteroduplex yields b (or c) is $1/2$. If heteroduplexes are symmetric, this gives factors of $1/2$ and $(1/2)(1/2)$ for the events in (8) and (9), respectively. If heteroduplexes are asymmetric, for both events heteroduplex formation must be initiated by the DNA molecule that contains c; this has conditional probability $1/2$, because we have posited parity for the initiation of heteroduplex formation. Then $c \not\rightarrow b$ automatically, and $b \rightarrow c$ has conditional probability $1/2$. Consequently,

$$p = \sigma(1/2) + (1 - \sigma)(1/2)(1/2) = \rho/4, \tag{10}$$

where $\rho = 1 + \sigma$ $(1 \leq \rho \leq 2)$, and

$$q = \sigma(1/2)(1/2) + (1 - \sigma)(1/2)(1/2) = 1/4. \tag{11}$$

We are now prepared to relate (f', g', l') to (f^{**}, g^{**}, l^{**}). In Figure 1, we focus attention on two repeats on each chromosome; d, e and i denote the genes a, b and c after conversion, respectively. We choose e at random. Solely for the calculation of f', if b interacts, define c as the gene with which it does so. Bearing in mind that there is at most one interaction per individual and appealing to (4), we deduce

$$\begin{aligned} f' &= P(d = e) \\ &= P(d = e \mid I_a, I_b)P(I_a, I_b) + 2P(d = e \mid I_a, I_b)P(I_a, I_b) \\ &= P(a = b)\left(1 - \frac{2\mu}{n}\right) + 2P(a = e \mid I_{bc})\frac{\mu}{n} \\ &= f^{**} + \frac{2\mu}{n}[P(a = e \mid I_{bc}) - f^{**}]. \end{aligned} \tag{12}$$

From (8) we obtain

$$\begin{aligned} &P(a = e \mid I_{bc}) \\ &\quad = P(a = e \mid b = c, I_{bc})P(b = c \mid I_{bc}) \\ &\qquad + P(a = e \mid b \neq c, b \not\rightarrow c, I_{bc})P(b \neq c, b \not\rightarrow c \mid I_{bc}) \\ &\qquad + P(a = e \mid b \neq c, b \rightarrow c, I_{bc})P(b \neq c, b \rightarrow c \mid I_{bc}) \\ &\quad = P(a = b \mid b = c)P(b = c) \\ &\qquad + P(a = b \mid b \neq c)P(b \not\rightarrow c \mid b \neq c, I_{bc})P(b \neq c \mid I_{bc}) \\ &\qquad + P(a = c \mid b \neq c)P(b \rightarrow c \mid b \neq c, I_{bc})P(b \neq c \mid I_{bc}) \\ &\quad = P(a = b \mid b = c)P(b = c) + P(a = b \mid b \neq c)(1 - p)P(b \neq c) \\ &\qquad + P(a = c \mid b \neq c)pP(b \neq c) \\ &\quad = P(a = b) + p[P(a = c \mid b \neq c)P(b \neq c) + P(a = c \mid b = c)P(b = c) \\ &\qquad - P(a = b \mid b = c)P(b = c) - P(a = b \mid b \neq c)P(b \neq c)] \\ &\quad = f^{**} + p[P(a = c) - P(a = b)] \\ &\quad = f^{**} + p(l^{**} - f^{**}). \end{aligned}$$

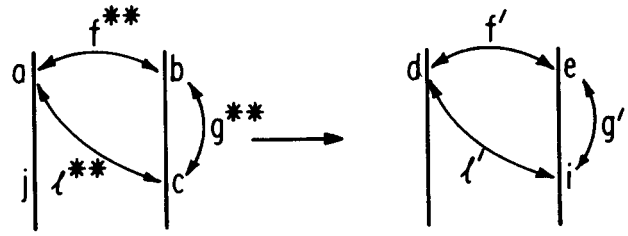


FIGURE 1.—The probabilities of identity before and after gene conversion.

Therefore, (7), (10) and (12) yield

$$f' = [1 - (n-1)\rho\alpha]f^{**} + (n-1)\rho\alpha l^{**}. \tag{13a}$$

Henceforth, i designates a repeat chosen at random, subject only to the condition that e and i be on the same chromatid at distinct loci. We invoke (5), (7), (9) and (11):

$$\begin{aligned}
g' &= P(e = i) \\
&= P(e = i \mid \not{I}_{bc})P(\not{I}_{bc}) + P(e = i \mid I_{bc})P(I_{bc}) \\
&= P(b = c)\left[1 - \frac{\mu}{n(n-1)}\right] + \frac{\mu}{n(n-1)}\,[P(e = i \mid b = c,\, I_{bc})P(b = c \mid I_{bc}) \\
&\quad + P(e = i \mid b \neq c;\, b \to c \text{ and } c \to b, \text{ or } b \not\to c \text{ and } c \not\to b;\, I_{bc})\cdot \\
&\quad \cdot P(b \neq c;\, b \to c \text{ and } c \to b, \text{ or } b \not\to c \text{ and } c \not\to b \mid I_{bc}) \\
&\quad + 2P(e = i \mid b \neq c,\, b \to c,\, c \not\to b,\, I_{bc})P(b \neq c,\, b \to c,\, c \not\to b \mid I_{bc})] \\
&= g^{**}\left[1 - \frac{\mu}{n(n-1)}\right] + \frac{\mu}{n(n-1)}\,[(1)g^{**} + 0 \\
&\quad + 2(1)P(b \to c,\, c \not\to b \mid b \neq c,\, I_{bc})P(b \neq c \mid I_{bc})] \\
&= g^{**}\left[1 - \frac{\mu}{n(n-1)}\right] + \frac{\mu}{n(n-1)}\,[g^{**} + 2q(1 - g^{**})] \\
&= \alpha + (1 - \alpha)g^{**}.
\end{aligned} \tag{13b}$$

For l', (5) gives

$$\begin{aligned}
l' &= P(d = i) \\
&= P(d = i \mid \not{I}_{aj}, \not{I}_{bc})P(\not{I}_{aj}, \not{I}_{bc}) + 2P(d = i \mid \not{I}_{aj}, I_{bc})P(\not{I}_{aj}, I_{bc}) \\
&= P(a = c)\left[1 - \frac{2\mu}{n(n-1)}\right] + \frac{2\mu}{n(n-1)}P(a = i \mid I_{bc}) \\
&= l^{**} + \frac{2\mu}{n(n-1)}\,[P(a = i \mid I_{bc}) - l^{**}].
\end{aligned}$$

From (8) we get

$$\begin{aligned}
P(a = i \mid I_{bc}) &= P(a = i \mid b = c, I_{bc})P(b = c \mid I_{bc}) \\
&\quad + P(a = i \mid b \neq c, c \not\to b, I_{bc})P(b \neq c, c \not\to b \mid I_{bc}) \\
&\quad + P(a = i \mid b \neq c, c \to b, I_{bc})P(b \neq c, c \to b \mid I_{bc}) \\
&= P(a = c \mid b = c)P(b = c) \\
&\quad + P(a = c \mid b \neq c)P(c \not\to b \mid b \neq c, I_{bc})P(b \neq c \mid I_{bc}) \\
&\quad + P(a = b \mid b \neq c)P(c \to b \mid b \neq c, I_{bc})P(b \neq c \mid I_{bc}) \\
&= P(a = c \mid b = c)P(b = c) + P(a = c \mid b \neq c)(1 - p)P(b \neq c) \\
&\quad + P(a = b \mid b \neq c)pP(b \neq c) \\
&= P(a = c) + p[P(a = b \mid b \neq c)P(b \neq c) + P(a = b \mid b = c)P(b = c) \\
&\quad - P(a = c \mid b = c)P(b = c) - P(a = c \mid b \neq c)P(b \neq c)] \\
&= l^{**} + p[P(a = b) - P(a = c)] \\
&= l^{**} + p(f^{**} - l^{**}).
\end{aligned}$$

Hence, (7) and (10) lead to

$$l' = \rho\alpha f^{**} + (1 - \rho\alpha)l^{**}. \tag{13c}$$

Substituting (3) into (13) and then (1) into the result produces the exact recursion relations for our model:

$$\begin{aligned}
f' &= v\{\theta[1 - (n - 1)\rho\alpha] + (1 - \theta)[1 - (n - 1)\rho\alpha]f \\
&\quad + (n - 1)\rho\alpha\theta g + (n - 1)\rho\alpha(1 - \theta)l\},
\end{aligned} \tag{14a}$$

$$g' = \alpha + v(1 - \alpha)[(1 - r)g + rl], \tag{14b}$$

$$l' = v[\rho\alpha\theta + \rho\alpha(1 - \theta)f + \theta(1 - \rho\alpha)g + (1 - \theta)(1 - \rho\alpha)l]. \tag{14c}$$

The system (14) involves six parameters: v (or u), θ (or N), n, α (or μ or λ), ρ (or σ) and r (or β). It depends on the order of the evolutionary forces in the life cycle. However, our assumptions concerning recombination are plausible only if both crossing over and gene conversion have low probabilities, and we lose no biological generality by positing weak mutation and random drift. If u, $n\alpha$ (or λ), θ (or $1/N$) and r (or $n\beta$) are all much less than unity, (14) becomes

$$f' = \theta + [1 - 2u - \theta - (n - 1)\rho\alpha]f + (n - 1)\rho\alpha l, \tag{15a}$$

$$g' = \alpha + (1 - 2u - \alpha - r)g + rl, \tag{15b}$$

$$l' = \rho\alpha f + \theta g + (1 - 2u - \theta - \rho\alpha)l. \tag{15c}$$

Here and later, we simplify writing by not indicating explicitly that (15) and all subsequent equations are approximate. In the approximate system (15), the evolutionary forces are additive, and this system is independent of the order of these forces.

If all heteroduplexes are asymmetric ($\rho = 1$), (15) reduces to the equations of OHTA (1982, 1983a, b). Hence, (6) reveals the precise biological interpretation of her parameter λ.

EQUILIBRIUM

The solution of (15) converges to the unique equilibrium given by

$$\hat{f} = [\theta + (n - 1)\rho\alpha\hat{I}]/[2u + \theta + (n - 1)\rho\alpha], \tag{16a}$$

$$\hat{g} = (\alpha + r\hat{I})/(2u + \alpha + r), \tag{16b}$$

$$\hat{I} = \alpha\theta[2u(1 + \rho) + \theta + \rho r + n\rho\alpha]/D, \tag{16c}$$

$$D = (2u + \theta)(2u + \alpha)(2u + \theta + n\rho\alpha) + r[\theta\rho\alpha + 2u(2u + \theta + n\rho\alpha)]. \tag{16d}$$

If all heteroduplexes are asymmetric ($\rho = 1$), (16) simplifies to OHTA's (1982, 1983b) result. The equilibrium (16) depends on five independent parameters: n, ρ, $N\alpha$, Nr and Nu. Some limiting cases of (16) are instructive and provide checks. As $u \to 0$, all variability disappears: $\hat{f}$, $\hat{g}$, $\hat{I} \to 1$. We also obtain the expected limits in the absence of gene conversion: as $\alpha \to 0$, (16) yields $\hat{f} \to f_0$ and $\hat{g}$, $\hat{I} \to 0$, where

$$f_0 = \theta/(2u + \theta) = 1/(1 + 4Nu). \tag{17}$$

Equation (17) is just the standard result (MALÉCOT 1946, 1948; KIMURA and CROW 1964) for the balance between mutation and random drift. Obviously, without gene conversion, the interlocus homologies must vanish at equilibrium. If $\theta \to 0$ ($N \to \infty$), we find $\hat{f}$, $\hat{I} \to 0$ and $\hat{g} \to g_1$, where

$$g_1 = \alpha/(2u + \alpha + r). \tag{18}$$

In an infinitely large population, there is no homology at equilibrium between chromosomes.

Some simple bounds illuminate the rather complicated solution (16). Since $\hat{I} > 0$, we have $\hat{f} > f_1$ and $\hat{g} > g_1$, where

$$f_1 = \theta/[2u + \theta + (n - 1)\rho\alpha] = 1/[1 + 4N(u + \rho\lambda)]. \tag{19}$$

By discarding terms independent of α in (16d) and noting that $n > 1$, we obtain a lower bound on D; inserting this into (16c) leads to $\hat{I} < f_0$. This upper bound on $\hat{I}$ and (16a) reveal that $\hat{f} < f_0$ and $\hat{f} - \hat{I} > 0$. Finally, $\hat{I} < 1$ and (16b) imply $\hat{g} < g_0$, where

$$g_0 = (\alpha + r)/(2u + \alpha + r). \tag{20}$$

Thus, we have shown

$$f_1 < \hat{f} < f_0, \qquad g_1 < \hat{g} < g_0, \qquad 0 < \hat{I} < \hat{f}, \tag{21}$$

where f_0, f_1, g_0 and g_1 are given by (17) to (20).

It is intuitively expected that interaction with other loci should lower the probability of intralocus identity ($\hat{f} < f_0$). The result $\hat{I} < \hat{f}$ is also reasonable.

Perhaps surprisingly, however, $\hat{l} > \hat{g}$ can occur; in fact, algebraic manipulation of (16) shows that it does so if, and only if,

$$2u + n\rho\alpha < (\rho - 1)\theta, \tag{22}$$

which requires some admixture of symmetric heteroduplexes ($\rho > 1$). For example, if $n = \rho = 2$, $\lambda = \beta = u$ and $\theta = 20u$, then for any $u > 0$, $\hat{g} = 70/139 = 0.5036$ and $\hat{l} = 72/139 = 0.5180$. This phenomenon is not caused by the fact that (16) is the lowest-order approximation in u, α, r and θ: since the homologies given by (16) are homogeneous functions of degree zero in these parameters, by multiplying all four parameters by the same constant, we can make (16) arbitrarily accurate without altering it. From Tables 1 and 2 we see that $\hat{f} > \hat{g}$ and $\hat{g} > \hat{f}$ both occur. Hence, the possible relative magnitudes are $\hat{f} > \hat{g} > \hat{l}$, $\hat{g} > \hat{f} > \hat{l}$, and $\hat{f} > \hat{l} > \hat{g}$.

OHTA (1982, 1983b) has tabulated some numerical examples for asymmetric heteroduplexes. In Tables 1 and 2, we exhibit for various parameter values the extreme cases of purely asymmetric ($\rho = 1$) and purely symmetric ($\rho = 2$) heteroduplexes. Unless otherwise specified, the parameters in Table 1 are $\lambda = 5 \times 10^{-6}$, $u = 10^{-8}$, $\beta = 10^{-3}$, $n = 50$ and $N = 5 \times 10^4$; the default values in Table 2 read $\lambda = 10^{-3}$, $u = 10^{-6}$, $\beta = 10^{-3}$, $n = 50$ and $N = 5 \times 10^4$. Much more extensive computations indicate the following.

All three probabilities of identity decrease as n (with λ fixed) and u increase. The intralocus homology ($\hat{f}$) has an intermediate minimum, whereas the interlocus homologies ($\hat{g}$ and $\hat{l}$) increase, as λ increases; $\hat{f}$ decreases, whereas $\hat{g}$ and $\hat{l}$ increase, as the proportion of symmetric heteroduplexes (σ) increases. The behavior of the homologies as a function of n with α fixed is generally similar to their dependence on n with λ fixed. At least for small n and $\rho > 1$, however, $\hat{g}$ and $\hat{l}$ can increase as n increases with α fixed: take $\rho = 2$, $\alpha = 10^{-6}$, $u = 5 \times 10^{-6}$, $\beta = 10^{-5}$ and $N = 5000$; then for $n = 2$ and 3, $(\hat{f}, \hat{g}, \hat{l}) = (0.8947, 0.0961, 0.1018)$ and $(0.8808, 0.0973, 0.1026)$, respectively. Although it appears that the homologies usually decrease as β increases, at least for $\rho > 1$ and relatively weak conversion, they can increase: choose $n = 10$ and ρ, α, u and N as above; then for $\beta = 10^{-5}$ and 10^{-4}, $(\hat{f}, \hat{g}, \hat{l}) = (0.7960, 0.1019, 0.1052)$ and $(0.7988, 0.1239, 0.1249)$, respectively. Finally, although the homologies usually decrease as N increases, $\hat{g}$ and $\hat{l}$ can increase if $\rho > 1$, n is fairly small and $0 < N\alpha, N\beta \ll 1$: for example, take $\rho = 2$, $n = 2$, $\alpha = 10^{-6}$, $u = 5 \times 10^{-7}$ and $\beta = 10^{-6}$; then for $N = 5000$ and 10,000, $(\hat{f}, \hat{g}, \hat{l}) = (0.9807, 0.5022, 0.5066)$ and $(0.9627, 0.5039, 0.5117)$, respectively.

The results of this section show that, especially if symmetric heteroduplexes occur ($\rho > 1$) and conversion is relatively weak, random drift and gene conversion can interact to produce complex, intuitively surprising equilibrium behavior.

CONVERGENCE

We now study the rate of convergence to the equilibrium (16). From (15) we infer that the column vector $\mathbf{k} = (f - \hat{f}, g - \hat{g}, l - \hat{l})^T$ satisfies $\mathbf{k}' = B\mathbf{k}$, where

TABLE 1

The probabilities of identity at equilibrium

Parameter	$\rho = 1$			$\rho = 2$		
	$\hat{f}$	$\hat{g}$	$\hat{l}$	$\hat{f}$	$\hat{g}$	$\hat{f}$
$10^4\lambda$						
0.01	0.937	0.629	0.629	0.925	0.743	0.743
1	0.910	0.906	0.906	0.910	0.908	0.908
100	0.956	0.957	0.956	0.956	0.957	0.956
10^7u						
0.1	0.916	0.835	0.835	0.913	0.870	0.870
1	0.660	0.334	0.334	0.596	0.400	0.400
10	0.475	0.044	0.044	0.350	0.059	0.059
$10^4\beta$						
0.1	0.920	0.842	0.842	0.915	0.873	0.873
1	0.917	0.835	0.835	0.913	0.871	0.871
10	0.916	0.835	0.835	0.913	0.870	0.870
$10^{-4}N$						
0.1	0.998	0.909	0.909	0.998	0.951	0.951
1	0.982	0.894	0.894	0.981	0.935	0.935
10	0.846	0.770	0.770	0.840	0.801	0.801
n						
10	0.981	0.963	0.963	0.981	0.972	0.972
100	0.857	0.715	0.715	0.846	0.770	0.770
1000	0.599	0.200	0.200	0.500	0.250	0.250

$B = C - 2uI$, I is the 3×3 identity matrix and

$$C = \begin{pmatrix} 1 - \theta - (n-1)\rho\alpha & 0 & (n-1)\rho\alpha \\ 0 & 1 - \alpha - r & r \\ \rho\alpha & \theta & 1 - \theta - \rho\alpha \end{pmatrix}. \tag{23}$$

The eigenvalues $\tilde{\xi}$ of B are given by $\tilde{\xi} = \xi - 2u$, where ξ represents the eigenvalues of C. We use $\tilde{T} = 2/(1 - \tilde{\xi}_m)$, where $\tilde{\xi}_m$ designates the maximal eigenvalue of the non-negative matrix B, as the characteristic convergence time. If there is no mutation, this reduces to OHTA's (1983a) measure,

$$T = 2/(1 - \xi_m) = 2/\epsilon_m, \tag{24}$$

where $\epsilon = 1 - \xi$. Since we may calculate $\tilde{T}$ from

$$\tilde{T} = T/(1 + Tu), \tag{25}$$

we focus our attention on T, the typical time to loss of genetic variability in the absence of mutation.

From the row sums of the non-negative matrix C we obtain for the maximal eigenvalue ξ_m the inequalities (GANTMACHER 1959, pp. 63, 68)

$$1 - \max(\theta, \alpha) \leq \xi_m \leq 1, \tag{26}$$

TABLE 2

The probabilities of identity at equilibrium

Parameter	$\rho = 1$			$\rho = 2$		
	$\hat{f}$	$\hat{g}$	$\hat{l}$	$\hat{f}$	$\hat{g}$	$\hat{l}$
$10^3\lambda$						
0.01	0.350	0.059	0.059	0.248	0.072	0.072
1	0.104	0.102	0.100	0.102	0.102	0.100
100	0.512	0.606	0.512	0.512	0.606	0.512
10^7u						
0.1	0.918	0.918	0.917	0.918	0.918	0.918
1	0.529	0.528	0.526	0.528	0.528	0.527
10	0.104	0.102	0.100	0.102	0.102	0.100
$10^5\beta$						
0.01	0.790	0.947	0.790	0.790	0.947	0.790
1	0.500	0.590	0.498	0.499	0.590	0.498
100	0.104	0.102	0.100	0.102	0.102	0.100
$10^{-4}N$						
0.1	0.853	0.818	0.817	0.849	0.831	0.830
1	0.368	0.354	0.353	0.363	0.357	0.355
10	0.055	0.055	0.053	0.054	0.055	0.053
n						
10	0.432	0.462	0.429	0.431	0.462	0.430
100	0.055	0.051	0.050	0.052	0.051	0.050
1000	0.010	0.005	0.005	0.007	0.005	0.005

whence

$$0 \leq \epsilon_m \leq \max(\theta, \alpha). \tag{27}$$

Therefore,

$$T \geq \min(4N, 2/\alpha). \tag{28}$$

Returning to (23), we find that ϵ satisfies

$$[\epsilon - \theta - (n-1)\rho\alpha][(\epsilon - \alpha - r)(\epsilon - \theta - \rho\alpha) - r\theta] - (n-1)\rho^2\alpha^2(\epsilon - \alpha - r) = 0, \tag{29}$$

which becomes

$$\epsilon^3 - a_1\epsilon^2 + a_2\epsilon - a_3 = 0, \tag{30a}$$

where

$$a_1 = r + 2\theta + (1 + n\rho)\alpha, \tag{30b}$$

$$a_2 = \theta[r + \theta + (2 + n\rho)\alpha] + n\rho\alpha(r + \alpha), \tag{30c}$$

$$a_3 = \alpha\theta(\theta + \rho r + n\rho\alpha). \tag{30d}$$

If all heteroduplexes are asymmetric ($\rho = 1$) and we replace ϵ by $1 - \xi$, then (30) simplifies to the result of OHTA (1983a). Putting $\eta = N\epsilon$ and multiplying (30a) by N^3, we see that η depends on four independent parameters: n, ρ, $\alpha_0 = N\alpha$ (or $\lambda_0 = N\lambda$) and $r_0 = Nr$ (or $\beta_0 = N\beta$). Therefore, so does the scaled characteristic convergence time

$$\tau = T/N = 2/\eta_m. \tag{31}$$

Equation (29) factors in the absence of gene conversion, random drift or crossing over. If $\alpha = 0$, (29) yields

$$\epsilon = 0, \theta, r + \theta, \tag{32}$$

in agreement with the work of KIMURA (1963). If $\theta = 0$ ($N = \infty$), (29) gives

$$\epsilon = 0, n\rho\alpha, \alpha + r. \tag{33}$$

If $\beta = 0$, we have

$$\epsilon = \theta, \alpha, \theta + n\rho\alpha, \tag{34}$$

which was derived by OHTA (1983a) for $\rho = 1$. Even without mutation, only in the last of these special cases is there convergence to complete homogeneity. From (24) and (34) we obtain the characteristic time

$$T = \max(4N, 2/\alpha), \tag{35}$$

which is, of course, at least as great as the lower bound (28). If $2N \geq 1/\alpha$, then $T = 4N$, which is both the exact characteristic convergence time (WRIGHT 1931; MALÉCOT 1946, 1948; KIMURA 1963) and the approximate mean conditional fixation time (KIMURA and OHTA 1969) of a new neutral mutant under pure random drift in a large population. If $2N < 1/\alpha$,

$$T = 2/\alpha = 4n(n - 1)/\mu. \tag{36}$$

Equations (4b) and (12a) of NAGYLAKI and PETES (1982) show that, in a single chromosome lineage with parity and one interaction per chromosome per generation, the mean conditional fixation time of a new mutant is $2(n - 1)^2$. For an average of $\mu/2$ interactions per chromosome per generation, as in this paper, this scales to

$$\bar{T}_1^* = 4(n - 1)^2/\mu. \tag{37}$$

Thus, $T \approx \bar{T}_1^*$ if the number of repeats is large ($n \gg 1$). Observe that the proportion of symmetric heteroduplexes does not influence (37). From (4b) and (12a) of NAGYLAKI and PETES (1982) we can see that this conclusion is false if sister-chromatid interactions are incorporated.

From (30) we can derive the asymptotic behavior of T in the limits $\alpha \to 0$ and $\theta \to 0$. In both cases, according to (32) and (33), $\epsilon_m \to 0$, whence (30) implies that $\epsilon_m \sim a_3/a_2$. Therefore, from (24) and (30) we infer

$$T \sim \frac{2}{\alpha}\left(\frac{1 + 2Nr}{1 + 2N\rho r}\right)$$

as $\alpha \to 0$ and

$$T \sim \frac{4nN(r + \alpha)}{r + n\alpha}$$

as $\theta \to 0$ (*i.e.*, $N \to \infty$).

OHTA (1983a) has presented for $\rho = 1$ some graphs of τ as a function of λ_0 (with n and β_0 fixed), as a function of β_0 (with n and λ_0 fixed) and as a function of n (with λ_0 and β_0 fixed). In Figures 2, 3 and 4 we display such graphs for $\rho = 2$. We exhibit τ as a function of n (with $\rho = 2$ and α_0 and β_0 fixed) and as a function of ρ (with n, λ_0 and β_0 fixed) in Figures 5 and 6. Figure 7 shows T as a function of N (with $\rho = 2$ and n, λ_0 and β_0 fixed). Extensive calculations as well as these and other plots indicate the following. The unscaled characteristic convergence time T decreases as λ (with ρ, n, N and β fixed) and ρ (with n, N, λ and β fixed) increase; T may increase or decrease as β increases (with ρ, n, N and λ fixed), the decrease occurring if $\rho > 1$ and $\alpha_0 \ll 1$. The time T increases as n increases if ρ, N, λ and β are fixed; T increases more slowly with increasing n if α is fixed instead of λ and even decreases for $\rho > 1$, $\alpha_0 \ll 1$, $\beta_0 \lesssim 1$ and n fairly small. Finally, T generally increases with increasing N if ρ, n, λ and β are fixed, but T can decrease for $\rho > 1$ and $\alpha_0 \ll 1$.

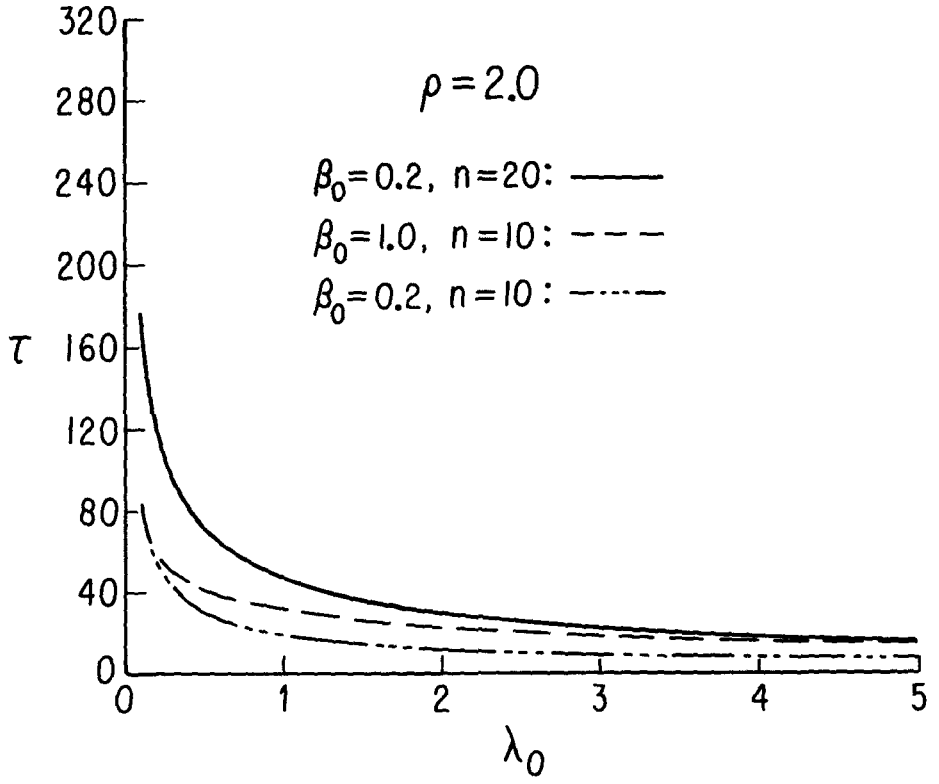


FIGURE 2.—The scaled convergence time as a function of the scaled conversion rate.

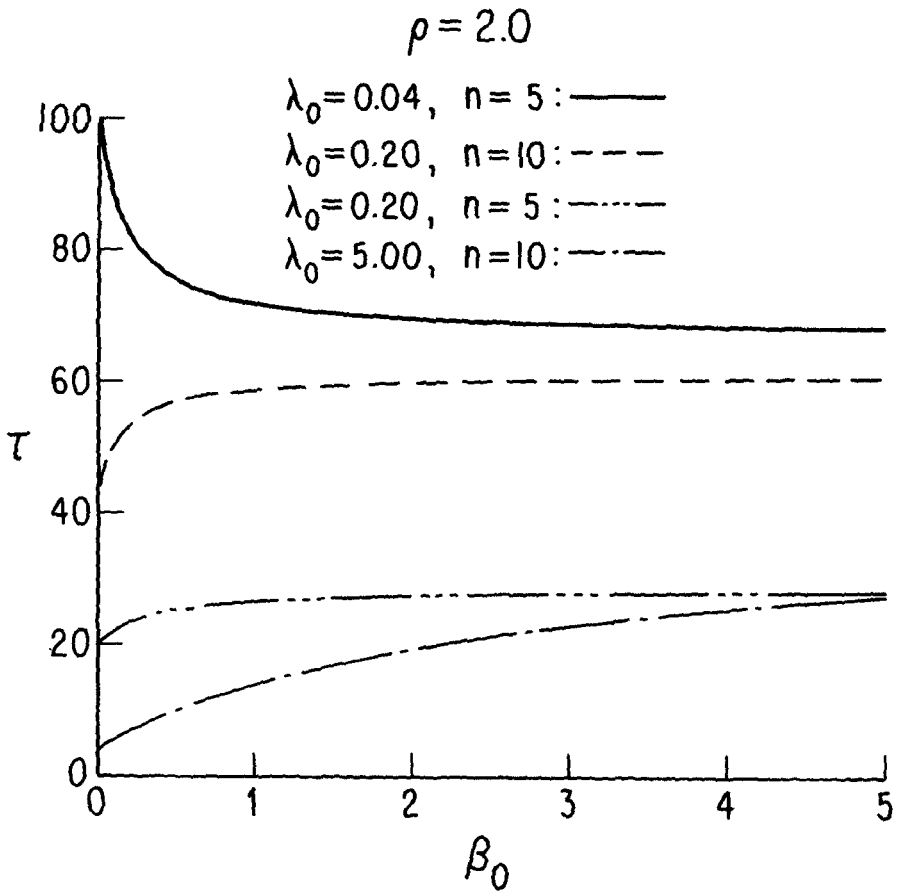


FIGURE 3.—The scaled convergence time as a function of the scaled crossover rate.

Thus, the convergence time exhibits complex, intuitively surprising behavior if symmetric heteroduplexes occur and gene conversion is relatively weak. The same conclusion was drawn at the end of EQUILIBRIUM for the equilibrium values of probabilities of identity.

DISCUSSION

We have formulated and investigated a model for the evolution of the probabilities of genetic identity within and between loci of a multigene family in a finite population. These probabilities converge globally to an equilibrium, which corresponds to complete homology within and between loci if, and only if, there is no mutation. The amount of homology at equilibrium ranges from zero to one; it depends sensitively on the mutation, conversion and crossover rates, the population size and the number of repeats, and less sensitively on the proportion of symmetric heteroduplexes. The characteristic convergence time depends strongly and less strongly on the same respective parameters; in

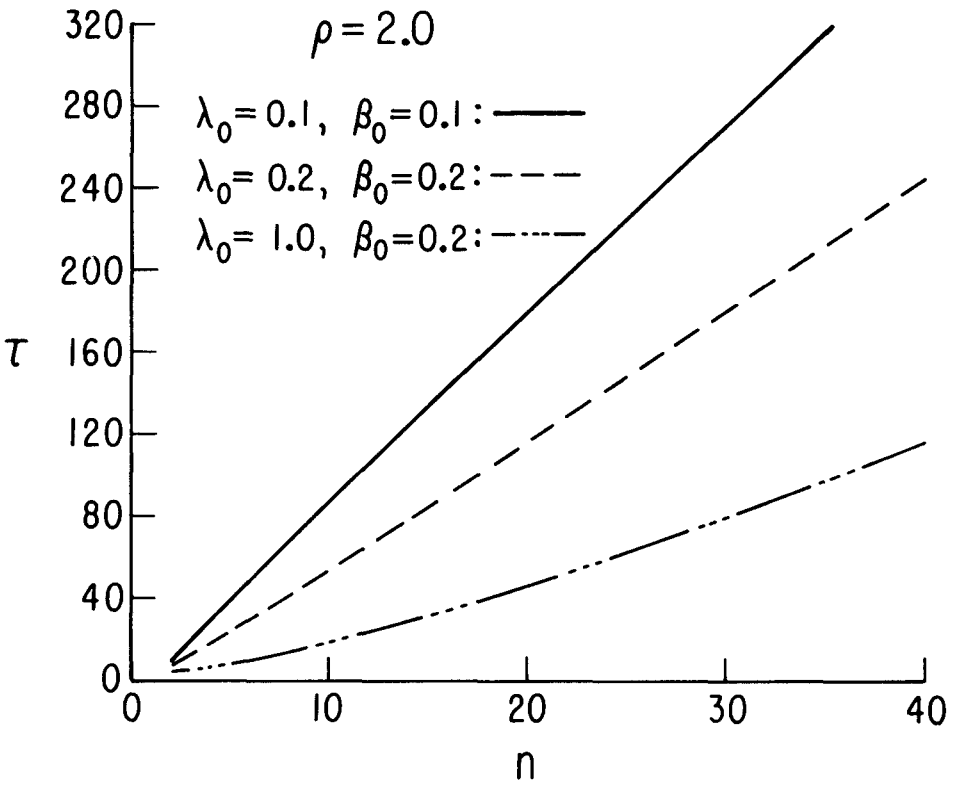


FIGURE 4.—The scaled convergence time as a function of the number of repeats, with λ_0 fixed.

many cases, if mutation is negligible, essentially total sequence and population homogeneity will be attained in an evolutionarily short time.

If there is no mutation, we can easily evaluate the probability that a repeat of type A_i is ultimately fixed. Let $\nu_i(t)$ represent the total number (summed over all loci and all individuals) of A_i alleles just before gametogenesis. Detailed consideration of our model confirms that the expectation of ν_i' conditional on the state of the population in the previous generation is simply ν_i, as expected intuitively. Hence, $\nu_i(t)$ is a martingale, and a straightforward application of the Optional Stopping Theorem (see, *e.g.*, KARLIN and TAYLOR 1975, p. 261) to our finite, absorbing Markov chain proves that A_i is fixed with probability $\nu_i(0)/(2Nn)$, which is precisely its initial frequency (regardless of position) in the population.

In the remainder of this section, we shall discuss extensions of and alternatives to our model.

SZOSTAK *et al.* (1983) have recently proposed a double-strand-break repair model of recombination. A reinterpretation of our parameters α and ρ shows that our formulation and results apply to a model that incorporates asymmetric heteroduplexes (probability γ), symmetric heteroduplexes (probability σ) and double-strand-break repair conversion (probability δ; $\gamma + \sigma + \delta = 1$). To deduce this generalization, we must recalculate p and q, defined by (8) and

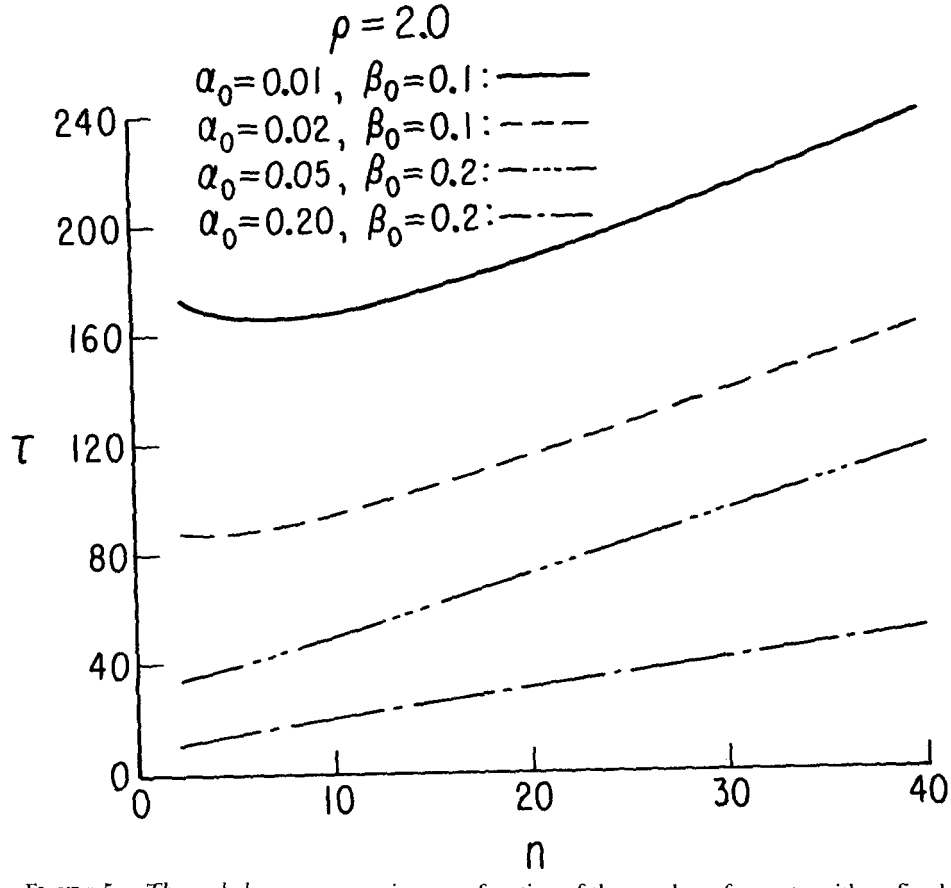


FIGURE 5.—The scaled convergence time as a function of the number of repeats, with α_0 fixed.

(9); we find

$$p = \tfrac{1}{4}(2 - \gamma), \qquad q = \tfrac{1}{4}(1 + \delta). \tag{38}$$

Therefore, instead of (13) we obtain

$$f' = [1 - (n - 1)(2 - \gamma)\alpha]f^{**} + (n - 1)(2 - \gamma)\alpha l^{**}, \tag{39a}$$

$$g' = (1 + \delta)\alpha + [1 - (1 + \delta)\alpha]g^{**}, \tag{39b}$$

$$l' = (2 - \gamma)\alpha f^{**} + [1 - (2 - \gamma)\alpha]l^{**}. \tag{39c}$$

Comparing (39) with (13) reveals that all our expressions hold if we make the replacements

$$\alpha \to \tilde{\alpha} = (1 + \delta)\alpha, \qquad \rho \to \tilde{\rho} = (2 - \gamma)/(1 + \delta). \tag{40}$$

Thus, the effective conversion rate is increased, but not more than twice. Since $1 \leq \tilde{\rho} \leq 2$, the range of this parameter is unaltered.

If the population does not reproduce in the ideal manner of our model (*i.e.*, by sampling from an infinite gametic pool), we must replace everywhere the

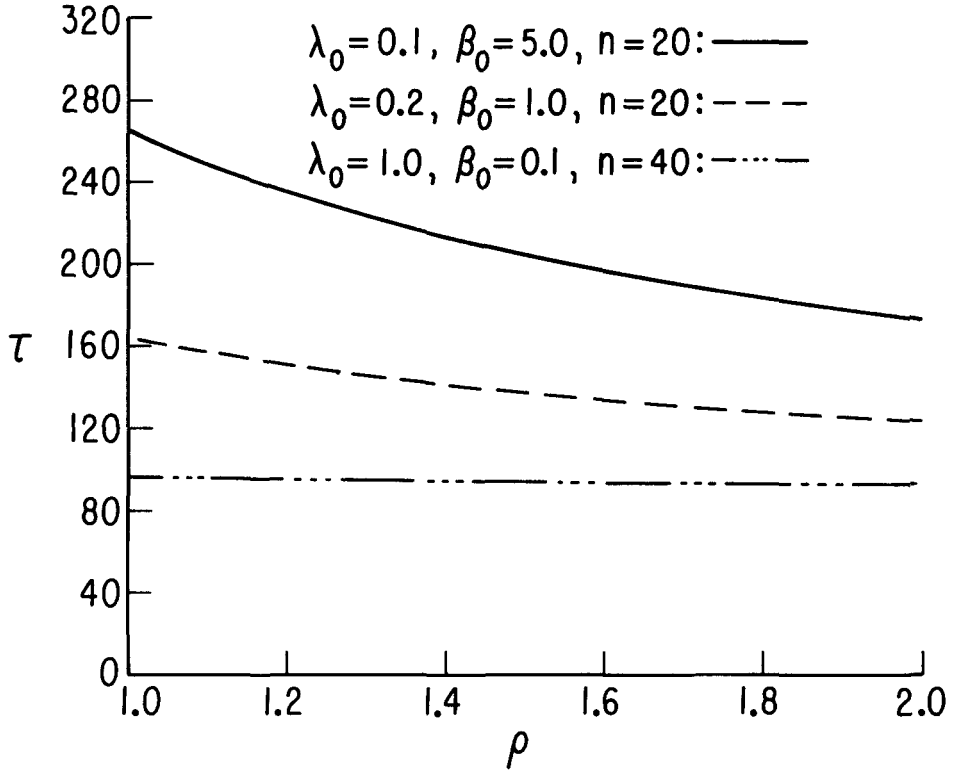


FIGURE 6.—The scaled convergence time as a function of ρ (= 1 + σ, where σ is the proportion of symmetric heteroduplexes).

actual population number, N, by the inbreeding effective population number (CROW and KIMURA 1970, pp. 345–352, 361–364), N_e. Thus, $\theta = 1/(2N_e)$.

We assumed that gene conversion occurs before chromosome duplication. We could equally reasonably posit conversion after chromosome duplication. If we neglect second-order terms that arise because sister chromatids may differ after conversion, then (1) and (3), which describe gametogenesis and mutation, remain unaltered. If μ is still the interaction probability per individual, since conversion now occurs at the four-strand stage, α must be replaced by $\alpha/2$ in (13) and all the ensuing equations.

We have supposed, for the sake of simplicity, that conversion is due solely to intrachromatid interactions. The method of FORMULATION also yields the equations corresponding to (39) for sister-chromatid interactions. In the life cycle of FORMULATION, we assume that chromosome duplication occurs immediately after mutation. As in the previous paragraph, μ is the interaction probability per individual. If b and c are on sister chromatids, instead of (4) and (5) we obtain

$$P(I_b) = \mu/(2n), \qquad P(I_{bc}) = \mu/(2n^2). \tag{41}$$

We again incorporate the model of SZOSTAK *et al.* (1983), so (38) applies.

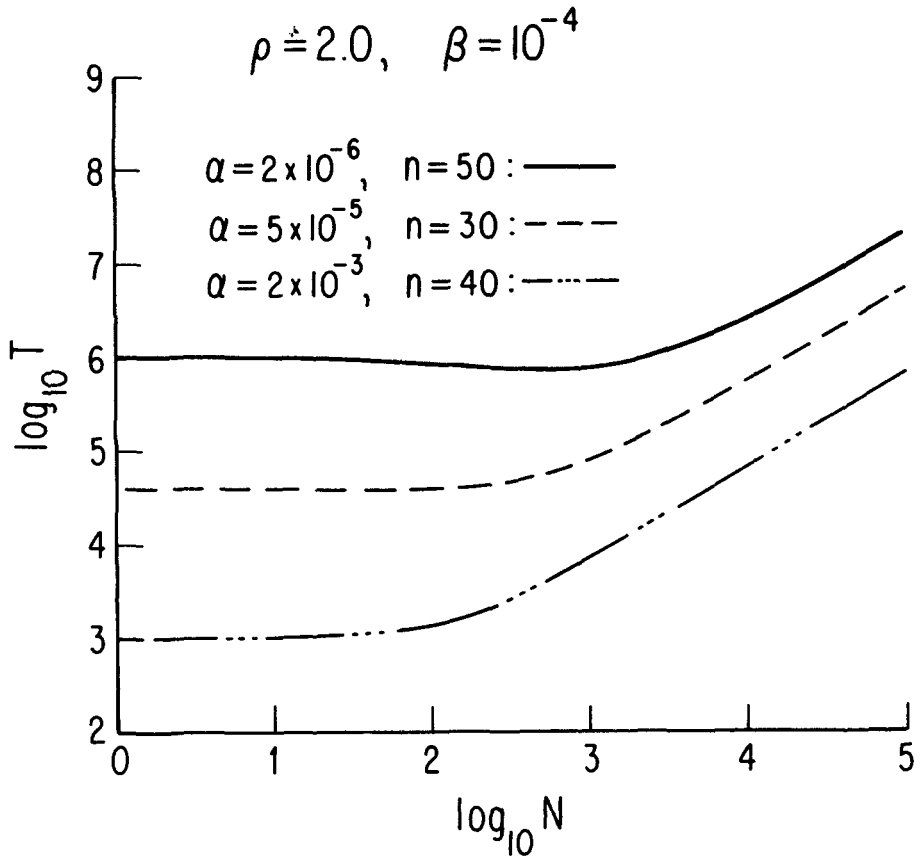


FIGURE 7.—The dependence of the unscaled convergence time on the population number.

Instead of (39), now we find

$$f' = [1 - (n - 1)\bar{\alpha}]f^{**} + (n - 1)\bar{\alpha}l^{**}, \tag{42a}$$

$$g' = \bar{\alpha} + (1 - \bar{\alpha})g^{**}, \tag{42b}$$

$$l' = \bar{\alpha}f^{**} + (1 - \bar{\alpha})l^{**}, \tag{42c}$$

where $\bar{\alpha} = (2 - \gamma)\mu/(4n^2)$. Comparing (42) with (13) informs us that the results for sister-chromatid conversion may be obtained from those for intrachromatid conversion by replacing ρ by 1 and α by $\bar{\alpha}$. Since $\alpha/4 \leqslant \bar{\alpha} < \alpha$, sister-chromatid conversion behaves like intrachromatid conversion with purely asymmetric heteroduplexes and a somewhat lower conversion rate.

We have neglected the dependence of the homologies on the positions of the genes sampled. Since the probability of crossing over increases with separation and g and l differ from each other, this position dependence is absent if, and only if, there are two repeats ($n = 2$) or there is no crossing over ($\beta = 0$). Let the homologies f_i, g_{ij} and l_{ij} refer to distinct loci i and j and denote the corresponding crossover probability by $r_{ij} = \beta|i - j|$. Then for

intrachromatid conversion and $n \geq 3$, the techniques used in FORMULATION lead to the "exact" linearized system

$$f_i' = \theta + [1 - 2u - \theta - (n-1)\rho\alpha]f_i + (n-1)\rho\alpha\bar{l}_i, \tag{43a}$$

$$g_{ij}' = \alpha + \{1 - 2u - [1 + (n-1)\rho]\alpha - r_{ij}\}g_{ij} + \tfrac{1}{2}(n-1)\rho\alpha(\bar{g}_i + \bar{g}_j) + r_{ij}l_{ij}, \tag{43b}$$

$$l_{ij}' = \tfrac{1}{2}\rho\alpha(f_i + f_j) + \theta g_{ij} + (1 - 2u - \theta - n\rho\alpha)l_{ij} + \tfrac{1}{2}(n-1)\rho\alpha(\bar{l}_i + \bar{l}_j), \tag{43c}$$

where

$$\bar{g}_i = \frac{1}{n-1}\sum_{k:k\neq i} g_{ik}, \qquad \bar{l}_i = \frac{1}{n-1}\sum_{k:k\neq i} l_{ik}. \tag{43d}$$

The system (43) is *not* translation invariant, *i.e.*, it does not have a solution of the form $f_i = f$, $g_{ij} = g_{i-j}$, $l_{ij} = l_{i-j}$. Since $\frac{1}{2}n(n+1)$ independent homologies appear in (43), the analytical and numerical investigation of this model is unlikely to be trivial. Studies of unequal crossing over (KIMURA and OHTA 1979; OHTA 1981, 1983b) suggest that (15) may be a reasonable approximation to (43).

Recall that our measure of the characteristic time to genetic homogeneity (in the absence of mutation) is based on the dominant eigenvalue of the recursion relations satisfied by the probabilities of identity. Hence, it is independent of the initial state of the population. A more complete analysis of our stochastic process would yield more information; the mean absorption and mean conditional fixation times would be of particular interest.

Since random drift leads to intralocus homogeneity (in the absence of mutation), in a finite population, interchromosomal gene conversion also produces sequence homogeneity. This process has been analyzed for unbiased conversion; the results will be reported in another publication.

Since biased gene conversion probably has considerable evolutionary importance (LAMB and HELMI 1982; NAGYLAKI and PETES 1982; NAGYLAKI 1983a, b; WALSH 1983), the most important (and difficult) extension of the work presented here is in that direction.

I thank NICHOLAS BARTON, TOMOKO OHTA, THOMAS PETES and BRUCE WALSH for helpful communication and PAUL FORD for the numerical calculations. This work was supported by National Science Foundation grant DEB81-03530.

LITERATURE CITED

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

DOVER, G., 1982 Molecular drive: a cohesive mode of species evolution. Nature **299:** 111–117.

FOGEL, S., R. K. MORTIMER and K. LUSNAK, 1981 Mechanisms of meiotic gene conversion, or "wanderings on a foreign strand." pp. 289–339. In: *The Molecular Biology of the Yeast Saccha-*

romyces: Life Cycle and Inheritance, Edited by J. N. STRATHERN, E. W. JONES and J. R. BROACH. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

GANTMACHER, F. R., 1959 *The Theory of Matrices,* Vol. II. Chelsea, New York.

HOLLIDAY, R., 1964 A mechanism for gene conversion in fungi. Genet. Res. **5:** 282–304.

KARLIN, S. and H. M. TAYLOR, 1975 *A First Course in Stochastic Processes,* Ed. 2. Academic Press, New York.

KIMURA, M., 1963 A probability method for treating inbreeding systems. Biometrics **19:** 1–17.

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

KIMURA, M. and T. OHTA, 1969 The average number of generations until fixation of a mutant gene in a finite population. Genetics **61:** 763–771.

KIMURA, M. and T. OHTA, 1979 Population genetics of multigene family with special reference to decrease of genetic correlation with distance between gene members on a chromosome. Proc. Natl. Acad. Sci. USA **76:** 4001–4005.

LAMB, B. C. and S. HELMI, 1982 The extent to which gene conversion can change allele frequencies in populations. Genet. Res. **39:** 199–127.

MALÉCOT, G., 1946 La consanguinité dans une population limitée. C. R. Acad. Sci. **222:** 841–843.

MALÉCOT, G., 1948 *Les mathématiques de l'hérédité.* Masson, Paris. (Extended translation: *The Mathematics of Heredity.* Freeman, San Francisco, 1969).

MESELSON, M. S. and C. M. RADDING, 1975 A general model for genetic recombination. Proc. Natl. Acad. Sci. USA **72:** 358–361.

NAGYLAKI, T., 1983a Evolution of a large population under gene conversion. Proc. Natl. Acad. Sci. USA **80:** 5941–5945.

NAGYLAKI, T., 1983b Evolution of a finite population under gene conversion. Proc. Natl. Acad. Sci. USA **80:** 6278–6281.

NAGYLAKI, T. and T. D. PETES, 1982 Intrachromosomal gene conversion and the maintenance of sequence homogeneity among repeated genes. Genetics **100:** 315–337.

OHTA, T., 1981 Further study on the genetic correlation between members of a multigene family. Genetics **99:** 555–571.

OHTA, T., 1982 Allelic and nonallelic homology of a supergene family. Proc. Natl. Acad. Sci. USA **79:** 3251–3254.

OHTA, T., 1983a Time until fixation of a mutant belonging to a multigene family. Genet. Res. **41:** 47–55.

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

OHTA, T., 1984 Some models of gene conversion for treating the evolution of multigene families. Genetics **106:** 517–528.

OHTA, T. AND G. DOVER, 1983 Population genetics of multigene families that are dispersed into two or more chromosomes. Proc. Natl. Acad. Sci. USA **80:** 4079–4083.

SZOSTAK, J. W., T. L. ORR-WEAVER, R. J. ROTHSTEIN and F. W. STAHL, 1983 The double-strand-break repair model for recombination. Cell **33:** 25–35.

WALSH, J. B., 1983 Role of biased gene conversion in one-locus neutral theory and genome evolution. Genetics **105:** 461–468.

WRIGHT, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

WRIGHT, S., 1949 Genetics of populations. pp: 111–112. In: *Encyclopaedia Britannica,* Vol 10, Ed. 14. Encyclopaedia Britannica, Chicago.

Corresponding editor: W. J. EWENS