Sequence-Dependent Gene Conversion: Can Duplicated Genes Diverge Fast Enough to Escape Conversion?

J. Bruce Walsh

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 86721

Manuscript received May 22, 1987
Revised copy accepted July 11, 1987

ABSTRACT

Conversion between duplicated genes limits their independent evolution. Models in which conversion frequencies decrease as genes diverge are examined to determine conditions underwhich genes can "escape" further conversion and hence escape from a gene family. A review of results from various recombination systems suggests two classes of sequence-dependence models: (1) the "k-hit" model in which conversion is completely inactivated by a few (k) mutational events, such as the insertion of a mobile element, and (2) more general models where conversion frequency gradually declines as genes diverge through the accumulation of point mutants. Exact analysis of the k-hit model is given and an approximate analysis of a more general sequence-dependent model is developed and verified by computer simulation. If μ is the per nucleotide mutation rate, then neutral duplicated genes diverging through point mutants are likely to escape conversion provided $2\mu/\lambda \gg 0.1$, where λ is the conversion rate between identical genes. If $2\mu/\lambda \ll 0.1$, the expected number of conversions before escape increases exponentially so that, for biological purposes, the genes never escape conversion. For single mutational events sufficient to block further conversions, occurring at rate ν per copy per generation, many conversions are expected if $2\nu/\lambda \ll 1$, while the genes essentially evolve independently if $2\nu/\lambda \gg 1$. Implications of these results for both models of concerted evolution and the evolution of new gene functions via gene duplication are discussed.

MULTIGENE families are common, if not ubiquitous, features of eukaryotic genomes. Sequence exchange processes, such as gene conversion, act to homogenize these families so that individual members are not evolving independently. It is unclear, however, as to how long any given member remains part of a gene family due to gene conversion with other family members. One might imagine that, as genes diverge, the probability of conversion decreases, raising the possibility that duplicated genes may "escape" from conversion.

Besides implications for the structure of multigene families, sequence-dependent gene conversion models have potentially important implications in the formation of new gene functions following a gene duplication. Evolution of new functions via gene duplication is an important concept (*e.g.*, BRIDGES 1935, MULLER 1936, OHNO 1970), but raises several questions. Clearly, some amount of functional divergence must occur between the duplicated copies in order for each to be maintained by selection. Gene conversion may sufficiently retard divergence of duplicated genes to place a major constraint on the rate of new gene formation.

Here we explore the consequences of sequence-dependent gene conversion, focusing on the expected number of conversions between a pair of duplicated genes before they diverge sufficiently to avoid further conversions.

This paper is dedicated to the memory of Larry Sandler.

MOLECULAR DATA ON SEQUENCE-DEPENDENT CONVERSION

Information on the sequence-dependence of conversion comes from three separate sets of data. The most direct are those few studies of conversion between duplicated genes, while less direct information comes from systems examining conversion between different alleles at the same locus and sequence similarity requirements for homologous recombination. Current molecular models (HOLLIDAY 1964; MESELSON and RADDING 1975; SZOSTAK *et al.* 1983) posit that conversion between alleles and reciprocal recombination are both manifestations of the same phenomena. It is unclear if these conversion models are reasonable for events occurring between duplicated genes (FINK and PETES 1984). Often conversions between duplicated genes occurs in the absence of reciprocal recombination (KLEIN and PETES 1981, KLEIN 1984; KLAR and STRATHERN 1984; JACKSON and FINK 1985, JINKS-ROBERTSON and PETES 1985), but there are exceptions (LISKAY, STACHELEK and LETSOU 1984; JINKS-ROBERTSON and PETES 1986; LICHTEN, BORTS and HABER 1987). Recently, SMITHIES has suggested that some conversions between duplicate

genes may be byproducts of processes aligning homologous chromosomes during meiosis (SMITHIES and POWERS 1985; POWERS and SMITHIES 1986). Until the biochemistry of conversion between duplicated genes is better understood, inferences based on conversion between alleles at the same locus, or on homologous recombination, should be viewed as tentative.

The dependence of conversion frequency on amount of sequence similarity potentially acts at (at least) two stages: conversion initiation and the subsequent elongation of conversion tracts. We examine these separately below.

Initiation of conversions: Several different systems examining homologous recombination and gene conversion suggest that recombination events likely initiate at, or near, special sites (reviewed in STAHL 1979; WHITEHOUSE 1982; SZOSTAK *et al.* 1983). If special sites are required for the initiation of conversion, mutational events removing such sites inactivate conversion. The 8 base pair (bp) *Chi* recombination initiation sequence in prokaryotes can be inactivated by single base pair changes (SMITH 1983). Likewise, a single point mutation in a conserved 6-bp sequence greatly reduces recombination during pneumococcal transformation (LEFEVRE *et al.* 1984). Chi-like sequences have been found in eukaryotic systems (KENTER and BIRSHTEIN 1981), and less well characterized hotspots for homologous recombination are known for lower eukaryotes (KEIL and ROEDER 1984; VOELKEL-MEIMAN, KEIL and ROEDER 1987) and mammals (JEFFREYS, WILSON and THEIN 1985; STEINMETZ, STEPHAN and LINDAHL 1986; STEINMETZ, UEMATSU and LINDAHL 1987). Fungal data on conversion between alleles at homologous loci (reviewed in MARKHAM and WHITEHOUSE 1982; WHITEHOUSE 1982, 1983; SZOSTAK *et al.* 1983) also suggest that conversion often (perhaps always) initiates at or near special sites, and proceeds in a polar direction from these sites.

There is suggestive evidence that conversion between duplicated loci may also initiate at specific sites. Intrachromosomal conversion resulting in mating type switching in yeast is initiated by a double-strand cut at a defined site (KOSTRIKEN *et al.* 1983; EGEL, BEACH and KLAR 1984). HESS, SCHMID and SHEN (1984) observed a gradient in sequence divergence in the human $\alpha 1$-$\alpha 2$ globin duplication unit, reminiscent of gradients in conversion between alleles in fungal systems, presumably caused by initiation at a defined site followed by polarity in heteroduplex migration. Stretches of $(TG)_n$ are highly correlated with conversions between primate G_γ and A_γ genes (SLIGHTOM, BLECHL and SMITHIES 1980; SLIGHTOM *et al.* 1985), although conversions can occur away from these sequences. Sequences with the potential to form Z-DNA (such as $(TG)_n$) have been suggested by a number of workers to greatly faculate recombination. Homologous pairing *in vitro* by *Ustilao Rec1* protein (a eukaryotic analogue to *RecA*) is greatly promoted by Z-DNA (KMIEC, ANGELIDES and HOLLOMAN 1985; KMIEC and HOLLOMAN 1986). Palindromic sequences are associated with sites of conversion between immunoglobulin V_H genes (KRAWINKEL, ZOEBELEIN and BOTHWELL 1986), suggesting that other types of DNA structures may also faculate conversions between duplications. Finally, transcription may faculate recombination, possibly by providing a more open chromatin structure (BLACKWELL *et al.* 1986). The most direct example of this is the *HOT1* recombination-stimulator sequence in yeast, which has recently been shown to correspond to polymerase I transcription regulatory sequences (VOELKEL-MEIMAN, KEIL and ROEDER, 1987).

The picture that emerges is there may indeed exist certain sites (either precisely defined sequences or more loosely defined structures of DNA), which are either required for conversion initiation, or at least greatly enhance its rate. If such sites are common and play the dominant role in conversion events between duplicated loci, single mutational events may be sufficient to block conversion by altering (or removing) these sites, preventing initiation.

Even if special sites are necessary for conversion initiation, they are probably not sufficient. Current models of conversion require some stabilization of the initiation complex by sequence homology. In single-stranded conversion models (HOLLIDAY 1964; MESELSON and RADDING 1975) the invading strand(s) must form a stable heteroduplex, while in double-strand-gap-repair models (SZOSTAK *et al.* 1983), repair of the gapped duplex requires donor DNA to be stabilized by flanking regions of homology. This latter requirement is supported by JASIN *et al.* (1985), who found that while the introduction of double-stranded gaps into donor plasmids in COS cells promotes homologous recombination, the highest levels of recombination occurred when gaps are at positions of uninterrupted homology between donor and target DNAs.

Heteroduplex formation, branch migration and lengths of conversion tracts: Studies on how the amount of homology alters homologous recombination provide insight into the relationship between homology and heteroduplex formation, and hence migration of conversion tracts. The basic theme from prokaryotic studies is that the frequency of homologous recombination seems to be linearly related to the amount of sequence similarity. Phage T4 requires a minimum sequence of 50 bp of homology for recombination, above this length, recombination increases linearly with homology (SINGER *et al.* 1982). This same pattern—a minimal required sequence length, with a linear increase in recombination with homology above

that length—is also seen in *Escherichia coli* (WATT *et al.* 1985; SHEN and HUANG 1986). The caveat with these experiments is that blocks of exact homology were used, rather than blocks of less than perfect homology. An unresolved issue is whether the minimal length sequence required must have exact homology, or whether imperfect homology is acceptable.

The limited data for mammalian systems generally support a linear relationship between amount of homology and rate of homologous recombination. AYARES *et al.* (1986), studying recombination between plasmids, observed a linear relationship between recombination frequency and sequence similarity in COS cells, but observed a biphasic curve when using human EJ cells. RUBNITZ and SUBRAMANI (1984) also found a biphasic curve relating length of sequence homology and recombination. The most direct study is that of LISKAY, LETSOU and STACHELEK (1987) who examined sequence requirements for mitotic conversions between duplicated genes in mouse L cells, and found an approximate linear decrease in conversion frequency as sequence similarity decreases.

In addition to nonhomologies being introduced by numerous point mutations, deletion and insertion events can produce, in a single event, a large region of nonhomology. The importance of such regions in altering levels of recombination and conversion is unclear. Large blocks of nonhomology can be included in stable heteroduplexes—LICHTEN and FOX (1984) recovered phage λ heteroduplexes containing insertions/deletions of 700 and 1300 bp. *In vitro* studies show that *RecA* can pair DNAs showing limited stretches of homologies, but at a slower rate than regions with perfect homology (GONDA and RADDING 1983; BIANCHI and RADDING 1983). However, several observations suggest that regions of nonhomology can, in some cases, prevent conversion between duplicated genes. Three independent studies suggest that insertion of *Alu* elements prevents conversions between duplicated globin genes (HESS *et al.* 1983; MICHELSON and ORKIN 1983; SCHIMENTI and DUNCAN 1984). One explanation of this is conversion initiates at (or near) particular sites, conversion tracts proceed outward from these sites in a polar fashion, with heteroduplex migration blocked by the nonhomology introduced by the insertion. The assumed polarity in direction of conversion tracts is consistent with both the finding of a gradient in sequence conversion in between adult globin genes (HESS, SCHMID and SHEN 1984) and the properties of characterized eukaryotic recombination enzymes (*Rec1* from the fungus *Ustlago* and a recently purified recombinase from human B lymphoblasts) which exhibit polarity in strand displacement (HSIEH, MEYN and CAMERINI-OTHERO 1986; KMIEC and HOLLOMAN 1986).

Small insertions or deletions do not always act as a barrier to conversion between duplicated genes (MICHELSON and ORKIN 1983; HILL *et al.* 1985). Double-strand gap repair models allow for the removal of blocks of nonhomology before heteroduplex formation. BRENNER, SMIGOCK and CAMERINI-OTHERO (1986) estimate that human L cells can generate double stranded gaps of between 150 and 250 bases, allowing for the possibility that regions of nonhomology this size may not be an important block to conversion.

Another important modeling issue, the average length of a conversion tract, appears to be highly system dependent. LISKAY and STACHELEK (1986) found mitotic conversion tracts between 1.2-kb duplicated elements could be as long as 360 bases. Conversion tracks between duplicated *Trypanosoma brucei* VSG genes average about 3.5 kb, but can be much smaller (PAYS *et al.* 1985). Mammalian globin genes often have conversion tracts in the neighborhood of 1.2–1.5 kb (SLIGHTOM, BLECHL and SMITHIES 1980; ERHART, SIMONS and WEAVER 1985; SLIGHTOM *et al.* 1985), but a detailed study of the human G_γ and A_γ fetal globin genes found evidence for numerous very small conversion tracts, often on the order of tens of base pairs, or less (SMITHIES and POWERS 1985; POWERS and SMITHIES 1986). Very short conversion tracts are also seen in chicken V_λ genes (10–120 nucleotides; REYNAUD *et al.* 1987), class I MHC genes (10–50 nucleotides; FLAVELL *et al.* 1986) and class II MHC genes (≈20 nucleotides; MCINTYRE and SEIDMAN 1984; GORKSI and MACH 1986). It is highly likely that several different pathways produce conversion-like events. The very short conversion tracts seen in both globin and immune genes may result from different pathways than the longer conversion tracts seen between other duplicated genes. Indeed, short conversion tracts between duplicated immune genes play an important role in generating diversity for the immune response, suggesting that the structure of conversion processes in these regions has been under strong selection (BORST and GREAVES 1987). An additional complication is introduced by the finding that the length of a conversion tract may be significantly increased by recombination-stimulating sequences (VOELKEL-MEIMAN, KEIL and ROEDER 1987).

The data reviewed above suggest two different classes of sequence-dependent conversion models. First, a single (or few) mutational events might irreversibly block conversion. The insertion of a mobile element downstream from a polar initiation site, or the deletion of an initiation site, are possible examples. Second, both prokaryotic and mammalian data suggest that conversion rates decline roughly linearly with sequence similarity. In this case, conversion rates gradually decline as duplicated genes accumulate point mutants.

"k-HIT" MODELS OF SEQUENCE-DEPENDENT CONVERSION

As discussed above, it may be biologically reasonable to suppose that conversion between duplicated genes can be completely stopped by one, or a few, mutational events, such as the insertion of an *Alu* element. Here we assume that k mutants are sufficient to completely block conversion. Note that these "mutants" are generally not point mutants, but rather may be larger events, such as insertions or deletions. The effects on conversion rates from the gradual accumulation of point mutants are examined in the next section and are ignored here.

The model follows a single chromosomal lineage through time. At time zero, a gene duplication occurs, with copy number subsequently stable. Intrachromosomal conversion occurs between the two copies, provided that k specified mutational events have not occurred. Each of the k mutants occur independently with mutation rate ν per generation per gene, implying the waiting time for k such mutants to appear follows a gamma distribution with parameters k and 2ν. If $k-1$ or fewer mutants have occurred, conversion occurs at rate λ per generation, else if k or more mutants have occurred since the last conversion event, conversion is completely stopped. When conversion occurs, it homogenizes both copies, and the mutation accumulation process starts over again. Implicit in these assumptions (for $k > 1$) is that conversion tracts cover the mutant sites.

Let π be the probability that the duplicated pair escapes conversion on a given trail (that is, it accumulates k mutants before a conversion event occurs). Under the above assumption of restarting after each conversion event, the probability that exactly i conversion events occur follows a geometric distribution, and is given by $(1-\pi)^i\pi$. Likewise, the expected number of conversions is $(1-\pi)/\pi$ and the variance in number of conversions is $(1-\pi)/\pi^2$.

We obtain π as follows: let t be the time until a conversion event occurs (in the absence of mutation) and τ the time until k mutants arise (in the absence of any conversion). Denote $\Xi = (1-\pi) = \text{Prob}(t < \tau)$, the probability that a conversion event occurs; $p(t = x)$ the probability density that $t = x$; and $P(\tau > x)$ the probability that $\tau > x$. Then,

$$\Xi = \int_0^\infty p(t = x)P(\tau > x)dx. \tag{1}$$

t follows an exponential distribution with mean λ, and τ follows a gamma distribution with parameters $(k, 2\nu)$ giving:

$$\Xi = \int_0^\infty \lambda e^{-\lambda x} \int_x^\infty [2\nu/(k-1)!](2\nu z)^{k-1}e^{2\nu z}dzdx. \tag{2}$$

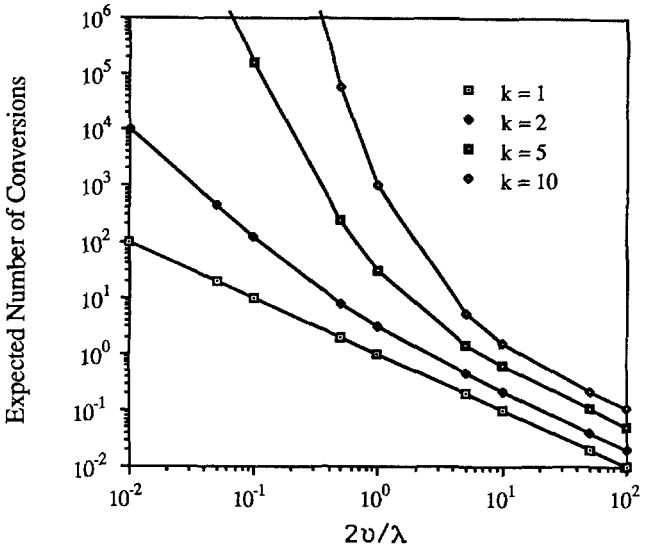

FIGURE 1.—Expected number of conversions between a pair of duplicated genes, assuming the k-hit model. k is the number of sites which must be mutated, λ the conversion rate and ν the per site mutation rate.

Changing the order of integration,

$$\Xi = \int_0^\infty [2\nu/(k-1)!](2\nu z)^{k-1}e^{-2\nu z} \cdot \int_0^z \lambda e^{-\lambda x}dxdz, \tag{3}$$

giving:

$$\Xi = \int_0^\infty [2\nu/(k-1)!](2\nu z)^{k-1}e^{-2\nu z}(1-e^{-z\lambda})dz, \tag{4}$$

which simplifies using the gamma function to

$$\Xi = 1 - [2\nu/(2\nu+\lambda)]^k \tag{5}$$

or,

$$\pi = [2\nu/(2\nu+\lambda)]^k. \tag{6}$$

Let $E(n)$ be the expected number of conversions. Recalling $E(n) = (1-\pi)/\pi$, from (6):

$$E(n) = (1 + (\lambda/2\nu))^k - 1 \tag{7a}$$

$$= \lambda/2\nu \quad \text{for} \quad k = 1. \tag{7b}$$

Since $(1+x)^k - 1 = kx + 0(x^2)$, (7) implies few conversions occur if $k\lambda/2 < \nu$, while multiple conversions occur if $k\lambda/2 > \nu$. This is seen in Figure 1, which plots $E(n)$ for various values of k and $2\nu/\lambda$. Mitotic conversion rates between mammalian duplicated genes are typically in the range of 10^{-5} to 10^{-7} (RUBNITZ and SUBRAMANI 1986; LISKAY, LETSOU and STACHELEK 1987), while higher rates (both mitotic and meiotic) can occur in fungal systems (JACKSON and FINK 1985; JINKS-ROBERTSON and PETES 1986). If the frequency of inactivating mutational events is comparable to the frequency of point mutants, *i.e.*, $2\nu \approx 2(5 \times 10^{-9})$ (based on substitution rates in pseudo-

genes, KIMURA 1983; LI, LUO and WU 1985; however, rates may be higher, e.g., LI and TANIMURA 1987), the expected number of conversions is ≥ 10 (assuming $k = 1$ and $\lambda = 10^{-7}$). If multiple bases/sites each have to be changed by single mutants (*i.e.*, $k > 1$), or if conversion rates exceed 10^{-7}, the expected number of conversions greatly exceeds this.

Another useful measure of constraint is the expected number of generations from the initial duplication until the last conversion event occurs. This is simply $E(n)E(T)$, where $E(T)$ is the expected time between conversions, *conditional* on a conversion occurring. From the same arguments leading to (1), the density of conditional conversion times is given by $p(t = x)P(T > x)/\Xi$, yielding:

$$E(T) = (1/\Xi) \int_0^\infty x\lambda e^{-\lambda x} \cdot \int_x^\infty [2v/(k-1)!](2vz)^{k-1}e^{-2vz}dzdx \tag{8}$$

$$= (1/\Xi) \int_0^\infty [2v/(k-1)!](2vz)^{k-1}e^{-2vz} \cdot \int_0^z x\lambda e^{-\lambda x}dxdz. \tag{9}$$

Upon evaluation of the inner integral and subsequent use of the gamma function,

$$E(T) = (1/\Xi)\{(1/\lambda)[1 - (2v/(2v + \lambda))^k] - [k/(2v + \lambda)][2v/(2v + \lambda)]^k\} \tag{10}$$

which reduces to

$$E(T) = 1/\lambda - [k/(2v + \lambda)][\pi/(1 - \pi)]. \tag{11}$$

The first term in (11) is the unconditional mean time, which is an overestimate since we condition on a conversion occurring, resulting in a correction term. The expected number of generations at which the last conversion event occurs following the original duplication is then:

$$E(n)E(T) = E(n)/\lambda - k/(2v + \lambda) \tag{12a}$$

$$= (1/\lambda)[E(n) - k/(1 + 2v/\lambda)]. \tag{12b}$$

Figure 2 plots the expected time constrained in λ^{-1} generations for various values of k and $2v/\lambda$. For $k = 1$, (12) simplifies to $E(n)/(2v + \lambda) = (1/\lambda)(\theta/(1 + \theta^{-1}))$, where $\theta = \lambda/2v = E(n)$.

MORE GENERAL SEQUENCE-DEPENDENT CONVERSION FUNCTIONS

Consider the more general situation where the conversion rate between duplicated genes can be written as $\lambda f(x)$, with x the fraction of sequence similarity between the two genes. (Note that this is not the most general formulation—if conversion is strictly dependent on the size of blocks of *exact* homology, knowledge of simply the amount of sequence similarity is insufficient.) Unlike the previous model, conversion is not all or none, rather declines steadily as sequence similarity declines. Impose $f(1) = 1$, so that λ is the conversion rate between two identical sequences. Clearly, λ varies for different duplications, depending on local chromatin configuration, the distance separating genes, etc.

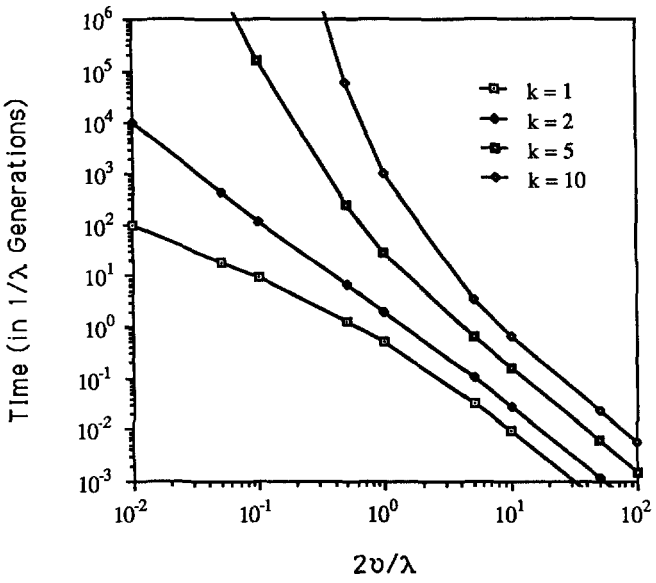

FIGURE 2.—Expected time (in $1/\lambda$ generations) following the initial gene duplication until the last conversion occurs for the k-hit model.

The basic structure of this model is similar to the previous k-hit model: a single chromosome lineage, which underwent a single duplication event at time zero, is followed. Intrachromosomal conversion events are assumed to convert an entire gene, restarting the process. As above, this assumption implies the total number of conversions follows a geometric distribution, with parameter π, the probability that the pair diverges sufficiently on any given trail to avoid further conversions. To compute π, consider $\phi(t)$, the probability that no conversion occurs in the first t generations after a duplication (or a conversion event), and define π as the limit of $\phi(t)$ as $t \to \infty$. Observe that conversion is a time-dependent Poisson process, with the time-dependent variable being random. Thus,

$$\text{Prob}[\text{conversion } \epsilon\ (t, t + \Delta t) \mid x_t] = \lambda f(x_t)\Delta t + o(\Delta t). \tag{13}$$

Conditioning on the sample path for x_t, we have a standard time-dependent process (COX and MILLER 1965, pp. 153–154):

$$\text{Prob[no conversion } \epsilon\ (0, t,) | x_s \quad \text{for} \quad s \leq t] = \left[\exp\left(-\lambda\left\{\int_0^t f(x_s)ds\right\}\right)\right] \tag{14a}$$

Taking the expectation over all sample baths of the process x_s in $(0, t)$ gives:

$$\phi(t) = E\left[\exp\left(-\lambda\left\{\int_0^t f(x_s)ds\right\}\right)\right] \tag{14b}$$

(14b) is also the expression for $\phi(t)$ under the more general formulation, where x_s is a markov process describing the divergence of two duplicated genes. This allows for modeling conversion based on the length distribution of blocks of exact homology between two sequences, as well as other general measures of sequence similarity.

Given the approximate linear relationship between rates of homologous recombination and amount of sequence similarity in both prokaryotic and mammalian systems, we first consider a linear conversion function:

$$\begin{aligned} f(x) &= (x - x_{\text{crit}})/(1 - x_{\text{crit}}) \quad \text{for} \quad x \geq x_{\text{crit}} \\ &= 0 \quad \text{for} \quad x \leq x_{\text{crit}} \end{aligned} \tag{15}$$

$f(1) = 1$ as prescribed, and conversion does not occur when sequences show less than x_{crit} percent similarity.

Another candidate conversion function is a quadratic. Our motivation is the observation of AYARES *et al.* (1986) of a biphasic relationship (a fairly smooth linear decrease changing into a very sharp dropoff) between the frequency of homologous recombination and length of sequence homology in certain mammalian cells lines. Consider:

$$\begin{aligned} f(x) &= 1 - (1 - x)^2/(1 - x_{\text{crit}})^2 \quad \text{for} \quad x \geq x_{\text{crit}} \\ &= 0 \quad \text{for} \quad x \leq x_{\text{crit}}. \end{aligned} \tag{16}$$

As above, conversion ceases when the genes show less that x_{crit} percent similarity. Note that (16) falls off more slowly that (15), given the same value for x_{crit}.

A deterministic approximation for π: Our general expression for π is:

$$\pi = E[\exp(-\Lambda)], \qquad \Lambda = \lambda \int_0^\infty f(x_t)dt. \tag{17}$$

The expectation in (17) being over all possible paths for the random variable x_t, the fraction of identical bases shared by the two duplicated genes at time t. In practice, (17) is very difficult to evaluate directly, however if a diffusion can be associated with x_t, the method of KAC functionals can be used (see the APPENDIX). In what follows, we compute π using a deterministic approximation of (17), which computer simulations show gives excellent results for the forms of $f(x)$ examined here.

The deterministic approximation is motivated by considering a long DNA sequence. For such a sequence, the stochastic process x_t is expected to follow fairly closely the deterministic trajectory, given that x_t is the average over all bases. For example, consider a gene of length n bases, where each base diverges by an independent, identically distributed process. If p_t is the probability a given base (at time t) is identical in the duplicated genes, then $E[x_t] = p_t$, and $\text{Var}[x_t] = p_t(1 - p_t)/n$. For large n, the variance in the process is small and paths closely follow the deterministic trajectory. We approximate π by $\exp(-\Lambda^*)$, with

$$\Lambda^* = \lambda \int_0^\infty f\{z(t)\}dt \tag{18}$$

and $z(t) = E(x_t)$, the expected amount of similarity at time t. Thus, we replace the expectation over all sample paths by the integral of the deterministic process. The validity of this approximation, as examined by simulations, is discussed below.

Before proceeding, a technical issue needs to be clarified. Consider the integral in (17) as x_t approaches its stationary distribution, x. For any biologically reasonable form of mutation, the distribution of x has support on (0, 1), and thus has some nonzero probability of being above x_{crit}, implying that the integral is unbounded, and hence $\pi = 0$ (*i.e.*, conversion always occurs). However, the expected waiting time for a conversion once stationarity is approached is likely to be on such a large time scale as to be biologically meaningless. Assuming at equilibrium that all bases are independent and each has probability ¼ of showing similarity, the DeMoivre-Laplace theorem states that the probability two sequences of length n show more than x_{crit} similarity is approximately given by the area under a normal distribution above the value 4 $(x_{\text{crit}} - \frac{1}{4})\sqrt{[n/3]}$. For $n = 1000$, and $x_{\text{crit}} = 0.4$, this probability is approximately 10^{-27} (using standard approximations for large normal values, ABRAMOWITZ and STEGUN 1964, Eq. 26.2.12).

It remains to specify x_t, the amount of similarity between two sequences, which were identical at time $t = 0$. Assuming selection is not acting on either duplicated gene, standard results for divergence of neutral genes (JUKES and CANTOR 1969) give the expected similarity at time t as,

$$z(t) = (\tfrac{1}{4})(1 + 3\exp(-\theta t)), \tag{19}$$

where $\theta = (\tfrac{4}{3})(2\mu)$. This assumes that mutation occurs at rate μ per base per gene, and that each base mutates to the other three at equal rates.

While we focus on the strict neutral case, selection on one gene reduces the rate of divergence. In the most extreme case (no change in one gene), the divergence is equivalent to assuming divergence of two neutral genes each with mutation rate $\mu/2$. Strong positive selection may increase the joint divergence over the strict neutral value—this has generally been

viewed as highly unlikely (KIMURA 1983), but apparent examples are beginning to appear (LEIGH BROWN 1987; HILL and HASTIE 1987).

Consider (16), the linear conversion function, first. Under the deterministic process, no conversion occurs after x_{crit} is reached. The upper limit of the integral in (18) is given by τ, which satisfies $x_{crit} = (1/4)(1 + 3\exp(-\theta\tau))$. Substitution of (19) into (18), with the integral truncated at τ gives:

$$\Lambda^* = [\lambda/(1 - x_{crit})][\tau(1/4 - x_{crit}) + (3/4)(1/\theta)(1 - \exp(-\theta\tau))]. \quad (20)$$

Using $\tau = (-1/\theta)\ln[(4/3)(x_{crit} - 1/4)]$, and the definition of θ, (20) reduces to

$$\Lambda^* = (\lambda/2\mu)(3/4)[1 + (x_{crit} - 1/4)/(1 - x_{crit})\ln\{(4/3)(x_{crit} - 1/4)\}]. \quad (21)$$

Thus, out deterministic approximation for the linear conversion function (15) is

$$\pi = \exp\{-(\lambda/2\mu)h(x_{crit})\},$$
$$h(x) = (3/4[1 + [(x - 1/4)/(1 - x)] \cdot \ln\{(4/3)(x - 1/4)\}]. \quad (22)$$

Proceeding in the same fashion for the quadratic conversion function (16) gives:

$$\pi = \exp\{-(\lambda/2\mu)g(x_{crit})\},$$
$$g(x) = (3/4)[1/2 + (3/4)/(1 - x) - (1 - (3/4)^2/(1 - x)^2)\ln\{(4/3)(x - 1/4)\}]. \quad (23)$$

Figure 3 plots π, the probability of escaping conversion, for both the linear and quadratic conversion function, assuming x_{crit} values of 0.8, 0.6 and 0.4. Data are lacking on x_{crit} values in nature, but they are likely in this range.

Given π, the expected number of conversions is, as above, $(1 - \pi)/\pi$. The validity of our deterministic approximation for π was checked by comparing the expected number of conversions using (22) and (23) with the mean number of conversions in a series of simulations, assuming gene sequences of length 100 and 1000. Details of the simulation proceedure are given in the APPENDIX. As seen in Tables 1 and 2, and in Figures 4 and 5, the approximation is excellent. Further, the simulation shows no difference in mean number of conversions for gene sequences of length 100 and 1000 (Figures 4B and 5B), and the variance in number of conversions is also comparable (Tables 1 and 2).

Both (22) and (23) demonstrate that π depends strongly on the ratio $2\mu/\lambda$, so that the relevant parameters to compare are the conversion rate and the *per nucleotide* mutation rate. Given a per nucleotide mutation rate of 5×10^{-9} and conversion rates of $\lambda = 10^{-7}$ to $\lambda = 10^{-5}$ gives $2\mu/\lambda$ values ranging from 0.1 to 0.001. As seen in Figures 4 and 5, if $2\mu/\lambda \leq 0.01$ (i.e. $\lambda \geq 10^{-6}$), regardless of the form of conversion function or value of x_{crit}, the expected number of conversions is very large. Conversion rates in excess of 10^{-6} are sufficient to prevent members of a gene family from "escaping" one another, unless conversion falls off very sharply (*i.e.*, x_{crit} close to 1) or significant divergence occurs by mutational event other than accumulation of point mutants. In this later realm, the *k*-hit model may be more important. Likewise, if conversion rates are low (*i.e.*, $\lambda < 10^{-7}$), the expected number of conversions between a pair of duplicated genes is small, and members of a gene family are likely to escape conversion and start to evolve independently.

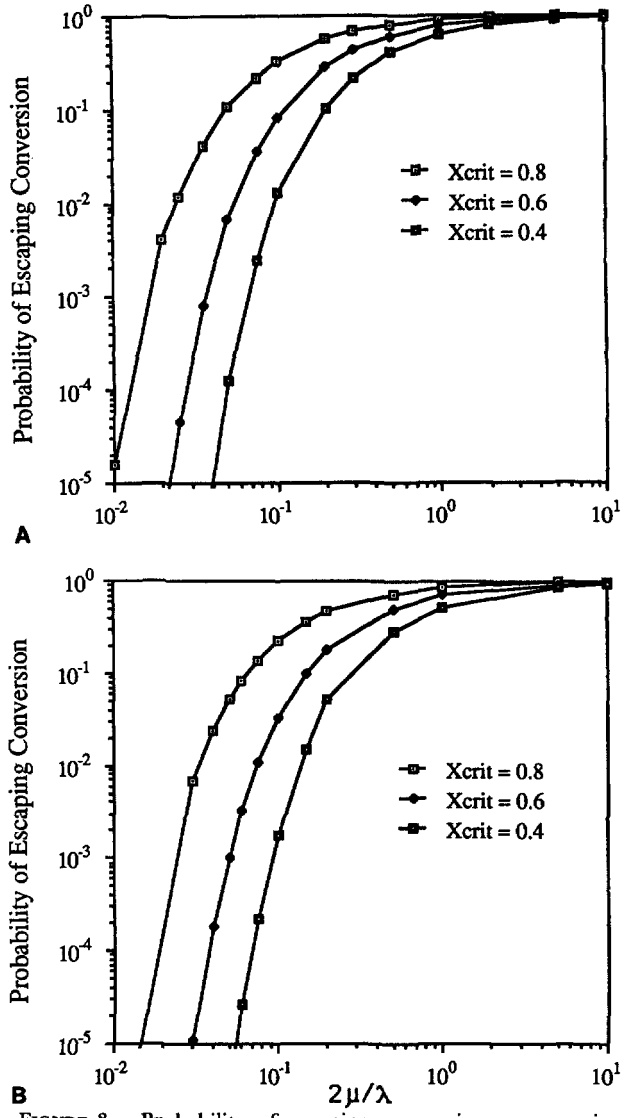

FIGURE 3.—Probability of escaping conversion on any given trail, π, for linear (A) (15) and quadratic (B) (16) conversion functions, $f(x)$. Sequences showing similarity of x_{crit} or less do not undergo conversion, λ is the conversion rate between identical sequences and μ is the per nucleotide mutation rate. Both graphs use the deterministic approximations for π given by (22) and (23) for the linear, and quadratic, conversion functions (respectively).

TABLE 1

Simulation results for linear conversion function mean number of conversions

	$X_{critical} = 0.8$		$X_{critical} = 0.6$		$X_{critical} = 0.4$	
$2\mu/\lambda$	$n = 1000$	$n = 100$	$n = 1000$	$n = 100$	$n = 1000$	$n = 100$
0.025	Expected = 81.5		Expected = 2.19×10^4		Expected = 6.12×10^7	
	89.8	63.4	>5000	>5000	>5000	>5000
	(7.7×10^4)	(1.5×10^4)				
0.05	Expected = 8.08		Expected = 147		Expected = 7.82×10^3	
	10.7	7.98	124.4	135.7	>5000	>5000
	(144)	(62.9)	(1.4×10^4)	(1.5×10^4)		
0.1	Expected = 2.01		Expected = 11.1		Expected = 87.5	
	1.7	2.6	11.5	10.9	89.1	74.6
	(4.02)	(7.12)	(93.2)	(162)	(4.5×10^3)	(6.8×10^3)
0.5	Expected = 0.25		Expected = 0.65		Expected = 1.45	
	0.21	0.31	0.62	0.64	1.52	1.30
	(0.24)	(0.45)	(0.76)	(1.10)	(5.25)	(2.97)
1.0	Expected = 0.12		Expected = 0.28		Expected = 0.57	
	0.14	0.13	0.44	0.28	0.53	0.53
	(0.17)	(0.13)	(0.49)	(0.34)	(0.87)	(0.98)
5.0	Expected = 0.022		Expected = 0.051		Expected = 0.094	
	0.016	0.016	0.056	0.060	0.088	0.120
	(0.016)	(0.016)	(0.053)	(0.056)	(0.080)	(0.162)

Number of conversions before duplicated genes diverge sufficiently to avoid further conversions, assuming a linear conversion function for $f(x)$ (15). Expected number of conversions is the deterministic approximation given by (22), while simulation results assume two different gene sizes: 100 and 1000 nucleotides per gene. The variance in number of conversions per trial is given in parenthesis below the mean observed value for each trial. For expected number of conversions >5, 50 trials per parameter set were used, while if <5, 250 trials used. Runs mean denoted by >5000 had most (all) of the trials exceed 5000 conversions (the cutoff point in the simulation program). Further simulation details given in APPENDIX.

TABLE 2

Simulation results for quadratic conversion function mean number of conversions

	$X_{critical} = 0.8$		$X_{critical} = 0.6$		$X_{critical} = 0.4$	
$2\mu/\lambda$	$n = 1000$	$n = 100$	$n = 1000$	$n = 100$	$n = 1000$	$n = 100$
0.025	Expected = 385		Expected = 9.20×10^5		Expected = 1.01×10^{11}	
	353	178	>5000	>5000	>5000	>5000
	(1.5×10^5)	(1.5×10^4)				
0.05	Expected = 18.7		Expected = 958		Expected = 3.18×10^5	
	17.4	13.3	762	702	>5000	>5000
	(470)	(199)	(4.1×10^5)	(3.9×10^5)		
0.1	Expected = 3.44		Expected = 30.0		Expected = 563	
	3.20	4.10	29.5	31.0	645	384
	(12.1)	(28.1)	(1.8×10^3)	(1.1×10^3)	(3.3×10^5)	(1.8×10^5)
0.5	Expected = 0.35		Expected = 0.99		Expected = 2.55	
	0.32	0.34	1.07	1.00	2.64	2.50
	(0.55)	(0.54)	(1.96)	(1.88)	(11.5)	(9.40)
1.0	Expected = 0.16		Expected = 0.41		Expected = 0.88	
	0.21	0.18	0.49	0.41	0.79	0.99
	(0.25)	(0.25)	(0.61)	(0.74)	(1.37)	(2.04)
5.0	Expected = 0.30		Expected = 0.071		Expected = 0.135	
	0.028	0.028	0.080	0.100	0.156	0.148
	(0.035)	(0.027)	(0.082)	(0.098)	(0.172)	(0.150)

Number of conversions before genes escape conversion, assuming a quadratic conversion function (16). Simulation results are compared to deterministic approximation given by (23). Other details the same as for Table 1.

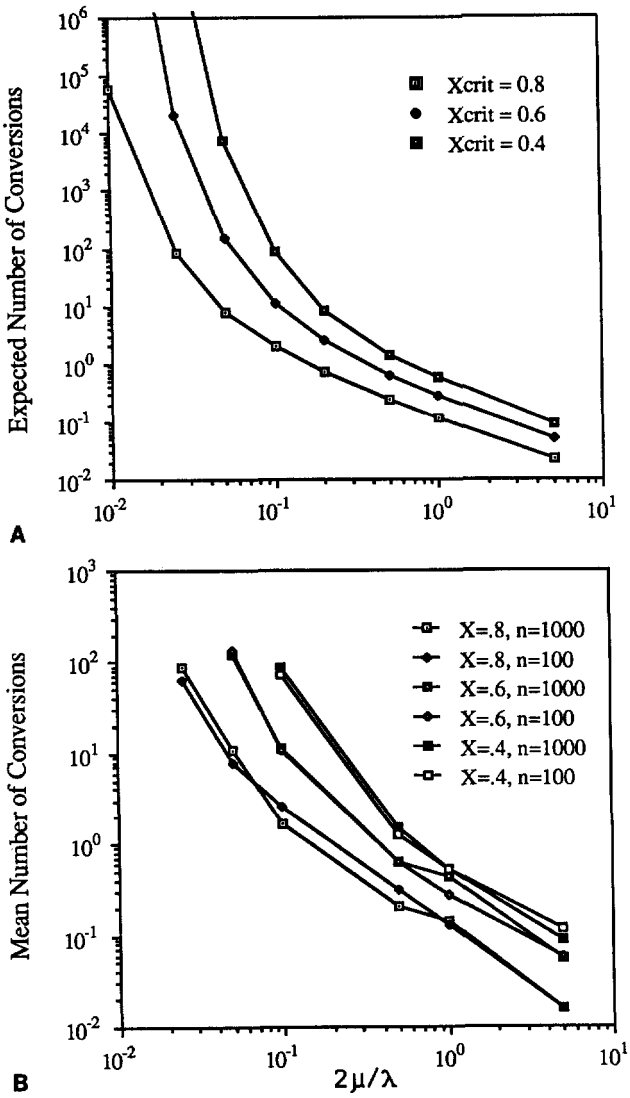

FIGURE 4.—Expected number of conversions between a pair of duplicated genes, assuming a linear conversion function (15). X_{crit} and x both denote the amount of sequence similarity below which conversion does not occur. Predicted results using the deterministic approximation (22) are shown in (A), while (B) plots the mean number of conversions per trail from a simulation assuming a gene size of $n = 100$ or $n = 1000$ nucleotides. Note in (B) that the mean number of conversions appears independent of gene size and agrees well with the approximation given by (22). Also see Table 1.

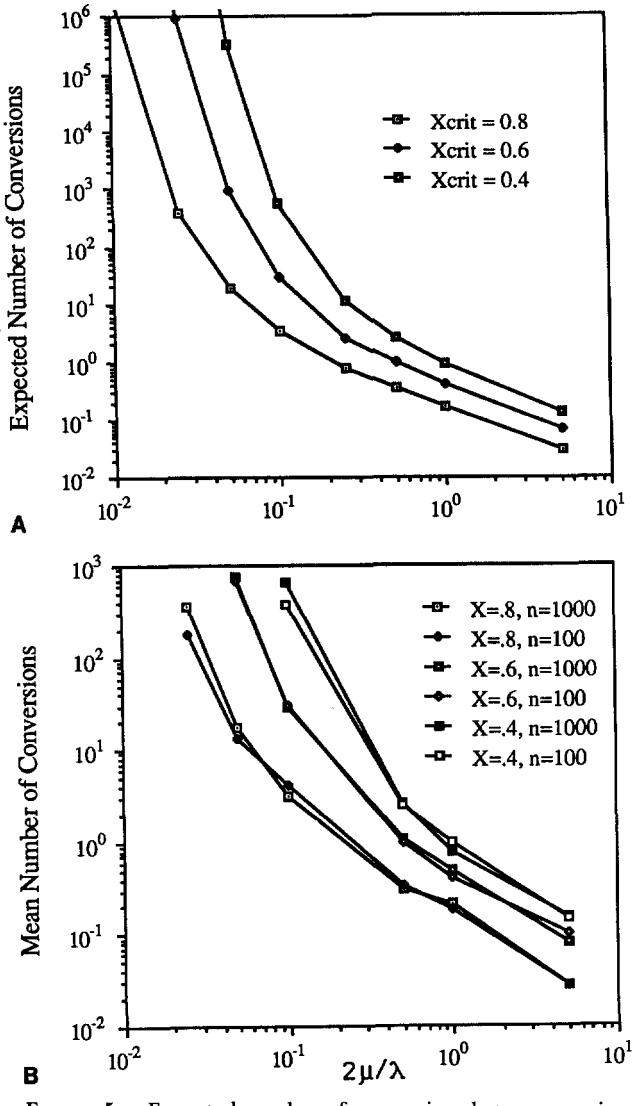

FIGURE 5.—Expected number of conversions between a pair of duplicated genes, assuming a quadratic conversion function (16). X_{crit} and x both denote the amount of sequence similarity below which conversion does not occur. Predicted results using the deterministic approximation (23) are graphed in (A), while (B) plots mean number of conversions per trail from a simulation assuming a gene size of $n = 100$ or $n = 1000$ nucleotides. Agreement with the approximation is excellent. Also see Table 2.

DISCUSSION

Exchange of information between duplicated genes by conversion constrains the independent evolution of each copy, a phenomenon called concerted evolution. The amount of similarity between members of multigene families induced by concerted evolution has been the subject of considerable population genetics modeling (reviewed in OHTA 1980, 1983; NAGYLAKI 1984a,b). These models assume conversion rates are independent of the amount of sequence similarity. As a first step toward extension of these models, we examined conditions under which genes can diverge sufficiently to avoid further conversion, effectively removing these copies from the rest of the gene family. Our analysis is initially restricted to selectively neutral genes.

Two different classes of sequence-dependence are suggested by a review of the genetic and molecular data on gene conversion. First, a single or a few events may be sufficient to prevent further conversions. This is modeled by the "k-hit" process, wherein conversion occurs at rate λ between duplicated genes unless k mutational events have occurred since the last conversion, each mutation occurring with rate ν per gene. These mutational events are likely unusual, such as insertion of *Alu* elements or deletion of conversion initiation regions. Another way to model sequence-dependent conversion is to assume that as genes diverge through the accumulation of point mutations

(occurring at rate μ per base pair per generation), conversion rates also decline. Specifically, conversion occurs at rate $\lambda f(x)$, where x is the amount of sequence similarity between copies and λ is the rate of conversion between identical genes. Linear and quadratic forms for $f(x)$ were examined. For both classes of models, the probability that no conversions occur, as well as the expected number of conversions, depends critically on the mutation rate to conversion rate ratio ($2\nu/\lambda$ for the k-hit model, $2\mu/\lambda$ for the more general model).

A strong result is obtained from the point mutation model—if $2\mu/\lambda \gg 0.1$, few conversions occur, and the duplicated members essentially evolve independently. If, however, $2\mu/\lambda \ll 0.1$, very large numbers of conversions occur and the duplicated copies remain a family. Given $\mu \approx 5 \times 10^{-9}$, $\lambda \gg 10^{-7}$ is sufficient to keep duplicated genes from diverging enough to escape conversion, while if $\lambda \ll 10^{-7}$, it is highly likely that no conversions occur between duplicated copies. Conversion rates between pairs of linked duplicated mammalian genes are typically in the range of 10^{-7} to 10^{-5}, suggesting that divergence simply through the accumulation of point mutations is not sufficient to allow linked copies to escape conversion.

As a numerical example, consider $\lambda = 10^{-6}$, and make the conservative choice of a linear conversion function for $f(x)$ (Eq. 15) and a large x_{crit} value—0.8 (*i.e.*, conversions only occur between sequences with more than 80% similarity). For this combination, the expected number of conversions between a duplicated pair before they evolve independently is $\approx 6 \times 10^4$. The expected number of conversions is much higher if we assume either a quadratic conversion function (16), or lower x_{crit} values. If the expected number of conversions is very large, we expect that models for the amount of concerted evolution which assume constant conversion rates are likely good approximations.

This does not imply that neutral genes cannot escape conversion. Unusual events, such as large insertions/deletions, insertion of mobile elements, or small deletions in regions critical for conversion, may be sufficient to block further conversions. The k-hit model is more appropriate here, and it suggests (assuming $k = 1$) that if $2\nu \gg \lambda$, few conversions occur, while if $2\nu \ll \lambda$, the genes evolve as a family. Given our almost complete ignorance on the nature of events sufficient to block conversion and the rate at which such events occur, we are currently not in a position to evaluate whether or not selectively neutral duplicated genes can escape conversion via this mechanism.

Duplications where the new copy is placed on a different chromosome from the original copy have a better chance of escaping conversion that do linked duplicated genes. In addition to a reduction in conversion rates compared with linked copies, the processes generating duplicate copies on nonhomologous chromosomes may also speed up the divergence process. Reverse-transcription provides one example of a mechanism for spreading duplicate copies to different chromosomes (for reviews, see BALTIMORE 1985; TEMIN 1985; VANIN 1985; WALSH 1985). Reverse-transcription removes introns, which many greatly reduce (or even block) the amount of conversion with the parental gene. Further, processed *Pol*II genes (chromosomally integrated reverse-transcribed mRNAs) are usually nonfunctional, allowing for rapid divergence. However, there are exceptions (SOARES *et al.* 1985; STEIN *et al.* 1983; LEWIS and COWAN 1986; MCCARREY and THOMAS 1987), and the suggestion has been made that reverse-transcription of an illegitimate *Pol*III transcript which spans a *Pol*II gene can result in the new copy being functional (VANIN 1985; SOARES *et al.* 1985). Such a process may be important in the evolution of new gene functions, by placing genes in potentially new regulatory regions.

Caveats and directions for extension: It is important to reiterate the tentative nature of our models relating conversion rates and sequence similarity. Most experiments examining sequence requirements looked at either homologous recombination or conversion between alleles at identical positions on homologous chromosomes. Given the uncertain connection between homologous recombination/conversion at single loci and conversion between different loci, these results may be misleading. A further caution is that even in those few experiments using duplicated genes, blocks of *exact* homology were used, and conversion as a function of the total length of exact homology was examined. If blocks of exact homology, rather than regions of high (but not perfect) homology, are required for conversion, modeling conversion as a function of the fraction of sequence similarity may give misleading results.

A key assumption in our models is that conversion covers an entire gene and the amount of homology in that region, and not elsewhere, sets conversion rates. This allows us to use π, the probability that no conversions occur following a new duplication, to obtain the expected number of conversions before genes "escape" further conversions. The reasoning is that each conversion event resets the process, so that the expected number of trails before an escape follows a geometric distribution with parameter π. An important extension would be to incorporate conversion tracts not covering the entire gene. It should be noted that our calculation of π, the probability no conversions occur following a duplication, is correct regardless of the length of an average conversion tract. Gene pairs under conditions giving a high value of π in our model are likely to experience only a few conversions, regardless of the average length of a conversion tract.

Role of selection: If one (or both) members of a duplicated pair are under selection, both the rate of divergence and the acceptability of certain conversion events are altered from the neutral model. Generally, one expects the rate of divergence between genes under selection to be slower than the divergence of neutral genes (but see below). For the same conversion rate, reduced rates of divergence imply (all else being equal) that more conversions occur between a duplicated pair than predicted from the neutral analysis. However, all else is likely not equal. If there is selection to maintain at least one functional copy, conversion events replacing functional copies with nonfunctional copies will be selected against. Thus, with selection to retain a functional copy, both divergence rate and effective conversion rate are reduced. As a crude bracket for extremes, divergence rates likely range from 2μ (strictly neutral case) to μ (one gene does not change at all), while effective conversion rates vary from λ (all conversions acceptable) to $\lambda/2$ (conversions from one gene not accepted, *e.g.*, a pseudogene converting a functional copy). Under these bounds, the mutation/conversion ratio ($2\mu/\lambda$ in the strictly neutral case) ranges from μ/λ to $4\mu/\lambda$. Recall from (22) and (23) that we can write $\pi = \exp\{-\theta h(x)\}$, where θ is the mutation/conversion ratio. Our general conclusion is that if $\theta \gg 0.1$, duplicated copies can escape multiple conversions, while if $\theta \ll 0.1$ the duplicated pair is strongly constrained by conversion. Given that θ likely ranges from μ/λ to $4\mu/\lambda$ for a duplication under selection to retain a functional copy, $\lambda \gg 40\mu$ is sufficient for conversion to constrain independent evolution, while $10\mu \gg \lambda$ is sufficient to allow the pair to escape multiple conversions.

In some cases, conversion events may be selected for—events generating diversity, such as in the immune system genes, are possible examples. Increases in diversity via conversion may be fairly widespread—frequent conversions observed between members of the cytochrome P-450 superfamily have been suggested to increase the diversity of substrates acted upon by family members (AFFOLTER and ANDERSON 1984; ADESNIK and ATCHISON 1986; ATCHISON and ADESNIK 1986).

Ultimately, if the duplicated pair has diverged functionally, only those conversions not disrupting selectable differences are likely to be fixed. Conversions may still occur between functionally distinct genes, for example conversions have been observed between the mammalian hormone genes oxytocin and vasopressin (IVELL and RICHTER 1984; RUPPERT, SCHERE and SCHUTZ 1984). However, we generally expect conversions (that are to become fixed in a population) between distinct genes to be limited to short regions not interfering with gene function.

Conversions between functionally distinct genes are expected to be deleterious, and, not surprisingly, it has been argued that selection should favor decreased conversion between such genes. SCHIMENTI and DUNCAN (1984) have suggested that certain *Alu* insertions are selected for by this very reason. Many others have argued a similar role for introns. Are such scenarios reasonable? Consider the extreme case where all conversions produce lethals, and consider two "alleles": allele A is a duplicated pair with a conversion blocker (such as *Alu*), allele a is a normal duplicated pair, both cases assume complete linkage. Given intrachromosomal conversion occurs in allele a at a maximum rate of λ (and usually lower, allowing for divergence), under the assumption of lethality, the genotypes AA:Aa:aa have fitnesses $1{:}1-\lambda{:}1-2\lambda$. From standard one-locus theory (KIMURA 1957), A behaves like a neutral allele unless $4N_e\lambda \gg 1$, where N_e is the variance effective population size. Even in the most extreme case (*i.e.*, complete lethality), for typical λ values (*i.e.*, $\lambda \approx 10^{-6}$), N_e must exceed 250,000 for selection to be effective. Given a more reasonable fitness decrease for a conversion (λs, with say $s = 0.1$), N_e must exceed 2,500,000. For typical mammalian populations, such large values of N_e are unlikely.

Conversion and evolution by gene duplication: Evolution by gene duplication is generally regarded as a primary mechanism for the creation of new genes (OHNO 1970; LI 1983; DOOLITTLE 1985), yet our understanding of this process is still at a very primitive stage. A major conceptual problem is genes must diverge enough to become functionally distinct, and hence independently maintained by selection, but must not diverge too much, else they become inactive pseudogenes. This rather fine balance in the amount of divergence can be greatly altered by conversion. Conversions at a sufficiently frequent rate may pose a major constraint to the evolution of new genes, but, on the other hand, conversions can potentially reactivate pseudogenes, increasing their chances of subsequent functional divergence.

To resolve these roles, a much better understanding than we currently possess on the amount of divergence sufficient to create a new, selectable, gene function is required. Regulatory divergence is likely to be the initial step toward evolution of a new function, with subsequent selection for structural divergence (WILSON, CARLSON and WHITE 1977; HALL 1983). Much regulatory divergence occurs following gene duplication (FERRIS and WHITT 1979), and simple differences in tissue-specific regulation may allow the same enzyme to play different roles. A striking example is ϵ-crystallin, comprising up to 23% of total eye lens protein in birds and crocodiles, which has recently been shown to be identical to the glycolytic enzyme lactate dehydrogenase (WISTOW, MULDERS and DE JONG 1987). A more direct example bearing on reg-

ulatory evolution and gene duplication is given by the alcohol dehydrogenase (ADH) gene in Drosophila. In most species, differences in tissue-specific levels of expression are due to the use of different 5′ promoters of the same structural gene (reviewed in FISCHER and MANIATIS 1986). However, in the *mulleri* subgroup, the ADH gene has duplicated and tissue-specificity is due to differences in expression of the duplicated copies (MILLS *et al.* 1986).

Given our ignorance in predicting regulatory changes from nucleotide changes, inferences based on sequence analysis focus on changes in amino acid sequence, potentially overlooking significant regulatory differences in favor of functionally trivial amino acid substitutions. However, some studies indicate that amount of structural change required for a new function may be slight—HILL *et al.* (1984) and CARRELL (1984) found that single amino acid changes are sufficient to create new functions in the mammalian plasma protease inhibitor genes. Significantly, active site regions in duplicate human serine protease inhibitor genes evolve at a much faster rate than that seen for presumed neutral regions (HILL and HASTIE 1987; LEIGH BROWN 1987), presumably reflecting selection favoring certain mutants following a duplication event.

Although usually not as extreme as the protease example, rapid divergence following duplication is a commonly observed feature (reviewed in LI 1983). This has been interpreted as either a removal from selective constraints (KIMURA 1983) or positive selection for new functional sites (GOODMAN 1981). Rapid rates of evolution of pseudogenes support a neutralist explanation, while the serine proteases and calmodulin (the latter showing rapid evolution of sites that subsequently become invariant, BADA *et al.* 1984), support a selectively-driven rate at least during parts of the initial divergence. The exact cause of divergence is of importance to modeling issues. As discussed above, given the same actual conversion rates, a pair of duplicated genes experiencing divergence through positive selection is likely to escape conversion more readily than a pair where only one copy is under selection. This introduces an asymmetry into the conversion escape process wherein genes which have not functionally diverged are more likely to undergo repeated conversions than are genes diverging under selection for new function.

An extensive population genetics theory on the rate of silencing of duplicated loci exists (reviewed in LI 1980; TAKAHATA 1982; WATTERSON 1983). In the absence of conversion, silencing of a duplicate locus can occur reasonably fast, suggesting that genes are more likely to be silenced than form new functions. Conversion can potentially both decrease the rate of silencing as well as reactivating silenced copies, provided they have not diverged sufficiently to escape further conversions. KOCH (1972) and RIGBY, BURLEIGH and HARLEY (1974) have suggested reactivation of a silenced duplicate gene is the mode for creation of new function, but there are problems with this idea as these authors present it (LI 1983). If a reactivating conversion event covers the entire pseudogene, any divergence accumulated will be lost. Even if conversion achieves reactivation by converting only a small region, the remaining divergence has accumulated randomly with respect to function, and are thus not likely to be important.

The exact role of conversion in the evolution of new functions from duplicate genes is unresolved, although theoretical analysis of the success or failure of duplicated genes in the absence of conversion has begun (OHTA 1987). OHTA's analysis stressed the importance of the ratio the different kinds of mutations (deteriorating *vs.* useful) in the dynamics of whether a duplicated gene acquires a new function or becomes a pseudogene. When conversion is considered, the spatial distributions of mutants also becomes very important. A short conversion tract is sufficient to reactivate a pseudogene if all the deteriorating mutant sites are clustered, but if such sites are scattered throughout the gene, conversion is unlikely to reactivate it.

The view we favor for the role of conversion in the evolution of new gene function is a variation of the ideas of KOCH and RIGBY *et al.* Given (1) the possibility that regulatory divergence is the initial driving step behind evolution by gene duplication, (2) the high risk of duplicate genes degenerating into pseudogenes, and (3) the above-mentioned asymmetry in the conversion escape process, we feel that conversion may enhance the chance of a duplicated pair giving rise to a new function. If the events required to initiate the evolution of a new function are few, such as regulatory divergence (in the absence of any initial change in amino acid sequence) and/or the change of one or a few key amino acids, the amount of divergence required for their appearance need not be great. Conversion can potentially enhance this process by reducing the rate of pseudogene formation, allowing duplicate copies more of a chance to acquire a selectable difference. Once significant functional divergence occurs, selection is likely sufficient to maintain the independence of each copy, allowing further evolution and refinement of the new function.

Thanks to MICHAEL LISKAY for helpful discussions on homology and conversion rates, and to TOM NAGYLAKI and TOMOKO OHTA for constructive comments. This paper is dedicated to the memory of LARRY SANDLER, outstanding teacher, colleague, and friend.

LITERATURE CITED

ABRAMOWITZ, M., and I. A. STEGUN (Editors), 1964 *Handbook of Mathematical Functions.* National Bureau of Standards, Washington, D.C.

ADESNIK, M., and M. ATCHISON, 1986 Genes for cytochrome P-450 and their regulation. CRC Crit. Rev. Biochem. **19:** 247–305.

AFFOLTER, M., and A. ANDERSON, 1984 Segmental homologies in the coding and 3′ non-coding sequences of rat liver cytochrome P-450e and P-450b cDNAs and cytochrome P-450e-like gene. Biochem. Biophys. Res. Commun. **118:** 655–662.

ATCHISON, M., and M. ADESNIK. 1986 Gene conversion in a cytochrome P-450 gene family. Proc. Natl. Acad. Sci. USA **83:** 2300–2304.

AYARES, D., L. CHEKURI, K.-Y. SONG and R. KUCHERLAPATI, 1986 Sequence homology requirements for intermolecular recombination in mammalian cells. Proc. Natl. Acad. Sci. USA **83:** 5199–5203.

BADA, M. L., M. GOODMAN, J. BERGER-COHN, J. G. DEMAILLE and G. MATSUDA, 1984 The early adaptive evolution of calmodulin. Mol. Biol. Evol. **1:** 442–455.

BALTIMORE, D., 1985 Retroviruses and retrotransposons: the role of reverse transcription in shaping the eukaryotic genome. Cell **40:** 481–482.

BIANCHI, M. E., and C. M. RADDING, 1983 Insertions, deletions and mismatches in heteroduplex DNA made by RecA protein. Cell **35:** 511–520.

BLACKWELL, T. K., M. W. MOORE, G. D. YANCOPOULOS, H. SUH, S. LUTZKER, E. SELSING and F. W. ALT, 1986 Recombination between immunoglobulin variable region gene segments is enhanced by transcription. Nature **324:** 585–589.

BORST, P., and D. R. GREAVES, 1987 Programmed gene rearrangements altering gene expression. Science **235:** 658–667.

BRENNER, D. A., A. C. SMIGOCK and R. D. CAMERINI-OTERO, 1986 Double-strand gap repair results in homologous recombination in mouse L cells. Proc. Natl. Acad. Sci. USA **83:** 1762–1766.

BRIDGES, C. B., 1935 Salivary chromosome maps. J. Hered. **26:** 60–64.

CARRELL, R., 1984 Therapy by instant evolution. Nature **312:** 14.

COX, D. R., and H. D. MILLER, 1965 *The Theory of Stochastic Processes.* Chapman & Hall, New York.

DOOLITTLE, R. F., 1985. The genealogy of some recently evolved vertebrate proteins. Trends Biochem. Sci. **10:** 233–237.

EGEL, R., D. H. BEACH and A. J. S. KLAR, 1984 Gene required for initiation and resolution of mating-type switching in fission yeast. Proc. Natl. Acad. Sci. **81:** 3481–3485.

ERHART, M. A., K. S. SIMMONS and S. WEAVER, 1985 Evolution of the mouse β-globin genes: a recent gene conversion in the Hbb^s haplotype. Mol. Biol. Evol. **2:** 304–320.

FERRIS, S. D., and G. S. WHITT, 1979 Evolution of the differential regulation of duplicate genes after polyploidization. J. Mol. Evol. **12:** 267–317.

FINK, G. R., and T. D. PETES, 1984 Gene conversion in the absence of reciprocal recombination. Nature **310:** 728–729.

FISCHER, J. A., and T. MANIATIS, 1986 Regulatory elements involved in *Drosophilia Adh* gene expression are conserved in divergence species and separate elements mediate expression in different tissues. EMBO J. **5:** 1275–1289.

FLAVELL, R. A., H. ALLEN, L. C. BURKLY, D. H. SHERMAN, G. L. WANECK and G. WIDERA, 1986 Molecular biology of the H-2 histocompatibility complex. Science **233:** 437–443.

GONDA, D. K., and C. M. RADDING, 1983 By searching processively RecA protein pairs DNA molecules that share a limited stretch of homology. Cell **34:** 647–654.

GOODMAN, M., 1981 Decoding the pattern of protein evolution. Prog. Biophys. Mol. Biol. **37:** 105–164.

GORSKI, J., and B. MACH, 1986 Polymorphism of human Ia antigens: gene conversion between two DR β loci results in a new HLA-D/DR specificity. Nature **322:** 67–70.

HALL, B. G., 1983 Evolution of new metabolic functions in laboratory organisms. pp. 234–257. In: *Evolution of Genes and Proteins,* Edited by M. NEI and R. K. KOEHN. Sinauer, Sunderland, Mass.

HESS, J. F., C. SCHMID and C.-K. J. SHEN, 1984 A gradient of sequence divergence in the human adult α-globin duplication units. Science **226:** 67–70

HESS, J. F., M. FOX, C. SCHMID and C.-K. J. SHEN, 1983 Molecular evolution of the human α-globin-like gene region: Insertion and deletion of *Alu* family repeats and non-*Alu* DNA sequences. Proc. Natl. Acad. Sci. USA **80:** 5970–5974.

HILL, A. V. S., R. D. NICHOLLS, S. L. THEIN and D. R. HIGGS, 1985 Recombination within the human embryonic ζ-globin locus: a common ζ-ζ chromosome produced by gene conversion of the $\psi\zeta$ gene. Cell **42:** 809–819.

HILL, R. E., and N. D. HASTIE, 1987 Accelerated evolution in the reactive centre region of serine protease inhibitors. Nature **326:** 96–99.

HILL, R. E., P. H. SHAW, P. A. BOYD, H. BAUMANN and N. D. HASTIE, 1984 Plasma protease inhibitors in mouse and man: divergence within the reactive centre regions. Nature **311:** 175–177.

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

HSIEH, P., M. S. MEYN and R. D. CAMERINI-OTERO, 1986 Partial purification and characterization of a recombinase from human cells. Cell **44:** 885–894.

IVELL, R., and D. RICHTER, 1984 Structure and comparison of the oxytocin and vasopressin genes from rat. Proc. Natl. Acad. Sci. USA **81:** 2006–2010.

JACKSON, J. A., and G. R. FINK, 1985 Meiotic recombination between duplicated elements in *Saccharomyces cerevisiae.* Genetics **109:** 303–332.

JASIN, M., J. DE VILLIERS, F. WEBER and W. SCHAFFNER, 1985 High frequency of homologous recombination in mammalian cells between endogenous and introduced SV40 Genomes. Cell **43:** 695–703.

JEFFREYS, A. J., V. WILSON and S. L. THEIN, 1985 Hypervariable "minisatellite" regions in human DNA. Nature **314:** 67–73.

JINKS-ROBERTSON, S., and T. D. PETES, 1985 High frequency meiotic gene conversion between repeated genes on nonhomologous chromosomes in yeast. Proc. Natl. Acad. Sci. USA **82:** 3350–3354.

JINKS-ROBERTSON, S., and T. D. PETES, 1986 Chromosomal translocations generated by high-frequency meiotic recombination between repeated yeast genes. Genetics **114:** 731–752.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules. pp. 21–132. In *Mammalian Protein Metabolism III,* Edited by H. N. MUNRO. Academic Press, New York.

KARLIN, S., and TAYLOR, H. M. 1981 *A Second Course in Stochastic Processes.* Academic Press, New York.

KEIL, R. L., and G. S. ROEDER, 1984 *Cis*-acting, recombination-stimulating activity in a fragment of the ribosomal DNA of *S. cerevisiae.* Cell **39:** 377–386.

KENTER, A. L., and B. K. BIRSHTEIN, 1981 Chi, a promoter of generalized recombination in λ phage, is present in immunoglobulin genes. Nature **293:** 402–404.

KIMURA, M., 1957 Some problems of stochastic processes in genetics. Ann. Math. Stat. **28:** 882–901.

KIMURA, M. 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

KLAR, A. J. S., and J. N. STRATHERN, 1984 Resolution of the recombination intermediates generated during yeast mating type switching. Nature **310:** 744–748.

KLEIN, H. L., 1984 Lack of association between intrachromosomal gene conversion and reciprocal exchange. Nature **310:** 748–753.

KLEIN, H. L., and T. D. PETES, 1981 Intrachromosomal gene conversion in yeast. Nature **289:** 144–148.

KMIEC, E. B., and W. K. HOLLOMAN, 1986 Homologous pairing

of DNA molecules by *Ustilago* Rec1 protein is promoted by sequences of Z-DNA. Cell **44:** 545–554.

KMIEC, E. B., K. J. ANGELIDES and W. K. HOLLOMAN, 1985 Left-handed DNA and the synaptic pairing reaction promoted by *Ustilago* Rec1 protein. Cell **40:** 139–145.

KOCH, A. L., 1972 Enzyme evolution. I. The importance of untranslatable intermediates. Genetics **72:** 297–316.

KOSTRIKEN, R., J. N. STRATHERN, A. J. S. KLAR, J. B. HICKS and F. HERRON, 1983 A site-specific endonuclease essential for mating-type switching in *Saccharomyces cerevisiae*. Cell **35:** 167–174.

KRAWINKEL, U., G. ZOEBELEIN and A. L. M. BOTHWELL, 1986 Palindromic sequences are associated with sites of DNA breakage during gene conversion. Nucleic Acids Res. **14:** 3871–3882.

LEIGH BROWN, A., 1987 Positively darwinian molecules? Nature **326:** 12–13.

LEFEVRE, J. C., A. M. GASC, A. C. BURGER, P. MOSTACHFI and A. M. SICARD, 1984 Hyperrecombination at a specific DNA sequence in pneumococcal transformation. Proc. Natl. Acad. Sci. USA **81:** 5148–5188.

LEWIS, S. A., and N. J. COWAN, 1986 Anomalous placement of introns in a member of the intermediate filament multigene family: an evolutionary conundrum. Mol. Cell. Biol. **6:** 1529–1534.

LI, W.-H., 1980 Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fish. Genetics **95:** 237–258.

LI, W.-H., 1983 Evolution of duplicated genes and pseudogenes. pp. 14–37. In: *Evolution of Genes and Proteins*, Edited by M. NEI and R. K. KOEHN. Sinauer, Sunderland, Mass.

LI, W.-H., and M. TANIMURA, 1987 The molecular clock runs more slowly in man than in apes and monkeys. Nature **326:** 93–96.

LI, W.-H., C.-C. LUO and C.-I. WU, 1985 Evolution of DNA sequences. pp. 1–94. In: *Molecular Evolutionary Genetics*, Edited by R. J. MACINTYRE. Plenum, New York.

LICHTEN, M., R. H. BORTS and J. E. HABER, 1987 Meiotic gene conversion and crossing over between dispersed homologous sequences occurs frequently in *Saccharomyces cerevisiae*. Genetics **115:** 233–246.

LICHTEN, M., and M. S. FOX, 1984 Evidence for inclusion of regions of nonhomology in heteroduplex products of bacteriophage λ recombination. Proc. Natl. Acad. Sci. USA **81:** 7180–7184.

LISKAY, R. M., and J. L. STACHELEK, 1986 Information transfer between duplicated chromosomal sequences in mammalian cells involves contiguous regions of DNA. Proc. Natl. Acad. Sci. USA **83:** 1802–1806.

LISKAY, R. M., A. LETSOU and J. L. STACHELECK, 1987 Homology requirements for efficient gene conversion between duplicated chromosomal sequences in mammalian cells. Genetics **115:** 161–167.

LISKAY, R. M., J. L. STACHELECK and A. LETSOU, 1984 Homologous recombination between repeated chromosomal sequences in mouse cells. Cold Spring Harbor Symp. Quant. Biol. **49:** 183–189.

MARKHAM, P., and H. L. K. WHITEHOUSE, 1982 A hypothesis for the initiation of genetic recombination in eukaryotes. Nature **295:** 421–423.

MCCARREY, J. R. and K. THOMAS, 1987 Human testis-specific PGK gene lacks introns and possess characteristics of a processed gene. Nature **326:** 501–504.

MCINTYRE, K. R., and J. G. SEIDMAN, 1984 Nucleotide sequence of mutant I-$A\beta^{bm12}$ gene is evidence for genetic exchange between mouse immune response genes. Nature **308:** 551–553.

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

MICHELSON, A. M., and S. H. ORKIN, 1983 Boundaries of gene conversion within the duplicated human α-globin genes. J. Biol. Chem. **258:** 15245–15254.

MILLS, L. E., P. BATTERHAM, J. ALEGRE, W. T. STARMER and D. T. SULLIVAN, 1986 Molecular genetic characterization of a locus that contains duplicate ADH genes in *Drosophila mojavensis* and related species. Genetics. **112:** 295–310.

MULLER, H. J., 1936 Bar duplication. Science **83:** 528–530.

NAGYLAKI, T., 1984a Evolution of multigene families under intrachromosomal gene conversion. Genetics **106:** 524–548.

NAGYLAKI, T., 1984b Evolution of muligene families under interchromosomal gene conversion. Proc. Natl. Acad. Sci. USA **81:** 3796–3800.

OHNO, S., 1970 *Evolution by Gene Duplication*. Springer-Verlag, New York.

OTHA, T., 1980 *Evolution and Variation in Multigene Families*. Springer-Verlag, New York.

OTHA, T., 1983 On the evolution of Multigene families. Theor. Popul. Biol. **23:** 216–240.

OTHA, T., 1987 Simulating evolution by gene duplication. Genetics **115:** 207–213.

PAYS, E., S. HOUARD, A. PAYS, S. VAN ASSEL, F. DUPONT, D. AERTS, G. HUET-DUVILLIER, V. GOMES, C. RICHET, P. DEGAND, N. VAN MEIRVENNE and M. STEINERT, 1985 *Trypanosoma brucei*: the extent of conversion in antigen genes may be related to the DNA coding specificity. Cell **42:** 821–829.

POWERS, P. A., and O. SMITHIES, 1986 Short gene conversions in the human fetal globin gene region: a by-product of chromosome pairing during meiosis? Genetics **112:** 343–358.

REYNARD, C.-A., V. ANQUEZ, H. GRIMAL and J.-C. WEILL, 1987 A hyperconversion mechanism generates the chicken light chain preimmune repertoire. Cell **48:** 379–388.

RIGBY, P. W. J., B. D. BURLEIGH, JR. and B. S. HARLEY, 1974 Gene duplication in experimental evolution. Nature **251:** 200–204.

RUBNITZ, J., and S. SUBRAMANI, 1984 The minimum amount of homology required for homologous recombination in mammalian cells. Mol. Cell. Biol. **4:** 2253–2258.

RUBNITZ, J., and S. SUBRAMANI, 1986 Extrachromosomal and chromosomal gene conversion in mammalian cells. Mol. Cell. Biol. **6:** 1608–1614.

RUPPERT, S., G. SCHERER and G. SCHUTZ, 1984 Recent gene conversion involving bovine vasopressin and oxytocin precursor gene suggested by nucleotide sequence. Nature **308:** 554–557.

SCHIMENTI, J. C., and C. H. DUNCAN, 1984 Ruminant globin gene structures suggest an evolutionary role for *Alu*-type repeats. Nucleic Acids Res. **12:** 1641–1655.

SHEN, P., and H. V. HUANG, 1986 Homologous recombination in *Escherichia coli*: dependence of substrate length and homology. Genetics **112:** 441–457.

SINGER, B. S., L. GOLD, P. GAUSS and D. H. DOHERTY, 1982 Determination of the amount of homology required for recombination in bacteriophage T4. Cell **31:** 25–33.

SLIGHTOM, J. L., A. E. BLECHL and O. SMITHIES, 1980 Human fetal G_γ- and A_γ-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell **21:** 627–638.

SLIGHTOM, J. L., L.-Y. E. CHANG, B. F. KOOP and M. GOODMAN, 1985 Chimpanzee fetal G_γ and A_γ globin gene nucleotide sequences provide further evidence of gene conversions in Hominine evolution. Mol. Blol. Evol. **2:** 370–389.

SMITH, G. R., 1983 Chi hotspots of generalized recombination. Cell **34:** 709–710.

SMITHIES, O., and P. A. POWERS, 1985 Gene conversions and

their relationship to homologous chromosome pairing. Philos. Trans. R. Soc. Lond. B **312:** 291–302.

SOARES, M. B., E. SCHON, A. HENDERSON, S. A. KARATHANASIS, R. CATE, S. ZEITLIN, J. CHIRGWIN and A. EFSTRATIADIS, 1985 RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. Mol. Cell. Biol. **5:** 2090–2103.

STAHL, F. W., 1979 *Genetic Recombination: Thinking about It in Phage and Fungi.* Freeman, San Francisco.

STEIN, J. P., R. P. MUNJALL, L. LAGACE, E. C. LAI, B. W. O'MALLEY and A. R. MEANS, 1983 Tissue-specific expression of a chicken calmodulin pseudogene lacking intervening sequences. Proc. Natl. Acad. Sci. USA **80:** 6485–6489.

STEINMETZ, M., D. STEPHAN and K. F. LINDAHL, 1986 Genetic organization and recombinational hotspots in the murine major histocompatibility complex. Cell **44:** 895–904.

STEINMETZ, M., Y. UEMATSU and K. F. LINDAHL, 1987 Hotspots of homologous recombination in mammalian genomes. Trends Gene. **3:** 7–10.

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.

TAKAHTA, N., 1982 The disappearance of duplicate gene expression. pp. 169–190. In: *Molecular Evolution, Protein Polymorphism and the Neutral Theory*, Edited by M. KIMURA. Springer-Verlag, Berlin.

TEMIN, H. M. 1985 Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. Mol. Biol. Evol. **2:** 455–468.

VANIN, E. F., 1985 Processed pseudogenes: characteristics and evolution. Annu. Rev. Genet. **19:** 253–272.

VOELKEL-MEIMAN, K., R. L. KEIL and G. S. ROEDER, 1987 Recombination-stimulating sequences in yeast ribosomal DNA correspond to sequences regulating transcription by RNA polymerase I. Cell **48:** 1071–1079.

WALSH, J. B., 1985 How many processed pseudogenes are accumulated in a gene family? Genetics **110:** 345–364.

WATT, V. M., C. J. INGLES, M. S. URDEA and W. J. RUTTER, 1985 Homology requirements for recombination in *Escherichia coli.* Proc. Natl. Acad. Sci. USA **82:** 4768–4772.

WATTERSON, G. A. 1983 On the time for gene silencing at duplicate loci. Genetics **105:** 745–766.

WHITEHOUSE, H. L. K., 1982 *Genetic Recombination: Understanding the Mechanisms.* Wiley, New York.

WHITEHOUSE, H. L. K., 1983 Duplex breaks in DNA as recombination initiators. Nature **306:** 645–646.

WILSON, A. C., S. S. CARLSON and T. J. WHITE, 1977 Biochemical evolution. Annu. Rev. Biochem. **46:** 573–639.

WISTOW, G. J., J. W. M. MULDERS and W. W. DE JUNG, 1987 The enzyme lactate dehydrogenase is a structural protein in avian and crocodilian lenses. Nature **326:** 622–624.

Communicating editor: W.-H. LI

APPENDIX

1: Simulation procedure

Let n be the number of nucleotides in each gene of the duplicated pair. For each base position which matches in the duplicated pair, score a 1, else score a zero. Summing over all bases, a total score is computed, which ranges from 0 to n. Denote by x the fraction of similarity between the two genes, where x = (total score)/n. The divergence and conversion process is simulated by following the total score associated with a duplicated pair. Sequences change by either a conversion event, resetting the total score to n, or a single point mutation, which alters the score by at most one. Following arguments of the type presented for the k-hit model, the probability that a conversion event, rather than a mutational event, occurs is $\lambda f(x)/(2n\mu + \lambda f(x))$. If a mutational event occurs, with probability x the total score is decreased by one, else with probability $(1 - x)/3$ the score is increased by one. For most sets of parameter values ($\lambda/2\mu$ and n = 100, 1000), 50 trials were run, each trial recording the number of conversions before a run of $3n$ mutations without a conversion stops the trial. For $\lambda/2\mu$ values with predicted mean number of conversions [using (22) and (23)] less than 5, 250 traits were run.

2: A diffusion approach for computing $\phi(t)$

By associating a diffusion with our divergence process x_s, then we can obtain a pde for $\phi(t)$ using Kac Functionals (KARLIN and TALYOR 1981, pp. 222–224). Defining $\sigma(x, t)$ as the expected $\phi(t)$ value, given the process starts at x, we have directly from KARLIN and TALYOR's Eq. (5.39)

$$\partial\sigma(x, t)/\partial t = -\lambda f(x)\sigma(x, t) + M_{\delta x}\partial\sigma(x, t)/\partial x + (1/2)V_{\delta x}\partial^2\sigma(x, t)/\partial x^2 \quad \text{(A.1)}$$

with initial condition $\sigma(x, 0) = 1$, where $M_{\delta x}$ and $V_{\delta x}$ are the instantaneous mean and variance (respectively) of the associated diffusion process. Solving (A.1) yields $\phi(t)$ by considering $\sigma(1, t)$.