A MODEL FOR DNA SEQUENCE EVOLUTION WITHIN TRANSPOSABLE ELEMENT FAMILIES

J. F. Y. BROOKFIELD

*Department of Genetics, School of Biological Sciences, University of Leicester, University Road,
Leicester LE1 7RH, England*

Manuscript received February 14, 1985
Revised copy accepted October 19, 1985

ABSTRACT

A quantitative model is proposed for the expected degree of relationship between copies of a family of transposable elements in a finite population of hosts. Special cases of the model (in which the process of homogenization of element copies either is or is not limited by transposition rate) are presented and illustrated, using data on mobile sequences from different species. It is shown that transposition will be expected, in large populations, to result in only a rather distant relationship between transposable elements at different genomic sites. Possible inadequacies of the model are suggested and quantified.

APPROXIMATELY 15% of the genome of most eukaryotes consists of interspersed repetitive DNA sequences (BOUCHARD 1982). Many types of such repetitive sequences have been described. Some, such as the *Ty1* element in yeast (FINK *et al.* 1981; EIBEL *et al.* 1981), the *copia*-like elements of *Drosophila* (RUBIN *et al.* 1981) and the integrated proviruses of vertebrate retroviruses (VARMUS 1983) share a common structure with long, terminal, repetitive sequences and are mobile in the genome. Other sequences, such as the human *Alu* sequence (JELINEK and SCHMID 1982), are more constant in position, yet the interspersion of these sequences, in itself, suggests they can move to new genomic sites.

Much speculation has occurred concerning the functions of these sequences. Initially, it was suggested that the sequences are involved in the control of gene expression, by being used to mark [as a means of control of transcription (BRITTEN and DAVIDSON 1969) or processing (DAVIDSON and BRITTEN 1979)] genes expressed in differentiated cell types. Other authors have speculated that, in view of the probably replicative nature of the transposition process that moves DNA sequences to new sites, and as the consequent overreplication of mobile DNA sequences relative to the rest of the genome, such sequences could persist even if they were useless or even slightly harmful [so-called selfish or parasitic DNA (ORGEL and CRICK 1980; DOOLITTLE and SAPIENZA 1980; SAPIENZA and DOOLITTLE 1981)]. Problems exist with both functionality and parasitism as explanatory principles for these sequences.

In *Drosophila melanogaster* almost all interspersed repetitive DNA sequences

change genomic locations between strains (YOUNG 1979), and even in a wild population, *copia*-like sequences were found to vary greatly between individuals in position on the X chromosome (MONTGOMERY and LANGLEY 1983). Such data absolutely rule out the possibility that *Drosophila* interspersed repeats perform the functional roles envisaged for some repeats by BRITTEN and DAVIDSON. Similarly, while it is true that the property of replicative transposition could allow mobile sequences to spread through populations without any natural selection in their favor, and conditions for an equilibrium between the processes of transposition and selection have been calculated (CHARLESWORTH and CHARLESWORTH 1983), this does not explain why such overreplicating sequences do, in fact, exist in genomes and why they comprise 15% of the genome, rather than some other proportion. These questions are real ones, but they are evolutionary and, in a sense, are ecological questions of a type that biologists are used to being unable to answer.

Therefore, the forces which determine the presence and nature of mobile DNA sequences are unclear, and are likely to remain so. However, it is possible to take a more mechanistic view of eukaryotic transposable elements, concentrating on a more simple description of the expected population dynamics of sequences with given properties of transposition and deletion. Such an approach can produce testable predictions, most specifically about the expected frequency spectra of transposable element sites (LANGLEY, BROOKFIELD and KAPLAN 1983; CHARLESWORTH and CHARLESWORTH 1983). The predictions of these authors have yet to be tested empirically, as the only relevant data (MONTGOMERY and LANGLEY 1983) correspond to a rather uninteresting special case of the models (KAPLAN and BROOKFIELD 1983a).

In this paper, I propose to take an equally mechanistic approach to a related question, that of the evolutionary relationship between transposable elements at different genomic locations. It may be possible to elucidate the evolutionary mechanisms affecting transposable elements by comparing, using DNA sequencing techniques, different copies of transposable element families and inferring functional constraints on certain sequences from strong conservation of such sequences between copies. A major problem, of course, would arise in such studies. In many clusters of genes, such as the mammalian β -globins (JEFFREYS 1982), where the evolutionary processes of duplication, loss and silencing to produce pseudogenes occur at rates low enough for individual events to be dated by phylogenetic comparisons, phylogenetic trees of related genes within the genome can be produced, and evolutionary rates deduced, by dividing proportional base-pair divergence measurements by times derived from such trees. For transposable elements, no such inferences about the relationships in times to common ancestors of different sequence copies are possible. What is required is a prediction of expected times to a common ancestor for randomly chosen copies of a transposable element family from different genomic locations.

THE MODEL

LANGLEY, BROOKFIELD and KAPLAN (1983) proposed a model for the evolution of sites of transposable elements, which postulated that the evolutionary process consisted of the following steps:

1. Transposable elements are selectively neutral and transpose to new sites at a rate that varies inversely with the number of transposable elements already present in the genome.

2. When transpositions occur, the element is always inserted at a site not occupied by any transposable elements in any other individuals in the population. This requires that the number of available sites for transposable elements is very large.

3. Elements can be deleted precisely from their chromosomal locations at a rate μ per element per generation that is copy-number-independent.

4. Each generation, Wright-Fisher sampling takes place in a diploid population of effective size $2N_e$ at each site occupied by transposable elements in at least some genomes.

5. There is sufficient recombination between transposable element sites to bring all such sites into linkage equilibrium.

6. There is a very low rate of immigration of transposable elements into the population. Thus, the transposable elements never become extinct by stochastic loss.

LANGLEY *et al.* showed that, at stationarity, the expected frequency spectrum of sites of transposable elements can be described by a simple formula analogous to the infinite alleles frequency spectrum of single-locus population genetics theory (KIMURA and CROW 1964).

The expected number of transposable element sites with frequencies in the range from x to $x + \delta x$ is

$$\Lambda \theta x^{-1} (1 - x)^{\theta-1} \delta x \quad (1)$$

where Λ = the expected number of transposable elements per haploid genome at equilibrium. This will depend on the rate of deletion and on the dependence of transposition rate upon copy number: $\theta = 4N_e\mu$, where N_e and μ are as defined above.

This model assumes selective neutrality of transposable element sites, but the expected frequency spectrum will be approximately the same if it is selection against individuals with many transposable elements, rather than deletion, which balances the expected increase in mean copy number resulting from replicative transposition. This will be true if, and only if, selection is weak but still sufficiently strong to prevent any sites having high frequencies, and if the effects of selection do not vary between sites. If this latter condition does not hold, the variance in frequency between transposable element sites will be increased (KAPLAN and BROOKFIELD 1983b).

At equilibrium, therefore, sites will be constantly created by a transposition process that copies a transposable element at an old site, and sites will be lost at an equivalent rate by a combination of deletion and sampling drift. As a result, elements at diverse genomic sites will come to be identical by descent. A quantitative description of this process can be produced by adding a further assumption to the model. This assumption is that all copies of the transposable element family at all genomic sites are functionally equivalent, *i.e.*, their probabilities of transposition and deletion are identical.

I shall consider a population at stationarity described by the above model, and I shall assume that the mean copy number of the individuals in the population is closely regulated and that, for this population, $N_e = N$, the total number of diploid individuals in the population. Thus, the population contains a total of $2N$ Λ copies of the element at all times. I shall also assume complete linkage equilibrium between transposable element sites. Simulations performed by LANGLEY, BROOKFIELD and KAPLAN indicate that such linkage equilibrium is likely to hold in nature. The following analysis will be approximately correct.

Consider a site with population frequency $i/2N$, i.e., that site is occupied by a transposable element in i of the $2N$ haploid genomes in the population. For such a site the expected time to a common ancestor for randomly chosen copies of the transposable element from different haploid genomes will be an unknown quantity, which can be called $t(i)$. At stationarity, the expected frequency spectrum of sites of transposable elements will be approximately given by a discrete version of (1). The expected number of sites of transposable elements with frequency $i/2N$ in the population will be

$$\Lambda \frac{\theta}{i} \left(\frac{2N - i}{2N} \right)^{\theta-1}$$

This will, henceforth, be represented as $f(i)$. A population with this expected frequency spectrum can be defined as being at time 0. If this population is randomly sampled by picking transposable elements from different sites, the expected time to a common ancestor can be defined as $\bar{T}$.

Consider now the population after one generation of transposition, deletion and sampling. After that generation, an expected proportion μ of the population of transposable elements will be at new sites generated by transposition events in that generation. There will be a compensating loss of an expected proportion of μ elements at all sites which had nonzero frequencies in generation 0. The population in generation 1 can be sampled randomly. If the population is at stationarity, the expected time to a common ancestor of randomly chosen element copies in generation 1 will still be $\bar{T}$. As two element copies are chosen, there will be three outcomes that can be defined.

1. Two new sites will be chosen. This will occur with probability μ^2 .
2. An old site and a new site will be chosen. This will occur with probability $2(1 - \mu)\mu$.
3. Two old sites will be chosen. This will occur with probability $(1 - \mu)^2$.

As $\mu \ll 1$, these probabilities are approximately 0, 2μ and $1 - 2\mu$.

If two old sites are chosen, we can consider the expected time to a common ancestor. The copies chosen at these sites are random samples of copies at old sites in generation 1. Furthermore, the copies at old sites in generation 1 are randomly sampled from the copies at those sites in generation 0. Thus, the copies sampled in generation 1 have descended from elements which, in generation 0, had an expected time to a common ancestor of $\bar{T}$. Thus, their expected time to a common ancestor in generation 1 is $\bar{T} + 1$.

The other significant cases are where an old site and a new site are sampled. They can be divided into cases where the old site sampled is a different one

from that which gave rise to the new site, and cases where the old site is that from which the new site was derived by transposition. To calculate the probability of the latter situation it can be noted that the probability that the new site is derived from an old site with frequency $i/2N$ is just $\frac{f(i)i}{2\Lambda N}$, since the transposition process is assumed to be equivalent to random sampling. The probability of picking this very same old site is, of course, the expected proportion of all transposable elements at that site, or $\frac{i(1-\mu)}{2N\Lambda}$.

Thus, the chance of picking a new site and the same old site from which it was derived, given that an old site and a new site are picked, is

$$\sum_{i=1}^{2N} \frac{f(i) \cdot i^2}{(2N\Lambda)^2}.$$

At this point, it is instructive to note that (1) represents an infinite-alleles distribution multiplied by a constant Λ , and that the homozygosity of the discrete analog of this distribution

$$\sum_{i=1}^{2N} \frac{f(i) \cdot i^2}{4N^2\Lambda}$$

will follow from standard theory (EWENS 1979) as approximately equal to $\frac{1}{1+\theta}$, a term variously referred to in the context of transposable elements as the "homozygosity" (LANGLEY, BROOKFIELD and KAPLAN 1983) or as "allelism" (OHTA 1984) of transposable element sites.

Thus, the probability of picking the old site that the new site was derived from is $\frac{1}{\Lambda(1+\theta)}$. The probability of picking a different old site is thus $1 - \frac{1}{\Lambda(1+\theta)}$. Clearly, if a new site and a different old site are picked, this is equivalent, in terms of mean time to a common ancestor, of sampling two old sites, *i.e.*, time = $\bar{T} + 1$. However, a new site and the old site from which it is derived being sampled is equivalent to sampling two copies at the same site in different individuals, which gives a time of $t(i)$ if the old site has frequency $i/2N$. The weighted mean of the $t(i)$'s,

$$\sum_{i=1}^{2N} \frac{f(i) \cdot i^2 t(i)}{(2N\Lambda)^2} \bigg/ \sum_{i=1}^{2N} \frac{f(i) i^2}{(2N\Lambda)^2} = \sum_{i=1}^{2N} f(i) i^2 t(i) \bigg/ \sum_{i=1}^{2N} f(i) i^2,$$

provides the approximate time in this case.

This time is hard to compute, but at most it will only be $2N$, as $\theta \rightarrow 0$, and for higher values of θ will be much less. Thus, it will be much less than $\bar{T}$, provided $\Lambda \gg 1$. I shall treat this quantity as 0. Proportional errors introduced in estimates of $\bar{T}$ will only be of the order of $1/\Lambda$.

Thus, the mean time for a common ancestor for elements sampled from different sites in generation 1 will be

$$(1 - 2\mu)(\bar{T} + 1) + 2\mu \left(1 - \frac{1}{(1 + \theta)\Lambda}\right)(\bar{T} + 1)$$

which, at stationarity, must equal $\bar{T}$. Thus,

$$\begin{aligned} 1 &= \frac{2\mu(\bar{T} + 1)}{\Lambda(1 + \theta)} \\ \bar{T} &= \frac{\Lambda(1 + \theta)}{2\mu} - 1 \simeq \frac{\Lambda(1 + \theta)}{2\mu} \end{aligned} \quad (2)$$

OHTA (1984) used a series of similar steps to those which I have used above in order to calculate the identity coefficients between transposable elements, both in the same genome and in different genomes. She showed these quantities to be very nearly the same for the case of free recombination between transposable element sites that I have modeled above. She further showed the transposition process to be mathematically similar in its effects to simple models of gene conversion (OHTA 1982). A more general model has been used to calculate the effect on the identity coefficients of any combination of transposition and unbiased gene conversion (OHTA 1985).

The value of $\bar{T}$, calculated above, can be illustrated by considering special cases.

1. θ large: This corresponds to the *D. melanogaster copia*-like elements (MONTGOMERY and LANGLEY 1983). As $\theta \gg 1$, $1 + \theta \simeq \theta$, and, as $\theta = 4N_e\mu$, $\bar{T} \simeq 2N_e\Lambda$. This is equivalent to the homogenization by genetic drift expected if all the copies of the transposable element were at a single locus in a host population of size $2N_e\Lambda$. Thus, the fact that the elements are at different genomic sites is not limiting to the homogenization process if site frequencies are low, a result also shown by SLATKIN (1985), and as a special case of the model of OHTA (1985).

2. θ is very small: If there are many sites where transposable elements have very high frequencies, θ estimates will be much less than 1 and $1 + \theta \simeq 1$. Thus, $\bar{T} \simeq \Lambda/2\mu$. This is the case where the rate of homogenization of the family is limited by a low transposition rate. In the extreme, as $\mu \rightarrow 0$, $\bar{T} \rightarrow \infty$, as then there would be no transposition, and elements at diverse genomic locations would be completely unrelated.

These values for times to common ancestors can be used to generate estimates for base-pair divergences between element copies. If the mutation rate for transposable elements to functionally equivalent transposable element copies is called v base-pair changes per base pair per generation, the proportion of bases diverged between randomly chosen element copies will be $2\bar{T}v$. (This estimate will be accurate only if it is much less than 1. It is based on an infinite site model without recombination).

APPLICATION OF THE MODEL

The model quantifies the straightforward prediction that, since identity by descent of elements at different sites in the genome arises by element trans-

position, the higher the rate of transposition, μ , the more closely elements will be related. This rate of transposition is, itself, reflected in the frequency spectrum of transposable element sites. Generally, all else being equal, the more variation there is in transposable element position in the population, the closer will be the relationship between transposable elements from different sites.

The *copia*-like sequences of *D. melanogaster* have high θ values, and Λ values of around 30–50. $\bar{T}$ can be calculated from (2), using the value of N_e of 3×10^6 calculated by KREITMAN (1983) from synonymous base pair heterozygosity in the alcohol dehydrogenase gene. This gives $\bar{T} = 0.9\text{--}1.5 \times 10^8$ generations, or around 10^7 yr. The rate of DNA sequence evolution in the *Drosophila* genus is little known. LANGLEY, MONTGOMERY and QUATTLEBAUM (1982) report estimated divergences of around 5% between *D. melanogaster* and *D. mauritiana* in the *Adh* flanking regions. This is the result of around 2 million years of independent evolution. Thus, if transposable elements evolved at the same rate, we would expect 25% divergence between copies. If, however, there was stronger sequence conservation of transposable elements than *Adh* flanking sequences, the observed value of 5% DNA sequence divergence within *copia*-like sequence families (SPRADLING and RUBIN 1981) could be consistent with the above divergence times. Indeed, STEPHENS, KREITMAN and NEI (1984) calculated a value of 4×10^5 for N_e , using KREITMAN's data but different assumptions, so the model may be entirely consistent with the data.

Unlike *Drosophila copia*-like sequences, most interspersed repetitive DNA sequences do not show variation in position between individuals within populations. An example of such a sequence is the human *Alu* sequence (JELINEK and SCHMID 1982), which is around 290 bp in length and is found repeated around 300,000 times in the human genome. The different repeat copies are diverged from each other by around 20% in base sequence. There is some evidence that the *Alu* sequence may be transposable via an RNA intermediate, which is transcribed by RNA Polymerase III, reverse transcribed and reinserted into the genome (JAGADEESWARAN, FORGET and WEISSMAN 1981). The circles of *Alu* DNA which would be expected to arise as intermediates in such a process have been isolated (KROWLEWSKI and RUSH 1984). Their very interspersed pattern itself implies that these sequences are mobile, and in the rat there is a polymorphism for the presence or absence of a sequence homologous to *Alu* near the prolactin gene (SCHULER, WEBER and GORSKI 1983). In the duplicated human α -globin genes, *Alu* sequence DNA has been inserted into DNA 5' to the α -2 gene (or removed from 5' to the α -1 gene) at some time since the genes duplicated (HESS *et al.* 1983). Despite the evidence of *Alu* sequence mobility, individuals from human populations appear to have their *Alu* sequences in the same places. For example, there are eight copies of the *Alu* sequences in the normal β -like globin cluster (ALLAN and PAUL 1984), and in each of at least 250 haplotypes for the cluster examined to date (JEFFREYS 1979; ANTONORAKIS *et al.* 1984), all these copies, but no others, have been found.

This result can be incorporated into the above model in a very trivial way by the use of a θ value very much less than 1, such that almost all transposable

element sites are fixed in the population. This conforms to special case 2, where transposition rate is low enough to limit homogenization within a family.

In this case, $\bar{T} \approx \Lambda/2\mu$, and since μ is unknown but, if $\theta \ll 1$, must be very much less than $1/4N_e$, $\bar{T} \gg 2\Lambda N_e$.

As $\Lambda = 300,000$, and N_e is an estimate of the effective human population size over evolutionary time, which must be at the very least 10^4 , $\bar{T}$ estimates will range upwards from 10^{10} generations. Such absurdly high estimates demonstrate that the observed sequence conservation of *Alu* sequences cannot be due simply to the identity by descent expected to arise between copies as a result of the homogenizing effect of the replicative transposition of 300,000 independent and functionally equivalent element copies per genome. Thus, in this case, the model is grossly wrong.

INACCURACY IN THE MODEL

The model proposed is extremely simplistic and, thus, will inevitably be an inaccurate description of some, and probably most, interspersed repetitive sequence families. In particular, it hypothesizes the sequence identity arising by sampling the results of independent transpositions as the single mechanism for maintaining family homogeneity. At least four alterations to the model (which are not mutually exclusive) could include ways in which sequence homogeneity greater than that predicted above could arise, most of which have also been discussed by OHTA (1984):

1. If transposable elements at different genomic sites were capable of converting each other, then homogenization could arise by this mechanism. If gene conversion is unbiased, it would significantly increase homogenization rates only for those sequences with low θ values, where transposition to new sites is limiting to homogenization process. (This is provided the gene conversion rate per element copy per generation is very much less than 1, which is virtually certain). If gene conversion is biased in favor of some sequence variants, however homogenization could arise for reasons discussed below. *Ty1* transposable elements in yeast have been shown to convert each other (ROEDER and FINK 1982). There is no evidence that *copia*-like sequences are involved in gene conversion. The *Alu* sequences 5' to the δ -globins in man and the chimpanzee (MAEDA, BLISKA and SMITHIES 1983) show sequence conservation as strong as that for the noncoding sequences in which they are embedded, showing that they have not been differentially converted by extraneous *Alu* sequences in the few million years since they were separated. In the area 5' to the duplicated human α -globin genes, however, there exist *Alu* sequences at identical positions which show a 12% sequence divergence that is greater than that for most of the duplicated flanking sequences around these genes (HESS, SCHMID and SHEN 1984). This is a divergence figure comparable to that between random *Alu* sequences from the genome and, thus, is consistent with one of these sequences having been converted.

OHTA (1985) has extended her model for the identity coefficients, between both allelic and nonallelic copies of repetitive DNA sequences, expected to arise as a result of a combination of duplicative transposition and unbiased

gene conversion. She shows that transposition and unbiased gene conversion have very similar effects on the identity coefficients of nonallelic repetitive sequences, but that identity coefficients of allelic repetitive sequences will generally be lower if the homogenization process is gene conversion rather than transposition. This result is a consequence of the assumption that transposition does not repeatedly introduce different copies of a transposable element family into the same genomic site in different individuals.

2. The effective value of Λ could be less than the observed value. The model presented here assumes that all copies of a transposable element in the genome have equal transposition probabilities. The *Alu* sequence is homologous to outer parts of the 7SL RNA gene, which is known to function as part of the signal recognition particle (ULLU and TSCHUDI 1984). There are very many fewer such 7SL RNA genes in the genome than there are *Alu* sequences, and if it were the case that all *Alu* sequences had been derived by reintegration of 7SL RNA transcripts and not by transposition of other *Alu* sequences, then two consequences would be, first, that the effective value for Λ would be close to the number of 7SL RNA genes, rather than 300,000, and, second, that the neutral mutation rate, v , would be reduced.

3. Variation in μ , Λ , and N : If these parameters of the model are themselves time-dependent variables, then $\bar{T}$ will not be dependent on their current values, but on a quantity calculated from their values over a period of time. Suppose the population goes through a cycle of n states, with the states differing in their μ , Λ , and N and, therefore, θ values. I shall call these quantities values in the i th state μ_i , Λ_i , and θ_i . The population starts the cycle with a value for $\bar{T}$ of $\bar{T}_0$. Now allow a short period of time, δt , in which time $\bar{T}$ goes to $\bar{T} + \delta\bar{T}$. It is clear from the above arguments that

$$\bar{T}_0 + \delta\bar{T} = (1 - 2\mu\delta t)(\bar{T}_0 + \delta t) + 2\mu\delta t \left(1 - \frac{1}{\Lambda(1 + \theta)}\right)(\bar{T}_0 + \delta t).$$

If δt is one generation, and if, during this generation, μ , Λ , and N are μ_1 , Λ_1 , and N_1 , then

$$\bar{T}_0 + \delta\bar{T} = \bar{T}_1 \simeq \bar{T}_0 + 1 - \frac{2\mu_1\bar{T}_0}{\Lambda_1(1 + \theta_1)}.$$

$$\text{Let } \frac{2\mu_1}{\Lambda_1(1 + \theta_1)} = x_1$$

$$\bar{T}_1 \simeq 1 + \bar{T}_0(1 - x_1).$$

Thus,

$$\bar{T}_2 = 1 + \bar{T}_1(1 - x_2) = 1 + 1 - x_2 + \bar{T}_0(1 - x_1)(1 - x_2).$$

But as $x \ll 1$,

$$\bar{T}_2 \simeq 2 + \bar{T}_0(1 - x_1 - x_2),$$

generally

$$\bar{T}_j \simeq j + \bar{T}_o \left(1 - \sum_{i=1}^j x_i \right).$$

At the end of the cycle we have

$$\bar{T}_n \simeq n + \bar{T}_o \left(1 - \sum_{i=1}^n x_i \right).$$

But as it is the end of the cycle $\bar{T}_n = \bar{T}_o$, thus

$$\bar{T}_o = \bar{T}_o + n - \bar{T}_o \sum_{i=1}^n x_i$$

$$\frac{1}{\bar{T}_o} = \frac{\sum_{i=1}^n x_i}{n}.$$

Thus, approximately, the value of $\bar{T}$ is constant during the cycle and is equal to the reciprocal of the arithmetic mean value of $\frac{2\mu}{\Lambda(1+\theta)}$ during the cycle. In other words, $\bar{T}$ is approximately equal to the harmonic mean value of $\frac{\Lambda(1+\theta)}{2\mu}$.

This result will be true only if the variation in x is small and the cycle is limited in duration, such that the approximation

$$\prod_{i=1}^j (1 - x_i) \simeq 1 - \sum_{i=1}^j x_i \text{ holds.}$$

Expansions or contractions in the sizes of sequence families are probable, and the homogeneity of copies will be far more critically dependent on smaller family sizes than larger ones. It seems clear that such changes have occurred in the *Alu* family and related sequences in primate evolution (DANIELS *et al.* 1983). Thus, attempts to model sequence homogeneity by examining equilibrium states of dynamic transposition models with time-invariant parameters are inappropriate.

4. Transpositional or conversional advantage may occur: Occasionally it may occur that one copy of a transposable element in one individual is mutated in a way such that it has an increased probability of transposition or a decreased probability of loss. This advantage will tend to favor the element and its descendents relative to the rest of the copies in the population, and, if the total number of element copies in the population is regulated independent of mutational changes within the elements themselves, the higher "fitness" of the newly mutated element may cause it to replace the other elements in the population. If there is neutral sequence variation between the elements, the advantageous mutation will occur in only one of the neutral variants, and, as it becomes fixed, it will carry with it the neutral sequence variant in which it

originally arose. This is because the advantageous variant will be allelic to other variants in only a small proportion of cases during its spread through the population, so recombination will have little opportunity for creating linkage equilibrium between the advantageous mutation and the preexisting neutral sequence variation. Thus, the occasional replacement through transpositional advantage of all copies of the sequence by a new one will homogenize all the copies within the species. This could, alternatively, occur through an advantage in gene conversion, rather than in transposition, with similar results.

Quantitatively, however, this process is more complicated than it might initially appear. The main problem is that the postulated advantageous increase in transposition rate of a new sequence variant, $\delta\mu$, will be a very small quantity, of the order of μ itself, and certainly less than 10^{-3} . Since the new variant will be found initially as a rare variant at one or a few genomic sites, its numbers will be subject to sampling drift each generation, and, as its selective advantage over other variants is only $\delta\mu$, only a proportion $2\delta\mu$ of such variants will be fixed, a result noted by OHTA (1983) for the case of an advantage in conversion.

This will be true in transposition only if the new mutant is at low frequency wherever it occurs, which would be expected only if $\mu + \delta\mu > 1/2N$.

If advantageous variants occur rarely, it is possible to calculate the expected time between successive mutations in transposable elements, which, as a result of the transpositional advantage they confer, become fixed in the population of transposable elements. If the rate of mutation to new advantaged copies is r per copy per generation, the total rate of mutation across all copies will be $2NAr$, since there are $2NA$ copies of the transposable element in the population. Of these, only a proportion $2\delta\mu$ will come to be fixed; thus, the total rate of production of mutations which subsequently become fixed will be $4NAr\delta\mu$. Thus, the expected time between such substitutions will be $1/4NAr\delta\mu$, a result that would be expected from single-locus population genetics theory (KIMURA 1968).

This result will be true only if the expected time between substitutions is long compared to the time taken for a mutation to spread to near-fixation. Since the time taken for a new mutant to spread will be largely determined by size of the selective advantage $\delta\mu$, it is probable that the time between substitutions will be very much greater than the time taken for an individual mutation to spread if $4NAr \ll 1$.

The effect of an advantageous variant substitution will essentially be that, if copies of an element are sampled at random, then their common ancestor will have existed at approximately the last time an advantageous mutation arose, or more recently. The expected time since this event, assuming a Poisson substitution process, will be the expected interval between substitutions, which is $1/4NAr\delta\mu$. It is instructive to compare this with $\bar{T}$ for where $\theta \gg 1$, which is $2NA$. For selective homogenization to be more important than the drift homogenization postulated in the argument leading to T ,

$$\frac{1}{4NAr\delta\mu} < 2NA$$

$$r\delta\mu > \frac{1}{8(N\Delta)^2}.$$

DISCUSSION

This quantitative analysis makes two strong assumptions. The first is that individual element families are independently regulated. This is particularly evident in the discussion of new selected variants given above. If the opposite, that the copy numbers of individual transposable element families were free to vary, were true, then the concept of the homogenization of *copia* sequences—for example, by a new advantageous *copia* arising and selectively eliminating other *copia* sequences—would entail the evolved *copia* going on to replace all 412, 297, *Foldback* and other sequences. It is not known whether *Drosophila* transposable element families are independently regulated in their numbers. DOWSETT and YOUNG (1982) report large variation in the abundances of elements between *Drosophila* species, yet YOUNG (1979) finds conservation between *D. melanogaster* strains in the numbers of element copies of the various families. Although YOUNG does not interpret the latter result in this way, each of these observations is consistent with a lack of individual family copy number regulation, with the discrepancy between the two results being due to the different time periods of independent evolution of the populations being compared. If element families are not regulated and can drift up and down in abundance, modeling of the kind attempted here is not possible, as family extinctions will occur without a compensatory process generating new element families; thus, there will be no stationary distribution. The addition to the model of a description of such a family-generating process would be too speculative to be of any real value.

Evidence that *Drosophila* transposable element families are functionally distinct comes from discrepancies between the lengths of the short direct repeats of target site DNA generated by different elements when they insert (SPRADLING and RUBIN 1981), and from the fact that the suppressibility of transposable element insertion mutations by specific mutations at other loci depends on the transposable element inserted, not on the locus that is mutated (JACKSON 1984; MODOLELL, BENDER and MESELSON 1983).

The second major assumption is that transposable elements are mobile only within and not between genomes. The structural similarities between *Ty1* elements in yeast (FINK *et al.* 1981), *copia*-like sequences in *Drosophila* (RUBIN 1983) and vertebrate retroviruses have led to the suggestion that these various sequences are evolutionarily related (SHIMOTOHNO and TEMIN 1981; MAJORS *et al.* 1981; ELDER, LOH and DAVIS 1983). Thus, *copia*-like sequences may be capable of movement between *Drosophila* species in viral particles. If this is so, predictions of evolutionary divergence between family copies, based on models which do not include interspecific movement, inevitably will be wrong. However, the majority of *D. melanogaster* elements studied (DOWSETT 1983; MARTIN, WIERNASZ and SCHEDL 1983; BROOKFIELD, MONTGOMERY and

LANGLEY 1984) have been found only in other *Drosophila* species phylogenetically closely related to *D. melanogaster*. This implies that movement of sequences between species is very rare or absent for most families, although one sequence, the *P* factor, does appear to have moved into *D. melanogaster* horizontally from a distantly related species (DANIELS *et al.* 1984).

The main requirement now is for data revealing the extent of within-species divergence in *Drosophila* transposable elements compared to the incremental variation found between species. Such data will reveal the extent to which the modeling of transposable element evolution outlined here is adequate.

I would like to thank B. CHARLESWORTH, C. H. LANGLEY, M. SLATKIN, T. OHTA and an unknown referee for useful comments on this manuscript. I would also like to thank, in particular, M. SLATKIN and T. OHTA for showing me unpublished manuscripts.

LITERATURE CITED

- ALLAN, M. and J. PAUL, 1984 Transcription *in vivo* of the *Alu* family member upstream from the human ϵ -globin gene. *Nucleic Acids Res.* **12**: 1193–1200.
- ANTONORAKIS, S., C. D. BOEHM, G. R. SARGEANT, C. E. THEISEN, G. J. DOVER and H. KAZAZSIAN, JR., 1984 Origin of the β^{H} -globin gene in blacks: the contribution of recurrent mutation or gene conversion or both. *Proc. Natl. Acad. Sci. USA* **81**: 853–856.
- BOUCHARD, R. A., 1982 Moderately repetitive DNA in evolution. *Int. Rev. Cytol.* **76**: 113–193.
- BRITTEN, R. J. and E. H. DAVIDSON, 1969 Gene regulation for higher cells: a theory. *Science* **165**: 349–357.
- BROOKFIELD, J. F. Y., E. MONTGOMERY and C. H. LANGLEY, 1984 Apparent absence of transposable elements related to the *P* elements of *D. melanogaster* in other species of *Drosophila*. *Nature* **310**: 330–332.
- CHARLESWORTH, B. and D. CHARLESWORTH, 1983 The population dynamics of transposable elements. *Genet. Res.* **42**: 1–27.
- DANIELS, G. R., G. M. FOX, D. LOEWENSTEINER, C. W. SCHMID and P. L. DEININGER 1983. Species-specific homogeneity of the primate *Alu* family of repeated DNA sequences. *Nucleic Acids. Res.* **11**: 7579–7593.
- DANIELS, S. B., L. D. STRAUSBAUGH, L. EHRMAN and R. ARMSTRONG, 1984 Sequences homologous to *P* elements occur in *Drosophila paulistorum*. *Proc. Natl. Acad. Sci. USA* **81**: 6794–6797.
- DAVIDSON, E. M. and R. J. BRITTEN, 1979 Regulation of gene expression: possible role of repetitive sequences. *Science* **204**: 1052–1059.
- DOOLITTLE, W. F. and C. SAPIENZA, 1980 Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**: 601–603.
- DOWSETT, A. P., 1983 Closely related species of *Drosophila* can contain different libraries of middle repetitive DNA sequences. *Chromosoma* **88**: 104–108.
- DOWSETT, A. P. and M. W. YOUNG, 1982 Differing levels of dispersed repetitive DNA among closely related species of *Drosophila*. *Proc. Natl. Acad. Sci. USA* **79**: 4570–4574.
- EIBEL, H., J. GAFNER, A. SLOTZ and P. PHILLIPSEN, 1981 Characterization of the yeast mobile element *Ty1*. *Cold Spring Harbor Symp. Quant. Biol.* **45**: 609–618.
- ELDER, R. T., E. Y. LOH and R. W. DAVIS, 1983 RNA from the yeast transposable element *Ty1* has both ends in the direct repeats, a structure similar to retrovirus RNA. *Proc. Natl. Acad. Sci. USA* **80**: 2432–2436.

- EWENS, W. J., 1979 *Mathematical Population Genetics*. Biomathematics Series, Vol. 9, p. 325. Springer-Verlag, New York.
- FINK G., P. FARABAUGH, G. S. ROEDER and D. CHALEFF, 1981 Transposable elements (*Tyl*) in yeast. Cold Spring Harbor Symp. Quant. Biol. **45**: 575-580.
- HESS, J. R., M. FOX, C. SCHMID and C-K. J. SHEN, 1983 Molecular evolution of the human adult 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.
- HESS, J. F., C. W. SCHMID and C-K. J. SHEN, 1984 A gradient of sequence divergence in the human adult α -globin duplication units. Science **226**: 67-70.
- JACKSON, I., 1984 Transposable elements and suppressor genes. Nature **309**: 751-752.
- JAGADEESWARAN, P., B. G. FORGET, and S. M. WEISSMAN, 1981 Short interspersed repetitive DNA elements in eucaryotes: transposable DNA elements generated by reverse transcription of RNA PolIII transcripts? Cell **26**: 141-142.
- JEFFREYS, A. J., 1979 DNA sequence variants in the $\epsilon\gamma$ -, $\alpha\gamma$ -, δ -, and β -globin genes in man. Cell **18**: 1-10.
- JEFFREYS, A. J., 1982 Evolution of globin genes. pp. 157-176. In: *Genome Evolution*, Systematics Association Series: Vol. 20, Edited by G. A. DOVER and R. B. FLAVELL. Academic Press, New York.
- JELINEK, W. R. and C. W. SCHMID, 1982 Repetitive sequences in eukaryotic DNA and their expression. Annu. Rev. Biochem. **51**: 813-844.
- KAPLAN, N. L. and J. F. Y. BROOKFIELD, 1983a Transposable elements in Mendelian populations. III. Statistical results. Genetics **104**: 485-495.
- KAPLAN, N. L. and J. F. Y. BROOKFIELD, 1983b The effect of selective differences between sites of transposable elements on homozygosity. Theor. Pop. Biol. **23**: 273-280.
- KIMURA, M., 1968 Evolutionary rate at the molecular level. Nature **217**: 624-626.
- KIMURA, M. and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics **49**: 725-738.
- KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304**: 412-416.
- KROWLEWSKI, J. J. and M. G. RUSH, 1984 Some extrachromosomal circular DNAs containing the *Alu* family of dispersed repetitive sequences may be reverse transcripts. J. Mol. Biol. **174**: 31-40.
- LANGLEY, C. H., J. F. Y. BROOKFIELD and N. L. KAPLAN, 1983 Transposable elements in Mendelian populations. I. A theory. Genetics **104**: 457-472.
- LANGLEY, C. H., E. MONTGOMERY and W. R. QUATTLEBAUM, 1982 Restriction map variation in the *Adh* region of *Drosophila*. Proc. Natl. Acad. Sci. USA **79**: 5631-5635.
- MAEDA, N., J. B. BLISKA and O. SMITHIES, 1983 Recombination and balanced chromosome polymorphism suggested by DNA sequences 5' to the human δ -globin gene. Proc. Natl. Acad. Sci. USA **80**: 5012-5016.
- MAJORS, J. E., R. SWANSTROM, W. J. DELORBE, G. S. PAYNE, S. H. HUGHES, S. ORTIZ, N. QUITRELL, J. M. BISHOP and H. E. VARMUS, 1981 DNA intermediates in the replication of retroviruses are structurally (and perhaps functionally) related to transposable elements. Cold Spring Harbor Symp. Quant. Biol. **45**: 719-730.
- MARTIN, G., D. WIERNASZ and P. SCHEDL, 1983 Evolution of *Drosophila* repetitive-dispersed DNA. J. Mol. Evol. **19**: 203-213.
- MODELELL, J., W. BENDER and M. MESELSON, 1983 *Drosophila melanogaster* mutations suppressible by the suppressor of hairy wing are insertions of a 7.3-kilobase mobile element. Proc. Natl. Acad. Sci. USA **80**: 1678-1682.

- MONTGOMERY, E. A. and C. H. LANGLEY, 1983 Transposable elements in Mendelian populations. II. Distribution of three *copia*-like elements in a natural population of *Drosophila melanogaster*. *Genetics* **104**: 473-483.
- OHTA, T., 1982 Allelic and nonallelic homology of a supergene family. *Proc. Natl. Acad. Sci. USA* **79**: 3251-3254.
- OHTA, T., 1983 Theoretical study on the accumulation of selfish DNA. *Genet. Res., Camb.* **41**: 1-15.
- OHTA, T., 1984 Population genetics of transposable elements. *IMA J. Math. Appl. Med. Biol.* **1**: 17-29.
- OHTA, T., 1985 A model of duplicative transposition and gene conversion for repetitive DNA families. *Genetics* **110**: 513-524.
- ORGEL, L. E. and F. H. C. CRICK, 1980 Selfish DNA: the ultimate parasite. *Nature* **284**: 604-606.
- ROEDER, G. S. and G. R. FINK, 1982 Movement of yeast transposable elements by gene conversion. *Proc. Natl. Acad. Sci. USA* **79**: 5621-5625.
- RUBIN, G. M., 1983 Dispersed repetitive DNAs in *Drosophila*. pp. 329-361. In: *Mobile Genetic Elements*, Edited by J. A. SHAPIRO. Academic Press, New York.
- RUBIN, G. M., W. J. BROREIN, P. DUNSMUIR, A. J. FLAVELL, R. LEVIS, E. STROBEL, J. J. TOOLE and E. YOUNG, 1981 *Copia*-like transposable elements in the *Drosophila* genome. *Cold Spring Harbor Symp. Quant. Biol.* **45**: 619-628.
- SAPIENZA, C. and W. F. DOOLITTLE, 1981 Genes are things you have whether you want them or not. *Cold Spring Harbor Symp. Quant. Biol.* **45**: 177-182.
- SCHULER, L. A., M. J. L. WEBER and J. GORSKI, 1983 Polymorphism near the rat prolactin gene caused by insertion of an *Alu*-like element. *Nature* **305**: 159-160.
- SHIMOTOHNO, K. and H. M. TEMIN, 1981 Evolution of retroviruses from cellular movable genetic elements. *Cold Spring Harbor Symp. Quant. Biol.* **45**: 719-730.
- SLATKIN, M., 1985 Genetic differentiation of transposable elements under mutation and unbiased gene conversion. *Genetics* **110**: 145-158.
- SPRADLING, A. C. and G. M. RUBIN, 1981 *Drosophila* genome organization: conserved and dynamic aspects. *Annu. Rev. Genet.* **15**: 219-264.
- STEPHENS, J. C., M. KREITMAN and M. NEI, 1984 Phylogenetic analysis of the *Adh* "fast-slow" variation in *D. melanogaster*: age of the polymorphism. *Genetics* **107** (Suppl): s103.
- ULLU, E. and C. TSCHUDI, 1984 *Alu* sequences are processed 7SL RNA genes. *Nature* **312**: 171-172.
- VARMUS, H. E., 1983 Retroviruses, pp. 411-503. In: *Mobile Genetic Elements*, Edited by J. A. SHAPIRO. Academic Press, New York.
- YOUNG, M. W., 1979 Middle repetitive DNA: a fluid component of the *Drosophila* genome. *Proc. Natl. Acad. Sci. USA* **76**: 6274-6278.

Communicating editor: B. S. WEIR