A Source of Small Repeats in Genomic DNA

Dan Fieldhouse and Brian Golding

Department of Biology, York University, North York, Ontario M3J 1P3, Canada

Manuscript received December 10, 1990
Accepted for publication June 10, 1991

ABSTRACT

The processes of spontaneous mutation are known to be influenced by neighboring DNA. Imperfect nearby repeats in the neighboring DNA have been observed to mutate to form perfect repeats. The repeats may be either direct or inverted. Such a mutational process should create perfect direct and inverted repeats in intergenic DNA. A larger than expected number of direct repeats has generally been observed in a wide range of species in both coding and noncoding DNA. Simulations are carried out to determine how this process might influence the repetitive structure of genomic DNA. These simulations show that small repeats created by this kind of a mutational process can explain the excess number of repeats in intergenic DNA. The simulations suggest that this mechanism may be a common cause of mutations, including single-base changes. The influences of the distance between imperfect repeats and of their degree of similarity are investigated.

SPONTANEOUS mutations are caused by a variety of unspecified mechanisms. Most of the mutagenic agents that have been investigated *in vitro* will generate nonrandom and often highly unique spectra of mutations. These demonstrate that the common practice among many scientists of bestowing on mutations a regularity and consistency that is normally attributed to radioactive decay is quite misleading.

One unusual pattern of mutation that has been observed are mutations that involve short repeats in the DNA. A typical DNA sequence will contain many short repeats in its sequence. It has been found that when two imperfect repeats are near each other they may mutate to become perfect repeats. A typical observation would be that prior to mutation, an imperfect repeat exists that differs by one or more nucleotides and may additionally differ by insertions or deletions. After a mutational event the repeats are identical in sequence.

This pattern has been observed in a wide variety of organisms ranging from humans, yeast, *Escherichia coli* to T4 bacteriophage (RIPLEY and GLICKMAN 1983; DRAKE, GLICKMAN and RIPLEY 1983; HAMPSEY *et al.* 1988; PAPANICOLAOU and RIPLEY 1989; SOLOV'EV, ROGOZIN and KOLCHANOV 1989; HARTNETT *et al.* 1990). That imperfect repeats may mutate to become perfect repeats has also been observed to occur *in vitro* (KUNKEL and ALEXANDER 1986; KUNKEL and BEBENEK 1988). This mechanism is also reflected in the types of substitutions that are fixed through evolution (GOLDING and GLICKMAN 1985, 1986). These mutations are the results of a polymerase or some other repair enzyme using an incorrect but preexisting template. Hence this type of mutation has been termed sequence directed mutagenesis. The actual mechanisms behind these mutations are unknown but possibilities include incorrect error repair, template slippage during synthesis (STREISINGER *et al.* 1966), strand switch during replication (PAPANICOLAOU and RIPLEY 1989) or some kind of small conversion event. We will call these mutations either sequence directed or small repeat conversions but this terminology is meant to reflect only the results of the mutational event and is not intended to reflect a discrimination among mechanisms.

Mutations that occur via such a process have many consequences for the structure of DNA. For example, these processes imply that mutational events will often involve several base changes and occasionally a mixture of base changes coincidental with deletions/insertions. HAMPSEY *et al.* (1988) estimate that as much as 10% of yeast mutations involve multiple events and feel that the majority of these are candidates for this type of mutational mechanism.

Another consequence involves the rate of mutational change. A molecular clock was suggested to exist by ZUCKERKANDL and PAULING (1965) whereby the rate of evolution is constant and reflects a constant rate of mutation. This hypothesis gains support from the neutral theory of molecular evolution (KIMURA 1983). A large body of protein data was initially suggested to prove that such a clock exists (WILSON, CARLSON and WHITE 1977). However, GILLESPIE (1984) has suggested that the rate of evolution does not follow a clocklike behavior but rather that the long time spans involved simply mask the fluctuations. Analyses of molecular sequences have shown that the rate of evolution is significantly overdispersed, with a variance that is much larger than expected from simple clocklike behavior. Hence a neutral model (with constant rates of mutation) can not apply. If, however, neutral mutations do not occur via a simple mutational mechanism then the basic assumptions of a molecular clock can be questioned. When mutations occur via a sequence directed mechanism this leads to an inflated variance to mean ratio of substitutional changes (GOLDING 1987). Repeat induced point (RIP) mutations will cause a similar phenomenon (CAMBARERI *et al.* 1989). This suggests that some of the episodic nature of evolutionary change (GILLESPIE 1986; see also TAKAHATA 1987) may be explained simply by the pattern of mutation.

Still another consequence of this mechanism would be the steady production of repeats in intergenic DNA. These would be created by sequence directed mutagenesis and then would slowly be eliminated by other types of mutations. It has been observed for some time that there are often excess numbers of direct and inverted repeats in genomic DNA (KARLIN *et al.* 1983; KARLIN and GHANDOUR 1985). We examine here the number of repeats in intergenic DNA and confirm that the number of direct repeats is larger than expected. Simulations are used to determine if sequence directed mutagenesis could explain this excess and to determine what some of the properties of such a process might be.

MATERIALS AND METHODS

Calculation of the expected number: To confirm that repeats are present within intergenic DNA sequences in unusually large numbers, observed numbers were compared to expectations calculated from random sequences. Since the base bias in intergenic DNA can be significant and since this can affect the expected number of repeats, the observed base bias was matched in the randomly generated sequences.

This was done using biased choices for each nucleotide depending on the range of a random number. For example, inserting a G if the number is less than 0.25, an A if the number is between 0.25 and 0.5, a T if the number is between 0.5 and 0.75, and a C if the random number is larger than 0.75 will give a 1:1:1:1 ratio of G:A:T:C. A base bias similar to that found in human δ/β-globin intergenic DNA can be matched using random numbers such that a G is inserted if the random number is less than 0.194, an A if the number is between 0.194 and 0.493, a T if the number is between 0.493 and 0.800, and a C if the random number is larger than 0.800. This collection of numbers is not the only collection that would give the observed bias. This particular collection was found by trial and error using another program. This other program simply iterates a matrix of equations describing the frequency of bases over time with mutation. This collection of numbers is one which did not alter the observed base frequencies after many generations.

In this way sequences were randomly generated with the same base bias as an observed sequence. Five hundred such sequences were generated and the number of 100% identical direct repeats were counted for each sequence. The mean and standard error of the number of direct repeats was compared to the observed values.

Simulation of repeat conversions: The effect of small direct repeat conversion on a randomly mutating sequence was simulated by permitting both of these processes (in their simplest forms) to occur simultaneously. All simulations started with 1000 base pairs (bp) of observed intergenic DNA. The number of repeats in this DNA was followed over time as both point mutations and as small repeat conversions occur.

Again the simulation accounts for differences in overall base composition. The base composition is taken from the observed value of 1000 bp of intergenic DNA. Throughout the simulations, the number of point mutations and the number of mutations caused by sequence directed mutagenesis are recorded. The rates and the ratio of repeat conversion events to point mutations can be altered by varying parameters.

Point mutations are permitted to occur at a given rate each generation. This rate is set such that the average number of point mutations is either ten or four per generation. The simulation is continued for 2000 generations.

Every ten generations, all of the repeats in the sequence are found including those repeats which are 75% imperfect matches. Then small repeat conversions are permitted to occur between these repeats at a rate determined by the particular model considered (see below). Conversions between imperfect repeats will cause mutational changes to create perfect repeats.

The program finds all imperfect direct repeats ranging in size from 5 to 25 bp in a sequence and stores the location, length and percent identity of these repeats in an array. Repeats must be at least 75% identical to be considered as imperfect repeats and must have no terminal mismatches. Note that a repeat of length x is not considered if it is part of a repeat of length $x + 1$. In addition repeats must be within 100 bp of each other.

This repeat information is used to simulate sequence directed mutagenesis. Since the actual mechanism of this process is not known, it is difficult to determine the factors that might influence it. Two parameters that will probably affect the likelihood of mutation are the distance between two imperfect repeats and the degree of similarity between the two repeats. Again however, the actual functional relationship between these two parameters and the probability of mutation are unknown. As a first approximation we will assume that there is a linear relationship between the probability of mutation and the distance and the percent identity of the repeats.

Each individual repeat is considered for a possible repeat conversion event based on certain conditions. For each pair of repeats, if a random number is less than a critical value then either the first sequence of the repeat is changed to that of the second, or *vice versa*. Which repeat is changed is determined by a second random number being greater or less than a half. The critical value for the first random number is determined from three quantities: (i) the overall rate of mutation desired for sequence directed mutations, (ii) the degree of similarity between the repeats and (iii) the distance between the repeats. Hence, a mutational event will occur if the random number is less than μ where

$$\mu = K\{(3*G + 2*A)/(3*M) - (D/L)\}$$

where K is a parameter to determine the overall rate of mutation. If it is large more repeat conversions will occur than if it is small.

G and A are the number of G/C and A/T matched bases between the two repeats. This provides a measure of the percent identity between the two repeats.

M is the maximum repeat length permitted (25).

D is the distance between the first base of each repeat.

L is the length of the sequence (1000).

The first term (K) can be used to adjust the ratio of point mutations to sequence directed mutational events.

The second term measures the degree of similarity between two imperfectly matched repeats weighted by the number of hydrogen bonds. The ratio of hydrogen-bonds is calculated by multiplying the number of G/C's by 3 and the number of A/T's by 2 and dividing by the total number possible if all 25 (M) bases were G or C. This gives a linear relationship of the probability of a repeat conversion to the number of possible H-bonds which could form between the two repeats.

The last term is a measure of the distance between the repeats. Only repeats within 100 bp of each other are considered for potential conversion. The ratio of this distance to the total sequence length is subtracted from the critical value. This provides a linear relationship between distance and the probability of repeat conversion. The more distant the repeats, the less likely a repeat conversion.

This formula is not meant to be an accurate representation of the true biological situation. Rather it is meant only to provide flexibility to the model and to permit the influence of each of these factors to vary. To determine how the structure of genomic DNA may respond to variations in each of these factors, the simulations were run with different mixtures of dependencies on distance and percent identity between the repeats. Simulations were run with no repeat conversion ($K = 0$), with a dependence on both distance and percent identity, with a dependence only on identity (removing the last term of the equation), with a dependence on distance only (replacing the first term with 1.0), or with a random likelihood of repeat conversion that did not depend on either distance or percent identity (the only factor considered was K). The simulations were run ten times for each of these parameter combinations. For each run data are collected every ten generations for a total of 200 per run. The overall rates were adjusted such that there was a ratio of either 4:1, 10:1 or 10:0 point mutations for each repeat conversion. A total of 10 × 9 90 runs were done with each run taking approximately 5 hr CPU time on a Sun Sparcstation. The means of each run were tested for significant difference from either the starting values or from the grand mean of all runs without repeat conversion using Student's t-tests.

RESULTS

The expected and observed numbers of 100% identical repeats are shown in Table 1. The expected number is derived from 500 simulations of randomly generated sequences with the same base bias. The observed data are from human intergenic DNA near the amylase gene, from human intergenic DNA between the δ- and β-globins, from Drosophila intergenic DNA between the *hsp23* and *hsp27* heat shock loci, from intergenic DNA near the chick crystallin gene, intergenic DNA from *Xenopus laevis* histone gene, and intergenic DNA near the *Zea mays* tubulin gene. These regions show intergenic DNA with low, intermediate and high numbers of perfect direct repeats. A particular 1-kilobase (kb) region was chosen based on a visual inspection to exclude regions that have simple runs of nucleotides (*e.g.*, 'CACACACACAC') and are not known to have any coding or otherwise selected function.

It will be noted that the expected value for direct repeats longer than 10 bp is less than one yet there are several such repeats observed. In general the observed values are larger than the expected (30 out of 36 times for repeats less than 11 bp) and usually more than the expected plus one standard deviation (25 out of 36 times). Only the human amylase intergenic region shows numbers close to that expected and its repeats of length 5, 7 and 8 are all larger than expected.

Three examples of the data shown in Table 1 are plotted in Figure 1 on a semi-log plot. The observed values are indicated by the circles and the expected value by a "+" with associated standard deviations. A log linear relationship between the expected number and repeat length is easily observed. This is to be expected from the results of KARLIN *et al.* (1983). The excess of direct repeats is more easily seen here. A consistent excess of imperfect repeats (75% identical) is also observed (data not shown).

The simulations to observe the effects of mutations via repeat conversions generate data on repeats of length 5 and larger. The results from one simulation run for repeats of length 5, 8, and 10 are shown in Figure 2. These are generally representative of the results for repeats with other lengths. These graphs give the number of perfect repeats every 10 generations. The horizontal line indicates the observed values for human globin intergenic DNA and were used as the initial values for the simulation. The curves are an indication of trends in the data and are cubic splines drawn to the median of every 20 sequential points. Since these splines are forced to fit the data they may dip below zero. The first result that can be concluded from this data is that there is a great deal of variation in the number of repeats found in DNA over time. It can also be seen that an equilibrium expected number of repeats is very quickly reached. Repeats of length 10 do not usually occur in a 1000 bp sequence without sequence directed mutation and hence the value of many points on this graph is zero. This is clearly shown by the position of the spline. Repeats of length 5 seem to have a great deal more noise associated with their number due to their small length. The pattern shown by the repeats of length 5, 8 and 10 are typical of all runs with repeats of other lengths. Despite the fact that these simulations all begin with the observed number of repeats found in human globin intergenic DNA (shown by the horizontal line), in all cases they tend to quickly fall below this number. Clearly the repeats of length 5 and 8 are well below the initial value. The same is also true for repeats of length 10 but is less apparent due to the smaller number of such repeats. These results again indicate the excess number of repeats found in intergenic DNA and confirms the results of KARLIN *et al.* (1983).

TABLE 1

Comparison of the observed *vs.* expected number of repeats in 1000 bp of intergenic DNA

Repeat length	Human amylase intergenic region			Human globin intergenic region			Drosophila heat shock intergenic region		
	Obs.	Exp.	Std.	Obs.	Exp.	Std.	Obs.	Exp.	Std.
5	329	274.0	17.7	363	329.0	25.3	474	406.2	38.2
6	68	68.1	8.3	99	85.5	10.8	198	111.4	15.3
7	18	17.0	4.3	31	22.3	4.9	68	30.2	6.6
8	6	4.2	2.2	11	5.8	2.5	15.	8.2	3.1
9	0	1.1	1.1	1	1.6	1.3	15	2.2	1.5
10	0	0.3	0.5	1	0.4	0.6	8	0.6	0.8
11	0	0.1	0.3	0	0.1	0.3	7	0.2	0.5
12	0	0.0	0.1	0	0.0	0.2	1	0.1	0.2
13	0	NA	NA	0	NA	NA	0	0.0	0.1
14	0	0.0	0.1	0	NA	NA	1	NA	NA
15	0	NA	NA	0	NA	NA	2	NA	NA
	Chicken crystallin intergenic region			*Xenopus laevis* histone intergenic region			*Zea mays* tubulin intergenic region		
5	368	300.2	21.0	372	316.0	24.3	306	271.4	16.3
6	83	76.7	9.9	97	81.0	9.9	70	67.3	8.2
7	37	19.6	4.4	27	21.1	5.1	28	16.9	4.2
8	8	4.9	2.2	10	5.6	2.4	11	4.2	2.1
9	1	1.2	1.1	1	1.3	1.1	3	1.1	1.1
10	1	0.3	0.6	3	0.3	0.5	2	0.3	0.5
11	0	0.1	0.3	0	0.1	0.3	0	0.1	0.3
12	0	0.0	0.1	0	0.0	0.2	0	0.0	0.1
13	0	NA	NA	0	0.0	0.0	0	0.0	0.0
14	1	NA	NA	0	0.0	0.1	0	NA	NA
15	0	NA	NA	0	NA	NA	0	NA	NA

In addition, the Drosophila sequence also had one direct repeat of length 17 bp. An NA indicates that repeats of this length were not observed in 500 simulated sequences.

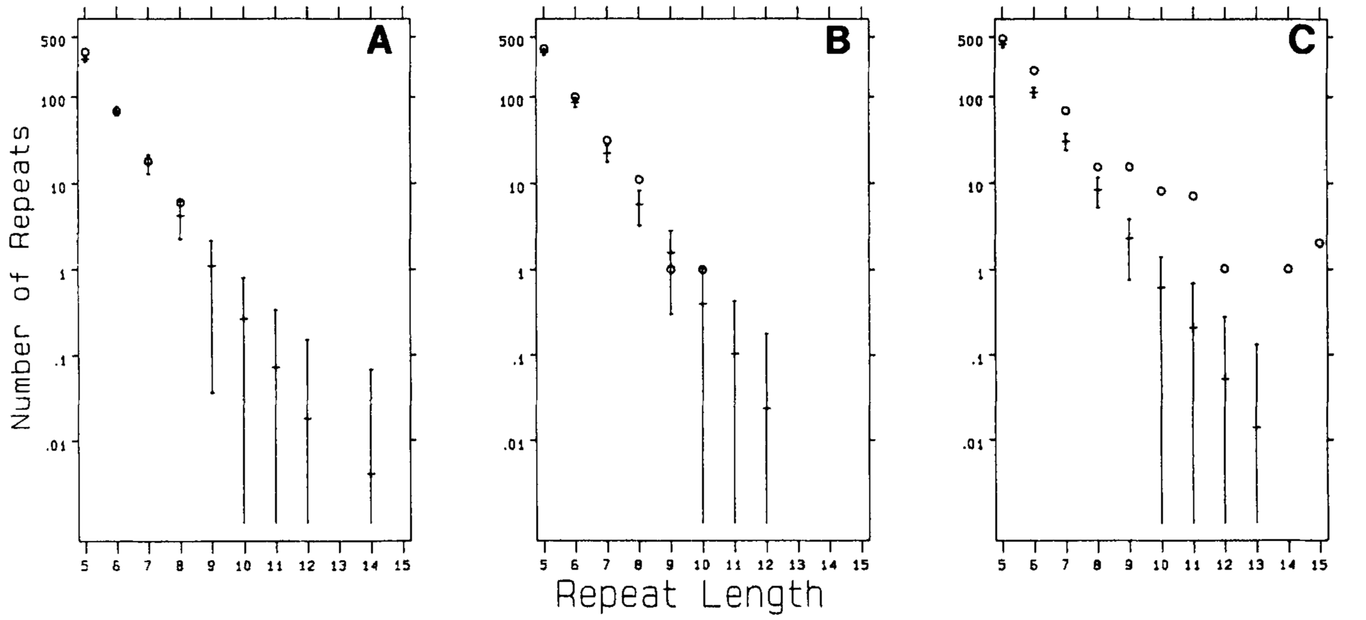

FIGURE 1.—Number of repeats observed and expected in the intergenic region around (A) human amylase genes, (B) human globin genes and (C) Drosophila heat shock genes. Drosophila heat shock genes also have one repeat of length 17. The observed values are indicated by circles and the expected values by a dash with standard deviations.

All ten independent runs of the simulation give similar results to those in Figure 2. The results for five of the runs for repeats of length 10 are shown in Figure 3. Note that there is variation over time within a single sequence and also variation in the number of repeats between runs of the simulation. Again the vast

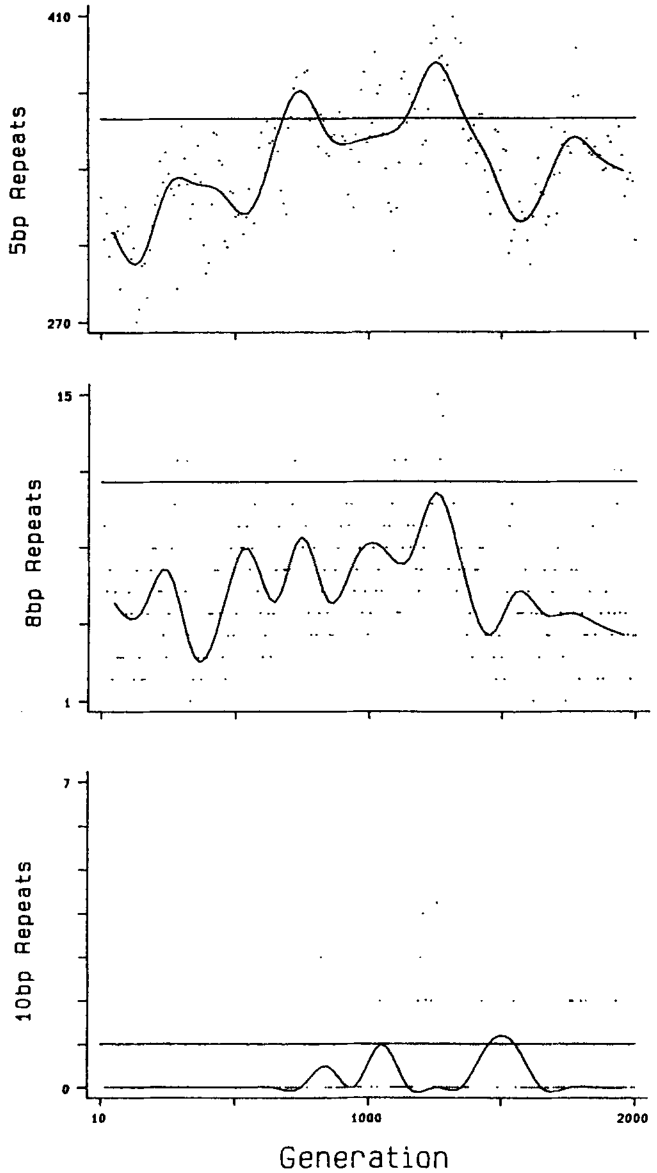

FIGURE 2.—The number of perfect repeats of length 5, 8 and 10 for a simulation which ran for 2000 generations. The horizontal line gives the number of perfect repeats observed in human globin intergenic DNA and the initial value for the simulation. The curve is a spline to show the general trends in the data. The number of repeats is tabulated every 10 generations.

majority of the data points fall well below the observed number of repeats in human globin intergenic DNA (the initial value of the simulations).

When sequence directed mutations are included the number of both imperfect and perfect repeats can be increased. Figure 4 gives the number of perfect repeats of length 10 for a single run when point mutation occurs at a rate of 10:1 relative to repeat conversions and at a rate of 4:1. Again, while the data presented are only from a single run, they are very representative of the other simulations. A ratio of 10:1 implies that repeat conversions are relatively common and represent approximately 10% of all mutations. Yet even at this level, the number of repeats still generally falls well short of the initial/observed value. Only when the repeat conversions represent one fifth of the mutations does the number increase to a level similar to that found in globins. The average number of repeats is shown in Figure 5 for a 10:1 ratio (solid line), a 4:1 ratio (short dashed line) and the observed values (diamonds, connected by a long dashed line excluding repeats of length 9 which appear to be under represented in the observed data). The number of repeats of all lengths are too few with a 10:1 ratio but closely approach the observed values with a 4:1 ratio. A 3:1 ratio of point mutations to repeat conversions begins to overshoot the observed numbers of repeats particularly for the larger repeat lengths. Since there appears to be a difference in the slope of these lines, the simulations appear to be biased toward preferentially increasing repeats of larger length and having less effect on smaller repeats.

Figure 6 gives the number of repeats of length 10 when there is no repeat conversion, with conversion dependent on distance and identity, dependent on distance alone, dependent on identity alone and with conversion simply a random choice among repeats. The relative rate of point mutations to repeat conversion mutations are 4:0, 4:1, 4:1, 4:1 and 4:1, respectively. In each case the number of repeats is increased by sequence directed mutations. In Figure 6D (a dependence only on identity) the simulated number is significantly larger than observed. In the other simulations with repeat conversion (Figure 6, B, C and E) the simulated number of repeats of length 10 are not significantly different from the observed value. Note that the rate at which repeat conversion must operate is quite large. With relative rates less than those in Figure 6, the number of repeats actually observed in intergenic DNA can not be matched.

The results of all runs for repeats of length 5, 8 and 10 are summarized in Table 2. Here the number of repeats for each simulation run are compared to the number actually observed in human globin intergenic DNA. Because the data represent time series data and hence are not statistically independent, a *t*-statistic is used to approximately compare the overall average from all 2000 generations to the number of repeats observed. The three numbers in each column of Table 2 represent the number of simulation runs that were significantly higher, not significantly different or significantly lower than the observed.

The repeats of length 5 are significantly less than observed for all parameter values used. Observed values fluctuate drastically but generally stay below their initial value. For repeats of length 6 or larger, repeat conversions will cause an increase in repeat numbers.

Without repeat conversion the number of repeats of length 8 and 10, in all ten runs, are significantly

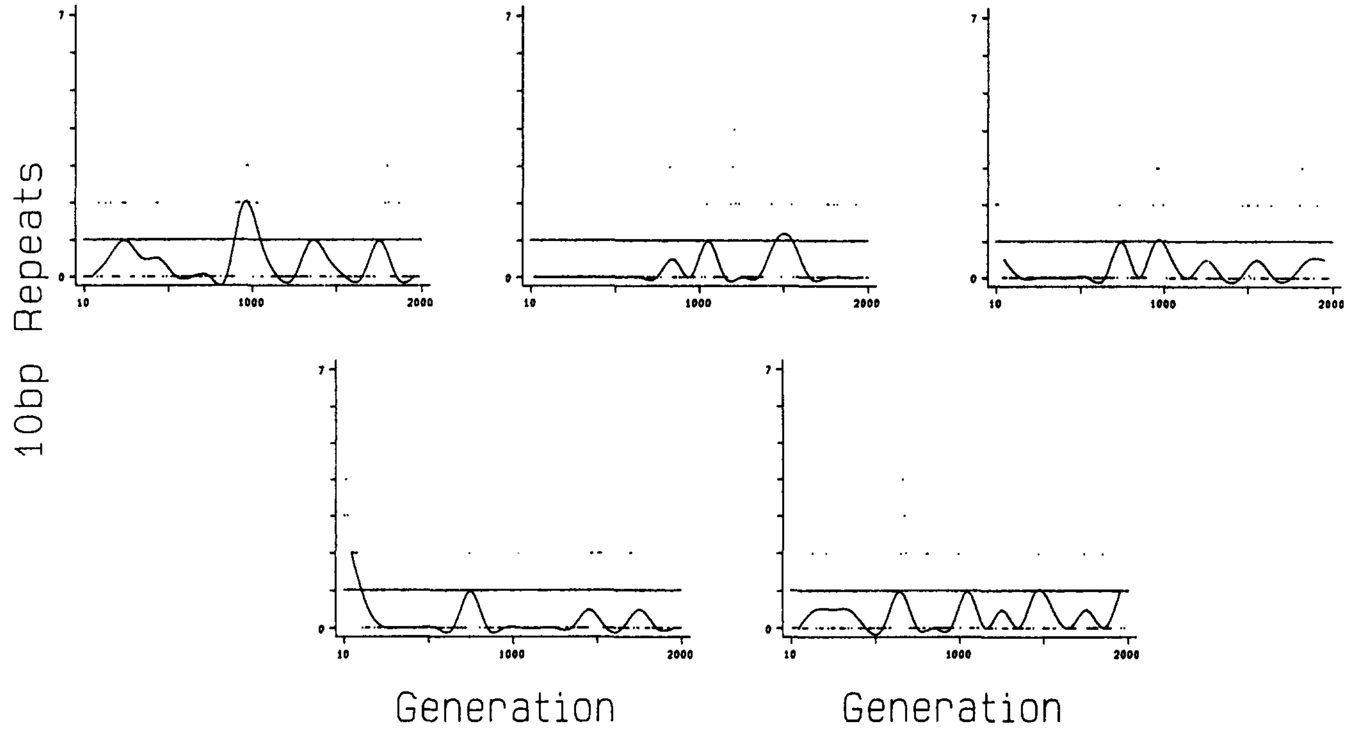

FIGURE 3.—The results of five of the simulations showing the number of perfect repeats of length 10 for 2000 generations. The horizontal line gives the number of perfect repeats observed in human globin intergenic DNA and the initial value for the simulation. The curve is a spline to show the general trends in the data. The number of repeats is tabulated every 10 generations.

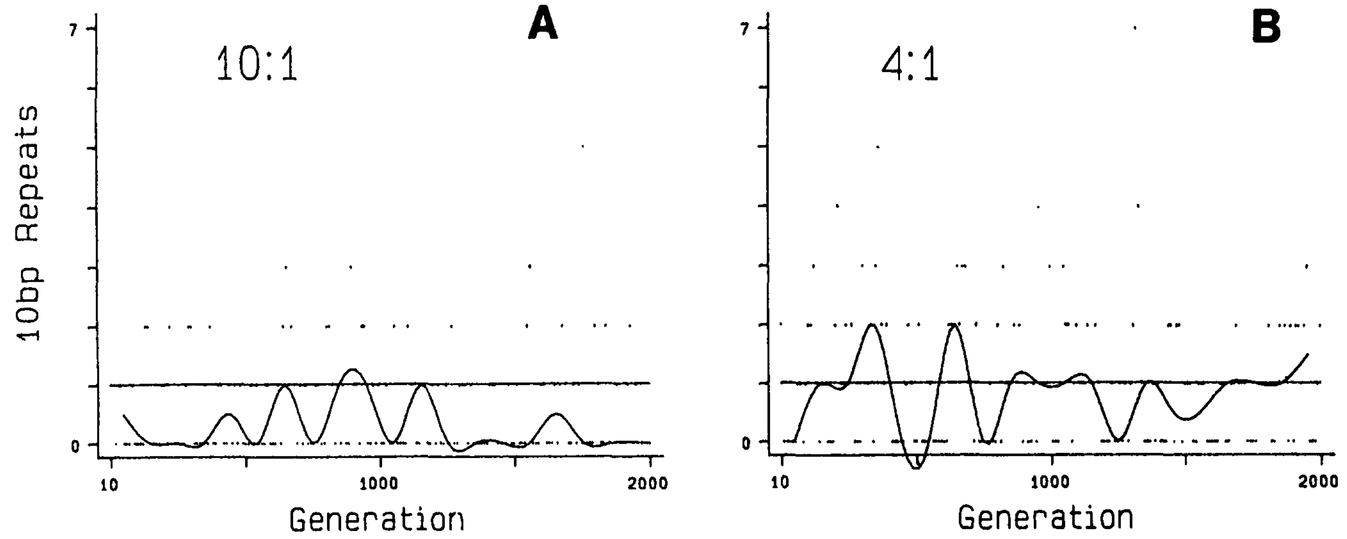

FIGURE 4.—The number of perfect repeats of length 10 when point mutations and repeat conversion occurs with a ratio of (A) 10:1 and (B) 4:1. The horizontal line gives the number of perfect repeats observed in human globin intergenic DNA and the initial value for the simulation. The curve is a spline to show the general trends in the data. The number of repeats is tabulated every 10 generations.

lower than the observed values. With repeat conversion the number of repeats of length 8 and 10 are not increased until the rate becomes quite large. Indeed with repeat conversion dependent on distance and identity the numbers of repeats of length 8 are still significantly smaller with both a 10:1 and 4:1 ratio. Repeat conversion events dependent only on the distance or only on the percent identity between repeats appear to increase the repeat number more effectively.

Table 3 compares the mean number of repeats with conversions to the mean number expected without conversions. These indicate that while the repeat conversions may not increase the number of repeats to such high levels as are actually observed, they do significantly increase the number of repeats beyond that expected in the absence of sequence directed mutagenesis. All of the repeats, including those of length 5, show a trend toward being significantly

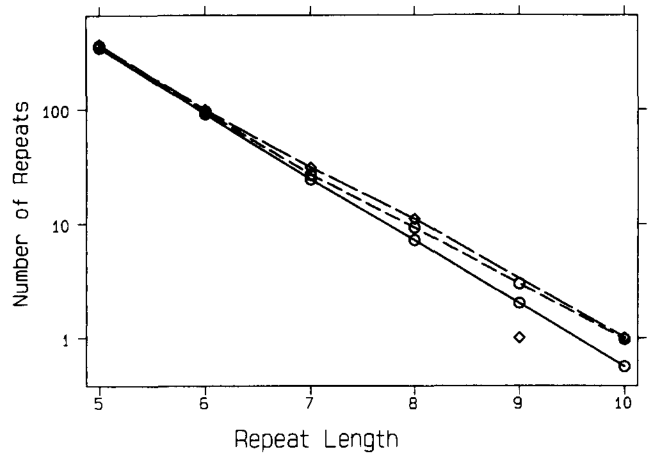

FIGURE 5.—The average number of repeats observed from the simulations with a 10:1 ratio of point mutations to repeat conversions (solid line), a 4:1 ratio (short dashed line) and the observed number of repeats (diamonds joined by a long dashed line–excluding observed repeats of length 9).

higher than expected in the absence of any repeat conversion.

The number of bases changed per repeat conversion event, when repeat conversions are dependent on both the distance between repeats and on their percent similarity, is shown in Table 4. Note the large number of events that alter only a single base between the two repeats. The distribution very quickly tails off with only a few repeat conversion events causing a large number of mutational changes.

DISCUSSION

Unfortunately very little is known about the details of the mechanism or mechanisms that might be responsible for sequence directed mutations. Therefore a computer simulation was used to study the generation of small repeats in DNA by this process. The simulation assumes that the repeats are generated in a simple fashion and ignores many of the other nonrandom features of spontaneous mutation. The simplistic assumptions made here about the nature of the processes involved may introduce unknown biases into the results. However, the results obtained should be quantitatively accurate for any process similar to that observed for sequence directed mutations. The two main assumptions are that identity between the repeats and the distance between repeats will affect the probability that a repeat conversion will occur. These assumptions are almost certainly true. It is also assumed that the sequences of the repeat do not shift during the repeat conversion event. Therefore, overlapping repeats are not considered. First and last base matches are arbitrary conditions used to establish the boundaries of a repeat. A 75% identity for imperfect repeats is also an arbitrary choice chosen simply to provide sufficient opportunity for repeat conversion without the necessity of keeping track of all imperfect repeats. The degree of identity will certainly affect the probabilities of mutational events. In a study of deletions caused by direct repeats, it was found that

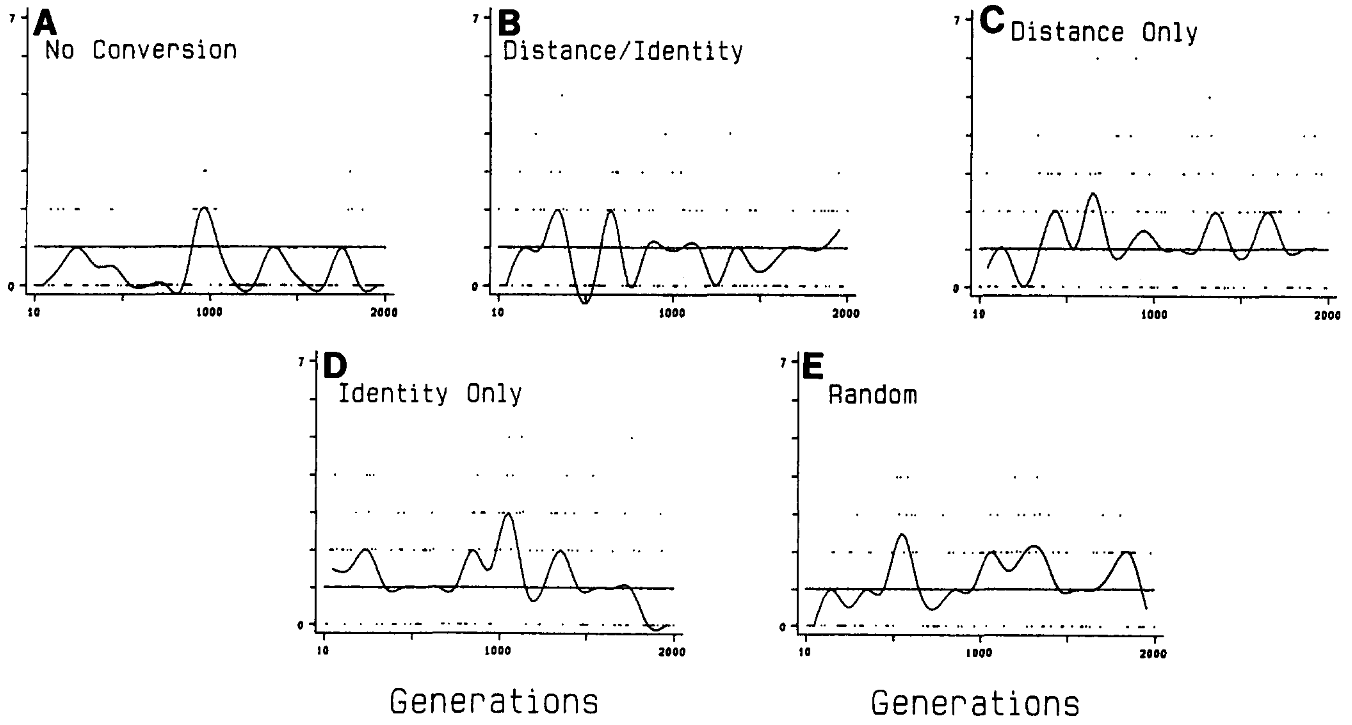

FIGURE 6.—The number of perfect repeats of length 10 when point mutations and repeat conversions occur with a ratio of (A) 4:0, (B) 4:1 with conversions dependent on repeat distance and identity, (C) 4:1 with conversions dependent only on distance between repeats, (D) 4:1 with conversions dependent only on the percent identity between repeats and (E) 4:1 with conversions randomly chosen between repeats.

TABLE 2

Results of Student's *t*-tests comparing the means of ten simulations to the observed number of repeats

	Ratio of point mutation to repeat conversion	Repeats of length 5	Repeats of length 8	Repeats of length 10
No repeat conversion	4:0	0/1/9	0/0/10	0/0/10
Distance + identity	10:1	0/0/10	0/0/10	0/0/10
	4:1	1/4/5	0/0/10	0/9/1
Distance alone	10:1	0/0/10	0/0/10	0/0/10
	4:1	1/2/7	0/2/8	3/6/1
Identity alone	10:1	0/0/10	0/0/10	0/0/10
	4:1	1/2/7	3/5/2	10/0/0
Random repeat conversion	10:1	0/0/10	0/0/10	0/0/10
	4:1	1/1/8	0/5/5	7/3/0

The first, second and third number in each column give the number of simulations that have significantly higher means, the number that are not significantly different and the number that are significantly lower than the observed number of repeats.

TABLE 3

Results of Student's *t*-tests comparing the means of ten simulations to the grand mean of all simulations conducted without repeat conversion

	Ratio of point mutation to repeat conversion	Repeats of length 5	Repeats of length 8	Repeats of length 10
Distance + identity	10:1	2/5/3	10/0/0	7/3/0
	4:1	8/1/1	10/0/0	10/0/0
Distance alone	10:1	2/6/2	10/0/0	9/1/0
	4:1	4/3/3	10/0/0	10/0/0
Identity alone	10:1	1/8/1	10/0/0	10/0/0
	4:1	6/2/2	10/0/0	10/0/0
Random repeat conversion	10:1	1/6/3	10/0/0	10/0/0
	4:1	8/1/1	10/0/0	10/0/0

The first, second and third number in each column give the number of simulations that have significantly higher means, the number that are not significantly different and the number that are significantly lower than the observed number of repeats.

altering 1 base of a 14/17 bp matching repeat decreased the occurrence of deletions by an order of magnitude (ALBERTINI *et al.* 1982).

The linear relationship used for the probability of repeat conversion occurring was chosen for its simplicity, but is supported by the literature. A linear relationship between the number of base matches and recombination has been found by STEPHAN (1989). Gene conversion events have also been shown to decline linearly with sequence similarity (WALSH 1987). A comparison between DNA melting thermodynamics and DNA fidelity also indicates a relationship between identity and the process of mutation (PETRUSKA *et al.* 1988). The equation given in the methods section is not meant to accurately represent what occurs in nature. Rather it is meant to include the major factors that might influence repeat conversions and to permit an examination of how repeats may depend on these factors. The results of Tables 2, 3, and Figure 5 suggest that these simulations are biased toward creating larger repeats at the expense of smaller repeats. The observed data suggest that a larger number of very short repeats are somehow generated and then destroyed by point mutations. The differences found in the simulations using different conditions of distance and identity might be usefully exploited in order to find a relationship that better fits the observed pattern.

This relationship might, in turn, be used to lend insights into the mechanisms that cause sequence directed mutations. Possible mechanisms include a slipped mispairing of repeats (see LEVINSON and GUTMAN 1987 for a review) during replication. For repeat conversion to occur, synthesis proceeds up to the sequence of the first repeat, at which time the sequence of the second repeat becomes aligned with the first. Synthesis continues using the second repeat sequence as a template. After synthesis the sequences are correctly realigned and synthesis continues normally. In this way, the two sequences of the repeat become 100% identical.

Another possible mechanism involves the mismatch repair process. The sequence of one repeat may be "repaired" using the sequence of the second repeat as a template. There might be less constraint on the distance between the sequences of the repeat in this

TABLE 4

Distribution of the number of mutations per repeat conversion event

No. of mutations	Run 1	2	3	4	5	6	7	8	9	10	Total (%)
0	88	81	98	80	87	82	88	96	83	88	871 (4)
1	1,076	1,074	1,067	1,056	1,079	990	966	1,060	1,079	1,041	10,488 (53)
2	694	688	671	753	678	729	652	694	754	695	7,008 (36)
3	113	104	121	101	118	97	85	98	92	114	1,043 (5)
4	27	18	27	14	19	22	31	24	32	20	234 (1)
5	14	3	8	5	7	9	1	4	5	3	59 (0)
6	2	0	1	7	3	1	4	3	1	3	25 (0)
	2,014	1,968	1,993	2,016	1,991	1,930	1,827	1,979	2,046	1,964	19,728 (100)

process, with the possibility of interhelical repeat conversion occurring (LEVINSON and GUTMAN 1987; HAMPSEY *et al.* 1988). Strand switching on the other hand would have a very strong dependence on the distance between repeats, since accidental strand switches during replication might not be expected to span large distances. Further studies of these processes will be required to determine the mechanism or mechanisms involved in small repeat mutations.

The larger than expected number of small direct repeats in the human sequence agrees with previous theoretical calculations of the expected values (KARLIN *et al.* 1983; KARLIN and GHANDOUR 1985). The observed number of repeats of each length generally exceeds the expected number of repeats of that length. In a sequence of 1000 bases, the longest expected 100% identical direct repeat will be of length 9 (KARLIN *et al.* 1983). This agrees with the expected values shown in Table 1/Figure 1 but not with the observed values which extend out to observations including repeats of length 17.

The graphs showing the change in the number of imperfect repeats over time (Figures 2 through 6) give an indication of the amount of variation involved. Further variation in nature may be caused by directional mutation pressures (SUEKA 1988). This level of variation is sufficiently high that caution should be exercised in overinterpreting the presence of a single direct repeat near some gene of interest.

In the presence of only point mutations the number of imperfect repeats falls well below that observed in intergenic DNA. The number of repeats is matched only when the rate of repeat conversion is quite large. A 4:1 ratio of repeat conversion to point mutation may seem high, but this is only one of the nonrandom patterns of spontaneous mutation and many of the repeat conversions do not cause multiple mutations. As shown in Table 4 more than half of the repeat conversions resulted in one or fewer base changes. Evidence for repeat conversions would be compelling when multiple base changes occur but this mechanism would not normally be invoked when single base changes occur. Many of the apparent single point mutations observed in nature may therefore be a result of a repeat conversion event rather than a point mutation event. That 10% of mutations observed in yeast involve multiple base changes (HAMPSEY *et al.* 1988) would suggest that a large number of single base changes may be the result of this type of mutational mechanism. Thus sequence directed mutagenesis may be much more common than previously anticipated.

We wish to thank B. WALSH for his very helpful suggestions and for his encouragement. This work was supported by Natural Sciences and Engineering Research Council grant number U0336 to B. G. Additional support was provided by the Canadian Institute for Advanced Research.

LITERATURE CITED

ALBERTINI, A. M., N. HOFER, M. P. CALOS and J. H. MILLER, 1982 On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. Cell **29:** 319–328.

CAMBARERI, E. B., B. C. JENSEN, E. SCHABTACH and E. U. SELKER, 1989 Repeat-induced G-C to A-T mutations in *Neurospora*. Science **244:** 1571–1575.

DRAKE, J. W., B. W. GLICKMAN and L. S. RIPLEY, 1983 Updating the theory of mutation. Am. Sci. **71:** 621–630.

GILLESPIE, J. H., 1984 The molecular clock may be an episodic clock. Proc. Natl. Acad. Sci. USA **81:** 8009–8013.

GILLESPIE, J. H., 1986 Variability of evolutionary rates of DNA. Genetics **113:** 1077–1091.

GOLDING, G. B., 1987 Multiple substitutions create biased estimates of divergence times and small increases in the variance to mean ratio. Heredity **58:** 331–339.

GOLDING, G. B., and B. W. GLICKMAN, 1985 Sequence-directed mutagenesis: evidence from a phylogenetic history of human alpha-interferon genes. Proc. Natl. Acad. Sci. USA **82:** 8577–8581.

GOLDING, G. B., and B. W. GLICKMAN, 1986 Evidence for local DNA influences on patterns of substitutions in the human alpha-interferon gene family. Can. J. Genet. Cytol. **28:** 483–496.

HARTNETT, C., E. L. NEIDLE, K. L. NGAI and L. N. ORNSTON, 1990 DNA sequences of genes encoding *Acinetobacter calcoaceticus* protocatechuate 3,4-dioxygenase: evidence indicating shuffling of genes and of DNA sequences within genes during their evolutionary divergence. J. Bacteriol. **172:** 956–966.

HAMPSEY, D. M., J. F. ERNST, J. W. STEWART and F. SHERMAN, 1988 Multiple base-pair mutations in yeast. J. Mol. Biol. **201:** 471–486.

KARLIN, S., and G. GHANDOUR, 1985 The use of multiple alphabets in kappa-gene immunoglobulin DNA sequence comparisons. EMBO **4:** 1217–1223.

KARLIN, S., G. GHANDOUR, F. OST, S. TAVARE and L. J. KORN, 1983 New approaches for computer analysis of nucleic acid sequences. Proc. Natl. Acad. Sci. USA **80:** 5660–5664.

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

KUNKEL, T. A., and P. S. ALEXANDER, 1986 The base substitution fidelity of eucaryotic DNA polymerases. J. Biol. Chem. **261:** 160–166.

KUNKEL, T. A., and K. BEBENEK, 1988 Recent studies of the fidelity of DNA synthesis. Biochim. Biophys. Acta **951:** 1–15.

LEVINSON, G., and G. A. GUTMAN, 1987 Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4:** 203–221.

PAPANICOLAOU, C., and L. S. RIPLEY, 1989 Polymerase-specific differences in the DNA intermediates of frameshift mutagenesis. *In vitro* synthesis errors of Escherichia coli DNA polymerase I and its large fragment derivative. J. Mol. Biol. **207:** 335–353.

PETRUSKA, J., M. F. GOODMAN, M. S. BOOSALIS, L. C. SOWERS, C. CHEONG and I. TINOCO, JR., 1988 Comparison between DNA melting thermodynamics and DNA polymerase fidelity. Proc. Natl. Acad. Sci. USA **85:** 6252–6256.

RIPLEY L. S., and B. W. GLICKMAN, 1983 Unique self-complementarity of palindromic sequences provides DNA structural intermediates for mutations. Cold Spring Harbor Symp. Quant. Biol. **47:** 851–861.

SOLOV'EV, V. V., I. B. ROGOZIN and N. A. KOLCHANOV, 1989 Somatic hypermutagenesis in immunoglobulin genes. I. Connection between somatic mutations and repeats. The method of statistical weights. Mol. Biol. **23:** 615–624.

STEPHAN, W., 1989 Tandem-repetitive noncoding DNA: forms and forces. Mol. Biol. Evol. **6:** 198–212.

STREISINGER, G, Y. OKADA, J. EMRICH, J. NEWTON, A. TSUGITA, E. TERZAGHI and M. INOUYE, 1966 Frameshift mutations and the genetic code. Cold Spring Harbor Symp. Quant. Biol. *31:* 77–84.

SUEKA, N., 1988 Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA **85:** 2653–2657

TAKAHATA, N., 1987 On the overdispersed molecular clock. Genetics **116:** 169–179.

WALSH, J. B., 1987 Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? Genetics **117:** 543–557.

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

ZUCKERKANDL, E., and L. PAULING, 1965 Evolutionary distance and convergence in proteins, pp. 97–166 in *Evolving Genes and Proteins*, edited by V. BRYSON and H. J. VOGEL. Academic Press, New York.

Communicating editor: W-H. LI