Volume 15 Number 13 1987 Nucleic Acids Research

High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12

Gene Levinson* and George A.Gutman

Department of Microbiology and Molecular Genetics, University of California, Irvine, CA 92717, USA

Received March 9, 1987; Accepted June 2, 1987

ABSTRACT

Slipped-strand mispairing (SSM) may play an major role in repetitive DNA sequence evolution by generating large numbers of short frameshift mutations within simple tandem repeats. Here we examine the frequency and size spectrum of frameshifts generated within poly-CA/TG sequences inserted into bacteriophage M13 in *Escherichia coli* hosts. The frequency of detectable frameshifts within a 40 bp tract of poly-CA/TG is greater than one percent and increases more than linearly with length, being lower by a factor of four in a 22 bp target sequence. The frequency increases more than 13-fold in *mutL* and *mutS* host cells, suggesting that a high proportion of frameshift events are normally repaired by methyl-directed mismatch repair. Of the 87 sequenced frameshifts in this study, 96% result from deletion or insertion of only one or two 2 bp repeat units. The most frequent events are 2 bp deletions, 2 bp insertions, and 4 bp deletions, the relative frequencies of these events being about 18:6:1.

INTRODUCTION

Slipped-strand mispairing (SSM) has been widely invoked as a mechanism that can readily generate short frameshift mutations. The mechanism involves mispairing of the two strands of duplex DNA within short tandem repeats. When a mispaired duplex serves as a primer for DNA synthesis, duplications or deletions may consequently occur (1-2). Previous evidence for SSM has come primarily from (a) *in vitro* experiments using synthetic polymers (1) and oligomers (3-9), (b) studies of spontaneous mutations in short tandem repeats in bacterial genes (10-18), and (c) inferences derived from eukaryotic sequence data (19-28). The mechanism(s) of SSM and our views of its possible roles in the evolution of repetitive sequences have been reviewed elsewhere (29).

Frameshifts produced by SSM are expected to exhibit the following characteristics: (a) deletions and insertions should occur at high frequency only in repetitive regions; (b) all size changes *via* SSM should involve deletion or duplication of an integral number of discrete repeat units; (c) their size distribution should be biased towards the shortest length compatible with mispairing (a single repeat unit), since mispairing involving minimal relative displacement of the two complementary strands of DNA should be most probable; and (d) the frequency of frameshifts should increase with the length of the repetitive region.

Previous *in vivo* studies of mutations attributable to SSM have been limited to genetic detection of frameshifts in fortuitous tandem repeats of bacterial genes. As a first step to a more controlled examination of *in vivo* frameshift events, we have employed the

well-known sequencing vector M13mp18 borne by *E. coli*. We have developed a simple but powerful procedure by which frameshift mutations in a variety of defined repetitive sequences can be rapidly screened and sequenced, similar to a procedure previously described (31). This approach has the following advantages: (a) the length and sequence composition of the tester sequence can be controlled and easily varied; (b) frequencies of frameshift events can be readily determined; (c) representatives of various frameshift events can be readily sequenced without any additional cloning steps; and (d) a variety of mutations can be introduced into the host (*E. coli*) which may help in the identification of the biochemical mechanisms involved.

In this report, we consider the frameshifts produced within tracts of poly-CA/TG, a simple repeat that is widely distributed in the eukaryotic genome (23; 31-35).

MATERIALS AND METHODS

Bacterial Media and Strains

LB medium contains 1% bacto-tryptone (DIFCO), 0.5 % yeast extract (DIFCO) and 0.5% NaCl. LM medium contains 1% bacto-tryptone (DIFCO), 0.5 % yeast extract (DIFCO), 0.06% NaCl and 0.25 % $MgSO_4 \cdot 7H_2O$ in glass distilled water. M9 plus B1 minimal medium is prepared by autoclaving 0.6 % Na_2HPO_4, 0.3 % KH_2PO_4, 0.05 % NaCl and 0.1 % NH_4Cl in glass distilled water, cooling, adjusting the pH to 7.4, and then adding 0.2 % glucose, 1 % $CaCl_2$ and 0.002 % thiamine, from sterile stocks. SOC medium is prepared by autoclaving 2% bacto-tryptone (DIFCO), 0.5 % yeast extract (DIFCO) and 0.05% NaCl in glass distilled water, cooling and then adding 0.5% $MgSO_4 \cdot 7H_2O$ and 0.2% glucose from sterile stocks. For bacterial plates, 1.5% and 0.7% Bacto Agar (Difco) were added to media for bottom and top agar, respectively.

The reported genotype of *E. coli* strain JM109 is *recA1, endA1, gyrA96, thi, hsdR17, supE44, relA1,* λ^-, Δ *(lac-proAB), [F', traD36, proA*$^+$*B*$^+$*, laci*^{q}*, Z* Δ*M15]* (36). The expected UV sensitive phenotype characteristic of *recA*$^-$ was confirmed in our strain by comparing its UV sensitivity against that of known *recA*$^+$ and *recA*$^-$ strains.

Strains BMH 71-18, 71-18-*mutL* and 71-18-*mutS* were kindly provided by H. Fritz, and share the following genotype: *recA*$^+$*, supE, thi,* Δ*(lac-proAB), [F', pro A*$^+$*B*$^+$*, laci*^{q}*, Z* Δ*M15]* (37); in addition, mutator strains have *Tn10 (tetR)* insertions in the *mutL* or *mutS* gene.

Poly-CA/TG-Bearing M13 Vectors

A synthetic insert, containing a 40 bp (base pair) tract of poly-CA/TG, and flanked by *Bam* H1 linkers, was removed as a *Bam* H1 fragment from another vector, p3'-40, kindly provided by H. Hamada (38), and isolated from an 8% acrylamide gel. The insert was ligated into the *Bam* H1 site of M13mp18 (36) by standard methods. This repetitive sequence, beginning at the *Bam* H1 site, is 5'...GATCCG $(AC)_{20}$ CG...3'.

An 84 bp control sequence containing no repeats longer than 6 bp, and no stop codons in any of the three reading frames, was obtained by adding *Bam* H1 linkers (New England Biolabs #1015) to a filled in *Eco* R1 fragment obtained from a pUC19 subclone of a 55 bp *Sal* I/ *Alu* I fragment, derived from a rat kappa chain gene; this fragment begins 5' to the J1 gene and extends into the coding region [bases 138 to 192; reference (39)]. The complete 84 bp insert at the *Bam* H1 site of M13mp9, confirmed by sequencing, is as

follows:
5'...GATCCCCACAGCCAGACAATGGAGAACTACCACTGTGGTGGACGTTCGGTGGAG
GCACCAAGGGGTACCGAGCTCGAATTCGCG...3'

The resulting circular double-stranded M13 plasmids were introduced into *E. coli* by transformation, using a modified high-efficiency protocol from Hanahan (40); this modification allowed screening and isolation of M13-infected transformants as plaques formed by slow-growing clones of infected cells. For each transformation, 250 ng of ligated construct DNA was added to transformation-competent cells. After the heat shock and addition of SOC medium as per Hanahan (40), the transformed cell suspension was plated out on small (100 x 15 mm) LM agar petri dishes by combining 100 to 500 ul of the cell suspension with 2 ml SOC top agar plus 25 ul of 4% Xgal (5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside) in fresh dimethylformamide, 10 ul IPTG (isopropyl-beta-D-thiogalactoside; a beta-galactosidase inducer) from a 100 mM aqueous stock, and 50 ul of a stationary phase cell suspension from an overnight 1.5 ml LB culture (always grown from a colony taken from a fresh M9 plus B1 minimal medium agar petri dish). Following overnight incubation at 37 C, colonies of transformed cells that either expressed or did not express the *lacZ* gene product were identified as blue or colorless plaques, respectively. Plaques were characterized by nucleotide sequence analysis as described below.

Using this protocol, we have routinely obtained transformation efficiencies of 2 x 10^7 plaque-forming-units per ug of M13 vector DNA. Following transformation, a variety of plaques with blue, light blue or colorless phenotypes were isolated. Subsequent sequence analysis showed that these phenotypes reflected a variety of mutations, including several long deletions, that were apparently generated as a consequence of transfection; previous studies have also reported high spontaneous mutation frequencies associated with transfection (41). Following initial plaque purification, these sequence variants remained relatively stable (except for the short frameshifts discussed below) during additional plaque purifications and platings.

Nucleotide Sequencing

Every mutated phage to be sequenced was isolated from a separate culture of a phage that had been previously plaque purified and sequenced; this ensured that (a) the nature of each size change could be unambiguously determined, and (b) each frameshift was derived from an independent event. Phage were plaque purified by plating out at low densities (approximately 50 plaques per small agar petri dish). Single plaques were then picked with the small end of a sterile pasteur pipette, and grown overnight as standard 1.5 ml LB cultures on a rotating wheel at 37 C to stationary phase (addition of fresh uninfected cells other than those in the small agar plug was unnecessary). Most of the M13-containing supernatant was saved for nucleotide sequencing, but a 1:100 dilution in TMG was also saved as an M13 stock for future infections (these are stable for at least several months at 4 C). TMG contains 10 mM Tris pH 7.5, 10 mM $MgSO_4$, and 0.1 % gelatin.

Template preparation and sequencing by the dideoxy method were carried out as per Sanger et al. (42). Up to 10 independent phage were sequenced per 50 x 100 cm sequencing gel (8% acrylamide).

Determination of Mutation Frequencies and Isolation of Mutants

1.5 ml LB cultures inoculated with single plaques and grown to stationary phase (as

described in the above section) were used to determine the frequency of frameshift mutations. Since our assay is based on a change in the phenotype of progeny derived from a parent plaque, it is possible that heterogeneity of the original parent plaque could lead to an overestimate of subsequent mutation frequencies. Such error was minimized by (a) use of three separate parent plaques for each frequency determination, and (b) microscopic examination of parent plaques for uniform color. Stationary phase M13 supernatants routinely provided titers of 3-5 x 10^{12} plaque-forming-units per ml of supernatant (excluding tiny plaques less than 0.3 mm in diameter, which make up less than 10% of the total and yield large plaques upon replating). A 1:100 dilution of each supernatant was saved as a stock. Two more serial 1:100 dilutions yielded a $1:10^6$ diluted suspension that was plated for frequency analysis. For each large (150 x 15 mm) petri dish, 3 ul of this suspension was preadsorbed (20 minutes at 37 C) by combining with 150 ul of a fresh overnight stationary phase bacterial culture. For plating phage to be counted, the preadsorption mixes were each mixed with 5 ml LB top agar (at 52-55 C) plus 75 ul Xgal and 30 ul IPTG, and then poured onto large LB agar petri dishes and incubated overnight at 37 C. This resulted in average plaque densities of 50 to 80 plaques per cm^2 (excluding tiny plaques).

Plaques were counted by placing the petri dish over a light box and superimposing a 9 x 9 cm translucent film grid marked with columns 1 cm wide. Each column was separately scored for the number of blue and colorless plaques, and the numbers were separately recorded. Plaques to be isolated were picked with sterile pasteur pipets, and grown as standard overnight 1.5 ml LB cultures as described above, then plated at low density for plaque purification. Plating conditions (age of plates and length of incubation time) and counting method were standardized to ensure consistent detection efficiencies.

Statistical comparison of mutation frequencies was carried out using the Fisher Exact Test, with a computer program kindly supplied by Dr. H. Tucker (Department of Mathematics, University of California, Irvine).

RESULTS

Frequency of Frameshifts in 40 bp CA/TG Repeat

The 40 bp poly-CA/TG sequence plus its flanking *Bam* H1 linkers together comprise a 48 bp insert, which is a multiple of three and contains no stop codons; consequently the alpha-complementing region of the *lacZ* gene, into which it is inserted, is

TABLE 1: FREQUENCY OF *LAC*$^+$ TO *LAC*$^-$ MUTATIONS IN 40 BP CA/TG TANDEM REPEAT

CLONE	MUTANTS	NUMBER SCREENED	*LAC*$^-$ FREQUENCY	FREQUENCY PER-BP
#1	28	2240	1.2×10^{-2}	
#2	24	2230	1.1×10^{-2}	
#3	28	2070	1.4×10^{-2}	
OVERALL:	80	6550	1.2×10^{-2}	3.0×10^{-4}

normally expressed, producing blue plaques when the infected bacteria are grown on medium containing Xgal. Frameshifts in the sequence disrupt the reading frame of the gene, resulting in colorless plaques easily distinguished from the blues for purposes of counting and/or isolation. Colorless plaques will result from any frameshift in the repetitive sequence, i.e. any size change that is not a multiple of 3 bp. Colorless plaques can also result from mutations outside of the repetitive sequence, but the rates of these events can be estimated from mutation rates in control phage.

The frequency of *Lac*$^-$ mutations for three phage containing the 40 bp repetitive CA/TG insert is shown in Table 1. An average of 1.2% of the plaques that were screened displayed mutations from *Lac*$^+$ to *Lac*$^-$ (blue to colorless).

To determine the frequency of *Lac*$^-$ mutations that are unrelated to the poly-CA/TG insert in our experimental system, the *Lac*$^-$ frequencies of two types of control phage were determined. Control A (Table 2A) consists of a phage containing an 84 bp insert of known sequence that has no tandem repeats longer than 6 bp, while control B (Table 2B) is a phage with no insert; both controls initially have a *Lac*$^+$ phenotype. The frequencies of *Lac*$^-$ mutations in these phage are 0.008% and 0.02%, respectively; dif-

TABLE 2: *LAC*$^+$ TO *LAC*$^-$ MUTATIONS IN CONTROL PHAGE

Table 2A
LAC$^+$ TO *LAC*$^-$ FREQUENCY IN 84 BP NONREPETITIVE CONTROL

CLONE	MUTANTS	NUMBER SCREENED	*LAC^- FREQUENCY	**CORRECTED FREQUENCY
#1	0	23220	0	
#2	1	20900	0.5×10^{-4}	
#3	1	17860	0.6×10^{-4}	
OVERALL:	2	61990	0.3×10^{-4}	0.8×10^{-4}

Table 2B
LAC$^+$ TO *LAC*$^-$ FREQUENCY WITH NO INSERT

CLONE	MUTANTS	NUMBER SCREENED	*LAC^- FREQUENCY	**CORRECTED FREQUENCY
#1	1	9820	1×10^{-4}	
#2	0	10180	0	
#3	1	9290	1×10^{-4}	
OVERALL:	2	29300	0.7×10^{-4}	2×10^{-4}

*Frequencies are approximate due to low number of *Lac*$^-$ events in controls.
**Plating density was increased to maximize number of observed *Lac*$^-$ events; correction compensates for underestimate of total plaques due to *Lac*$^+$ overlap at high density.

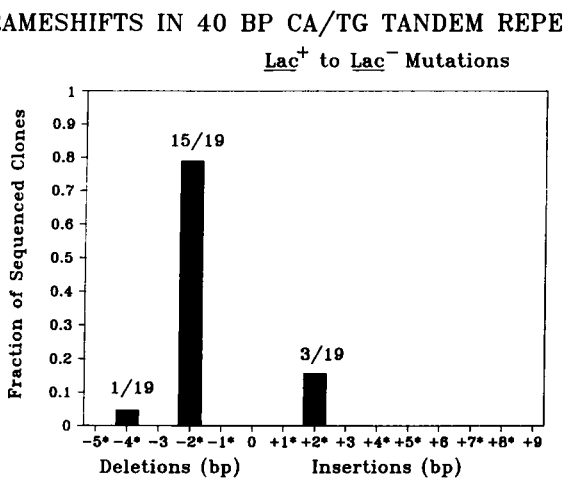

Figure 1. Size distribution of frameshifts in nineteen colorless plaques representing independent *Lac*$^+$ to *Lac*$^-$ mutations within 40 bp poly-CA/TG sequences borne by M13, in *E. coli* host strain JM109. The frequency distribution indicates the number and size of detectable deletions and insertions that occurred. Lengths of frameshifts that could be detected if they occurred are marked by asterisks; note that only even-length frameshifts are actually found. Size changes that are multiples of 3 bp cannot be detected by any of our screening procedure.

ferences between these values are not statistically significant [p = 0.39], and reflect the small number of frameshift events in the controls. These frequencies are about 50-150 times lower than the frequencies obtained with the 40 bp poly-CA/TG insert (Table 1). These results establish two points: first, they indicate that most of the events detected in our experimental system (greater than 98%) involve the inserted sequence; and second, they show that the high frequency of *Lac*$^-$ mutations is not an intrinsic property of all inserts, but is characteristic of the poly-CA/TG tandem repeat.

Size Distribution of Mutations within the 40 bp CA/TG Repeat

To characterize the spectrum of mutations among the *Lac*$^-$ plaques, twenty colorless plaques, each representing an independent mutation event, were sequenced. Nineteen of the twenty involved frameshifts within the repetitive insert; one apparently resulted from a mutation outside of the repetitive insert that was not characterized. The spectrum of the nineteen frameshift mutations is shown in Figure 1. All of them involve one (18/19) or two (1/19) discrete repeat units of CA/TG. Deletions of 2 bp (15/19) outnumber 2 bp insertions (3/19) by a 5:1 ratio in this small sample.

Size Distribution of *Lac-* to *Lac+* Mutations in 44 bp CA/TG Repeat

Since only a small minority of the frameshifts in our first sample were insertions, we increased our ability to detect insertions by using a phage containing a 44 bp poly-CA/TG sequence (one of the mutants derived from the 40 bp insert, described above) and a total insert length of 52 bp. This insert normally generates a colorless plaque phenotype because it is 1 bp out of the correct reading frame. Expression of the *lacZ* gene, detect-

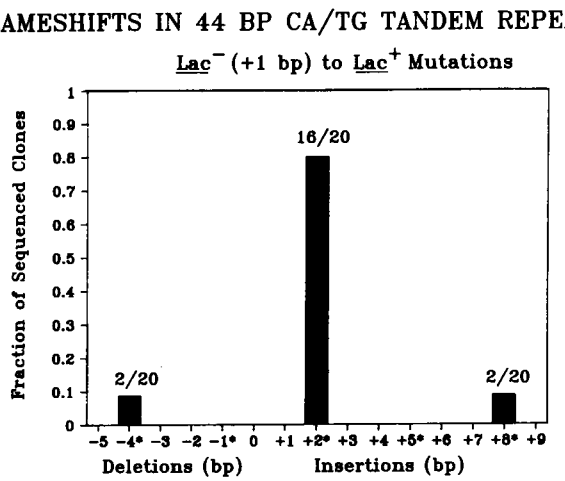

Figure 2. Size distribution of frameshifts in twenty blue plaques representing independent *Lac*⁻ (+1 bp) to *Lac*⁺ mutations within 44 bp poly-CA/TG. Other notations as in Fig 1.

able as blue plaque color, will occur only with subsequent frameshifts that restore the correct reading frame, limiting detection to insertions of 2,5,8,11,...(3n + 2) bp and deletions of 1,4,7,10,...(3n - 1) bp; 2 bp deletions, the most common events in the 40 bp insert, are therefore not detected with this selection scheme.

We sampled twenty independent *Lac*⁺ mutation events, and the spectrum of detectable mutations in these phage (blue plaques) is shown in Figure 2. As was the case for 40 bp sequence, the large majority of size changes in the 44 bp sequence (16/20) involve a single repeat unit; the remaining 4 events involve two or four repeat units, i.e. four or eight bp.

Relative Frequencies of Deletions and Insertions in Large Samples

Because the 44 bp sequence does not permit detection of 2 bp deletions, comparison of the frequencies of frameshift mutants in populations of phage containing the 40 bp and 44 bp poly-CA/TG inserts (Tables 1 and 3) allowed us to estimate the relative frequencies of deletions and insertions of 2 bp in large-sample frequency studies. Overall frequencies of *Lac*⁻ to *Lac*⁺ mutations were determined in three large samples of populations of phage bearing 44 bp poly-CA/TG inserts (Table 3), and indicate that an average of 0.4% of the plaques screened underwent *Lac*⁻ to *Lac*⁺ mutation events.

Our sequence data indicate that the majority of detectable frameshifts involve 2 bp deletions in phage populations bearing the 40 bp insert (15/19), and 2 bp insertions in those bearing the 44 bp insert (16/20; see Fig. 1 and 2). Since the results in Table 2 show that the contribution of mutations *not* involving the repetitive sequences is negligible, we can simplify the problem by assuming that the large-sample mutation frequencies in Tables 1 and 3 reflect primarily the frequencies of 2 bp deletions and 2 bp insertions, respectively. Under this assumption, the ratio of 2 bp deletions to insertions is 3:1 (1.2:0.4).

TABLE 3: FREQUENCY OF *LAC*$^-$ (+1 bp) TO *LAC*$^+$ MUTATIONS IN 44 BP CA/TG TANDEM REPEAT

CLONE	MUTANTS	NUMBER SCREENED	*LAC*$^+$ FREQUENCY	FREQUENCY PER-BP
#1	10	2650	0.4×10^{-2}	
#2	16	2750	0.6×10^{-2}	
#3	9	2360	0.4×10^{-2}	
OVERALL:	35	7760	0.5×10^{-2}	1.0×10^{-4}

This large-sample ratio is not very different from the 5:1 ratio estimated from the small-sample data in Figure 1.

Effect of Repetitive-Tract Length on Frameshift Frequency

To assess effects of repetitive-tract length on frameshift frequency, we determined mutation frequencies in a shorter (22 bp) poly-CA/TG tract by the same counting method described for 40 bp tracts (this clone was one of several variants isolated from the original transformation). The spectrum of common frameshift mutations in the 22 bp repeat was also examined by sequencing twenty-four independent mutant phage, and is shown in Figure 3. All of the frameshifts sampled involved deletions or insertions within the 22 bp repetitive sequence. Of these, 21 were 2 bp deletions and 3 were 2 bp insertions. Thus, the spectrum of frameshifts and the excess of deletions over insertions in the 22 bp CA/TG insert are similar to those observed in the 40 bp insert. (Although the ratio of 21:3 is higher than that of 15:3, this difference is *not* statistically significant [p = 0.79]).

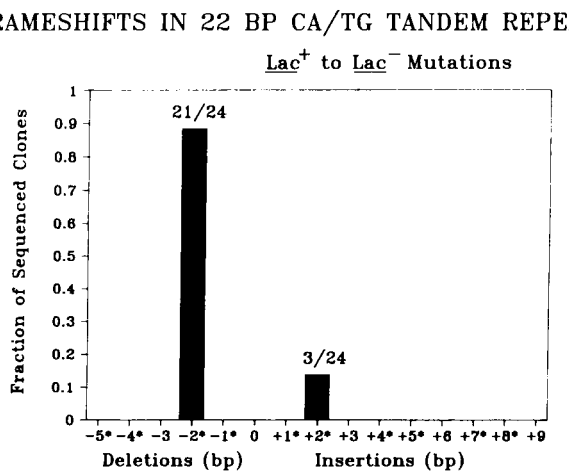

Figure 3. Size distribution of frameshifts in twenty-four colorless plaques representing independent *Lac*$^+$ to *Lac*$^-$ mutations within 22 bp poly-CA/TG. Other notations as in Fig 1.

TABLE 4: FREQUENCY OF *LAC*$^+$ TO *LAC*$^-$ MUTATIONS IN 22 BP CA/TG TANDEM REPEAT

CLONE	MUTANTS	NUMBER SCREENED	*LAC*$^-$ FREQUENCY	FREQUENCY PER-BP
#1	10	3060	0.3×10^{-2}	
#2	13	3510	0.4×10^{-2}	
#3	2	2400	0.1×10^{-2}	
OVERALL:	25	8970	0.3×10^{-2}	1.3×10^{-4}

Frameshift frequency data for large populations of phage bearing the 22 bp insert are shown in Table 4. Since we wished to determine the effect of overall tract length on frequencies of mutations per-bp of repetitive sequence, comparisons were obtained by dividing frequency values by repetitive tract lengths, which has been done in the right-most column. The 22 bp poly CA/TG tract displays a per-bp mutation frequency of 1.3×10^{-4}, and our sequence data show that most or all of these are due to frameshifts in the repetitive tract itself. The data therefore indicate that shortening the repetitive insert from 40 bp (data from Table 1) to 22 bp reduces the per-bp mutation frequency by a factor of about 2.4, a difference that is statistically significant [$p < 0.05$]. Possible reasons why the frequency per unit length could depend on the overall repetitive tract length are considered in the Discussion.

Role of Methyl-Directed Mismatch Repair

Mutations within DNA carried by *E. coli* hosts are subject to repair by a variety of mechanisms, and it is clearly of interest to determine what proportion of frameshift events are normally corrected by such mechanisms. To begin studying this question, we examined mutation frequencies in *mutL* and *mutS* strains, which are deficient for methyl-directed mismatch repair, a major pathway for correction of replication errors (43-46). A recent study (47) has shown that methyl-directed mismatch repair in *E. coli* can efficiently repair single unpaired bases due to frameshift mutations in newly synthesized DNA strands. To examine whether frameshifts of two or more bases in CA/TG repeats are subject to such repair, we compared large-sample frameshift frequencies in *mutL* and *mutS* hosts to those of isogenic *mut*$^+$ controls. To eliminate detection of point mutations, only frameshifts causing reversion to the *Lac*$^+$ phenotype were analyzed. Separate analysis of mutations that restored the *Lac*$^+$ phenotype in 42 bp and 38 bp CA/TG tracts allowed separate determination of frequencies of frameshifts of 3n -2 (or 3n +1) and 3n + 2 (or 3n -1) nucleotides, respectively. These data are shown in Tables 5A and 5B, respectively. The total detectable frameshift frequencies (obtained by adding the corresponding values from Tables 5A and 5B) show increases in frameshift frequency of over one order of magnitude (13-fold and 14-fold for *mutL* and *mutS*, respectively), when comparing the mutator strains to *mut*$^+$. These data suggest that over 90% of detectable frameshifts are normally corrected by methyl-directed mismatch repair.

The spectra of common frameshift mutations in the *mutL* host are shown in Figures 4 and 5. Sequences of 24 independent mutations in *mutL* hosts show that the size distribution of frequent frameshift mutations in *mutL* hosts is generally similar to that observed in 40-44 bp inserts borne by the standard JM109 host, which is *mut*$^+$ (Figures 1 and

TABLE 5: EFFECT OF *mutL* and *mutS* ON DETECTABLE FRAMESHIFT FREQUENCY

Table 5A
Lac$^-$ (+2 bp) to *Lac*$^+$ Mutations in 42 bp CA/TG

STRAIN	CLONE	MUTANTS	NUMBER SCREENED	*LAC*$^+$ FREQUENCY
71-18 *mutL*	#1	692	6720	10.3×10^{-2}
	#2	1467	11600	12.6×10^{-2}
	#3	889	6000	14.8×10^{-2}
	OVERALL	3048	24320	12.5×10^{-2}
71-18 *mutS*	#1	831	6960	11.9×10^{-2}
	#2	947	8080	11.7×10^{-2}
	#3	835	5440	15.3×10^{-2}
	OVERALL	2613	20480	12.8×10^{-2}
71-18 *mut*$^+$	#1	72	10480	0.7×10^{-2}
	#2	63	7840	0.8×10^{-2}
	#3	63	6400	1.0×10^{-2}
	OVERALL	198	24720	0.8×10^{-2}

Table 5B
Lac$^-$ (+1 bp) to *Lac*$^+$ Mutations in 38 bp CA/TG

STRAIN	CLONE	MUTANTS	NUMBER SCREENED	*LAC*$^+$ FREQUENCY
71-18 *mutL*	#1	204	9520	2.1×10^{-2}
	#2	157	7520	2.1×10^{-2}
	#3	175	7680	2.3×10^{-2}
	OVERALL	536	24720	2.2×10^{-2}
71-18 *mutS*	#1	277	8320	3.3×10^{-2}
	#2	109	3760	2.9×10^{-2}
	#3	186	8320	2.2×10^{-2}
	OVERALL	572	20400	2.8×10^{-2}
71-18 *mut*$^+$	#1	15	3440	0.4×10^{-2}
	#2	22	7520	0.3×10^{-2}
	#3	13	7440	0.2×10^{-2}
	OVERALL	50	18400	0.3×10^{-2}

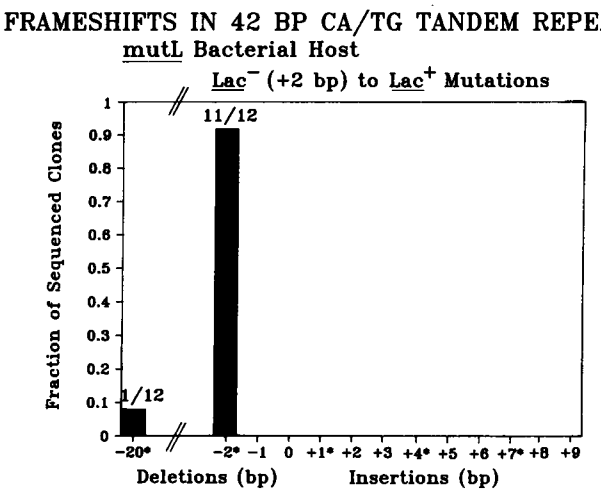

Figure 4. Size distribution of frameshifts in twelve blue plaques representing independent *Lac*⁻ (+2 bp) to *Lac*⁺ mutations within 42 bp poly-CA/TG, borne by *mutL* bacterial hosts. A single large deletion of 20 bp is indicated. Other notations as in Fig 1.

2). Assuming that the fractions of 2 bp deletions and insertions in the 24 sequenced clones are similar in the larger samples used for frequency studies (see above), the data show that the absence of methyl-directed mismatch repair does not appreciably alter the

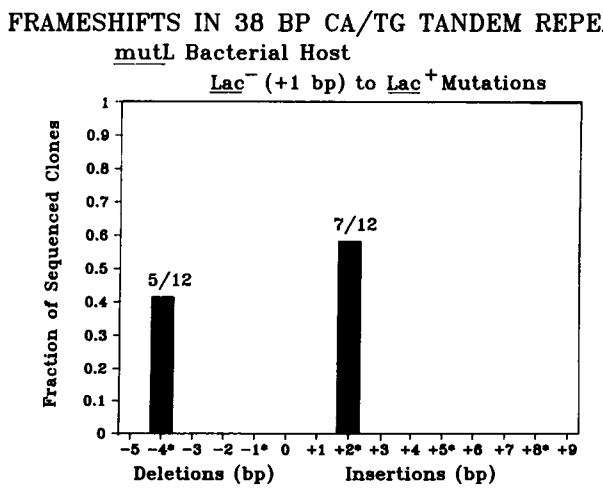

Figure 5. Size distribution of frameshifts in twelve blue plaques representing independent *Lac*⁻ (+1 bp) to *Lac*⁺ mutations within 38 bp poly-CA/TG, borne by *mutL* bacterial hosts. Other notation as in Fig 1.

relative proportions of deletions versus insertions, since 2 bp deletions still greatly outnumber 2 bp insertions as well as longer deletions. [The 5:7 ratio of 4 bp deletions to 2 bp insertions in the *mutL* strain (Figure 5) is appreciably higher than the 2:16 ratio in the wild-type (Figure 2), although this difference falls short of statistical significance at the .05 level [p = 0.07]. However, if this difference does turn out to be meaningful, it would suggest that short deletions are repaired more efficiently than short insertions in this system by methyl-directed repair].

DISCUSSION

Frameshifts in Long CA/TG Repeats are Consistent with SSM

In this study, we have shown that frameshift mutations in our experimental system exhibit four characteristics that would be expected if they are generated by a slipped-strand mispairing mechanism.

First, deletions and insertions occur at high frequency only in repetitive regions. We have observed very high frequencies of frameshifts within long poly-CA/TG sequences borne by an M13 vector in *E. coli* hosts, in excess of 1%, while control phage bearing a nonrepetitive sequence failed to exhibit this high frameshift frequency. The high frequencies we observed in the long CA/TG tracts are an extreme example of a feature common to a variety of simple repeats found in bacterial chromosomes, since a number of previous investigations (10-18) of much shorter runs (four to twelve bases) of other short repeat units (one or four bases) also found frameshift frequencies far in excess of frequencies of other types of mutation such as base substitutions (48). It therefore appears likely that our observed frameshift frequency is a characteristic of replication of this CA/TG repeat in *E. coli*, and is probably not peculiar to M13 replication.

Second, all of the 87 sequenced frameshifts in the CA/TG repeats that we have examined involve deletion or duplication of an integral number of repeat units (i.e. 2 bp). Although our selection scheme is potentially capable of detecting, for example, 1 bp and 5 bp size changes, as well as insertions other than CA/TG, no such events were found among any of the sequenced clones.

Third, the size distribution of frameshifts is strongly biased towards a single repeat unit, the shortest length compatible with mispairing. Of the 87 sequenced frameshifts in this study, 76 (87%) resulted from deletion or insertion of a single repeat unit; those involving deletion of only two repeat units accounted for an additional 9%, and the remainder included less than 4% of the total.

Fourth, the frameshift frequency clearly depends on the length of the repetitive region. The greater than 2-fold decrease in the frameshift frequency per unit length in the 22 bp CA/TG tract compared with that observed in the 40 bp tract suggests that longer tracts are more permissive for frameshift events than shorter ones. Previous studies of frameshifts in short poly-A runs in bacterial genes showed an even more dramatic length dependence for shorter tracts: decreasing the length of $[A]_5$ or $[A]_6$ by a single bp decreased the frameshift frequency by more than 10-fold (14).

This length effect may be at least partially understood by considering the structure of the mispaired intermediate assumed to be required for SSM (see reference [29]), which consists of a mispaired (although not mismatched) central element limited by unpaired regions on either side. If, for example, mispairing of ten repeat units is critical to the

stability of the structure, one can align this overall region of length 20 (ten 2 bp repeat units) in only two different base-paired registers along a total repeat length of 22 ([22-20/2 + 1), but in eleven different registers along a repeat length of 40 ([40-20]/2 + 1). If SSM is limited mainly by the stability of this structure, then the overall frequency of SSM in the 40 bp region should be 5.5 times greater than that in the 22 bp one (11/2). Dividing this value by 1.8 (i.e. 40/22) yields a frequency *per unit length* 3.0 times higher for the 40 bp length than for 22 bp. Since this is only slightly higher than the actual value we obtained (namely 2.4), this simple analysis suggests that availability of up to ten two-base repeat units for mispairing may contribute substantially to the stability of the structure, but additional pairing does not. (If the optimum length for a stable SSM intermediate is large compared to the length of the repetitive sequence in which it is to occur, then length-dependency for the frequency of frameshifts should be even more extreme; this could explain the greater than 10-fold differences seen in the work of Streisinger and Owen (14) referred to above, since they were studying much shorter stretches of single-base repeats.)

Although this analysis is highly speculative, more detailed studies with the system we have used could potentially shed considerable light on this question. Other explanations could also account for the greater-than-linear increase of frameshift frequency with length. One possibility is that in longer tracts, a larger number of the transitory single-stranded domains that characterize such repetitive regions (49-53) might be able to "fuse" into longer denatured regions that could be more prone to subsequent mispairing.

In summary, our results are consistent with the four characteristics expected of a slipped-strand mispairing mechanism, and we consider it likely that the frameshifts observed in this model system are consequences thereof. The high frameshift frequencies observed in this model system suggest that, at least in repetitive regions, SSM may play a major role in DNA sequence evolution (23, 29).

Methyl-Directed Repair Corrects Most Frameshifts

In our system, absence of methyl-directed mismatch repair (in *mutL* and *mutS* mutator strains) increases the frameshift frequency by over 13-fold, indicating that this repair pathway effectively corrects over 90% of occurring frameshifts. Our frequency and sequence data indicate that deletions and insertions in the CA/TG tracts are both subject to a high efficiency of repair by this pathway in our experimental system (although there is a suggestion that deletions may be corrected more efficiently than insertions). Since strand discrimination by the methyl-directed mismatch repair pathway is based on transient undermethylation of the newly synthesized strand during DNA replication (37; 43-47), the dramatic increase in frameshift frequency in these mutator strains is consistent with the notion that most CA/TG deletions and insertions occurred during DNA replication, in accord with earlier suggestions (5; 15). However, it is possible that mutations could also occur at other times during the cell cycle -- i.e. as a consequence of DNA synthesis during repair -- and, as has been suggested (20; 29), that repair mechanisms might play a role in the generation of certain frameshift events as well as in their correction.

Negligible Role of Recombination

There are several reasons for believing that M13 recombination events (54-56), including unequal crossovers, do not contribute significantly to our results. One is that the JM109 host carries the *recA*$^-$ mutation which is known to dramatically reduce the frequency of homologous recombination (57-60), although it does not completely abolish all

recombination events (61). In addition, since the 71-18 *mut*$^+$ strain is *recA*$^+$, the fact that combined frameshift frequencies in the 38 and 42 bp CA/TG sequences were almost identical to those obtained in JM109 hosts with the 40 bp repeat (1.1% *vs.* 1.2%) indicates that *recA*-mediated recombination does not contribute substantially to frameshift frequencies in this system.

Another reason is that the sample of frameshifts sequenced (Figures 1-5) contained only very short deletions and insertions (with the exception of a single large deletion), a pattern that is at odds with the far broader spectrum of deletion or insertion lengths expected from double-crossover events (62).

Hence, we consider it unlikely that recombination has contributed significantly to our data. Although there are reports that simple repetitive tracts may be "hotspots" for unequal crossing over (63), gene conversion (25) and illegitimate recombination (21) in eukaryotes, there is no reason to think that these mechanisms should apply to this model prokaryotic system.

Relative Frequencies of Deletions and Insertions

From frequency and sequence data, we have estimated that about three quarters of all 2 bp frameshifts are deletions while one quarter are 2 bp insertions, in the 40-44 bp poly CA/TG tracts carried by the prokaryotic tester strain. Equal or greater proportions of deletions were observed in phage that either bore the shorter insert or that were carried by mutator hosts. A similar bias towards deletions has been reported in frameshifts occurring in shorter tandem repeats, and has been predicted on thermodynamic grounds (14). However, a bias towards frameshifting insertions rather than deletions has been reported in the case of a tandem repeat of CTGG (10), while induction of frameshifts at other simple repeats has revealed some sites that favor insertions and others that favor deletions (14; 64). The ratio of deletions to insertions may therefore be a complex sequence-dependent phenomenon.

A general bias towards deletions is not unexpected in prokaryotic cells, since rapid cell division would be hindered by the presence of excess DNA in the genome; it may therefore be to prokaryotes' selective advantage to produce or tolerate insertions less often than deletions, and the genetic apparatus of these organisms may have evolved accordingly because of selective pressure against an increase in genome size. Clearly, additional studies are needed, since most prokaryotic studies have been limited to a small number of loci in lab strains of *Escherichia* and *Salmonella.*

In eukaryotes, the relative frequencies and size spectra of deletion and insertion events is an open question of considerable importance with respect to DNA sequence evolution. In any case, it is clear that the rate of accumulation of tandem repeats in a given organism will depend not only on the proportion of deletions and insertions that are produced, but also on other factors including repair and natural selection (23; 29). It may depend on stochastic factors as well, since simple repeats with no apparent function have appeared in a broad range of distantly related taxa; this has previously resulted in the erroneous assumption of homology based on hybridization experiments (see reference [65]).

The role played by SSM in DNA sequence evolution is a complex question. In this context, the significance of the present study is two-fold: first, it describes a method by which defined sequences can be analyzed in a systematic fashion; and second, it demonstrates that frameshift mutations in repetitive sequences may occur at extremely high frequencies *in vivo*, frequencies that are very sensitive to repeat tract length. SSM may there-

fore constitute an important source of genetic variation, whose ultimate evolutionary fate will depend upon a variety of forces that remain to be described.

ACKNOWLEDGEMENTS

We thank H. Hamada for providing the plasmid clone (p3'-40) containing a 40 bp poly-CA/TG sequence with *Bam* H1 linkers, T. Winkler for assistance in cloning of the non-repetitive control phage, R. Scherrer for providing us with a method of confirming the *recA*⁻ genotype of our strains, H. Tucker for assistance with statistical analysis, R. Besta for technical support, and K. Bertrand, R. Campbell, R. Kolodner and J. Manning for helpful suggestions and discussions during the course of these studies. This work was supported by U.S. Public Health Service grants AI-14774 and AI-21366, and an award from the Chancellor's Patent Fund (Univ. Calif. Irvine). G.L. was supported by N.I.H. Research Service Award HD07029 and Earle C. Anthony and Monsanto Company Fellowships.

*Present address: Division of Tumor Immunology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115, USA

REFERENCES

1. Fresco, J.R. and Alberts, B.M. (1960) Proc. Natl. Acad. Sci. USA 46, 311-321
2. Kornberg, A. (1980) DNA Replication, pp. 340-343, W. H. Freeman, San Francisco.
3. Kornberg, A., Bertsch, L.L., Jackson, J.F. and Khorana, H.G. (1964) Proc. Natl. Acad. Sci. USA 51, 315-323.
4. Radding, C.M., Josse, J. and Kornberg, A. (1962) J. Biol. Chem. 237, 2869-2876.
5. Schachman, H.K., Adler, J., Radding, C.M. Lehman, I.R. and Kornberg, A. (1960) J. Biol. Chem. 235, 3242-3249.
6. Swartz, M.N., Trautner, T.A. and Kornberg, A. (1962) J. Biol. Chem. 237, 1961-1967.
7. Wells, R.D. and Blair, J.E. (1967) J. Mol. Biol. 27, 273-288.
8. Wells, R.D., Buchi, H., Kossel, H., Ohtsuka, E. and Khorana, H.G. (1967a) J. Mol. Biol. 27, 265-272.
9. Wells, R.D., Jacob, T.M., Narang, S.A. and Khorana, H.G. (1967b) J. Mol. Biol. 27, 237-263.
10. Farabaugh, P., Schmeissner, U., Hofer, M. and Miller, J.H.. (1978) J. Mol. Biol. 126, 847-863.
11. Levin, D.E., Yamasaki, E. and Ames, B.N. (1982) Mutat. Res. 94, 315-330.
12. Owen, J.E., Schultz, D.W., Taylor, A. and Smith, G.R. (1983) J. Mol. Biol. 165, 229-248.
13. Pribnow, D., Sigurdson, D.C., Gold, L., Singer, B.S., Napoli, C., Brosius, J., Dull, T.J., and Noller, H.F. (1981) J. Mol. Biol. 149, 337-376.
14. Streisinger, G. and Owen, J. 1985. Genetics 109, 633-659.
15. Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzhaghi, E. and Inouye, M. (1966) Cold Spring Harbor Symp. Quant. Biol. 31, 77-84.
16. Roth, J.R. (1974) Ann. Rev. Genet. 8, 319-346.
17. Clark, C.H. and Johnson, A.W.B. (1976) Mutation Research 36, 147-164.
18. Yourno, J, Ino, I. and Kohno, T. (1971) J. Mol. Biol. 62, 233-240.
19. Efstradiatis, A., Posakony, J. W., Maniatis, T., Lawn, R.M., O'Connel, C., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980) Cell 21, 653-668.
20. Flanagan, J.G., Lefranc, M.-P., and Rabbitts, T.H. (1984) Cell 36, 681-688.
21. Hasson, J.- F., Mougneau, E., Cuzin, F. and Yaniv, M. (1984) J. Mol. Biol. 177, 53-68.
22. Jones, C.W. and Kafatos, F.C. (1982) J. Mol. Evol. 19, 87-103.
23. Moore, G.P. (1983) TIBS 8, 411-414.
24. Rodakis, G.C., Lecanidou, R. and Eickbush, T.H. (1984) J. Mol. Evol. 20, 265-273.

25. Slightom, J.L., Blechl, A.E. and Smithies, O. (1980) Cell 21, 627-638.
26. Tautz, D. and Renz, M. (1984a) J. Mol. Biol. 172, 229-235.
27. Tautz, D. and Renz, M. (1984b) Nucleic Acids Res. 12, 4127-4138.
28. Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E. and White, R. (1987) Science 235, 1616-1622.
29. Levinson, G. and Gutman, G.A. (1987) Mol. Biol. Evol. 4, 201-219.
30. Lorenzetti, R, Cesareni, G. and Cortese, R. (1983) Mol. Gen. Genet. 192, 515-516.
31. Hamada, H. and Kakunaga, T. (1982a) Nature 298, 396-398.
32. Hamada, H., Petrino, M.G. and Kakunaga, T. (1982b) Proc. Natl. Acad. Sci. USA 79, 6465-6469.
33. Jeang, K.-T. and Hayward, G.S. (1983) Mol. Cell. Biol. 3, 1389-1402.
34. Rogers, J. (1983) Nature 305, 101-102.
35. Schmid, C.W. and Shen, C-K. J. (1985) in Molecular Evolutionary Genetics, MacIntyre, R.J., ed., Plenum Press, New York, pp. 345-347.
36. Yanisch-Perrin, C., Viera, J. and Messing, J. (1985) Gene 33, 103-119.
37. Kramer, B., Kramer, W., and Fritz, H.-J. (1984) Cell 38, 879-887.
38. Hamada, H., Seidman, M., Howard, B.H. and Gorman, C.M.. (1984b) Mol. Cell. Biol. 4, 2622-2630.
39. Sheppard, H.W. and Gutman, G.A. (1982) Cell 29, 121-127.
40. Hanahan, D. (1983) J. Mol. Biol. 166, 557-580.
41. Lebkowski, J.S., Clancy, S., Miller, J.H. and Calos, M.P. (1985) Proc. Natl. Acad. Sci. 82, 8606-8610.
42. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.H.J. and Roe, B.A. (1980) J. Mol. Biol. 143, 161-178.
43. Glickman, B.W. (1981) in Molecular and cellular mechanisms of mutagenesis, Lemontt, J.F. and Generoso, W.M., eds., Plenum Press, N.Y., pp 65-87.
44. Grossman, L. (1981) Arch. Biochem. Biophys. 211, 511-522.
45. Kramer, W., Schughart, K. and Fritz, H-J. (1982) Nucleic Acids Res. 10, 6475-6485.
46. Lu, A-L., Clark, S. and Modrich, P. (1983) Proc. Natl. Acad. Sci. 80, 4639-4643.
47. Dohet, C., Wagner,R. and Radman, M. (1986) Proc. Natl. Acad. Sci. 83, 3395-3397.
48. Cox, E.C. (1976) Ann. Rev. Genet. 10, 135-156.
49. Hamada, H., Petrino, M.G., Kakunaga, T., Seidman, M. and Stollar, B.D. (1984a) Mol. Cell. Biol. 4, 2610-2621.
50. Hentschel, C.C. (1982) Nature 295, 714-716.
51. Mace, H.A.F., Pelham, H.R.B. and Travers, A.A. (1983) Nature 304, 555-557.
52 Nickol, J.M. and Felsenfeld, G. (1983) Cell 35, 467-477.
53. Weintraub, H. (1983) Cell 32, 1191-1203.
54. Dagert, M. and Ehrlich, S.D. (1983) EMBO J. 2, 2117-2122.
55. Dagert, M. and Ehrlich, S.D. (1983) EMBO J. 3, 87-89.
56. Michel, B. and Ehrlich, S.D. (1986) Proc. Natl. Acad. Sci. 83, 3386-3390.
57. Anderson, R.P. and Roth, J. (1977) Ann. Rev. Microbiol. 31, 473-505.
58. Clark, A.J. (1973) Ann. Rev. Gen. 7, 67-86.
59. Radding, C.M. (1982) Ann. Rev. Gen. 16, 405-437.
60. Walker, G.C. (1985) Ann. Rev. Biochem. 54, 425-457.
61. Matfield, M,R. Badawi and Brammar, W.J. (1985) Mol. Gen. Genet. 199, 518-523.
62. Smith, G.P. (1976) Science 191, 528-535.
63. Jeffreys, A. J., Wilson, V. and Thein, S. L. (1985) Nature 316, 76-79.
64. Calos, M. and Miller, J. H. (1981) J. Mol. Biol. 153, 39-66.
65. Levinson, G., Marsh, J.L., Epplen, J.T. and Gutman, G.A. (1985) Mol. Biol. Evol. 2, 494-504.