DNA mismatch correction by Very Short Patch repair may have altered the abundance of oligonucleotides in the *E.coli* genome

Ashok S.Bhagwat and Michael McClelland¹

Department of Chemistry, Wayne State University, Detroit, MI 48202 and ¹California Institute of Biological Research, La Jolla, CA 92037, USA

Received December 16, 1991; Revised and Accepted March 2, 1992

ABSTRACT

A base mismatch correction process in *E. coli* K-12 called Very Short Patch (VSP) repair corrects T:G mismatches to C:G when found in certain sequence contexts. Two of the substrate mismatches (5'-CTWGG/3'-GGW'CC; W = A or T) occur in the context of cytosine methylation in DNA and reduce the mutagenic effects of 5-methylcytosine deamination to thymine. However, VSP repair is also known to repair T:G mismatches that are not expected to arise from 5-methylcytosine deamination (example—CTAG/GGT-C). In these cases, if the original base pair were a T:A, VSP repair would cause a T to C transition. We have carried out Markov chain analysis of an *E. coli* sequence database to determine if repair at the latter class of sites has altered the abundance of the relevant tetranucleotides. The results are consistent with the prediction that VSP repair would tend to deplete the genome of the 'T' containing sequences (example—CTAG), while enriching it for the corresponding 'C' containing sequences (CCAG). Further, they provide an explanation for the known scarcity of CTAG containing restriction enzyme sites among the genomes of enteric bacteria and identify VSP repair as a force in shaping the sequence composition of bacterial genomes.

INTRODUCTION

A number of forces shape the nucleotide composition of bacterial genomes. For example, codon preferences and the abundance of certain amino acids in proteins are thought to cause the underrepresentation or overrepresentation of certain trinucleotides (1). Some of the processes influencing the genome composition work against each other and it is not always possible to predict the overall result. Frequent genetic exchanges between related bacteria followed by genetic recombination should lead to the homogenization of bacterial genomes, while the restriction of incoming DNA must lead to heterogeneity.

The K-12 strain of *E. coli* contains a single DNA cytosine methylase, Dcm, that transfers a methyl group to the second cytosine in both strands of 5'-CCWGG (W = A or T) sequences (2). Some of these sites have been shown to be 'hot-spots' for

C to T mutations (3, 4). This is believed to be due to the deamination of 5-methylcytosine (5-meC) to thymine. If the T:G mismatches resulting from the deamination are not correctly repaired, 50% of the progeny molecules would contain a C to T mutation. Such a process would tend to decrease the number of CCWGG sequences in the genome.

E. coli contains a base mismatch correction process called Very Short Patch (VSP) repair (5, 6) that tends to counteract the mutagenic effects of 5-meC deamination. It corrects 5'-CTAGG/3'-GGTCC and CTTGG/GGACC mismatches to CCWGG (6–11). Consistent with this antimutagenic role for VSP repair, 5-meC to thymine mutations at a site in the *E. coli* chromosome were much higher in a strain defective in VSP repair (4). Therefore, VSP repair would tend to reduce the depletion of CCWGG sites from the chromosome.

Interestingly, VSP repair also acts on T:G mismatches that are unlikely to have resulted from 5-meC deamination. It repairs T:G mismatches to C:G in CTAGA/GGTCT, TTAGG/AGTCC and GTAGG/CGTCC (7). In addition, *in vitro* studies with Vsr, a protein required for VSP repair (12), suggest that CTAGC/GGTCC may also be repaired to CCAGC/GGTCC by this process (13). Therefore, the sequence-specificity of VSP repair may be such that T:G mismatches in NTWGG/N'GW'CC and CTWGN/GGW'CN' (N = A, C, G or T and primes indicate complements of bases) are repaired to C:G by the process. Although the sequences resulting from repair at these sites are subsequences of the site of methylation, CCWGG, there is no evidence for methylation at these sites. Further, VSP repair can act on the mismatches described above in the absence of cytosine methylation in the 'G' containing strand (13). Therefore, some of the T:G mismatches arising as a result of replication errors must be subject to VSP repair.

The replacement of T:G with C:G by VSP repair in sequence contexts other than the sites of cytosine methylation is potentially mutagenic. If it were the only force changing the genome, it should deplete the genome of sequences of the type NTWGG, CTWGN and their complements, while enriching the genome with sequences of the type NCWGG, CCWGN and their complements. To test this hypothesis, we have examined the database of *E. coli* sequences for the frequencies of these and other related sequences by Markov chain analysis. The results are consistent with the predictions and provide insight into the evolution of the *E. coli* genome.

MATERIALS AND METHODS

EcoSeq5 is a compilation of *E. coli* sequences that has been processed to eliminate sequence overlaps, duplications and vector sequences (14). The 1,774,683 base combined sequence and the observed frequencies of mono-, di-, tri-, tetra- and pentanucleotides in this sequence were generously provided by K. Rudd (NIH). The observed frequencies of oligonucleotides in the *Saccharomyces cerevisiae* genome were taken from the work of Arnold et al. (15). Various predictors of oligonucleotide frequencies have been developed (16). We have used a simple method, Markov chain analysis (17, 18), to calculate the expected frequencies of various trinucleotides, tetranucleotides and pentanucleotides that are expected to be affected by VSP repair or by 5-meC deamination in *E. coli*. A comparison of the frequencies arrived at through Markov chain analysis ('expected' frequencies) with the frequencies observed in the database, is indicative of the depletion or the overabundance of specific oligonucleotide sequences.

The expected frequencies of tri-($N_1N_2N_3$), tetra-($N_1N_2N_3N_4$) and penta-($N_1N_2N_3N_4N_5$) nucleotides, respectively, were calculated by the first, second and third order Markov chains as shown below:

$$\begin{aligned} \text{Exp. \% Freq. of } N_1N_2N_3 &= \frac{(\text{Obs. \% Freq. of } N_1N_2)(\text{Obs. \% Freq. of } N_2N_3)}{(\text{Obs. \% Freq. of } N_2)} \\ \text{Exp. \% Freq. of } N_1N_2N_3N_4 &= \frac{(\text{Obs. \% Freq. of } N_1N_2N_3)(\text{Obs. \% Freq. of } N_2N_3N_4)}{(\text{Obs. \% Freq. of } N_2N_3)} \\ \text{Exp. \% Freq. of } N_1N_2N_3N_4N_5 &= \frac{(\text{Obs. \% Freq. of } N_1N_2N_3N_4)(\text{Obs. \% Freq. of } N_2N_3N_4N_5)}{(\text{Obs. \% Freq. of } N_2N_3N_4)} \end{aligned}$$

Obs. % freq. and exp. % freq. are the observed percent frequency and the expected percent frequency, respectively.

The significance of the deviation of the obs. % freq. from the exp. % freq. was judged in two ways. The ratio of obs. % freq./exp. % freq. was calculated for all tetranucleotides and the rank order (referred to as the 'deviation rank') was established based on the magnitude of this ratio. This number varies from 1.345 (for CTAC) to 0.2828 (for CTAG). Overrepresented

tetranucleotides have a rank close to 1, while tetranucleotides underrepresented from the genome have a rank close to 256.

The residual values were calculated using the observed and expected numbers for each oligonucleotide by the formula:

$e^2 = 2 [(\text{obs.}) \ln (\text{obs./exp.}) - (\text{obs.} - \text{exp.})]$; where 'e' is the residual and obs. and exp. are the observed and the expected numbers of the oligonucleotide in the database (19). The residual values reported in the tables have been rounded off to the nearest integer.

RESULTS AND DISCUSSION

The sequence-specificity of VSP repair is such that T:G mismatches in NTWGG/N'GW'CC and CTWGN/GGW'CN' are expected to be repaired to C:G by the process. The observed frequencies of tetranucleotides that are likely to be affected by this process and their expected frequencies based on Markov chain analysis are presented in Table I. All CTWG, TWGG, CCWA, and CWAG sequences are among the rarest tetranucleotides in the genome, while CCWG and CWGG are among the most common (Table I). Further, while the former sequences are present at less than expected frequencies in the genome, the latter sequences are present at higher than expected frequencies (Table I). The numbers reported in this paper were derived from a recent update of the *E. coli* sequence database which contains the sequence of approximately 38% of the *E. coli* genome and which had been processed to remove duplicate entries, overlapping sequences, sequences of cloning vectors and other artifacts (14). It should also be noted that we have carried out this analysis on three earlier updates of the *E. coli* sequence database containing 74,444, 217,128 and 1,233,609 bases respectively (data not shown) and in every case the results regarding the under- and over-representation have been qualitatively the same.

The 'expected' frequencies were calculated such that deviation of this frequency from the observed frequency is not due to a similar deviation of the frequency of a component subsequence. For example, the observed frequency of CTAG sequences is much lower than expected, but this is not due to the known underrepresentation of TAG sequences in the genome (16,

TABLE I
Frequencies of Tetranucleotides in *E. coli*

Sequence	Frequency(%) [*]		Rank [†]			Complement	Frequency(%) [*]		Rank [†]		
	Obs.	Exp.	Obs.	Dev.	Residuals [‡]		Obs.	Exp.	Obs.	Dev.	Residuals [‡]
CCAG	0.7311	> 0.6087	7	15	+20	CTGG	0.7320	> 0.6020	6	10	+22
CTAG	0.0210	< 0.0743	256	256	-31	CTAG	0.0210	< 0.0743	256	256	-31
CAGG	0.5450	> 0.4880	43	52	+11	CCTG	0.5387	> 0.4759	48	39	+12
TAGG	0.0884	< 0.1204	254	247	-13	CCTA	0.0875	< 0.1200	255	248	-13
CCTG	0.5387	> 0.4759	48	39	+12	CAGG	0.5450	> 0.4880	43	52	+11
CTTG	0.2103	< 0.3184	231	254	-27	CARG	0.2156	< 0.3155	224	252	-25
CTGG	0.7320	> 0.6020	6	10	+22	CCAG	0.7311	> 0.6087	7	15	+20
TTGG	0.2922	< 0.4259	185	251	-29	CCAA	0.2818	< 0.4174	191	253	-30

* - The observed (Obs.) frequency is the frequency of the tetranucleotide in the EcoSeq5 database. The expected (Exp.) frequency was calculated by the second order Markov-chain analysis as described in Materials and Methods.

† - The frequency (Freq.) rank is the rank of the observed frequency among the frequencies of all tetranucleotides. The most frequent tetranucleotide has the rank 1. The deviation (Dev.) rank was calculated by dividing the observed frequency by the frequency calculated by Markov-chain analysis and then ranking this ratio for all the tetranucleotides. The most overrepresented sequence has the rank 1, while the most underrepresented sequence has the rank 256.

‡ - Residuals were calculated as described in Materials and Methods.

19–21). The calculation of expected frequency for CTAG by the Markov chain analysis (17, 18) uses the observed frequencies of TAG, CTA and TA. In other words, the depletion of CTAG sequences seen in Table I is above and beyond the known depletion of NTAG sequences. A somewhat different method for predicting the frequencies of oligonucleotides than that used here has been described by Burge et al. (16). When the expected frequencies of tetranucleotides were calculated using the formula described by these researchers, the results were generally similar to those shown in Table I (not shown).

The statistical significance of the difference between the observed and the expected frequencies is represented by the 'residuals'. While the magnitude of the residual value represents the extent of deviation from randomness, the sign of the residuals indicates underrepresentation (negative sign) or overrepresentation (positive sign) of the sequences. The magnitudes of the residuals are so large (between 11 and 31) that it is very unlikely that the differences between the expected and the observed frequencies are the result of random fluctuations in the sample. A second measure of the extent of this deviation is the deviation rank. When the tetranucleotides were ranked based on the magnitude of the ratio, observed frequency/expected frequency, the 'C' containing tetranucleotides ranked high (between 6 and 52), while the 'T' containing tetranucleotides ranked low (between 247 and 256). In every case, the sign of the difference is consistent with the hypothesis that VSP repair does act on T:G mismatches that arise independently of 5-meC deamination. Presumably, the original base pair at some of these sites was T:A and the action of VSP repair changed it to C:G, thereby affecting the tetranucleotide frequencies.

The depletion of CTAG sequences from the *E. coli* genome has been noted before (16, 19, 20). It has also been pointed out that this may be the cause of the infrequency with which several restriction enzymes whose recognition sequences include CTAG cut *E. coli* genome (20). Further, a similar depletion of CTAG-containing restriction sites has been observed for several other enteric bacteria including *Salmonella typhimurium*, *Enterobacter aerogenes* and *Klebsiella pneumoniae* (20). Interestingly, the genomes of a number of enteric bacteria contain homologs of *dcm* and *vsr* genes (M. Lieb and A. Bhagwat, unpublished results) and have methylated Dcm sites [M. Lieb, unpublished results and (22)]. Therefore, it is possible that the depletion of CTAG-

containing restriction sites from non-*E. coli* enterobacterial genomes is also, in part, due to VSP repair.

Saccharomyces cerevisiae DNA does not contain detectable levels of 5-methyl-cytosine and homologues of *dcm* or *vsr* have not been detected in its genome by DNA hybridization (M. Lieb, personal communication). When the observed frequencies of the eleven tetranucleotides expected to be affected by VSP repair were compared to their observed frequencies in the yeast database (15), a consistent pattern of over- and under-representation was not found (Table II). Therefore, the overabundance of 'C' containing tetranucleotides and the depletion of 'T' containing tetranucleotides may be restricted to organisms that can carry out VSP repair.

It should be noted that the repair of T:G mismatches that occur outside the context of 5-meC deamination by VSP repair process to C:G is not mutagenic *per se*. T:G mismatches can arise due to errors in copying either A:T or C:G pairs. If it is assumed that equal numbers of T:G mismatches arise from each type of base pair, the total number of mutations (two out of four daughter molecules) are the same whether or not VSP repair acts on the mismatches prior to replication. This is illustrated in Figure 1. Therefore, VSP repair could alter the base composition of the *E. coli* genome without increasing the overall mutation frequency.

Although, T:G mismatches generated by 5-meC deamination are repaired by the VSP repair process, four Dcm sites in *lacI* gene and one site in *cl* gene in λ prophage have been found to be hot-spots for C to T mutations (3, 4). The level of C to T mutations varies significantly from site to site. If 5-meC deamination and VSP repair were the only forces shaping pentanucleotide frequencies, CCWGG should be somewhat underrepresented, while CTWGG and CCWAG sites should be somewhat overrepresented in the genome. As expected, CCWGG sites are significantly underrepresented in the genome and CTA-GG and CCTAG sites are slightly overrepresented in the genome (Table III). The calculations of the expected frequencies of the pentanucleotides were done using the third-order Markov chain analysis and they take into account the overrepresentation of CCWG and CWGG sequences and the underrepresentation of CTAGN and NCTAG sequences in the genome (Table I). These results suggest that VSP repair is unable to completely prevent C to T transitions at methylated cytosines in DNA. This is consistent with the observation that the Dcm site in *cl* gene in

TABLE II
Frequencies of Tetranucleotides in *S. cerevisiae*

Sequence	Frequency(%) [*]				Complement	Frequency (%) [*]			
	Obs.	≥	Exp.	Frequency Rank [†]		Obs.	≥	Exp.	Frequency Rank [†]
CCAG	0.295	>	0.287	157	CTGG	0.295	<	0.307	158
CTAG	0.224	>	0.223	195	CTAG	0.224	>	0.223	195
CAGG	0.225	<	0.241	193	CCTG	0.203	<	0.223	209
TAGG	0.181	<	0.237	224	CCTA	0.208	<	0.250	206
CCTG	0.203	<	0.223	209	CAGG	0.225	<	0.241	193
CTTG	0.387	<	0.413	105	CAGG	0.592	>	0.549	40
CTGG	0.295	<	0.307	158	CCAG	0.295	>	0.287	157
TTGG	0.551	<	0.554	50	CCAA	0.587	<	0.596	44

* - The observed (Obs.) frequencies were taken from Table 3 in reference 15. The expected (Exp.) frequencies were calculated by the second order Markov-chain analysis as described in Materials and Methods.

† - The frequency rank is the rank of the observed frequency among the frequencies of all tetranucleotides and is taken from Table 3 in reference 15. The most frequent tetranucleotide has the rank 1.

λ prophage is a hot-spot for C to T mutations, despite VSP repair of T:G mismatches at this site (4). However, contrary to expectations, the deamination products CTTGG and its complement CCAAG are underrepresented in the genome (Table III). Although the residual values for this underrepresentation are not as large as the values for CCWGG, they are statistically quite significant. Currently, we have no explanation for the underrepresentation of these pentamers. Perhaps some other processes also affect their abundance.

VSP repair may play a role in shaping the frequencies of trinucleotides. VSP repair-dependent correction of the T:G mismatch in TAG/GTC to C:G has been observed (M. Lieb, unpublished results). It is possible that VSP repair may act on all TWG/GW'C mismatches correcting them to CWG/GW'C. If this were the only process affecting trinucleotide frequencies, one would expect that the trinucleotides TWG and CWA would be underrepresented in the *E. coli* genome while CWG would be overrepresented. This is indeed the case [(19) and data not shown]. This result is difficult to interpret because the abundance of amino acids in proteins and the abundance of tRNAs can result in strong codon bias, which in turn can affect frequencies of trinucleotides in the DNA. Further, the predictions from the known codon bias in *E. coli* and from the strand bias of VSP repair are identical in every case. For example, both CUG and UUG code for leucine. While CUG is the most frequently used leucine codon in *E. coli*, UUG is among the less used codons [frequency of CUG:UUG is about 5:1; (23)]. The observed frequencies of the trinucleotides CTG and TTG in the genome are consistent with the relative frequencies of use of CUG and UUG codons [(21) and data not shown]. However, VSP repair at TTG/GTG mismatch would also predict a similar frequency distribution. Therefore, it is possible that VSP repair at these sites helps in maintaining the biased codon usage in the cell. It may also be the cause of the bias. It should be noted that organisms that are known to contain methylation at Dcm sites and contain *dcm* and *vsr* homologs, such as *Salmonella typhimurium* and *Klebsiella pneumonia* (M. Lieb., unpublished results), contain CUG:UUG bias similar to *E. coli* [5:1 and 15:1, respectively; (23)]; while some of the bacteria that are not known to contain methylation at Dcm sites and do not contain *dcm* and *vsr* homologs, such as *Bacillus subtilis* and *Staphylococcus aureus* (M. Lieb., unpublished results), contain a different bias [2:1 and 1:5, respectively; (23)].

There is presently no evidence that VSP repair can act on T:G mismatches in CTW/GGW' contexts, and the frequencies of the relevant trinucleotides are not consistent with the existence of such repair (data not shown). The symmetry of the Dcm

recognition site around the central base, W, may be a requirement for the mismatch recognition protein(s) and hence mismatches within subsequences of Dcm site such as CTW/GGW' that lack the symmetry may not be corrected by the process.

E. coli contains another mismatch correction system that can also repair T:G mismatches. This process repairs these (and other) mismatches in favor of the strand containing GATC sites methylated in their adenines and hence is referred to as methyl-directed mismatch correction (24). It is most active immediately following DNA replication, when the newly synthesized strand is not methylated at GATC sites. If T:G mismatches arise due to replication errors, VSP repair and methyl-directed correction process would be expected to compete for the correction of the mismatch. If the original base pair were T:A, then the two correction processes would tend to work against each other. In experiments involving transfection of phage heteroduplexes, Zell and Fritz (11) have shown that such competition does occur and that VSP repair is as efficient- or better-at correcting T:G mismatches as the methyl-directed process. Therefore, a large fraction of T:A base pairs should be converted to C:G by VSP repair when replication error results in the creation of a T:G mismatch in the appropriate sequence context. Cytosine

Figure 1. Mutagenic Consequences of VSP Repair. The results of replication of a T:G mismatch, with or without prior correction by VSP repair, are shown. In pathway I, replication is assumed occur without repair, while in pathway II the mismatch is corrected to C:G prior to replication.

TABLE III
Frequencies of Pentanucleotides in *E. coli*

Sequence	Frequency(%) [*]				Complement	Frequency(%) [*]			
	Observed	≥	Expected	Residuals [†]		Observed	≥	Expected	Residuals [†]
CCAGG	0.1295	<	0.1733	-15	CCTGG	0.1307	<	0.1750	-15
CTAGG	0.0041	>	0.0033	+ 2	CCTAG	0.0041	>	0.0032	+ 2
CCTGG	0.1307	<	0.1750	-15	CCAGG	0.1295	<	0.1733	-15
CTTGG	0.0287	<	0.0385	- 7	CCARG	0.0305	<	0.0385	- 6

^{*} - The observed (Obs.) frequency is the frequency of the tetranucleotide in the EcoSeq5 database. The expected (Exp.) frequency was calculated by the second order Markov-chain analysis as described in Materials and Methods.

[†] - Residuals were calculated as described in Materials and Methods.

deamination can occur at any time during the life cycle of the cell. Hence most T:G mismatches created due to 5-mC deamination will occur after the methylation of GATC sites in the daughter strand and will be subject to VSP repair alone.

There are mechanisms in addition to replication errors and 5-mC deamination that could give rise to mismatches that would be subject to VSP repair. This repair process was originally identified based on its ability to bias the results of certain recombinational crosses between phage λ mutants (5, 6). If it has a similar influence on chromosomal recombination, it would be a significant force in shaping the composition of genomes of bacteria that frequently exchange genetic material by conjugation or by other means. As most enteric bacteria contain methylation at GATC sites (25), the recombination intermediates resulting from such an exchange would not be subject to methyl-directed mismatch correction. Therefore, VSP repair may be a form of molecular 'drive' (26) for genome evolution.

Deamination of (unmethylated) cytosines result in U:G mismatches which are repaired by the *ung*-dependent process to C:G (27). VSP repair might compete with this process also to correct the mismatch to C:G. Deamination of adenine to hypoxanthine is less frequent than the deamination of cytosine, but it can be detected *in vitro* (27). Most of the T:I (I = inosine) mispairs resulting from deamination should also occur independent of DNA replication and hence should not be subject to methyl-directed correction. If VSP repair acts on these mismatches in a way similar to its action on T:G mismatch, a C:I pair should result from the repair. VSP repair, in effect, would fix the mutagenic product of deamination.

Our conclusion that the VSP repair process affects the frequencies of certain oligonucleotides in the *E. coli* genome has predictive value:

(1) As the DNA sequence databases increase in size, it should be possible to carry out an analysis similar to that presented here on sequences from other bacteria. It would be expected that enteric bacteria that contain *vsr* homologs and hence are likely to contain VSP repair would contain sequence biases similar to the ones described here for *E. coli* K-12. In contrast, non-enteric bacteria do not contain *vsr* homologs (M. Lieb and A. Bhagwat, unpublished results) and their genomes should not contain VSP repair-mediated bias.

(2) Mismatch correction processes other than VSP repair and methyl-directed mismatch correction are known. These include the T:G to C:G correction process in human cells (28) and the G:A to G:C correction system in *E. coli* K-12 (29, 30). Although, these processes act to avoid mutations during vegetative growth, during recombination they could bias the result in a manner analogous to VSP repair. In that case, they may also influence the composition of their host genome. Oligonucleotide frequency analysis of sequence databases may also reveal additional systematic sequence biases. Some of these may be the result of strand-biased mismatch correction processes. For example, GGCC is the second least frequent tetranucleotide in the *E. coli* genome. Using transfection of appropriate heteroduplexes as an assay, it may be possible to test biased repair in this sequence context.

(3) The data regarding tetranucleotide frequencies of *E. coli* K-12 are consistent with repair of T:G mismatches in all four combinations of tetranucleotide subsets of the Dcm recognition sequence (Table I). However, the existence of VSP repair has not been shown with CTTG/GGAC and TTGG/GACC mismatches. The data are also consistent with the repair of both

TWG/GW'C mismatches, but not with the correction of CTW/GGW' mismatches. Correction of TAG/GTC has been observed (M. Lieb, unpublished results). Possible correction in the remaining sequence contexts can be tested.

VSP repair differs from the methyl-directed mismatch correction in *E. coli* in that it appears to be unselective. It changes a T:G mismatch to a C:G even in situations where the T was the original correct base. As a result, it introduces T to C mutations at some sites, while preventing C to T mutations at other sites. It resembles *umuCD*-dependent repair in *E. coli* in that it can introduce mutations in the genome. It may not, however, affect the overall mutation rate in the cell significantly. If the *E. coli* genome is considered to contain coding regions throughout its length, then a CCWGG to CTWGG mutation has a 1 in 6 probability of creating an in-frame amber (TAG) codon. Perhaps, the prevention of nonsense mutations by VSP repair offsets the price the organism pays in terms of missense mutations at other sites. Consequently, the sequence composition of the genome is altered in a significant way.

ACKNOWLEDGEMENTS

The authors are grateful to K. Rudd (National Institutes of Health) for making EcoSeq5 database and its analysis available prior to its public distribution and to M. Lieb (University of Southern California School of Medicine) for communicating unpublished results. The comments given by M. Lieb and K. Rudd on the manuscript were also appreciated. A. S. B. would like to thank G. Phillips (Princeton University) and E. Raleigh (New England Biolabs) for help with the analysis of earlier versions of the database. He would also like to acknowledge his stimulating discussions with M. Marinus (University of Massachusetts School of Medicine) and M. Ehrlich (Tulane University) concerning this work. A.S.B. is the recipient of Research Career Development Award from the National Institutes of Health (NIH) and the research in his laboratory is supported by NIH grant GM40576 and by a grant from Boehringer Mannheim GmbH. M. M. is supported by NIH grants HG00106 and HG00456 and is co-P.I. on USDA grant 91-37305-6719.

REFERENCES

1. Grantham, R. C., Gautier, C., Gouy, M., Mercier, R. and Pavé, A. (1980) *Nucleic Acids Res.*, **8**, r49-r62.
2. Schlagman, S., Hattman, S., May, M. S. and Berger, L. (1976) *J. Bacteriol.*, **126**, 990-996.
3. Coulondre, C., Miller, J. H., Farabaugh, P. J. and Gilbert, W. (1978) *Nature*, **274**, 775-780.
4. Lieb, M. (1991) *Genetics*, **128**, 23-27.
5. Lieb, M. (1983) *Mol. Gen. Genet.*, **191**, 118-125.
6. Lieb, M. (1985) *Mol. Gen. Genet.*, **199**, 465-470.
7. Lieb, M., Allen, E. and Read, D. (1986) *Genetics*, **114**, 1041-1060.
8. Lieb, M. (1987) *J. Bacteriol.*, **169**, 5241-5246.
9. Jones, M., Wagner, R. and Radman, M. (1987) *J. Mol. Biol.*, **194**, 155-159.
10. Jones, M., Wagner, R. and Radman, M. (1987) *Cell*, **50**, 621-626.
11. Zell, R. and Fritz, H.-J. (1987) *EMBO J.*, **6**, 1809-1815.
12. Sohail, A., Lieb, M., Dar, M. and Bhagwat, A. S. (1990) *J. Bacteriol.*, **172**, 4214-4221.
13. Hennecke, F., Kolmar, H., Bründl, K. and Fritz, H.-J. (1991) *Nature*, **353**, 776-778.
14. Rudd, K. E., Miller, W., Werner, C., Ostell, J., Tolstoshev, C. and Satterfield, S. G. (1991) *Nucleic Acids Res.*, **19**, 637-647.
15. Arnold, J., Cuticchia, A. J., Newsome, D. A., Jennings III, W. W. and Ivarie, R. (1988) *Nucleic Acids Res.*, **16**, 7145-7158.
16. Burge, C., Campbell, A. M. and Karlin, S. (1992) *Proc. Natl. Acad. Sci (USA)*, **89**, 1358-1362.

17. Karlin, S. (1966) *A First Course in Stochastic Processes*. Academic Press.
18. Basawa, I. V. and Prakash-Rao, B. L. S. (1980) In (ed.), *Statistical Inference for Stochastic Processes*, Academic Press, Chap. 4.
19. Phillips, G. J., Arnold, J. and Ivarie, R. (1987) *Nucleic Acids Res.*, **15**, 2611–2626.
20. McClelland, M., Jones, R. and Nelson, M. (1987) *Nucleic Acids Res.*, **15**, 5985–6005.
21. Phillips, G. J., Arnold, J. and Ivarie, R. (1987) *Nucleic Acids Res.*, **15**, 2627–2638.
22. Gomez-Eichelmann, M. C., Levy-Mustri, A. and Ramirez-Santos, J. (1991) *J. Bacteriol.*, **173**, 7692–7694.
23. Wada, K.-n., Wada, Y., Doi, H., Ishibashi, F., Gojobori, T. and Ikemura, T. (1991) *Nucleic Acids Res.*, **19**, 1981–1986.
24. Radman, M., Wagner, R. E., Glickman, B. W. and Meselson, M. (1980) In M. Alacevic (ed.), *Progress in Environmental Mutagenesis*, Elsevier/North Holland, pp. 121–130.
25. Brooks, J. E., Blumenthal, R. M. and Gingeras, T. R. (1983) *Nucleic Acids Res.*, **11**, 837–851.
26. Dover, G. A. (1982) *Nature*, **299**, 111–117.
27. Lindahl, T. (1979) *Progress Nucleic Acid Res.*, **22**, 135–192.
28. Brown, T. C. and Jiricny, J. (1987) *Cell*, **50**, 945–950.
29. Radicella, J. P., Clark, E. A. and Fox, M. S. (1988) *Proc. Natl. Acad. Sci. (USA)*, **85**, 9674–9678.
30. Au, K. G., Clark, S., Miller, J. H. and Modrich, P. (1989) *Proc. Natl. Acad. Sci. (USA)*, **86**, 8877–8881.