Nucleic Acids Research, Vol. 20, No. 7 ***1657–1662***

Statistical evaluation and biological interpretation of non-random abundance in the *E.coli* K-12 genome of tetra- and pentanucleotide sequences related to VSP DNA mismatch repair

Rainer Merkl, Manfred Kröger[1], Peter Rice[2] and Hans-Joachim Fritz*

Institut für Molekulare Genetik, Georg-August-Universität Göttingen, Grisebachstraße 8, W-3400 Göttingen, [1]Institut für Mikrobiologie und Molekularbiologie, Justus-Liebig-Universität Gießen, Frankfurter Straße 107, W-6300 Gießen and [2]EMBL, Postfach 10.2209, Meyerhofstraße 1, W-6900 Heidelberg, Germany

Received December 16, 1991; Revised and Accepted February 26, 1992

ABSTRACT

The abundance of all tetra- and pentanucleotide sequences is calculated for a set of DNA sequence data comprising 767,393 nucleotides of the *E. coli* K-12 genome. Observed frequencies are compared to those expected from a Markov chain prediction algorithm. Systematic and extreme non-random representations are found for special sets of sequences. These are interpreted as arising from incorporation of a 2′-deoxyguanosine residue opposite thymidine during replication which, in special sequence contexts, leads to a T/G mismatch that is simultaneously substrate for two competing DNA mismatch repair systems: the *mutHLS* and the VSP pathway. Processing by the former leads to error correction, by the latter to mutation fixation. The significance of the latter process, as demonstrated here, makes it unlikely that VSP repair has evolved mainly as a mutation avoidance mechanism. It is proposed that in *E. coli* K-12, VSP repair, together with DNA cytosine methylation, constitutes a mutagenesis/recombination system capable of promoting gene-conversion-like unidirectional transfer of short stretches of DNA sequence.

INTRODUCTION

In *Escherichia coli* K-12, the Dcm DNA cytosine methyltransferase catalyzes transfer of a methyl group from S-adenosyl methionine (SAM) onto the 5-position of the inner cytosine residue of the target sequence CC$^{A}/_{T}$GG (Figure 1, structure [I])[1-3]. Such sites have been identified as hotspots of spontaneous transition mutation (5meC to T)[4-5]. As for now, hydrolytic deamination of 5meC residues provides the simplest explanation of these hotspots[4,5], despite the proven existence in *E. coli* K-12 of an efficient DNA mismatch repair mechanism (*very short patch* or VSP repair) acting on the T/G mismatch that is the primary product of the deamination reaction (Figure 1, structure [II]) [6]. The initial, fully methylated sequence [I] is restored *via* VSP repair product [IV]. Within this picture, the mutation process is interpreted as escape of the mismatched, premutagenic intermediate [II] from VSP repair into DNA replication, which yields structure [III].

VSP repair is initiated by an endonucleolytic cut on the 5′-side of the mismatched thymidine residue. This cut is catalyzed by the Vsr gene product, a strand- and sequence-specific DNA mismatch endonuclease[7]. The substrate requirements of Vsr endonuclease are defined by structures [II] and [V], Figure 1 and, more generally, by structures [VIII] and [XI], Figure 2. In other words, Vsr endonuclease recognizes a T/G mismatch in the specific context of the target sequence of Dcm methylation; the first or the last nucleotide pair of this sequence, however, may deviate. Presence of a cytosine-5-methyl group on the uncleaved strand is not essential[7]. This biochemical characterization of Vsr endonuclease is in complete accord with genetic data on VSP repair[8-11].

Previously, we have developed an assay for the quantitative assessment of DNA mismatch repair acting on a heteroduplex DNA molecule derived from the phage M13 genome[12]. With this assay, we demonstrated that VSP repair and *mutHLS* repair (the post-replicative error correction pathway of *E. coli*) can compete for one and the same substrate site[9]. If one assumes this competition not to be restricted to the experimental situation of transfecting *E. coli* with heteroduplex DNA, it must be expected to have profound consequences for the frequency of occurence of certain tetra- and pentanucleotide sequences in the genome of *E. coli* K-12. For an illustration of this point, consider the right branch of Figure 1. Structure [V] can arise from structure [III] by misincorporation of a 2′-deoxy-guanosine

* To whom correspondence should be addressed

residue opposite thymidine during replication. T/G mismatches are generally corrected by the *mutHLS* repair system with very good efficiency[12]. In the special case of structure [V], however, the T/G mismatch is at the same time a substrate of VSP repair. Successful competition of VSP repair for the mismatch[9] will result in active fixation of the mutation (route [III], [V], [VI], [I]). Hence, one must predict a higher frequency of T to C transition mutation for such cases, in which the mutation mechanism can proceed *via* a replication error leading to a mismatched intermediate that constitutes a substrate site of VSP repair (Figure 1, structure [V] and Figure 2 structures [VIII] and [XI]).

On an evolutionary timescale, therefore, the process illustrated in Figure 2 must be expected to result in progressive depletion of the *E. coli* K-12 genome of a special set of tetranucleotide sequences (Figure 3, Table 1, A−G) and, correspondingly, sequences of another set (Figure 3, Table 1, H−K) are predicted to accumulate. Within this set of tetranucleotide sequences, a special subset of pentanucleotide sequences is characterized by its tendency to undergo enhanced mutagenesis also in the reverse direction (Figure 1; Table 2, L−Q). For these pentanucleotide sequences, one expects the trend described above to be counteracted to a degree which depends on the relative rates of the forward and the backward process (designated 'gain' and 'loss', respectively, in Figure 1). Here we demonstrate by statistical analysis of the current DNA sequence data base of the *E. coli* K-12 genome that these predictions are indeed borne out and we discuss biological implications of this finding. In particular, we offer an explanation for the evolutionary significance of DNA cytosine methylation in *E. coli* K-12 and discuss the possible general role VSP-like DNA mismatch repair pathways may have in patchwise gene conversion. The extreme non-random occurence in the *E. coli* genome of some of the tetranucleotide sequences under consideration here has been noticed earlier by statistical analysis of a considerably smaller data set[13, 14]; to date, however, these sequences were neither systematically grouped together nor was the biological significance of the phenomenon explained.

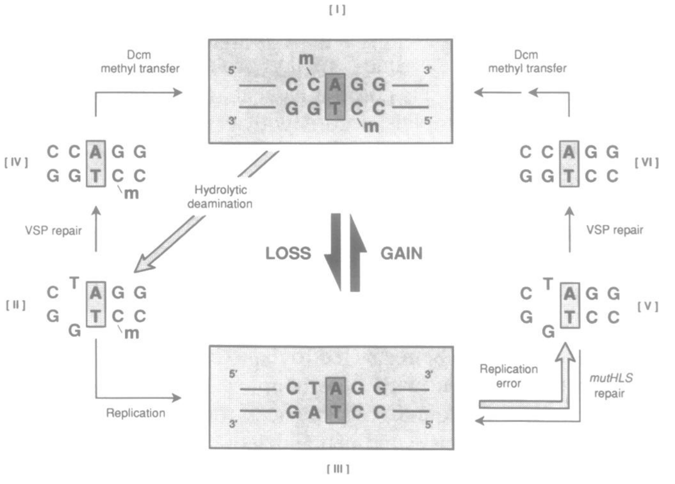

Figure 1. Loss and gain of DNA cytosine methylation sites in the *E. coli* K-12 genome by two processes of spontaneous mutagenesis working in opposite directions. The central A/T nucleotide pair indicated by shading in structures [I] to [VI] can be inverted.

MATERIALS AND METHODS

A data set of the *E.coli* K-12 genome containing 767,393 nucleotides of the strand running 5′ to 3′ in clockwise map direction was extracted from the ECD (*E. coli* sequence data base−release 9, EMBL Heidelberg; available on CD-ROM)[15]. The data set comprises all contiguous stretches of DNA sequence longer than 5.000 nucleotides; these are ordered with respect to map direction and are free of overlaps. This data set minimizes bias for coding regions; it represents roughly half of the total DNA sequence deposited in the ECD.

The data set of the *Bacillus subtilis* genome (194.634 nucleotides) was extracted from GenBank[16] (Release 67.0 3/91), selecting for *B. subtilis* strain 168. The DNA of this strain is not methylated in its $CC^{A}/_{T}GG$ sites[17].

As pointed out by Phillips *et al.*[13], Markov chains can be used to predict the frequency of any sequence motif from observed frequencies of shorter sequences of which the motif is made up in such a way that carry-over of non-randomness within the shorter sequences is eliminated and possible biological effects acting at the sequence length of the motif under consideration are highlighted.

We used the following equations to calculate the expected frequencies p_M of tetra- and pentamer sequences in the two data sets described above.

$$\text{I)}\quad p_{M\,4.2}\,(a_1a_2a_3a_4) \;:=\; \frac{p(a_1a_2a_3) * p(a_2a_3a_4)}{p(a_2a_3)}$$

$$\text{II)}\quad p_{M\,5.2}\,(a_1a_2a_3a_4a_5) \;:=\; \frac{p(a_1a_2a_3) * p(a_2a_3a_4) * p(a_3a_4a_5)}{p(a_2a_3) * p(a_3a_4)}$$

$$\text{III)}\quad p_{M\,5.3}\,(a_1a_2a_3a_4a_5) \;:=\; \frac{p(a_1a_2a_3a_4) * p(a_2a_3a_4a_5)}{p(a_2a_3a_4)}$$

where $a_i \in \{A, T, C, G\}$ and corresponding p-values are frequencies of dimer, trimer and tetramer sequences extracted from the data set.

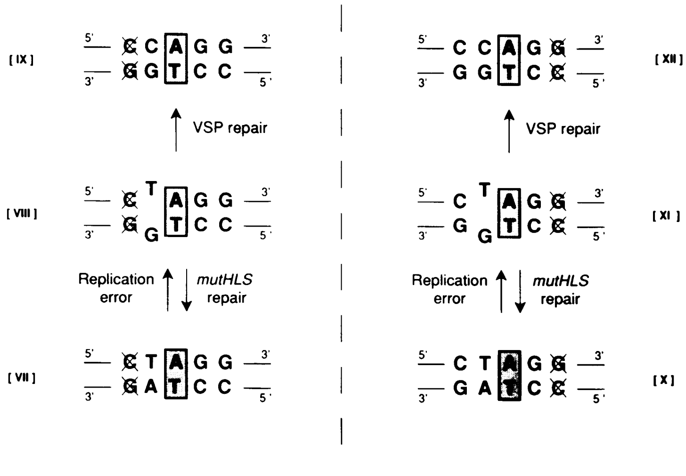

Figure 2. Spontaneous mutagenesis process leading to depletion from the *E. coli* K-12 genome of one set of tetranucleotide sequences and accumulation of another. For the unidirectional process illustrated, the crossed-out base pair is not allowed at the respective position indicated. The central A/T nucleotide pair indicated by shading in structures [VII] to [XII] can be inverted.

The ratio observed frequency p divided by expected frequency p_M is a measure of non-statistical over- or underrepresentation of the corresponding sequence. For computations, a MicroVAX 3200 was used under VMS. Programs were written in PASCAL and are available from the authors on request.

RESULTS

Gleaning information about biological processes from experimental DNA sequence data necessarily depends on detection and interpretation of non-random features of the nucleotide sequences under consideration. Extracting from a DNA sequence data base the frequency of a given oligonucleotide sequence motif is straightforward, not so the decision whether or not any such observed frequency deviates conspicuously enough from statistical expectation to make it worthy an attempt to underlay it with biological interpretation.

It is well known, for example, that the different trinucleotide sequences are represented quite differently in the *E. coli* genome[13], with some frequencies deviating drastically from values one might expect on the basis of the frequencies of individual nucleotides. As has been pointed out before[18], this distortion at the level of trinucleotides is due—if not alone so at least to some extent—to peculiarities of the genetic code and codon preferences of *E. coli*. If one now tries to discuss non-statistical occurences of sequence motifs of more than three nucleotides length, it is not trivial to separate any effect that is specific for, *e.g.*, a given subset of all possible tetranucleotide sequences from the distortion already present at the level of trinucleotides of which the tetranucleotides are made up. A statistical prediction procedure based on Markov chains (see Materials and Methods) takes such distortions into account and is therefore able to overcome this problem[13].

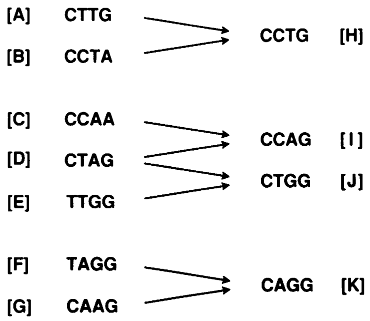

Figure 3. Correlation diagram showing the complete set of eight sequence transitions as they result from the processes illustrated in Figure 2.

Table 1. Selected tetranucleotide frequencies in *Escherichia coli* and in *Bacillus subtilis* 168.

		E. coli		*B. subtilis*	
		frequency * 100	$\text{ratio}_{4.2}$	frequency * 100	$\text{ratio}_{4.2}$
A	CTTG	0.21	0.68	0.47	1.17
B	CCTA	0.09	0.71	0.12	0.89
C	CCAA	0.27	0.67	0.28	1.00
D	CTAG	0.02	0.29	0.10	0.93
E	TTGG	0.29	0.68	0.38	0.97
F	TAGG	0.09	0.81	0.15	0.88
G	CAAG	0.21	0.66	0.45	0.97
H	CCTG	0.58	1.12	0.30	1.03
I	CCAG	0.69	1.20	0.19	0.79
J	CTGG	0.82	1.22	0.30	0.89
K	CAGG	0.54	1.13	0.40	0.98

Tetranucleotide sequences A–K and their correlation by mutation processes are illustrated in Figures 2 and 3. The frequency values are extracted from the DNA sequence data base. $\text{Ratio}_{4.2}$ is observed frequency divided by expected frequency $p_{M4.2}$ (see Materials and Methods).

Figure 3 summarizes the complete set of sequence transitions that result from the processes illustrated in Figure 2 (note that in all structures shown in Figure 2, the shaded A/T base pair can be inverted). Consequently, tetranucleotide sequences A–G are predicted as under-represented, sequences H–K as over-represented. These expectations are fully borne out by the calculations (see Table 1). Ratios of observed divided by predicted frequencies range from 0.29 to 0.81 for sequences A–G and from 1.12 to 1.22 for sequences H–K. As a control, the same calculations were carried out for the *Bacillus subtilis* data set. This bacterium does not methylate DNA cytosine residues within the $CC^{A}/_{T}GG$ sequence context[17] and is therefore predicted not to display the frequency pattern observed with *E. coli* K-12. Indeed, the ratio values in this case are generally closer to unity and there is no systematic trend discernible. The left panel of Figure 4 displays absolute frequencies and ratios for the entire set of 256 tetranucleotides as extracted from the *E. coli* data set in a two-dimensional fashion. The four tetranucleotide sequences predicted as over-represented are indicated individually. It is evident that sequences H–K are not only located above the unity value on the ordinate but also belong to the most frequent tetranucleotide sequences in absolute terms. CTGG is the third most frequent tetranucleotide sequence in the entire data base. The right panel of Figure 4 is an enlarged version of the lower left corner of the diagram, indicated in the left panel by shading. In this area of lowest absolute frequency and lowest ratio one finds all seven tetranucleotide sequences A–G (highlighted by filled circles and bold-face print). Note that absolute frequencies alone are not sufficient to appreciate the under-representation of these sequences. A fairly large number of tetranucleotide sequences are also quite rare, but solely because of each being composed of two rare trinucleotides (*i.e.* they have a ratio value close to unity); some are even at the same time rare in absolute terms and unexpectedly frequent compared to their frequency predicted by the Markov chain algorithm. An especially striking corroboration of our hypothesis comes from the extreme under-representation of CTAG. Note that due to its symmetry, this is the only tetranucleotide sequence that can be

Table 2. Selected pentanucleotide frequencies in *Escherichia coli* and in *Bacillus subtilis* 168.

		E. coli			*B. subtilis*		
		frequency * 100	$\text{ratio}_{5.2}$	$\text{ratio}_{5.3}$	frequency * 100	$\text{ratio}_{5.2}$	$\text{ratio}_{5.3}$
L	CTAGG	0.005	0.28	1.16	0.022	0.91	1.10
M	CCTAG	0.006	0.39	1.86	0.012	0.65	0.78
N	CTTGG	0.034	0.38	0.84	0.098	1.23	1.09
O	CCAAG	0.027	0.32	0.72	0.065	1.01	1.04
P	CCAGG	0.129	1.04	0.77	0.036	0.64	0.83
Q	CCTGG	0.147	1.03	0.75	0.035	0.60	0.66

Pentanucleotide sequences L–Q and their correlation by mutation processes are illustrated in Figure 1. $\text{Ratio}_{5.2}$ and $\text{ratio}_{5.3}$ are observed frequency divided by expected frequency $p_{M5.2}$ and $p_{M5.3}$ respectively. Also see legend to Table 1.

used by the same VSP repair-driven process in two different fashions (Figure 3); this leads to two different exits depleting the same pool. On the other hand, sequences H−K are over-represented to different degrees. Different efficiencies of processing the corresponding mismatched intermediates (Figure 2, structures [VIII] and [XI] by VSP repair could provide an explanation. Since purified Vsr endonuclease has recently become available[7], this working hypothesis is now amenable to experimental test.

The eleven tetranucleotide sequences A−K (Table 1) define a family of 42 pentanucleotide sequences which participate in the processes illustrated in Figures 1 to 3. Of these, 28 sequences are derived from tetramers A−G and 14 sequences from tetramers H−K. Within this set, pentanucleotide sequences L−Q (Table 2) are special in the sense that they fit the reaction scheme of Figure 1, *i.e.* for these sequences one not only has to take into consideration the mutation fixation process driven by VSP repair, but also the reverse mutation caused by hydrolytic deamination of 5meC. For this particular subset of pentanucleotide sequences, therefore, the trend of nonrandom occurences observed at the tetranucleotide level can be expected to be diminished to a smaller or larger extent, depending on the relative rates of the two processes operating in opposite directions. These expectations are borne out, as made evident by the data summarized in Table 2: Ratios calculated using second order Markov chains show essentially the same trend as for the tetranucleotides summarized in Table 1. If, however, third order Markov chains are used to calculate ratios for the same pentanucleotide sequences, a reversion (sequences L, M, P, Q) or at least a strong attenuation of this trend (sequences N and O) is observed. Again, values extracted from the *B. subtilis* data base serve as a control.

Observed frequencies and ratios as derived from third order Markov chain are plotted in Figure 5 for the entire set of 1024 pentanucleotide sequences. Overall, ratio values for this set are clustered much more closely around the unity value than for the set illustrated in Figure 4. The data points representing the 42 pentanucleotide sequences defined above are highlighted by a circle. The special subset of pentanucleotide sequences L−Q (Table 2) is indicated by bold-face print. In addition, three extreme examples of the 42 pentanucleotides set are given in italics. Note that observed frequencies for pentanucleotide set L−Q are in striking contrast to what would be expected by only considering increased mutagenesis by hydrolytic deamination of 5meC residues.

Sequences B, D and F are the three least abundant tetranucleotides in the entire data base (Table 1, Figure 4). If one calculates their respective predicted frequencies from first order Markov chains, the deviations between observed and predicted values are even more extreme (data not shown). This is due to

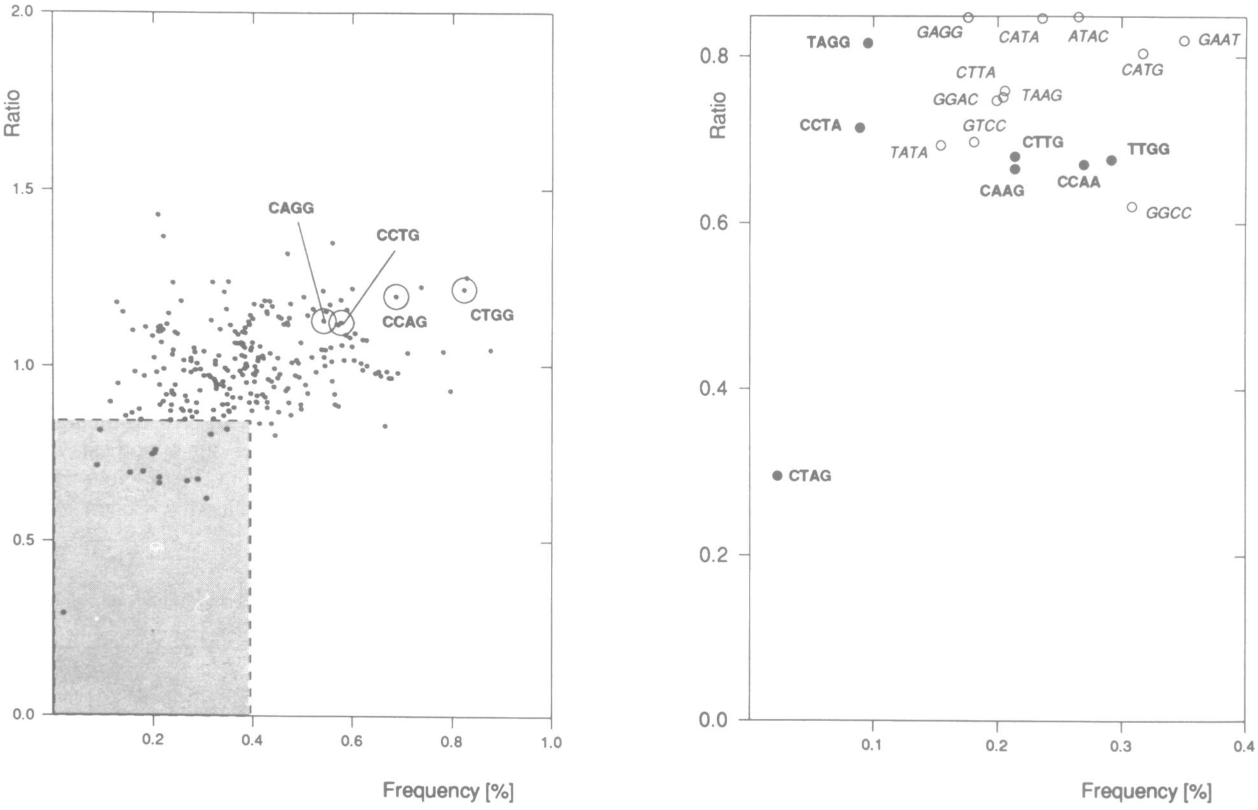

Figure 4. Two-dimensional plot of frequencies and ratios for the complete set of tetranucleotide sequences extracted from the *E. coli* K-12 sequence data set. Abscissa: Observed frequency in per cent. Ordinate: Observed frequency divided by frequency $p_{M4,2}$ predicted from second order Markov chain (I), see Materials and Methods. Subscript indeces to probabilities and ratios are used in the following sense: M points to the fact that this probability was derived from a Markov chain, the first number denotes the length of nucleotide string under consideration, the second number indicates the order of the Markov chain used. In the left panel, tetranucleotide sequences H−K (see Table 1) are indicated by a circle around the respective data point. The right panel is a blow-up of the shaded lower left corner of the left panel. Tetranucleotide sequences A−K, Table 1, are indicated by a filled circle and straight, bold-face print.

the strong under-representation of TAG and CTA, to which we have intentionally blinded the prediction procedure. If, however, the substrate requirements of Vsr endonuclease, and with it of VSP repair, were relaxed to the extent that significant activity would be exerted on sites with *both* nucleotide pairs flanking the central triplet degenerate (compare Figure 2), it would seem possible that this striking under-representation of TAG and C-TA itself could be caused by the described VSP repair-driven process. To date, however, we have not been able to detect any such activity of Vsr endonuclease in an *in vitro* cleavage assay (W. Gläsner, this laboratory, unpublished).

DISCUSSION

The data presented here lend strong support to the assumption that competition between the VSP and the *mutHLS* pathways of DNA mismatch repair is indeed important in *E. coli* K-12 and that this competition is a very significant source of spontaneous mutations. This notion makes it necessary to reconsider the evolutionary significance of VSP mismatch repair, formerly thought to be primarily responsible for mutation avoidance. Disposal of the entire *dcm/vsr* locus would not only make unnecessary any mechanism of counteracting the mutagenic effect of 5me-C deamination but would simultaneously avoid mutagenesis by VSP repair itself. Hence, maintenance of the *dcm/vsr* locus can only be explained on the basis of a biological function of Dcm methylation associated with a significant selective value. To date, the search for such a role has been notoriously unsuccessful.

By placing emphasis on mutagenesis rather than mutation avoidance, we can now propose a mechanism of action of the Dcm/Vsr enzyme couple that allows (on an evolutionary time scale) rapid interconversion of states [I] and [III], Figure 1. This interconversion, for which we provide statistical evidence, must necessarily lead to increased occurence of sequence polymorphisms associated with such sites in larger populations of *E. coli*. Similar sequence polymorphisms must accompany irreversible T/A to C/G transitions as illustrated in Figures 2 and 3.

As we have pointed out earlier[7], these polymorphisms have interesting implications for genetic recombination. Consider a recombination event between two cells whose genomes differ *i.a.* in one or more such sites. If strand exchange passes through that site, heteroduplex DNA is formed with a T/G mismatch (in one out of two possible strand combinations) that is a substrate of VSP repair. Vsr endonuclease will incise next to the mismatched thymidine residue[7] and DNA polymerase I commence repair synthesis[19] with the short synthesis tract typical for that enzyme. Any additional base/base mismatch located within the length of that synthesis tract will be passively co-repaired. As a result, a short stretch of DNA sequence will be copied in a complementary fashion from one strand onto the other.

In summary, VSP repair, rather than being a cellular device of mutation avoidance, may well constitute (together with Dcm-

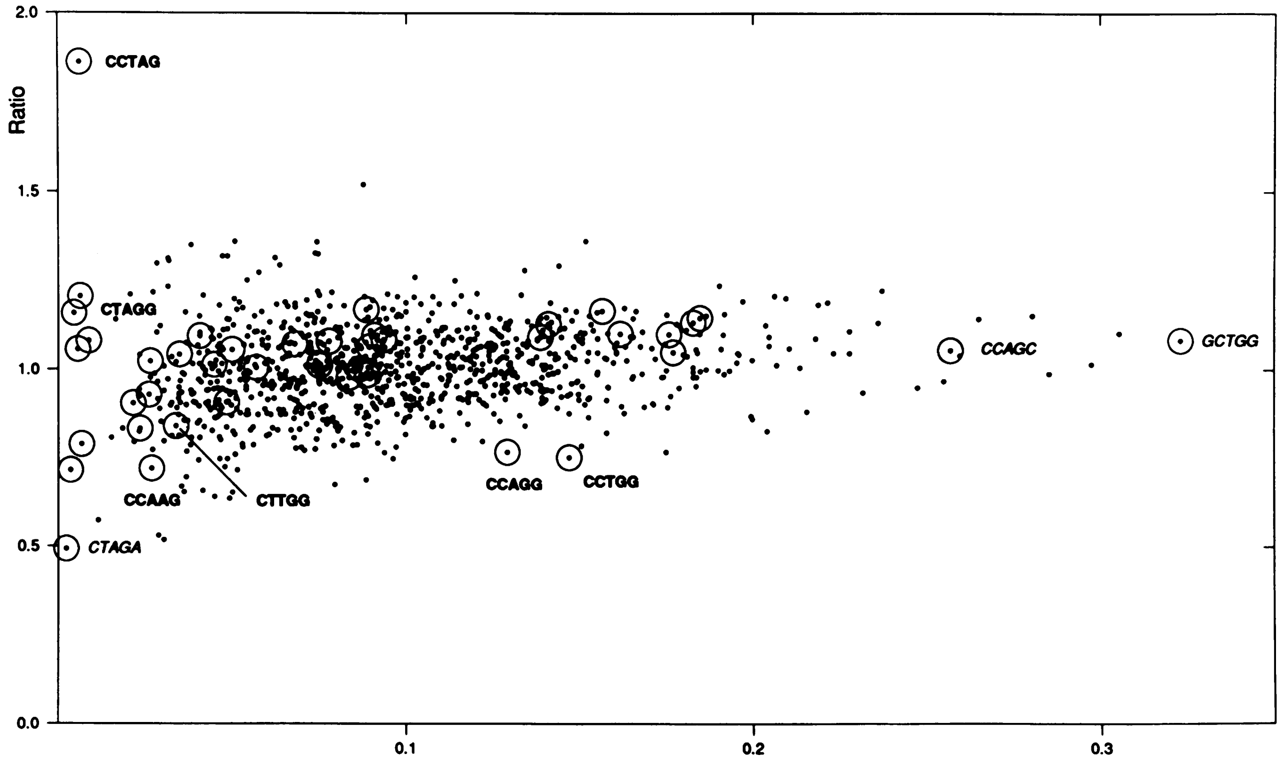

Figure 5. Two-dimensional plot of frequencies and ratios for the complete set of pentanucleotide sequences extracted from the *E. coli* K-12 sequence data set. Abscissa: Observed frequency in per cent. Ordinate: Observed frequency divided by frequency $p_{M5,3}$ predicted from third order Markov chain (III), see Materials and Methods. Pentanucleotide sequences L−Q, Table 2, are indicated by a filled circle and straight, bold-face print. The entire set of 42 pentanucleotides contained in the sequence family described by structures [VII], [IX], [X] and [XII], Figure 2, are indicated by a circle around the respective data point.

mediated DNA cytosine methylation) a mutagenesis/recombination system capable of promoting unidirectional transfer of short patches of DNA sequence.

In principle, such a mechanism could be sustained without DNA methylation. Only the latter, however, makes the mutation event reversible (see Figure 1) and can thus keep the process of creating sequence polymorphisms going without time limits. We propose that the biological significance of DNA cytosine methylation in *E. coli* K-12 may lie in this stimulation of a special type of recombination. Since this mechanism requires partners of recombination that are genetically separated by some distance, it would no longer seem surprising that under laboratory conditions, *i.e.* during work with closely related derivatives of one experimental *E. coli* isolate, it is difficult to identify a conspicuous phenotype associated with *dcm* mutations.

Unidirectional transfer of genetic information by a molecular mechanism as sketched above can result in gene conversion phenomena. The somatic diversification of chicken immunoglobulin genes, for example, is interpreted as resulting from gene conversion[20]. It thus seems possible that the mechanism outlined above provides a paradigm beyond *E. coli* and the prokaryotes and is at the core of some of such gene conversion phenomena.

ACKNOWLEDGEMENTS

Active help by Ralf Wahl and Gabriel Schachtel in preparing the ECD database is gratefully acknowledged. This work was supported by Deutsche Forschungsgemeinschaft through *Forschungsschwerpunkt Gentechnologie* and SFB 272.

REFERENCES

1. Boyer,H.W., Chow,L.T., Dugaiczyk,A., Hedgpeth,J. and Goodman,H.M. (1973) *Nature New Biology*, **244**, 40–43.
2. May,M.S. and Hattman,S. (1975) *J. Bacteriol.*, **122**, 129–138.
3. Schlagman,S., Hattman,S., May,M.S. and Berger,L. (1976) *J. Bacteriol.*, **126**, 990–996.
4. Coulondre,C., Miller,J.H., Farabaugh,P.J. and Gilbert,W. (1978) *Nature*, **274**, 775–780.
5. Duncan,B.K. and Miller,J.H. (1980) *Nature*, **287**, 560–561.
6. Lieb,M. (1991) *Genetics*, **128**, 23–27.
7. Hennecke,F., Kolmar,H., Bründl,K. and Fritz,H.-J. (1991) *Nature*, **353**, 776–778.
8. Lieb,M., Allen,E. and Read,D. (1986) *Genetics*, **114**, 1041–1060.
9. Zell,R. and Fritz,H.-J. (1987) *EMBO J.*, **6**, 1809–1815.
10. Jones,M., Wagner,R. and Radman,M. (1987) *J.Mol.Biol.*, **194**, 155–159.
11. Sohail,A., Lieb,M. Dar,M. and Bhagwat,A.S. (1990) *J. Bacteriol.*, **172**, 4214–4221.
12. Kramer,B., Kramer,W. and Fritz,H.-J. (1984) *Cell*, **38**, 879–887.
13. Phillips,G.J., Arnold,J. and Ivarie,R. (1987) *Nucleic Acids Res.*, **15**, 2611–2626.
14. McClelland,M., Jones,R., Patel,Y. and Nelson,M. (1987) *Nucleic Acids Res.*, **15**, 5985–6005.
15. Kröger,M., Wahl,R. and Rice,P. (1991) *Nucleic Acids Res.*, **19**, Supplement, 2023–2043.
16. Burks,C., Cassidy,M., Cinkosky,M.J., Cumella,K.E., Gilna,P., Hayden,J.E.-D., Keen,G.M., Kelley,T.A., Kelly,M., Kristofferson,D. and Ryals,J. (1991) *Nucleic Acids Res.*, **19**, Supplement, 2221–2225.
17. Dreiseikelmann,B. and Wackernagel,W. (1981) *J. Bacteriol.*, **147**, 259–261.
18. Phillips,G.J., Arnold,J. and Ivarie,R. (1987) *Nucleic Acids Res.*, **15**, 2627–2638.
19. Dzidic,S. and Radman,M. (1989) *Mol. Gen. Genet.*, **217**, 254–256.
20. Reynaud,C.-A., Anquez,V., Grimal,H. and Weill,J.-C. (1987) *Cell*, **48**, 379–388.