Polymorphism of genes encoding SOS polymerases in natural populations of *Escherichia coli*

Ivana Bjedov [a], Guillaume Lecointre [b], Olivier Tenaillon [c], Christelle Vaury [d],
Miroslav Radman [a], François Taddei [a], Erick Denamur [c], Ivan Matic [a,*]

[a] *INSERM U571, Faculté de Médecine Necker-Enfants Malades, Université Paris V, 156 rue Vaugirard, 75730 Paris Cedex 15, France*
[b] *Service de Systématique moléculaire (GDR CNRS 1005), Muséum National d'Histoire Naturelle, Paris, France*
[c] *INSERM U458, Hôpital Robert Debré, Paris, France*
[d] *Centre d'Etude du Polymorphisme Humain (CEPH), Hôpital Saint Louis, Paris, France*

Abstract

High fidelity replicative DNA polymerases can be blocked during DNA replication by various base damages, which represents a potentially lethal event. *Escherichia coli* possesses three DNA polymerases, PolII, PolIV and PolV, that can continue replication over such lesions in template DNA, thus allowing for cell survival. Genes coding for these enzymes, *polB*, *dinB*, and *umuCD* respectively, belong to the stress-inducible SOS regulon. We have analyzed the patterns of nucleotide sequence variability of genes encoding for three SOS polymerases from *E. coli* natural isolates in order to identify the nature of selective forces that determine their evolution. The frequency of inferred inter-strain recombination events, and the frequency of synonymous and non-synonymous base substitutions within these genes do not deviate significantly from those observed for the control group composed of 2 genes coding for DNA polymerases PolI and PolIII and 10 metabolic genes. This suggests that the loci coding for SOS polymerases are subject to selective pressure for the maintenance of their function and specificity. The fact that genes coding for translesion-synthesis (TLS) polymerases, particularly *dinB* and *umuC* homologs, have been conserved during evolution and the present analysis suggest that their activity is essential for the cellular survival and fitness.
© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Translesion-synthesis polymerases; Phylogeny; Nucleotide sequence polymorphism; Recombination; Evolution

1. Introduction

The functional integrity of genetic information is constantly threatened by the intrinsic instability of DNA as well as by damages inflicted by the re- active intermediates of cellular metabolism and by

environmental agents. All living organisms possess numerous DNA repair pathways which have evolved to assure survival and maintain mutation rates as low as possible. Some of DNA repair systems have been remarkably conserved throughout evolution [1]. For example, a new family of DNA polymerases (Y) has been recently identified in prokaryotes, archea and eukaryotes, based on the strong homology with *Escherichia coli* proteins [2]. The common char- acteristic of these enzymes is that they can bypass

* Corresponding author. Tel.: +33-1-40-61-53-25;
fax: +33-1-40-61-53-22.
E-mail address: matic@necker.fr (I. Matic).

DNA lesions that block chain elongation by replicative DNA polymerases. Hence, their designation as translesion-synthesis (TLS) polymerases. These proteins are so different from the five previously identified polymerase families, that their functions have been identified only in recent years, although *E. coli* genes coding for these proteins and some of their phenotypes have been known for decades [3].

E. coli possesses two DNA polymerases belonging to Y-family: PolIV and PolV, each processing a different type of DNA damage [4,5]. Genes coding for these two polymerases, *umuCD* and *dinB*, respectively, belong to the SOS regulon. In addition, another *E. coli* DNA polymerase, PolII (belonging to the B-family of polymerases), has also the capacity to participate in the bypass of some lesions [4–6]. PolII is encoded by *polB* gene, which is also a part of the SOS regulon. *E. coli* SOS system, composed of at least 40 genes of which the majority code for DNA repair functions [7,8], is induced by a wide variety of cellular stresses that damage DNA and/or block DNA replication (for review see [3]). All those stresses have one common characteristic; they all increase intracellular concentration of single-stranded DNA. The contact with single-stranded DNA activates co-protease activity of the RecA protein, which promotes the self-cleavage of the LexA protein, i.e. the SOS repressor, thus inducing the SOS response. When DNA lesions are repaired and replication restored, SOS functions are again repressed.

Therefore, unlike replicative polymerases, these three DNA polymerases are expected to be required only punctually, when bacterial cells are stressed by DNA damage. It is difficult to estimate how often this is happening in nature. It is not known to what extent these enzymes are important in the maintenance of genomic integrity of their hosts because they are also required for mutagenesis. It is not clear why bacteria possess several TLS polymerases which have poor fidelity, while possessing DNA repair pathways that can efficiently remove non-coding lesions from DNA in error-free fashion (e.g. recombination, base and nucleotide excision repair). For example, *Deinococcus radiodurans* is extremely resistant to the lethal and mutagenic effects of many physical and chemical DNA damaging agents inspite of the apparent absence of genes coding for SOS polymerases [9,10]. Finally, it is not known whether SOS polymerases have fixed

functions or whether they are fast evolving enzymes adapting to the wide diversity of DNA lesions that bacteria encounter in nature.

In order to address these questions, we have performed phylogenetic analysis of genes encoding all five *E. coli* polymerases: *polB*, *dinB* and *umuC* genes coding for three SOS polymerases, *polA* coding for PolI polymerase involved in gap-filling steps during DNA replication and nucleotide excision repair [3], and *dnaE* gene coding for alpha subunit of replicative DNA polymerase III [11]. Our data suggest that genes coding for *E. coli* SOS polymerases are under selective constraints similar to those observed for the housekeeping genes. Therefore, it may be concluded that their activity and specificity are fixed, and that they are very important for the survival and fitness of *E. coli* cells in nature.

2. Material and methods

A set of 30 strains representative of the genetic diversity of the *E. coli* reference (ECOR) collection of natural isolates [12] have been studied. Partial *polB* (977 bp), *dnaE* (985 bp), *dinB* (1010 bp), and *umuC* (1230 bp) genes were amplified by the polymerase chain reaction from the DNA of these strains and sequenced without interim cloning as in [19]. Three of these genes have been also sequenced from one *Escherichia fergusonii* strain, the closest species to *E. coli* [13], which served as an outgroup for phylogenetic inference. Sequences of the primers used were as follow: *polB*5′ TCAACGCCTGGTTTGCCAACTA, *polB*5′bis TGGAAAAACTCAACGCCTGGT, *polB*3′ TCCAGTTCTAATGCGCTGGTCA, *polB*3′bis TGTTGGCATCAGAAAACGGC, *dnaE*5′ ATGGAATTTATCCAGTGGTCGA, *dnaE*5′ bis GTGGCCTACGCGCTGAAAAT, *dnaE*3′ CTTTTTCTTACCCATCGCACG, *dinB*5′ TTGAGAGGTGAGCAATGCGTA, *dinB*5′bis TCATA/TCATGTG/TGATATGGACTG, *dinB*3′ GTATACATCATAATCCCAGCAC, *dinB*3′bis TACATCATAAC/TCCC/TAA/GCACCA, and *umuC*5′ CGCCCATTACCATCAGTAGTG, *umuC*3′ CCCTCAGTAAATCAGAACTTCG.

The sequences have been deposited in the Genbank database under accession numbers: AF483912–AF483939 for *polB*; AF483940–AF483969 for *dnaE*; AF483080–AF483106 for *dinB*; and AF483970–

AF483998 for *umuC*. In addition, 30 partial *polA* sequences (341 bp) originating from 30 ECOR strains were acquired from GenBank [14] and included in analysis.

Sequences were aligned using CLUSTAL [15] analysis from the Sequence Navigator[TM] package. Phylogenetic trees were constructed by the parsimony method with the heuristic search of PAUP*4.0 [16] using 100 random addition sequences. Character incongruence was measured between each polymerase gene data set and a large data set representative of the ECOR strain phylogeny [17] using the incongruence length difference (ILD) test [18], as in [19]. In this test, the null hypothesis of congruence is rejected when the *P*-value is below the threshold of 0.05. Iterative removal of single taxon or combination of taxa were performed and the analysis repeated until the incongruence was no longer significant.

Transferred segments among sequenced genes for a given strain were identified by a computer algorithm based on aligned DNA sequences and the previously established phylogeny as in [19]. The relative number of transfers (t') [19] was calculated to take into account the length of the sequenced region and the number of sequenced strains, allowing comparison of the numbers of transfers between genes.

Because a single nucleotide difference observed between aligned sequences can be due to more than one inferred substitution event, the percentage of non-synonymous substitutions was calculated using substitutions inferred onto the most parsimonious tree. The percentage of non-synonymous substitutions is the number of nucleotide changes leading to non-synonymous substitutions in the most parsimonious tree divided by the total number of substitutions onto the most parsimonious tree.

Besides the above-mentioned phylogenetic approaches, we also analyzed sequences with methods developed by population geneticist using the DnaSP software package [20] (http://www.bio.ub.es/~julio/DnaSP32Inf.html). We estimated θ ($4N_e \times \mu$, N_e being the effective population size and μ the mutation rate) with and without recombination and computed 95% confidence intervals by coalescence simulations. We performed Tajima [21] tests for natural selection on studied DNA sequences. To compare the non-synonymous and synonymous substitution rates, for each gene, we calculated the mean of all

pair-wise sequences comparisons K_A and K_S (without the outgroup).

3. Results and discussion

The nature of selective pressures acting on a given bacterial gene can be deduced from the pattern of its DNA sequences polymorphism. For example: (i) The genes whose protein products are under strong and permanent selection for conservation, like the housekeeping genes, have very low non-synonymous polymorphism and very low level of recombination (horizontal transfers). (ii) Genes under diversifying selection, like genes encoding surface antigens and antibiotic resistance targets, have both a high level of non-synonymous polymorphism and very high level of recombination. (iii) Some genes may have high level of recombination but low non-synonymous polymorphism, for example, mismatch repair genes [19]. Mismatch repair genes are under an alternative selection. Depending on the strength of selective pressure, the functionality of the mismatch repair system is selected or counter-selected. The alternation of non-functional and functional mismatch repair is facilitated by the restoration of mismatch repair proficiency through the reacquisition of functional genes by horizontal gene transfer, resulting in mosaic gene structure. (iv) Finally, genes coding for the functions that are subject to weak or infrequent selection may accumulate deleterious or inactivating mutations. For example, reactivated cryptic *cel* operons from *E. coli* natural isolates are temperature sensitive which suggests that mutations have been accumulated while operons were not under selection [22]. Furthermore, it was found that cryptic *bgl* operon is eliminated in 29% of *E. coli* natural isolates as a result of deletions [23]. In that case, the function can be restored only through reacquisition of the functional alleles via horizontal gene transfer. Consequently, such genes may have high level of recombination.

With these premises, we tried to determine whether genes coding for SOS polymerases can be assigned to one of above-mentioned categories, hoping also to get an insight in the nature of selective forces that govern their evolution. We thus sequenced *polB*, *dinB* and *umuC* genes, as well as *dnaE* gene, from a set of 30 natural *E. coli* isolates belonging to the ECOR

collection. The sequences of ECOR strain *polA* genes used in this analysis were acquired from GenBank. The control group of *E. coli* metabolic genes was composed of: *aceK*, *crr*, *icd*, *gapA*, *mdh*, *trpA*, *pabB*, *putP*, *trpB*, and *trpC* [17]. The strains chosen for this study belong to all *E. coli* phylogenetic groups. Two main questions have been addressed using these sequence data: (1) Is there any evidence of recombination in genes for SOS polymerases? (2) What are the relative levels and the nature of DNA sequence polymorphisms of these three genes?

3.1. Detection of recombination events

We looked for the hallmarks of recombination by comparing phylogenies of each gene with the ECOR strain phylogeny based on a whole genome data set [17] as in [19] (Fig. 1). The incongruences between gene and strain phylogenies suggest that horizontal gene transfer has occurred in recent evolutionary past [24,25]. According to the ILD test, significant character incongruences have been observed between the data sets of *polB*, *dnaE* and *dinB* genes and whole genome reference data set, while it was not significant for *umuC* and *polA* genes (Table 1). Since ILD test can discriminate between taxonomic incongruences, which are the consequence of recombination and those due to tree reconstruction artifacts, the observed incongruences are most likely the result of horizontal gene transfer. To estimate the extent of incongruence due to recombination, single strains or combinations of strains were progressively removed, and the analysis repeated, until the incongruence was no longer significant. For *polB*, *dinB* and *dnaE*, 2, 10, and 3 strains, respectively, have to be removed to raise the *P*-value above the significance threshold (Table 1).

3.2. Quantification of recombination events

In order to compare the frequency of inferred recombination events in genes coding for 5 *E. coli* DNA polymerases with that observed in a set of 10 previously studied *E. coli* metabolic genes [17], we quantified recombination events using the previously published method [19]. When the sequence of a gene in a given strain contained more than three de novo mutations compared to its inferred ancestral sequence, and this, in more than three contiguous informative

positions, we searched a homologue of a gene fragment defined by those mutations among the ancestral and the existing DNA sequences belonging to different phylogenetic groups. When such fragments were found, we assumed that a single genetic exchange has occurred, considering that a transfer event is more likely than at least three independent homoplastic mutations (mutations leading to similarity appeared independently rather than inherited from a common ancestor) exhibiting the same contradicting pattern in at least three contiguous sites informative for parsimony, which is a very unlikely scenario. This approach allowed us to confirm that all the incongruences detected by the ILD test in studied genes were results of recombination events (Table 1 and Fig. 2a). Additional recombination events were identified that were not detected by the ILD test, probably due to insufficient lengths of transferred stretches of DNA. For all recombination events detected by this approach, the phylogenetic group of the donor strain was clearly identified.

In order to be able to compare the number of recombination events in genes coding for SOS polymerases with a control group of genes, we have calculated the relative number of transfers, which corrects for the number of strains and different length of sequenced DNA for different genes as previously described [19]. The results clearly indicate that the frequency of recombination events is not significantly different (less than three standard deviations from the mean for all genes) for genes coding for SOS polymerases compared to a control group of genes, composed of *dnaE*, *polA* and 10 metabolic genes (Table 1 and Fig. 2a). For comparison, using the same 10 metabolic genes as a control, we have previously observed that mismatch repair genes exhibit high level of recombination as suggested by standard deviations of relative numbers of transfers ranging from 4.5 (*mutH*) to 11 (*mutS*) [19]. Therefore, it can be concluded that recombination events were not selected in the evolution of genes encoding SOS polymerases.

3.3. DNA sequence divergence

Relatively low frequency of recombination events in studied genes rendered more accurate the estimates of the DNA sequence divergence. In *polB*, *dinB*, *dnaE*, *umuC* and *polA* genes, we have observed that 20.0, 18.3, 11.2, 8.8 and 5.5% of nucleotides were variable,
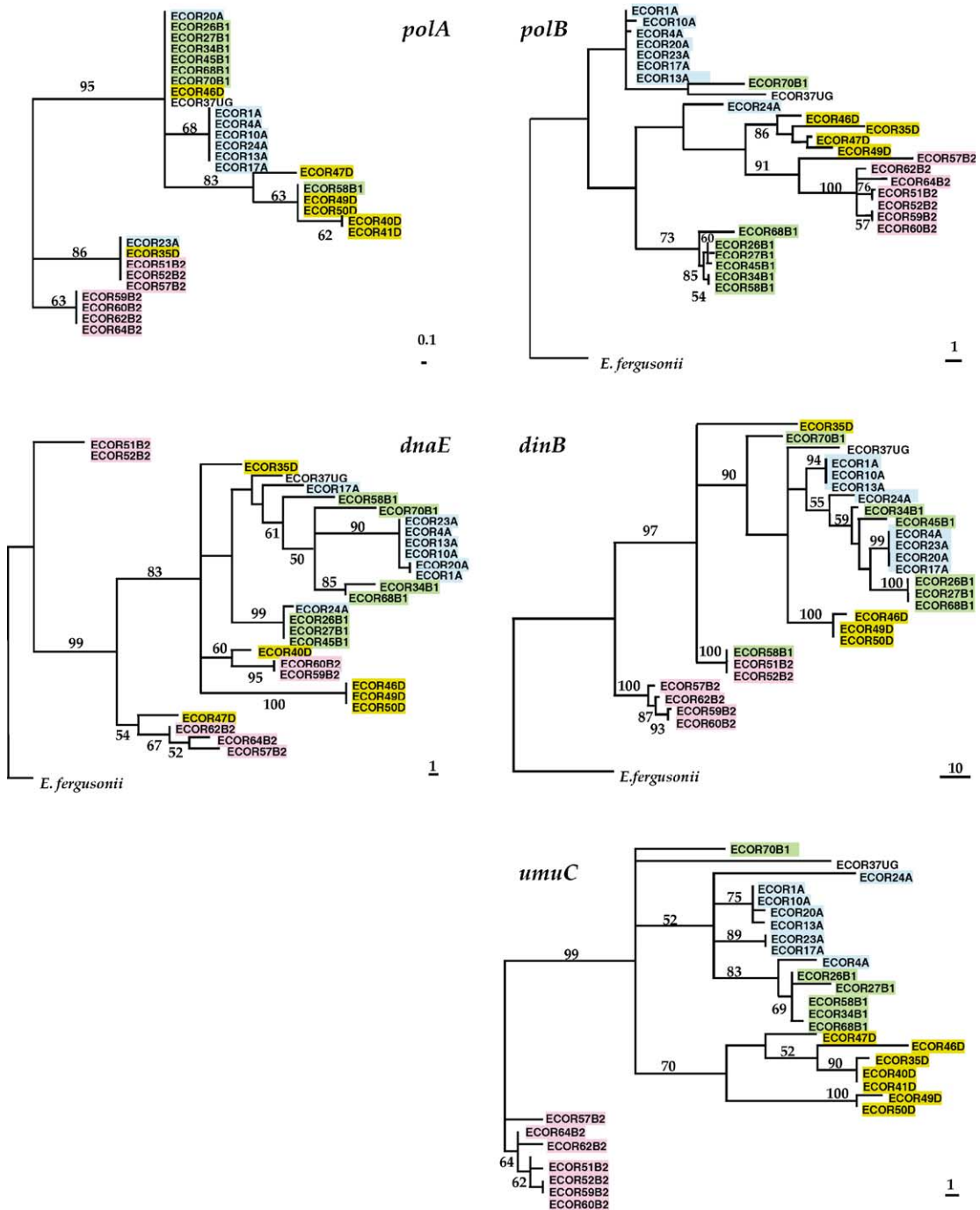
Fig. 1. Strict consensus phylogenetic trees of *polA*, *polB*, *dnaE*, *dinB* and *umuC* genes of 30 *E. coli* strains belonging to A, B1, B2, D and UG (ungrouped) lineages. Phylogenetic trees were constructed by the parsimony method with the heuristic search of PAUP*4.0. Branch lengths are given under ACCTRAN optimization. Numbers above branches correspond to bootstrap proportions obtained from 1000 iterations, indicated when above 50%. Scale bar corresponds to one step. The trees were rooted on *E. fergusonii* (for *polB*, *dnaE* and *dinB*) or on B2 group strains (for *polA* and *umuC*) when *E. fergusonii* sequences were not available. B2 group was used because it was shown to be the most basal within *E. coli* species phylogeny [17]. Colors denote the ECOR phylogenetic strain group as in [37]. The same topology was found from a maximum likelihood approach (data not shown). Tree statistics and homoplasy measurements are given in Table 1.

Table 1
Characteristics of the phylogenetic data of the *E. coli* polymerase genes

Gene	Length of the gene	Length of the analyzed sequence	Number of taxa	Outgroup	Number of variable sites without outgroup	Number of informative sites for parsimony without outgroup	Non-synonymous base substitution (%)[a]	Number of most parsimonious trees	Tree length	Consistency index	Retention index	P-value of the ILD test	Taxa responsible for significant incongruence	P-value without identified taxa	Number of transfers
polA	2787	341	30 *E. coli*	*E. coli* B2 strains	12	11	8.3	1	12	1	1	0.26	None	–	4
polB	2352	977	27 *E. coli*, 1 *E. fergusonii*	*E. fergusonii*	196	82	9.7	27	223	0.63	0.83	0.01	24,70	0.33	11
dnaE	3483	985	29 *E. coli*, 1 *E. fergusonii*	*E. fergusonii*	185	45	5.5	48	115	0.71	0.86	0.01	24,17,47	0.15	13
dinB	1056	1010	26 *E. coli*, 1 *E. fergusonii*	*E. fergusonii*	110	85	8.6	12	247	0.65	0.83	0.01	52,51,58,70, 4,23,20,17, 24,45	0.14	17
umuC	1269	1230	29 *E. coli*	*E. coli* B2 strains	109	48	23	28	115	0.73	0.86	0.06	None	–	3

[a] Non-synonymous base substitutions were inferred onto the most parsimonious tree.

respectively (Table 1 and Fig. 2b). The comparison with the metabolic genes showed that *polB* and *dinB* genes have the highest level of DNA polymorphism, *dnaE* and *umuC* genes are close to the average observed for the all genes, whereas the value for *polA* gene is very low (Fig. 2b). Nevertheless, the polymorphism of *polB* and *dinB* genes is not significantly different from control genes since they are 2.0 and 1.7 standard deviations from the average of all genes. Using a population genetics approach, we computed θ (which gives an estimation of the mutation rate) for all the genes and found similar results (data no shown).

Among all substitutions, the fraction of non-synonymous substitutions inferred onto the most parsimonious tree is 23, 9.7, 8.6, 8.3 and 5.5% for *umuC*, *polB*, *dinB*, *polA* and *dnaE* genes, respectively. These values are not significantly different from those of a set of 10 metabolic genes (Table 1 and Fig. 2c). We also used population genetic approaches, which take into account gene codon composition, in order to study non-synonymous (K_A) and synonymous (K_S) mutation rates. Neither K_A nor K_S nor the ratio K_A/K_S of SOS polymerase genes were significantly different from control group (data not shown). Furthermore, the results of Tajima test confirmed that the patterns of *polA*, *polB*, *dinB*, *dnaE* and *umuC* genes nucleotide variations do not deviate significantly from the patterns predicted by neutral theory (data not shown).

However, when only genes coding for polymerases were taken into account, *umuC* presented significantly higher K_A and K_A/K_S values. This suggest that UmuC protein is subject to lower selective constraint on amino-acid sequence than the other polymerases such as PolA for which we found a high degree of selective constraints as observed before [26]. This diversity in UmuC does not reflect diversifying selection because the observed nucleotide polymorphism is linked to the phylogeny with a diversification occurring between the phylogenetic groups A, B1, D and B2 (see Fig. 1 and ILD test in Table 1). A possible cause of this phenomenon could be that, since UmuC acts with multiple protein partners (UmuD and RecA; former being encoded by the same operon), all partners have co-evolved in each phylum in order to conserve specific protein–protein interactions [27].
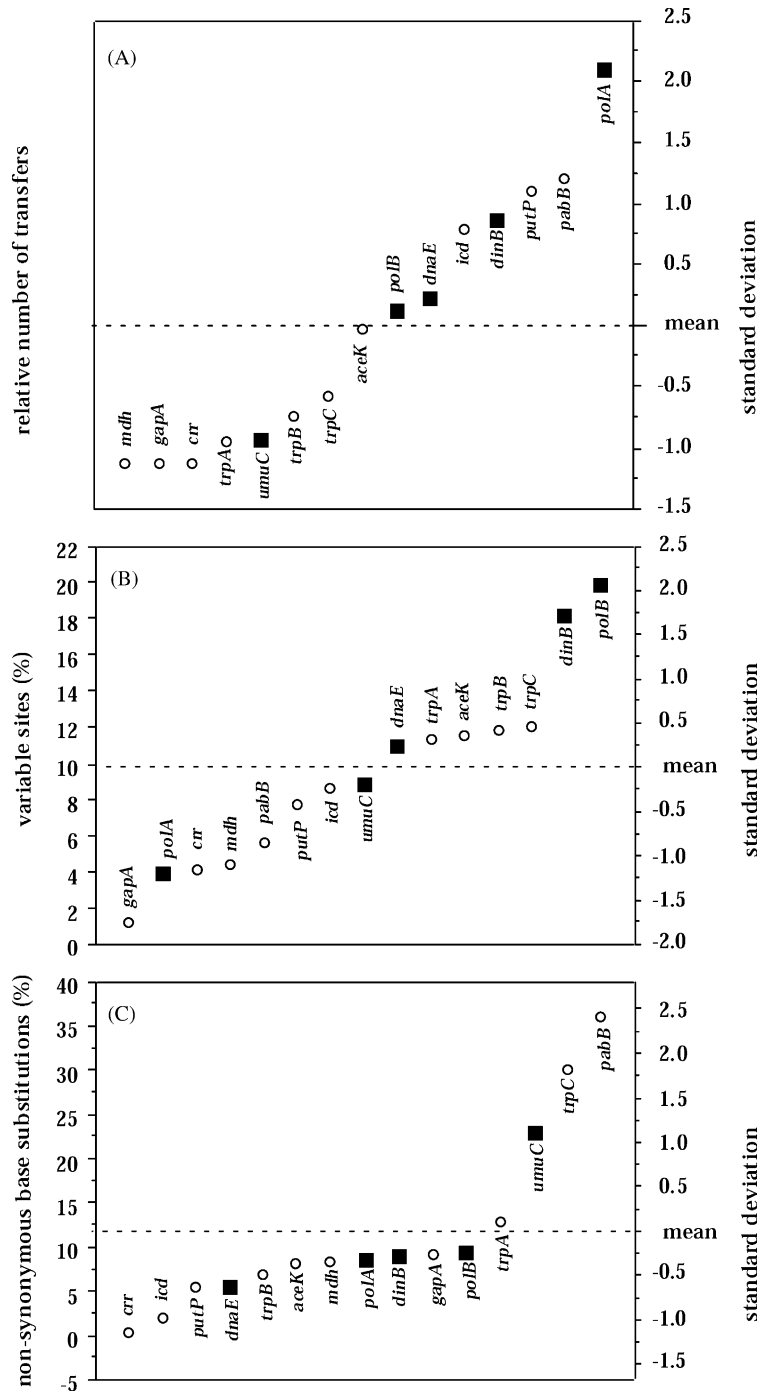
Fig. 2. Patterns of DNA sequence polymorphism of *polA*, *polB*, *dnaE*, *dinB* and *umuC* genes. (A) Relative number of recombination events is expressed in terms of σ, the standard deviation from the mean calculated using following values: *mdh*, 0; *gapA*, 0; *crr*, 0; *trpA*, 4.9×10^{-9}; *umuC*, 5.1×10^{-9}; *trpB*, 9.0×10^{-9}; *trpC*, 1.4×10^{-8}; *aceK*, 2.7×10^{-8}; *polB*, 3.2×10^{-8}; *dnaE*, 3.4×10^{-8}; *icd*, 4.9×10^{-8}; *dinB*, 4.9×10^{-8}; *putP*, 5.7×10^{-8}; *papB*, 5.9×10^{-8}; *pabB*, and *polA*, 8.3×10^{-8}. (B) Variable nucleotide sites has been inferred on the most parsimonious tree. (C) Number of non-synonymous nucleotide substitutions inferred on the most parsimonious tree.

3.4. Codon usage

Genes coding for TLS polymerases belong to the SOS regulon which is induced only in stressed bacterial cells (for review see [3]). When the SOS system is repressed in non-stressed *E. coli* cells, genes coding for SOS polymerases are expressed at very low level or not at all. How often SOS system is induced in bacterial cells and what is the level of SOS gene expression in natural environments is unknown. In order to learn more about selective pressures acting on genes coding for SOS polymerases, we compared codon adaptation index (CAI) of these three genes with all other SOS genes [7,8], and with all *E. coli* ORFs using database available for the sequenced genome of the *E. coli* MG1655 strain [28] (ftp://ftp.pitt.edu/dept/biology/lawrence/). Bacterial codon usage relates to the expression level and reflects the degree of selective constraint on the gene. The comparison indicated that CAIs of genes coding for TLS polymerases do not deviate significantly from those of other SOS genes nor from those of average *E. coli* genes. Therefore, it can be concluded that genes coding for SOS polymerases do not belong to the category of highly expressed genes. Of all SOS genes, only *recA* gene belongs to this category [29]. It is also interesting to note that, according to base composition and CAI pattern, none of studied genes have been acquired in *E. coli* MG1655 genome through horizontal gene transfer from a diverged bacterial species [28].

4. Conclusions

In this study, we have analyzed DNA sequences of *E. coli* genes coding for TLS polymerases, *polB*, *dinB*, and *umuC*, obtained from natural isolates with intention to try to identify the evolutionary forces involved in their evolution. The phylogenetic analysis of these genes indicates that there is low level of recombination among genes from strains belonging to different phylogenetic lineages. The frequency of inferred recombination events observed for the studied genes is not significantly different from that of a control group composed of *polA*, *dnaE* and 10 metabolic genes Figs. 1 and 2b; Table 1). The patterns of DNA sequence diversity established for *polB*, *dinB*, and *umuC* genes, i.e. percentage of synonymous and non-synonymous substitutions, are not different from that observed for the control group of genes (Fig. 2a and b; Table 1). Thus, our data suggest that the loci coding for TLS polymerases in *E. coli* are subject to functional constraints similar to those observed for housekeeping genes, i.e. they seems to be essential for the survival and fitness of bacterial cells. Furthermore, it may be concluded that PolII, PolIV and PolV are polymerases with fixed specificities rather than fast evolving enzymes in search of new specificities.

The ubiquity and evolutionary conservation of TLS polymerases also support their biological relevance [2]. The homologs of genes coding for *E. coli* TLS polymerases have been identified in chromosomes and in naturally occurring plasmids in different bacterial species. It has been reported that the presence of these genes, for example *impCAB* (carried by *Shigella flexneri* plasmid) and *mucAB* operons (carried by R plasmids) enhance survival of bacterial cells exposed to UV irradiation [30,31]. Since some of those plasmids are virulence plasmids, it may be that presence of genes coding for TLS polymerases contributes to the robustness of pathogenic strains. Homologs of genes coding for *E. coli* TLS polymerases have been found also in archea and eukaryotes [2]. The inactivation of mouse Rev3l gene (encoding the catalytic subunit of the TLS polymerase ζ) causes embryonic lethality [32]. Humans have at least seven members of this superfamily of proteins. Individuals carrying inactivated gene coding for polymerase η suffer from the genetic disorder, xeroderma pigmentosum variant, characterized by high cellular sensitivity to UV light and hyper-mutability [33,34]. In the view of the diversity of specific phenotypes of TLS polymerases [35], it can be concluded that their activity is extremely important for the cellular survival and fitness. Indeed, it has been recently reported that *E. coli* mutants lacking one or more SOS polymerase suffer an important fitness reduction when grown in competition with wild-type strain under starvation conditions [36].

References

[1] J.A. Eisen, P.C. Hanawalt, A phylogenomic study of DNA repair genes, proteins, and processes, Mutat. Res. 435 (1999) 171–213.

[2] H. Ohmori, E.C. Friedberg, R.P. Fuchs, M.F. Goodman, F. Hanaoka, D. Hinkle, T.A. Kunkel, C.W. Lawrence, Z. Livneh, T. Nohmi, L. Prakash, S. Prakash, T. Todo, G.C. Walker, Z. Wang, R. Woodgate, The Y-family of DNA polymerases, Mol. Cell 8 (2001) 7–8.

[3] E.C. Friedberg, G.C. Walker, W. Siede, DNA Repair and Mutagenesis, ASM Press, Washington, DC, 1995.

[4] R. Napolitano, R. Janel-Bintz, J. Wagner, R.P. Fuchs, All three SOS-inducible DNA polymerases (Pol II, Pol IV and Pol V) are involved in induced mutagenesis, EMBO J. 19 (2000) 6259–6265.

[5] X. Shen, J.M. Sayer, H. Kroth, I. Ponten, M. O'Donnell, R. Woodgate, D.M. Jerina, M.F. Goodman, Efficiency and accuracy of SOS-induced DNA polymerases replicating benzo[a]pyrene-7,8-diol 9,10-epoxide A and G adducts, J. Biol. Chem. 277 (2002) 5265–5274.

[6] C.A. Bonner, S.K. Randall, C. Rayssiguier, M. Radman, R. Eritja, B.E. Kaplan, K. McEntee, M. Goodman, Purification and characterization of an inducible *Escherichia coli* DNA polymerase capable of insertion and bypass at abasic lesions in DNA, J. Biol. Chem. 263 (1988) 18946–18952.

[7] J. Courcelle, A. Khodursky, B. Peter, P.O. Brown, P.C. Hanawalt, Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient *Escherichia coli*, Genetics 158 (2001) 41–64.

[8] A.R. Fernandez-De-Henestrosa, T. Ogi, S. Aoyagi, D. Chafin, J.J. Hayes, H. Ohmori, R. Woodgate, Identification of additional genes belonging to the LexA regulon in *Escherichia coli*, Mol. Microbiol. 35 (2000) 1560–1572.

[9] K.S. Makarova, L. Aravind, Y.I. Wolf, R.L. Tatusov, K.W. Minton, E.V. Koonin, M.J. Daly, Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics, Microbiol. Mol. Biol. Rev. 65 (2001) 44–79.

[10] O. White, J.A. Eisen, J.F. Heidelberg, E.K. Hickey, J.D. Peterson, R.J. Dodson, D.H. Haft, M.L. Gwinn, W.C. Nelson, D.L. Richardson, K.S. Moffat, H. Qin, L. Jiang, W. Pamphile, M. Crosby, M. Shen, J.J. Vamathevan, P. Lam, L. McDonald, T. Utterback, C. Zalewski, K.S. Makarova, L. Aravind, M.J. Daly, K.W. Minton, R.D. Fleischmann, K.A. Ketchum, K.E. Nelson, S. Salzberg, H.O. Smith, J.C. Venter, C.M. Fraser, Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1, Science 286 (1999) 1571–1577.

[11] K.J. Marians, Replication fork propagation, in: F.C. Neidhardt, R. Curtiss III, J.L. Ingraham, E.C.C. Lin, K.B. Low, B. Magasanik, W.S. Reznikoff, M. Riley, M. Schaechter, H.E. Umbarger (Eds.), *Escherichia coli* and *Salmonella* Cellular and Molecular Biology, ASM Publications, Washington, DC, 1996, pp. 749–763.

[12] H. Ochman, R.K. Selander, Standard reference strains of *Escherichia coli* from natural population, J. Bacteriol. 157 (1984) 690–693.

[13] J.G. Lawrence, H. Ochman, D.L. Hartl, Molecular and evolutionary relationships among enteric bacteria, J. Gen. Microbiol. 137 (1991) 1911–1921.

[14] E.W. Brown, J.E. LeClerc, M.L. Kotewicz, T.A. Cebula, Three R's of bacterial evolution: how replication, repair, and recombination frame the origin of species, Environ. Mol. Mutagen. 38 (2001) 248–260.

[15] D.G. Higgins, A.J. Bleasby, R. Fuchs, CLUSTAL V: improved software for multiple sequence alignment, CABIOS 8 (1992) 189–191.

[16] Swofford, D.L., PAUP* Phylogenetic Analysis Using Parsimony (And Other Methods), Sinauer Associates, Sunderland, MA, 1999.

[17] G. Lecointre, L. Rachdi, P. Darlu, E. Denamur, *Escherichia coli* molecular phylogeny using the incongruence length difference test, Mol. Biol. Evol. 15 (1998) 1685–1695.

[18] J.S. Farris, M. Källersjö, A.G. Kluge, C. Bult, Testing significance of incongruence, Cladistics 10 (1995) 315–319.

[19] E. Denamur, G. Lecointre, P. Darlu, O. Tenaillon, C. Acquaviva, C. Sayada, I. Sunjevaric, R. Rothstein, J. Elion, F. Taddei, M. Radman, I. Matic, Evolutionary implications of the frequent horizontal transfer of mismatch repair genes, Cell 103 (2000) 711–721.

[20] J. Rozas, R. Rozas, DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis, Bioinformatics 15 (1999) 174–175.

[21] F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism, Genetics 123 (1989) 585–595.

[22] B.G. Hall, W. Faunce III, Functional genes for cellobiose utilization in natural isolates of *Escherichia coli*, J. Bacteriol. 169 (1987) 2713–2717.

[23] B.G. Hall, Widespread distribution of deletions of the *bgl* operon in natural isolates of *Escherichia coli*, Mol. Biol. Evol. 5 (1988) 456–467.

[24] D.E. Dykhuizen, L. Green, Recombination in *Escherichia coli* and the definition of biological species, J. Bacteriol. 173 (1991) 7257–7268.

[25] J.J. Bull, J.P. Huelsenbeck, C.W. Cunningham, D.L. Swofford, P.J. Waddel, Partitioning and combining data in phylogenetic analysis, Syst. Biol. 42 (1993) 384–397.

[26] P.H. Patel, L.A. Loeb, DNA polymerase active site is highly mutable: evolutionary consequences, Proc. Natl. Acad. Sci. U.S.A. 97 (2000) 5095–5100.

[27] M.D. Sutton, B.T. Smith, V.G. Godoy, G.C. Walker, The SOS response: recent insights into *umuDC*-dependent mutagenesis and DNA damage tolerance, Annu. Rev. Genet. 34 (2000) 479–497.

[28] J.G. Lawrence, H. Ochman, Molecular archaeology of the *Escherichia coli* genome, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 9413–9417.

[29] S. Karlin, J. Mrazek, A. Campbell, D. Kaiser, Charac-
     terizations of highly expressed genes of four fast-growing
     bacteria, J. Bacteriol. 183 (2001) 5025–5040.

[30] L.J. Runyen-Janecky, M. Hong, S.M. Payne, The virulence
     plasmid-encoded *impCAB* operon enhances survival and
     induced mutagenesis in *Shigella flexneri* after exposure to
     UV radiation, Infect. Immun. 67 (1999) 1415–1423.

[31] S.G. Sedgwick, P.A. Goodwin, Differences in mutagenic and
     recombinational DNA repair in enterobacteria, Proc. Natl.
     Acad. Sci. U.S.A. 82 (1985) 4172–4176.

[32] J. Wittschieben, M.K. Shivji, E. Lalani, M.A. Jacobs, F.
     Marini, P.J. Gearhart, I. Rosewell, G. Stamp, R.D. Wood,
     Disruption of the developmentally regulated Rev3l gene
     causes embryonic lethality, Curr. Biol. 10 (2000) 1217–1220.

[33] R.E. Johnson, C.M. Kondratick, S. Prakash, L. Prakash,
     hRAD30 mutations in the variant form of xeroderma
     pigmentosum, Science 285 (1999) 263–265.

[34] C. Masutani, R. Kusumoto, A. Yamada, N. Dohmae, M.
     Yokoi, M. Yuasa, M. Araki, S. Iwai, K. Takio, F. Hanaoka,
     The XPV (xeroderma pigmentosum variant) gene encodes
     human DNA polymerase eta, Nature 399 (1999) 700–
     704.

[35] E.C. Friedberg, R. Wagner, M. Radman, Specialized DNA
     polymerases, cellular survival, and the genesis of mutations,
     Science 296 (2002) 1627–1630.

[36] B. Yeiser, E.D. Pepper, M.F. Goodman, S.E. Finkel, SOS-
     induced DNA polymerases enhance long-term survival and
     evolutionary fitness, Proc. Natl. Acad. Sci. U.S.A. 99 (2002)
     8737–8741.

[37] P.J. Herzer, S. Inouye, M. Inouye, T.S. Whittman,
     Phylogenetic distribution of branched RNA-linked multicopy
     single-stranded DNA among natural isolates of *Escherichia
     coli*, J. Bacteriol. 172 (1990) 6175–6181.