Comparative analysis of tandem repeats in the 44 outer membrane proteins of non-pathogenic (K12) and two pathogenic (O157) strains of *E. coli*

Sharmila S. Mande* and V. V. Raja Rao

Life Sciences Division, Advanced Technology Centre,
Tata Consultancy Services Limited, Deccan Park, 1 Software Units
Layout, Madhapur, Hyderabad 500 081, India

Repetitive sequences are common constituents of genomes of all living organisms contributing to strain characteristics, gene polymorphism and overall evolutionary dynamics of organisms. The present study deals with a comparative analysis of tandem repeats in the 44 outer membrane proteins of two pathogenic (*Escherichia coli* O157) and one non-pathogenic (*E. coli* K12) strains of *E. coli*. This study specifically addresses the role of the repeat sequences in strain characterization and implications in amino acid mutations leading to structural changes in the corresponding proteins. The results indicate a strong preference of these genes for short tandem repeats, with instances of longer repeats being unique to the pathogenic strain (*E. coli* O157). Tandem repeats in genes accounted for more than half of the mutations in the corresponding protein sequences.

Keywords: Cell envelope protein, mutation rate, outer membrane proteins, outer membrane genes, tandem repeats.

IN recent years, various types of repeated DNA sequences were discovered in genes, inter-genic sequences and transposable elements in many prokaryotes[1–3]. Two major classes of repeats have been identified, namely tandemly repeated sequences and interspersed repeats. A tandem repeat is made up of monomeric sequences (whose lengths can vary between one and several hundred nucleotides) repeated periodically, with the contiguous monomers arranged in a head-to-tail configuration. These low complexity repeats are abundantly distributed throughout the genomes of eukaryotes and have been widely studied[4]. Although less abundant in bacterial genomes, their origin and function have been studied[5,6]. Earlier studies on tandem repeats in bacteria focused on a given type of repeat in an organism.

Tandem repeats find application in characterizing various strains/serotypes of virus/bacteria as demonstrated in the studies on human cytomegalovirus, *Mycobacteria* and *Escherichia coli*[7–10]. Genes containing tandem repeats exhibit high mutation rates, allowing the bacterium to respond rapidly to challenging environmental conditions[11,12]. Tandem repeats are involved in phenotypic variation of surface expressed molecules of several major bacterial pathogens. Tandem repeats identified within the 5′-end of the translated reading frames of genes required for lipopolysaccharide (LPS) biosynthesis in *Haemophilus influenzae*, *Neisseria meningitidis*, *N. gonorrhoeae* and other pathogens are subject to loss or gain of one or more of the repeats, resulting in high frequency of phase variation of oligosaccharide core structures leading to variant antigens to evade host immune response[13–19]. Various applications of tandem repeats in microbial pathogenesis, evolution and molecular epidemiology have been reviewed[6]. Nevertheless, despite such discoveries, understanding of the biology of tandem repeats is far from complete in the case of bacterial genomes and there are only a few studies that have dealt with the comparison of tandem repeats.

With the advent and increased use of computational tools in deciphering the bacterial genomes, it has become possible to perform exhaustive detection of tandem repeats at the scale of complete genomes. The objective of the present study is to carry out a systematic investigation on the variability of tandem repeats in 44 outer membrane genes of the three strains of *E. coli*, one non-pathogenic strain (*E. coli* K12) and two pathogenic strains (*E. coli* O157), to find out their role in strain differentiation and study the conformational changes in protein structure due to mutations caused by tandem repeats.

One strain of non-pathogenic (*E. coli* K12) and the two strains of pathogenic *E. coli* (*E. coli* O157) were considered for the present study. Forty-four outer membrane genes common to both strains were selected from the *E. coli* cell envelope protein data collection (http://www.cf.ac.uk/bios/staff/ehrmann/tools/ecce/ecce.htm). The corresponding gene and protein sequences were retrieved from the NCBI database (http://www.ncbi.nlm.nih.gov).

The tandem repeats were identified using a program developed by us with the algorithm of Landau and Schmidt[20]. The tandem repeat search criteria was restricted to repeat length of 2 to 50 and a period (number of repeats) of 2 to 30 with zero mismatches. The gene sequences taken from the NCBI database file were used as inputs for the tandem repeats-finding program. The program gives the location of each of the tandem repeat identified and the sequence of the repeat. Comparison of the tandem repeats was done using the global alignment algorithm[21] adapted to our problem. The method for alignment of tandem repeats is described below and can be thought of as a generalized basic global alignment method. The basic units of comparison were taken as tandem repeats rather than nucleic acid symbols.

Let $\wp_1 = \{\tau_1, \tau_2, \ldots ,\tau_m\}$ be the set of all tandem repeats in the gene g of *E. coli* K12 strain with tandem repeat lengths $|\tau_1|, |\tau_2|, \ldots |\tau_m|$ (length of repeating unit times number of repeats). Similarly, let $\wp_2 = \{\eta_1, \eta_2, \ldots ,\eta_m\}$ be the set of all tandem repeats in the corresponding gene g of *E. coli* O157 strain. Then the two strings that were aligned were $\tau_1\tau_2 \ldots \tau_m$ and $\eta_1\eta_2 \ldots \eta_n$. Considering the indices of the tandem repeats in each of the strings to be 1, 2, … m and 1,

*For correspondence. (e-mail: sharmila@atc.tcs.co.in)

2, …, n respectively, the edit distance matrix $D(i, j)$ for $1 \le i \le m$, $1 \le j \le n$ could be defined as

$$D(0, 0) = 0,$$

$$D(i, 0) = D(i-1, 0) + |\tau_i|,$$

$$D(0, j) = D(0, j-1) + |\eta_j|,$$

$$D(i, j) = \min\{D(i-1, j) + |\tau_i|),$$

$$D(i, j-1) + |\eta_j|,$$

$$D(i-1, j-1 + p(i, j)\}.$$

Here $p(i, j)$ was defined to have value $|\tau_i| - |\eta_j|$ if repeating unit (τ_i) = repeating unit (η_j), and $p(i, j)$ had value $|\tau_i| + |\eta_j|$ otherwise.

Having defined the distance function, the standard global alignment algorithm was used to obtain an alignment of the tandem repeats. This procedure was repeated for all pairs of genes between *E. coli* K12 and *E. coli* O157.

The corresponding gene sequences of the three strains were aligned using ClustalW[22] with default parameters. The corresponding protein sequences were also aligned. The unique tandem repeats in the genes of *E. coli* K12 and two strains of *E. coli* O157 were analysed to check whether they led to any amino acid variations in their respective proteins. The overall mutation rate for each amino acid (ratio of the number of times an amino acid is mutated to the total number of the same amino acid across all genes), the mutation rate for each amino acid due to tandem repeats and that for each amino acid not due to tandem repeats in their corresponding genes were calculated.

The three available 3D structures out of 44 proteins studied were analysed. All these structures correspond to proteins of *E. coli* K12 strain. For each of these available structures, the native amino acids of *E. coli* K12 were replaced with the new amino acids in *E. coli* O157. Energy minimization was performed on the new structure and a superimposition of the native (*E. coli* K12) and mutated (*E. coli* O157) structures was done.

Comparison of alignment of tandem repeats and alignment of gene sequences using ClustalW[22] revealed that tandem repeat alignment algorithm finds some of the matches missed by the ClustalW alignment. For example, the tandemly repeated sequence *catcat* occurs in the *ampC* gene at positions 891 and 894 in *E. coli* O157 and *E. coli* K12 respectively. This match was picked up by tandem repeats alignment algorithm, but was missed by the sequence alignment.

A total of 5079 tandem repeats were identified across 44 outer membrane genes studied in the two strains of *E. coli* O157 and one strain of *E. coli* K12 (Table 1). The tandem repeats of type (2,2) (dinucleotide repeating twice) constituted the major component of the total tandem repeats accounting for 64.0% (3247/5079), followed by the repeats of type (3,2) (trinucleotide repeating twice), which contributed 25.7% (1304/5079) of the total identified repeats. Both together contributed to about 90% of the total repeats and the remaining 10% of the contribution was accounted by other repeats of type (2,3), (2,4), (3,3), (4,2), (5,2) and (6,2), contributing 3.3, 0.1, 0.6, 4.1, 1.3 and 0.9% respectively. The tandem repeats of type (7,2), (8,2) and (12,2) had only one occurrence each, while other repeat lengths were totally absent in all the genes studied.

Data for the occurrence of tandem repeats and their uniqueness to each strain are summarized in Table 1. Out of a total of 3247 repeats of type (2,2), 87% was common in location and occurrence in the three strains of *E. coli*. While 6.7% repeats was unique in *E. coli* K12 strain, 5.9% of repeats in both the strains of *E. coli* O157 had no corresponding occurrence in *E. coli* K12. In the repeats of type (3,2), 79.9% of a total of 1304 was common across all the three strains. In contrast to the repeats of type (2,2), the number of unique repeats of (3,2) was more in *E. coli* O157 strains (10.7%) compared to *E. coli* K12

Table 1. Type of tandem repeats (TRs) and their occurrence across 44 outer membrane genes in the *E. coli* K12 and two strains of *E. coli O157*

TR Type (n,m)*	TRs common in *E. coli* K12 and *E. coli* O157**	TRs unique in *E. coli* K12	TRs unique in *E. coli* O157**	Total TRs
(2,2)	2838	217	192	3247
(2,3)	143	10	15	168
(2,4)	5	0	1	6
(3,2)	1043	121	140	1304
(3,3)	24	1	5	30
(4,2)	165	21	24	210
(5,2)	47	2	17	66
(6,2)	33	9	2	44
(7,2)	0	0	1	1
(8,2)	0	0	1	1
(12,2)	0	0	1	1

**n* is the length of the repeat and *m* is the number of occurrences.

**Common in both strains *E. coli* 0157.

Table 2. Changes in gene sequences due to tandem repeats resulting in mutations in the corresponding protein sequences

TR unique in *E. coli* 0157			TR unique in *E. coli* K12		
Location	Repeat	Change in protein sequence	Location	Repeat	Change in protein sequence
ampC					
	$(AT)_2$	254L→M		$(CA)_2$	105T→A
758		255K→N	312		
	$(CTG)_2$	248R→C		$(AA)_2$	254L→M
741			762		255 K→N
892	$(ATC)_2$	298S→I	781	$(GA)_2$	261E→D
1098	$(CGC)_3$	367D→A			
btuB					
324	$(GG)_3$	110V→G	631	$(CAGA)_2$	212T→P
	$(ACA)_2$	324V→I	986	$(GTA)_2$	330S→N
965					331I→V
1019	$(AG)_2$	341T→S	1024	$(ACG)_2$	343T→A
yciD					
	$(TAC)_2$	190G→V		$(CA)_2$	191A→T
		192Q→T			192Q→T
570		191A→T	572		
579	$(AA)_2$	194H→K	579	$(AC)_2$	194H→K
nlpA					
24	$(CGGG)_2$	10T→A	295	$(GC)_2$	99A→T
232	$(AAT)_2$	79H→N	622	$(CAG)_2$	208Q→E
			692	$(TT)_2$	216H→N
yehB					
59	$(CA)_2$	21Y→H	114	$(CAG)_2$	39Q→E
297	$(GG)_2$	100S→R	127	$(AT)_2$	44I→L
301	$(AA)_2$	101G→E	252	$(AG)_2$	84S→T
412	$(AA)_2$	139S→N	360	$(GG)_2$	122V→I
651	$(TT)_2$	219A→S	382	$(GT)_2$	128S→A
879	$(TT)_2$	295A→S	612	$(GG)_2$	206V→E
1281	$(TT)_2$	429V→L	675	$(CG)_2$	227V→I
1698	$(ACG)_2$	567S→T	881	$(CG)_2$	295A→S
2019	$(ATGC)_2$	675N→H	1166	$(GGCT)_2$	390A→T
2089	$(TGG)_2$	699I→V	1283	$(TG)_2$	429V→L
2133	$(ATG)_2$	713L→M	1394	$(TGA)_2$	467E→K
2461	$(AA)_2$	821R→K	1633	$(ATC)_2$	547H→A
			1816	$(GT)_2$	607V→I
			2017	$(TGA)_2$	675N→H
			2094	$(AT)_2$	699I→V
			2190	$(GCC)_2$	731A→T
fhuA					
646	$(GC)_2$	216S→A	–	–	–
fhuE					
2147	$(GC)_2$	717T→A	945	$(GCA)_2$	317Q→R
–	–	–	2149	$(AC)_2$	717T→A
fimD					
88	$(CT)_2$	31V→F	113	$(CA)_2$	38A→V
1618	$(TCA)_2$	540T→S	–	–	–
hofQ					
362	$(AT)_2$	121S→N	–	–	–
mltB					
–	–	–	196	$(AA)_2$	66K→R
mltE					
–	–	–	648	$(AC)_2$	216E→D

(*contd...*)

Table 2. (*contd...*)

TR unique in *E. coli* 0157		Change in	TR unique in *E. coli* K12		Change in
Location	Repeat	protein sequence	Location	Repeat	protein sequence
nlpB					
–	–	–	20	(AA)$_2$	7Q→H
sfmD					
1311	(CGA)$_2$	439A→D	–	–	–
slp					
–	–	–	252	(CT)$_2$	85S→A
vacJ					
–	–	–	505	(TT)$_2$	169F→L
ybhC					
1121	(GG)$_2$	375A→G	1181	(ACG)$_2$	394N→S
yeaF					
693	(TG)$_2$	231M→I	692	(TGG)$_2$	231M→I
fepA					
205	(GA)$_2$	69K→E	934	(GC)$_2$	312A→S
929	(ACT)$_2$	312A→S	1082	(AT)$_2$	362I→N
			1258	(ACC)$_2$	420T→A
			1632	(GGCG)$_2$	547V→I
ygiG					
–	–	–	137	(ATG)$_2$	47:D→N
yiaD					
	(GAAGC)$_2$	87N-K	–	–	–
253		88IV			
yicP					
681	(TGC)$_2$	228T→A	233	(GCG)$_2$	78R→H
1299	(GC)$_2$	433D→E	1125	(GA)$_2$	376R→K
1309	(GT)$_2$	438S→C	1303	(GAT)$_2$	435D→N
ylcB					
214	(GA)$_2$	72:V→E	–	–	–
yraJ					
1446	(TGA)$_2$	483N→D	803	(AC)$_2$	269T→N
1582	(GA)$_2$	528D→E	949	(GT)$_2$	318S→P
yejO					
136	(AGT)$_2$	47N→S	133	(GTAA)$_2$	45V→I
	(AAGC)$_2$	148T→A	138	(TAA)$_2$	47N→S
440		149T→S			
852	(GA)$_2$	285K→R	186	(CGC)$_2$	63A→T
	(GATCA)$_2$	298Y→D		(GG)$_3$	115G→I
887			340		116A→T
	(GC)$_2$	385V→A		(ATTTGC)$_2$	134F→L
1151			399		136F→I
	(GC)$_2$	410D→G		(AA)2	148T→A
1229			439		149T→S
			494	(AT)$_2$	165D→G
			891	(AT)$_2$	298Y→D
			1095	(TT)$_2$	366F→L
			1197	(CCT)$_2$	400L→R
			1913	(AT)$_2$	639I→V
ypaJ					
586	(AC)$_2$	196I→T	223	(ACT)$_2$	76T→P
883	(ATT)$_2$	296V→I	224	(CTA)$_2$	77N→D
900	(CTC)$_2$	301T→S	272	(AA)$_2$	92K→R

(*contd...*)

Table 2. (*contd...*)

TR unique in *E. coli* 0157			TR unique in *E. coli* K12		
Location	Repeat	Change in protein sequence	Location	Repeat	Change in protein sequence
947	(GTG)$_2$	316R→S	387	(TA)$_2$	130I→V
2670	(AAC)$_2$	892K→T	405	(TT)$_2$	136F→Y
3577	(TT)$_2$	1194T→S	884	(TTG)$_2$	296V→I
	(ATCCCAAC	1232I→N	947	(GG)$_2$	316R→S
3689	CCAA)$_2$	1236K→N			
4044	(GG)$_2$	1349T→A	3185	(CT)$_2$	1063L→I
			4201	(AA)$_2$	1401K→E
			4386	(GAAATC)$_2$	1463K→E

(9.3%). The tandem repeats that were common in the three strains clearly outnumbered those unique in one of the three strains. Among all the 44 genes studied, 41 genes had unique tandem repeats in one of the three strains. The three genes *ycdS*, *lpp* and *pal* were however significant in that they did not possess any unique tandem repeats. Tandem repeats with repeat size less than 7 had more tandem repeats common in all strains than the unique repeats of this type in either of the three strains of *E. coli*. It was significant that the repeats of type (7,2), (8,2) and (12,2) had no common occurrence. These occurred only in the O157 strains of *E. coli*. Sequences of these unique tandem repeats, each occuring twice, were *tgttctg*, *gacaccgt* and *atcccaacccaa*.

Out of a total of 779 unique tandem repeats identified in the 44 outer membrane genes, only 155 led to amino acid variations in their corresponding protein sequences, the remaining 624 resulted in silent mutations. The 155 tandem repeats resulted in 116 amino acid variations, with two or more tandem repeats leading to a similar variation in certain cases. Table 2 lists changes in the amino acids due to the tandem repeats.

A comparison of the protein sequences corresponding to the 44 outer membrane genes studied highlighted a total of 230 amino acid variations in the two strains of *E. coli*. Out of these 230 mutations, 116 were due to the occurrence of tandem repeat(s), at the corresponding region in the DNA sequence of the gene, contributing 54.3% towards the total mutations. Analysis of the mutation rate of individual amino acids (Table 3) revealed that the average mutation rate of the amino acids was 1.2 per 100 with a range from 0.3 to 2.7 per every 100 amino acids, whereas the rate of mutation of amino acids due to tandem repeats had a value of 0.6 per 100 with a range of 0.1 to 1.3. An in-depth analysis of the mutation rates of individual amino acids led to the observation that five amino acids, viz. histidine, threonine, isoleucine, valine and lysine had a higher mutation rate compared to other amino acids and four amino acids methionine, glutamic acid, tyrosine and glycine had mutation rates less than 0.5. Histidine had the highest mutation rate of 2.7, whereas glycine had the lowest mutation rate of 0.3. When the contribution of each amino acid was studied, it was found that valine, serine, threonine and alanine contribute to about 43.5% of the total mutations. Cysteine and tryptophan were significant in that they did not undergo any mutations in the 44 genes studied, leading to a mutation rate of zero.

Out of the 44 genes, the protein structures were available corresponding to only three genes, viz. *fhuA*, *mltB* and *fepA* of *E. coli* K12. Structural analysis indicated that in the case of the fhuA protein, the intramolecular hydrogen bond involving serine in *E. coli* K12 was lost due to its mutation to alanine in both strains of *E. coli* O157, whereas the neighbouring residues were unaffected. In the case of the mltB protein, which had a mutation from lysine in *E. coli* K12 to arginine in *E. coli* O157, as expected no structural changes were observed. Structural analysis of the fepA protein revealed that four out of the five mutations were on the surface of the protein molecule, while the most significant mutation converting lysine in *E. coli* K12 to glutamic acid in *E. coli* O157 was buried deep inside the hydrophobic core, resulting in a considerable change in the conformation around this region of the protein. Of the five mutations in *fepA*, the conversion of alanine in *E. coli* K12 to serine in both the strains of *E. coli* O157 resulted in the formation of new hydrogen bond, while the conversion of serine in *E. coli* K12 to alanine in *E. coli* O157 resulted in the loss of hydrogen bond. Mutation of valine in *E. coli* K12 to isoleucine in *E. coli* O157 did not result in any changes. The mutation converting isoleucine in *E. coli* K12 to asparagine in *E. coli* O157 resulted in the formation of two new hydrogen bonds making the molecule energetically more stable. For the mutation converting Lys47 in *E. coli* K12 to glutamic acid in *E. coli* O157, a conformational change in the region around this residue has been observed. The side chain of this lysine was at a distance of 3.7 Å from Asp 676 and a distance of 8.78 Å from Arg 50. On the other hand, the side chain of the mutated residue glutamic acid moved away from Asp 676 by a distance of 3.2 Å and closer towards Arg 50 (Figure 1), thus resulting once again in a more favourable conformation.

Tandem repeats are common constituents across all genomes. Earlier studies on tandem repeats concentrated either on a single organism or single type of repeat. The current study involves a comparative evaluation of the tandem

Table 3. Mutation rate and percentage contribution of each amino acid towards overall mutations and mutations due to tandem repeats (TRs) across the 44 outer membrane genes of *E. coli* K12 and E.coli O157* strains of *E. coli*

Amino acid	Total mutations	Mutation rate[1]	Per cent contribution[2]	Mutations due to TRs[3]	Mutation rate due to TRs[4]	Per cent contribution due to TRs	Mutations due to non TRs	Mutation rate due to non TRs	Per cent contribution due to non TRs
Phenylalanine	8	1.15	3.48	5	0.72	4.27	3	0.43	2.65
Leucine	15	0.88	6.52	4	0.24	3.41	11	0.64	9.73
Isoleucine	18	1.93	7.83	8	0.85	6.89	10	1.07	8.77
Methionine	2	0.48	0.87	1	0.24	0.85	1	0.24	0.88
Valine	24	1.77	10.43	14	1.03	11.9	10	0.73	8.84
Serine	24	1.64	10.43	12	0.82	10.2	12	0.81	10.6
Proline	4	0.52	1.74	0	0	0	4	0.52	3.53
Threonine	29	2.07	12.61	16	1.14	13.6	13	0.92	11.5
Alanine	23	1.33	10	12	0.69	10.2	11	0.63	9.73
Tyrosine	3	0.34	1.3	2	0.22	1.7	1	0.11	0.88
Histidine	8	2.66	3.48	4	1.31	3.41	4	1.31	3.53
Glutamine	10	0.98	4.35	5	0.49	4.27	5	0.49	4.42
Asparagine	16	1.25	6.96	6	0.46	5.12	10	0.78	8.84
Lysine	13	1.65	5.65	9	1.14	7.69	4	0.5	3.53
Aspartic acid	14	1.19	6.09	7	0.59	5.98	7	0.59	6.19
Glutamic acid	4	0.48	1.74	3	0.36	2.56	1	0.11	0.88
Cysteine	0	0	0	0	0	0	0	0	0
Tryptophan	0	0	0	0	0	0	0	0	0
Arginine	9	0.96	3.91	5	0.53	4.27	4	0.42	3.53
Glycine	6	0.33	2.61	3	0.16	2.56	3	0.16	2.65
	Total: 230	Avg: 1.15		Total: 116	Avg: 0.55		Total: 114	Avg 0.52	

*Common in both strains of *E. coli* O157.

[1] Mutation rate of an amino acid X = Number of amino acids of type X mutated in all 44 proteins × 100/total number of amino acids of type X in all 44 proteins.

[2] Per cent contribution of amino acid = Number of mutations of the amino acid × 100/total number of mutations.

[3] Mutation rate due to TRs of an amino acid X = Number of amino acids, of type X, mutated due to tandem repeats in all 44 proteins × 100/total number of amino acids, of type X in all 44 proteins.

[4] Per cent contribution due to TRs of an amino acid = Number of mutations of the amino acid due to tandem repeats × 100/total number of mutations due to tandem repeats.

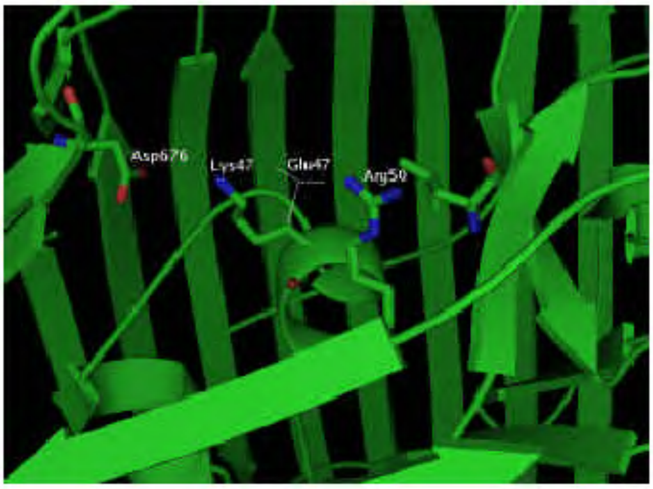

Figure 1. Conformational change in the fepA protein due to mutation of Lys 47 in *E. coli* K12 strain to Glu47(shown in thin lines) in *E. coli* O157 strain. Salt bridge between Lys 47 and Asp 676 is broken and a new salt bridge between Glu 47 and Arg 50 is formed in the pathogenic *E. coli* O157 strain.

repeats across 44 outer membrane genes of *E. coli* K12 and *E. coli* O157 strains of *E. coli*.

There is a strong preference for short tandem repeats in the 44 genes studied. The absence of repeats of length 9, 10 and 11, single occurrence of repeats of type (7,2), (8,2) and (12,2) specifically in pathogenic strains and the fact that the dinucleotide and trinucleotide repeats account for about 85% of total repeats indicate the common occurrence of short tandem repeats. The preference of *E. coli* genome for short repeats of up to five base pairs was clearly evident from an earlier study[10] in which it was demonstrated that the repeats of six or more base pairs in length comprised just 2.4% of the *E. coli* genome. Our results are in agreement with the earlier work[6] that reported common occurrence of short repeat unit lengths in prokaryotes. The results demonstrate the presence of repeats of type (7,2), (8,2) and (12,2) only in the pathogenic strain *E. coli* O157 but not in *E. coli* K12. Similar observations were made in other organisms such as *H. influenzae*[23], *Salmonella* and *E. coli*[12].

Tandem repeats accounted for 54% of the total mutations across all the 44 proteins studied. The occurrence of tandem repeats in genes resulted in two kinds of mutations in the corresponding proteins, namely silent mutations whereby a codon is modified to an alternate codon for the same amino acid or in amino acid mutations, whereby an amino acid is replaced by a new amino acid. The silent mutations and the codons they affect will have to be examined in detail to estimate their effect on protein folding.

Analysis of the three available structures, namely, fhuA, mltB and fepA for amino acid mutations resulted in interesting conclusions. While the mutations in fhuA and mltB did not lead to any conformational changes in their respective new structures in the pathogenic strains of *E. coli* O157, analysis of fepA protein revealed that the mutations leading to conversion of lysine to glutamic acid and isoleucine to asparagine resulted in significant conformational changes in the protein structure in the pathogenic strains of *E. coli.* In the native protein, the side chain of Lys 47 formed a salt bridge with that of Asp 676. The conversion of positively charged lysine to negatively charged glutamic acid at position 47 in the pathogenic strain resulted in the movement of the side chain of the mutated residue, thereby breaking the above salt bridge. However, this movement of Glu 47 facilitated formation of a new salt bridge between its side chain and that of Arg 50. The conversion of isoleucine, a hydrophobic residue to a polar residue, asparagine on the surface of the protein resulted in a large change in energy. Due to its hydrophobic nature isoleucine was less stable on the surface of the molecule, while the conversion to asparagine, a polar amino acid, resulted in the formation of two additional hydrogen bonds resulting in an energetically more stable structure.

1. Ohno, S. and Epplen, J. T., The primitive code and repeats of base oligomers as the primordial protein-encoding sequence. *Proc. Natl. Acad. Sci. USA*, 1983, **80**, 3391–3395.
2. Versalovic, J., Koeuth, T. and Lupski, J. R., Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. *Nucleic Acids Res.*, 1991, **19**, 6823–6831.
3. Bachellier, S., Clement, J. M., Hofnung, M. and Gilson, E., Bacterial Interspersed Mosaic Elements (BIMEs) are a major source of sequence polymorphism in *Escherichia coli* intergenic regions including specific associations with a new insertion sequence. *Genetics*, 1997, **145**, 551–562.
4. Katti, M. V., Ranjekar, P. K. and Gupta, V. S., Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol.*, 2001, **18**, 1161–1167.
5. Levinson, G. and Gutman, G. A., High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. *Nucleic Acids Res.*, 1987, **15**, 5323–5338.
6. van Belkum, A., Scherer, S., Van Alphen, L. and Verbrugh, H., Short-sequence DNA repeats in prokaryotic genomes. *Microbiol. Mol. Biol. Rev.*, 1998, **62**, 275–293.
7. Walker, A., Petheram, S. J., Ballard, L., Murph, J. R., Demmler, G. J. and Bale Jr J. F., Characterization of human cytomegalovirus strains by analysis of short tandem repeat polymorphisms. *J. Clin. Microbiol.*, 2001, **39**, 2219–2226.
8. Bifani, P., Moghazeh, S., Shopsin, B., Driscoll, J., Ravikovitch, A. and Kreiswirth, B. N., Molecular characterization of *Mycobacterium tuberculosis* H37Rv/Ra variants: Distinguishing the mycobacterial laboratory strain. *J. Clin. Microbiol.*, 2000, **38**, 3200–3204.
9. Shin, Y. C., Lee, H., Walsh, G. P., Kim, J. D. and Cho, S. N., Variable numbers of TTC repeats in *Mycobacterium leprae* DNA from leprosy patients and use in strain differentiation. *J. Clin. Microbiol.*, 2000, **38**, 4535–4538.
10. Gur-Arie, R., Cohen, C. J., Eitan, Y., Shelef, L., Hallerman, E. M., and Kashi Y., Simple sequence repeats in *Escherichia coli*: Abundance, distribution, composition, and polymorphism. *Genome Res.*, 2000, **10**, 62–71.
11. Moxon, E. R., Rainey, P. R., Nowak, M. A. and Lenski, R. E., Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.*, 1994, **4**, 24–33.
12. Rocha, E. P., Matic, I. and Taddei, F., Over-representation of repeats in stress response genes: A strategy to increase versatility under stressful conditions? *Nucleic Acids Res.*, 2002, **30**, 1886–1894.
13. Weiser, J. N., Love, J. M. and Moxon, E. R., The molecular mechanism of phase variation of *H. influenzae* lipopolysaccharide. *Cell*, 1989, **59**, 657–665.
14. Jarosik, G. P. and Hansen, E. J., Identification of a new locus involved in expression of *Haemophilus influenzae* type b lipooligosaccharide. *Infect Immunol.*, 1994, **62**, 4861–4867.
15. Stern, A., Brown, M., Nickel, P. and Meyer, T. F., Opacity genes in *Neisseria gonorrhoeae*: Control of phase and antigenic variation. *Cell*, 1986, **47**, 61–67.
16. Sarkari, J., Pandit, N., Moxon, E. R. and Achtman, M., Variable expression of the Opc outer membrane protein in *Neisseria meningitidis* is caused by size variation of a promoter containing polycytidine. *Mol. Microbiol.*, 1994, **13**, 207–217.
17. Jonsson, A. B., Nyberg, G. and Normark, S., Phase variation of gonococcal pili by frameshift mutation in pilC, a novel gene for pilus assembly. *EMBO J.*, 1991, **10**, 477–488.
18. Gotschlich, E. C., Genetic locus for the biosynthesis of the variable portion of *Neisseria gonorrhoeae* lipopolysaccharide. *J. Exp. Med.*, 1994, **180**, 2181–2190.
19. Jennings, M. P., Hood, D. W., Peak, I. R., Virji, M. and Moxon, E. R., Molecular analysis of a locus for the biosynthesis and phase-variable expression of the lacto-N-neotetraose terminal lipopolysaccharide structure in *Neisseria meningitidis. Mol. Microbiol.*, 1995, **18**, 729–740.
20. Landau, G. M. and Schmidt, J. P., An algorithm for approximate tandem repeats. In *Proceedings of the 4th Annual Symposium on Combinatorial Pattern Matching* (eds Apostolico, A. *et al.*), Padova, Italy, Springer-Verlag, Berlin, 1993, 684, pp. 120–133.
21. Gusfield, D., Core string edits, alignments, and dynamic programming. In *Algorithms on Strings, Trees, and Sequences*, Cambridge University Press, 1997, chapter 11.
22. Thompson, D., Higgins, G. and Gibson, J., ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 1994, **22**, 4673–4680.
23. Field, D. and Wills, C., Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 1647–1652.

ACKNOWLEDGEMENTS. We thank Tata Consultancy Services Limited, Hyderabad for providing facilities to carry out the present work. We also thank Pradeep Reddy for valuable inputs.

Received 16 April 2005; revised accepted 29 October 2005