http://www.pjbs.org

PJBS

ISSN 1028-8880

Pakistan Journal of Biological Sciences

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Pakistan Journal of Biological Sciences 11 (20): 2372-2381, 2008
ISSN 1028-8880

In silico Comparison of Simple Sequence Repeats in High Nucleotides-Rich Genomes of Microorganism

Fakher Rahim
Physiology Research Center, Ahwaz Jondishpour University of Medical Sciences, Ahwaz, Iran

Abstract: This study determined the distribution of a specific group of Simple Sequence Repeats (SSRs), in genome sequences of 7 chromosomes (*Shigella flexneri* 2a str 301 and 2457 T, *Shigella sonnei*, *E. coli* K12, *M. tuberculosis*, *M. leprae* and *S. saprophytycus*) have downloaded from the GenBank database for identifying abundance, distribution and composition of SSRs. The data obtained in the present study show that: (i) Tandem repeats are widely distributed throughout the genomes. (ii) SSRs are differentially distributed among coding and non-coding regions in investigated *Shigella* genomes. (iii) Total frequency of SSRs in non-coding regions is higher than coding regions. (iv) In all investigated chromosomes ratio of Tri-nucleotide SSRs are much higher than randomized genomes and Di nucleotide SSRs are lower. (v) Ratio of total and mono-nucleotide SSRs in real genome is higher than randomized genomes in *E. coli* K12, *Sh. flexneri* str 301 and *S. saprophyticus*, while it is lower in *Sh. flexneri* str 2457T, *Sh. sonnei* and *M. tuberculosis* and it is approximately same in *M. leprae*. (vi) Frequency of codon repetitions are vary considerably depending on the type of encoded amino acid.

Key words: Simple sequence repeats, GenBank, genome, chromosome

INTRODUCTION

Repetitive DNA consists of homopolymeric tracts of single nucleotides or of small or large numbers of multimeric classes of repeats. These can either be homogeneous (i.e., built from identical units) or heterogeneous (i.e., built from mixed units) (Van Belkum *et al*., 1998). A special category of repeats are tandem repeats, which are made up of monomeric sequences of variable length, repeated periodically, with contiguous monomers arranged in a head-to-tail fashion (Yeramian and Buc, 1999). Simple Sequence Repeats (SSRs) refer to the sequences that are one to six-nucleotides (nt) repeated in tandem in a genome. SSRs have many advantageous features for various biological studies: SSRs are ubiquitous and abundant in a genome, highly variable and suitable for high-throughput applications (Ellegren, 2004; Lawson and Zhang, 2006; Selkoe and Toonen, 2006; Choi *et al*., 2004; Yu *et al*., 2004a, b; Suwabe *et al*., 2006; Dettman and Taylor, 2004). In addition to practical usages of SSRs for biological studies, the SSRs have also been under the intense scrutiny of researchers to elucidate the evolution of genomes: (1) why are they ubiquitously present in a genome, (2) how do they arise, (3) why are they are unusually polymorphic and (4) what are their biological or structural functions are (Ellegren, 2004; Buschiazzo and Gemmell, 2006)? The evolutionary dynamics of SSRs have been actively discussed and hypotheses for experimental confirmation have been reviewed in the recent literature (Ellegren, 2004; Buschiazzo and Gemmell, 2006; Li *et al*., 2002, 2004). The variability in repeat number of these small tandem repeats (also called simple sequence repeats or SSRs) is caused by slipped-strand mispairing, in which the tertiary structure of the repetitive DNA allows mismatching of the neighboring repeats and repeats can then be inserted or deleted during DNA polymerase-mediated DNA duplication (Van Belkum *et al*., 1998; Levinson and Gutman, 1987; Van Belkum *et al*., 1999). The genus *Shigella* an etiological agent of bacillary dysentery, identified in 1890's, a very important member of the family Enterobacteriaceae is classified into four etiologically important species viz., *Shigella flexneri*, *Shigella dysenteriae*, *Shigella sonnei* and *Shigella boydii* (Hale, 1991). Simple sequence repeats (SSRs), or microsatellites, are the genetic loci where one or a few bases are tandemly repeated for varying numbers of times (Levinson and Gutman, 1987). Repetitive DNA consists of simple homopolymeric tracts of a single nucleotide type [poly (A), poly (G), poly (T), or poly(C)] or of large or small numbers of several multimeric classes of repeats. These multimeric repeats are built from identical units (homogeneous repeats), mixed units (heterogeneous repeats), or degenerate repeat sequence motifs

(Jeffreys *et al*., 1985). SSRs have been extensively studied in eukaryote genomes and are well-established targets for pedigree analysis (Jeffreys *et al*., 1986). But little is currently known about microsatellites in simple organisms (Field and Wills, 1998). Bacterial SSR-type DNA can be divided into four main categories. First, dispersed repeat motifs that generally do not occur in tandem have been identified. Although these repeats occur throughout genomes of a multitude of microorganisms, they are sometimes organized in tandem as well. The homopolymeric tracts form a second class. Multimers of one of the four nucleotides are peculiar sequence elements that are frequently encountered in the genome of *S. cerevisiae*, for instance. These homogeneous stretches can amount to as much as 42 nucleotides. Third, short-motif SSRs are identified. With repeat units differing from 2 to 6 bases, it is this class of repeats that is most liable to unit number variation at a given locus. Particularly, when these short-motif repeats are located within genes and are not 3 or 6 nucleotides long, they can drastically affect the coding potential of a given transcript. Fourth, repeats harboring more than 8 nucleotides per unit form a separate category. (Van Belkum *et al*., 1998). Some investigators considered SSRs to be selectively neutral sequences randomly or almost randomly distributed over the euchromatic genome (Schlotterer and Wiehe, 1999; Schlotterer, 2000). Initial studies of humans reported a higher mutation rate of tetra-nucleotide repeats (Weber and Wong, 1993), whereas a later study that compared microsatellite variability in different human populations found strong evidence for an inverse correlation of microsatellite repeat unit length and mutation rate (Chakraborty *et al*., 1997). Prokaryotic and eukaryotic repeat families are clustered to nonhomologous proteins. This may indicate that repeated sequences emerged after these two kingdoms had split. The eukaryotes incorporating more repeats may have an evolutionary advantage of faster adaptation to new environments (Kashi *et al*., 1997; King and Soller, 1999; Wren *et al*., 2000). In a variety of organisms, it has been demonstrated that microsatellite mutation rates are positively correlated with repeat number (Wierdl *et al*., 1997; Schlotterer *et al*., 1998). In prokaryotes, strong positive selective pressures are associated with highly mutable microsatellite tracts that control pathogenicity (Moxon *et al*., 1994). The increasing availability of prokaryotic genome sequences has shown that SSRs are also widespread in prokaryotes and that there is extensive variation in their length, number and distribution (Cox and Mirkin, 1997; Field and Wills, 1998; Gur-Arie *et al*., 2000; Coenye and Vandamme, 2003; Yang *et al*., 2003). The present study attempt to analyze distribution and composition of SSRs in the entire genomes of three strain of *Shigella* and compared with *E. coli* K12, GC rich (*M. tuberculosis* and *M. leprae*) and also AT rich genomes (*S. saprophytycus*).

MATERIALS AND METHODS

This study was conducted in the physiology research center, Ahwaz Jondishapour University of Medical Sciences, Ahwaz, Iran, from July 2006-July 2007.

DNA sequences: The whole genome sequence of *Sh. flexneri* 2a str 301 (NC_004337), *Sh. flexneri* 2a str 2457T (NC_004741), *Shigella sonnei* Ss046 (NC_007384), *Escherichia coli* K12 (NC_000913), *Mycobacterium tuberculosis* CDC1551(NC_002755.2), *Mycobacterium leprae* TN (NC_002677) and *Staphylococcus saprophyticus* subsp. saprophyticus ATCC 15305(NC_007350) were downloaded from the GenBank database (Table 1).

Analysis of SSRs: In this study two softwares for identifying SSRs have been used. The first one was developed by Gur-Arie *et al*. (2000) to screen the entire genome of the organisms included in this study for SSRs with minimal number of three repeats for chromosomes, minimal motif length of one and minimal length of whole SSR array two. This software can be downloaded from ftp://ftp.technion.ac.il/supported/biotech/ssr.exe. The second software was MICAS (microsatellite Analysis Server) an Interactive web-based server to find

Table 1: Whole-genome sequences used in this study

Organism	Accession No.	GC mol (%)	Genome size	CDS No.	CDS Bases	CDS Genome (%)	Repeat density	Observed No. of repeats (O)	Expected No. of repeats (E)	O/E
Sh.f 2a str 301	NC_004337	50.9	4 607 203	4031	2 897 451	62.8	0.0476	35828	14908	2.40
Sh.f 2a str 2457T	NC_004741	50.9	4 599 354	3927	2 796 275	60.7	0.0222	19024	13511	1.41
Sh. sonnei Ss046	NC_007384	51.0	4 825 265	4108	2 954 622	61.2	0.0638	5779	3104	1.86
E. coli K12	NC_000913	50.8	4 639 221	3857	2 415 482	52.1	0.0222	19237	13620	1.41
M. tuberculosis CD1551	NC_002755	65.6	4 411 532	3764	2 625 357	59.5	0.0079	6767	23608	0.29
M. leprae TN	NC_002677	57.8	3 268 203	3051	2 641 589	80.8	0.0095	5925	11526	0.51
S. saprophyticus ATCC 15305	NC_007350	32.1	2 516 575	2134	1 961 106	77.9	0.0542	38709	25981	1.49

non-redundant microsatellites in coding and non-coding region of genome sequence. This software also can be downloaded from http://210.212.212.7/MIC/gr-ve.html or (http://www.cdfd.org.in/micas).

Statistical analysis: To determine difference between the observe and the expected number of tandem repeats in entire genome of the organisms included in this study, distribution of SSRs between coding and non-coding regions of the genome and compare SSRs distributions with random expectations in coding and non coding regions, SPSS 11.0.1, SAS 9.1 and Sequence Shuffling Tool (http://bcf.arl.arizona.edu/resources/online_tools/shuffle.php) have been used. Statistical significance was tested with χ^2 test and two-tailed t-tests.

RESULTS

Distribution of SSRs: By a computer-based screen of genome sequence of three chromosome of *Shigella*, we found large number of SSRs with motif length 1-9 bp scattered through out genome (Table 2). The number of mono-nucleotide SSRs decreased rapidly with increasing size of the repeat unit and there is an almost perfect and highly significant linear relationship between the logarithm of the number of mono-nucleotide repeats and the repeat size ($p<0.0001$ for all genomes). Mono-nucleotide SSRs constituted the majority of SSRs in all 3 *Shigella* genomes, with the majority of mono-nucleotide SSRs being = 6 bp. As mono-nucleotide repeats number became higher, there is more and more representation of SSRs in non-coding regions, but it is no markedly difference in di and tri nucleotide SSRs (Table 4 and Fig. 2). In 2 strain of *Shigella flexneri* (301 and 2457T), coding regions contain less di-nucleotide and tetra-nucleotide SSRs than tri-nucleotide SSRs. In *Sh. sonnei* tri-nucleotide and tetra-nucleotide SSRs are more represented in coding regions than di-nucleotide repeats (Table 2).

Frequency of SSRs: In all investigated *Shigella* chromosomes, total frequency of SSRs in non-coding regions is higher than coding regions. Frequency of total SSRs in whole genome and coding regions of *Sh.f* 2457T is more than *sh.f* 301 and Sh.sonnei, however in non-coding region it is more in *Sh. sonnei* (Table 2). There is significant difference between frequency of total SSRs and also mono-nucleotide SSRs in coding regions and non-coding regions of 3 chromosomes of *Shigella* by χ^2 test ($p = 0.0001$). Frequency of total, mono-nucleotide and Di-nucleotide SSRs are higher in genome of *Staphylococcus saprophyticus* (AT-rich) than other genomes, it is 24, 21.68 and 1.2% of total nucleotides of the genome, respectively. Frequency of total and mono-nucleotide SSRs is lower in *Mycobacterium tuberculosis* (GC-rich) 15.5 and 13.04% and frequency of di-nucleotide SSRs is lower in *Sh.f1a* str 301. Frequency of triplet SSRs in *Mycobacterium tuberculosis* is much more than other genomes (Table 3). The distribution of mono-nucleotide SSRs over different length categories are significantly different between investigated genomes by χ^2 test ($p = 0.0001$).

The upper limits: The upper limits for mono-nucleotide SSRs are; 29 nt poly (T) in *Sh. sonnei*, 22 nt poly (G) in *M. leprae*, 17 nt poly (G) in *Sh.f* 301, 14 nt poly (G) in *Sh.f* 2457T,10 nt poly (G) in *E. coli* K12, 9 nt poly(G) in *M. tuberculosis* and 9 nt poly (A/T) in *S. saprophyticus*. The upper limits for any given SSRs are 108 bp in *Sh.f* 301, 98 bp in *Sh. sonnei*, 63 bp in *M. tuberculosis*, 58 bp in *Sh.f* 2457T, 48 bp in *E. coli* K12, 42 bp in *M. leprae* and 28 bp in *S. saprophyticus*.

Table 2: Frequency of SSRs > 3 bp in 3 chromosomes of *Shigella*

		Total SSRs		SSRs in coding region		SSRs in non-coding region		
Chromosome	Genomic length (bp)	N	%	N	%	N	%	GC (%)
Sh. sonnei	4825260	822647	17.0	639622	16.4	183025	19.7	50.8
Sh .f 2457 T	4599554	768345	16.7	565897	14.8	202448	21.2	50.9
Sh .f 301	4607203	789224	17.1	584896	15.9	204328	22.2	50.9

Table 3: The ratio of SSRs in real genome/randomized genome in different chromosomes

			Ratio of SSRs in real genome/randomized genome for				
Chromosomes	Length (bp)	GC (%)	Total SSRs	Mono nucleotides SSRs	Di nucleotides SSRs	Tri nucleotides SSRs	Tetra nucleotides SSRs
E. coli K12	463975	50.0	1.05	1.06	0.76	3.13	0.84
Sh .f 2a st.301	4607203	50.9	1.01	1.02	0.72	2.81	1.10
Sh.f 2a st. 2457 T	4599554	50.9	0.99	0.99	0.72	2.68	0.86
Sh. sonnei	4825260	50.8	0.98	0.99	0.72	2.65	0.79
M. tuberculosis	4403837	65.0	0.72	0.73	0.70	3.80	0.59
M. leprae	3268203	57.0	1.00	1.01	0.68	2.86	0.89
S. saprophyticus	2516575	33.0	1.08	1.09	0.68	2.39	0.54

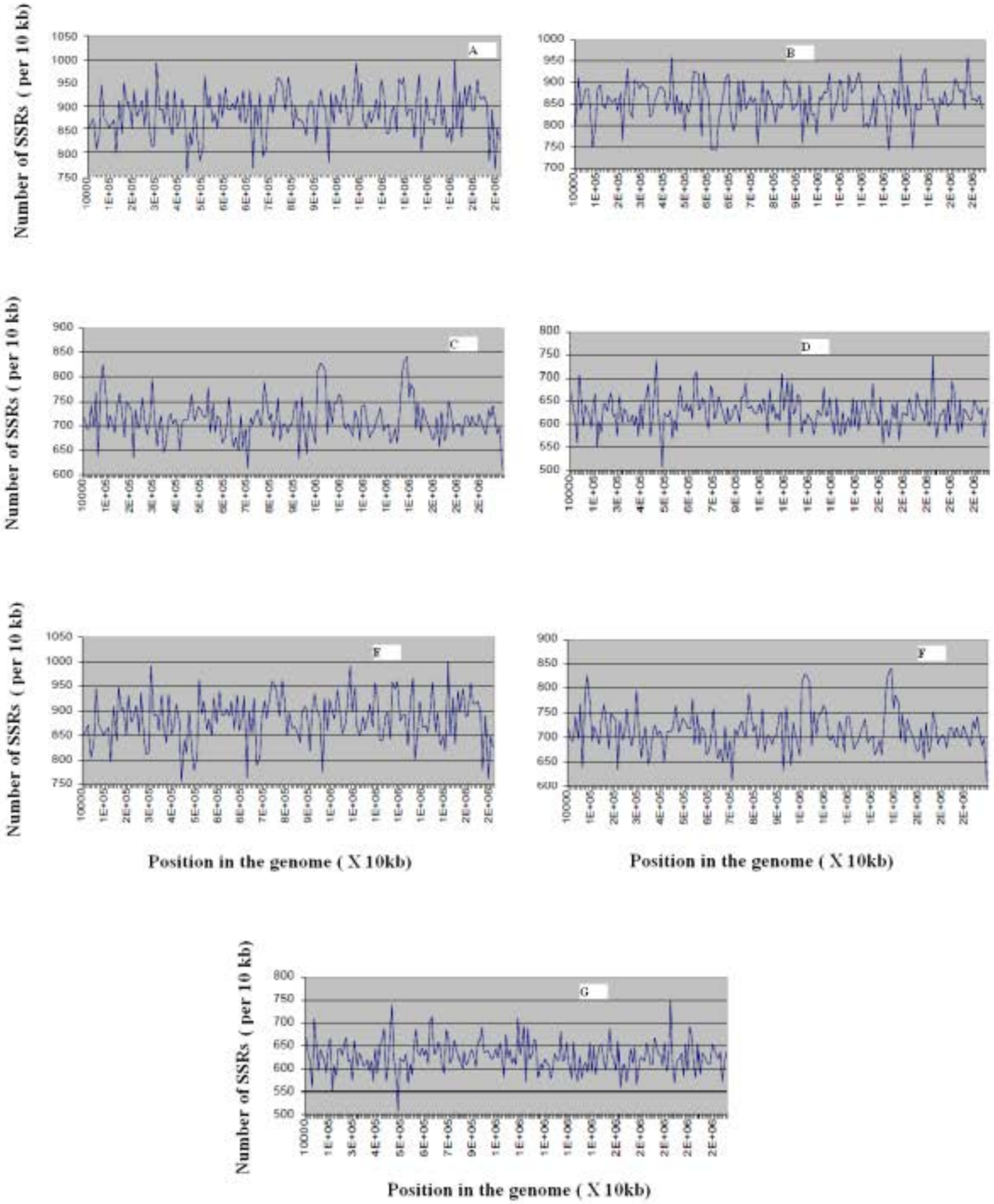

Fig. 1: Distribution of SSRs in the genomes of selected microorganisms, (A) Sh.f 2457T, (B) Sh.f 301, (C) *Sh. sonnei*, (D) *E. coli* K12, (E) *M. tuberculosis*, (F) *M. leprae* and (G) *S. saprophyticus*

Table 4: Comparison of No. of SSRs in 3 *Shigella* genome, *E. coli* K12, GC and AT rich genomes

	Genomes length bp N (%)						
No. of copy per nucleotide	Sh.f 2a St. 2457 T 4599554	Sh.f 2a St. 301 4607203	*E. coli* K12 4639675	*Sh. sonnei* 4825260	*M. tuberculosis* 4403837	*M. leprae* 3268203	*S. saprophytcus* 2516575
Mono	211892(15.72)	218725(16.1)	225446(16.52)	222843(15.73)	177013(13.04)	144775(14.16)	156616(21.68)
3	151802(9.9)	157238(10.2)	163345(10.6)	158475(9.85)	141748(9.7)	124828(11.46)	105847(12.6)
4	41614(3.6)	42441(3.7)	42901(3.7)	45633(3.8)	28472(2.6)	13633(1.7)	33105(5.3)
5	12916(1.4)	13613(1.5)	13837(1.5)	13178(1.37)	5830(0.44)	5266(0.81)	12562(2.5)
6	3950(0.52)	4071(0.5)	4123(0.53)	4481(0.56)	818(0.11)	851(0.16)	3591(0.86)
7	1357(0.2)	1142(0.17)	1000(0.15)	848(0.12)	138(0.022)	161(0.034)	1360(0.38)
8	221(0.038)	188(0.02)	217(0.037)	201(0.033)	6	30(0.007)	138(0.04)
9	32(0.006)	24(0.003)	22(0.004)	25(0.0047)	1	3	13
> 10	8	8	1	2	0	3	0
Di	7297(0.98)	7289(0.96)	7575(1.06)	7610(0.97)	8526(1.19)	5263(0.98)	4955(1.2)
3	6863(0.90)	6824(0.88)	7081(0.92)	7112(0.88)	7880(1.07)	4978(0.91)	4641(1.1)
4	412(0.072)	437(0.075)	465(0.08)	469(0.078)	606(0.11)	261(0.063)	296(0.09)
5	21(0.0046)	23(0.005)	28(0.006)	28(0.006)	40(0.009)	24(0.007)	16(0.006)
> 6	1	1	1	1	0	11	2
Tri	1770(0.356)	1812(0.356)	2401(0.468)	1845(0.365)	3998(0.738)	1542(0.426)	1297(0.469)
3	1715(0.34)	1756(0.34)	2335(0.45)	1789(0.33)	3794(0.78)	1502(0.41)	1263(0.45)
> 4	55(0.016)	56(0.016)	66(0.018)	56(0.015)	204(0.058)	40(0.016)	34(0.019)
Tetra	49	49	43	35	64	32	35
Penta	2	2	0	0	7	7	2
Hepta	10	9	3	9	18	4	3
3	7	6	0	9	16	3	3
> 4	3	3	0	0	2	1	
Hepta	0	0	0	2	0	1	1
Octa	0	0	2	1	0	0	1
Nona	3	3	0	0	55	0	0

Frequency of SSRs in real genome/randomized genome in different chromosomes: Total number of SSRs observed in 7 computer generated randomized genomes (with the same overall nucleotide frequency as the original genome) were higher than expected by chance alone in *E. coli* K12, *sh.f* 301 and *Staphylococcus saprophyticus* but it is lower in *Sh.f* 2457T, *Sh. sonnei* and specially *Mycobacterium tuberculosis* and it is approximately same in *Mycobacterium leprae* (Table 2). There is significant difference between frequency of total SSRs in real genome and randomized genome of investigated chromosomes by χ^2 test ($p<0.0001$). Ratio of mono-nucleotide composition in real genome/randomized genome show that there is overrepresentation of A/T mono-nucleotide SSRs in *Shigella* species and *E. coli* K12 (with 50-51 GC%), however in *M. tuberculosis* (with 65% GC) A/T mono-nucleotide SSRs in real genome is approximately same as randomized genome (Table 2).

With increasing size of mono-nucleotide motif length from 3-8 nt, in 3 investigated chromosomes of *Shigella* and *E. coli* K12 ratio of A/T mono-nucleotide in real genome/randomized genome are increased and G/C mono-nucleotide are decreased, except in motif length of 7 and 8 in *Shigella sonnei* (Fig. 3). There is significant difference between frequency of mono-nucleotide repeats of *Shigella* genomes and *E. coli* K12 with GC-rich and AT-rich genomes by χ^2 test ($p<0.0001$). In all investigated chromosomes ratio of real genome to randomized genome

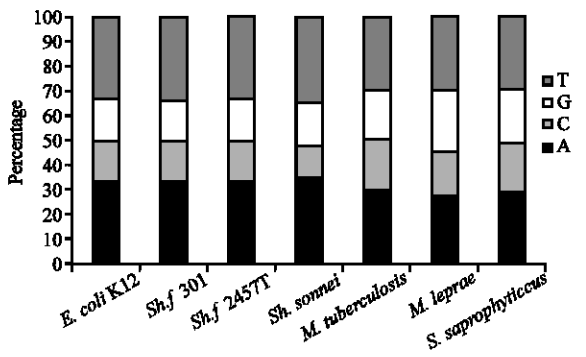

Fig. 2: Nucleotide composition of the mono-nucleotide SSRs≥6 bp in the genomes of investigated microorganisms

for Tri-nucleotide SSRs are much higher than 1 and Di nucleotide SSRs are lower and also tetra-nucleotide SSRs are lower than 1. Average of ratio of real genome to randomized genomes in 7 chromosomes for Di, Tri and tetra-nucleotide SSRs is 0.71, 2.9 and 0.8, respectively (Table 2). There is significant difference between ratio of real genome to randomized genomes in 7 chromosome for Di, Tri and tetra-nucleotide SSRs by χ^2 test ($p<0.0001$).

Composition of mono-nucleotide SSRS: The A/T composition of mono-nucleotide repeats in *Shigella genomes* is significantly higher than the overall

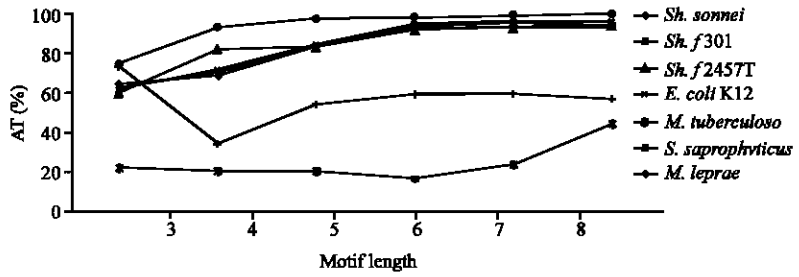

Fig. 3: AT% of mono-nucleotide SSRs

composition (and, consequently, an under representation of G and C mono-nucleotide SSRs), however this difference can exclusively be attributed to non-coding regions, difference is significant with χ^2 test ($p<0.0001$). Frequency of C mono-nucleotide SSRs in coding and non-coding regions of *Sh. sonnei* is more than *Sh. flexneri* 301 and 2457T and frequency of T mono-nucleotide SSRs in coding regions of *Sh. flexneri* 2457T is more than *Sh. flexneri* 301 and *Sh. sonnei*. In the genome of Sh.f 301, Sh.f 2457T, *Sh. Sonnei*, *E. coli* K12 and *S. saprophyticus* as repeat number became higher, frequency of A and T became more and more represented and difference is significant with χ^2 test ($p<0.001$). But no such trend is observable for *M. tuberculosis* and *M. leprae* (Table 2).

Frequency of di-nucleotide SSRs: In all 3 genome of *Shigella*, frequency of GC/CG in coding region is higher and frequency of AT/TA is lower. But frequency of GC/CG in *Sh. sonnei* is higher than 2 strain of *Shigella flexneri*. Difference is significant with two-tailed t-test ($p<0.01$). Frequency of AC/CA in non-coding region is higher but in *Shigella sonnei* difference is more. Frequency of CT/TC in non-coding region is higher than coding regions in all chromosomes (Table 2).

Frequency of tri-nucleotide SSRs: The tri-nucleotide SSRs are predominant in coding regions of *Sh. flexneri* str 301 and 2457T and *Sh. sonnei*. The tri-nucleotide SSRs can be grouped into 10 motif subclasses, each representing six overlapping and complementary unit patterns. Analysis of present data also indicates that (i) there is a tremendous overrepresentation of A and T in mononucleotide SSRs = 6 bp (and, consequently, an underrepresentation of G and C) (Fig. 1). The tri-nucleotide SSRs Groups number 9 and 10 in coding and non-coding regions of chromosomes, are over represented and group's number 5 and 6 are under represented in coding and non-coding regions of all *Shigella* chromosomes (Table 2). Between distribution of tri-nucleotide SSRs in coding region and non-coding region in both strain of *Shigella flexneri* is significant difference by two-tailed t-test ($p<0.001$).

Codon repetitions in complete genome sequences: In *Sh.f* 2a.str 2457T and 301, *Sh. sonnei* and *E. coli* K12 repetitions of Alanine (271, 237, 298, 318 time, respectively) are predominant, followed by Arginine (236, 220, 273, 246 times, respectively), Glutamine (174,173,161,163 time, respectively), Leucine and Valine. In *Mycobacterium tuberculosis* Arginine repetitions (1310 times) are predominant, followed by Alamine (958 times), Valine (287 time) and Serine (235 times). In *Mycobacterium leprae* Arginine repetitions (276 times) are predominant, followed by Alanine (268 times), Valine (177 times), Tereonin (117 times) and Serine (104 times). In *Staphylococcus saprophyticus* Isoleucine repetitions (267 times) a re predominant, followed by Tyrosine (133 times), Serine (96 times) and Leucine (66 times).

Frequency of tetra-nucleotide SSRs: In *Sh.f* 2a st 301 most frequency of tetra nucleotide SSRs are GCTG (6 times), CAGC (5 times), TGCC (5 times), CCAG (4 times) and CTGG (4 times). But in *sh.f* 2a str 2457T most frequency of tetra nucleotide SSRs is GCTG (6 time), TGGC (6 times), CTGG (5 times) and CCGA (4 times) and most of them are in coding region. The tetra-nucleotide SSRs are predominant in coding regions of *Sh. sonnei* and *sh.f* 2a str 2457T and non-coding regions of *Sh.f* 2a str. 301.

Frequency of longer unit SSRs: Frequency of Penta-nucleotide repeats was 7 in *M. tuberculosis* and *M. leprae* and 2 in Sh.f 2a str 2457T and 301 and *S. saprophyticus*. Frequency of hexa-nucleotide repeats was 17 in *M. tuberculosis*, 10 in Sh.f 2a str 2457T, 9 in Sh.f 2a str 301, 8 in *Sh. sonnei*, 4 in *M. leprae* and 3 in *S. saprophyticus*. Frequency of hepta-nucleotide repeats was 3 in *Sh. sonnei*, 1 in *M. leprae* and *S. saprophyticus*. Frequency of octa-nucleotide repeats was 1 in *Sh. sonnei*

and *S. saprophyticus*. Frequency of nonanucleotide repeats was 36 in *M. tuberculosis*. There are no penta-nucleotide repeats in *Sh. sonnei* and *E. coli* K12, heptanucleotide repeats in Sh.f 2a str 2457T and 301, *E. coli* K12 and *M. tuberculosis*, octa-nucleotide repeats in all investigated genomes except *Sh. sonnei* and nona-nucleotide repeats in all investigated genomes except *M. tuberculosis*. In *Sh.f* str 301 and 2457 T the hexa-nucleotide SSRs are predominant in coding regions but in *Sh. sonnei* it is predominant in non-coding regions.

DISCUSSION

Present data show that the investigated *Shigella* chromosomes contain numerous SSRs, with a motif length between 1- 9 nt, which are distributed almost evenly over the genome. This confirms similar findings reported in earlier studies for other organisms and *Shigella flexneri* 301 (Field and Wills, 1998; Gur-Arie *et al*., 2000; Coenye and Vandamme, 2003; Yang *et al*., 2003). As mono-nucleotide repeat number became higher, there is more and more representation of SSRs in non-coding regions in 3 investigated *Shigella* genome, which can be due to the fact that longer mono-nucleotide SSRs has more opportunity to undergo slipped-strand mispairing and there will be more mutability in their length than in shorter mono-nucleotide SSRs. This could help to explain why these are overrepresented in non-coding regions of the genome as selection has ample opportunity to operate against these larger repeats that would cause frame shift and non-sense mutations in coding regions (Coenye and Vandamme, 2005). The observation that in some genomes (including the genomes of the e-Proteobacteria *Campylobacter jejuni*, *Helicobacter pylori*, *Helicobacter hepaticus*, *Wolinella succinogenes* and those of *Haemophilus ducreyi*, *Neisseria meningitidis* and *Synechocystis* sp.) larger mono-nucleotide SSRs are not (or to a lesser extent) excluded from coding regions, suggest they may play an important role in phase variation as this process has been observed in these organisms (Henderson *et al*., 1999; Linton *et al*., 2001; Saunders *et al*., 1998). DNA strand slippage can occur during transient dissociation and reannealing in the repeat region and this could be a deceptive event for DNA processing machinery leading to expansions or deletions in the repeat tracks. It has been suggested that if the nucleotides on the single strand are self-complementary, they can base pair to form loops or hairpins and stabilize strand slippage (Gacy *et al*., 1995; Moore *et al*., 1999). The upper limits for length of any given SSRs was higher in *sh.f* 301 (108 bp) and for mono-nucleotide SSRs was higher in *Sh. sonnei* (29 bp). The upper limits for length of any given SSRs and mono-nucleotide SSRs in *Staphylococcus saprophyticus* was lower (28 and 9, respectively). It has been proposed that these limits to repeat lengths are evidence for the fact that the increase of repeat length by mutations is counteracted by selection (through a mechanism acting on the length of the SSR sequence itself and/or through a mechanism acting on gene expression as affected by the SSR) (Gur-Arie *et al*., 2000). If this is true, present data suggest that, these mechanisms are less active in *Shigella* genomes than *Staphylococcus saprophyticus*.

The over representation of poly (A) and poly (T) mono-nucleotide repeats in all *Shigella* sp. can be explained by the fact that strand separation for these poly (A) and poly (T) tracts is considerably easier than for poly (G) or poly (C) tracts, increasing the possibility of slipped strand mispairing. In this study in 3 investigated *Shigella* genomes, *E. coli* K12, *M. tuberculosis* and *M. leprae* CG/GC Di-nucleotide SSRs are more frequent compared with other diucleotide repeats followed by GT/TG Di-nucleotide repeats and AT/TA Di-nucleotide repeats are extremely rare. In *S. Saprophyticus* AT/TA are predominant followed by AC/CA and GT/TG Di-nucleotide repeats (AT reach genome). It is evident that in human and *Drosophila* chromosomes, AC Di-nucleotide repeats are more frequent, followed by AT and AG repeats. In contrast, *Arabidopsis* chromosomes contain more AT repeats, followed by AG repeats. However, in the yeast genome, AT repeats seems to be predominant compared with other Di-nucleotide repeats. Interestingly, GC Di-nucleotide repeats are extremely rare in all of the eukaryotic genomes studied. Lower frequencies of CpG di-nucleotides in vertebrate genomes have been attributed to methylation of cytosine, which, in turn, increases its chances of mutation to thymine by deamination (Schorderet and Gartler, 1992). However, CpG suppression by this mechanism cannot explain the rarity of (CG)n Di-nucleotide repeats in yeast, *C. elegans* and *Drosophila*, since they do not show cytosine methylation. However, it has been observed that similar to present study: TA is underrepresented in almost all prokaryotic genomes; which could be due to the fact that (i) TA forms the thermodynamically least stable DNA (allowing unwinding of the helix), (ii) RNases preferentially degrade UA Din-ucleotides in mRNA and/or (iii) TA is part of many regulatory sequences. This may explain why TA/AT in di-nucleotide SSRs is lower than GC/CG. Tri-nucleotide SSRs in coding regions of three investigated *Shigella* genome are overrepresented, whereas Di-nucleotide and tetra-nucleotide repeats are underrepresented. It has been reported that triplet repeats show approximately twofold greater frequency in exonic

regions than in intronic and intergenic regions in all human chromosomes except the Y chromosome (Subramanian *et al*., 2003). Such dominance of triplets over other repeats in coding regions may be explained on the basis of the suppression of nontrimeric SSRs in coding regions, possibly caused by frame shift mutations (Metzgar *et al*., 2000).

Frequency of codon repetitions in complete genome sequences of Sh.f 2a .str 2457T and 301, *Sh. sonnei* and *E. coli* K12 are approximately same. Codon repetitions are comparatively more numerous in *Mycobacterium tuberculosis* (Arginine 1310 time) than in other investigated genomes, (even *Mycobacterium leprae* with Arginine repetition 276 time) since the comparatively frequency of microsatellites is very low. Frequencies of codon repetitions are low in *Staphylococcus saprophyticus* since the microsatellites are more frequent. While in all investigated genome except *Staphylococcus saprophyticus* Arginine and Alanine are predominant in *Staphylococcus saprophyticus* Isoleucine and Tyrosine are predominant and Arginine and Alanine are very low abundant. Within a Tri-nucleotide repeat class, frequencies of different codon repeats vary considerably depending on the type of encoded amino acid. More frequency of small/hydrophilic basic amino acids repetitions than hydrophobic amino acids in investigated genomes (except *Staphylococcus saprophyticus*) and hydrophobic than small/hydrophilic amino acids in *Staphylococcus saprophyticus* might play an important role in the structure and function of the encoded proteins in these genomes.

In *Drosophila* chromosomes, AGC repeats are predominant, followed by AAC repeats. The *Arabidopsis* and *C. elegans* chromosomes have comparatively higher frequencies of AAG tri-nucleotide repeats. In contrast, the yeast genome contains more AAT, AAG, AAC, ATG and AGC repeats. Present data indicate that when the GC content of mono-nucleotide SSRs is high and pathogenicity is more, the average and standard deviation of repeat density is lowest. This is confirmed by the observation of Coenye and Vandamme (2005), who have shown that the GC content of mono-nucleotide SSRs is highest when the repeat density is lowest and repeat density is significantly higher in organisms with an intracellular or strictly parasitic lifestyle. These observations suggest that the higher energy cost of G and C over A and T/U could be the reason for the high variation seen in genomic C+G content and it might be responsible for the marked differences observed in G+C content of these mono-nucleotide SSRs, as it would be too costly to have many poly (G) and/or poly(C) SSRs in genomes with a high density of mono-nucleotide SSRs (Coenye and Vandamme, 2005). While density of SSR in *E. coli* is more than *Sh. flexneri* str 301 and 2457T and *Sh. sonnei* and it is very low in *M. tuberculosis*, there is similarity between distributions of SSR during the genome of these organisms in most of positions.

This observation indicates that, like *Staphylococcus aureus*, e-*Proteobacteria*, *E. coli* and *R. solanacearum*, the *Shigella* sp. contains a large number of SSRs. The observed similarities such as distribution of SSRs in the genome and representation of various types of SSRs indicate that investigated *Shigella* genomes and *E. coli* K12 have shared a similar evolutionary history. Although there are some differences between investigated *Shigella* genomes and *E. coli* K12 such as frequency of total SSRs in whole genome, coding regions and non-coding regions. The upper limits for mono-nucleotide SSRs and any given SSRs, average of SSR density, composition of mono-nucleotide SSRs and frequency of di-nucleotide SSRs. These variations be attributable to differences in gene expression and regulation of gene expression. This study also suggest that, genomic distribution of SSR is nonrandom across coding and non-coding regions and differential distributions of various repeats observed in different genome sequences suggest that apart from the nucleotide composition of repeats, the characteristic DNA replication/repair/recombination machinery might have an important role in the evolution of SSRs.

ACKNOWLEDGMENTS

This study (Grant No. BIM 255) was supported by Bioinformatics Center of University of Pune and Physiology Research Center of Ahwaz Jondishapour University of Medical Sciences. We would like to thanks the Ahwaz Jondishapour University of Medical Sciences to give us opportunity to do this research.

REFERENCES

Buschiazzo, E. and N.J. Gemmell, 2006. The rise, fall and renaissance of microsatellites in eukaryotic genomes. Bioassays, 28: 1040-1050.

Chakraborty, R., M. Kimmel, D.N. Stivers, L.J. Davison and R. Deka, 1997. Relative mutation rates at di-, tri- and tetra-nucleotide microsatellite loci. Proc. Natl. Acad. Sci. USA., 94: 1041-1046.

Choi, H.K., D. Kim, T. Uhm, E. Limpens and H. Lim *et al*., 2004. A sequence-based genetic map of *Medicago truncatula* and comparison of marker colinearity with *M. sativa*. Genetics, 166: 1463-1502.

Coenye, T. and P. Vandamme, 2003. Simple sequence repeats and compositional bias in the bipartite *Ralstonia solanacearum* GMI1000 genome. BMC Genom., 4: 10-10.

Coenye, T. and P. Vandamme, 2005. Characterization of mono-nucleotide repeats in sequenced prokaryotic genomes. DNA Res., 12: 221-233.

Cox, R. and S. Mirkin, 1997. Haracteristic enrichment of DNA repeats in different genomes. Proc. Natl. Acad. Sci. USA., 94: 5237-5242.

Dettman, J.R. and J.W. Taylor, 2004. Mutation and evolution of microsatellite loci in *Neurospora*. Genetics, 168: 1231-1248.

Ellegren, H., 2004. Microsatellites: Simple sequences with complex evolution. Natl. Rev. Genet., 5: 435-445.

Field, D. and C. Wills, 1998. Abundant microsatellite polymorphisms 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., 95: 1647-1652.

Gacy, A.M., G. Goellner, N. Juranic, S. Macura and McMurray, 1995. Tri-nucleotide repeats that expand in human disease form hairpin structures *in vitro*. Cell, 81: 533-540.

Gur-Arie, R., C.J. Cohen, Y. Eitan, L. Shelef, E.M. Hallerman and Y. Kashi, 2000. Simple sequence repeats in *Escherichia coli*: Abundance, distribution, composition and polymorphism. Genet. Res., 10: 62-71.

Hale, T., 1991. Genetic basis of virulence in *Shigella* species. Microbial Rev., 55: 206-224.

Henderson, I.R., P. Owen and J.P. Nataro, 1999. Molecular switches the ON and OFF of bacterial phase Variation. Mol. Microbiol., 33: 919-932.

Jeffreys, A.J., V. Wilson and S.L. Thein, 1985. Hypervariable minisatellite regions in human DNA. Nature, 314: 67-73.

Jeffreys, A.J., V. Wilson, S.L. Thein, D.J. Weatheral and B.A.J. Ponder, 1986. DNA fingerprints and analysis of multiple markers in human pedigrees. Am. J. Hum. Genet., 39: 11-24.

Kashi, Y., D.G. King and M. Soller, 1997. Simple sequence repeats as a source of quantitative genetic variation. Trends Genet., 13: 74-78.

King, D.G. and M. Soller, 1999. Variation and Fidelity: The Evolution of Simple Sequence Repeats as Functional Elements in Adjustable Genes. In: Evolutionary Theory and Processes: Modern Perspectives, Wasser, S.P. (Ed.). ISBN: 9783904144230, pp: 65-82.

Lawson, M.J. and L. Zhang, 2006. Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. Genome Biol., 7: R14-R14.

Levinson, G. and G.A. Gutman, 1987. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. Mol. Biol. Evol., 4: 203-221.

Li, Y.C., A.B. Korol, T. Fahima, A. Beiles and E. Nevo, 2002. Microsatellites: Genomic distribution, putative functions and mutational mechanisms: A review. Mol. Ecol., 11: 2453-2453.

Li, Y.C., A.B. Korol, T. Fahima and E. Nevo, 2004. Microsatellites within genes: Structure, function and evolution. Mol. Biol. Evol., 21: 991-1007.

Linton, D., A.V. Karlyshev and B.W. Wren, 2001. Deciphering *Campylobacter jejuni* cell surface interactions from the genome sequence. Curr. Opin. Microbiol., 4: 35-40.

Metzgar, D., J. Bytof and C. Wills, 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res., 10: 72-80.

Moore, H., P.W. Greenwell, C.P. Liu, N.T. Arnheim and D. Petes, 1999. Triplet repeats form secondary structures that escape DNA repair in yeast. Proc. Natl. Acad. Sci. USA., 96: 1504-1509.

Moxon, E., P. Rainey, M. Nowak and R. Lenski, 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr. Biol., 4: 24-33.

Saunders, N.J., J.F. Peden, D.W. Hood and E.R. Moxon, 1998. Simple sequence repeats in the *Helicobacter pylori* genome. Mol. Microbiol., 27: 1091-1098.

Schlotterer, C., R. Ritter, B. Harr and G. Brem, 1998. High mutation rate of a long microsatellite allele in *Drosophila melanogaster* provides evidence for allele-specific mutation rates. Mol. Biol. Evol., 15: 1269-1274.

Schlotterer, C. and T. Wiehe, 1999. Microsatellites, a Neutral Marker to Infer Selective Sweeps. In: Microsatellites: Evolution and Applications, Goldstein, D.B. and C. Schltterer (Eds.). ISBN: 978-0-19-850407-8, pp: 238-247.

Schlotterer, C., 2000. Evolutionary dynamics of microsatellite DNA. Chromosoma, 109: 365-371.

Schorderet, D.F. and S.M. Gartler, 1992. Analysis of CpG suppression in methylated and nonmethylated species. Proc. Natl. Acad. Sci. USA., 89: 957-961.

Selkoe, K.A. and R.J. Toonen, 2006. Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers. Ecol. Lett., 9: 615-629.

Subramanian, S., R.K. Mishra and L. Singh, 2003. Genome wide analysis of microsatellite repeats in humans: Their abundance and density in specific genomic regions. Genome Biol., 4: R13-R13.

Suwabe, K., H. Tsukazaki, H. Iketani, K. Hatakeyama and M. Kondo *et al*., 2006. Simple sequence repeat-based comparative genomics between *Brassica rapa* and *Arabidopsis thaliana*: The genetic origin of clubroot resistance. Genetics, 173: 309-319.

Van Belkum, A., S. Scherer, L. Van Alphen and H. Verbrugh, 1998. Short-sequence DNA repeats in prokaryotic genomes. Microbiol. Mol. Biol. Rev., 98: 275-293.

Van Belkum, A., W. Van Leeuwen, S. Scherer and H. Verbrugh, 1999. Occurrence and structure-function relationship of pentameric short sequence repeats in microbial genomes. Res. Microbiol., 150: 617-626.

Weber, J.L. and C. Wong, 1993. Mutation of human short tandem repeats. Hum. Mol. Genet., 2: 1123-1128.

Wierdl, M., M. Dominska and T.D. Petes, 1997. Instability in yeast: Dependence on the length of the microsatellite. Genetics, 146: 769-779.

Wren, J.D., E. Forgacs and J.W. Fondon, 2000. Repeat polymorphisms within gene regions: Phenotypic and evolutionary implications. Am. J. Hum. Genet., 67: 345-356.

Yang, J., J. Wang, L. Chen, J. Yu, J. Dong, Z. Yao, Y. Shen, Q. Jin and R. Chen, 2003. Identification and characterisation of simple sequence repeats in the genomes of *Shigella* species. Gene, 322: 85-92.

Yeramian, E. and H. Buc, 1999. Tandem repeats in complete bacterial genomes sequences. Sequence and structural analysis for comparative studies. Res. Microbiol., 150: 745-754.

Yu, J.K., M. La Rota, R.V. Kantety and M.E. Sorrells, 2004a. EST derived SSR markers for comparative mapping in wheat and rice. Mol. Gen. Genom., 271: 742-751.

Yu, J.K., T.M. Dake, S. Singh, D. Benscher, W.L. Li, B. Gill and M.E. Sorrells, 2004b. Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat. Genome, 47: 805-818.