BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net *Hypothesis*

Volume 6(9)

Analysis of distribution and significance of simple sequence repeats in enteric bacteria *Shigella dysenteriae* SD197

Batwal Saurabh[1], Sitaraman Sneha[1], Ranade Suvidya[2]*, Khandekar Pramod[3], Bajaj Shailesh[2]

[1]Sinhagad College of Engineering, Wadgaon BK, Pune - 411041; [2]Department of chemistry, University of Pune, Maharashtra India; [3]Biotechnology Chair, University of Pune; Ranade Suvidya - Email: suvidya@chem.unipune.ac.in; *Corresponding author

Received June 13, 2011; Accepted June 28, 2011; Published July 19, 2011

Abstract:
We have explored the possible role of SSR density in genome to generate biological information. In our study, we have checked the SSR (simple sequence repeats) status in virulent and non virulent genes of enteric bacteria to see whether the SSRs distribution contributes to virulence. The genome, plasmid and virulent genes sequences in fasta format were downloaded from NCBI GenBank and VFDB. The sequences were subjected to SSR analysis using software tool ssr.exe. The resulting data was pasted in excel sheet and further analyzed for percentage of each type of SSR. Higher nucleotide repeats have been observed in our study. Overall high density of SSRs can enhance antigenic variance of the pathogen population in a strategy that counteracts the host immune response. Frequency of A and T repeats is higher in the chromosome, plasmid and the virulence genes. However, in dinucleotide repeats the frequencies of GC/CG repeats are higher in genome, whereas plasmid has more of AT/TA repeats. Genome has trinucleotide repeats having predominantly G and C whereas plasmid has trinucleotide repeats having predominantly A and T. The repeat number obtained and percentage of repeats is higher in virulence genes as compared to other gene families. Due to the presence of this large number of SSRs, the organism has an enormous potential for generating this genomic and phenotypic diversity.

Background:
Simple Sequence Repeats (SSRs) in DNA sequence are composed of tandem iterations of short oligonucleotides. SSRs may have functional and structural properties that distinguish them from general DNA sequences. SSRs are found abundantly in eukaryotic and prokaryotic genomes **[1, 2]**. SSRs are ubiquitously distributed in the genomes, both in protein coding and non-coding regions **[3]**. The SSRs consist of simple homopolymeric tracts of a single nucleotide base (poly (A), poly (C), poly (T) or poly (G) or of large or small numbers of several multimeric classes of repeats. Several classes of SSRs exist. The genus *Shigella* is an important human pathogen and is responsible for the majority of cases of endemic bacillary dysentery. Moreover, variability in the number of repeat units at a given genomic site, i.e. the sequence heterogeneity, among individual strains can be used to assess intra-species diversity. There is accumulating evidence that SSRs serve a functional role, affecting gene expression, and that polymorphism of SSR tracts may be important in the evolution of gene regulation **[4, 5, 6]**. Mutation mechanisms have been studied in some detail in eukaryotes, essentially human and yeast. The data obtained so far indicates that SSRs mutate by replication slippage process caused by mismatches between DNA strands while being replicated during meiosis **[7]**. Typically, slippage in each SSRs occur about once per 1,000 generations **[8]**. Molecular analysis of changes in SSRs allows epidemiological studies on the spread of pathogenic bacteria. In pathogens, SSRs can enhance antigenic variance of the pathogen population in a strategy that counteracts the host immune response **[9]**. In this scenario, SSRs located in protein coding regions or in upstream regulatory regions can reversibly deactivate or alter genes involved in interactions with the host. Some SSRs may also affect local structure of the DNA molecule. SSRs are informative markers for the identification of pathogenic bacteria, and may serve as indicators for the adaptation of pathogens in vivo and ex vivo environments **[10]**. SSR-mediated variation has important implications for bacterial pathogenesis and evolutionary fitness. In our study, we have analyzed the distribution and composition of SSRs of entire genome *of Shigella dysenteriae* SD197 and compared with the virulence factors of the genome and the virulence plasmid. We have also made an attempt to show how SSR studies are useful to generate new biological information.

Methods:

DNA Sequences:
All the DNA sequences were downloaded in FASTA format from (http://www.ncbi.nlm.nih.gov/genbank/). The details of genome/gene sequences, their lengths and other features are as follows. Genome of *Shigella dysenteriae* Sd197: Chromosome: (NCBI Entrez Genome) Genbank Accession Number- NC_007606, Size: 4369232 bp, Gene Count: 4660; Proteins: 4270. Plasmid pSD1_197:Genbank Accession Number: NC_007607; Size: 182726 bp, Gene Count: 224, Proteins: 223.

Databases:
The various databases used for downloading the genome, plasmid and genes include NCBI GenBank, Virulence Factor for Pathogenic Bacteria (VFDB), ShiBASE (details given in Supplementary material available with authors).

Analysis of SSRs:
In this study, we have used two software for identifying SSRs. Software developed by Gur-Arie i.e.Ssr.exe **[11]** downloadable from (ftp://ftp.technion.ac.il/pub/supported/biotech/) and MICAS (Microsatellite Analysis Server) available at http://www.cdfd.org.in/micas to screen the genomes, plasmids and virulent genes of the organism included in this study **[12]**. Virulent genes are shown in Table 2 (Supplementary material available with authors). Parameters set for extensive study of SSR analysis using ssr.exe include minimal number of repeats = 2, minimal motif length = 1, length of whole SSR array = (2*1) = 2. This software searches for all of the SSRs with motif lengths up to 10 bp; records motif, repeat number, and genomic location; and reports the results in an output file. The second software was MICAS (Microsatellite Analysis Server) an interactive web-based server to find the non-redundant microsatellites.

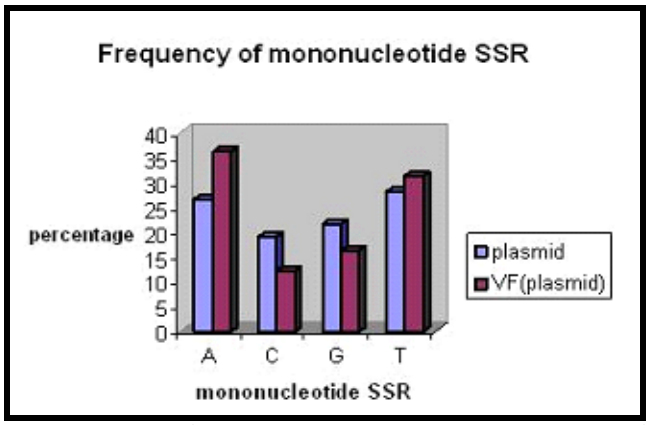

Figure 1: Frequency of mononucleotide SSR in plasmid and virulence genes of plasmid

Results and Discussion:
Large numbers of SSR were found to be scattered in whole genome, plasmid and all the virulence factor families. Our study shows high density of SSR in virulent genes and regulatory regions in SD197. The repeat number obtained and percentage of repeats obtained is more in virulence genes compared to structural genes. The SSR mononucleotide repeats were found to be in large number followed by dinucleotide, trinucleotide and higher motif repeats. This is shown in **Table 1 (see Supplementary material)**. Shorter repeats were found to be more abundant than the longer repeats. The environmental changes cause stress reactions such as change in copy number of tandem repeats. With this high density of SSR stress response genes and virulent genes may undergo such change resulting in changed activity of additionally relevant genes and further relaxation of stress by adapting to changed environment **[13]**.

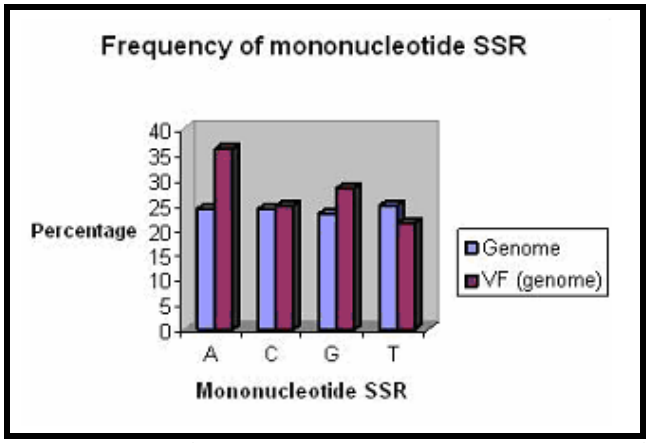

Figure 2: Frequency of mononucleotide SSR in genome and virulence genes of genome

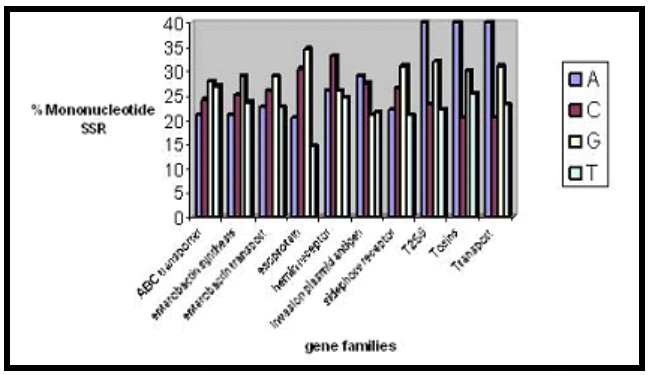

Figure 3: Frequency of mononucleotide SSR in virulence gene families of genome

The A and T repeats are overrepresented as compared to C and G in plasmid and genome. The frequency of A and T mononucleotide SSR is higher in the virulence genes of plasmid as compared to entire plasmid **(Figure 1)**. The overrepresentation of A and T mononucleotide SSR can be explained by different ways like slipped strand mispairing, is more likely for poly A or poly T as strand separation is energetically more favorable compared to poly GC. Similarly the higher energy cost of synthesis of CG dNTPs by the cells **[14]**. Longest SSR in plasmid was 10bp for A, 8bp for C, 11bp for G, and 19bp for T. There is overrepresentation of A and G mononucleotide SSR in virulence genes of genome as compared to entire genome **(Figure 2)**. Virulence gene families of genome show higher percentage of G mononucleotide SSR followed by C and A **(Figure 3)**. Presence of higher frequency of A and G repeats in virulence gene indicates presence of secondary structure in DNA. The longest SSR in genome was 23bp long A, 11bp long for C, 17bp for G, and 9bp for T. Longer mononucleotides SSRs have more opportunity to undergo slipped-strand mispairing and there will be more mutability in their length than in shorter mononucleotide SSRs. Another possible reason is the involvement of repeated sequences in the formation of non-canonical DNA structures, including triple stranded H-DNA. These structures are more easily formed in GA-rich regions **[15, 16]** and could block transcription by RNA polymerase **[17]**. Genome encoded virulence genes showed higher frequency of GC/CG dinucleotide repeats but it is lower as compared to entire genome. In contrast there was higher frequency of AC/CA, AT/TA, CT/TC repeats than the entire genome (Figure 4, Supplementary material available with authors). The frequency of AT/TA dinucleotide repeats was higher in plasmid encoded virulence genes compared to entire plasmid. It has been proposed that GT, CA, CT, GA GC or AT repeats binding proteins could participate in recombination process by inducing Z conformation of DNA or other alternative secondary DNA structures. Presence of GC/CG dinucleotide in genome more frequently compared to AT/TA could be due to the fact that TA forms thermodynamically least stable DNA. RNases preferentially degrade UA dinucleotides in mRNA. The large number of trinucleotide repeats was found in genome and plasmid. The motifs containing predominantly G and C are found to be over represented in genome. Whereas, the motifs containing predominantly A and T, are found to be over represented in plasmid. Some motifs were completely absent such as: AAC, AAG, AAT, CCA, CCG, CCT, GGA, GGT, GGC, TTA, TTG, and TTC. Highly repeated motif in genome was CAG and in plasmid was ATT. The distribution of tri and hexa-nucleotide repeats reflects codon repetition and that of amino acids suggesting these repeats are strongly selected and shows its association with protein function. The tetra nucleotide repeats were over represented in plasmids whereas penta-nucleotide repeats are slightly over represented in genome. The hexa nucleotide repeats are present more in genome. SSRs were also found in structural genes where the motif A was over represented (Figure 13, Supplementary material available with authors) However, the repeat number obtained and percentage of repeats obtained is less than that of virulence genes. Our study suggests that genomic distribution of SSR is non random and apart from nucleotide composition of repeats the characteristic DNA replication, repair and recombination machinery might have important role in the evolution of SSR. Our analyses performed on the genome, plasmid and large number of genes of *Shigella dysenteriae SD197* clearly indicates that, due to the presence of this large number of SSRs, the organism has an enormous potential for generating this genomic and phenotypic diversity.

Conclusion:

SSR of many types are found in prokaryotic genomes as well. These are present in functional domains and play an important role in functional alterations and implications in mutation helping the organism to adapt to its surroundings. Higher nucleotide repeats have been observed in our study. The repeat number obtained and percentage of repeats obtained is higher in virulence genes as compared to other gene families. We found that frequency of A and T repeats are higher in the chromosome, plasmid and the virulence genes. However, in dinucleotide repeats there is a significant difference in the motifs obtained as we observed that the frequencies of GC/CG repeats are higher in genome whereas plasmids harbor more of AT/TA repeats. Genome has trinucleotide repeats having predominantly G and C whereas plasmid has trinucleotide repeats having predominantly A and T. There is overrepresentation of mononucleotide repeats A and T and dinucleotide repeats AT/TA in the type III secretion system of plasmid which is composed of the mxi-spa group. This study will help in in-depth analysis and understanding of the elements that control and regulate the pathogenicity and survival of a microbe. This can also be used as a foundation for development of sophisticated molecular tools and diagnostic kits.

References:

[1] Weber JL. *Genomics* 1990 **7**: 524 [PMID: 1974878]
[2] Bayliss CD *et al. J Clin Invest.* 2001 **107**: 657 [PMID: 11254662]
[3] Toth G *et al. Genome Res.* 2000 **10**: 967 [PMID: 10899146]
[4] Kashi Y *et al. Trends Genet.* 1997 **13**: 74 [PMID: 9055609]
[5] King DG *et al*. *Endeavour* 1997 **21**: 36
[6] Belkum VA *et al. Res Microbiol*. 1999 **150**: 617 [PMID: 10673001]
[7] Tautz D. *Nucleic Acids Res.* 1989 **17**: 6463 [PMID: 2780284]
[8] Weber JL & Wong C. *Hum mol genet.* 1993 **2**: 1123 [PMID: 8401493]
[9] Mrazek J *et al. J Bacteriol.* 2010 **192**: 3763 [PMID: 20494989]
[10] van Belkum A *et al. Infect Immun.* 1997 **65**: 5017 [PMID: 9393791]
[11] Gur-Arie R *et al. Genome Res.* 2000 **10**: 62 [PMID: 10645951]
[12] Hosseini A *et al. DNA sequence*. 2008 **19**: 167 [PMID: 18464038]
[13] Trifonov EN. *Ann N Y Acad Sci*.1999 **870**: 330 [PMID: 10415494]
[14] Rocha EP *et al. Nucleic Acids Res.* 2002 **30**: 1886 [PMID: 11972324]
[15] Schroth GP & Ho PS. *Nucleic Acids Res*. 1995 **23**: 1977 [PMID: 7596826]
[16] Hoyne PR *et al. J Mol Biol*. 2000 **302**: 797 [PMID: 10993724]
[17] Ashely C & Lee JS. *DNA Cell Biol*. 2000 **19**: 235 [PMID: 10798447]

Edited by P Kangueane
Citation: Saurabh *et al.* Bioinformation 6(9): 348-351 (2011)

BIOINFORMATION

open access

Supplementary material:

Table 1: Frequency of SSR's In Genome and Plasmid

Type of Simple Sequence repeats	Genome		Plasmid	
	N	%	N	%
Mononucleotide	805904	80.26036936	33471	79.5167843
Dinucleotide	131678	13.11387574	5923	14.0712232
Trinucleotide	54001	5.377985723	2083	4.94856627
Tetranucleotide and above	12529	1.247769173	616	1.46342622
Total	1004112	100.000	42093	100.000