SIMPLE SEQUENCE REPEATS IN SPECIFIC GENE GROUPS OF SHIGELLA GENOME

Hosseini Ashraf1,2, Indira Ghosh3, Pramod Khandekar4, Mohammad Hiresh Ayoubian5
1Tehran University of Medical Science, School of Allied Medical Sciences, Tehran, IRAN
2Institute of Bioinformatics and Biotechnology, University of Pune, Pune, INDIA
3School of Information Technology, JNU, New Delhi, INDIA
4Dept. Of Biotechnology, Sinhgad Engineering College, Sinhgad Institute of Technology, Vadguon, Pune, India.
5Dept. Of Biotechnology, University of Pune, Pune, INDIA

ABSTRACT

In this paper we attempt to analyze the phenomenon of simple sequence repeats (SSRs) variations in Clusters of Orthologous Groups of proteins (COGs) and horizontal transfer genes (HGT) of Shigella flexneri. We have performed a detailed comparative study of the distribution of SSRs in different gene clusters. According to our finding SSR elements in Shigella pathogenicity islands (PAIs) are significantly overrepresented than in other gene clusters of Shigella particularly in sflI islands which have implications in Shigella virulence and also in virulence genes of 2 Shigella plasmids. The trinucleotide groups of SSRs, the codon repetitions and the amino acid repeats have been found in this study are subject to; (I) Strong selection of SSRs in PAIs which have important roles in Shigella virulence and (II) important roles of SSRs in determining protein function and genetic development.

Corresponding author: Email: ashosaini@yahoo.co.in; ash-hosseini@tums.ac.ir; Tel: +91-989189995351; Fax: 0982188622533

[1] INTRODUCTION

Simple sequence repeats (SSRs) may provide an evolutionary advantage; they may function as evolutionary tuning knobs by allowing fast adaptation to new environments [1, 2]. Numerous lines of evidence have demonstrated that genomic distribution of simple sequence repeats (SSRs) is nonrandom, presumably because of their effects on chromatin organization, regulation of gene activity, recombination, DNA replication, cell cycle, mismatch repair (MMR) system, etc [3]. Recently, however, many reports have demonstrated that a large number of SSRs are located in transcribed regions of genomes, including protein-coding genes and expressed sequence tags (ESTs) [4], although in general, repeat numbers and total lengths of SSRs in these regions are small [5,6]. Debates over whether SSRs play any functional role in organism development, adaptation, survival, and evolution are never-ending. The currently available information on the location of specific SSRs in known genes and ESTs permits the unraveling of the biological significance of SSR distribution, expansion, and contraction in the functioning of the genes themselves [7].

Prokaryotic and eukaryotic repeat families are clustered to non homologous 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 [8]. In a variety of organisms, it has been demonstrated that microsatellite mutation rates are positively correlated with repeat number [9]. In prokaryotes, strong positive selective pressures are associated with highly mutable microsatellite tracts that control pathogenicity [10].

The presence of SSRs in prokaryotes is rare, but most that do occur are related to pathogenic organisms; their variation in repeat numbers can also cause phenotypic changes [11]. Haemophilus influenzae (Hi), an obligate upper respiratory tract commensal/pathogen, uses phase variation (PV) to adapt to host environment changes. Switching occurs by slippage of SSR repeats within genes coding for virulence molecules. When SSR repeats lie within protein coding regions, UTRs, and introns, any changes by replication slippage and other mutational mechanisms may lead to changes in protein function [12].

Shigella is an important human pathogen, responsible for the majority of cases of endemic bacillary dysentery prevalent in developing nations [13]. Shigellosis is common among children less than five years of age in developing countries and in persons who travel from industrialized to less developed countries [14].
In this paper, we attempt to analyze the phenomenon of SSR variation in clusters of orthologous groups of proteins (COGs) and horizontal transfer genes (HGT) of *Shigella flexneri* as it is common among children in developing countries. We have performed a detailed comparative study of the distribution of SSRs in different gene clusters.

**II] MATERIALS AND METHODS**

**2.1. DNA sequences**


Pseudogenes, plasmid virulence genes, and chromosome virulence genes were determined by using SHIBASE, an integrated database for comparative genomics of *Shigella*, at http://www.mgc.ac.cn/ShiBASE/VFs.htm.

**2.2. Analysis methods**

We used the software developed by Gur-Arie et al [15] to screen the COGs, *Shigella* islands, pseudogenes, and virulence genes of *Shigella* for SSRs with a motif length between 1 and 10 bp and a minimal number of three repeats and an entire SSR array length of at least two. This software can be downloaded from ftp://ftp.genomics.inb.rug.nl/otpath/.

Filter of DNA is done by using http://puma.icb.usp.br/sms/filter_dna.html, for remove non-DNA characters from text including digits and blank spaces from a sequence.

Shuffle nucleic acid sequence has done to produce a randomized sequence with the same overall composition as the original sequence by using sequence shuffling tool downloaded from http://bfc.arl.arizona.edu/resources/online_tools/shuffle.php.

**2.3. Calculation of the expected number of mononucleotide SSRs**

To compare the observed number of SSRs with the expected number, we calculated the expected number of homo-oligomer tracts of *t* bases in a sequence of length *N* using the formula given by De Wachter T

\[
T_1.t = \sum \frac{p_i \cdot (1-p_i) \cdot (N+i-1)}{1-p_i + (N+i-1)}
\]

for *i* = 1 to 4, where *p* is the frequency of each base in the sequence (R).

**[III] RESULTS**

The Using above mentioned computer programs, we analyzed most important gene groups/mechanisms of *Shigella* which have a important role in *Shigella* function and pathogenicity. The frequency of total SSRs (%5.49) and mononucleotide SSRs (%5.28), with a minimal repeat of 3 in *Shigella* islands, is higher than the other clusters. The frequency of dinucleotide SSRs in *Shigella* islands and transcription genes (%0.17) as well as trinucleotide SSRs in translation genes (%0.06) are higher, followed by intracellular trafficking and secretion genes (%0.053). The frequency of total SSRs (%4.23) and mononucleotide SSRs (%4.18) is lower in transport genes Table 1. The ratio of observed mononucleotide SSRs to expected mononucleotide SSRs for A and T is greater than 1 and for C and G smaller than 1, except for IS elements and A8 in cell wall/membrane genes and T8 in translation genes. The ratios of observed mononucleotide SSRs to expected mononucleotide SSRs for A3–A5 and T3–T5 for different gene clusters are nearly identical but increases with increasing motif length differences. The ratio of A and T mononucleotide SSRs with increasing motif length was shown to increase, except for A8 and T8, which decrease. The ratios of A6–A8 of *Shigella* islands are much higher than those of other gene groups. The ratio of observed mononucleotide SSRs to expected mononucleotide SSRs in IS elements is totally different from other gene groups Figure 1. Genes with a high GC content have a lower mononucleotide and dinucleotide SSR density than genes with a low GC content. The numbers of mononucleotide and dinucleotide SSRs are negatively correlated with GC content (Figure 2). The A/T compositions of mononucleotide repeats in all investigated genes are much higher than the C/G compositions of those repeats. Differences were significant by $X^2$ test ($p < 0.0001$) Figure 3. In dinucleotide SSRs the frequency of GC/CG in all investigated genes except *Shigella* islands, are higher while frequency of AT/TA in most of genes are lower. Most difference between frequency of CG/GC and AT/TA has observed in the amino acid transport genes (%62.9 and %4.3 respectively) and least difference has observed in the *Shigella* islands (%23.8 and %15.0 respectively). The frequency of AC/GT dinucleotide repeats is higher in intracellular trafficking and secretion genes and *Shigella* islands than other genes. The frequency of AG/CT dinucleotide repeats is higher in posttranslational modification, protein turnover, chaperones genes and signal transduction mechanisms genes. The frequency of TC/GA dinucleotide repeats is higher in signal transduction mechanisms genes and repair, replication genes. The frequency of TG/CA dinucleotide repeats is higher in repair, recombination, replication genes and *Shigella* islands Figure 4.

The codon repetition of CAG for (Gln) and GCG for (Ala) are most abundant in all gene clusters except translation genes. In translation genes ACG for (Thr) and CGC for (Arg) are the most abundant. Alanine and Arginine repetition are strongly overrepresented in most of the gene clusters. Alanine and serine repetition in the signal transduction mechanism genes as well as alanine and glutamine in the replication recombination and repair genes are strongly overrepresented. Arginine and glutamine in the inter trafficking and secretion genes are overrepresented.

In contrast the frequency of amino acid repeats in *Shigella* islands is nearly identical and there is no overrepresentation of alanine or arginine such as other gene groups. Differences between frequency of amino acid repeats in *Shigella* islands...
and other gene groups were significant by $\chi^2$ test (P<0.001) [Figure–5].
The frequency of mononucleotide SSRs in Sf II island are higher than other SI gene groups and it is lower in Sci islands [Table–2]. The percentage of mono, di, tri and total SSRs $\geq$3bp in mxi-spa genes are more than vir genes in both investigated plasmids [Table–3].

### Table 1. Frequency of SSRs in different gene clusters of *Shigella.f 2a str 301*

| Genes | Length, bp | N | % | N | % | N | % | N | % | N | % |
|-------|------------|---|---|---|---|---|---|---|---|---|---|---|
| IS    | 303097     | 6.6| 13447| 4.43| 13080| 4.31| 322| 0.11| 45| 0.015|
| SI    | 430105     | 9.3| 23620| 5.49| 22731| 5.28| 741| 0.17| 141| 0.032|
| PS    | 363034     | 7.9| 17330| 4.77| 16720| 4.6| 476| 0.13| 131| 0.036|
| RRRR  | 423084     | 9.3| 18387| 4.34| 17587| 4.16| 446| 0.11| 93| 0.022|
| TrS   | 135834     | 3.0| 5976 | 4.4| 5669 | 4.2| 197| 0.15| 75| 0.06 |
| TrC   | 192078     | 4.2| 9021 | 4.7| 8621 | 4.5| 322| 0.17| 74| 0.04 |
| CWMM  | 244116     | 5.4| 11843| 4.85| 11349| 4.65| 384| 0.16| 97| 0.04 |
| ET    | 292344     | 6.3| 13151| 4.5| 12555| 4.3| 446| 0.15| 134| 0.046|
| TR    | 1093803    | 23.7| 46376| 4.23| 44479| 4.18| 1645| 0.15| 468| 0.043|
| InCTS | 95382      | 2.1| 4468 | 4.66| 4405 | 4.6| 107| 0.11| 51| 0.053|
| PTMPTC| 113073     | 2.4| 5167 | 4.57| 4812 | 4.3| 136| 0.12| 58| 0.05 |
| STM   | 137145     | 3.0| 6478 | 4.73| 6232 | 4.53| 191| 0.14| 53| 0.04 |
| rRNA  | 7475       | 0.16| 326 | 4.36| 317 | 4.24| 8| 0.11| 1| 0.013|
| Total | 32821      | 0.71| 1723| 5.25| 1675| 5.1| 43| 0.13| 5| 0.015|

IS: Insertion Sequences; PS: Pseudogenes; SI: Shigella islands; RRR: Repair, recombination, replication; TrS: Translation; TrC: Transcription; CWMM: Cell wall/ Membrane Mechanism; EP: Energy production; TR: Transport Genes; InCTS: Intracellular trafficking and secretion; PTMPTC: Posttranslational modification, protein turnover, chaperones; STM: Signal transduction mechanisms

### Table 2. SSRs in 4 groups of *Shigella* Islands with sizes >1 kb in chromosome of *Shigella.f 2a str 301*

<table>
<thead>
<tr>
<th>Mono nucleotide Repeats</th>
<th>Sci islands 21440bp</th>
<th>ipaH islands 98767bp</th>
<th>SHI-1 &amp; 2 80483bp</th>
<th>Sf II island 28913bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td>Mono</td>
<td>834</td>
<td>3.89</td>
<td>4918</td>
<td>4.98</td>
</tr>
<tr>
<td>Di</td>
<td>30</td>
<td>0.14</td>
<td>138</td>
<td>0.14</td>
</tr>
<tr>
<td>Tri</td>
<td>8</td>
<td>0.037</td>
<td>35</td>
<td>0.035</td>
</tr>
<tr>
<td>Tetra</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>872</td>
<td>4.1</td>
<td>5093</td>
<td>5.16</td>
</tr>
</tbody>
</table>

### Table 3. Frequency of SSRs in Vir genes and Mxi Spa genes of plasmids pCP301 and pSD1_197

<table>
<thead>
<tr>
<th>Plasmids</th>
<th>pCP301</th>
<th>pSD1_197</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virulence genes</td>
<td>Mxi Spa genes 25551bp</td>
<td>Vir genes 32551bp</td>
</tr>
<tr>
<td>N</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td>SSRs</td>
<td>Mono ≥3bp</td>
<td>1571</td>
</tr>
<tr>
<td></td>
<td>Di ≥3bp</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Tri-Hexa ≥3bp</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Total ≥3bp</td>
<td>1653</td>
</tr>
</tbody>
</table>

Fig: 1. Ratio of observed Mononucleotide SSRs /expected Mononucleotide SSRs

Fig: 2. Correlations between mono, di and trinucleotide SSRs with GC% in different gene clusters
**[IV] DISCUSSION**

**4.1. Association of SSRs with protein function**

Although in prokaryotes SSRs are not as abundant as in eukaryotes, most of the SSRs in bacteria are located in virulence genes and/or regulatory regions, and they affect pathogenesis and bacterial adaptive behavior, indicating the signature of natural selection [12, 16]. This hypothesis has been supported by our finding. For instance in our finding SSR in *Shigella* Pathogenicity islands are overrepresented than in other gene groups of *Shigella*, particularly ipaH islands, sf II island and SHI-1&2 which have implications in *Shigella* virulence. Thus, phage-mediated horizontal DNA transfer appears to be one of the major routes by which *Shigella flexneri* gains virulence determinants.

However regarding to overrepresentation of SSRs in ipaH, there is evidence that *S. flexneri* expresses more IpaH within host cells, and the proteins penetrate the host cell nuclei [17]. This, and the fact that all IpaH proteins have a leucine-rich repeat region found in a diverse group of proteins from bacteria and eukaryotes [18], implies that IpaH might be involved in manipulating host gene expression.

Biased distribution of Codon repetition and amino acid repeats...
have been found in different gene groups, suggesting that repeats of these kinds are subject to strong selection. Functional associations of amino acid repeats for such a scenario to be valid, amino acid repeats of this kind must be associated in some way with protein function.

Our data has shown the overrepresentation of SSRs in ipaH genes and mxi- spa genes of Plasmids pCP301 and pSD1_197. While the Ipa proteins are essential for the invasion of epithelial cells, and their secretion is mediated by the proteins encoded at the mxi and spa loci [19, 20], the SSRs overrepresentation indicates opportunity of adaptability under different system environment.

**Fig: 5. Frequency of Codon repetition in different gene groups of Shigella.f 2a str 301**

### 4.2. SSRs in Shigella Pathogenicity islands >1 kb

The Overrepresentation of total SSRs (%5.49) and mononucleotide SSRs (%5.28) in Shigella islands compare with other gene groups (%4.47 and %4.28 respectively), more occurrence of SSRs with increasing size of motif length in Shigella islands compare to other gene groups, differences on amino acid repeats pattern and codon repeats frequency between Shigella islands and other gene groups, high frequency of AT/TA dinucleotide and low frequency of GC/CG dinucleotide SSRs than the other gene clusters, shows differences between genetic structure of Shigella islands (horizontal genes transfer) and other COG genes. These differences may be associated with their evolution or may be generated after integration of PAI-specific DNA regions into the host genome via recombination [21]. It has been speculated that changes in length of repeats in such systems could alter their behavior and therefore contribute to their evolutionary diversification [22, 23], perhaps involving molecular co evolution between proteins [23].

Strongly biased distributions of the all SSR elements in Shigella islands have been found in this study emphasize the importance of SSRs in these PAIs which have important roles in Shigella virulence.

### 4.3. SSRs in specific gene groups located in large Shigella Islands>1 kb

Our investigation of SSRs in Sci islands, ipaH islands, SHI-1 and 2 islands and sf II islands, and their implications on the role of virulence, clearly shows differences between frequencies of total and mononucleotide SSRs and the composition of mononucleotide and dinucleotide SSRs between them, particularly between Sci islands and sf II island. This may be
associated with their evolution because these islands share homology with genes of different phages. For example, chromosomal ipaH islands are originally linked with phage P27. The Sci island possesses a typical structure of PAI—inserted at an asp-tRNA, and ends with an IS629 on the other side. It also carries paralogs of the Salmonella sci CDEF operon of unknown function and of phages P22 and HK620. The Sf II island has been demonstrated to be a lysogenic phage required for the expression of the type II antigen. It appears to be one of the major routes by which S. flexneri gains virulence determinants. Significant differences occur between SSR elements in the Sf II island with other ISs gene groups by X2 test (P=0.01), including overrepresentation of mononucleotide SSRs, a higher frequency of AT/TA mononucleotide SSRs and higher frequency of AT/TA dinucleotide SSRs. This may be associated with gene function. By analyzing the Sf II island we found that this island involves 37 proteins, including 12 hypothetical proteins (%.5.1 SSR), 12 IS elements (%.5.6 SSR), four phage integrases (%.6.0 SSR), two putative glucosyl tranferases (%.6.9 SSR), and seven other genes (%.4.5 SSR). Our investigation has shown that SSR in putative glucosyl tranferases and phage integrases is higher than the others, implying the involvement of these genes in pathogenic activities.

4.4. Correlation of GC content and SSRs

Our data clearly indicate that the GC content of mononucleotide SSRs is highest when the repeat density is lowest. This suggests that there could be other reasons for the tremendous overrepresentation of poly (A) and poly (T) mononucleotide SSRs. It has been suggested 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 G+C content. Indeed, the synthesis of GTP requires an additional NAD compared with AMP, while the synthesis of CTP from UTP requires an additional ATP molecule. In addition, due to its central role in metabolism, ATP is abundantly present in the cell [24, 25].

[V] CONCLUSION

In conclusion, the present study has shown biased distribution of trinucleotide groups of SSRs, Codon repetition and amino acid repeats in different gene clusters in Shigella. Significant differences between SSR patterns in Shigella Pathogenicity islands with other gene groups of Shigella are also manifested. The overrepresentation of SSRs in ipaH genes and Mxi, Spa gene particularly plasmid ipaH genes are correlated with pathogenicity of Shigella.

Strongly biased distributions of the all SSR elements in Shigella islands have been found in this study, emphasize the importance of SSRs in these PAIs which have important roles in Shigella virulence.

Study has suggested that SSRs in different positions of a gene can play important roles in determining protein function, genetic development, and regulation of gene expression.

CONFLICT OF INTERESTS

Authors declare no conflict of interests

FINANCIAL DISCLOSURE

The work was carried out without any financial support

ACKNOWLEDGEMENT

Help and support of Director, research fellows, faculty members and staff of Bioinformatics Centre, University of Pune is gratefully acknowledged.

REFERENCES


