THE NUCLEOTIDE SEQUENCE OF *SHIGELLA FLEXNERI* 1A : A COMMON INDIAN ISOLATE

Suvidya H Ranade, Ashraf Hossani***, D N Deobagkar*, D D Deobagkar*, B A Chopade** and Pramod S Khandekar****

Departments of Chemistry, *Zoology, **Microbiology and ***IBB, University of Pune, Pune.

**** Sinhadadh College of Engineering, Pune

ABSTRACT

We report complete sequences of 2 genes of *S.flexneri* 1a an Indian isolate for the first time. *Shigella* is causative agent of shigellosis or bacillary dysentery. Genomic library was constructed by shotgun approach. Sequencing was carried out using Big Dye terminator chemistry using ABI 3730 48 capillary DNA analyzer. 170 recombinants were subjected to nucleotide sequencing. Sequence data was analyzed, of these 2 clones showed presence of complete genes out of the total clones sequenced. Annotations were done using various bioinformatics tools. Gene Sfo676 on contig SF21B11, 513 bp long codes for a protein 170 aa long with molecular weight of 18836.5 daltons. The protein is 99 % identical to *S. flexneri* 2a 301 and not with any other strain of *Shigella*. It has 7 different sites for phosphorylation, myristoylation and glycosylation. Predicted cellular localization is cytoplasmic membrane. SF0368 is another full-length gene SF0368 on contig SF69C1 is a 312 nucleotide long. It is 103 aa long with molecular weight 11394.0 daltons. Protein is 100% identical to *S. flexneri* 2a 301 and 99% with *S. sonnei* strain 046. The gene shows difference when compared with *S.sonnei* in mono and dinucleotide frequency as well as amino acid composition.

KEY WORDS

Shigella Flexneri 1a, Shotgun sequencing, Cytoplasmic membrane, Shibase.

INTRODUCTION

Shigella infection (shigellosis) is an intestinal disease caused by a family of bacteria called *shigella*. The main sign of shigella infection is diarrhoea, which often is bloody. *Shigella* can be acquired through direct contact with the bacteria in the stool or through contaminated food or by drinking or swimming in contaminated water. Children between ages 2 and 4 are most likely to get shigellosis. *Shigella* infection is far more common in developing nations, especially where there's overcrowding and poor sanitation. *Shigella* are highly invasive in the colon and the rectum, and are able to proliferate in the host cell cytoplasm, triggering an inflammatory reaction. The clinical manifestations of *Shigella* infection vary from short-lasting

watery diarrhea to acute inflammatory bowel disease characterized by fever, intestinal cramp and bloody diarrhea with mucopurulent feces. Currently, no vaccine is available that can provide adequate protection against the many different serotypes of *Shigella*. Since the current preventive and treatment strategies are found inadequate the WHO has placed an anti- *Shigella* vaccine at priority (1, 2).

Shigella species are non-speculating, gram negative, facultative anaerobes. *Shigella* was recognized as a causative agent of bacillary dysentery in 1890s and sub grouped into 4 species: *S.dysenteriae*, *S.flexneri*, *S.boydii* and *S sonnei*. World wide annual episodes due to *shigella* infection are 160 million of which 1.1 million deaths occur in children below 5Yr.of age (3, 4). Out of the 332 complete genome sequences, the sequences available in database for *shigella* species are *S flexneri* 2a strain 301, *S flexneri* 2a strain 2457T, *S.sonnei* strain 046, *S.boyddii* 4 strain 227, *S. dysenteriae*1 strain 197, which are more prevalent in the Western world. There is no information available on *S. flexneri* 1a in database, which is more prevalent in Western and Eastern India. Since the current

Address for Correspondence :

Dr. S H Ranade

Department of Chemistry,

University of Pune, Pune 411007, Maharashtra,

M: 91- 9822617330

E-mail: suvidya@chem.unipune.ernet.in

preventive and treatment strategies are found inadequate the WHO has placed an anti- *Shigella* vaccine at priority (1).

During early 1984, various areas of India, including the eastern region, witnessed an extensive epidemic of bloody dysentery, predominantly caused by multidrug-resistant *S. dysenteriae* type 1, which swept through the districts of West Bengal from north to south (5). Prevalence of *Shigellae* serotypes in Bombay was studied from June 1988 to May 1991 (6). The thirty-four strains of *Shigella* spp. were isolated in Calcutta India in 1998-2000 (7).

It has been suggested that the four species of *Shigella* are so closely related to *E. coli* that all these bacteria could be considered members of a single species (8). They share greater than 90% homology.

The most important components of virulence in *Shigella* infection include adherence, invasiveness and toxigenicity. The genes for these virulence factors reside on either plasmids or the bacterial chromosome or on lysogenic bacteriophages (9). *Shigella* strains have large plasmids that encode many virulence genes (6) especially genes involved in invasion and intracellular movement and these products are known as invasion plasmid-associated antigens. The molecular analysis of the clinical samples in Bangladesh has shown that all *S. dysenteriae* type 1 isolates belong to ribotype pattern A, *S. flexneri* strains produced two ribotypes, C and D. Among the *Shigella* species, *Shigella dysenteriae* is especially known as a cause of bacillary dysentery; this has also been shown with *S. flexneri*. Another virulence factor related to *S. dysenteriae* is its capacity to produce an exotoxin called Shiga toxin. Two

new enterotoxins have recently been described in *S. flexneri* 2a. The first toxin is called *Shigella* enterotoxin 1 (ShET-1), The second enterotoxin, called *Shigella* enterotoxin 2 (ShET-2), is encoded in the *sen* gene. ShET-1 was found in 45% of *S. flexneri* strains but was not found in either *S. sonnei* or *S. dysenteriae*. Studies on pathogenicity showed the high prevalence of ShET-1 and ShET-2 in *S. flexneri* strains and the high prevalence of ShET-2 in *S. sonnei* and *S. dysenteriae* strains isolated from patients with traveler's diarrhea (10). Following table shows comparison of general features of *shigella* species and *E.coli* MG 1655.

As there is no sequence data available for *S. flexneri* 1a, which is more common in India we have chosen this strain for sequencing. We report here two complete gene sequences of *S. flexneri* 1a. Sequencing data can be used for genetic research and for clinical challenge studies. Comparative genomics studies help to identify differences or similarity with other *shigella* species and strains.

MATERIALS AND METHODS

***S. flexneri* 1a** strain was obtained from Dept of Microbiology, Armed Forces Medical College, Dept of Microbiology, Pune, which was characterized microbiologically and biochemically by standard methods given in Bergey's manual (11). Enzymes Sau3 a, BamH1, T4 DNA ligase, as well as other biochemicals like IPTG, X Gal, M13 sequencing primers were procured from Bangalore Genie, Bangalore, India. ABI Big dye terminator chemistry kit was obtained from Lab India. *Shigella flexneri* 1a was grown in Luria Broth at 37°C O/N on orbital shaker.

Table showing comparison of general features of shigella species

Chromosomes	MG1655	Sd197	Sf301	Sb227	Ss046
Total length (bp)	4639675	4369232	4607203	4519823	4825265
No. of total ORFs	4254	4557	4434	4353	4434
No. of pseudogenes	12	285	254	217	210
Average length of ORFs (bp)	954	812	891	894	931
Percentage of CDS (%)	87.3	77.2	80.4	80.5	80.5
G + C content (%)	50.79	51.25	50.89	51.21	51.01
CDS regions (%)	51.85	52.30	51.95	52.30	52.20
RNA genes (%)	54.84	54.71	54.79	54.89	54.99
Intergenic regions (%)	42.28	44.75	46.07	43.92	43.33
No. of rRNA (16S/23S/5S)	7/7/8	7/7/8	7/7/8	7/7/8	7/7/8
No. of transfer RNA	86	85	97	91	97
IS elements (partial copies)	44 (7)	623 (94)	314 (67)	403 (89)	394 (67)

METHOD

(A) Method of shotgun cloning and sequencing:

- *Shigella flexneri* 1a was grown in Luria Broth at 37°C O/N on orbital shaker. Genomic DNA of *S. flexneri* was isolated according to method of Chen and Kuo (12).
- Plasmid isolation was carried out according to the method of Birnboim and Doly (13).
- Genomic DNA of *S. flexneri* 1a was subjected to partial digestion with restriction endonuclease *Sau*3a. Genomic DNA was cleaved into fragments using the enzyme *Sau*3a(1.2U in 40µl reaction). The digestion was set up for different time period from 5-20 min. The digestion product was purified and ethanol precipitated.
- PUC 18 was used as a vector for shotgun cloning. Purified PUC 18 was digested with *Bam*H1 enzyme for 4 hrs. Dephosphorylation of digested PUC 18 was done using CIP(10U in 40µl reaction). The CIP treated digestion product was purified and ethanol precipitated.
- The purified genomic DNA digest as well as the digested vector was visualized on agarose gel using ethidium bromide staining.
- DNA fragments with average size between 1-2 Kb were cloned into *bam* H1 site of plasmid pUC 18. (Fig1).
- The ligation reaction was set up using T4 DNA ligase 5U, 10X cohesive end buffer and vector:insert ratio 1 :3. calculated as :

$$\frac{[\text{ng of vector} \times \text{size of insert(Kb)} / \text{size of vector(Kb)}] \times 3 = \text{ng of inset.}}$$
- The ligation mix was incubated at 15°C for 24 Hrs. The ligation reaction was set up alongwith the blank reaction using uncut vector.
- Competant cells were prepared using *Dh5α* culture by standard CaCl_2 method.
- Transformation was set up using the competent cells and the ligation mixture. Recombinants were selected using IPTG X gal selection method
- The presence of colonies of transformed cells on SOB (ampicillin, Xgal, IPTG) plates were marked next day , no colonies seen with controls.
- White colonies were picked up and inoculated into LBamp.
- Plasmid isolation was carried out using Birnboim and Dolly method. Those clones which showed presence of insert (1-2Kb) on agarose gel electrophoresis as the higher size band than that of PUC 18 vector were considered for sequencing.(shown photograph) (14).
- Sequencing was done using Big Dye terminator chemistry using ABI 3730 48 capillary sequencer .Cycle sequencing was done using 1pmol of M13 (F,R) primers,

300ng plasmid and the reaction ready mix by LabIndia.

- Cycling parameters for cycle sequencing were as follows.

94°C	1 min
96°C	10 sec
50°C	5 sec
60°C	4 min

For 25 cycles.
- Each sample was sequenced three times from both the ends using M13 sequencing primers. 48 Samples were processed in one plate using forward and reverse primers within 120 min by this machine. The best Sequence of all 3 runs was selected depending on Q value.
- The software available at studden package Gap4 was used for vector removal, repeat masking and partial assembly (7).

(B) DNA sequence analysis:

Nucleotide sequence analysis was performed using Big Dye terminator chemistry using ABI 3730 48 capillary DNA analyzer. Annotations were done using various bioinformatics tools like Blast, Gene Mark, Is finder, ShiBASE etc. Length and location of each partial/ complete gene was identified using Blastn and ShiBASE, which is newly, developed software for comparative genomics of *shigella* genomes (3). Six frame ORF was found using ORF finder at NCBI. Various proteomic tools, ProtParam for determination of physicochemical parameters, Prosite for prediction of motif, PSORTb's for analysis of bacterial sub cellular localization of protein, HMMTOP for prediction of transmembrane helices and topology of protein,. The dinucleotide frequency in SF0368 and SF0676 genes of *S. flexneri* 1a *S. flexneri* 2a and *S. sonnei* was calculated using the tool DNA stat available at 123 genomics, and compared the frequencies. The nucleotide sequences as well as amino acid sequences have been deposited in the gene bank at NCBI. (Accession numbers DQ512801, DQ512802).

The details about bioinformatics tools used for sequence annotations are given below.

1) GAP4 is a part of the Staden package of sequence assembly, editing and analysis tools. Performs assembly, contig joining, assembly checking, repeat searching, experiment suggestion, read pair analysis and contig editing. It has graphical views of contigs, templates, readings and traces which all scroll in register. Contig editor searches and experiment suggestion routines use phred confidence values to calculate the confidence of the consensus sequence.

2) Blast: The Basic Local Alignment Search Tool (BLAST) for comparing gene and protein sequences against others in public databases. In bioinformatics, Basic Local Alignment Search Tool, or BLAST, is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold. BLAST will find subsequences in the database which are similar to subsequences in the query. In typical usage, the query sequence is much smaller than the database, e.g., the query may be one thousand nucleotides while the database is several billion nucleotides. The BLAST program was designed by Eugene Myers, Stephen Altschul, Warren Gish, David J. Lipman and Webb Miller at the NIH and was published in J. Mol. Biol. in 1990.

3) ORF finder: Six frame ORF was found using ORF finder at NCBI. The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. Available at NCBI site <http://www.ncbi.nlm.nih.gov/>. Under the heading Hot Spots. One has to paste the sequence in format in the given window or enter the GI or Accession number of the sequence already present in database. This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

4) VecScreen: It is a system for quickly identifying segments of a nucleic acid sequence that may be of vector origin. NCBI developed VecScreen to combat the problem of vector contamination in public sequence databases. This Web page has been designed to help researchers identify and remove any segments of vector origin before sequence analysis or submission.

5) Genmark: GeneMark developed in 1993 was the first gene finding method recognized as an efficient and accurate tool for genome projects. The GeneMark algorithm uses species specific inhomogeneous Markov chain models of protein-coding DNA sequence as well as homogeneous Markov chain models of non-coding DNA. Parameters of the models are estimated from training sets of sequences of known type. The

major step of the algorithm computes a posterior probability of a sequence fragment to carry on a genetic code in one of six possible frames (including three frames in complementary DNA strand) or to be "non-coding". GeneMark.hmm algorithm was designed to improve the gene prediction quality, particularly to improve GeneMark in finding exact gene starts. From 1998 until now GeneMark.hmm and its self-training version, GeneMark S, are the standard tools for gene identification in new prokaryotic genomic sequences, including metagenomes.

6) Is finder: This database provides a list of insertion sequences isolated from eubacteria and archae. It is organised into individual files containing their general features (name, size, origin, family.....) as well as their DNA and potential protein sequences. Although most of the entries have been identified as individual elements, a growing number are included from their description in sequenced bacterial genomes. The search engine permits the retrieval and display of individual and groups of ISs based on a combination of their general features. Two levels of search are available. The simple search option enables the user to sort elements using a limited number of basic items whereas the extensive search offers an additional set of possibilities such as comparisons of the sequences of terminal inverted repeats and a variety of different layout displays. Built in links are provided to: the EMBL sequence database, the NCBI taxonomy database and to the ESF plasmid database. At present, only individual sequences can be downloaded one by one for comparison. An on-line BLAST facility is available.

7) ShiBASE: Length and location of each partial/ complete gene was identified using Blastn and ShiBASE which is newly developed software for comparative genomics of *Shigella* genomes (3). In China scientist have sequenced representative strains of 4 *Shigella* species. They have also done comparative genome hybridization of 43 serotypes of *Shigella* strains by microarray. In order to present this data in better way they developed integrated database for comparative genomics of *Shigella* which is ShiBase. It provides workbench for the comparative genomics studies of *Shigella*, which is one of the most efficient ways to reveal the extremely diversity and dynamics features of *Shigella* genomes. The intra-species comparison of *Shigella* genomes are presented at several different levels in the database, including not only basic genome features, structure of genomes and orthologs ordering, but also metabolic pathways and virulence factors. Moreover, the CGH results on 43 different serotypes of *Shigella* strains were also integrated into ShiBASE.

8) ProtParam: used for determination of physicochemical parameters. This program is available at <http://www.expasy.org/tools/protparam>. It computes various physico-chemical properties that can be deduced from a protein sequence. The parameters computed by ProtParam include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY). Molecular weight and theoretical pI are calculated as in Compute pI/Mw. The amino acid and atomic compositions are self-explanatory

9) PROSITE is a database of protein families and domains. It was used for motifs identification in protein. It consists of entries describing the domains, families and functional sites as well as amino acid patterns, signatures and profiles in them. It is used in identifying possible functions of newly discovered proteins and analysis of known proteins for previously undetermined activity. PROSITE offers tools for protein sequence analysis and motif detection (see sequence motif). It is part of the ExPASy proteomics analysis servers. Available at <http://www.expasy.org/tools/prosite>. PROSITE is complemented by ProRule, a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids. Output includes graphical view and feature detection.

10) PSORTb's: Used for analysis of bacterial sub cellular localization of protein. It is developed by Kenta Nakai in 1991; PSORT is an algorithm which assigns a probable localization site to a protein given an amino acid sequence alone. Originally developed for prediction of protein localization in Gram-negative bacteria, PSORT was expanded into a suite of programs (PSORT, PSORT II, iPSORT) capable of handling proteins from all classes of organisms. The modules may act as a binary predictor, classifying a protein as either belonging or not belonging to a particular localization site, or they may be multi-category, able to assign a protein to one of several localization sites. When analyzing a Gram-negative organism, possible localization sites are: cytoplasm, cytoplasmic membrane, periplasm, outer membrane and extracellular space. Gram-positive localization sites include: cytoplasm, cytoplasmic membrane, cell wall and extracellular space.

11) HMMTOP for prediction of transmembrane helices and topology of protein. It is an automatic server for predicting transmembrane helices and topology of proteins, developed by G.E. Tusnády, at the Institute of Enzymology. In the output HMMTOP prints the number of amino acids and the name of

the predicted sequence in one line beginning with a '>HP:' string, and following by the localization of the N terminal amino acid (IN or OUT), the number of the predicted transmembrane helices and the begin and end positions of each transmembrane helix. The method is based on the hypothesis that the localizations of the transmembrane segments and the topology are determined by the difference in the amino acid distributions in various structural parts of these proteins rather than by specific amino acid compositions of these parts. A hidden Markov model with special architecture was developed to search transmembrane topology corresponding to the maximum likelihood among all the possible topologies of a given protein."

12) DNASTAT: It is a collection of Pascal routines for researchers who develop their own application programs for statistical analysis of DNA and protein sequences. Dynamic and file-based data structures allow users to process sets of sequences by simple loop control without limitations on the number of sequences and their individual sizes. The dinucleotide frequency in SF0368 and SF0676 genes of *S. flexneri* 1a, *S. flexneri* 2a and *S. sonnei* was calculated using the tool DNA stat available at 123 genomics, and compared the frequencies.

RESULTS AND DISCUSSION

Nucleotide sequence analysis was done on 170 recombinants. Total 99.4kb Sequence data was generated from both ends (49.7kb with either forward/reverse). Total 375 protein products have been identified for *S. flexneri* 1a. The average length of

Fig 1: Recombinants showing 1-2 KB inserts

- Well 19 Puc18 vector Plasmid, Well 20 1 Kb ladder, Well 21-32 Clones showing 1-2 KB insert, Well 33 Puc18 vector Plasmid, Well 34 1 Kb ladder
- Well 1-18 Clones showing 1-2 KB insert.

Fig 2: Linear sequence alignment result obtained by ShiBase software for S.flexnei 1a sequence for matches with sequence length above 500 bases

sequence was 550 bases. Total 60 kb data in form of contigs was analyzed through homology search using various bioinformatic tools. Amongst 170 recombinants sequenced and analyzed we obtained 150 partial gene sequences and two complete genes, which are analyzed in details. The nucleotide sequences as well as amino acid sequences have been deposited in the gene bank at NCBI. (Accession numbers DQ512801, DQ512802).

The sequence analyzed so far showed 99% identity with *S. flexneri* 2a 301 and thus was taken as reference strain for localization of the various DNA fragments on genome. Fig 1 shows recombinants having 1-2 Kb fragments. Fig 2 shows few of the identified gene sequences which are homologous to different strains of Shigella, Total AT content 48.2%; total GC content 51.8% was found for the total sequenced data.

Contig No. SF1a25B11, which is 664 bases long showed presence of SF0676 gene. It is 513 bases long. It is present at genomic position 709099 to 709625. The gene is 100% identical to SF 0676 gene of *S. flexneri* 2a 301 and 2457t strains. G+C content of the gene is 42.11%. As seen from the Fig 2 the color of the arrow indicates functional category of ORF, black colour indicates its function is unknown. As the *ShiBASE* and Blastn results shown that it is identical to only *S. flexneri* 2a 301 strain but not with any other strain it could

Table I: Results obtained using various bioinformatic tools for SF0676, SF0368 genes from contig SF1a25B11, SF1a69C1

1. Organism	<i>S.flexneri1a</i>	<i>S.flexneri1a</i>
2. Contig name	SFa25B11	SF1a69C1
3. Gene Symbol	SF0676	SF0368
4. DNA Molecule Name	Chromosome of <i>S. flexneri1a</i>	Chromosome of <i>S. flexneri1a</i>
5. Protein length	170aa	103aa
6. pI	9.76	9.63
7. Putative identification	Conserved hypothetical protein	Conserved hypothetical protein
8. Position	709099 to 709625	379789 to 380306.
9. Gene length	513	312
10. Molecular Weight	18836.5	18836.5
11. % GC	42.11	49.36
12. Aliphatic index	128.47	86.41
13. Hydrophobicity	0.888	-0.565
14. Formula	C ₈₈₄ H ₁₃₈₆ N ₂₁₄ O ₂₂₃ S ₈	C ₄₉₀ H ₈₂₃ N ₁₄₉ O ₁₅₄ S ₄
15. Predicted motifs	58 – 61 CK2_PHOSPHO_SITE, 69 – 74, 73 – 78, 81 – 86 MYRISTYL N-myristoylation site, 95 – 97, 118 – 120 PKC_PHOSPHO_SITE, 110 – 113 ASN_GLYCOSYLATION	14 - 16:TIK PKC_PHOSPHO_SITE 24 – 27, 57 - 60:TspD CK2_PHOSPHO_SITE 69 - 72: NQSW ASN_GLYCOSYLATION 95 - 98: RKnS CAMP_PHOSPHO_SITE 40 - 45:GVdISA MYRISTYL N-myristoylation site
16. Motif Scan	DedA family protein	Ribon helix helix protein copG family
17. Cellular localization	CytoplasmicMembrane	Multiple sites
18. Structure prediction	• N-terminus: OUT • Number of transmembrane helices: 5 Transmembrane helices: 4-27 34-53 62-85 112-136 145-169	Ribbon-helix-helix domain, super family SSF47598

Table 2: Comparison of SF 0368 on contig SF1a69C1 of *S.flexneri 1a* with those of *S.flexneri 2a* and *S.Sonnei*

	<i>S. flexneri 1a</i>	<i>S. sonnei 046</i>	<i>S. flexneri 2a</i>
Gene name	SF0368	SS00410	SF0368
Location		379105..379416	380283-379972
Amino acid replaced	7 A, 9 Y	7 T, 9 P	no
% Identity with S flexneri 2a/S.sonnei		98%	100%
GC%	49.36	48.06	49.36
Dinucleotides with similar freq	GT AC CC	GT AC CC	Same as 1a
Dinucleotide with different frequency	GA, AG, AT, TA, TT, TC, CA GG, GC, AA, TG, CG, CT	GA, AG, AT, TA, TT, TC, CA GG, GC, AA, TG, CG, CT	—
Mononucleotide variation	G 24.68% A 30.77% T 19.87% C 24.68%	G 23.33% A 31.11% T 20.83% C 24.72%	Same as 1a

be a strain specific gene. It codes for conserved hypothetical protein, which is 170 aa long. There is a substitution of one amino acid at first position, which is M to L. The protein is 99 % identical to *S. flexneri 2a* 301 and not with any other strain of *Shigella*. It has 7 different sites for phosphorylation, myristoylation and glycosylation. The protein is rich in aliphatic Ile, Leu and neutral hydroxyl Ser amino acid residues. Motif scans results show that it belongs to DedA family of proteins. This family combines the DedA related proteins and YIAN/YGIK family. Members of this family are not functionally characterized. The protein has maximum i.e.81 hydrophobic and 52 hydrophilic residues of the total 170 amino acids. Table 1 summarizes details of the result. There was not a significant difference in the dinucleotide frequency of gene Sf0676 of *S. flexneri 1a* and *2a*.

SF0368 is another full-length gene identified in a contig SF1a69C1 that was 580 bases long. Actual gene length is a 312 bp. The gene is 100% identical with SF0368 of *S. flexneri 2a* 301 and 98% to *S. sonnei strain 046*. The gene is present at genomic position 379789 to 380306. G+C content is 49.36%. It codes for a conserved hypothetical protein, which is 301 amino acid long. It is 100% identical to that of *S. flexneri 2a* 301 and 98% with *S. sonnei strain 046*. There is a substitution of amino acids at position 7 and 9 from A to T and Y to P with respect to *S. sonnei*. There is inter-strain difference in the SF0368 gene sequence as indicated by the dinucleotide frequency variation. Between *S. flexneri 1a*, *S. flexneri 2a* 301 and *S. sonnei*. The protein is rich in aliphatic Ala, Leu and basic Lys amino acids residues. The protein has maximum i.e.53 hydrophilic residues and 41 hydrophobic residues of

the total 103. Motif scan shows that the protein has Ribbon-helix-helix domain, super family SSF47598. (Table 2). Such domains are found in several bacterial and phage repressors, including the Escherichia coli methionine repressor (MetJ), the bacterial plasmid-encoded repressors CopG (IPR002145), the bacterial omega transcription repressor and the phage repressors Arc and Mnt (8, 15, 16).

Fig 3: Variation in dinucleotide composition of SF 0368 and similar gene found in *S. sonnei* and *S. flexneri 2a* 301

The comparative analysis of gene sequence SF0368 was done. As the blast results shows hit with the corresponding gene from three different strains, these were taken for comparison. The graph shown dinucleotide comparison for Sf0368 of *S. flexneri 1a* with *S. flexneri 2a* and *S. sonnei*. Except GG, GT, AC, TA, CT and CC all other dinucleotides show different frequency of occurrence in *S sonnei* than the other two strains. Which indicated that these 6 dinucleotides are conserved during strain differentiation.

The comparative analysis of gene sequence SF0368 was done. As the blast results shows hit with the corresponding gene from three different strains, these were taken for comparison. The graph shown dinucleotide comparison for Sf0368 of *S. flexneri* 1a with *S. flexneri* 2a and *S. sonnei*. Except GG, GT, AC, TA, CT and CC all other dinucleotides show different frequency of occurrence in *S. sonnei* than the other two strains. Which indicated that these 6 dinucleotides are conserved during strain differentiation. GA, AG, AT, TA, TT, TC, CA dinucleotides in Sf 0368 gene of *S. sonnei* showed higher frequency of occurrence than that of *S. flexneri* 1a and 2a. GG, GC, AA, TG, CG, CT dinucleotides in the same gene of *S. sonnei* showed lower frequency of occurrence than that of *S. flexneri* 1a and 2a. GT AC CC dinucleotides showed equal frequency of occurrence in all the three strains of shigella. There is no difference between dinucleotide frequency of SF0368 of *S. flexneri* 1a and 2a since the gene sequences are 100% identical. Table 2 and Fig 3 show the comparative results of all 3.

The results indicate that the sequence of *S. flexneri* 1a is very much identical to *S. flexneri* 2a 301 strains. Both the genes identified code for conserved hypothetical proteins. For *S. flexneri* 2a 301 453 (10.8%) genes belong to such category. Sf0676 does not show the presence in *S. sonnei* but SF0368 does. This indicates that SF0676 can be a strain specific gene. We have generated a sequence data for the two genes and their products their function is still unknown. This information has been submitted in a database. The prediction of topology or structure or basic physicochemical properties can give us clues which will be useful not only for function prediction but also for studying the gene expression. As the more and more data will be available it will help to study several aspects of *shigella* genomics, Pathogenicity and prevention of this strain, which is common in India.

ACKNOWLEDGEMENT

Work is supported by grant from Potential for Excellence Programme, UGC; University Of Pune.

REFERENCES

- Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, et al. Genome sequence of *Shigella flexneri* 2a: insights into Pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Research* 2002; 30 (20): 4432-41.
- Wei J, Goldberg MB, Burland V, Venkatesan MM, Deng W. Complete Genome Sequence and Comparative Genomics of *Shigella flexneri* Serotype 2a Strain 2457T. *Infect Immun* 2003; 71(5): 2775-86.
- Yang J, Chen L, Yu J, Sun L, Jin Q. Shibase: integrated software for comparative genomics of *Shigella*. *Nucleic Acid Research* 2006; 34: D398-D401.
- Hossain MA, Rahman M, Ahmed QS, Malek MA, Sack RB, Albert MJ. Increasing frequency of mecillinam-resistant *shigella* isolates in urban Dhaka and rural Matlab, Bangladesh: a 6 year observation, *J Antimicrobial Chemotherapy* 1998; 42: 99-102.
- Dutta S, Ghosh A, Ghosh K, Dutta D, Bhattacharya SK, Nair GB, et al. Newly emerged multiple-antibiotic-resistant *Shigella dysenteriae* type 1 strains in and around Kolkata, India, are clonal. *J Clin Microbiol* 2003; 41: 5833-40.
- Sonawala M, Saraswathi K, Deodhar LP., Serogroup prevalence of *Shigella* in Bombay. *J Postgraduate Med* 1995; 41 (4): 104-6.
- Bonfield JK, Smith KF, Staden R. A new DNA sequence assembly program, *Nucl. Acids Res* 1995; 23: 4992-9.
- Weissbach H, Brot N. Regulation of methionine synthesis in *Escherichia coli*. *Mol Microbiol* 1991; 5: 1593-7.
- Faruque SM, Khan R, Kamruzzaman M, Yamasaki S, Ahmad QS, Azim T, et al. Isolation of *Shigella dysenteriae* Type 1 and *S. flexneri* Strains from Surface Waters in Bangladesh: Comparative Molecular Analysis of Environmental *Shigella* Isolates versus Clinical Strains. *Appl Environ Microbiol* 2002; 68 (8): 3908-13.
- Vargas M, Gascon J, Anta MTJD, Vila J. Prevalence of *Shigella* Enterotoxins 1 and 2 among *Shigella* Strains Isolated from Patients with Traveler's Diarrhoea. *Clin Microbiol* 1999; 37(11): 3608-11.
- Garrity GM, Brenner DJ, Krieg NR, Staley JT. Science - pages Bergey's Manual of systematic bacteriology, 2005, 1368 pages
- Chen WP, Kuo TT. A simple and rapid method for the preparation of gram-negative bacterial genomic DNA. *Nucleic Acids Res* 1993; 21 (9): 2260.
- Birnboim H, Doly J. A rapid extraction procedure for screening of recombinant plasmid DNA. *Nucleic Acids Res* 1979; 7: 1513-25.
- Sambrook J, Fritsch EF, Maniatis T. Molecular Cloning: a Laboratory Manual, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory, 1989.
- Murayama K, Orth P, De la hoz AB, Alonso JC, Saenger W. Crystal structure of omega transcriptional repressor encoded by *Streptococcus pyogenes* plasmid pSM19035 at 1.5 Å resolution. *J Mol Biol* 2001; 314: 789-96.
- Schildbach JF, Karzai AW, Raumann BE, Sauer RT. Origins of DNA-binding specificity: role of protein contacts with the DNA backbone. *Proc Natl Acad Sci USA* 1999; 96: 811-7.