BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(24)

Comparison of simple sequence repeats in Staphylococcus strains using *in-silico* approach

Sunil Thorat[1]* & Prashant Thakare[2]

[1]Institute of Bioresources and Sustainable Development, Imphal – 795001, Manipur, India; [2]Department of Biotechnology, S.G.B. Amravati University, Amravati – 444602, Maharashtra, India; Thorat Sunil – Email: sunilsthorat.ibsd@nic.in; *Corresponding author

Received November 15, 2012; Accepted November 16, 2012; Published December 08, 2012

Abstract:
Staphylococci are Gram-positive bacteria which play an important role in infectious disease and are major causes of community-acquired and hospital-acquired infections. Strains of *Staphylococcus aureus* are reported as genomically and phenotypically highly heterogeneous; hence *in-silico* based comparison of genomic data on simple sequence repeats may provide valuable information for understanding the pathogenicity and control measures. This study determined the distribution of a specific group of Simple Sequence Repeats (SSRs), in genome sequences of six Staphylococcus strains (*Staphylococcus aureus* COL, *S.aureus* MRSA252, *S.aureus* MSSA476, *S.aureus* Mu50, *S.aureus* MW2, *S.aureus* N315) and plasmid sequences of four Staphylococcus strains (*Staphylococcus aureus* COL pT181, *Staphylococcus aureus* MSSA pSAS, *Staphylococcus aureus* VRSAp, *Staphylococcus aureus*, *Staphylococcus aureus* pN315 DNA) downloaded from the GenBank database for identifying abundance, distribution and composition of SSRs. The data obtained in the present study shows that (i) a large number of tandem repeats are distributed throughout the genome and plasmid sequences. (ii) Number of mononucleotide SSRs decreased rapidly with increase in size of repeat unit. (iii) Total frequency of SSRs in plasmid regions is less than genomic regions. (iv) In all investigated strains, ratios of AT/TA repeats are dominating over GC/CG repeats in genomics as well as plasmid sequences, and (v) Dinucleotide combination of AT is dominated in all the six Staphylococcus genome sequences.

Background:
Staphylococcus is a genus of Gram-positive bacteria, a major human pathogen. The Staphylococcus genus includes at least forty species, of which; nine have two subspecies and one has three subspecies. Most are harmless and reside normally on the skin and mucous membranes of humans and other organisms. Found worldwide, they are a small component of soil microbial flora. The continuous emergence and spread of antibiotic resistant strains (e.g. MRSA, VRSA) leads to limited treatment options. Being a commensal organism living asymptomatically in the nasal cavities of a large number of the human population, *S.aureus* is also responsible for a raft of different infections that range in both anatomical site and severity. These infections are assisted by a vast array of different virulence factors like adhesins, invasins, toxins and modulins, which enables evasion of host immune responses and also contribute to colonization, dissemination, tissue damage and transmission. Though some infections are superficial and self-limiting, *S.aureus* is also responsible for serious invasive diseases. *S.aureus* is a leading cause of sepsis and infective endocarditis. Colonization of the heart and subsequent formation of vegetations involves a number of complex interactions **[1]**.

The first bacterial species in which tandem repeats were identified was *Mycobacterium tuberculosis*, being described as mycobacterial interspersed repeat units **[2, 3]**. The sequences of six *Staphylococcus aureus* genomes; COL, MRSA252, MSSA476, Mu50, MW2 and N315 were studied along with four *Staphylococcus aureus* plasmids; COL pT181, MSSA pSAS, VRSAp, pN315 DNA which were downloaded from the GeneBank database. Five strains (COL, MRSA252, Mu50, MW2 and N315) are MRSA, while MSSA476 is methicillin sensitive. N315 and Mu50 are closely related strains that are hospital-acquired and vancomycin-intermediate resistance respectively, while MRSA252 is an epidemic hospital strain. MW2 is a hypervirulent community strain. The sequencing of these

 © 2012 Biomedical Informatics

genomes has allowed us to identify novel multiple tandem repeats that are similar to MIRUs in *M.tuberculosis*, and the present study describes the nature and distribution of these repeats, and their potential as a novel tool for understanding the micro-evolution of hospital MRSA.

Microsatellites are also called simple sequence repeats (SSRs) or short tandem repeats (STRs) which are a group of tandem repeated sequences. SSR are comprised of mono-, di-, tri-, tetra-, penta-, or hexa-nucleotide units and are widely present in plant and animal genomes. These repeats can be either perfect tandem repeats or interrupted by several non-repeat nucleotides or compound repeats called compound SSRs **[4]**. These sequences experience frequent mutations that alter the number of repeats. The distribution of particular motif classes within a genome can vary substantially among different species **[5, 6]**. 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 **[7]**. SSRs are highly polymorphic characterized by high rates of insertion and deletion (INDEL) mutations of their repeat units. INDELs of repeat units in a SSR arise due to slipped strand mispairing during the process of DNA replication. Slippage on the template strand tends to contraction of SSRs whereas slippage on the growing strand manifests into expansion of SSRs. The bias in SSRs either toward expansion or contraction is referred to as the directionality of SSR evolution. Ever since it became known that several hereditary diseases are associated with expansion of triplet repeats and colon cancer is associated with instability of certain mono and di-nucleotide repeats, there has been a lot of interest on the discovery of the mechanisms behind directionality of SSR mutations. Despite some investigations into the directional evolution of the SSRs, there is still an insufficient understanding of the factors influencing directionality of SSR mutations **[8]**.

SSRs have been extensively studied in eukaryote genomes and are well-established targets for pedigree analysis **[9]**; though there is insufficient information about microsatellites in simple organisms **[10]**. Bacterial SSR-type DNA can be divided into four main categories. First, dispersed repeat motifs that generally do not occur in tandem. These are repeats that occur throughout genomes of a mass of microorganisms, and are sometimes organized in tandem. A second class is formed by the homopolymeric tracts. Multimers of one of the four nucleotides that are frequently encountered in the genome, for instance the homogenous stretches of *S.cerevisiae* have occurred for as much as 42 nucleotides. The third category is short-motif SSRs with repeat units differing from 2 to 6 bases, this class of repeats are more likely to unit number variation at a given locus. Mostly, when these short-motif repeats are located within genes and are not 3 or 6 nucleotides long, they are likely to upset the coding potential of a given transcript. Fourth, the repeats of more than 8 nucleotides per unit, form a separate category. In addition, the longer repeat unit, the greater the chance that point mutations will be introduced **[11]**. The pathogenic phenotypes caused by unstable tandem repeats have been described in human beings. However, a few notable examples of diseases caused by repeats have been described in other species. Canine epilepsy is reported to be caused by expansion of 12-nucleotide repeats in the single exon of the *Epm 2b* gene **[12]**.

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 **[13-17]**. The present study is an attempt to analyze distribution and composition of SSRs among whole genomes of six strains of *Staphylococcus aureus*.

Methodology:

Tools for analysis of SSRs

A tool for Simple Sequence Repeats was developed to find out the short tandem repeats from the input sequences, the SSR tool is hosted on a local webserver i.e. *wampserver* which has a combination of Apache, MySQL and PHP to run the scripts for analysis. The front-end of SSR tool has Perl based links for di-nucleotides, tri-nucleotides, tetra-nucleotides, penta-nucleotides and hexa-nucleotides. This tool is used to find out the number of nucleotide repeats of all the possible (mono, di, tri, tetra and pentanucleotide) combinations out of input nucleotide sequence. For combined analysis of all the five repeats; a separate script is developed where an input sequence is screened for repeats from di- to penta-nucleotides. A script for finding repeats for selective combinations for all the five SSR is also developed where a user can select a particular repeat (di-, tri-, etc.) for selective combinations (di-AA, tri-ATC, etc.). This program can also analyse whole genome sequences. Apart from finding occurrences of all possible combinations of repeats, graphical presentation along with numerical data for research analysis can be viewed using SSR tool.

Mononucleotide	AAAA
Dinucleotide	CACACACA
Trinucleotide	ATGATGATGATG
Tetranucleotide	GTATGTATGTATGTAT
Pentanucleotide	CGTAGCGTAGCGTAGCGTAG

For the present study, we have used two tools, a webserver based Simple Sequence Repeats (SSR) tool as mentioned above and software developed by Gur-Arie (Ssr.exe) downloaded from (ftp://ftp.technion.ac.il/pub/supported/biotech) to screen the genomes and plasmids of the selected Staphylococcus strains for comparative analysis. Using web based SSR, possible combinations of repeats were searched from the input sequence and results were displayed with its frequency occurrences. On comparing the repeats of six genomes and four plasmids of Staphylococcus strains, the result showed frequency of particular repeat conserved in all the selected genomes and plasmids and also showed which particular combination occur the most. The offline tool Ssr.exe was used for extensive study by setting parameters with 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 10bp; records motif, repeat number and genomic location and reports the results in an output file.

DNA sequences

The whole genome sequences of *Staphylococcus aureus* COL (NC_002951), *S.aureus* MRSA252 (NC_002952), *S.aureus*

© 2012 Biomedical Informatics

MSSA476 (NC_002953), *S.aureus* Mu50 (NC_002758), *S.aureus* MW2 (NC_003923) and *S.aureus* N315 (NC_002745) were downloaded from the GeneBank database. Similarly plasmid sequences of *Staphylococcus aureus* COL pT181 (NC_006629), *Staphylococcus aureus* MSSA pSAS (BX571858), *Staphylococcus aureus* VRSAp (NC_002774) and *Staphylococcus aureus* pN315 DNA (AP003139) were downloaded from NCBI GenBank database.

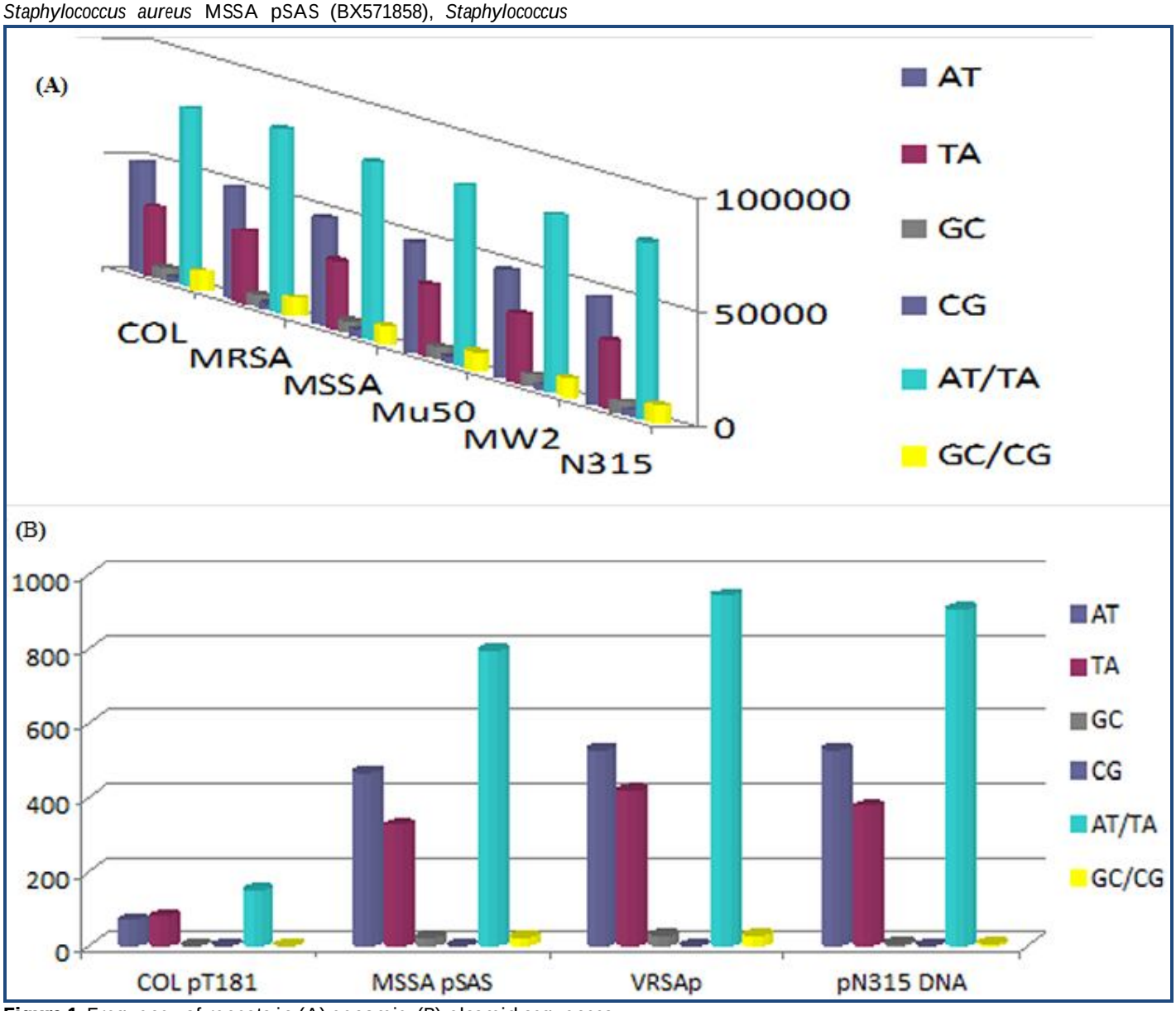

Figure 1: Frequency of repeats in (A) genomic; (B) plasmid sequences.

Results and Discussion:

From the investigated six whole genomes and four plasmids of *Staphylococcus aureus* strains, the observations were as under: (1) Large numbers of Simple Sequence Repeats were found to be scattered in whole genome and plasmid sequences of *Staphylococcus aureus*. Shorter repeats were found to be more than longer repeats. High density of repeats was occurred in genomic sequences; (2) The strain *Staphylococcus aureus* MRSA252 and VRSAp shows high density and more percentage of SSR in genomic and plasmid sequences. The SSR mononucleotide repeats were found to be large in number followed by dinucleotide, trinucleotide and higher motif repeats; (3) The number and percentage of repeats obtained were comparatively more in genomic regions as compare to plasmid sequences of all the studied strains of Staphylococcus; (4) Number of mononucleotide SSRs decreased rapidly with increase in size of repeat unit in all six genomic and four plasmid sequences of Staphylococcus; (5) Total frequency of SSRs in plasmid regions is less than genomic regions; (6) In all investigated strains, ratios of AT/TA repeats were overrepresented over GC/CG repeats in genomics as well as plasmid sequences; (7) Dinucleotide combination of AT dominated all the six Staphylococcus genome sequences.

For the comparative genomic analysis of Staphylococcus genomes and plasmids; we searched for SSRs with minimal number of repeats 2 and motif length from 1 to 10 base pairs. The length of whole SSR array was multiples of 2 *ie*. 2*1 =2. No significantly large SSR with mononucleotide motif was identified among the studied genomic and plasmid sequences.

© 2012 Biomedical Informatics

The longest repeat in genomic sequences was T_{11} and $A_{10,}$ which showed similarity with hypothetical proteins and transposase genes. The over-representation of nucleotides was tested among six genomic and four plasmid sequences and the results **Table 1 (see supplementary material)** showed that AT/TA repeats were dominating.

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 environmental changes cause because of 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 a change in activity of additionally relevant genes and further relaxation of stress by adapting to changed environment **[18]**. The overrepresentation of A and T mononucleotide SSR can be explained by different ways like slipped strand mispairing, which 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 **[19]**. Longest SSR in sequences was 11bp for T and 10bp for A, whereas plasmids of selected four strains showed 8bp for A, 8bp for T, 6bp for G, and 6bp for T. There is overrepresentation of A and T mononucleotide SSR in plasmids as well as genomic sequences. The frequency of AT/TA dinucleotide repeats was higher in genomic sequences as compared to plasmids **(Figure 1)**. 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 motifs containing predominantly A and T are found to be over represented in genome. Whereas, the motifs containing predominantly G and C, are found to be over represented in plasmids except in plasmid of strain COL pT181. The over representation of poly (A) and poly (T) mononucleotide repeats in all the Staphylococcus strains could 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. 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 mono nucleotide repeats were over represented in plasmids whereas penta-nucleotide repeats are slightly over represented in genome. The penta nucleotide repeats are present more in genome.

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 and plasmid of genes of *Staphylococcus aureus* 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. Though SSRs play an important role in the dynamics of eukaryotes, their presence in bacteria is rare, and mostly reduced to pathogenic organisms. The results found from the present study are difficult to analyse and predict the instability of such small SSRs from sequence alone. Nevertheless, these results can be useful for further experimental studies.

References:

[1] Edwards AM *et al*. *PLoS Pathog*. 2010 **6**: 6 [PMID: 20585570]
[2] Supply P *et al*. *Mol Microbiol*. 1997 **991**: 1003 [PMID: 9426136]
[3] Supply P *et al*. *Mol Microbiol*. 2000 **36**: 762 [PMID: 10844663]
[4] Xin D *et al*. *Mol Biol Rep*. 2012 **39**: 9047 [PMID: 22744420]
[5] Levinson G *et al*. *Mol Biol Evol*. 1987 **4**: 203 [PMID: 3328815]
[6] Fodon JW *et al*. *Trends Neurosci*. 2008 **7**: 328 [PMID: 18550185]
[7] Jeffreys AJ *et al*. *Biotechnology.* 1985 **24**: 467 [PMID: 1422054]
[8] Kumar P *et al*. *J Mol Evol*. 2012 **74**: 127 [PMID: 22415400]
[9] Jeffreys AJ *et al*. *Am J Hum Genet*. 1986 **39**: 11 [PMID: 3019128]
[10] Field D *et al*. *Proc Biol Sci*. 1996 **263**: 209 [PMID: 8728984]
[11] Belkum A *et al*. *Microbiol Mol Biol Rev*. 1998 **62**: 275 [PMID: PMC98915]
[12] Gemayel R *et al*. *Annu Rev Genet*. 2010 **44**: 445 [PMID: 20809801]
[13] Cox R *et al*. *Proc Natl Acad Sci USA*. 1997 **94**: 5237 [PMID: 9144221]
[14] Field D *et al*. *Proc Natl Acad Sci USA*. 1998 **95**: 1647 [PMID: 9465070]
[15] Gur-Arie R *et al*. *Genome Res*. 2000 **10**: 62 [PMID: 10645951]
[16] Coenye T *et al*. *BMC Genomics.* 2003 **4**: 10 [PMID: 12697060]
[17] Yang J *et al*. *Gene.* 2003 **322**: 85 [PMID: 14644500]
[18] Trifonov EN, *Ann N Y Acad Sci.* 1999 **870**: 330 [PMID: 10415494]
[19] Rocha EP *et al*. *Nucleic Acids Res*. 2002 **30**: 1886 [PMID: 11972324]

Edited by P Kangueane

Citation: Thorat & Thakare, Bioinformation 8(24): 1182-1186 (2012)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited

BIOINFORMATION

open access

Supplementary material:

Table 1: Frequency of SSRs in genomic and plasmid of Staphylococcus strains.

S. aureus strain	Length (bp)	Mononucleotide repeats					
		A	T	G	C	AT/TA	GC/CG
COL	2809422	531230	533764	131207	129213	78470	8387
MRSA252	2902619	545161	556828	133374	136131	80933	8638
MSSA476	2799802	528429	532723	130962	129417	77826	8416
Mu50	2878040	541945	548420	134024	133895	79791	8508
MW2	2820462	529682	539670	130255	131624	78557	8457
N315	2814816	528833	538350	130166	131467	78204	8316
COL pT181	4440	975	879	252	137	151	2
MSSA pSAS	20652	4469	4339	885	714	793	24
VRSAp	25107	4582	5750	900	1083	941	30
pN315 DNA	24653	5405	5081	1131	868	904	4

© 2012 Biomedical Informatics