BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 6(5)

Comparison and correlation of Simple Sequence Repeats distribution in genomes of *Brucella* species

Jangampalli Adi Pradeep Kiran, Veeraraghavulu Praveen Chakravarthi, Yellapu Nanda Kumar, Somesula Swapna Rekha, Srinivasan Shanthi Kruti, Matcha Bhaskar*

Division of Animal Biotechnology, Department of Zoology, S. V. University, Tirupati-517502, Andhra Pradesh, India; Matcha Bhaskar - Email: matchabhaskar2010@gmail.com; *Corresponding author

Received May 03, 2011; Accepted May 07, 2011; Published May 26, 2011

Abstract:
Computational genomics is one of the important tools to understand the distribution of closely related genomes including simple sequence repeats (SSRs) in an organism, which gives valuable information regarding genetic variations. The central objective of the present study was to screen the SSRs distributed in coding and non-coding regions among different human *Brucella* species which are involved in a range of pathological disorders. Computational analysis of the SSRs in the *Brucella* indicates few deviations from expected random models. Statistical analysis also reveals that tri-nucleotide SSRs are overrepresented and tetra-nucleotide SSRs underrepresented in *Brucella* genomes. From the data, it can be suggested that over expressed tri-nucleotide SSRs in genomic and coding regions might be responsible in the generation of functional variation of proteins expressed which in turn may lead to different pathogenicity, virulence determinants, stress response genes, transcription regulators and host adaptation proteins of *Brucella* genomes.

Keywords: simple sequence repeats, overrepresented, underrepresented, *Brucella* genomes.

Abbreviations: SSRs = Simple Sequence Repeats; ORFs = Open Reading Frames.

Background:
Brucellosis is one of the bacterial zoonoses caused by organisms belonging to the genus *Brucella*, gram-negative, non-spore-forming, facultative, intracellular bacteria. The *Brucella* genome consists of two circular chromosomes without plasmids, suggesting a remarkable difference compared to the single chromosome of many bacteria **[1]**. Among different types of *Brucella* species, *B. melitensis* 16M, *B. suis*, *B. abortus* S19 and *B.canis* ATCC23365 is the common human pathogenic species which are mainly involved in osteoarthritis, heart problems, and several neurological disorders **[2]**. It is worth to notify that the DNA sequences of these *Brucella* species share greater than 90% identity **[3]** and relatively, a small number of differences are responsible for the host preference and virulence restriction of *Brucella* species **[4]**. Many genome sequence projects of *Brucella* species provided tremendous information to decipher the mechanisms underlying *Brucella* pathogenicity **[5]**; however, the potential information at greater resolution by expressing different pathogenicity of closely related genomes like *Brucella* species are little attracted.

Microsatellites or simple sequence repeats are the DNA regions with tandemly repeated bases. They are present in both coding and non-coding regions. It is well known that, variations in the repeat numbers of microsatellites in the coding regions leads to premature termination and frame shift mutations, which in turn causes drastic changes in the gene products **[6, 7, 8, 9]**. The number of repeats at a locus can change by mutation and the rate of mutation depends on the number of tandem units within the repeat **[10]**. Some microsatellites occurring in flanking regions of coding sequences are believed to play significant roles in regulation of gene expression by forming various DNA secondary structures and offering a mechanism of unwinding **[11]**. The variations in SSR numbers can alter the spacing between structurally important domains like the -35 and -10 promoter regions and affect promoter strength. Further, the promoter strength is also affected by unusual repeat structures which have ability to alter normal pattern of DNA **[12]**. SSRs also interfere with replication elongation **[13]** and also integrity of open reading frames **[14]**. SSRs are present virtually in genome and also exhibits polymorphism **[15]**. The primary cause of microsatellite polymorphism is thought to be strand slippage during DNA replication **[16]**. These microsatellites provide a framework for crucial genetic rearrangements with their reversible frame-shift mutations that can confer a certain degree of selective advantage of pathogenic bacteria. Although these microsatellites act as gene regulators, the loss or gain of repeats in the promoter region can regulate transcriptional activity **[17]**. The influence of SSRs on gene regulation, transcription and protein function typically depends on the number of repeats, while mutations that add or subtract repeat units are both frequent and reversible. SSRs have a major role in generating the genetic variation underlying adaptive evolution **[18]**. Despite the wealth of scientific studies carried out over the years on the pathogenicity of *Brucella*, there are still fundamental gaps in knowledge related to the distribution of SSRs among different *Brucella* species with closely related genome. Considering the facts that a) *Brucella* species viz., *B. melitensis* 16M, *B. suis*, *B. abortus*S19 and *B. canis*ATCC23365, are involved in many hazardous health effects in humans **[1]**, b) these species share DNA sequences of about 90% similarities **[2]** c) SSRs are primarily involved in virulence and pathogenic nature of bacteria **[5]** and d) studies related to distribution pattern of SSRs are little exploited in *Brucella* species, the present study was an attempt to

understand the importance and influence on SSRs role in generating genomic diversity among the closely related genomes of human pathogenic *Brucella* species.

Methodology:

DNA Sequences:

The genome sequence of *Brucella melitensis*16M NC_003317, NC_003318, *Brucella suis* NC_004310, NC_004311, *Brucella abortus* S19 NC_010742, NC_010740 and *Brucella canis*ATCC23365 NC_010104, NC_01010 were downloaded from database www.ncbi.nlm.nih.gov **(Table 1 see Supplementary material)**, and the coding sequences of respective genomes were generated from the AMIGene tool (http://www.genoscope.cns.fr/agc/tools/amigene). The expected count was obtained by 10 randomized versions of Homogeneous Bernoulli or Markov models which are widely used in the analyses of DNA sequences.

Analysis of SSRs:

MISA tool used to screen the four genomic and coding sequences of *Brucella* and executes different class of SSRs from 1-4 (mono, di, tri and tetra) with different combinations. These data was used for the analysis of microsatellites with some modifications are made to accommodate for larger Sequences. To determine whether the observed SSR frequencies of a given motif length and repeat number occurred as expected by chance and comparing the mean frequencies of observed or randomly occurred genomes. Ten randomized sequences were generated for each genome with "RSAT" Regulatory Sequence Analysis Tool (http://rsat.ccb.sickkids.ca/). The four *Brucella* Coding sequences were generated using the AMI Gene tool from complete genomes of each species and statistical data shows the overrepresented or underrepresented SSRs types among the four genomes.

Statistical Analysis:

One-way analysis of variance ANOVA was done for mean O/E ratios of different class of SSRs among the four species of *Brucella* genomes. By using SPSS 16.0.1 to determine the distribution of SSRs reports motif, motif length, and repeat number of genomic and coding regions of all SSRs of the four species of *Brucella* genomes.

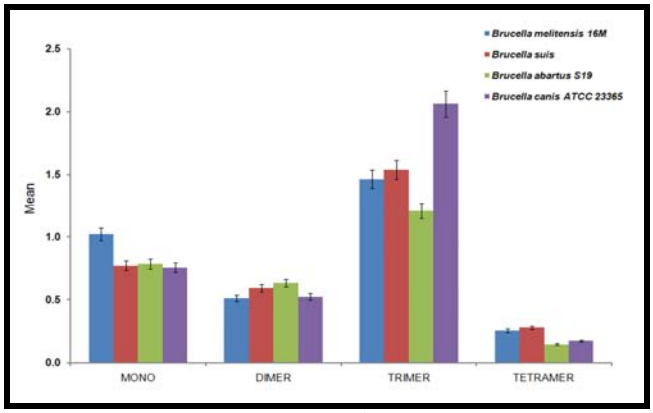

Figure 1: Bars represent the abundance of SSRs in the genomic regions of Brucella species.

Results:

The findings of the present study revealed that there are different types of nucleotides distributed among the genomes of selected *Brucella sp*. Four different types of nucleotides *viz.*, mono-, di-, tri-and tetra-nucleotides were determined by using the computational tools such as MISA, AMIGene and RSAT. Further, comparison of the index of nonrandomness of SSRs of each repeat type showed very similar correlations between SSR length and nonrandomness. SSRs of trimers showed the strongest nonrandomness, followed by SSRs of monomers. The patterns of nonrandomness exhibited by SSRs of dimers and tetramers were in the order of dimmers>tetramers. Results pertaining to the mononucleotide repeat type indicate that little or no variations were observed among the four genomes of *Brucella* species. However, *B.melitensis* 16M and shows a little abundance with mean values of 1.02 when compared to the mean values 0.78, 0.77 and 0.75 of *Brucella abortus*S19, *Brucella suis* and *Brucella canis* ATCC23365, respectively in the genomic region. Whereas, in the coding regions of *Brucella* species there was no much variation observed in mono and di-nucleotide SSRs when compared to tri- and tetra nucleotide SSRs (**Figure 2**). Similar to mononucleotide repeats, di-nucleotide repeat SSRs also showed a little abundance with mean values of 0.63, 0.59, 0.52 and 0.51 of *Brucella abortus*S19, *Brucella suis*, *Brucella canis* ATCC23365 and *Brucella melitensis*16M respectively in the genomic regions (**Figure 1**) On contrary, the tri-nucleotide SSRs showed a distinct variability with extensive level of overrepresentation in both genomic and coding regions of four *Brucella* species with mean values of 1.46, 1.53, 1.20, 2.06.and 2.07 , 2.43, 1.82, 3.39.while in tetra-nucleotide SSRs showed under representation with mean values of 0.25, 0.28, 0.14, 0.17 in genomic and 0.19, 0.22, 0.09, 0.06 and coding region.

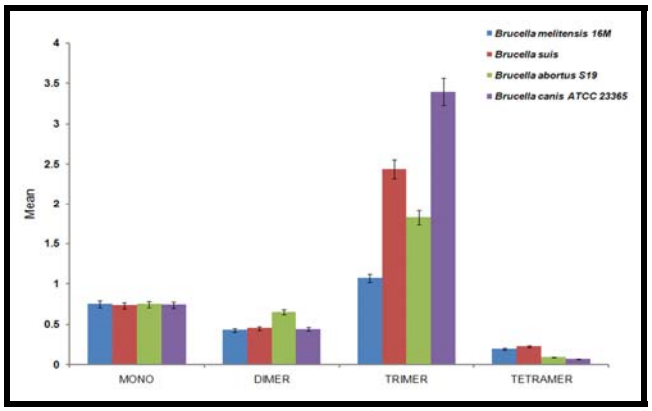

Figure 2: Bars represent the abundance of SSRs in the coding regions of Brucella species.

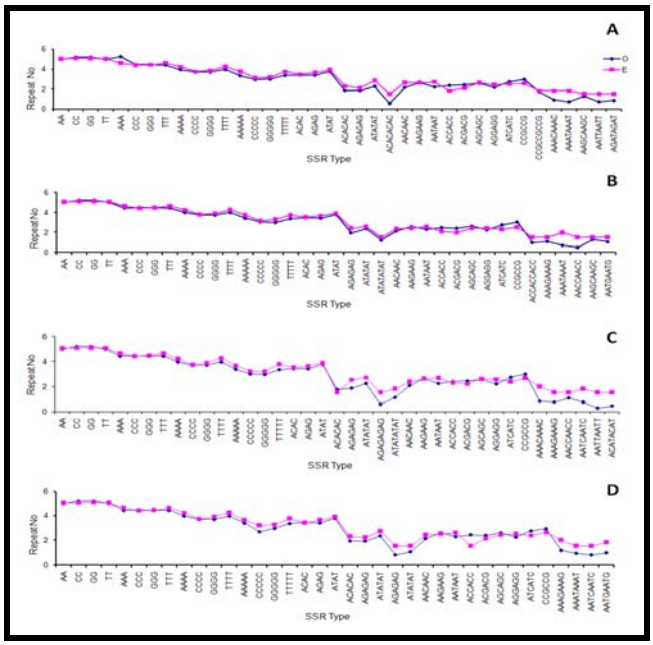

Figure 3: Distribution of observed (filled diamond)/expected (filled square) ratio of SSRs in genomic region of *Brucella* species. **(A)** *Brucella melitensis*16M; **(B)** *Brucella suis*; **(C)** *Brucella abortus*S19; **(D)** *Brucella canis*ATCC23365

Expected frequencies of SSRs of given motif length and repeat number were determined by computer generated genomes to construct random ordering of nucleotides based on their overall frequencies in the genome, and their departures were tested using parametric statistics. Among different types of nucleotides, the O/E ratio performed with computational tool SPSS 16.0.1 indicates that tri-nucleotide repeat tracts were overrepresented in both coding and non-coding regions. The O/E values also reveal that tri-nucleotide microsatellites were related to the genomes of *Brucella melitensis*16M, *Brucella suis* and *Brucella abortus*S19 but not *Brucella canis*ATCC23365 (**Figure 3, 4**). On the other hand, tetra nucleotide repeat type SSRs are underrepresented (**Figure 1, 2**) in both coding and non-coding regions. The magnitude of distribution of tetra nucleotide SSRs were in the order of *Brucella canis*ATCC23365 > *Brucella abortus*S19> *Brucella melitensis*16M > *Brucella suis* (**Figure 1, 2**).

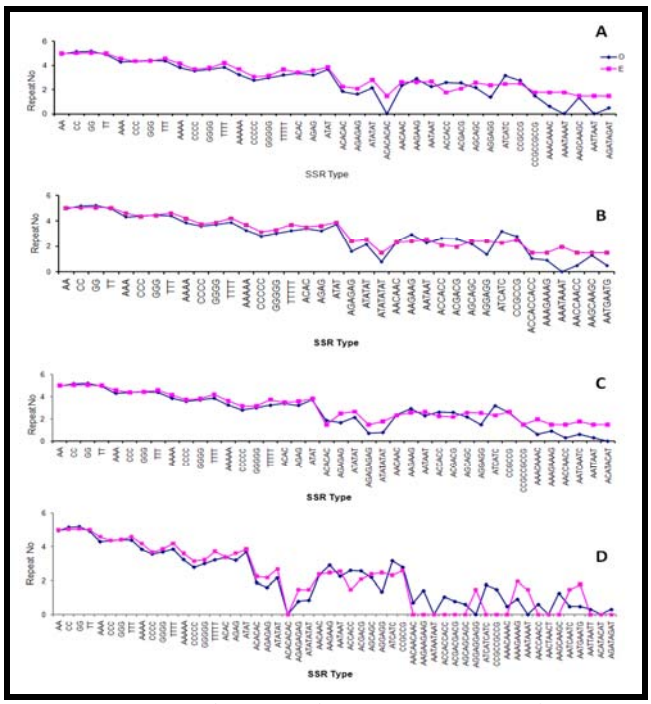

Figure 4: Distribution of observed (filled diamond)/expected (filled square) ratio of SSRs in coding region of Brucella species. **(A)** Brucella melitensis16M; **(B)** Brucella suis; **(C)** Brucella abortusS19; **(D)** Brucella canisATCC2336

Discussion:

Computer based analysis of genome-wide screening of SSR frequencies of given motif length and repeat number showed a significant ($P < 0.05$) variation of tri- and tetra nucleotide SSRs relative to expectations among the four *Brucella* species selected (**Figure 3, 4**). The picture represents that the tri nucleotide repeats were overrepresented and tetra nucleotide repeats were underrepresented in both genomic and coding region of the selected species (**Figure 1, 2**). Expected frequencies of SSRs of given motif length and repeat number were determined by computer generated genomes to construct random ordering of nucleotides based on their overall frequencies in the genome, and their departures were tested using Parametric statistics. The number of SSRs decreased rapidly with increasing size of the repeat unit. Similarly, mono nucleotide SSRs was slightly underrepresented in all the organisms except in *B.melitensis* 16M (1.02). In mono-nucleotide SSRs, the mean value of the average of four genomes in total and coding region are 0.83 and 0.74, respectively (**Figure 1, 2**), whereas in case of di-nucleotide SSRs the average of four genomes and the coding region are 0.5 and 0.4, respectively, indicating the underrepresentation pattern in these organisms (**Figure 1, 2**). On contrary, the tri-nucleotide SSRs showed a distinct variability with extensive level of overrepresentation in both genomic and coding regions of four *Brucella* species. The tetra-nucleotide repeats exhibited underrepresentation as that of mono-and di-nucleotide SSRs. The observed/expected (O/E) ratio gives the valuable information regarding the distribution pattern of virtually existing SSRs in the organisms and the computationally based analysis of SSRs. The O/E ratio of all the means of SSRs Vs species indicate that, the distribution pattern of SSRs in genomic and coding regions were either under-or over-represented in different species. In *B.suis* and *B.abortus* S19, slight underrepresentation pattern in the SSRs were observed in the genomic region whereas, in the coding region a slight overrepresentation was observed in *B.canis* ATCC23365 and underrepresentation was observed in *B.abortus* S19. The O/E ratio indicates that, there is a tremendous overrepresentation of tri nucleotide repeats on an average r4 >1.5, $P < 0.05$ in genomic and r4>2.4, $P < 0.05$ in coding region of four *Brucella* genomes. Whereas O/E value of tetra-nucleotide repeats indicate under-represented pattern in *B. canis*ATCC23365 and *B. abortus* S19 (**Figure 4**).

Further, statistical tools like Poisson distribution and normal distribution analysis were performed to evaluate whether SSRs of a given length are over- or under-represented in genomic DNA sequences or not by comparing observed SSR counts with random sequences **[17]**. In general, as SSR counts were decreased, there was an increase in the SSR length indicating an exponential trend among all the *Brucella* species tested (**Figure 1, 2**). However, in tri-nucleotide SSRs, a direct proportional pattern was observed in the SSR counts and SSR length which might be responsible for the over-represented distribution of four genomes. Herein, we suggest that tri-nucleotide SSRs in both genomic and coding regions might be an important approach to generate functional variation of proteins in the *Brucella* species. In the present study, even the organisms share >90% genomic homology, the over-representation of tri-nucleotide SSRs might be potential agents to cause wide range of pathogenicity among the selected human *Brucella* species. Since, SSRs act as stress response genes, transcription regulators and virulence factors. We also postulate that all these properties of SSRs are accompanied by tri-nucleotide SSRs which are embedded in repeat motifs. These repeat rich regions act as reservoirs of genes, which are capable of bringing about certain variability in virulence, antigenicity, and host adaptation. The results of the study suggests that SSRs can be used as tool for studies of genetic variation, determine the virulence and diagnostic purposes in species of *Brucella*.

Conclusion and Future directions:

To conclude, deviations from expected values can be interpreted as over- or under-representation of SSRs of a given type and length. There is preferential distribution of repeat motifs, and some motifs are organism specific. Although the four genome models can deviate from exponential relationship, the deviations are experimental, and the differences between random models can be used as benchmarks in assessing the significance of variations observed in the genomic DNA sequence. The ability to identify SSRs with nonrandom distribution profiles, as demonstrated in this article, is of interest since it may assist the elucidation of disease associated SSRs and the development of novel indicators of adverse health conditions with high prevalence among minority groups. The development of indicators of these health conditions will assist the diagnosis, treatment, and management of affected individuals and will contribute to the alleviation of extant health disparities. Further, monitoring SSR type, distribution and abundance has paramount importance to know the inn near future short-term variability in the genome of a large number of medically important microorganisms. Further investigation on the analysis of SSR composition in clinical isolates may in the end be an important prognostic marker of a patient's risk of developing severe infections. The present study may be helpful in understanding the influence of SSRs and their effects on human beings.

Acknowledgements:

I specially thank Professor. K.V.S. Sharma. Dept of Statistics, S.V. University, Tirupati from whom I receive great help to analyses the Statistical data.

References:

[1] Xiang *et al. BMC Bioinformatics*. 2006 **7**: 347 [PMID: 16842628]
[2] Young EJ. *Clin Infect Dis*. 1995 **21**: 283 [PMID: 8562733]
[3] Halling SM *et al. J Bacteriol*. 2005 **187**: 2715 [PMID: 15805518]
[4] Gandara B *et al. J Clin Microbiol*. 2001 **39**: 235 [PMID: 11136777]
[5] Van Belkum A *et al. Microbiol Mol Biol Rev*. 1998 **62**: 275 [PMID: 9618442]
[6] Moxon ER *et al. Curr Biol*. 1994 **4**: 24 [PMID: 7922307]
[7] Coenye T & Vandamme P. *DNA Res*. 2005 **12**: 221 [PMID: 16769685]
[8] Sreenu VB *et al. Nucleic Acids Res*. 2003 **31**: 106 [PMID: 12519959]
[9] Sreenu VB *et al. BMC Genomics*. 2006 **7**: 78 [PMID: 16603092]
[10] Wierdl M *et al. Genetics*. 1997 **146**: 769 [PMID: 9215886]
[11] Catasti P *et al. Genetica*. 1999 **106**: 15 [PMID: 10710707]
[12] Perez-Martin J *et al. Microbiol Rev*. 1994 **58**: 268 [PMID: 8078436]
[13] Krasilnikov MM *et al. EMBO J*. 1999 **17**: 5095 [PMID: 9724645]
[14] Henderson IR *et al. Mol Microbiol*. 1999 **33**: 919 [PMID: 10476027]
[15] Ellegren H. *Nat Rev Genet*. 2004 **5**: 435 [PMID: 15153996]
[16] Levinson G & Gutman GA. *Mol Biol Evol*. 1987 **4**: 203 [PMID: 3328815]
[17] Van Ham SM *et al. Cell* 1993 **73**:1187 [PMID: 8513502]
[18] Kashi Y & King DG. *Trends Genet*. 2006 **22**: 253 [PMID: 16567018]

Edited by P Kangueane

Citation: Kiran *et al.* Bioinformation 6(5): 179-182 (2011)

BIOINFORMATION

open access

Supplementary material:

Table 1: DNA sequences with accession numbers

Genome	Ac No	Genome Size(bp)	GC Content
Brucella melitesis 16M	NC_003317 NC_003318	3,269,159	57%
Brucella suis	NC_004310 NC_004311	3,315,175	57%
Brucella abortus S19	NC_010742 NC_010740	3,283,936	57%
Brucella canis ATCC 23365	NC_010104 NC_010103	3,312,769	57%