Austin Journal of Computational Biology and Bioinformatics

Open Access

Austin Publishing Group

Research Article

Comparative Genome Analysis of Short Sequence Repeats in Pathogenic and Non Pathogenic *Leptospira*- A Statistical Approach

Suresh KP[1]*, Kamatchi BI[1], Rahman H[2], Abraham S[3], Victor AAR[3] and Santra1 S[3]

[1]Department of Epidemiology and Biostatistics, National Institute of Veterinary Epidemiology and Disease Informatics (NIVEDI), India

[2]Director, ICAR-National Institute of Veterinary Epidemiology and Disease Informatics, India

[3]School of Biological Sciences, Madurai Kamaraj University, India

***Corresponding author:** Suresh KP, Department of Epidemiology and Biostatistics, National Institute of Veterinary Epidemiology and Disease Informatics (NIVEDI), Bangalore. Karnataka, India

Received: May 02, 2016; **Accepted:** June 03, 2016; **Published:** June 08, 2016

Abstract

Leptospirosis is a zoonotic disease caused by the genus, *Leptospira*. In this study the available genomic sequence of *Leptospira interrogans Lai* (Pathogenic) and *Leptospira biflexa Patoc* (Non-Pathogenic) of Chromosome I and Chromosome II were examined for the presence of short sequence repeats (n=1,2,3). Short Sequence Repeats (SSRs) or microsatellites extensively exist in genomes of prokaryotes and eukaryotes. Simple sequence repeats are the genetic loci where the bases are tandemly repeated for varying number of times. Comparative genome analysis will provide an opportunity to identify features that are unique to pathogenic species. SSR of the four sequences (pathogenic and non pathogenic) was found using R package. It was observed that, the pathogenic sequence contain more number of tandem repeats in both the chromosomes. Among them, mononucleotide frequency is very high in compared microsatellites. Meanwhile the occurrence of C/G or G/C has shown to be more difference in frequencies between pathogenic and non pathogenic. In both the chromosomes sequences, dinucleotide repeats is more frequent in pathogenic sequence and the frequency of TC in chromosome I and II and GA in chromosome II are shown to be less significant. Additionally, trinucleotide repeats are longer in pathogenic sequence. Moreover, the statistical analysis of microsatellites in both the sequences indicates the highly significant pattern of nucleotides. This suggests that, these short microsatellite repeats plays a major role in the gene expression, genetic diversity, gene evolution and understanding genomic instability.

Keywords: Short sequence repeats; *Leptospira*; Chromosome; R; Statistical analysis

Introduction

Leptospira and its approach to tandem repeat

Leptospirosis occurs both in developing and industrialized countries. Wild and domestic animals are important carriers of the disease [1]. Today, leptospirosis is recognized as a re-emerging infectious disease; therefore, understanding its epidemiology is a vital issue for designing intervention programs and diminishing its transmission. The disease has peak incidence during rainy seasons in warm climate regions and in summer or fall in temperate regions [2,3]. Leptospirosis has a global distribution and is more prevalent in tropical regions than in temperate zone. This is due to longer survival of leptospirosis [4,5]. The genus *Leptospira* contains 17 genome species as shown by DNA-DNA hybridization [6]. Under the current genotypic classification system, pathogenic and non pathogenic serovars may reside within the same genomopecies [7].

Experimental analysis of Variable Number Tandem Repeat (VNTR) can provide information relating to both the evolutionary and functional areas of bacterial diversity [8]. The ability to detect VNTRs in microorganisms has been greatly enhanced by the availability of whole genomic sequences and software http://tandem.bu.edu/ [9] that can search for VNTR loci from the sequences [10].

Tandem repeat: A repeat is recurrence of a pattern whereby DNA exhibits recurrence of many features [11]. Tandem repeats are copies of repetitive DNA sequences that lie adjacent to each other in a genomic sequence. Tandem Repeats (TRs) are DNA sequence motifs that contain at least two adjacent repeating units. According to the conservation of the repeated sequence, TRs are classified as identical/perfect TRs or degenerated/imperfect repeats [12]. They are common in both prokaryote and eukaryotes genome [13] and found in both protein coding and non-coding regions of the genome [14,15]. Microsatellites are abundant in genomes of eukaryotes and less abundant in bacterial genomes [16]. They are involved in recombination activity like unequal crossing over or unequal sister chromatin exchange [17]. Changes in copy number of repeats in satellite DNA could be accounted by biological processes, such as unequal crossing over [18].

Differentiation of repeats: Repeats are differentiated into microsatellites [unit size: 1–6 or 1–10 bp; also known as Simple Sequence Repeats (SSR)], minisatellites (unit size: 10–60 or 10–100 bp) and macrosatellites (unit size >100 bp) [19,20]. Microsatellite repeats are found to be significant in biological and medical fields [21,22]. These repeats have been implicated in many neurogenetic and other diseases [23]. Recent studies show that these repeats have

Citation: Suresh KP, Kamatchi BI, Rahman H, Abraham S, Victor AAR and Santra S. Comparative Genome Analysis of Short Sequence Repeats in Pathogenic and Non Pathogenic *Leptospira*- A Statistical Approach. Austin J Comput Biol Bioinform. 2016; 3(1): 1013.

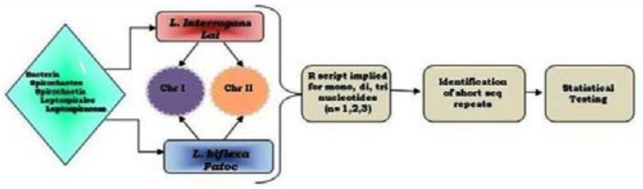

Figure 1: Showing the diagrammatic representation of work flow.

many functional roles to play [24]. DNA repeats exist in one of the following patterns, such as, Forward or direct repeat, Reverse repeat, Complement repeat and Palindromic repeat. Usually in Bacteria, repeats are divided into two subclasses: short repeats and longer repeats. In our study we have focused on the first category which is constituted of short sequence repeats ranging from mononucleotides to trinucleotides (n=1, 2, 3).

The Advancement of Biology

The advancement of biology and computational analysis represents a major endeavor in the post-genomic era. The increasing number of whole-genome sequencing projects has provided an enormous amount of information which leads to the need of new tools and string processing algorithms to analyze and classify the obtained sequences [25].

Understanding of repeats

Recently, advances in sequencing technology have contributed to the availability of lots of organisms. Theoretical studies have shown that evolution of these repeats is intended for the mechanisms of rigorous evolution which includes unequal crossing over and gene expression [18]. The analysis of these repeats in the pathogenesis of Leptospirosis will help in understanding the distribution of different patterns of repeats which are at close proximity to the variation sites which can make a mechanism of concerted evolution and unequal crossing over. The repeats having significant function has a role in causing many Neurodegenerative disease and even these repeats can act as markers [26].

In the current post genomic era the application of biology with informatics (Bioinformatics) will not only be able to analyze the proposition by deciphering the relations between the nucleotides in the chromosome. Here, we present a approach to calculate the simple sequence repeats by using an R package and statistical testing by χ^2 test which automates the process of biological-term classification and easier analysis of simple sequence repeats. A flowchart is shown to describe the outline of the research work (Figure 1).

Materials and Methods

Leptospira genome sequence retrieval

The complete genomic sequence of Pathogenic *Leptospira interrogans Lai* (Ref seq: NC_004342.2 and NC_004343.2) and non pathogenic *Leptospira biflexa Patoc* (Ref Seq: NC_010602.1 and NC_010843.1) of Chromosome I and Chromosome II was obtained from http://www.ncbi.nlm.nih.gov [27] (Table 1). The sequences of four chromosomes of *Leptospira* genome were used as input sequences. The R programming script was used to analysis the number of Singlet, doublets and triplets and its frequencies were analyzed for further calculations.

Table 1: Total Length Genomic sequences of Pathogenic and Non pathogenic *Leptospira.*

Serovar	Source	Length
Interrogans Lai	Chromosome I	4338762
Interrogans Lai	Chromosome II	359372
Biflexa Patoc	Chromosome I	3599677
Biflexa Patoc	Chromosome II	277655

Table 2: R programming script used to analyze the tandem repeat sequences.

```
library(ape)
library(seqinr)
dbs<- ("d:/fasta/Interrogans_chr1.fasta","d:/fasta/Biflexa_chr1.fasta",
"d:/fasta/Interrogans_chr11.fasta","d:/fasta/Biflexa_chr11.fasta")
numdbs <- length(dbs)
e1<-3*numdbs
p<-1
for (i in 1:numdbs)
{  db <- dbs[i]
temp<-read.dna(db,format="fasta")
x<-1
while(x<4)
{
e<-count(temp,x)
sink("d:/Result.txt", append=TRUE)
e1[p]<-list(e)
print(e1[p])
sink()
x<-x+1
p<-p+1
}
}
```

In *silico* identification of short sequence repeats

The program to find the short sequence repeats such as mono, di and trinucleotides repeats was implemented in R an open-source programming environment [28] for both the pathogenic and non pathogenic sequences of chromosome I and II (Table 2).

The difference in counts between pathogenic and non-pathogenic sequences of chromosome I and chromosome II of mono, di and trinucleotides (n=1, 2, 3) were analyzed. In mono and dinucleotides the Forward, Reverse, Complimentary and Palindromic sequences were grouped together as they represent the same sequence. In case of mononucleotide, A repeat is same as T repeat on a complementary stand whereas in case of Dinucleotides, (AG) is also equivalent to (GA, TC, CT) and in case of trinucleotides (AGT) it is equivalent to (TGC, TCA, ACG) in different reading frames or on a complementary strand. The amino acid specified by two or more synonyms was grouped together.

Statistical analysis of tandem repeat frequencies

The frequencies of short sequence repeats were determined. The frequencies obtained were used to calculate the Chi-Square test including P-value and Cramer's value using Vassar stat software http://vassarstats.net/ [29] to find the significant repeats in the whole genome.

Mode of calculation

The below mentioned calculation is applied to find the SSR in *Leptospiral* genome.

First step: the frequency of occurrence P(X) of the base residues such as mono, di and trinucleotide consecutive bases from the

Table 3: Different repeats of Mononucleotide count in Pathogenic and Non Pathogenic sequences of chromosome I.

S.No	Mononucleotide	Pathogenic (n=4.338762)	Non Pathogenic (n=3.599677)	χ^2 Value	P Value	Cramer's Value
1	A/T	1.411	1.102	3370.49	<.0001	0.0206
2	C/G	0.756	0.703	5784.11	<.0001	0.0270
3	G/C	0.763	0.697	4150.70	<.0001	0.0229
4	T/A	1.408	1.098	3478.72	<.0001	0.0209

Table 4: Different repeats of Mononucleotide count in Pathogenic and Non Pathogenic sequences of chromosome II.

S.No	Mononucleotide	Pathogenic (n= 0.359372)	Non Pathogenic (n=0.277655)	χ^2 Value	P Value	Cramer's Value
1	A/T	0.117	0.084	323.62	<.0001	0.0225
2	C/G	0.064	0.054	330.77	<.0001	0.0228
3	G/T	0.063	0.055	572.34	<.0001	0.03
4	T/A	0.116	0.084	295.3	<.0001	0.0215

analyzed dataset were used to calculate. Whereas the X represents the base residues (i.e.) A, T, G, C for Singlet; AT, GC, CG for doublets; CAG, ATT, GAC for triplets. The observed frequency of the bases were calculated with the formula [Pobser (AB) =P (A) x P (B)] [26].

Second step: the estimated value (Expected Value) and the frequency of the base repeats of short range interactions of the consecutive base repeats were calculated using the formula as follows [26].

Third step: the calculations were also presented to calculate the value for all singlet, doublet and triplet using the following formula. [X2 (AB) = (Pobser (AB) –Pexpect (AB)) 2/Pexpect (AB)]. The high significant x2 values were selected from the datasets for further analysis and were used to predict the significant frequency.

Results and Discussion

During the past 10 years a number of studies have aimed to clarify the virulence factors of leptospires on the basis of known genomic sequences of some serovars of *Leptospira* interrogans and one serovar of *Leptospira* biflexa. Most of these studies compare the proteome similarities between pathogenic and saprophytic leptospires, detecting a number of proteins present only in the pathogenic, virulent serovars [30]. The *leptospiral* outer membrane lipoproteins act as the main virulence factor towards host tissues. The genome of *Leptospira* interrogans encode more lipoproteins than non-spirochetes genome: approximately 145 genes have been detected which encode putative lipoproteins in addition to putative extracellular and outer membrane proteins [31].

The Lipoproteins which are contributing to the major virulence factors are exclusively found only in pathogenic *Leptospira* species. On the basis of genome sequence information used in this study the acquisition of virulence associated genes by the pathogenic leptospires during the course of evolution may be contributing to the larger genome size in pathogenic *Leptospira* than non pathogenic species. The chromosome I and II of pathogenic *Leptospira* has huge variation rate in the size of genome. This major sequence composition differences is observed in both the chromosome. It has additional 7,39,085 base pairs in Chromosome I and 81,717 base pairs in Chromosome II when compared to non pathogenic *Leptospira* genome.

Mononucleotides frequency of chromosome I and II

Complete genome sequences of *Leptospira* Interrogans and *Leptospira* biflexa of chromosome I and II (pathogenic and non pathogenic) for the occurrence of mononucleotides revealed that, the pathogenic sequence contain higher frequency of A and T when compared to non pathogenic sequences, i.e. A/T repeat units to be in longer tail. In contrast, the frequencies of C and G are found to be similar in both the cases with minor difference. In Chi-square test, C/G and G/C has more difference in frequencies of nucleotides in pathogenic and nonpathogenic with higher chi-square values of 5784.11 & 4150.70 respectively in chromosome I (Table 3) and 572.34 & 330.77 (Table 4) in chromosome II. Similarly, the Cramer's value is also high in these nucleotides. The repeats of Mononucleotides of the pathogenic and non pathogenic *Leptospira* by Chi-square analysis reveals that poly (A) and Poly (T) were found in all the chromosomes when compared to C/G repeat units (Figure 2). The length distribution of the mononucleotide A/T repeat units seems to have longer (Figure 3).

Dinucleotides frequency of chromosome I and II

Dinucleotide repeats reveals that the repeat AA was seen rich in the majority of the chromosomes and from the statistical analyzed data, the repeats GG, CG and TC found to be more significant. The dinucleotides frequency of chromosome I and II reveals that, the pathogenic sequences contain higher frequency of certain homomeric dinucleotides such as AA which are found to be 13.55% in pathogenic and 11.98% in non pathogenic sequences (Figure 4).

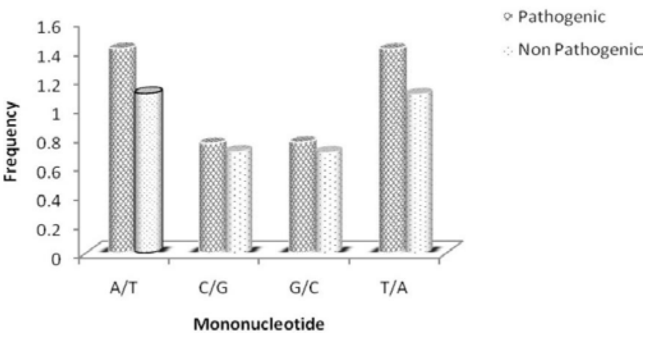

Figure 2: Significance of Mononucleotide counts in pathogenic and Non pathogenic Sequences of chromosome 1.

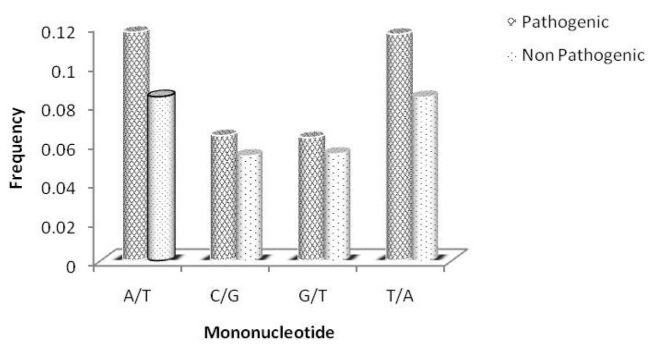

Figure 3: Significance of Mononucleotide count in pathogenic and Non Pathogenic sequences of chromosome 2.

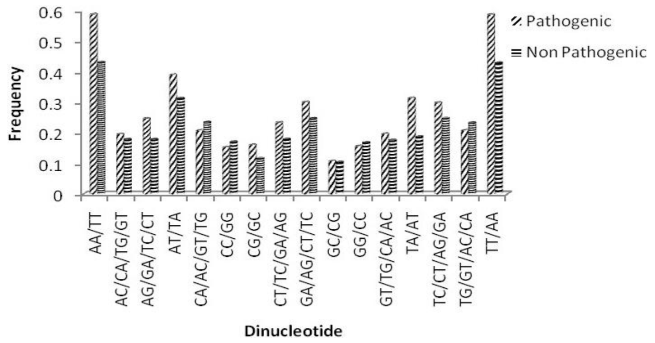

Figure 4: Significance of Dinucleotide count in pathogenic and non pathogenic sequences of chromosome 1.

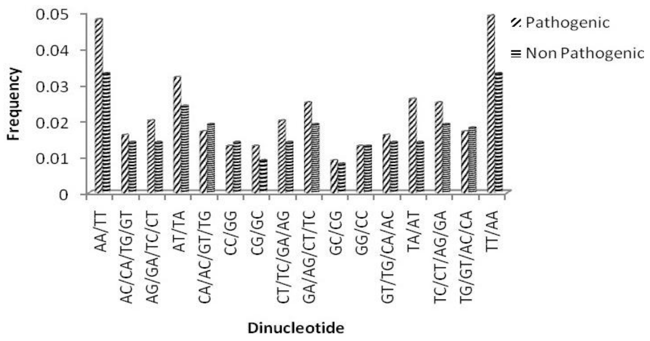

Figure 5: Significance of Dinucleotide count in pathogenic and non pathogenic sequences of chromosome 1.

Similarly frequency of TT in pathogenic are (13.54%) and in case of non pathogenic sequences (11.92%). In contrast, CC and GG are frequently repeated in nonpathogenic sequences (Table 5). The GG repeats act as an intra molecular G-G base pairing between telomere repeats stabilizes the hairpin DNA [32]. Among all the dimer repeats, AT was found to be predominant. In case of heteromeric dinucleotides, AT is more frequent followed by GA, TA, CT and TC (Figure 5). In all these five nucleotides, pathogenic sequences contain more number of frequencies than non pathogenic (Table 6). However, the frequency of CA and TG is also predominant in non pathogenic sequence. Interestingly the frequency of AC, AG, CG, GC and GT are less frequent in both the chromosome sequences.

Triplet codon repetitions in chromosome I and II

Similarly like mono and di, trinucleotides also contains more number of frequency in pathogenic when compared to non pathogenic sequences in both the chromosomes (Figure 6). These tri nucleotides have a significant role in the biology of diseases. All the codons are coding for amino acids Trimer repeats reveals that ACG, CCG/CAG, CGT, TCG were found to be maximum in majority of the chromosome (Figure 7). In case of chromosome I, CAT/CAC which codes for the amino acid Histidine; has the highest χ^2 value and Cramer's value of 10206.22 and 0.0359 (Table 7). Even in case of chromosome II, CAT/CAC has the highest χ^2 value and Cramer's value as 864.07 and 0.0368 (Table 8). CAC/CAT is also called as His-tag repeating sequence, whereas it is helpful in purification

Table 5: Different repeats of Dinucleotide count in Pathogenic and Non-Pathogenic sequences of chromosome I.

S.No	Dinucleotide	Pathogenic (n=4.338762)	Non Pathogenic (n=3.599677)	χ^2 Value	P Value	Cramer's Value
	AA/TT	0.588	0.431	4326.30	<.0001	0.0230
	AC/CA/TG/GT	0.196	0.179	897.24	<.0001	0.0106
	AG/GA/TC/CT	0.237	0.179	1068.31	<.0001	0.0116
	AT/TA	0.390	0.313	205.21	<.0001	0.0051
	CA/AC/GT/TG	0.208	0.235	11285.47	<.0001	0.0377
	CC/GG	0.153	0.171	7435.48	<.0001	0.0306
	CG/GC	0.161	0.117	1202.83	<.0001	0.0123
	CT/TC/GA/AG	0.234	0.180	638.75	<.0001	0.0090
	GA/AG/CT/TC	0.301	0.247	19.17	<.0001	0.0016
	GC/CG	0.108	0.105	1285.26	<.0001	0.0127
	GG/CC	0.157	0.169	5925.77	<.0001	0.0273
	GT/TG/CA/AC	0.197	0.176	557.08	<.0001	0.0084
	TA/AT	0.314	0.188	13491.90	<.0001	0.0412
	TC/CT/AG/GA	0.299	0.248	0.97	0.3247	0.0003
	TG/GT/AC/CA	0.208	0.233	10387.04	<.0001	0.0362
	TT/AA	0.587	0.429	4624.36	<.0001	0.0241

Table 6: Different repeats of Dinucleotide count in Pathogenic and Non-Pathogenic sequences of chromosome II.

S.No	Dinucleotide	Pathogenic (n=0.359372)	Non Pathogenic (n=0.277655)	χ^2 Value	P Value	Cramer's Value
1	AA/TT	0.048	0.033	335.51	<.0001	0.0230
2	AC/CA/TG/GT	0.016	0.014	49.66	<.0001	0.0088
3	AG/GA/TC/CT	0.020	0.014	80.73	<.0001	0.0113
4	AT/TA	0.032	0.024	23.11	<.0001	0.0060
5	CA/AC/GT/TG	0.017	0.019	929.39	<.0001	0.0382
6	CC/GG	0.013	0.014	851.72	<.0001	0.0366
7	CG/GC	0.013	0.009	97.02	<.0001	0.0123
8	CT/TC/GA/AG	0.020	0.014	60.38	<.0001	0.0097
9	GA/AG/CT/TC	0.025	0.019	6.28	0.0122	0.0031
10	GC/CG	0.009	0.008	141.52	<.0001	0.0149
11	GG/TT	0.013	0.013	410.28	<.0001	0.0254
12	GT/TG/CA/AC	0.016	0.014	64.68	<.0001	0.0101
13	TA/AT	0.026	0.014	1151.54	<.0001	0.0425
14	TC/CT/AG/GA	0.025	0.019	0	1	0
15	TG/GT/AC/CA	0.017	0.018	884.20	<.0001	0.0373
16	TT/AA	0.049	0.033	432.94	<.0001	0.0261

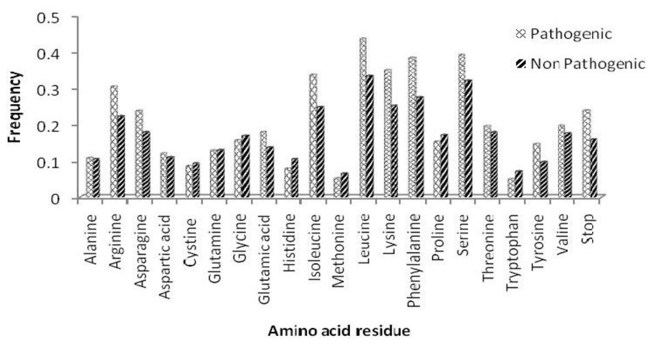

Figure 6: Significance of trinucleotide count in pathogenic and non pathogenic sequences of chromosome 1 encoded by amino acid residue.

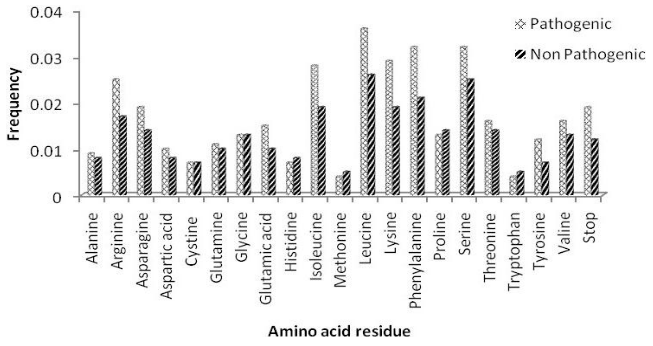

Figure 7: Significance of trinucleotide count in pathogenic and non pathogenic sequences of chromosome 2 encoded by amino acid residue.

of recombinant DNA [32]. Report says that when there is increase in CAG repeats then the individual is affected with Huntington's diseases and when CTG repeats ranges from 50 to 5000 times in the gene which may leads to Myotonic dystrophy [26]. In case of *Leptospira* it also has the repeat of CAG and CTG which codes for the codes Glutamine and Leucine. In future many genetic, Leptospirosis and neurodegenerative disorder can be cured by analysis of Triplets.

Chi-square result of mononucleotide, dinucleotide and trinucleotides

In all the microsatellite repeats, the majority of the frequencies are occupied by pathogenic sequences in both the chromosomes. Among the three SSR, the frequency of mononucleotides repeats are higher than di and trinucleotide. The p value of all the mononucleotides shows that all are highly significant. The chi-square of dinucleotides shows that, the highest differences are associated with TA (χ^2 =13491.90) followed by CA and TG. In terms of p value, all the dinucleotides are highly significant except TC showing (χ^2 =0.97and p=0.3247) in chromosome I and (p=1) and in chromosome II. Some amino acid repeats are more frequent in pathogenic but less frequent in non pathogenic, but the percentage shows that the higher frequency is in non pathogenic. For example, the frequency of amino acid alanine in chromosome II shows the repeats are more frequent in pathogenic (0.108) and less frequent in non pathogenic (0.105) but the percentage is 2.4 in pathogenic and 2.9 in non pathogenic. The amino acid histidine shows highest chi-square and Cramer's value followed by Tryptophan and proline. It was observed that, the percentage difference is more in these three amino acids when compared to others. All the amino acids are highly significant in both the chromosomes except the amino acid serine (χ^2 =0.47 and p= 0.493) in chromosome II.

Conclusion

The computational tools and statistical analysis made a formulation towards the analysis of repetitive DNA sequences in *Leptospira* Interrogans and *Leptospira* biflexa of chromosome I and II (pathogenic and non pathogenic). These tools, methods and approaches have been briefly highlighted in the study. The result of chi-square test, in case of mononucleotides the p value indicates to be highly significant which proves the test in which this is asymptotically

Table 7: Total occurrence of codon repeats in Pathogenic and Non-pathogenic sequences of Chromosome I.

S.No	Trinucleotide	Encoded Amino acid residue	Pathogenic (n=4.338762)	Non Pathogenic (n=3.599677)	χ^2 Value	P Value	Cramer's Value
1	GCA/ GCC/ GCG /GCT	Alanine	0.108	0.105	1285.26	<.0001	0.0127
2	AGA/ CGA/ CGC/ CGG/ CGT/ AGG	Arginine	0.305	0.223	2220.51	<.0001	0.0167
3	AAT/AAC	Asparagine	0.238	0.179	1029.85	<.0001	0.0114
4	GAT/GAC	Aspartic acid	0.121	0.110	541.68	<.0001	0.0083
5	TGT/TGC	Cystine	0.086	0.093	3338.26	<.0001	0.0205
6	CAA/CAG	Glutamine	0.129	0.130	2671.45	<.0001	0.0183
7	GGT/GGC/GGA/GGG	Glycine	0.157	0.169	5925.77	<.0001	0.0273
8	GAA/GAG	Glutamic acid	0.180	0.137	658.25	<.0001	0.0091
9	CAT/CAC	Histidine	0.079	0.105	10206.22	<.0001	0.0359
10	ATT/ATC/ATA	Isoleucine	0.337	0.248	2266.36	<.0001	0.0169
11	ATG	Methonine	0.052	0.065	4831.85	<.0001	0.0247
12	TTA/TTG/CTT/ CTC/ CTA/CTG	Leucine	0.437	0.334	1492.26	<.0001	0.0137
13	AAA/AAG	Lysine	0.350	0.252	3143.26	<.0001	0.0199
14	TTT, TTC	Phenylalanine	0.384	0.275	3750.49	<.0001	0.0217
15	CCT/CCC/CCA/CCG	Proline	0.153	0.171	7435.16	<.0001	0.0306
16	TCT/TCC/TCA/TCG/AGT/AGC	Serine	0.392	0.321	33.75	<.0001	0.0021
17	ACT/ACC/ ACA/ACG	Threonine	0.196	0.179	897.24	<.0001	0.0106
18	TGG	Tryptophan	0.050	0.071	8460.73	<.0001	0.0326
19	TAT/TAC	Tyrosine	0.147	0.097	3129.28	<.0001	0.0199
20	GTT/ GTC/ GTA/GTG	Valine	0.197	0.176	557.08	<.0001	0.0084
21	TAA, TGA, TAG	Stop	0.239	0.159	4798.65	<.0001	0.0246

Table 8: Total occurrence of codon repeats in Pathogenic and Non-pathogenic sequences of Chromosome II.

S.No	Trinucleotide	Amino acid residue	Pathogenic (n=0.359372)	Non Pathogenic (n=0.277655)	χ^2Value	P Value	Cramer's Value
1	GCA/ GCC/ GCG /GCT	Alanine	0.009	0.008	141.52	<.0001	0.0149
2	AGA/ CGA/ CGC/ CGG/ CGT/ AGG	Arginine	0.025	0.017	208.33	<.0001	0.0181
3	AAT/AAC	Asparagine	0.019	0.014	88.97	<.0001	0.0118
4	GAT/GAC	Aspartic acid	0.010	0.008	42.56	<.0001	0.0082
5	TGT/TGC	Cystine	0.007	0.007	306.99	<.0001	0.022
6	CAA/CAG	Glutamine	0.011	0.010	209.86	<.0001	0.0182
7	GGT/GGC/GGA/GGG	Glycine	0.013	0.013	410.28	<.0001	0.0254
8	GAA/GAG	Glutamic acid	0.015	0.010	78.36	<.0001	0.0111
9	CAT/CAC	Histidine	0.007	0.008	864.07	<.0001	0.0368
10	ATT/ATC/ATA	Isoleucine	0.028	0.019	203.41	<.0001	0.0179
11	ATG	Methonine	0.004	0.005	387.4	<.0001	0.0247
12	TTA/TTG/CTT/ CTC/ CTA/CTG	Leucine	0.036	0.026	131.18	<.0001	0.0144
13	AAA/AAG	Lysine	0.029	0.019	230.68	<.0001	0.019
14	TTT, TTC	Phenylalanine	0.032	0.021	366.95	<.0001	0.024
15	CCT/CCC/CCA/CCG	Proline	0.013	0.014	852.07	<.0001	0.0366
16	TCT/TCC/TCA/TCG/AGT/AGC	Serine	0.032	0.025	0.47	0.493	0.0009
17	ACT/ACC/ ACA/ACG	Threonine	0.016	0.014	49.66	<.0001	0.0088
18	TGG	Tryptophan	0.004	0.005	612.63	<.0001	0.031
19	TAT/TAC	Tyrosine	0.012	0.007	364.37	<.0001	0.0239
20	GTT/ GTC/ GTA/GTG	Valine	0.016	0.013	64.68	<.0001	0.0101
21	TAA, TGA, TAG	Stop	0.019	0.012	305.31	<.0001	0.0219

true which can be made to approximate a chi-square distribution as closely as desired. In addition the dinucleotides also has the highest differences showing the value associated with TA, CA and TG (χ^2 =13491.90), the p value are also shown to be highly significant (χ^2 =13491.90). In the process of testing for codon repetitions tri nucleotides also contain higher frequencies in both the chromosome I and II. The p value is defined as the probability of obtaining a result equal to than what was actually observed, assuming that the hypothesis under consideration is true.

Analysis of these repeats helps in finding the markers for many dreadful diseases [26]. In this study we have shown the occurrence of Singlet, Doublet and Triplet of pathogenic and non-pathogenic *Leptospira* chromosomes I and II. The repeats have significant function. In future it may help to improve the studies in Microsatellite, Gene switch in non-coding DNA etc. The composition biases of the chromosome strongly influence the rate of tandem repeat and the repeat of amplication [32]. This will certainly provide new clues in deciphering the dynamics of repeats in bacterial genomes and also will provide much information on evolutionary implications.

References

1. Rafiei A, Hedayati Zadeh-Omran A, Babamahmoodi F, Alizadeh Navaei R, Valadan R. Review of Leptospirosis in Iran [in persian]. J Mazandaran Univ Med Sci. 2012; 22: 114-124.
2. Supply P, Magdalena J, Himpens S, Locht C. Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. Mol Microbiol. 1997; 26: 991-1003.
3. Zavitsanou A, Babatsikou F. Leptospirosis: epidemiology and preventive measures. Health Sci J. 2008; 2: 75-82.
4. Zuerner RL, Alt DP. Variable nucleotide tandem-repeat analysis revealing a unique group of *Leptospira* interrogans serovar Pomona isolates associated with California sea lions. J Clin Microbiol. 2009; 47: 1202-1205.
5. Soltanimajd N, Khodaverdidarian E, Khaki P, Moradi Bidhendi S, Yahaghi E. Epidemiological patterns of *Leptospira* spp among slaughterhouse workers in Zanjan- Iran. Asian Pac J Trop Dis. 2012; 2: 550-552.
6. Bharti AR, Nally JE, Ricaldi JN, Matthias MA, Diaz MM, Lovett MA, et al. Leptospirosis: a zoonotic disease of global importance. Lancet Infect Dis. 2003; 3: 757-771.
7. Levett PN. Leptospirosis. Clin Microbiol Rev. 2001; 14: 296-326.
8. Van Belkum A, Scherer S, Van Alphen L, Verbrugh H. Short-sequence DNA repeats in prokaryotic genomes. Microbiol Mol Biol Rev. 1998; 62: 275-293.
9. Laboratory for Biocomputing and Informatics.
10. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999; 27: 573-580.
11. Roa SR, Trivedi S, Emmanuel D, Merita K, Hynniewta M. DNA repetitive sequences-types, distribution and function: A review. Journal of Cell and Molecular Biology. 2010; 7: 1-11.
12. Lopes J, Ribeyre C, Nicolas A. Complex minisatellites rearrangements generated in the total or partial absence of Rad27/hFEN1 activity occur in a single generation and are Rad51 and Rad52 dependent. Mol Cell Biol. 2006; 26: 6675-6689.
13. Usdin K, Kumari D. Repeat-mediated epigenetic dysregulation of the FMR1 gene in the fragile X-related disorders. Front Genet. 2015; 6: 192.
14. Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 2000; 10: 967-981.
15. Katti MV, Ranjekar PK, Gupta VS. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol. 2001; 18: 1161-1167.
16. Le Fleche P, Hauck Y, Onteniente L, Prieur A, Denoeud F, Ramisse V, et al. A tandem repeats database for bacterial genomes: application to the genotyping of Yersinia pestis and Bacillus anthracis. BMC Microbiol. 2001; 1: 2.
17. Richard GF, Kerrest A, Dujon B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev. 2008; 72: 686-727.
18. UgarkoviÄ D, Plohl M. Variation in satellite DNA profiles--causes and effects. EMBO J. 2002; 21: 5955-5959.
19. Mayer C1, Leese F, Tollrian R. Genome-wide analysis of tandem repeats in Daphnia pulex--a comparative approach. BMC Genomics. 2010; 11: 277.
20. Gemayel R, Cho J, Boeynaems S, Verstrepen KJ. Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. Genes (Basel). 2012; 3: 461-480.
21. Gulcher J. Microsatellite markers for linkage and association studies. Cold Spring Harb Protoc. 2012; 2012: 425-432.
22. Yan HM, Dong C, Zhang EL, Tang CF, A XX, Yang WY, et al. [Analysis of genetic variation in rice paddy landraces across 30 years as revealed by microsatellite DNA markers]. Yi Chuan. 2012; 34: 87-94.
23. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999; 27: 573-580.
24. Gangwal K, Lessnick SL. Microsatellites are EWS/FLI response elements: genomic “junk” is EWS/FLI’s treasure. Cell Cycle. 2008; 7: 3127-3132.
25. Cariaso M, Folta P, Wagner M, Kuczmarski T, Lennon G. IMAGEne I: clustering and ranking of I.M.A.G.E. cDNA clones corresponding to known genes. Bioinformatics. 1999; 15: 965-973.
26. Pavithra V, Surendar, Mugilan S. Analysis of Tandem repeats in human genome by computational and statistical approach. Journal of Pharmacy Research. 2014; 8: 359-362.
27. Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BF, Rapp BA, et al. GenBank. Nucleic Acids Res. 1999; 27: 12-17.
28. Ihaka R, Gentleman RR. A language for data analysis and graphics. J Comp Graph Stat. 1996; 5: 299-314.
29. Cinco M. New insights into the pathogenicity of leptospires: evasion of host defences. New Microbiol. 2010; 33: 283-292.
30. Setubal JC, Reis M, Matsunaga J, Haake DA. Lipoprotein computational prediction in spirochaetal genomes. Microbiology. 2006; 152: 113-121.
31. Achaz G, Rocha EP, Netter P, Coissac E. Origin and fate of repeats in bacteria. Nucleic Acids Res. 2002; 30: 2987-2994.
32. Hengen P. Purification of His-Tag fusion proteins from Escherichia coli. Trends Biochem Sci. 1995; 20: 285-286.

Austin J Comput Biol Bioinform - Volume 3 Issue 1 - 2016
ISSN : 2379-7967 | www.austinpublishinggroup.com
Suresh et al. © All rights are reserved

Citation: Suresh KP, Kamatchi BI, Rahman H, Abraham S, Victor AAR and Santra S. Comparative Genome Analysis of Short Sequence Repeats in Pathogenic and Non Pathogenic *Leptospira*- A Statistical Approach. Austin J Comput Biol Bioinform. 2016; 3(1): 1013.