International Journal of Genetics
ISSN: 0975-2862 & E-ISSN: 0975-9158, Volume 8, Issue 1, 2016, pp.-180-185.
Available online at http://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BPJ0000226

COMPARATIVE GENOME ANALYSIS OF SHORT SEQUENCE REPEATS IN PATHOGENIC AND NON PATHOGENIC *LEPTOSPIRA*- A STATISTICAL APPROACH

KAMATCHI INDUMATHI B.[1], ABRAHAM SUNIL[2], VICTOR RAMASAMY ANGELA ASIR[2], SANTRA S.[1], RAHMAN H.[3], GORTI RAVI KIRAN[4] AND SURESH K.P.*[1]

[1]Department of Epidemiology and Biostatistics, National Institute of Veterinary Epidemiology and Disease Informatics (NIVEDI), Bangalore. Karnataka, India.
[2]School of Biological Sciences, Madurai Kamaraj University, Madurai-625019, Tamilnadu, India.
[3]ICAR-National Institute of Veterinary Epidemiology and Disease Informatics (NIVEDI), Bengaluru.
[4]ICAR-National Institute of Animal Nutrition and Physiology, Bangalore.
*Corresponding Author: Email- indumathi_28@yahoo.com

Received: December 02, 2015; Revised: April 07, 2016; Accepted: April 08, 2016

Abstract- *Leptospira* is a pathogenic bacteria, which causes Leptospirosis in humans and animals. The genome sequence of *Leptospira interrogans Lai* (Pathogenic) and *Leptospira biflexa Patoc* (Non-Pathogenic) were retrieved and examined for the presence of Short Sequence Repeats (SSR) (n=1,2,3)in Chromosome I and Chromosome II. 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 a better understanding in the identification of the evolutionary relationship between pathogenic and non-pathogenic species. SSRs in genome sequences (pathogenic and non- pathogenic) were found using 'R' script, which was implemented in R packages. It was observed that the pathogenic sequence contains a number of tandem repeats in both the chromosomes. Meanwhile, the occurrence of C/G or G/C has more difference in their frequencies between pathogenic and non-pathogenic sequences. In both the chromosomes, dinucleotide repeats are frequent, but in the case of TC in both the chromosomes and GA in chromosome II are found to be less significant. Additionally, trinucleotide repeats are longer in pathogenic, whereas in non-pathogenic they are shorter. The statistical analysis of the microsatellites in both the sequences indicates the highly significant pattern of nucleotide repeats. The more number of genes in the pathogenic species may be acquired to the virulence in their course of evolution. This work partially suggests that SSRs plays a major role in genetic diversity, gene evolution and also in understanding the genomic instability.

Keywords- Short sequence repeats; Leptospira; Chromosome I and II; microsatellites; Statistical analysis.

Citation: B. Kamatchi Indumathi, et al., (2016) Comparative Genome Analysis of Short Sequence Repeats in Pathogenic and Non Pathogenic *Leptospira*- A Statistical Approach. International Journal of Genetics, ISSN: 0975- 2862 & E-ISSN: 0975-9158, Volume 8, Issue 1, pp.-180-185.

Introduction

Leptospirosis occurs both in developing and industrialized countries. [1]. Leptospirosis is now a re-emerging and infectious, it occurs in both tropical and temperate regions. [2,3] The genus *Leptospira* contains 17 genomospecies as shown by DNA-DNA hybridization studies [6]. Under the current genotypic classification system, pathogenic and non pathogenic serovars may reside within the same genomospecies [7].

Experimental analysis of Variable Number Tandem Repeat (VNTR) can provide related information to both the functional and evolutionary areas of bacterial diversity [8]. The ability to detect VNTRs in microorganisms has been greatly enhanced by the availability of whole genome 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 repetition of many features. [11]. Tandem repeats (TRs) are copies of repetitive DNA sequences that lie adjacent to each other in a genomic sequence. TRs are DNA sequence motifs and containing two adjacent repeating units at least. According to the conservation of the repeated sequence, these repeats are classified as perfect or imperfect repeats [12]. They are common in both prokaryotic and eukaryotic genomes. [13] Also found in both protein coding and non-coding regions of the genome. [14, 15] Microsatellites are abundant in eukaryotic whereas less 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 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 many functional roles to play [24]. DNA repeats exist in one of the following patterns such as, forward repeat, reverse repeat, complement repeat and palindromic repeat. Usually in bacteria, repeats are divided into two subclasses: they are short repeats and longer repeats. In our study we have focused on the first category which constitutes 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 number of whole genome sequencing projects provides a large amount of information which leads to the need of processing algorithms and new tools to examine and classify the obtained sequences [25].

Understanding of repeats

Recent, advances in sequencing technology have contributed to the availability of several organisms. Theoretical studies have shown that evolutions of these repeats are 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 at close proximity to the variation sites, which can make a mechanism of concerted evolution and unequal crossing over. The repeats having significant functional role in causing many neurodegenerative disease and also it is widely used 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 an approach to calculate the simple sequence repeats by using an 'R' packages 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. [Fig-1].

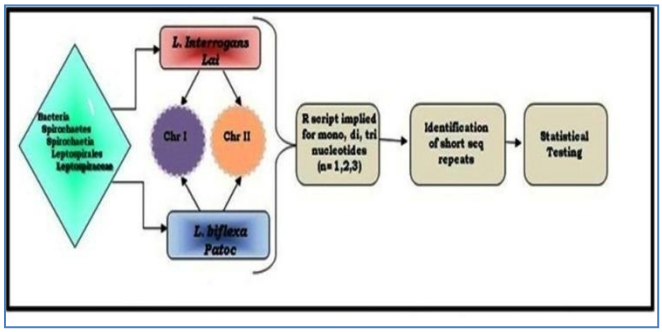

Fig-1 The flowchart showing the outline of the research work

Materials and Methods

***Leptospira* genome sequence retrieval and analysis**

The complete genome 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' script was used to analyse the number of singlet, doublets, triplets and its frequencies for further calculations.

Table-1 *Total Length genomic sequences of pathogenic and non pathogenic Leptospira obtained from NCBI*

Serovar	Source	Length
L. interrogans Lai	chromosome I	4338762
L. interrogans Lai	chromosome II	359372
L. biflexa Patoc	chromosome I	3599677
L. biflexa Patoc	chromosome II	277655

***In silico* identification of short sequence repeats**

Short sequence repeats such as mono, di and trinucleotide repeats were 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 II in mono, di and trinucleotides (n=1, 2, 3) were analyzed. In mononucleotide, A is same as T on a complementary stand, whereas in dinucleotide, (AG) is also equivalent to (GA, TC, CT) and in case of trinucleotide, (AGT) it is equivalent to (TGC, TCA, ACG), in different reading frames or on a complementary strand. In mono and dinucleotide the forward, reverse, complimentary and palindromic sequences were grouped together as they represent the same sequences. The amino acid specified by two or more synonyms was grouped together.

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
 }
}
```

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 was applied to find the SSR in *Leptospira* genome.

a. First step: The calculation was based on the frequency of occurrence P(X) of the base residues such as mono, di and trinucleotides from the analyzed dataset. Whereas 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 [P_{obser} (AB) =P (A) x P (B)]. [26].

$$P_{expect}(AB) = \left[\frac{\text{freq. of occurrence of individual base repeats}}{\text{Sum of freq. of occurrence of all base repeats}} \right] X \text{ Total occurrence of consecutive base repeats}$$

b. Second step: The estimated value (Expected Value) and the frequency of

base repeats of short range interactions were calculated using the formula as follows [26].

c. **Third step:** The formula is $[X^2 (AB) = (P_{obser} (AB) - P_{expect} (AB))^2 / P_{expect} (AB)]$ whereas the calculation is to present the value of all the singlet, doublet and triplets The highly signified x^2 values from the datasets were selected for further analysis and were used to predict the significant frequency.

Results and Discussion

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

The Lipoproteins contribute to the major virulence factors, which 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 their genome size. On comparison of non-pathogenic with pathogenic chromosomes, it is found that chromosome I has an additional of 7, 39,085 base pairs in and 81,717 base pairs in chromosome II when compared to non-pathogenic *Leptospira* genome.

Mononucleotide frequency of chromosome I and II

Complete genome sequences of *Leptospira interrogans* and *Leptospira biflexa* of chromosome I and II (both pathogenic and non-pathogenic) for the occurrence of mononucleotides revealed that, the pathogenic sequences containing higher frequency of A and T when compared to non pathogenic sequences, i.e. A/T repeat units found 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 the frequencies of nucleotides in pathogenic and non-pathogenic 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 repeat of mononucleotides in 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 [Fig-2]. The length and the distribution of the mononucleotide repeats such as A/T seem to have longer. [Fig-3]

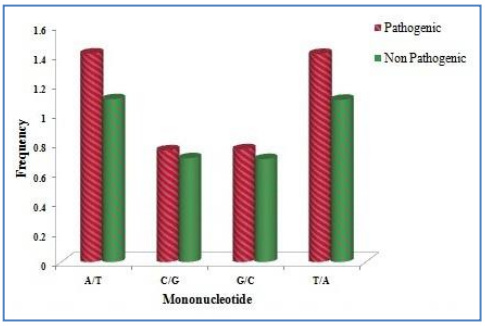

Fig-2 showing the significance of mononucleotide count in pathogenic and non pathogenic sequences of chromosome I: The repeats in the mononucleotide has poly A and T in both the chromosome in comparison to other repeats like C/G

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^2Value	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^2Value	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	0.0300
4	T/A	0.116	0.084	295.30	<.0001	0.0215

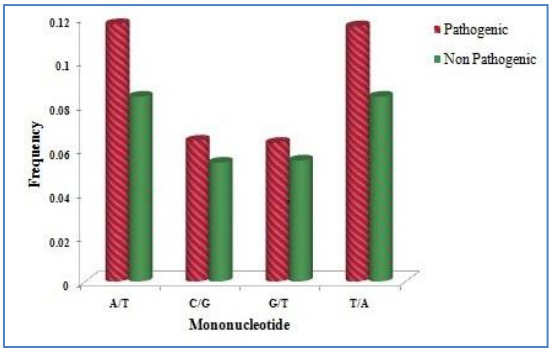

Fig-3 showing the significance of mononucleotide count in pathogenic and non pathogenic sequences of chromosome II: whereas the analysis reveals poly A and T similar to Chromosome I

Dinucleotides frequency of chromosome I and II

Dinucleotide repeats reveals that AA was seen rich in the majority of the chromosomes from the statistical analyzed data; the repeats GG, CG and TC are also found to be more significant. The dinucleotides frequency of chromosome I and II reveals that, the pathogenic sequences contain higher frequency of certain homomeric dinucleotide such as AA, which are found to be 13.55% in pathogenic and 11.98% in non-pathogenic sequences [Fig-4]. Similarly the dinucleotide frequencies of TT in pathogenic are 13.54% and 11.92% in non-pathogenic sequences. In contrast, CC and GG are frequently repeated in non-pathogenic sequences. [Table-5] The GG repeats act as an intra molecular; G-G base pairing between telomere repeats and also 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 [Fig-5]. In all of these five nucleotides, the pathogenic sequences contain more frequencies than non-pathogenic. [Table-6] However, the frequency of CA and TG is also predominant in non-pathogenic sequence. The frequency of AC, AG, CG, GC and GT are less frequent in both the chromosome sequences, whereas AT repeats are the most frequent in general for all the eukaryotic genomes especially in embryophytes, yeast, and fungi. [33]

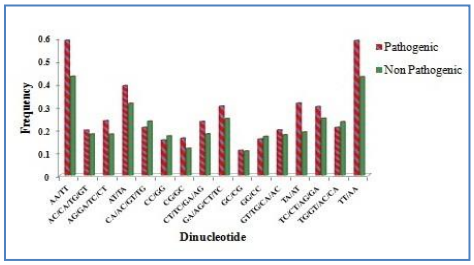

Fig-4 showing the significance of dinucleotides count in pathogenic and non pathogenic sequences of Chromosome I: whereas the pathogenic sequences contain higher frequency of homomeric dinucelotide

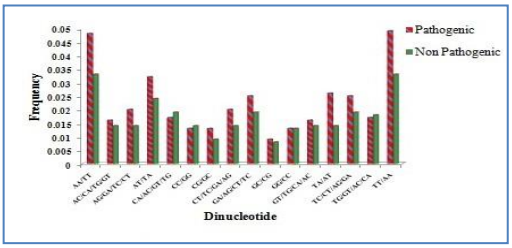

Fig-5 showing the significance of dinucelotide count in pathogenic and non pathogenic sequences of chromosome II: Among all the dimer repeats, AT was found to be prominent. AT is more frequent followed by other repeats similar to chromosome I.

Table-5 *Different repeats of dinucleotides count in pathogenic and non pathogenic sequences of chromosome I*

S. No	Dinucleotide	Pathogenic (n=4.338762)	Non Pathogenic (n=3.599677)	χ^2Value	P Value	Cramer's Value
1.	AA/TT	0.588	0.431	4326.30	<.0001	0.0230
2.	AC/CA/TG/GT	0.196	0.179	897.24	<.0001	0.0106
3.	AG/GA/TC/CT	0.237	0.179	1068.31	<.0001	0.0116
4.	AT/TA	0.390	0.313	205.21	<.0001	0.0051
5.	CA/AC/GT/TG	0.208	0.235	11285.47	<.0001	0.0377
6.	CC/GG	0.153	0.171	7435.48	<.0001	0.0306
7.	CG/GC	0.161	0.117	1202.83	<.0001	0.0123
8.	CT/TC/GA/AG	0.234	0.180	638.75	<.0001	0.0090
9.	GA/AG/CT/TC	0.301	0.247	19.17	<.0001	0.0016
10.	GC/CG	0.108	0.105	1285.26	<.0001	0.0127
11.	GG/CC	0.157	0.169	5925.77	<.0001	0.0273
12.	GT/TG/CA/AC	0.197	0.176	557.08	<.0001	0.0084
13.	TA/AT	0.314	0.188	13491.90	<.0001	0.0412
14.	TC/CT/AG/GA	0.299	0.248	0.97	0.3247	0.0003
15.	TG/GT/AC/CA	0.208	0.233	10387.04	<.0001	0.0362
16.	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^2Value	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

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 in both the chromosomes. [Fig-6] These tri nucleotides have a significant role in the biology of diseases. All the codons codes for amino acids sequences. Trimer repeats such as ACG, CCG/CAG, CGT, and TCG were found to be the maximum ones in both the chromosomes. [Fig-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 of recombinant DNA [34]. Report says that when there is increase in CAG repeats, the individual is affected with Huntington's diseases and if CTG repeats ranges from 50 to 5000 times in the gene it may also lead to Myotonic dystrophy [26, 32]. In case of *Leptospira* it also has the repeat of CAG and CTG which codes for the amino acid Glutamine and Leucine. In future many genetic, Leptospirosis and neurodegenerative disorder can be cured by analysis of Triplets.

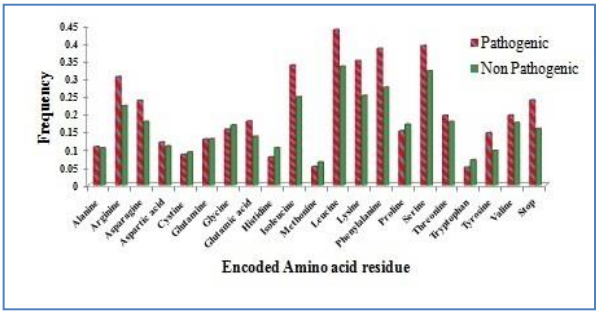

Fig-6 showing the significance of trinucleotides count in pathogenic and non pathogenic sequences of chromosome I: Frequency of Leucine is high compared to other amino acid residues in both the chromosome

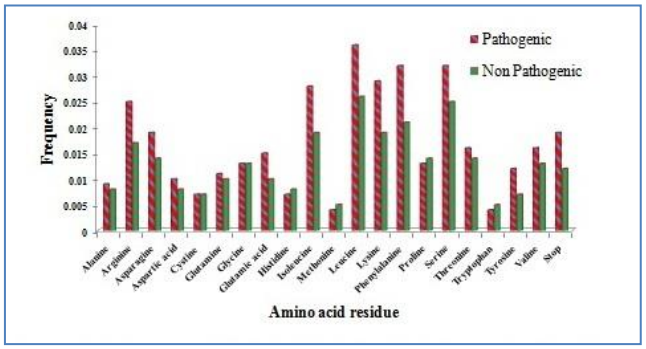

Fig-7 showing the significance of Trinucleotides count in Pathogenic and Non pathogenic sequences of chromosome II encoded by amino acid residue. Frequency of Leucine is high compared to other amino acid residues in both the chromosome

Chi-square result of Mononucleotide, Dinucleotide and Trinucleotides

In all the microsatellite repeats, the majority of frequencies are occupied by pathogenic sequences in both the chromosomes. Among the three SSR, the frequency of mononucleotide repeats are higher than di and trinucleotides. The 'p' value of all the mononucleotide shows they 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 (χ^2=0.97 and p=0.3247) in chromosome I. 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 show 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 differences are more, in these three amino acid sequences. 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.

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^2Value	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	Tri-nucleotide	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

Conclusion

The computational tools and the statistical techniques 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, techniques and approaches have been briefly highlighted in the study. The result of chi-square test, indicates, the 'p' value of mononucleotides to be highly significant which proves the test is asymptotically true, which can be made to an approximate chi-square distribution as closely 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 process of testing for codon repetitions the trinucleotides 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 which it is actually observed and assumed that the hypothesis under consideration is true.

Analysis of the 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 in improving the studies of 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 amplification [32, 33]. This will certainly provide new ideas in deciphering the dynamics of repeats in bacterial genomes and also will provide much information on evolutionary implications.

Conflict of Interest: None declared

References

[1] Rafiei A., Hedayati Z.A., Babamahmoodi F., Alizadeh N.R. and Valadan R. (2012) *J Mazandaran Univ Med Sci*, 22(94), 114–24.
[2] Supply P., Magdalena J., Himpens S. and Locht C. (1997) *Mol Microbiol*, 26(5), 991–1003.
[3] Zavitsanou A. and Babatsikou F. (2008) *Health Sci J*, 2(2), 75–82.
[4] Zuerner R.L. and Alt D.P. (2009) *J Clin Microbiol,* 47(4), 1202–5.
[5] Soltanimajd N., Khodaverdidarian E., Khaki P., Moradi B.S., Yahaghi E. and Mirnejad R. (2012) *Asian Pac J Trop Dis*, 2, 550–52
[6] Bharti A.R., Nally J.E., Ricaldi J.N., Matthias M.A., Diaz M.M. and Lovett M.A. (2003) *Lancet Infect Dis*, 3, 757–71.
[7] Levett P.N. (2001) *Clin Microbiol Rev,*14,296–326
[8] VanBelkum A., Scherer S., Van Alphen L. and Verbrugh H. (1998) *Microbiol Mol Biol Rev*, 62, 275–93.
[9] Available from http://tandem.bu.edu/
[10] Benson G. (1999). *Nucleic Acids Res*, 27, 573–80.
[11] Roa S.R., Trivedi S., Emmanuel D., Merita K. and Hynniewta M. (2010) *Journal of Cell and Molecular Biology, 7* (2) & 8 (1), 1–11.
[12] Lopes J., Ribeyre C. and Nicolas A. (2006) *Mol Cell Biol*, 26, 6675–89.
[13] Usdin K. and Kumari D. (2015) *Frontiers in Genetics*, 6, 1–7
[14] Toth G., Gaspari Z. and Jurka J. (2000) *Genome Res*, 10, 967–81.
[15] Katti M.V., Ranjekar P.K. and Gupta V.S. (2001) Mol. *Biol. Evol*, 18(7), 1161– 67.
[16] Le Fleche P., Hauck Y., Onteniente L., Prieur A., Denoeud F., Ramisse V., Sylvestre P., Benson., Ramisse F. and Vergnaud G. (2001) *BMC Microbiol*, 1:2.
[17] Richard F.G. Kerrest A. and Dujon B. (2008) *Microbiol. Mol. Biol*, 72(4), 686–727.
[18] Ugarkovic D. and Plohl M. (2002) *The EMBO journal*, 21, 5955–9.
[19] Mayer C., Leese F. and Tollrian R. (2010) *BMC Genomics*, 11, 277.
[20] Gemayel R., Cho J., Boeynaems S. and Verstrepen K.J. (2012) *Genes*, 3, 461–80.
[21] Gulcher J. (2012) *Cold Spring Harb Protoc*, 4, 425–32.
[22] Yan H.M., Dong C., Zhang E.L., Tang CF., A X.X., Yang W.Y, Yang Y.Y., Zhang F.F. and Xu F.R. (2012) *Yi Chuan*, 34(1), 87–94.
[23] Bensen G. (1998) *Nucleic Acids Res*, 27(2), 573–80.
[24] Gangwal K. and Lessnick S.L. (2008) *Cell Cycle*, 7(20), 3127–32.
[25] Gusfield D. (1999) *Cambridge University Press*.
[26] Pavithra V., Surendar and Mugilan S. (2014) *Journal of Pharmacy Research*, 8(3), 359–62.
[27] Available from http://www.ncbi.nlm.nih.gov.
[28] Ihaka R. and Gentleman R. (1996) *J Comp Graph Stat*, 5, 299–314.
[29] Available from http://vassarstats.net/
[30] Cinco M. (2010) *New Microbiologica,* 33, 283–92.
[31] Setubal J.C., Reis M.G., Matsunaga J. and Haake D.A. (2006) *Microbiology*, 152, 113–21.
[32] Achaz G., Rocha E.P.C., Netter P. and Coissac E. (2002) *Nucleic acids research,* 30(13), 2987–94.
[33] Gabor Toth, Zoltan Gaspari and Jerzy Jurka (2000) *Genome Res.,* 10(7): 967–981.
[34] Hengen P. (1995) *Trends in Biochemical Sciences*, 20(7), 285–6.