ISSN No: - 2456 – 2165

CpG Frequency Analysis in Human Genome Using R Programming

S. Balamurugan* and Dr. S Prasanna
Department of Computer Applications, Vels University,
Pallavaram, Chennai – 600 117, Tamil Nadu, India
sivabala76@gmail.com, prasanna.scs@velsuniv.ac.in

Abstract:-In the present information society, biological data are enormously increasing. After the successful achievement of Human Genome Project (HGP), the complete human genome sequence (reference genome) was made available in online resources for bioinformatics tools and services. Those data sets are large and complex (Big data), but analyzing and comparing these massive amount of genomic sequences paves way to understand the complex diseases to find the personalized medicines. Consequently, it is very important that every individual should know about human genome. Resources to access human genome are enormous, but only the researchers can access those resources by providing appropriate keywords and is not possible for a biologist with little computer knowledge. Therefore, a convenient and instinctual data interfaces are important to easily access, download, visualize and analyze human genomic data. With these requirements in mind, the present study proposed a web applicationdevelopment using "R programming along with Bioconductor packages" for CpG site frequency analysis of CpG Island, CpG non-island and CpG Island Shores & shelves (Downstream and Upstream) in human reference genome.Later the App will be hosted for the user's further analysis.

Keywords: Human Genome; Human Genome Project; CPG Island; R and Bioconductor.

I. INTRODUCTION

Frequency analysis, that the possibility of four different nitrogenous base pairs (A, T, G, C) on 22 body and XY sex chromosomes of human reference genome is the first and basic analysis.[1]The second important basic feature is dinucleotide frequencies such as GC and AT content analysis.[2]Analyzing the frequency of GC content is important, because the GC-content determines the stable nature of DNA and length of the coding sequence is directly proportional to higher GC content.[3] Similarly,CpG sites are the regions of DNA where a Cytosine is followed by a Guanine linearly.[4] Studying methylation in CpG site is important that they act as gene markers and involve in gene regulation. They are playing key role in disease onset through silencing or expressing particular gene action. Based onCpG frequencies, CpGs are classified as CpG Island (CGI), Non CpG Island (non-CGI), CpG Island shores (CGI shores) and CpG Island shelves (CGI shelves).[5]

A. CpG Island (CGI)

CGIs are the genomic regions (~1000 base pairs long)with high frequencies of CpG sites in a GC-rich sequence. The "p" in CpG refers to the phosphodiester bond between Cytosine and Guanine, which indicates that C and G are next to each other in a sequence. [6,7,8] In humans, 40% of CGIs are found in gene promoters and the remaining 60% have been termed "orphan CGIs" in the remaining portion of the sequence. [9]

B. Algorithms for CGI Identification

Gardiner-Garden and Frommer's (1987)[10] algorithm criteria include: length over 200 base pairs, over 50% GC pairs, and a ratio of observed to expected number of CpG dinucleotides over 0.60.Takai and Jones (2002)[11] algorithm criteria include:length over 500 base pairs, over 55% GC pairs, and a ratio of observed to expected number of CpG dinucleotides over 0.65. Another algorithm, CpGcluster detects CpG clusters through statistical significance based on the physical distance between neighboring CpGdinucleotides in a chromosome. [12] According to the introduction of the three algorithms above, the present study considered Gardiner-Garden and Frommer algorithm as a major algorithm.

C. CGI Shores and Shelves (Up and Downstream)

CGIs are interspaced by long stretches of highly methylated CpG-poor regions that are found both within and between genes. One can find the "CGI shore" from 0 to 2 kb on either side of a CGI flanking regions and the "CGI shelf" from 2 to 4kb on either side of a CGI flanking regions. [9]

D. Non CpG Island (Non-CGI)

Non CpG Island is the upstream and downstream regions to the CpG Island where CpG sites are absent or not with the specified conditions as in the above said algorithms.In non-

ISSN No: - 2456 – 2165

CGI, CpG site is not located in a promoter, a gene body, a CGI, a CGI shore or a CGI shelf and are located in the "open sea". As like in CpG island, DNA methylation is also found at Non-CpG island and are also important in Epigenetic Gene Regulation and Brain Function. [13]

There are number of software available to predict CpG islands. CpGProDis for identifying mammalian promoter regions associated with CpG islands in large genomic sequences. [14]Newcpgseek is another tool to scores each position of CpG site in the sequence using a running sum calculated from all positions in the sequence, starting with the first and ending in the last.[15] The tools cpgplot and newcpgreportis to identifyCpG islands in one or more nucleotide sequences. [15]CgiHunter is a tool used for CpG island annotation and it has been proven to identify all genome regions (http://cgihunter.bioinf.mpi-inf.mpg.de).CpGPAP (CpG island Predictor Analysis Platform) is a web-based application that provides an interface for predicting CpG islands in genome sequences or in user input sequences. [16]But these tools can be accessed mostly with the bioinformaticians. To our knowledge, there is no separate web application available for visualizing CpG sites in different regions of Human genome. The present study proposed a web application development to provide visualization and comparison of CGI vs non-CGI, CGI vs CGI shores and CGI vs CGI shelves.

II. MATERIALS AND METHODS

The proposed web application development in the present study is written entirely in the open-source R programming language. [17] The methodology adopted for the SAFA-HG App development description is as follows.

A. Data Resources

The experimental data, used for the present study were retrieved from UCSC for the CpGfrequency analysis. For CpG Island analysis the table in *.txt ("cpgislandExtUnmasked.txt") format was downloaded. Similarly, to analyze CpG Island Shores and Shelves, the same table in *.bed ("cpgislandExtUnmasked.bed)" file format was downloaded from UCSC genome browser (https://genome.ucsc.edu/cgi-bin/hgTables) using the Library (BSgenome.Hsapiens.UCSC.hg19).

a). R Programming Language Related Interfaces and Packages

R is an open source programming language and software environment for statistical computing and graphics. RStudio is an integrated development environment (IDE) for R that provides an alternative interface to R. In the present study, RStudio and its dependent Bioconductor packages are used for nucleotide frequency (CpG) analysis on Human reference genome. Bioconductor is a collection of R packages for the analysis and comprehension of high-throughput genomic data. It consists of 1296 software packages, 309 experiment data packages, and 933 up-to-date annotation packages. The Bioconductor project provides a data packages like "Biostrings" and "BSgenome" that helps to access the full genome sequences of a given organism from data resources for sequence analysis. These packages are called Biostrings-based genome data packages and require the BSgenome package to work properly. As mentioned earlier, the full genome sequences for Human as provided by UCSC (hg19) stored in Biostrings objects. [18]Biostrings is a memory efficient string container, string matching algorithms, and other utilities, for fast manipulation of large biological sequences or sets of sequences. RSQLite embeds the SQLite database engine, providing a DBI-compliant interface. The DBI package defines a common interface between the R and database management systems (DBMS). Sqldf () transparently sets up a database, imports the data frames into that database. These packages assist in managing data in R environment.

III. RESULTS AND DISCUSSION

After Mendel's discovery in Genetics, more than 6000 genetic disorders have been studied, but still we do not have a clear understanding of many of their roles in health and diseases. [19]Due to the revolution of Human Genome Project and Personal Genomics, for the past few years, the size of Human Genome information in the data resources like NCBI and PGP has grown exponentially. Due to this increase, the data of human genome, is one of the important Big data source exist today. Research contribution of the fields such as computational biology, bioinformatics and systems biology mostly involve in analyzing these genomic data to improve human health. With this genomic revolution, it is possible in future that, all the individuals may have their own genome information as their personal medical card and genetic passport which will be the reference to the physician to have personalized medicine. [20,21]So, it is important to know that everyone should have knowledge on Human genome and its features. In this regard, the present study proposed an interactive web application development on CpG analysis in Human genome using R programming.

This application has the power and flexibility to be resident on a local computer or serve as a web-based environment, enabling easy sharing and visualization of data to the biological researchers with little computer knowledge. Unlike the traditional system (download the data and stored in a local hard drive for analysis), the present study acquired data (real time data) from online as the size of human genome data is too big. The results in the form of tables and plot images

(histograms) of nucleotide (CpG) frequency analysis are shown and described as follows.

A. CpG Island and Non-Island Tables

User can view the list of CpG Island details for minimum 10 and maximum 50 entries in a sliding window screen. The user can also find the details by providing this keyword in the search box. In such a way, Chromosome 1 has 4332 entries covered in 280 pages,Chromosome 2 has 3464 entries covered in 231 pages, Chromosome 3 has 2681 entries covered in 179 pages,Chromosome 4 has 2473 entries covered in 165 pages, Chromosome 5 has 2637 entries covered in 176 pages,Chromosome 6 has 2647 entries covered in 177 pages,Chromosome 7 has 2841 entries covered in 190 pages, Chromosome 8 has 1982 entries covered in 133 pages,Chromosome 9 has 2308 entries covered in 154 pages,Chromosome 10 has 2095 entries covered in 140 pages, Chromosome 11 has 2295 entries covered in 153 pages,Chromosome 12 has 2429 entries covered in 162 pages, Chromosome 13 has 1429 entries covered in 96 pages,Chromosome 14 has 1540 entries covered in 103 pages,Chromosome 15 has 1456 entries covered in 98 pages, Chromosome 16 has 2176 entries covered in 146 pages,Chromosome 17 has 2505 entries covered in 167 pages, Chromosome 18 has 1089 entries covered in 73 pages,Chromosome 19 has 3275 entries covered in 219 pages, Chromosome 20 has 1288 entries covered in 86 pages, Chromosome 21 has 682 entries covered in 46 pages,Chromosome 22 has 1036 entries covered in 70 pages, Chromosome X has 1945 entries covered in 130 pages and the Chromosome Y has 424 entries covered in 29 pages. The end users can be easily visualizing the CpG site in the CGI regions. For a reference, the screenshot of the first pages of CGI and non-CGI of chromosome 1 are shown in Figure 1 and 2, respectively.

B. CpG Island Plots

In the obtained histograms, the Island frequencies are plotted in red and non-island frequencies are plotted in blue color. The comparative plot for the chromosomes 1 is shown in Figure 3. In the plot, the frequencies are not normally distributed. The graphs show the density plots, which can allow us to easily review the whole distribution of CGI and non-CGI data. The graph was plotted after satisfied with Gardiner-Garden and Frommeralgsorithm. The users can use these outputs (both graphical and statistical) for the further CGI and non-CGI analysis.

C. CpG Island and CpG Island Shores (Up and Downstream) Plot

Shores are the regions immediately flanking CpG islands (CGI) the consensus definition of a CpG shore is up to 2kbp away from the CGI. Methylation in CGI shores (both upstream and downstream) is more responsible for many diseases. Therefore, the identification and analyzing the CGI shores compare with CpG Island is important. In the proposed web application development, the present study compared CpG Island and CpG Island shores (Up and Downstream) and the plot for Chromosome 1 and is shown in Figure 4. In the density plot, red color indicates CGI frequencies, dark blue color indicates CGI shores downstream and light blue color indicates CGI shores upstream. As like previous, in these plots also the frequency distribution is not normalized.CpG sites information in shores region can be retrieved and visualized and the significance of methylation in context with many diseases can be studied by the end users.

D. CpG Island and CpG Island Shelves (Up and Downstream) Plot

CpG Shelves are defined as the 2 kb outside of a shore's flanking regions. As like the previous plots, the comparison of CGI frequencies with CGI shelves (both up and downstream) is constructed and the plot for chromosome 1 is shown in Figure 5. In the density plot, red color indicates CGI frequencies, dark blue color indicates CGI shelves downstream and light blue color indicates CGI shelves upstream. In these plots, also the frequency distribution is not normalized. Similar to CpG Island shores, the users (biologist) who are interested to work with major diseases like cancer can also concentrate on CpG Island shelves regions in both upstream and downstream.

IV. CONCLUSION

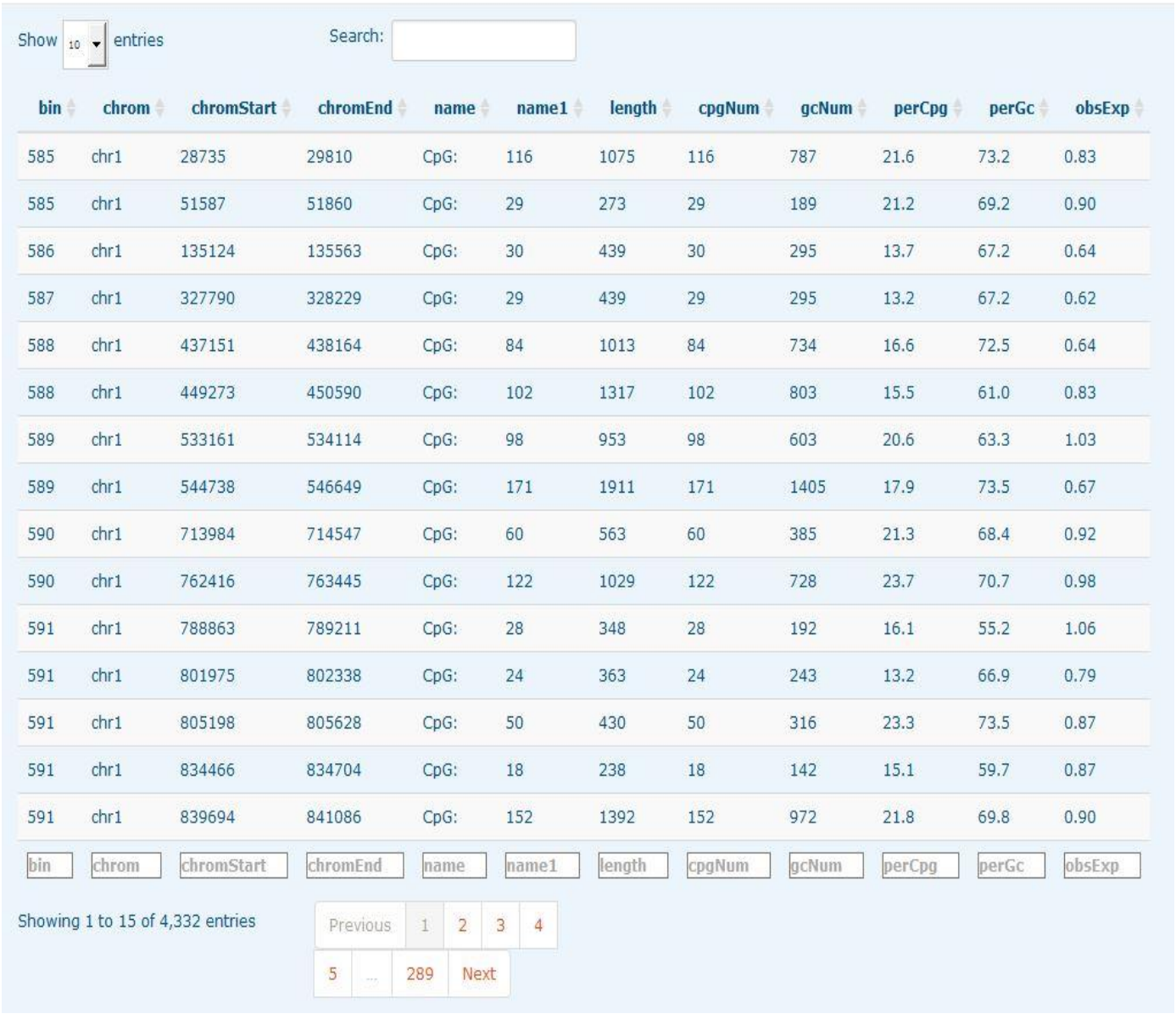
In the present study, the web application is developed is proposed for frequency analysis on CpG sites in Human reference genome. The methodology followed is summarized in the section "Materials and Methods". Through this App, visualization, download and analysis of CpG Island data can be done by the end user for their further analysis. The development of this App is ongoing and we intend to add to improve upon the visualization and analysis features. The tool will be well developed with the improved facilitates that to predict CpG sites responsible for diseases like cancer.Apart from CGI and its shores, shelves and Non-CGI, the work on CpG canyon, CpG ocean, Gene body and Gene desert will be done.In future, the work will be continued on the above-mentioned regions and the App will be maintained.

REFERENCES

[1]. Louie E et al., "Nucleotide Frequency Variation Across Human Genes", Genome Research. 2003; 13(12):2594-2601.
[2]. Beleza Yamagishi ME and Shimabukuro AI, "Nucleotide Frequencies in Human Genome and Fibonacci Numbers", Bulletin of Mathematical Biology, 2008; 70(3): 643.
[3]. Vinogradov AE., "DNA helix: the importance of being GC-rich", Nucleic Acids Research, 2003; 31(7):1838-1844.
[4]. Sharif J et al., "Divergence of CpG island promoters: a consequence or cause of evolution?", Dev Growth Differ, 2010; 52(6):545-554.
[5]. Edgar R et al., "Meta-analysis of human methylomes reveals stably methylated sequences surrounding CpG islands associated with high gene expression", Epigenetics & Chromatin, 2014; **7**:28,1-12.
[6]. Bird AP, "CpG rich islands and the function of DNA methylation", Nature, 1986; 321: 209-213.
[7]. Larsen F et al., "CpG islands as gene markers in the human genome", Genomics,1992; 13: 1095-1107.
[8]. Han L et al., "CpG island density and its correlations with genomic features in mammalian genomes", Genome Biology, 2008; 9(5):R79:1-12.
[9]. Cooper DN et al., "Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides", Human Genomics, 2010; 4(6):406-410.
[10]. Gardiner-Garden M and Frommer M, "CpG islands in vertebrate genomes", J. Mol. Biol.1987; 196(2), 261.
[11]. Takai D and Jones P, "Comprehensive analysis of CpG islands in human chromosomes 21 and 22", Proc. Natl Acad. Sci.,2002; 99(6), 3740–3745.
[12]. Hackenberg M et al., "CpGcluster: a distance-based algorithm for CpG-island detection", BMC Bioinform, 2006; 7, 446.
[13]. Jang HS et al., "Review on CpG and Non-CpGMethylation in Epigenetic Gene Regulation and Brain Function", genes,2017; 8(148):1-20.
[14]. Ponger L and Mouchiroud D, "CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences", Bioinformatics,2002; 18(4), 631–633.
[15]. Rice P et al., "EMBOSS: the European Molecular Biology Open Software Suite", Trends Genet,2000; 16(6):276-277.
[16]. Chuang LY et al., "CpGPAP: CpG island predictor analysis platform", BMC Genetics, 2012; 13:13.
[17]. R Core Team, "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna, Austria, 2015.
[18]. Pagès H et al., "Biostrings: String objects representing biological sequences, and matching algorithms", R package version, 2017; 2.44.1.
[19]. Rehm HL et al., ClinGen--the Clinical Genome Resource, N Engl J Med. 2015; 4;372(23):2235-2242.
[20]. Akgun M, "Privacy preserving processing of genomic data: A survey", Journal of Biomedical Informatics,2015; 56: 103-111.
[21]. Baranov VS, "Genome paths: A way to personalized and predictive medicine", Acta Naturae, 2009.

ISSN No: - 2456 – 2165

LIST OF FIGURES



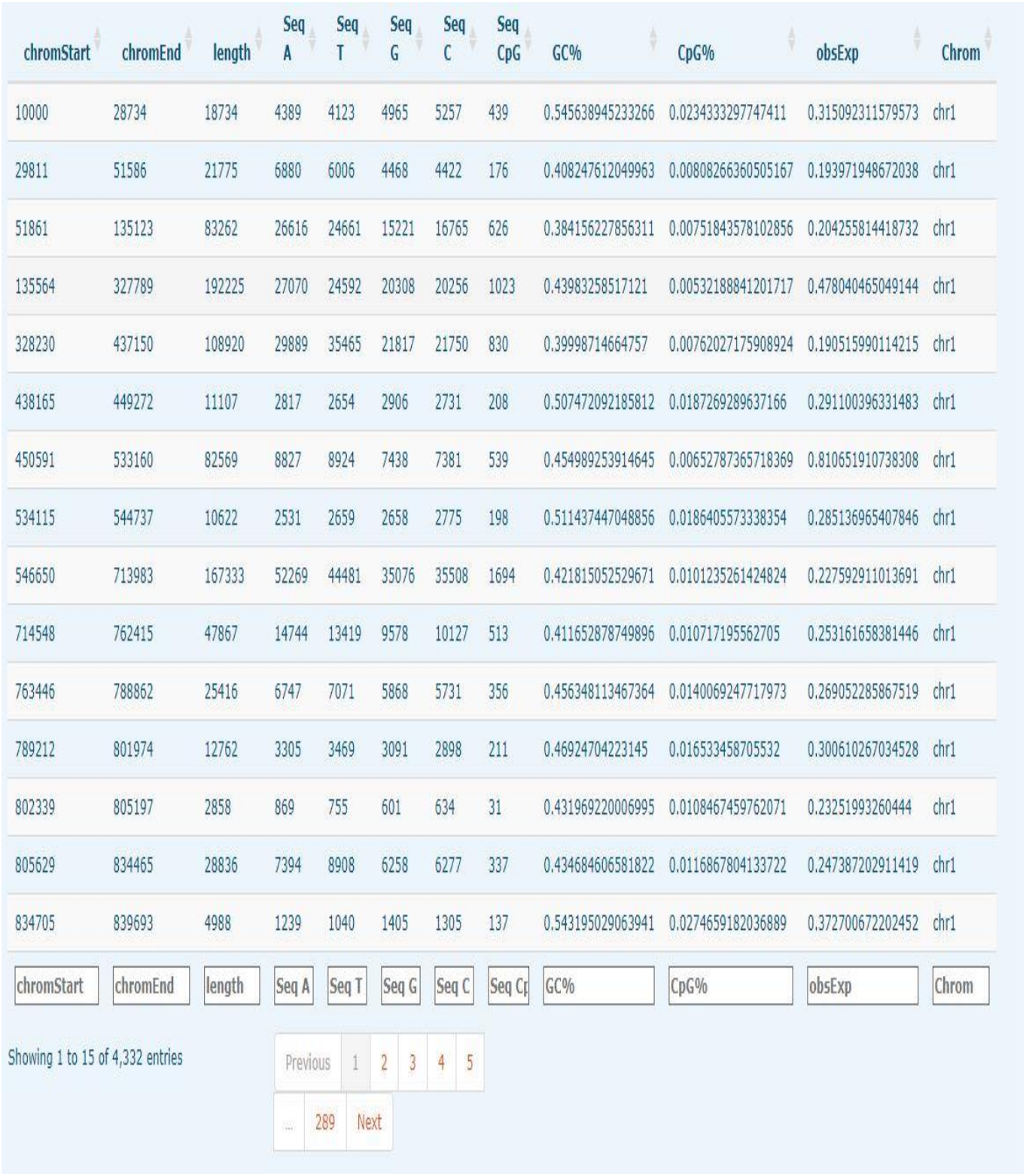
Show 10 entries Search:

bin	chrom	chromStart	chromEnd	name	name1	length	cpgNum	gcNum	perCpg	perGc	obsExp
585	chr1	28735	29810	CpG:	116	1075	116	787	21.6	73.2	0.83
585	chr1	51587	51860	CpG:	29	273	29	189	21.2	69.2	0.90
586	chr1	135124	135563	CpG:	30	439	30	295	13.7	67.2	0.64
587	chr1	327790	328229	CpG:	29	439	29	295	13.2	67.2	0.62
588	chr1	437151	438164	CpG:	84	1013	84	734	16.6	72.5	0.64
588	chr1	449273	450590	CpG:	102	1317	102	803	15.5	61.0	0.83
589	chr1	533161	534114	CpG:	98	953	98	603	20.6	63.3	1.03
589	chr1	544738	546649	CpG:	171	1911	171	1405	17.9	73.5	0.67
590	chr1	713984	714547	CpG:	60	563	60	385	21.3	68.4	0.92
590	chr1	762416	763445	CpG:	122	1029	122	728	23.7	70.7	0.98
591	chr1	788863	789211	CpG:	28	348	28	192	16.1	55.2	1.06
591	chr1	801975	802338	CpG:	24	363	24	243	13.2	66.9	0.79
591	chr1	805198	805628	CpG:	50	430	50	316	23.3	73.5	0.87
591	chr1	834466	834704	CpG:	18	238	18	142	15.1	59.7	0.87
591	chr1	839694	841086	CpG:	152	1392	152	972	21.8	69.8	0.90
bin	chrom	chromStart	chromEnd	name	name1	length	cpgNum	gcNum	perCpg	perGc	obsExp

Showing 1 to 15 of 4,332 entries Previous 1 2 3 4 5 ... 289 Next

Figure 1: Table View of CGI Details Along With Number of Cpgs, Obs/Exp Ratio Etc of Chromosome 1.

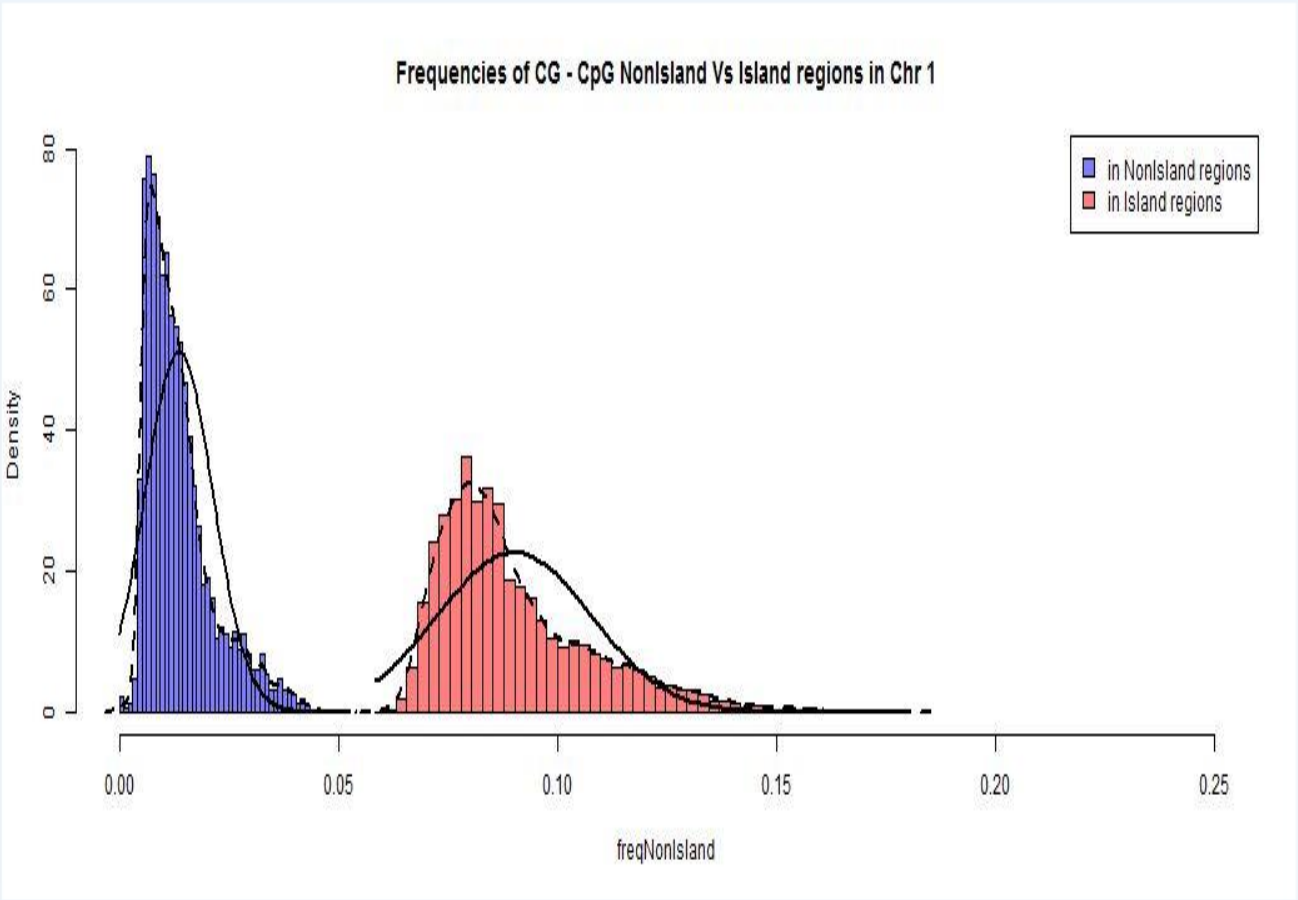
ISSN No: - 2456 – 2165



chromStart	chromEnd	length	Seq A	Seq T	Seq G	Seq C	Seq CpG	GC%	CpG%	obsExp	Chrom
10000	28734	18734	4389	4123	4965	5257	439	0.545638945233266	0.0234333297747411	0.315092311579573	chr1
29811	51586	21775	6880	6006	4468	4422	176	0.408247612049963	0.00808266360505167	0.193971948672038	chr1
51861	135123	83262	26616	24661	15221	16765	626	0.384156227856311	0.00751843578102856	0.204255814418732	chr1
135564	327789	192225	27070	24592	20308	20256	1023	0.43983258517121	0.00532188841201717	0.478040465049144	chr1
328230	437150	108920	29889	35465	21817	21750	830	0.39998714664757	0.00762027175908924	0.190515990114215	chr1
438165	449272	11107	2817	2654	2906	2731	208	0.507472092185812	0.0187269289637166	0.291100396331483	chr1
450591	533160	82569	8827	8924	7438	7381	539	0.454989253914645	0.00652787365718369	0.810651910738308	chr1
534115	544737	10622	2531	2659	2658	2775	198	0.511437447048856	0.0186405573338354	0.285136965407846	chr1
546650	713983	167333	52269	44481	35076	35508	1694	0.421815052529671	0.0101235261424824	0.227592911013691	chr1
714548	762415	47867	14744	13419	9578	10127	513	0.411652878749896	0.010717195562705	0.253161658381446	chr1
763446	788862	25416	6747	7071	5868	5731	356	0.456348113467364	0.0140069247717973	0.269052285867519	chr1
789212	801974	12762	3305	3469	3091	2898	211	0.46924704223145	0.016533458705532	0.300610267034528	chr1
802339	805197	2858	869	755	601	634	31	0.431969220006995	0.0108467459762071	0.23251993260444	chr1
805629	834465	28836	7394	8908	6258	6277	337	0.434684606581822	0.0116867804133722	0.247387202911419	chr1
834705	839693	4988	1239	1040	1405	1305	137	0.543195029063941	0.0274659182036889	0.372700672202452	chr1

Figure 2: Table View of Non-CGI Details Along With Number of Cpgs, Obs/Exp Ratio Etc of Chromosome 1.

ISSN No: - 2456 – 2165

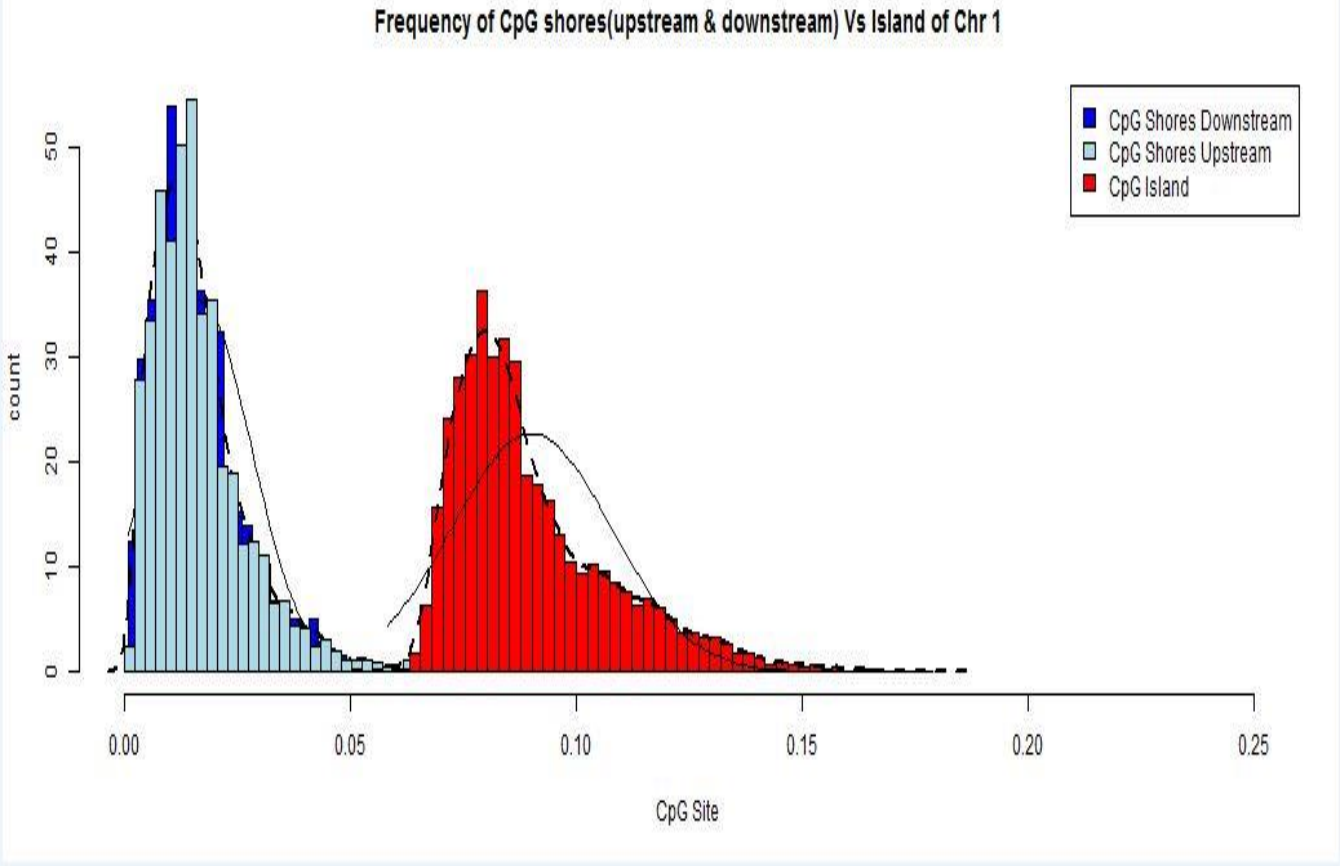


Summary & Structure of table observation:

```
[1] "The Structure & Summary of FreqIsland"
 num [1:4332] 0.1078 0.1058 0.0682 0.0659 0.0828 ...
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.05825 0.07767 0.08519 0.09015 0.09877 0.17870
[1] "The Structure & Summary of FreqNonIsland"
 num [1:4333] 0.02343 0.00808 0.00752 0.00532 0.00762 ...
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.000000 0.007872 0.011620 0.013630 0.016680 0.050690
```

Figure 3: Comparative Histogram Plot of CGI and Non-CGI of Chromosome 1.

ISSN No: - 2456 – 2165

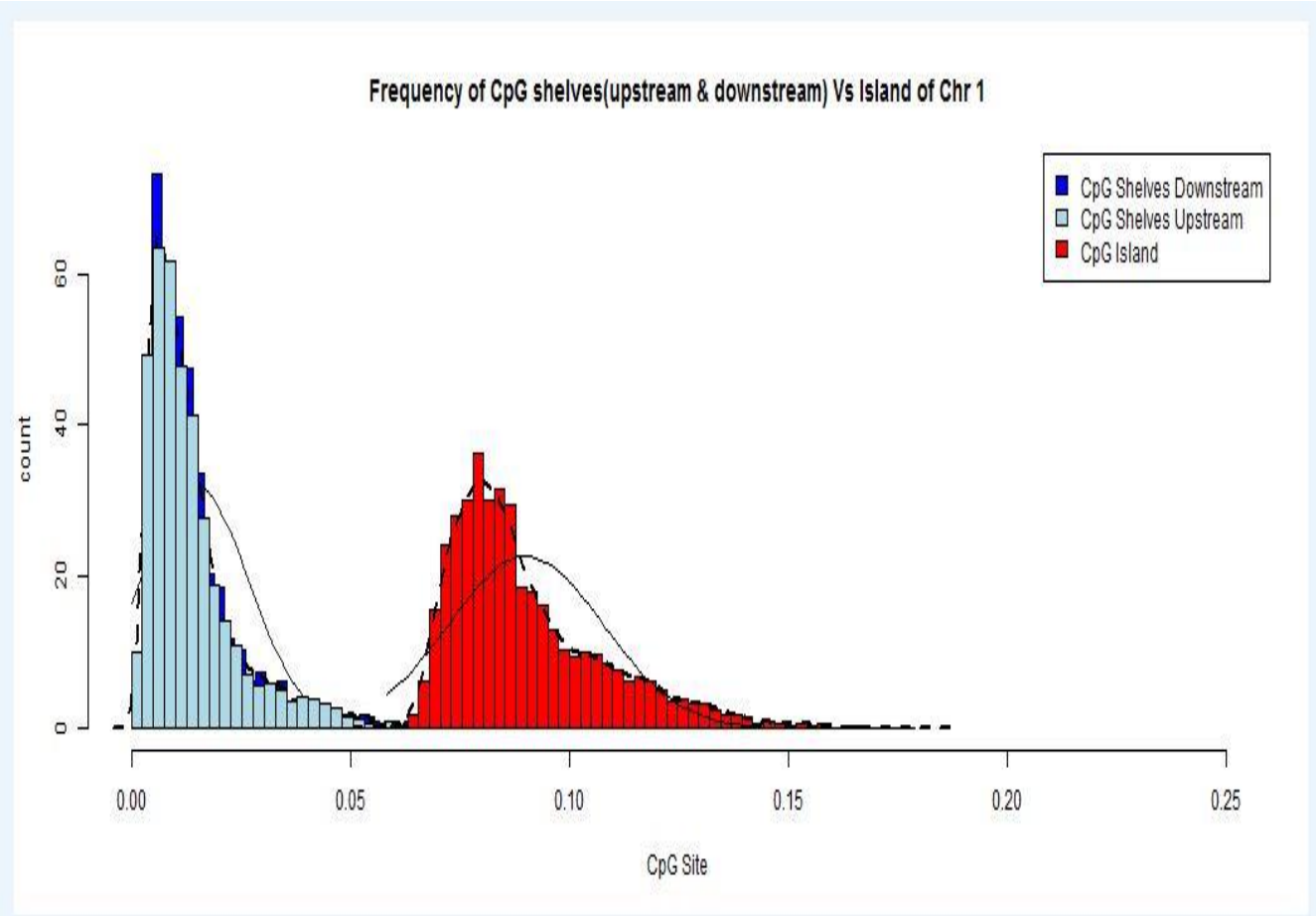


Summary & Structure of table observation:

```
[1] "The Structure & Summary of CpG Shore Downstream"
 num [1:4334] 0.007 0.0215 0.0085 0.008 0.0145 ...
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0009995 0.0089960 0.0144900 0.0168800 0.0214900 0.1039000
[1] "The Structure & Summary of CpG Shore Upstream"
 num [1:4334] 0.004 0.0175 0.0085 0.004 0.0165 ...
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.000000 0.008996 0.014490 0.016920 0.020990 0.111900
```

Figure 4: Comparative Histogram Plot of CGI and CGI Shores (Both Up and Downstream) of Chromosome 1.

ISSN No: - 2456 – 2165



Summary & Structure of table observation:

```
[1] "The Structure & Summary of CpG Shelves Downstream"
 num [1:4334] 0.005 0.0045 0.0175 0.003 0.0125 ...
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.000000 0.006497 0.010990 0.014350 0.017490 0.115900
[1] "The Structure & Summary of CpG Shelves Upstream"
 num [1:4334] 0.0035 0.0105 0.01 0.019 0.0125 ...
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.000000 0.006997 0.010990 0.014450 0.017490 0.123900
```

Figure 5: Comparative Histogram Plot of CGI and CGI Shelves (Both Up and Downstream) of Chromosome 1.