Comparison of methylation resistant and sensitive CpG islands using DNA sequence information content

A Dissertation

Submitted in partial fulfillment of the requirement

For the award of degree of

Masters of Technology

In

Biotechnology

Under the guidance of

Dr. Vikas Handa

Assistant Professor

Submitted by

Shivangi

Roll no. 601404019

DEPARTMENT OF BIOTECHNOLOGY

THAPAR UNIVERSITY

PATIALA-147004

July 2016

CANDIDATE DECLARATION

I hereby declare that the work being presented in the M.Tech dissertation entitled "**Comparison of methylation resistant and sensitive CpG islands using DNA sequence information content**" has been carried out by me under the guidance of Dr. Vikas Handa, Associate Professor, Department of Biotechnology, Thapar University, Patiala. Further, I declare that I have not submitted the matter embodied in this dissertation for the award of any other degree or any other qualification of any university or examining body in India/elsewhere.

Shivangi

Shivangi
M.Tech Biotechnology
Roll. No. 601404019

Date: 15.7.16
Place: Patiala

CERTIFICATE

This is to certify that dissertation entitled **"Comparison of methylation resistant and sensitive CpG islands using DNA sequence information content"** submitted by **Shivangi (601404019)** in partial fulfillment of the requirements for the award of Masters of Technology in Biotechnology to Thapar University, Patiala is an authentic work carried out by her under my supervision and guidance.

To the best of our knowledge, the matter embodied in this dissertation has not been submitted to award of any Degree or certificate in any other university/institute.

15.07.2016

Dr. Vikas Handa
Assistant Professor
Department of Biotechnology
Thapar University
Patiala

Dr. Dinesh Goyal
Head of Department
Department of Biotechnology
Thapar University
Patiala

Dr. S.S Bhatia
Dean (Academic Affairs)
Thapar University
Patiala

CONTENTS

CANDIDATE DECLARATION……..……i

CERTIFICATE…….ii

CONTENTS…….iii

ABBREVIATIONS.....iv

LIST OF FIGURES.....v

LIST OF TABLES.....vi

ABSTRACT.....vii

Chapter 1. Introduction.....1-6

Chapter 2. Review of Literature.....7-11

Chapter 3. Scope of Study.....12-13

Chapter 4. Objectives.....14-15

Chapter 5. Materials and Methods.....16-34

Chapter 6. Result.....35-45

Chapter 7. Discussion.....46-48

Chapter 8. Conclusion.....49-50

Chapter 9. References.....51-53

ABBREVIATIONS

Abbreviations	Full Forms
A	Adenine
C	Cytosine
C_5	5- position of Cytosine
CpA	Cyotsine binded to Adenine by a phosphodiester bond
CpG	Cytosine binded to Guanine by a phosphodiester bond
DNA	Deoxyribonuclic Acid
Dnmt	DNA methyltransferases
mC	Methyl cytosine
PCR	Percentage compression ratio

LIST OF FIGURES

Figure 1: DNA methylation.....2

Figure 2: Mechanism of methylation.....3

Figure 3: Sequence downloaded from NCBI.....22

Figure4: Macro recorded for the split of RY using Notepad++.....23

Figure 5: Macro recorded for the split of YR using Notepad++.....23

Figure 6: Calculation o f μ, σ and z.....25

Figure 7: Pairwise sequence alignment using Needleman Wunsch algorithm.....31

Figure 8: Compression length for the compression scheme.....31

Figure 9: Compression scheme.....35

LIST OF TABLES

Table 1: Nucleotide sequence of chromosome 21 and their length and the methylation status.....17

Table 2: Microsoft excel tools.....21

Table 3: Notepad++ tools.....21

Table 4: Replacement of Nucleotides.....22

Table 5: Compression of motif repeats.....25

Table 6: Replacement scheme for the motif repeats.....26

Table 7: Compression scheme.....30

Table 8: Binary replacement of nucleotides.....32

Table 9: Description of figure 3.....33

Table 10: Z-values for resistant CpG islands.....38

Table 1: Average z-values of resistant CpG islands and sensitive CpG Islands.....42

Table 12: PCR for sensitive CpG islands.....43

Table 13: PCR for resistant CpG islands.....44

Table 14: t-test result.....45

ABSTRACT

The organization of eukaryotic DNA is highly diversified and it includes certain CpG rich sequences called CpG islands. These CpG islands are found to be unmethylated usually. The methylation is occurs at the dinucleotide CpG. Most of the CpG islands are known to be associated with the promoter region of the genes and are usually unmethylated. Lately, it has been found out that CpG Islands exhibit varying propensity to get methylated. They can be fully methylated or unmethylated, thereby may be considered as methylation sensitive or methylation resistant. In the present work, a comparative study of the methylation sensitive and methylation resistant is done based on the two attributes *i.e.*, randomness and compression. Both the attributes were statistically analyzed and thus the result can be speculated to influence the methylation of CpG Islands.

CHAPTER-1

INTRODUCTION

INTRODUCTION

DNA is the genetic material present in all living organism (except RNA viruses) that carries the genetic information from one generation to next generation in the form of sequence of four bases *i.e.* Adenine, Cytosine, Guanine and Thiamine. Point mutations (insertion, deletion, substitution, inversion and translocation) lead to alteration of DNA sequences which causes the genetic changes. There are some heritable changes that are not related to alterations of the DNA sequences and such changes come under Epigenetics. The word Epigenetics was given by Conrad Waddington in 1940. Epigenetics is defined as heritable changes in gene activity and expression without the change in DNA sequence (Holliday, 2006). In multicellular organisms all cells contain the same genetic information, not all of them are expressed by the cell types. Epigenetic involves several modifications of histone proteins *i.e.* acetylation, phosphorylation and methylation etc which further affects the gene silencing or gene activation (Moore *et al.*, 2012). DNA methylation plays a vital role in the control of gene expression, development, genomic imprinting, carcinogenesis, X chromosome inactivation and genomic stability. However the erroneous DNA methylation often contributes to the cancer development and multifactorial disease (Zhang, 2009). DNA methylation occurs at N_6 position of Adenine, N_4 and C_5 position of Cytosine, and out of all only it is found at C_5 position of Cytosine in eukaryotes. In eukaryotic DNA, a methyl group from S-adenyl methionine (SAM) is transferred to the fifth carbon of Cytosine base forming the 5mC, and this reaction is catalyzed by DNA methyltransferases enzyme(DNMTs) (Herman *et al.*,2004).

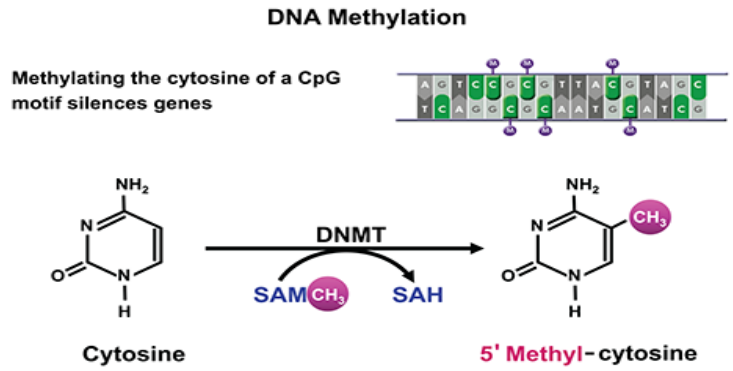


Figure 1: DNA methylation, where the methyl group of SAM is transferred to 5-methyl cytosine position by the Dnmts

http://pubs.niaaa.nih.gov

In eukaryotes, DNA methylation is catalyzed by the family of DNMTs: Dnmt1, Dnmt2, Dnmt3A, Dnmt3B and Dnmt3L. Dnmt1 is responsible for copying the methylation pattern to the daughter cells during the post replication process, exhibiting high preference for a hemimethylated DNA substrate. It acts mainly as "maintenance" methyltransferases during synthesis of DNA. Dnmt3a and Dnmt3b acts mainly as *de novo* enzyme with specific roles. Dnmt3a and Dnmt3b methylate the hemimethylated as well as unmethylated DNA, correlating their high expression in embryonic stem cells, early embryos and developing germ cell (Jeltsch *et al.*, 2005).

DNA methylation in DNA is catalyzed by DNA methyltransferases and methyl group is transferred from S-adenosyl-L-methionine (AdoMet) to the DNA bases. There is a nucleophilic attack on the C_6 of the target Cytosine. The attack is by the thiol group of the Cysteine residue. The covalent bond formation activates the C_5 to the electrophilic attack and further leading to the addition of the methyl group to the C of the Cytosine followed by the removal of the C_5 position proton and resolution of covalent intermediate (Jeltsch, 2004).

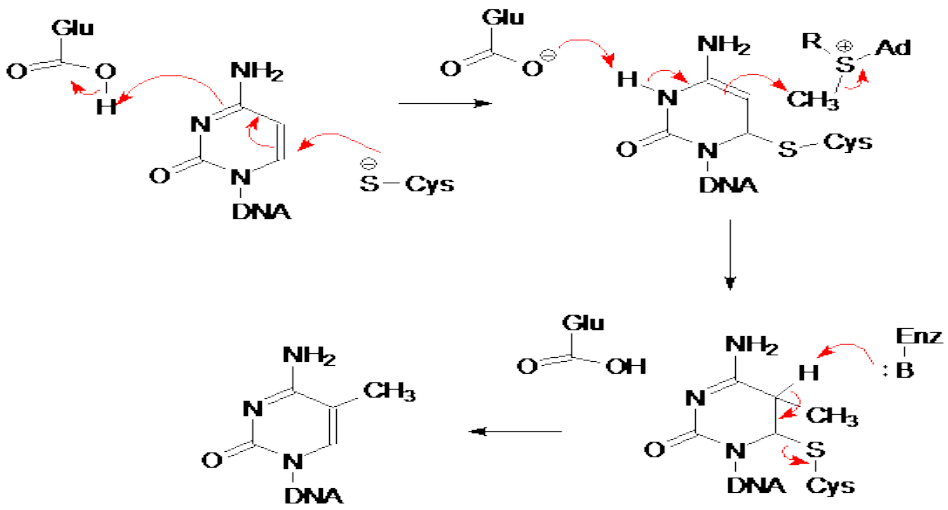


Figure 2: Mechanism of Methylation

http://chemistry.umeche.maine.

In case of prokaryotes methylation is usually at GATC or certain short sequences in context of restriction modification system. However in case of mammalian genomes they are methylated to certain specific CG sites, in methylation patterns. Largely methylation occurs at CG dinucleotide sequences. Not all CGs are methylated in the genome and this uneven distribution of methylated CGs generates patterns carrying epigenetic information. Such patterns of methylated CGs are significantly conserved among individual of same species

within corresponding cell types and different in case of different cell types of the same organism. DNA at each CG site carries specific methylation state. Methylation state is inherited and is maintained after cell division. Pattern of methylated and unmethylated CG sites are transformed to pattern of hemimethylated and methylated sites by the process of DNA replication. Methylation pattern is read and then copied to daughter strand by the maintenance methyltransferases which have a high specificity for methylation at hemimethylated sites (Jeltsch, 2004)

DNA methylation is highly regulated in mammals. There is absence of unregulated *de novo* methylation activity which could further harm the methylation pattern during the cellular and developmental cycles. This is under the control of methyltransferases. True *de novo* methylation expressed at very low levels in the differentiated tissues activity. Due to conformational changes the catalytic activity of Dnmt3a and Dnmt3b gets activated. The interaction of stimulatory proteins and the methyltransferases increases methyltransferases activity to target region (Handa, 2005).

In mammals, the function of cytosine methylation is found to be complex and difficult. While the role of Dnmt1 and Dnmt2 were demonstrated through the knockout models suggesting their role in mouse development and in the general epigenetic phenomenon such as genomic imprinting, X chromosome inactivation and transposon control. However in epigenetic gene regulation the function of methyl cytosine is highly specific. Through the genomic mapping approach the function of various individual Dnmt genes has been studied. For example, Dnmt3a was abundant in postnatal neuron stem cells and used for the postnatal neurogenesis. However other Dnmts interacted with other transcribed gene bodies preserving that gene regulation promotes methylation. It was shown that Dnmt3b in mouse ESCs was depended on the presence of Histone H3 lysine in the same region.

Methylation can silence gene by modifying DNA as well as by modifying histones. There are specific sites in the DNA and histones where methylation can occur. Methylation inhibits transcription by one of the two ways. It can either inhibit transcription directly by blocking the binding site of activator protein or it recruits the MBD (Methyl-CpG binding protein), thus no further activator protein can bind. This MBD protein recruits another type of protein called HDAC (Histone deacetylase). It causes deactylation of histones and thus the DNA start coiling on to histones tightly, leading to the formation of heterochromatin, which leads to gene inactivation (Tatematsu *et al*., 2000).

When DNA methylation occurs at 5 position of Cytosine it gets converted to 5 mC which may further undergoes deamination reaction resulting in the formation of thiamine. The Thiamine Guanine mismatch resulting from this reaction cannot be repaired by any of the DNA repair pathway. So 5mC site acts as hotspot for the mutation. On the other hand when deamination occurs in normal cytosine, it changes the cytosine to Uracil. As Uracil is not the nucleotide already present into DNA, and this is further notices by the DNA repair proteins is repaired by one of the DNA repair enzyme named as Uracil deglycosylase. This phenomenon has contributed to mammalian genome DNA evolution and resulted in representation of CG dinucleotide in the genomes.

There are certain regions in higher eukaryotes that contain high content of CpG , these dinucleotide clusters of CpG are called as CpG islands present in the promoter and exonic regions mammalian genes. These regions are rich in GC context. CpG islands are usually found to be unmethylated. CpG is referred as dinucleotide in which cytosine is joined to guanine and p refers to the phosphodiester bond between them. The proposed definition of CpG Island was given by Gardiner-Garden and Frommer in 1987 described CpG islands as a long stretch of DNA of 200 bp having CG content of about 50% and observed CpG/expected CpG in excess of 0.6. Regions of DNA greater than 500 bp having G+C content greater than or equal to 55% and having observed CpG/expected CpG of 0.65 were more likely to be associated with 5' regions of genes and it excluded the most Alu-repetitive elements. They are useful markers for genes in the organism having 5mC in their genomes. It often plays important roles in the gene silencing during process such as X chromosome inactivation, imprinting, and silencing of intragenomic parasites (Takai, 2002).

Due to the lack of DNA methylation and absence of CpG deficiency, CGIs are distinct in vertebrates. In some invertebrates the CpG occurs at the expected frequency due to little or no methylation. CGIs are not detected in these genomes. Whereas, many plant genomes are highly methylated and directly linked to genes. Most invertebrate have methylated genomes comprising alternating methylated and non methylated domains.

CGIs can be located at annotated TSS i.e. within gene bodies (Intragenic) or between annotated genes (Intergenic). These intragenic and intergenic CGIs of unknown function are classes as orphan also characterized transcriptional unit. Orphans are also the site of transcription and often represent novel promoters. These orphans are also the site for

transcription and a novel promoter. It often lacks the TATA box. Due to the high GC content they function as promoter.

Feltus *et al* in 2003, attempted to find out why certain genes are more methylated than others. They over expressed DNA cytosine-5-methyltransferase 1(Dnmt 1) and studied the behaviour of CpG Islands towards methylation. The overall phenomenon of CpG Island methylation was increased but not uniformly. Majority of CpG Islands remain unmethylated and a small fraction was affected by it. The methylation prone and methylation resistant CpG Islands so obtained were similar with respect to size, G+C content, CpG frequency and chromosomal location but are found to be different based on its sequence context.

There are various reports where the methylation of DNA has been associated with the sequence attributes of DNA. Studies carried by Handa and Jeltsch showed that there is crystal clear relation between the tendency of a CpG site to undergo methylation and its flanking sequence. They further proposed that there are distinct statistically significantly consensus sequences flanking CpG sites which induces various levels of methylation. Intrinsic sequence preference of *de novo* MTases could be one of the potential parameters that influences the generation of the DNA methylation patterns of mammalian genomes, a process that is not well understood (Handa and Jeltsch *et al.,* 2005).

Not all the CpG islands are found unmethylated always. There are reports of highly methylated CGIs .It may be inferred that CGIs can be classified in two categories, namely methylation resistant and methylation sensitive. Feltus *et al.,* used MEME (Motif-based sequence analysis tool) and MAST(Motif Alignment and Search Tool) algorithm and derived 5 motifs from methylation prone sequence and 8 motifs from methylation resistant sequences. These motifs successfully discriminated between methylation prone and methylation resistant CpG islands with 87% accuracy (Feltus *et al.,* 2006).

The present study aimed at investigating if sequence complexity differs between the methylation sensitive and resistant CGIs. The sequence complexity has been measured as percentage compression of the CGIs sequence or degree of randomness of the bases in the DNA sequence of CGIs. Compression percentage is based on the size of compressed data and the original data. Compression is handling huge data's. Here we have used the statistical tool for analyzing the compressed data. Similarly randomness is explained using the runs test. It was done to check out the regularity.

CHAPTER-2

REVIEW

OF

LITERATURE

REVIEW OF LITERATURE

All cells contain the same genetic information, not all of them are expressed by the cell types. This diversified gene expression is mediated by the Epigenetics. Epigenetics is the heritable changes that are not associated with the change in DNA sequences. In the mammalian genome the most frequently occurring epigenetic event is DNA methylation. It occurs at N_6 position of Adenine and N_4 and C_5 positions of cytosine but in mammalian genome it occurs only at C_5 position of Cytosine residues. This reaction is catalyzed by enzyme named as (Dnmt)(Moore *et al*., 2012).

Waddington coined the term Epigenetics which linked the developmental biology and genetics. It described all those mechanisms which were required for the unfolding of the genetic programme for the development process. Later on it was suggested that 5mC was primarily responsible in controlling the gene expression. It was inferred that gene silencing also associated with the methylation phenomenon. However, there are various other important epigenetic mechanisms involving chromatin and histone modification, and in the regulation of RNAs (Holliday *et al.,* 2006).

Holliday in his work explained that DNA methylation changes the gene activity. He found out that both DNA methylation and gene expression are related to each other. Majority of DNA methylation occurs at CpG sites. In mammalian genomes those CpG sites are depleted that may cause the mutagenic potential of 5mC, which is formed by the addition of methyl group in Cytosine residue at 5 position that deaminate to form Thymine. The rest CpG sites are spread across the genome where they get heavily methylated with the exception of CpG islands .CpG islands mainly located at promoter region of genes that regulates various processes i.e. gene silencing, X chromosome inactivation, genomic imprinting. The first computational analysis of CpG islands was done by Gardiner-Garden and Frommer using the vertebrate sequences from the GenBank. They defined CpG island as DNA region of 200 bp having GC content more than 50% and having CpG ratio(observed/expected) greater than 0.6 or equal to 0.6.This definition of CpG island was further improved by Takai and Jones in 2001,which was based on the study of chromosome 21 and 22, in which they redefined CpG island as sequence of more than or equal to 500 bp with GC content more or equal to 55% and observed to expected CpG ratio equal to 0.65 or more than 0.65 (Takai and Jones *et al*., 2002).

During evolution there is a conversion of mCpG to TpG and CpA. Change in one 5mC leads to the loss of 2 CpG and gain of one CpA and one TpG. In 1980 Bird had found that DNA methylation is related to the deficiency of CpG dinucleotide. In many mammals the presence of CpGs are less as expected to the observed base composition. In this study restriction enzyme HapII and Msp1 was used to analyze the degree of methylation. The collected sample of vertebrate genome was run for the gel electrophoresis, molecular weight of the digests and the CCGG sites uncut by Hpa II were analyzed. After the digestion with the HpaII it was found that the DNA bands were of the same molecular weight as the Hpa II. It was observed that the organism with the most extreme CpG deficiency had the highest levels of DNA methylation *viz* mutation of 5mC site is relatively more frequent compared to the other dinucleotides and those genomes which are poorly methylated can have the deficiency of CpG.(Bird *et al.*,1980)

DNA sequence adjusts CGIs for promoter function which destabilizes nucleosome and attracts proteins which create a transcriptional permissive chromatin state. Using unique DNA sequence composition, silencing of CGI promoters is done through CpG methylation. Therefore CGIs influence local chromatin structure and gene activity regulation. CpG acts as a substrate for DNA methyltransferases which promotes the regulation of chromatin structure (Deaton & Bird *et al.,* 2011).

Handa and Jeltsch in 2005 suggested that in preferences of the flanking sequences of *denovo* DNMTs plays a significant role in the process of DNA methylation and also in the cause of CG islands. They determined the effect of flanking sequence on the DNA methylation phenomenon. The catalytic activity of Dnmt3a and Dnmt3b was studied using the synthetic oligonucleotide substrate that covered all the possible flanks. Through the methylation kinetics, it was observed that there was a 13 fold difference between the preferred and disfavoured flanking sequences. It was further statistically analyzed and showed that the AT rich flanks preferred over the GC rich flanks.

Shelenkov and Korotkov in 2008 used the novel method of Runs Test to determine the regularity of DNA sequences. In this test they determined the Adenine distribution (non randomness) with the length of 3 nucleotide. The sequences were further divided into separate periods by using the symbol F after each nucleotide. Then in the sequence the F symbol was further changed to 0 and a to 1, after which the number of runs were calculated. Run is basically a sequence of identical elements that is preceded and followed by different

elements or no elements. In the periodic sequence the number of runs should be greater or equal to non periodic sequence. Now using the number of runs as a periodicity measure tells us about the insensitivity to nucleotide insertion and deletions. By this study they found that in eukaryotic promoters more than 60% sequences shows the regularity property that was based on statistical significant level.

Bock and Paulsen in 2006 suggested an epigenetic tool that was designed to discriminate the CpG Islands that were prone to methylation and the unmethylated. A scoring prediction tool was designed which covered various aspects which referred to sequence, repeats, predicted structure, CpG islands, genes, predicted binding sites, conservation, and single nucleotide polymorphisms. The analysis was carried out on the 132 sequences of the chromosome 21. The result showed that the three attributes *i.e.*, specific DNA repeats, particular DNA structure and certain sequence patterns were highly correlated with the CpG Island methylation.

Vass and Wilson (1984) had performed the runs test for the scanning of DNA sequences which ere potential for forming the Z-DNA regions. They have developed a FORTRAN programme for this purpose. It also detected the non-random arrangements of the purines and pyrimidine within same strand. Graphical output was obtained through the statistical test and was further used to search for the ß-type globin DNA sequences which have the Z-DNA regions.

Compression of DNA sequence is done in order to retrieve the information. Better compression provides better understanding .Several algorithms have been designed for the compression like Biocompress-2, GenCompress and CTW+LZ (Chen *et al.,* 2002). Grumabch and Tahi had proposed a lossless algorithm in their work in order to compress the information stored in the DNA sequence. Their method was based on the presence of palindromes. Two approaches are used for the compression purpose: one is statistical and the other one is substitutional.

DELIMINATE, a novel compression algorithm was designed by Mohammad. The algorithm compressed the data into the loss- less manner. It helped in efficient compression, storage, retrieval and the transmission of data. The compression was performed in 2 phases. In the first phase, all non ATCG characters and regions with low complexity were recorded. In the second phase the sequences with highest frequencies of occurrence were recorded and the bases with lowest frequencies were represented in the binary code. They have used four

different datasets *i.e*. FNA, FFN, Eukaryotic and Next generation sequencing for the evaluation of compression efficiency. Compression ratio was calculated. The result was compared with other compression algorithm such as gzip, bzip2 and lzma. The results showed that the data obtained from the DELIMINATE algorithm was 7-27% better when compared with the other algorithm (Mohammed *et al.*, 2012).

CHAPTER-3

SCOPE

OF

STUDY

SCOPE OF STUDY

DNA sequences are the carriers of genetic information. Although coding sequences are the most vital sequences, many other types of sequences also play very important role. One of such sequences is CpG islands. Functional significance is that CpG islands usually do not get methylated. However there have been several reports where CpG islands have been found to be methylated. Many scientists have attempted to explain why some CpG islands get methylated while others do not. The present study attempts to explore some new parameters to answer this question. These parameters are related to information theory and thus it can be linked with biological phenomena.

CHAPTER-4

OBJECTIVES

OBJECTIVES

(i) To design a Run Length Encoding (RLE) algorithm for the DNA sequence compression.

(ii) To check the randomness of the sequences of sensitive and resistant CpG islands using runs test.

(iii) To compare the DNA sequence compressibility of the sensitive and resistant CpG islands.

CHAPTER-5

MATERIALS

AND

METHODS

MATERIALS AND METHODS

Data Source: The data were collected from the published work of Yamada *et al.* It comprised of DNA sequences of two different sets of CpG Island: Fully methylated (M) and unmethylated(U).It was further classified into four categories that is fully methylated, unmethylated, incomplete and compositely methylated(Yamada *et al*., 2004). Here we had analyzed the fully methylated sequences and the methylated sequences. Below is the given list of sequences along with their chromosome number, length and methylation status:

Table 2: Nucelotide sequences of chromosome 21 and their length and their methylation status

S.No	Sequences and regions	Chromosome	Length	Methylation status
1	(NT_002836.4 740746-742525)	21	1780	fully methylated
2	(NT_002836.4 798428-799837)	21	1410	fully methylated
3	(NT_002836.4 1099894-1101335)	21	1442	fully methylated
4	(NT_002836.4 2100365-2102278)	21	1914	unmethylated
5	(NT_002836.4 2765740-2767861)	21	2122	unmethylated
6	(NT_002836.4 4548804-4550994)	21	2191	unmethylated
7	(NT_002836.4 4648389-4650557)	21	2169	unmethylated
8	(NT_002836.4 4854926-4857042)	21	2117	unmethylated
9	(NT_002836.4 12511562-12513060)	21	1499	unmethylated
10	(NT_002836.4 12684911-12686522)	21	1612	unmethylated
11	(NT_002836.4 13119202-13121651)	21	2450	unmethylated
12	(NT_002836.4 13793868-13795756)	21	1889	unmethylated
13	(NT_002836.4 13915193-13916938)	21	1746	unmethylated
14	(NT_002836.4 13917154-13918663)	21	1510	unmethylated
15	(NT_002836.4 15834745-15836237)	21	1493	unmethylated
16	(NT-002836.4 16246692-16248676)	21	1985	unmethylated
17	(NT-002836.4 18607855-18609856)	21	2002	unmethylated
18	(NT_002836.4 18679791-18682057)	21	2267	unmethylated
19	(NT_002836.4 18821198-18823391)	21	2194	unmethylated
20	(NT_002836.4 19227134-19228707)	21	1574	unmethylated
21	(NT_002836.4 19248734-19250388)	21	1655	fully methylated
22	(NT_002836.4 19359982-19362161)	21	2180	unmethylated
23	(NT_002836.4 19561116-19562518)	21	1403	unmethylated
24	(NT_002836.4 19675812-19677620)	21	1809	unmethylated

S.No	Sequences and regions	Chromosome	Length	Methylation status
25	(NT_002836.4 19719511-19721351)	21	1841	unmethylated
26	(NT_002836.4 19972011-19973417)	21	1407	unmethylated
27	(NT_002836.4 19975198-19977216)	21	2019	unmethylated
28	(NT_002836.4 20018303-20020533)	21	2231	unmethylated
29	(NT_002836.4 20178164-20179883)	21	1720	unmethylated
30	(NT_002836.4 20273122-20274794)	21	1673	unmethylated
31	(NT_002836.4 20351550-20353202)	21	1653	unmethylated
32	(NT_002836.4 20427755-20429718)	21	1964	unmethylated
33	(NT_002836.4 20536702-20538471)	21	1770	unmethylated
34	(NT_002836.4 20590829-20592722)	21	1894	unmethylated
35	(NT_002836.4 21021661-21023784)	21	2124	unmethylated
36	(NT_002836.4 21323771-21325344)	21	1574	unmethylated
37	(NT_002836.4 21562903-21564940)	21	2038	unmethylated
38	(NT_002836.4 21617860-21620045)	21	2186	unmethylated
39	(NT_002836.4 21740221-21742136)	21	1916	fully methylated
40	(NT_002836.4 21835266-21836717)	21	1452	unmethylated
41	(NT_002836.4 22835347-22836774)	21	1428	fully methylated
42	(NT_002836.4 23008324-23010418)	21	2095	unmethylated
43	(NT_002836.4 23018419-23020042)	21	1624	unmethylated
44	(NT_002836.4 23083543-23085110)	21	1568	unmethylated
45	(NT_002836.4 23104500-23106945)	21	2446	unmethylated
46	(NT_002836.4 23268373-23270063)	21	1691	unmethylated
47	(NT_002836.4 23333456-23335039)	21	1584	unmethylated
48	(NT_002836.4 23646833-23649367)	21	2535	unmethylated
49	(NT_002836.4 23648887-23650365)	21	1479	unmethylated
50	(NT_002836.4 23657131-23658548)	21	1418	unmethylated
51	(NT_002836.4 23695691-23697656)	21	1966	unmethylated
52	(NT_002836.4 23914335-23916288)	21	1954	unmethylated
53	(NT_002836.4 23938177-23939758)	21	1582	unmethylated
54	(NT_002836.4 24021123-24022718)	21	1596	unmethylated
55	(NT_002836.4 24215237-24217598)	21	2362	unmethylated
56	(NT_002836.4 24314183-24317285)	21	3103	unmethylated
57	(NT_002836.4 24511883-24513479)	21	1597	unmethylated
58	(NT_002836.4 25608278-25609973)	21	1696	unmethylated
59	(NT_002836.4 25753617-25755681)	21	2065	unmethylated
60	(NT_002836.4 26130818-26133115)	21	2298	unmethylated
61	(NT_002836.4 26259167-26262261)	21	3095	unmethylated
62	(NT_002836.4 26295243-26297124)	21	1882	unmethylated

S.No	Sequences and regions	Chromosome	Length	Methylation status
63	(NT_002836.4 26390768-26392267)	21	1500	unmethylated
64	(NT_002836.4 26556991-26558660)	21	1670	unmethylated
65	(NT_002836.4 27791317-27792866)	21	1550	unmethylated
66	(NT_002836.4 28112154-28114765)	21	2612	unmethylated
67	(NT_002836.4 28452242-28454382)	21	2141	unmethylated
68	(NT_003545.2 122339-124366)	21	2028	unmethylated
69	(NT_003545.2 178197-181181)	21	2922	unmethylated
70	(NT_003545.2 387867-389371)	21	1505	unmethylated
71	(NT_003545.2 403601-405007)	21	1407	unmethylated
72	(NT_003545.2 665089-666509)	21	1421	unmethylated
73	(NT_003545.2 682430-684731)	21	2302	unmethylated
74	(NT_003545.2 756256-760387)	21	4132	fully methylated
75	(NT_003545.2 822395-823837)	21	1443	unmethylated
76	(NT_003545.2 854975-856401)	21	1427	fully methylated
77	(NT_003545.2 1017136-1019217)	21	2082	fully methylated
78	(NT_003545.2 1047902-1049365)	21	1464	unmethylated
79	(NT_003545.2 1142956-1145447)	21	2492	unmethylated
80	(NT_003545.2 1275925-1277810)	21	1886	unmethylated
81	(NT_002835.3 158046-161231)	21	3186	unmethylated
82	(NT_002835.3 297286-298768)	21	1483	fully methylated
83	(NT_002835.3 389981-391414)	21	1434	fully methylated
84	(NT_002835.3 391442-393133)	21	1692	unmethylated
85	(NT_002835.3 473197-474680)	21	1484	fully methylated
86	(NT_002835.3 508085-509942)	21	1858	unmethylated
87	(NT_002835.3 521399-523051)	21	1653	unmethylated
88	(NT_002835.3 596448-599173)	21	2726	unmethylated
89	(NT_002835.3 743918-746243)	21	2326	unmethylated
90	(NT_002835.3 839100-840858)	21	1759	unmethylated
91	(NT_002835.3 865543-867041)	21	1499	unmethylated
92	(NT_002835.3 972548-974376)	21	1829	unmethylated
93	(NT_002835.3 974194-975798)	21	1605	unmethylated
94	(NT_002835.3 1031958-1033740)	21	1783	unmethylated
95	(NT_002835.3 1070778-1072793)	21	2016	unmethylated
96	(NT_002835.3 1082093-1084166)	21	2074	unmethylated
97	(NT_002835.3 1101487-1103181)	21	1695	fully methylated
98	(NT_002835.3 1187382-1188869)	21	1488	unmethylated
99	(NT_002835.3 1441617-1443042)	21	1426	unmethylated
100	(NT_002835.3 1533742-1535201)	21	1460	unmethylated

S.No	Sequences and regions	Chromosome	Length	Methylation status
101	(NT_002835.3 1549573-1551858)	21	2286	unmethylated
102	(NT_002835.3 1604933-1607527)	21	2595	unmethylated
103	(NT_002835.3 1664660-1666157)	21	1498	unmethylated
104	(NT_002835.3 1671866-1673671)	21	1806	unmethylated
105	(NT_002835.3 1680257-1681681)	21	1425	fully methylated
106	(NT_002835.3 1710450-1712127)	21	1678	fully methylated
107	(NT_002835.3 1750441-1752578)	21	2138	unmethylated
108	(NT_002835.3 1806394-1808604)	21	2211	unmethylated
109	(NT_002835.3 1865795-1867783)	21	1989	fully methylated
110	(NT_002835.3 2020037-2021768)	21	1732	unmethylated
111	(NT_002835.3 2077997-2079517)	21	1521	fully methylated
112	(NT_002835.3 2112166-2114090)	21	1925	fully methylated
113	(NT_002835.3 2136861-2139459)	21	2599	unmethylated
114	(NT_002835.3 2286418-2288635)	21	2218	fully methylated
115	(NT_002835.3 2364595-2366958)	21	2364	fully methylated
116	(NT_002835.3 2371534-2374424)	21	2891	unmethylated
117	(NT_002835.3 2475023-2477222)	21	2200	fully methylated
118	(NT_002835.3 2653007-2655508)	21	2502	fully methylated
119	(NT_002835.3 2716963-2718791)	21	1829	fully methylated
120	(NT_002835.3 2732562-2734239)	21	1538	fully methylated
121	(NT_002835.3 2735023-2736682)	21	1660	fully methylated
122	(NT_002835.3 2827157-2828726)	21	1570	unmethylated
123	(NT_002835.3 2841396-2842933)	21	1538	fully methylated
124	(NT_002835.3 2855101-2856552)	21	1772	fully methylated
125	(NT_002835.3 2861362-2862966)	21	1605	fully methylated
126	(NT_002835.3 2957438-2959877)	21	2440	unmethylated
127	(NT_002835.3 014936-3016783)	21	1848	unmethylated
128	(NT_002835.3 3052859-3055543)	21	2685	unmethylated
129	(NT_002835.3 3132299-3134479)	21	2181	fully methylated
130	(NT_002835.3 3187879-3189962)	21	2084	Unmethylated
131	(NT_002835.3 3364581-3366420)	21	1840	Unmethylated
132	(NT_002835.3 3396443-3398312)	21	1870	Unmethylated

Total of 132 sequences were selected, among which 29 fully methylated and 103 unmethylated. The fully methylated were the methylation sensitive and the unmethylated were methylation resistant.

Sequence Analysis Tools

Microsoft Excel

Microsoft excel tool was used for the statistical and computational analysis of the data. Following tools were used:

Table 3: Microsoft Excel Tools

S.No.	Tools	Class	Functions
1.	LEN	Text	Find the number of characters in the region
2.	MIN	Statistical	Find the smallest number in data
3.	AVERAGE	Statistical	Find the arithmetic mean of the numerical data
4.	SUM	Math & Trig	Find the sum of the selected region
5.	STDEV	Statistical	Find the deviation from the mean

Notepad++

It was used for sequence manipulation and analysis

Table 4: Notepad++ Tools

S.No.	Macros used	Function
1.	Macro recording	Used for recording a action
2.	Run a macro multiple time	To replay the recorded macro multiple times.

Procedure

The nucleotide sequences were obtained from the NCBI (www.ncbi.nlm.nih.gov/) with the given accession number in the FASTA format.

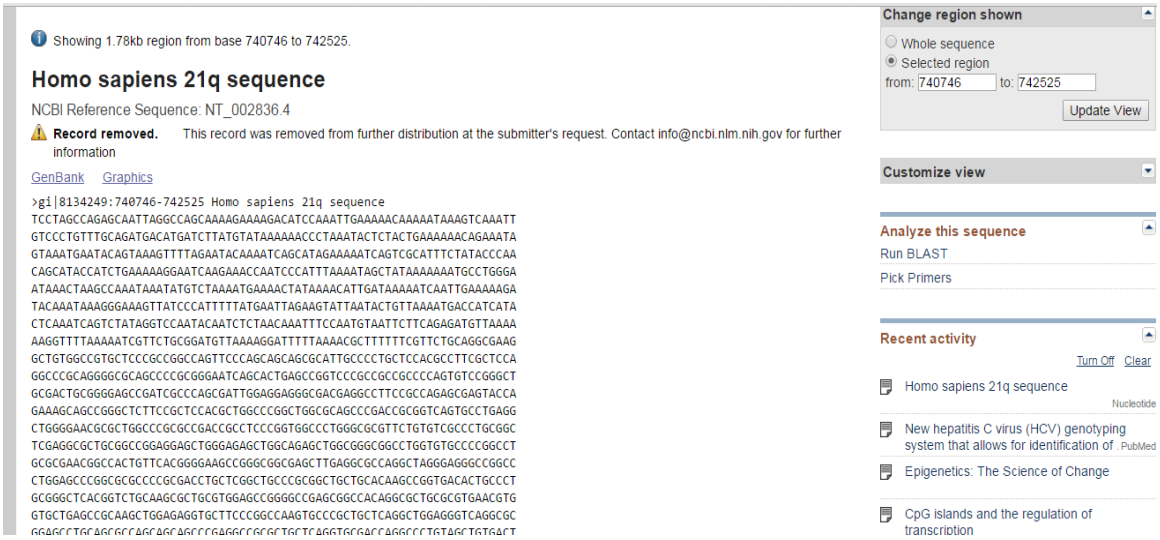


Figure 3: Sequences downloaded from NCBI

Likewise all the sequences were obtained and two separate files were made. One included all the sensitive sequences and the other with all resistant sequences. Following are the steps:

1. All the sensitive sequences were made into a single string and each sequence was separated from each other by J. Similarly all the resistant sequences were made into a single string and were separated from each other by X.
2. Furthermore conversions were done using Notepad++ where all the nucleotides (A, T, C and G) were replaced with certain alphabet and separate files were made accordingly.
3. 7 files of resistant and 7 files of sensitive sequences were made. Following conversions were done:

Table 5: Replacement of Nucleotides

S.No.	File Name	Ist nucleotide conversion	IInd nucleotide conversion
1.	_RY	G and A replaced with R	C and T replaced with Y
2.	_SW	G and C replaced with S	A and T replaced with W
3.	_MK	A and C replaced with M	G and T replaced with K

4.	_AB	G , T and C replaced with B
5.	_CD	G , A and T replaced with D
6.	_GH	A , T and C replaced with H
7.	_TV	G , A and C replaced with V

4. After the replacement of nucleotides with RY, we have used algorithm to split the R and Y fragments individually. Two macros were recorded for the fragmentation first of all macro with name RY was designed in order to split between R and Y.

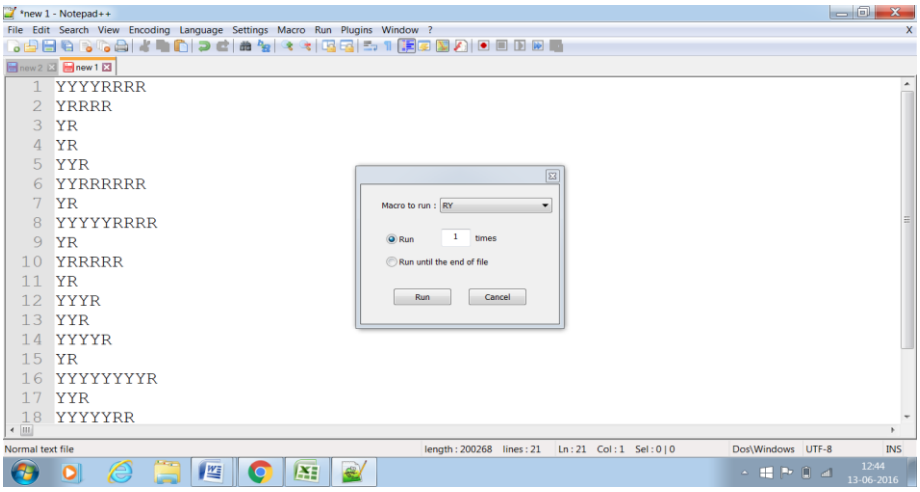


Figure 4: Macro recording for the split of RY using Notepad++

5. This algorithm was run until the end of sequence splitting R and Y.
6. After this another macro was designed for the same file, named YR to split between Y and R and was run until the end.

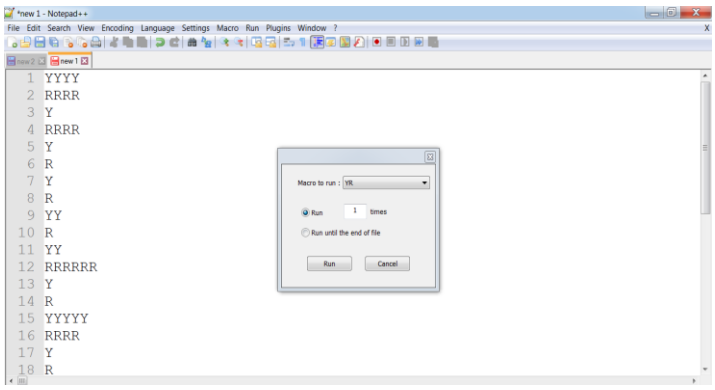


Figure 5: Macro recording for the split of YR

7. This algorithm was run until the end of sequence splitting R and Y.

After this another macro was designed for the same file, named YR to split between Y and R and was run until the end

Randomness

In order to check the randomness we carried out the runs test. Runs test a statistical sample test that is used to find the number of the two alternatives. It is a parametric test. First of all we have calculated the *runs.* Runs can be defined as a sequence of like elements bounded on either side by either unlike elements or no elements. Runs test was firstly run in the sensitive sequences.

As shown below, in column B, we have used the function =LEN (A) to find the corresponding length of the elements in the row. In the column C, 1 is placed if Ys present in column A and 0 for the Rs. Similarly in column D, 1 was placed for the Rs and 0 for the Ys. Column E shows the product of column B and C. In column F we have calculated the product of B and D. In column J we have calculate the Value of Ys using the function =SUM (E1: EN). Similarly the values of Rs were calculated in column K using function =SUM (F1: FN). In column L, we have calculated the sum of column J and column K to find out N that is total number of characters.

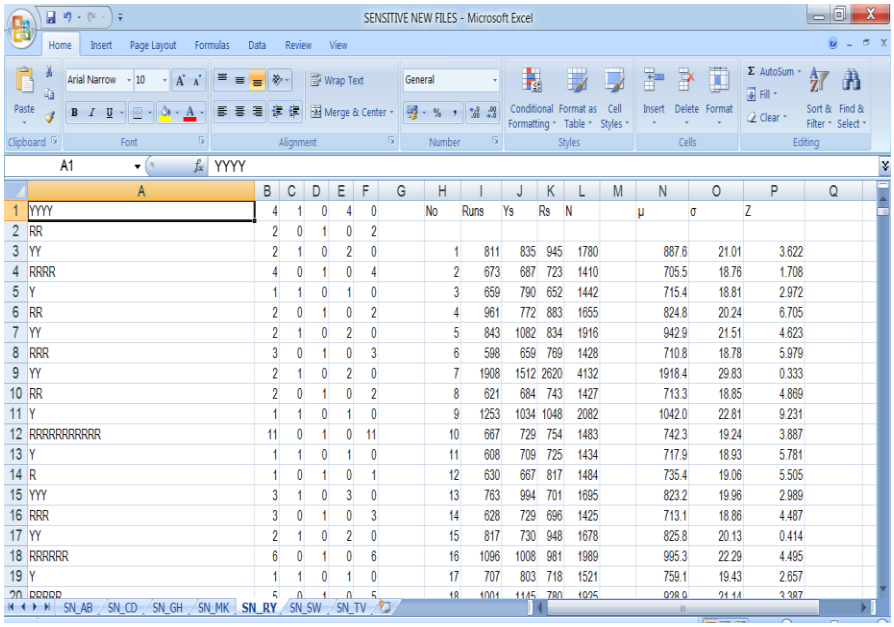
We have calculated μ (mean), σ (standard deviation) and Z (critical value) by using the following formulae:

$$\boldsymbol{\mu = \frac{2n_1n_2}{N} + 1}$$

$$\boldsymbol{\sigma = \sqrt{\frac{2n_1n_2(2n_1n_2 - N)}{N^2(N-1)}}}$$

$$\boldsymbol{Z = Z = \frac{|u-\mu| - 0.5}{\sigma}}$$

Here b is μ, n_1 and n_2 are Ys and Rs respectively. N is the sum of Ys and Rs.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	YYYY	4	1	0	4	0		No	Runs	Ys	Rs	N		μ	σ	Z	
2	RR	2	0	1	0	2											
3	YY	2	1	0	2	0		1	811	835	945	1780		887.6	21.01	3.622	
4	RRRR	4	0	1	0	4		2	673	687	723	1410		705.5	18.76	1.708	
5	Y	1	1	0	1	0		3	659	790	652	1442		715.4	18.81	2.972	
6	RR	2	0	1	0	2		4	961	772	883	1655		824.8	20.24	6.705	
7	YY	2	1	0	2	0		5	843	1082	834	1916		942.9	21.51	4.623	
8	RRR	3	0	1	0	3		6	598	659	769	1428		710.8	18.78	5.979	
9	YY	2	1	0	2	0		7	1908	1512	2620	4132		1918.4	29.83	0.333	
10	RR	2	0	1	0	2		8	621	684	743	1427		713.3	18.85	4.869	
11	Y	1	1	0	1	0		9	1253	1034	1048	2082		1042.0	22.81	9.231	
12	RRRRRRRRRR	11	0	1	0	11		10	667	729	754	1483		742.3	19.24	3.887	
13	Y	1	1	0	1	0		11	608	709	725	1434		717.9	18.93	5.781	
14	R	1	0	1	0	1		12	630	667	817	1484		735.4	19.06	5.505	
15	YYY	3	1	0	3	0		13	763	994	701	1695		823.2	19.96	2.989	
16	RRR	3	0	1	0	3		14	628	729	696	1425		713.1	18.86	4.487	
17	YY	2	1	0	2	0		15	817	730	948	1678		825.8	20.13	0.414	
18	RRRRRR	6	0	1	0	6		16	1096	1008	981	1989		995.3	22.29	4.495	
19	Y	1	1	0	1	0		17	707	803	718	1521		759.1	19.43	2.657	
20	RRRRR	5	0	1	0	5		18	1001	1145	780	1925		928.9	21.14	3.387	

Figure 6: Calculation of μ, σ and z

Similarly we have calculated the runs test for all the sensitive and resistant files.

Compressibility

We had used the run length encoding compression algorithm to compress the sequence data. It is performed to look for the similarity across the whole sequence. First of all, macros were recorded in the Notepad++ to make a single stretch of the whole sequences. Here we had used the binary conversion. All the motif repeats to be compressed are as follows:

Table 6: Compression of motif repeats

S.No.	Motif repeats	Compressed to
1.	1	A
	0	
2.	10	B
3.	100	C
	110	
4.	1000	D
	1100	
	1110	
5.	11000	E

	11100	
	11110	
	10000	
	11010	
	10110	
6	100001	F
	100010	
	100101	
	111010	
	111100	
	111110	
	100000	
	110100	
	111000	

Repeats were replaced in the following manner

Table 7: Replacement scheme for the motif repeats

S.No.	Motif repeats	Replacement
	A	
(i)	1111111111111111	[16J
(ii)	111111111111111	[15J
(iii)	11111111111111	[14J
(iv)	1111111111111	[13J
(v)	111111111111	[12J
(vi)	11111111111	[11J
(vii)	1111111111	[10J
(viii)	111111111	[9J
(ix)	11111111	[8J
(x)	1111111	[7J
(xi)	111111	[6J
(xii)	11111	[5J

(xiii)	1111	[4J
(xiv)	0000000000000000	[16O
(xv)	000000000000000	[15O
(xvi)	00000000000000	[14O
(xvii)	0000000000000	[13O
(xviii)	000000000000	[12O
(xix)	00000000000	[11O
(xx)	0000000000	[10O
(xxi)	000000000	[9O
(xxii)	00000000	[8O
(xxiii)	0000000	[7O
(xxiv)	000000	[6J
(xxv)	00000	[5J
(xxvi)	0000	[4J

S.No.	Motif repeats	Replacement
B		
(i)	1010101010101010	[8JO
(ii)	10101010101010	[7JO
(iii)	101010101010	[6JO
(iv)	1010101010	[5JO
(v)	10101010	[4JO
(vi)	101010	[3JO
C		
(i)	100100100100100100	[6JOO
(ii)	100100100100100	[5JOO
(iii)	100100100100	[4JOO
(iv)	100100100	[3JOO
(v)	100100	[2JOO
(i)	1101101101101101 10	[6JJO

(ii)	110110110110110	[5JJO
(iii)	110110110110	[4JJO
(iv)	110110110	[3JJO
(v)	110110	[2JJO
D		
(i)	1000100010001000	[4JOOO
(ii)	100010001000	[3JOOO
(iii)	10001000	[2JOOO
(i)	1100110011001100	[4JJOO

(ii)	110011001100	[3JJOO
(iii)	11001100	[2JJOO
(i)	1110111011101110	[4JJJO
(ii)	111011101110	[3JJJO
(iii)	11101110	[2JJJO
E		
(i)	11000110001100011000	[4JJOOO
(ii)	110001100011000	[3JJOOO
(iii)	1100011000	[2JJOOO
(i)	11100111001110011100	[4JJJOO
(ii)	111001110011100	[3JJJOO
(iii)	1110011100	[2JJJOO
(i)	11110111101111011110	[4JJJJO
(ii)	111101111011110	[3JJJJO
(iii)	1111011110	[2JJJJO
(i)	10000100001000010000	[4JOOOO

(ii)	100001000010000	[3JOOOO
(iii)	1000010000	[2JOOOO
(i)	11010110101101011010	[4JJOJO
(ii)	110101101011010	[3JJOJO
(iii)	1101011010	[2JJOJO
(i)	10110101101011010110	[4JOJJO
(ii)	101101011010110	[3JOJJO
(iii)	1011010110	[2JOJJO
	F	
(i)	100001100001100001	[3JOOOOJ
(ii)	100001100001	[2JOOOOJ
(i)	100010100010100010	[3JOOOJO
(ii)	100010100010	[2JOOOJO
(i)	100101100101100101	[3JOOJOJ
(ii)	100101100101	[2JOOJOJ
(i)	111010111010111010	[3JJJOJO
(ii)	111010111010	[2JJJOJO
(i)	111100111100111100	[3JJJJOO
(ii)	111100111100	[2JJJJOO
(i)	1111101111110111110	[3JJJJJO
(ii)	111110111110	[2JJJJJO
(i)	100000100000100000	[3JOOOOO
(ii)	100000100000	[2JOOOOO

(i)	110100110100110100	[3JJOJOO
(ii)	110100110100	[2JJOJOO
(i)	111000111000111000	[3JJJOOO
(ii)	111000111000	[2JJJOOO

Compression scheme done in the following ways

: Table 8: Compression scheme

S. No	**Steps**
1.	A
2.	A+B
3.	A+B+C
4.	A+B+C+D
5.	A+B+C+D+E
6.	A+B+C+D+E+F
7.	F+E
8.	F+E+D
9.	F+E+D+C
10.	F+E+D+C+B
11.	F+E+D+C+B+A

In order to check whether all the algorithm that we have used were correct, we have carried out the multiple sequence alignment of a bacteriophage lambda sequence. First of all, a bacteriophage lambda sequence of 48502 bp was acquired from NCBI in the FASTA format. Then the set of algorithms with compression and decompression were run. At last, both the compressed sequence and the original sequence were aligned using pair wise sequence alignment, showing 100% similarity and proved the accuracy of designed algorithm.

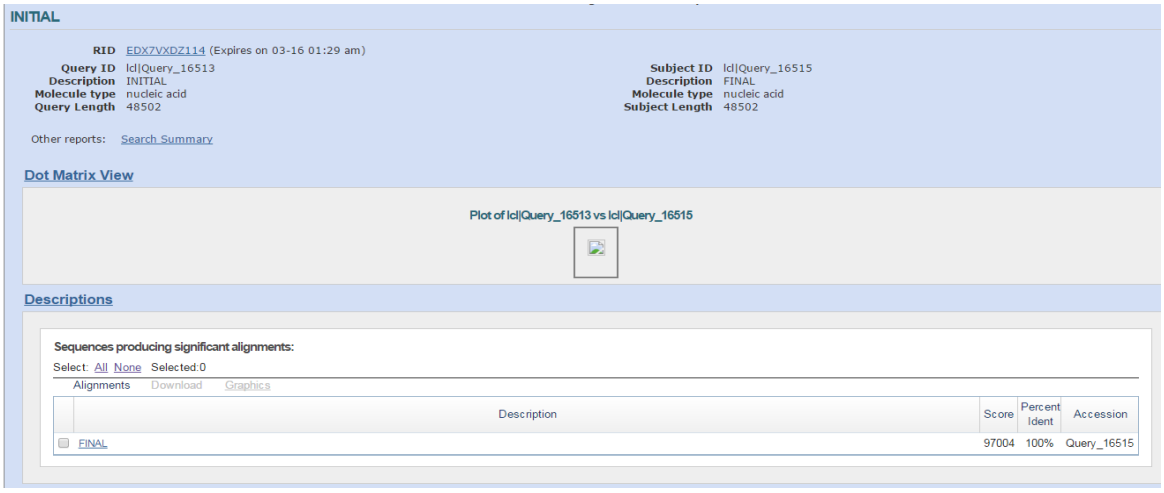


Figure 7: Pairwise alignment with NEEDLEMAN WUNSCH method

We have used the t-test to calculate the compressibility.

t-test is used for the statistical analysis. It compares the mean of two populations. It indicates whether or not the difference between two group's averages most likely reflects a real difference in the population from which the groups were sampled. Below is the excel sheet which shows how the data were stored and t-test was carried out.

Chart 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	
1	SENSITIVE	initial length	binary	A	B	C	D	E	F	A+B	A+B+C	A+B+C+D	A+B+C+D+E	A+B+C+D+E+F	F+E	F+E+D	F+E+D+C	F+E+D+C+B	F+E+B+A
2																			
3	Sequence 1	1780	3560	3025	3324	3486	3464	3518	3492	2861	2848	2848	2845	2837	3450	3368	3300	3106	
4	Sequence 2	1410	2820	2333	2742	2747	2736	2766	2780	2291	2275	2267	2267	2263	2726	2644	2582	2510	
5	Sequence 3	1442	2884	2265	2848	2807	2798	2830	2844	2247	2238	2224	2224	2224	2793	2709	2653	2619	
6	Sequence 4	1655	3310	2954	3212	3142	3140	3271	3286	2896	2814	2804	2804	2800	3247	3085	2962	2884	
7	Sequence 5	1916	3832	3090	3652	3749	3712	3725	3716	2964	2938	2916	2908	2908	3615	3505	3431	3267	
8	Sequence 6	1428	2856	2468	2712	2781	2806	2817	2804	2382	2357	2355	2355	2355	2768	2720	2648	2522	
9	Sequence 7	4132	8264	6623	7858	8021	8159	8231	8106	6361	6342	6292	6292	6284	8076	7976	7792	7398	
10	Sequence 8	1427	2854	2381	2702	2763	2782	2783	2822	2269	2248	2248	2245	2241	2751	2683	2602	2468	
11	Sequence 9	2082	4164	3729	4052	3889	3808	4104	4144	3663	3637	3567	3564	3560	4087	3735	3560	3478	
12	Sequence 10	1483	2966	2609	2764	2884	2908	2916	2928	2479	2456	2448	2448	2444	2878	2822	2753	2587	
13	sequence 11	1434	2868	2366	2776	2808	2786	2814	2844	2316	2305	2291	2285	2285	2793	2717	2667	2585	
14	sequence 12	1484	2968	2199	2956	2908	2906	2932	2948	2197	2185	2185	2185	2185	2912	2854	2804	2792	

Figure 8: Compression lengths for compression scheme

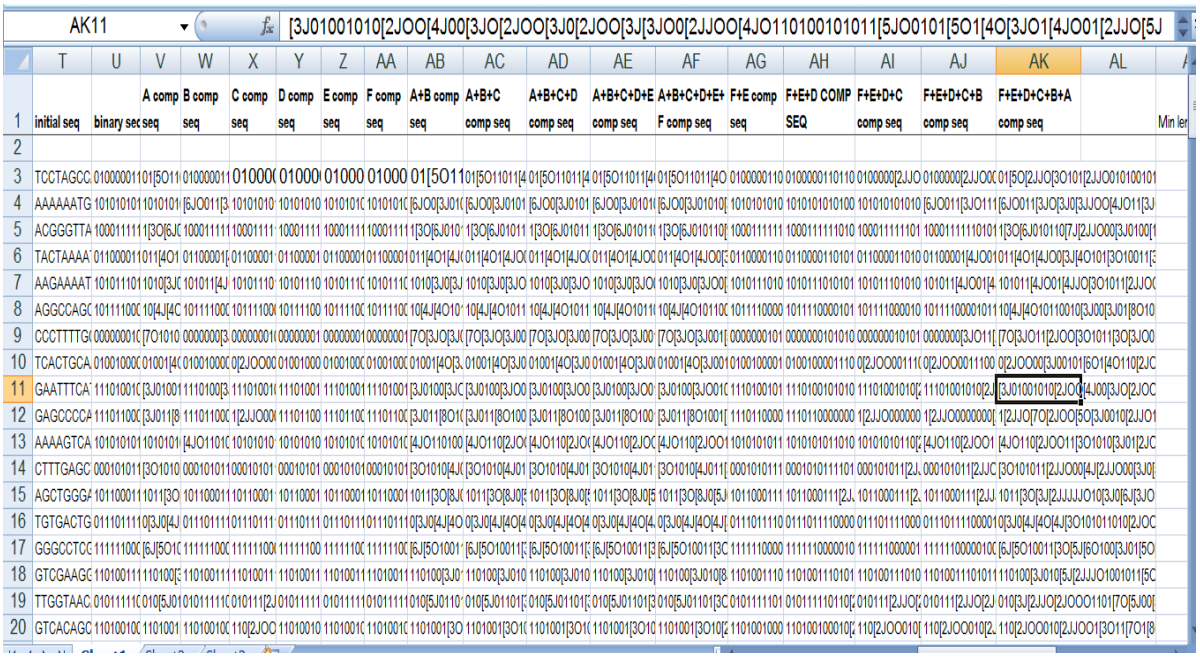


Figure 9: Compression Scheme

Here we have shown the compressibility in sensitive sequences.

1. First of all we have recorded a macro using Notepad++ to make a single stretch of whole sequence.
2. Now using a single stretch of the first sensitive sequence as shown in column T3, we have calculated the initial length of this sequence using the function=LEN (T3).
3. In the column C, we have converted the initial sequence to the binary using the recorded macro and its length was calculated using the LEN function. Here we are using the following binary conversions:

Table 9: Binary replacement of nucleotides

S.No.	Nucleotide	Replacement
1.	A	11
2.	T	10
3.	C	01
4	G	00

4. Similarly the length of all compressed sequences was calculated using the LEN function respectively as shown in the table below:

Table 9: Description of figure 3

S.No.	Motif repeats	Columns of the Compressed length	Columns for the Compressed sequence
1.	Initial length	B	T
2.	Binary length	C	U
3.	A	D	V
4.	B	E	W
5.	C	F	X
6.	D	G	Y
7.	E	H	Z
8.	F	I	AA
9.	A+B	J	AB
10.	A+B+C	K	AC
11.	A+B+C+D	L	AD
12.	A+B+C+D+E	M	AE
13.	A+B+C+D+E+F	N	AF
14.	F+E	O	AG
15.	F+E+D	P	AH
16.	F+E+D+C	Q	AI
17.	F+E+D+C+B	R	AJ
18.	F+E+D+C+B+A	S	AK

5. Similarly we have calculated the compression for the resistant and sensitive sequences.
6. After calculating all the length, we have now used the minimum function MIN in order to find the motif having the minimum length across all and F->E->D->C->B->A scheme showed the best results.
7. Now we have calculated the compression ratio in column AN, using the formula

$$\text{Percentage compression ratio} = \frac{\text{Minimum length}}{\text{Initial length}} \text{x}100 = \frac{\text{C}}{\text{AM}} \text{x}100$$

8. Once the PCR was calculated we have analysed the arithmetic mean, standard deviation, pooled variance.
9. And at last t-test was calculated for the PCR.

CHAPTER-6

RESULTS

RESULTS

Measurement of DNA sequence randomness using RUNS TEST

Runs test was done to check the randomness of the given sequence. We have calculated the Z (critical value), μ(mean) and σ (standard deviation) for the different CGIs. Here we have analysed the z values of the sensitive and the resistant files.

Sensitive CpG islands sequences:

Z-values of all the 29 sensitive CpG island sequences were calculated by determining runs of Adenines and non-Adenines, Cytosine and non-Cytosines, Guanines and non-Guanines, Thymines and non-thymines, purines and pyrimidines, GCs and ATs, GTs and ACs & all four different bases (G, A, T & Cs).

Table 10: Z values for the sensitive CpG islands

S.No.	A & B	C & D	G & H	M & K	R & Y	S & W	T & V	N
1.	7.120	2.721	0.189	1.643	3.622	4.305	3.011	5.458
2.	2.723	0.813	0.566	1.782	1.708	1.138	0.896	1.367
3.	0.006	1.190	3.841	1.865	2.972	1.554	2.712	3.617
4.	3.337	1.448	0.081	0.635	6.705	3.620	8.889	6.325
5.	3.413	4.484	5.296	2.976	4.623	4.576	2.405	6.934
6.	1.932	1.425	0.564	0.015	5.979	2.040	1.146	2.249
7.	6.314	6.391	0.738	5.290	0.333	6.749	13.702	7.021
8.	2.526	2.308	1.838	3.812	4.869	0.662	4.062	4.618
9.	6.870	12.655	2.362	0.564	9.231	8.879	2.581	10.017
10.	3.983	0.798	2.364	2.992	3.887	0.595	2.617	4.306
11.	4.171	1.450	1.506	0.117	5.781	1.528	2.839	4.113
12.	0.582	0.469	7.336	3.359	5.505	2.348	1.866	3.894
13.	1.049	3.279	0.324	5.150	2.989	8.314	2.927	0.217

14.	2.262	2.162	2.885	2.830	4.487	4.818	0.602	1.622
15.	3.732	3.356	0.472	6.529	0.414	4.076	2.972	1.755
16.	0.457	8.293	6.093	10.347	4.495	8.714	4.518	1.613
17.	2.222	0.592	3.911	8.027	2.657	2.021	4.268	4.890
18.	3.678	5.180	0.784	3.063	3.387	9.826	2.861	5.672
19.	2.440	3.663	7.486	4.946	6.812	2.676	5.426	8.336
20.	2.614	2.284	5.024	3.639	1.077	8.359	2.842	2.000
21.	5.178	3.751	5.447	5.159	7.617	11.228	3.660	7.804
22.	0.300	2.335	2.530	5.515	2.660	7.499	0.938	2.574
23.	1.651	7.080	4.857	5.266	7.337	5.218	2.551	4.437
24.	1.343	0.458	1.629	1.921	2.556	5.781	1.478	0.363
25.	1.894	3.766	3.372	3.624	5.821	4.374	0.574	3.266
26.	3.023	0.724	3.952	3.981	3.998	6.234	0.068	1.360
27.	2.075	0.431	1.078	17.575	3.023	6.799	2.227	0.6140
28.	0.841	0.696	0.337	3.120	2.169	7.077	1.471	0.537
29.	4.757	0.048	5.037	7.842	4.886	4.841	1.183	1.1505

Resistant CpG islands sequences:

Z-values of all the 103 resistant CpG island sequences were calculated by determining runs of Adenines and non-Adenines, Cytosine and non-Cytosines, Guanines and non-Guanines, Thymines and non-thymines, purines and pyrimidines, GCs and ATs, GTs and ACs & all four different bases (G, A, T & Cs).

Table 11: Z values for resistant CpG Island

S.No.	A & B	C & D	G & H	M & K	R & Y	S & W	T & V	N
1.	4.742	2.239	3.380	0.825	6.814	2.636	4.320	5.727
2.	4.210	3.737	2.375	0.387	6.538	7.260	8.842	8.101
3.	4.704	1.565	1.887	0.621	5.343	3.056	4.618	5.148
4.	5.946	2.082	2.779	1.350	7.206	3.062	6.031	6.595
5.	4.760	1.394	4.134	2.685	5.514	2.514	4.265	6.181
6.	7.794	2.065	2.428	3.389	6.526	1.585	6.124	6.638
7.	2.253	2.897	1.920	1.232	6.494	0.932	4.619	4.996
8.	4.672	3.390	2.725	1.472	9.840	1.363	7.452	7.277
9.	4.276	0.213	0.191	0.801	5.041	0.035	3.173	2.388
10.	2.675	0.410	0.395	1.888	2.043	0.050	2.400	2.226
11.	5.001	2.427	3.009	1.358	7.959	2.320	5.177	6.706
12.	4.460	0.948	0.290	2.257	1.873	2.308	2.544	3.710
13.	4.649	1.810	0.822	2.122	3.947	1.169	3.776	4.123
14.	6.630	3.625	2.509	4.413	5.110	4.237	6.122	7.921
15.	3.610	1.930	1.022	0.031	6.304	1.056	4.750	4.232
16.	2.931	1.563	2.997	0.229	7.232	1.092	4.237	4.544

17.	2.055	1.052	1.027	0.830	5.319	0.268	3.899	3.360
18.	1.391	1.110	2.321	2.454	2.987	1.290	2.795	3.157
19.	2.540	0.716	0.308	1.034	4.086	1.123	2.147	2.295
20.	5.439	2.405	3.354	1.670	8.394	0.831	3.994	6.275
21.	0.567	1.042	1.822	4.290	8.587	2.452	4.907	1.036
22.	1.173	0.912	2.326	50.146	3.069	-0.001	4.969	3.724
23.	0.549	1.727	0.347	1.472	2.999	3.519	1.310	0.953
24.	4.424	1.755	0.599	-0.008	3.321	3.043	2.364	3.486
25.	1.408	3.894	2.398	0.158	6.198	2.096	2.659	4.613
26.	3.812	1.207	0.626	0.329	3.840	0.725	0.737	0.0147
27.	5.228	3.218	4.178	51.837	7.490	1.055	1.399	5.936
28.	3.810	1.869	1.772	0.074	6.993	2.293	5.866	5.224
29.	4.124	1.882	2.467	0.414	6.443	2.478	4.702	5.330
30.	4.156	3.220	3.595	0.510	6.725	4.550	5.053	6.629
31.	5.000	1.264	0.451	0.347	3.272	4.222	5.771	4.176
32.	0.944	2.894	1.422	2.935	4.960	3.007	1.236	2.914
33.	3.990	5.149	3.566	2.409	4.315	5.014	4.249	6.636
34.	1.878	4.651	1.252	0.072	7.473	1.506	5.353	3.246
35.	2.950	1.946	4.151	2.126	4.741	2.329	3.174	5.319
36.	5.150	2.536	2.552	1.756	4.557	1.788	-0.003	4.658
37.	3.822	1.935	2.932	3.841	4.780	1.641	5.065	5.941
38.	3.598	0.840	1.972	1.691	4.620	1.489	4.907	4.476
39.	2.464	5.351	1.888	3.007	6.613	-0.001	2.905	5.547
40.	3.758	1.123	3.216	1.489	4.871	2.453	4.005	5.026
41.	5.316	4.091	4.424	2.530	5.377	4.544	2.770	7.155
42.	3.627	2.421	4.980	0.395	12.376	1.611	9.233	8.067

43.	2.117	1.461	3.558	0.215	10.429	0.642	7.503	6.273
44.	0.413	3.316	2.925	1.606	6.694	0.784	5.560	5.226
45.	4.968	3.612	2.115	4.052	2.537	4.990	5.103	6.590
46.	2.949	2.909	4.492	4.341	5.312	2.035	5.605	6.666
47.	0.857	0.968	0.303	0.367	3.615	0.715	4.767	2.677
48.	3.892	1.018	1.131	1.048	1.085	0.689	0.264	0.870
49.	5.566	4.304	2.960	3.374	5.824	4.239	5.343	7.731
50.	6.239	3.603	2.408	0.770	6.805	6.288	5.065	6.628
51.	0.835	0.924	0.015	1.271	4.381	4.139	0.579	0.990
52.	2.928	2.055	1.270	0.390	3.810	0.470	2.827	2.1447
53.	4.285	2.916	3.020	0.500	7.405	2.574	4.169	5.940
54.	5.896	2.360	1.004	3.603	3.819	3.521	5.756	6.315
55.	5.962	1.419	4.288	2.575	5.200	3.305	3.369	6.269
56.	6.259	3.770	2.590	0.876	5.594	6.520	5.379	7.229
57.	4.413	3.265	1.377	2.399	5.215	3.420	6.107	6.365
58.	5.180	1.338	3.222	3.676	3.033	1.803	2.005	4.905
59.	4.469	1.620	0.122	0.979	3.757	3.788	2.820	3.733
60.	7.014	4.324	4.878	1.736	9.452	3.354	3.673	8.1957
61.	1.579	1.994	5.259	2.589	8.213	1.730	3.624	5.342
62.	2.627	3.266	2.471	4.464	2.248	3.811	4.626	0.662
63.	7.098	1.028	2.952	0.374	6.251	5.254	6.097	6.661
64.	1.461	2.599	2.942	2.623	6.440	1.405	3.927	4.510
65.	3.113	2.007	2.265	2.378	3.859	0.644	-0.036	3.237
66.	1.691	-0.008	4.567	0.593	6.613	3.087	0.760	2.466
67.	5.981	5.713	4.002	3.837	8.156	5.040	7.884	9.610
68.	8.607	3.639	3.695	0.372	8.010	0.884	3.538	5.232

69.	1.428	2.104	0.819	1.332	2.709	1.151	1.421	1.667
70.	4.712	1.591	3.317	1.888	4.064	3.392	3.536	5.153
71.	5.230	2.374	2.754	1.070	5.037	5.152	4.995	6.447
72.	6.263	4.964	4.015	4.074	6.458	6.769	8.853	9.618
73.	3.508	-0.010	0.252	0.355	2.394	0.605	1.345	1.922
74.	0.671	1.085	4.367	1.650	6.651	2.588	8.209	6.179
75.	3.109	1.661	1.012	1.925	3.603	0.533	2.930	3.492
76.	6.886	7.416	3.807	3.810	8.933	6.015	7.919	10.308
77.	5.620	3.738	1.265	2.757	6.151	3.119	6.904	6.820
78.	1.703	1.281	1.051	1.933	2.997	0.561	3.637	3.143
79.	2.700	2.471	1.433	0.973	6.807	1.433	2.331	3.713
80.	1.084	3.733	1.880	1.073	6.315	1.948	0.074	3.117
81.	2.174	4.720	2.651	2.133	7.479	0.874	2.055	5.055
82.	2.472	3.492	3.590	2.792	6.939	1.319	1.742	4.952
83.	4.869	3.155	3.742	0.673	8.376	0.905	1.722	5.653
84.	1.117	0.693	4.012	3.250	5.007	3.724	0.075	2.907
85.	1.462	4.740	2.937	3.353	5.457	0.564	2.784	5.358
86.	1.097	2.692	1.754	2.030	5.049	2.668	0.240	2.710
87.	28.325	0.645	2.433	1.379	6.414	0.841	3.096	4.035
88.	0.671	0.258	3.171	2.296	2.757	1.584	1.662	2.137
89.	5.074	1.048	0.276	0.118	4.673	0.365	1.906	3.125
90.	0.223	5.257	3.127	1.604	9.027	1.400	3.500	5.391
91.	3.547	2.560	0.547	0.281	5.564	2.955	5.597	5.033
92.	3.373	4.754	3.257	5.247	4.411	0.792	2.441	5.966
93.	5.454	0.948	0.692	-0.017	1.872	2.721	2.936	2.3755
94.	2.240	7.148	5.122	3.850	6.852	2.003	0.236	7.200

95.	3.658	4.880	3.359	2.604	7.532	0.823	0.392	5.423
96.	5.892	4.419	4.168	3.235	7.702	6.028	9.774	10.687
97.	1.122	3.263	6.403	3.107	8.305	3.686	1.596	5.526
98.	4.214	2.857	3.596	3.526	6.469	0.375	2.280	6.501
99.	5.583	3.305	1.580	1.023	6.275	4.006	4.869	4.767
100.	4.178	1.629	1.462	0.690	5.938	1.702	5.455	5.366
101.	2.694	3.420	3.433	1.953	5.921	1.642	3.124	6.277
102.	4.763	3.355	2.508	3.267	5.709	1.919	4.196	4.616
103.	4.291	1.826	1.964	2.608	5.451	0.137	2.821	5.100

Average z-values of sensitive CpG islands and the resistant CpG Islands was calculated using AVERAGE function.

Table 12: Average z-values of resistant CpG islands and sensitive CpG Islands

Average Z value	**Column1**	**Column2**
	Resistant	Sensitive
AB	3.925164	2.84449
CD	2.5378358	2.9051498
GH	2.4698287	2.8240408
MK	2.7933507	4.2615515
RY	5.6625924	4.1931487
SW	2.3440177	5.0292042
TV	3.8532222	3.0100639
N	4.9705493	3.7289559

Z-values obtained from the sensitive CpG islands and resistant CpG island sequences were exhibiting high variation amongst different categories. Therefore clear results of sequence randomness cannot be concluded from the data.

Comparision of Compressibilty of sensitive and resistant CpG islands

In order to find the compression all the sequences of sensitive and resistant CpG islands were subjected to the compression algorithms. Dividing compressed length by original length, Percent Compressed Ratio were determined for all the sequences. A two sample t-test was applied to the data. It compares the mean of two populations. It indicates whether or not the difference between two group's averages most likely reflects a real difference in the population from which the groups were sampled. Various compression schemes were applied and the one with the consistent minimum value was selected to calculate the compression ratio. Here we have used the one sample two tailed t-test.

PCR (Percentage compression ratio) Percentage Compression Ratio was calculated by dividing minimum compressed length by the total length multiplied by 100.

PCR of sensitive CpG islands

Table 13: PCR for sensitive CpG Islands

S.No.	PCR
1	77.247
2	78.404
3	75.347
4	79.879
5	73.826
6	79.622
7	75.666
8	75.893
9	77.161
10	79.535
11	77.301
12	72.978
13	75.044
14	75.263
15	74.136

S.No.	PCR
16	72.624
17	75.444
18	78.545
19	72.933
20	76.876
21	79.786
22	79.177
23	70.858
24	75.626
25	71.958
26	75.553
27	75.826
28	77.414
29	77.625

Resistant CpG islands:

Table 14: PCR for resistant CpG Islands

S.No.	PCR
1.	73.197
2.	75.919
3.	76.107
4.	74.55
5.	77.94
6.	75.65
7.	75.496
8.	74.694
9.	77.74
10.	76.031
11.	77.152
12.	79.136
13.	75.617
14.	75.475
15.	75.607
16.	73.245
17.	76.906
18.	75.711
19.	78.51
20.	76.092
21.	74.579
22.	76.297
23.	75.607
24.	76.154
25.	76.076
26.	77.824
27.	71.748
28.	76.349
29.	75.876
30.	74.894
31.	74.835
32.	76.112
33.	75.27
34.	73.742

S.No.	PCR
35	76.653
36.	76.874
37.	75.8
38.	75.606
39.	74.734
40.	76.671
41.	74.763
42.	74.241
43.	75.558
44.	75.317
45.	74.669
46.	74.028
47.	76.106
48.	74.687
49.	77.032
50.	72.865
51.	77.802
52.	74.381
53.	74.843
54.	77.502
55.	75.202
56.	75.239
57.	75.7
58.	74.222
59.	77.935
60.	72.224
61.	74.334
62.	74.458
63.	75.736
64.	71.528
65.	75.515
66.	75.651
67.	74.218
68.	70.998
69.	77.254

S.No.	PCR
70	73.997
71.	76.988
72.	72.583
73.	77.571
74.	73.062
75.	75.529
76.	70.69
77.	75.623
78.	75.043
79.	75.584
80.	73.592
81.	71.713
82.	72.042
83.	74.306
84.	73.867
85.	72.245
86.	78.331
87.	75.103
88.	71.96
89.	76.358
90.	72.397
91.	77.685
92.	72.895
93.	75.057
94.	71.074
95.	71.643
96.	71.688
97.	68.153
98.	74.447
99.	76.515
100.	74.618
101.	71.857
102.	76.332
103	76.711

ONE SAMPLE TWO TAILED T-TEST:

H_0: $\mu = 0$

H_A: $\mu \neq 0$

Table 15: t-test results

	Sensitive	Resistant
N	29	103
Am	76.122	74.988
S	2.467753157	1.979872867
Ss	170.514558	399.82945
Pooled variance		4.3872616

SAMPLE	T
PCR	0.005545938
	0.012516237

This hypothesis was rejected as calculated t value is less than the critical p value.

The PCR value of sensitive CpG islands was found to be significantly higher than that of resistant CpG islands. So compressibility of CpG methylation resistant CpG islands was greater than that of methylation sensitive ones.

CHAPTER-7

DISCUSSION

DISCUSSION

All the genetic information is stored in the DNA. However there are certain heritable changes that are not encoded in the DNA sequences and such modifications are termed Epigenetics modification. There is a change in the phenotype not in the genotype. DNA methylation, histone modification and the non coding RNA (ncRNA) are the common epigenetic mechanism. Currently, DNA methylation is one of the studied and well discussed epigenetic modifications. DNA methylation plays a vital role in the control of gene expression, development, genomic imprinting, carcinogenesis, X chromosome inactivation and genomic stability.DNA methylation involves the transfer of the methyl group to the 5^{th} position of the Cytosine residue forming 5mC, reaction is catalyzed by the Dnmts. In eukaryotes three different types of Dnmts (Dnmt 1, Dnmt3a and Dnmt 3b) required for the requirement and establishment of methylation. Along with these three Dnmts, two additional Dnmts are there *i.e.,* with the related functions.

In eukaryotes, certain regions contain high regions of dinucleotide CpG and are referred as CpG Islands. These islands are short stretches of GC rich DNA, mostly associated with the promoter regions of genes and are not methylated. Gardiner-Garden and Frommer in 1987 as a long stretch of DNA of 200 bp having CG content of about 50% and observed CpG/expected CpG in excess of 0.6.This definition of CpG island was modified by Takai and Jones in 2001, as a long region of 500 bp DNA with a G+C content more than or equal to 55% and observed over expected CpG value more than or equal to 0.65.

Certain genes are often more methylated than others. While most of the CpG sites in the human genome are methylated, those in the CpG Islands are unmethylated in the normal tissue. Certain CpG islands are more prone to methylation while some are less methylated. The methylation prone and methylation resistant CpG Islands are similar in respect to size, C+G context, CpG frequency and chromosomal location but are found to be different based on its sequence context.

Earlier work done in this field, suggested an epigenetic tool that was designed to discriminate the CpG Islands that were prone to methylation and the unmethylated. A scoring prediction tool was designed which covered various aspects which referred to sequence, repeats, predicted structure, CpG islands, genes, predicted binding sites, conservation, and single

nucleotide polymorphisms. The analysis was carried out on the 132 sequences of the chromosome 21. The result showed that the three attributes *i.e.*, specific DNA repeats, particular DNA structure and certain sequence patterns were highly correlated with the CpG Island methylation.

It has been found out that in the genomic DNA, distribution of CpG dinucleotide is non uniform. It is found to occur with more frequency in CpG Island as compared to bulk genome and within CpG Island itself, the distribution of CpG dinucleotides varies. The dissimilarity of sequences in methylated and unmethylated Islands may be attributed to differences in CpG distribution. Here we are using two attributes: compression and randomness to determine the methylation prone and methylation resistant sequences. Compression is inversely proportional to the complexity of sequence. The complexity of the sequence can be measured by using Kolmogorov complexity. Kolmogorov complexity is based on algorithmic information theory considering objects as individual symbol strings.

Another attribute randomness was checked using the runs test. Here we had compared the z value of all the sensitive and the resistant sequences. However z values were highly variable in all the eight different methods of runs test analysis. Actually there is mutually conflicting results when mean Z-values are compared.

Run Length Encoding (RLE) was designed to analyse the compression and decompression of the sequences. Both the results were checked using the pairwise alignment to check the correctness of the designed algorithm. The algorithm which gave the minimum compression was selected for the best result. Here F->E->D->B->A compression was selected for the further analysis. Percentage Compression Ratio (PCR) was calculated in both the sensitive and resistant sequence. Average PCR for resistant was 74.988 and for sensitive it was 76.122. More the CGIs less will be methylation. It is a known fact that sequences with higher GC content and higher Obs/Exp GpG values have lower probability of methylation. Both the attributes indicate lower complexity of the sequences and as a result higher expected compressibility. The data is in good accordance with this observation.

It is evident that information theory can reveal a lot about functional roles of different types of DNA sequences. In this work, a novel attribute has been reported which appears to differentiate subtly different CpG islands by comparing their compressibility.

CHAPTER-8

CONCLUSION

CONCLUSION

Through the present work we have analysed the information complexity in the sensitive and resistant sequences. Here we have analysed the sequences using the two parameters: randomness and compression. We tried to figure what makes some CpG island sequences highly methylated whereas some unmethylated. Through compression it was concluded that more the compression ratio lesser the complexity in the sequences. Resistant sequences were highly complexed with better compression percentage. The randomness test shows that the insignificant z value are highly scattered and with large variations. So compressibility of the CpG islands is a more reliable attribute to differentiate between methylation sensitivity and resistance.

CHAPTER-9

REFERENCES

REFERENCES

1. Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic acids research, 8(7), 1499-1504.
2. Bock, C., Paulsen, M., Tierling, S., Mikeska, T., Lengauer, T., & Walter, J. (2006). CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. PLoS Genet,2(3), e26.
3. Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. Genes & development, 25(10), 1010-1022
4. Feltus, F. A., Lee, E. K., Costello, J. F., Plass, C., & Vertino, P. M. (2006). DNA motifs associated with aberrant CpG island methylation. Genomics,87(5), 572-579.
5. Gardiner-Garden, M., & Frommer, M. (1987). CpG islands in vertebrate genomes. Journal of molecular biology, 196(2), 261-282.
6. Grumbach, S., & Tahi, F. (1994). A new challenge for compression algorithms: genetic sequences. Information Processing & Management,30(6), 875-886.
7. Handa, V., & Jeltsch, A. (2005). Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome. Journal of molecular biology, 348(5), 1103-1112.
8. Hermann, A., Gowher, H., & Jeltsch, A. (2004). Biochemistry and biology of mammalian DNA methyltransferases. Cellular and Molecular Life Sciences CMLS, 61(19-20), 2571-2587.
9. Holliday, R. (2006). Epigenetics: a historical overview. Epigenetics, 1(2), 76-80.
10. Mohamadlou, H., Shope, J. C., & Flann, N. S. (2014). Maximizing Kolmogorov Complexity for accurate and robust bright field cell segmentation. BMC bioinformatics, 15(1), 1.
11. Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. Neuropsychopharmacology, *38*(1), 23-38.
12. Takai, D., & Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proceedings of the national academy of sciences, 99(6), 3740-3745.
13. Tatematsu, K. I., Yamazaki, T., & Ishikawa, F. (2000). MBD2-MBD3 complex binds to hemi-methylated DNA and forms a complex containing DNMT1 at the replication foci in late S phase. Genes to Cells, 5(8), 677-688.

14. Vass, J. K., & Wilson, R. H. (1984). ‘ZSTATS”-a statistical analysis for potential Z-DNA sequences. Nucleic acids research, 12(1Part2), 825-832.
15. Yamada, Y., Watanabe, H., Miura, F., Soejima, H., Uchiyama, M., Iwasaka, T., & Ito, T. (2004). A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. Genome research, 14(2), 247-266.
16. Zar, J. H. (1999). Biostatistical analysis. Pearson Education India.