Comparison of CpG distribution in methylation sensitive and methylation resistant CpG Islands and study the effect on methylation

Submitted in partial fulfillment of the requirement of the degree of

MASTERS OF SCIENCE IN BIOTECHNOLOGY

Under the guidance of:

Dr. Vikas Handa

Assistant Professor

Submitted by:

Shivangi

Roll no. 301201020

DEPARTMENT OF BIOTECHNOLOGY AND ENVIRONMENTAL SCIENCES

THAPAR UNIVERSITY, PATIALA

CERTIFICATE

This is to certify that the report entitled "**Comparison of CpG distribution in methylation sensitive and methylation resistant CpG islands and study the effect on methylation**" submitted by Shivangi (301201020) in partial fulfilment of the requirement for the award of Degree of Masters in Science in Biotechnology to Thapar University, Patiala is a record of student's own work carried out by her under my supervision and guidance. The report has not been submitted for the award of any other degree or certificate in this or any other university.

Dr. Vikas Handa
Assistant Professor
Department of Biotechnology

Dr. Dinesh Goyal
Head of Department
Department of Biotechnology

Dr. S.K. Mahopatra
Dean
(Academic Affairs)
Thapar University
Patiala

CANDIDATE'S DECLARATION

I hereby declare that the work being presented in the report entitled "**Comparison of CpG distribution in methylation sensitive and methylation resistant CpG islands and study the effect on methylation**" in partial fulfilment of the requirement for the award of Degree of Masters in Science in Biotechnology to Thapar University, Patiala is my own work during the period of six months from January to June 2014, under the supervision of Dr. Vikas Handa, Associate Professor, Department of Biotechnology, Thapar University, Patiala. I have not submitted the matter embodied in this report for the award of any other degree.

Date: 18-7-2014

Shivangi

Shivangi
Roll No. 301201020

It is certified that the above statement made by the student is correct to the best of my knowledge and belief.

Date: 18-7-2014

Dr. Vikas Hands
Assistant Professor
Department of Biotechnology

ACKNOWLEDGEMENT

Nothing concrete can be achieved without a combination of inspiration and perspiration. Although writing a few words on a piece of paper is not a proper way to acknowledge those people who had helped me in the completion of this project, yet the words coming from our heart and soul need no mode of communication.

I take the opportunity to present a vote of thanks to all those guideposts who really acted as lighting pillars to enlighten my way throughout this project that has led to the completion of this study.

I find it a matter of honor in showering my gratitude, indebtedness and thankfulness to my guide respected **Dr. Vikas Handa**, Assistant Professor, Department of Biotechnology and Environmental Sciences, Thapar University, Patiala.

A profound acknowledgement goes to **Dr. Dinesh Goyal**, Head, Department of Biotechnology and Environmental Sciences, Thapar University, Patiala.

The work was not possible without the cooperation of the entire faculty members of DBTES.

A special thanks to my friends and coworkers Japnjyot Saini and Tajeshwar Preet Kaur for their support.

Shivangi

Date: 18-7-2014

Place: Patiala

Shivangi

301201020

CONTENTS

Page No.

ABBREVATIONS

$\Delta G_{folding}$	Gibb's free energy of folding
5-mC	5-methylcytosine
Adomet	S-adenosyl L-methionine
C_5	Carbon at 5th position
C_6	Carbon at 6_{th} position
CpA	Cytosine and Adenine
CpG	Cytosine and Guanine
DNMT	DNA methyltransferase
DNMT 1	DNA methyltransferase 1
DNMT3a	DNA methyltransferase 3a
DNMT3b	DNA methyltransferase 3b
DNMT3L	DNA methyltransferase 3L
DNMT2	DNA methyltransferase 2
ENV	Glutamic acid, Asparagines, Valine
ExpCpG	Expected frequency of CpG dinucleotides
HDAC	Histone deacetylase
IR	Inverted repeats
M	Fully methylated CpG Islands
MBD	Methyl-CpG binding protein
mCpG	Methylated CpG
N_6	Nitrogen at 6th position
N_4	Nitrogen at 4th position

ObsCpG	Observed frequency of CpG dinucleotides
ObsCpG/ExpCpG	Ratio of observed frequency of CpG over the expected frequency of CpG dinucleotide
PCQ	Proline, Cysteine, Glutamine
SD	Standard deviation
TpG	Thymine and Guanine
U	Unmethylated CpG Islands
VMR	Variance by mean ratio

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Abstract

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 and occasionally methylated. The methylation is centered on 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 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 based on DNA sequence attributes has been performed to find out why certain CpG Islands tend to get methylated. The abundance or under-representation of tri- and tetranucleotide permutations has been investigated in the two classes of CpG Islands *viz.*,fully methylated and unmethylated. Further $\Delta G_{folding}$ and relative abundance of inverted repeats has also been studied to find any bearing on CpG Island methyation. Finally distribution of CpGs in the CpG Islands has been analyzed. All the four factors were statistically analyzed and they appear to be significantly different in the two classes of CpG Islands and thus can be speculated to influence the methylation of CpG Islands.

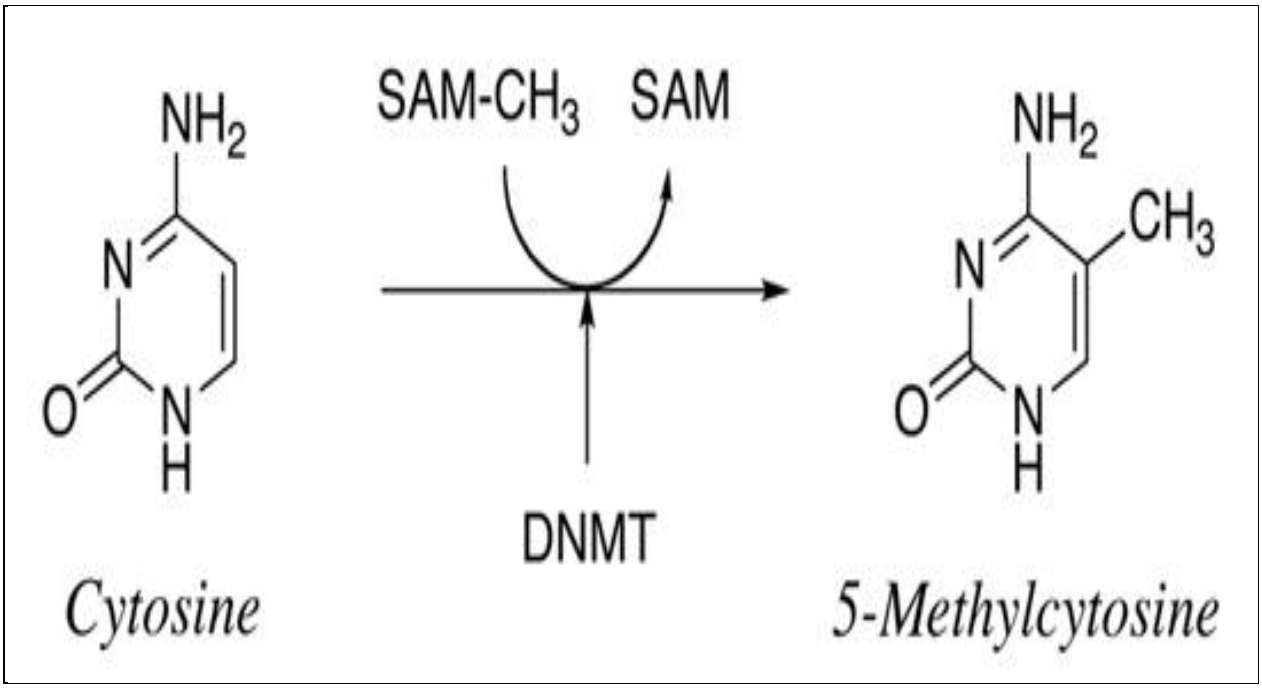
CHAPTER 2

Introduction

Introduction

DNA is genetic material in most of the living organisms that carries genetic information in the form of long sequences of four bases namely, Adenine, Cytosine, Guanine and Thymine. The genetic information is faithfully inherited by daughter cells after cell division as a result of replication. This makes the molecular basis of Mendelian inheritance. However the expression of genetic information is governed by certain other factors that come under Epigenetics. Epigenetics involves genetic control by factors other than an individual's DNA sequence *i.e.,* a change in phenotype without a change in DNA sequence. Epigenetic changes can switch genes on or off and determine which proteins are transcribed under given conditions of space or time. It can contribute to differential expression of genes in different tissues. Within cells, there are three systems that can interact with each other to silence genes: DNA methylation, histone modifications, and RNA-associated silencing (Simmons, 2008). Histone modification is an important component of epigenetics and it comprises of acetylation, ubiquitination, phosphorylation and methylation of histones that affects gene activation/silencing. DNA methylation occurs at N_6 position of adenine, N_4 and C_5 position of cytosine. Among these three types of methylation only one is found at large in higher Eukaryotes *i.e.,* at C_5 position of cytosine (Hotchkiss, 1948). This modification of the DNA alters gene expression in cells. These changes in gene expression are stable and the cell does not revert back to a stem cell or another type of cell during its differentiation from embryonic stem cells into cells of a particular tissue. DNA methylation is vital to healthy growth and development and is linked to various processes such as genomic imprinting, X-chromosome inactivation, carcinogenesis, suppression of repetitive elements and cellular differentiation *etc.*. DNA methylation is implicated in the state of the chromatin structure, which is attributed to cell growth and its differentiation into a complex multi cellular organism made up of different tissues and organs (Hermann *et al.,* 2004). In addition, DNA methylation plays an important part in the development of cancer and is a key regulator of gene transcription. Aberrant methylation of DNA has been associated with an increased rate of malignancy. DNA hypermethylation results in gene silencing and DNA hypomethylation has been associated with the development of cancer (Das and Singal, 2004). DNA methylation also suppresses the expression of retroviral genes, along with other potentially dangerous sequences of DNA that have entered the host genome and may damage the host.

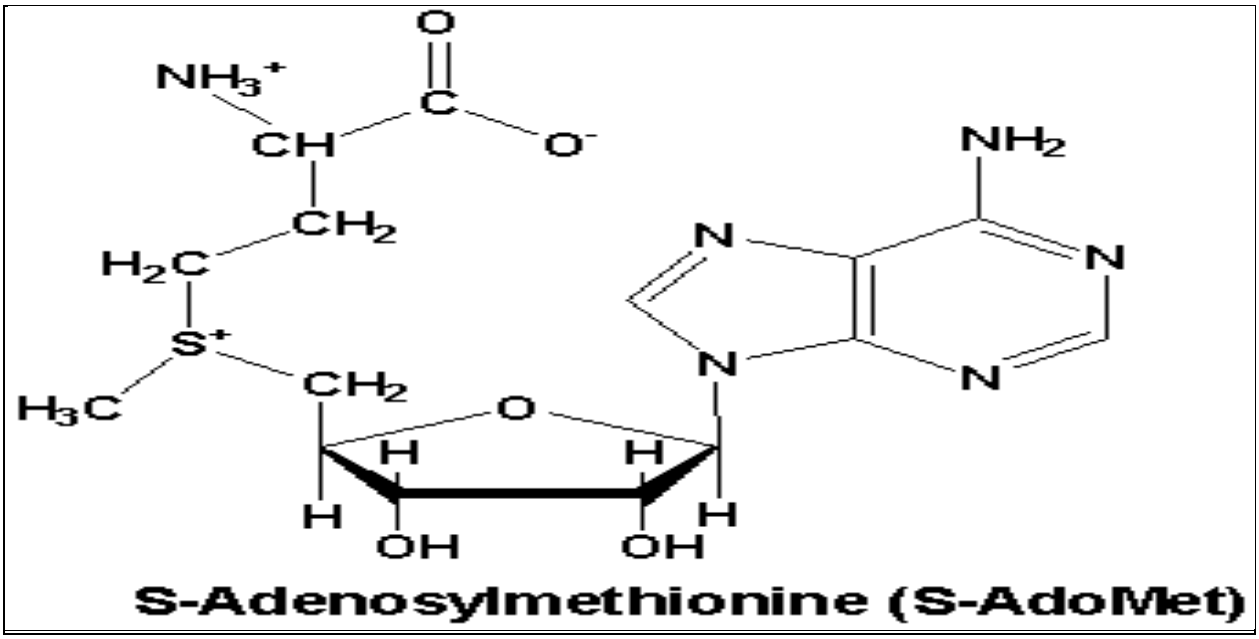
5-methylcytosine DNA methylation is more active in higher eukaryote and occurs at cytosine of dinucleotide CpG (Jeltsch, 2002). It occurs by the covalent addition of a methyl (CH_3) group at the 5-carbon of the cytosine ring resulting in 5-methylcytosine (5-mC). This covalent addition of methyl group is catalyzed by the action of enzymes belonging to the protein family of DNA methyltransferases (DNMT) (Roberts *et al.*, 2003).



http://www-medchem.ch.cam.ac.uk/lab_rotations/murrell.php

Fig. 1. Methylation of Cytosine mediated by DNA methyltransferase.

In eukaryotes three different DNMTs (DNMT1, DNMT3a and DNMT3b) are required for establishment and maintenance of DNA methylation patterns. Two additional enzymes (DNMT2 and DNMT3L) may also have more specialized but related functions. DNMT1 appears to be responsible for the maintenance of established patterns of DNA methylation, while DNMT3a and DNMT3b seem to mediate establishment of new or *de novo* DNA methylation patterns. In order to correctly convey the epigenetic information from one cell to the daughter cells after replication, methylation of DNA occurs and this phenomenon is brought about by transfer of methyl group from S-adenosyl L-methionine (Adomet), which acts as a source of methyl group to the cytosine of CpG dinucleotides (Wu *et al.*, 1985).



http://www.pearsonhighered.com/mathews/ch21/sadomet.htm

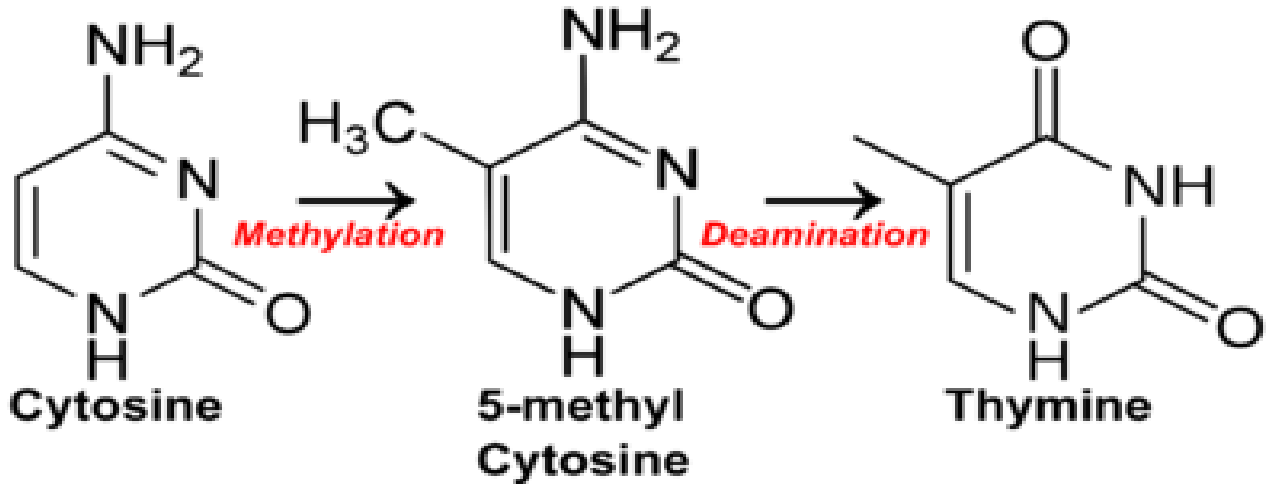
Fig. 2. S-Adenosylmethionine: methyl group donor

Since this methyl group in Adomet is bound to sulphonium atom, it makes the molecule thermodynamically less stable and thus allowing methylthiole to undergo a nucleophillic attack very easily. The enzyme, cytosine-C_5 methyltransferase, consists of certain common motifs in prokaryotes and eukaryotes, *viz.*, Motif IV (PCQ) and Motif VI (ENV). The PCQ motif is present in the active site of the enzyme and the thiol group present in the cystine residue carries out a nucleophilic attack on C_6 of cytosine and thus activating the C_5 atom towards an electrophilic attack. It is accompanied by addition of methyl group to C_5, elimination of 5 position proton and resolution of covalent intermediate. On the other hand, the glutamic acid present in ENV motif stabilizes the DNA-protein complex (Jeltsch, 2002).

Methylation can silence gene by modifying DNA as well as by modifying histones. There are specific sites in 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 deacetylation of histones and thus the DNA starts coiling

on to histones tightly, leading to the formation of heterochromatin, which leads to gene inactivation (Tatematsu *et al.,* 2000).

The CpG dincleotides are under-represented in mammalian genomes. The reason for this suppression can be attributed to CpG methylation. The basis for this connection is the tendency of mCpG to mutate to TpG. Methylated Cytosines are mutational hotspots and they frequently get converted to Thymine causing CpG dinucleotide suppression (Bird, 1980). The transformation of cytosine into thymine is a two step process in which the cytosine bases first gets methylated due to action of DNA methyltransferases. The methylated cytosine then gets deaminated to form thymine. The CpG Island erosion was shown by Matsuo *et al.* in mouse genomes where accumulation of TpGs and CpAs were observed and this was presumably due to higher rate of deamination (Matsuo *et al.,* 1993).



http://en.wikipedia.org/wiki/CpG _site

Fig. 3. Mutation in methylcytosine.

There are certain regions in higher eukaryotic genomes which contain high frequency of dinucleotide CpG and are called CpG Islands. CpG Islands are short stretches of GC rich DNA sequence, mostly associated with promoter region of genes and are usually not methylated. Historically, they were discovered owing to large number of Hpa II restriction sites (CCGG) that resulted in generation of unusually small restriction fragment upon digestion with the restriction enzyme. Formerly they were known as Hpa II tiny fragments (Craig and Bickmore, 1994). In 1987 Gardiner-Garden defined CpG Islands as 200bp or longer stretches of DNA with at least

50% GC content and observed over expected CpG ratio of greater than or equal to 0.6 (Gardiner-Garden and Frommer, 1987). Takai and Jones, in 2001, redefined CpG Islands based on analysis of complete genomic sequence of human chromosome 21 and 22. The new definition gave prediction results matching accurately with the functional features of CpG Islands. They defined CpG Islands as regions of DNA at least 500bp long with C+G content more than or equal to 55% and observed over expected CpG value to be more than or equal to 0.65 (Takai and Jones, 2001).

While most CpG sites in the human genome are methylated, those in CpG Islands are typically unmethylated in normal tissue. However, in human cancers, *de novo* methylation of CpG Island sequences is accompanied by gene silencing and can serve as an alternative to mutation or deletion in the inactivation of tumor suppressor and other genes (Herman *et al.,* 1994). In the bulk of genomic DNA, most CpG sites are heavily methylated while CpG Islands in germ-line tissues and located near promoters of normal somatic cells remain unmethylated, thus allowing gene expression to occur. When a CpG Island in the promoter region of a gene is methylated, expression of the gene is repressed. The present work is an attempt to reason why certain CpG Islands are easy target for methylation and not the others. Much work has been done to find out the reason behind this unusual difference in the enormity of methylated and unmethylated CpG Islands. CpG Island methylation analysis revealed that some CpG Islands are more frequently methylated than others and such CpG Islands are often found to be associated with promoter region of genes.

Feltus *et al.* in 2003, attempted to find out why certain genes are more often methylated than others. They over expressed DNA cytosine-5-methyltransferase 1 (DNMT 1) and studied the behavior of CpG Islands towards methylation. The overall phenomenon of CpG Islands 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, C+G content, CpG frequency and chromosomal location but are found to be different based on its sequence context (Feltus *et al.,* 2003).

Again in 2006, 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 sequences 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).

Further in 2006, Bock *et al.* also attempted to discriminate between CpG Islands that are prone to methylation from those that remain unmethylated by using a computational epigenetic approach. They used 8 attributes to distinguish between the two classes of CpG Islands *viz.*, single nucleotide polymorphism, evolutionary conservation, predicted transcription factor binding site, gene and exon distribution, predicted DNA structure, CpG Island frequency and distribution, DNA sequence properties and patterns, repeat frequency and distribution. Three groups of DNA attributes were found to be highly correlated with CpG Island methylation *viz.*, sequence patterns, specific DNA repeats and particular DNA structure (Bock *et al.*, 2006)

Yamada *et al.* scrutinized methylation status of CpG Islands in human chromosome 21q. Almost all of the human house-keeping genes have CpG Islands associated with them and majority of those islands escape methylation. They developed HpaII–McrBC PCR method for analyzing the extent of methylation. It was found that most of the CpG Islands remain unmethylated and a small fraction is fully methylated. Also, some of the CpG Islands are compositely methylated. Thus they divided the CpG Islands into four groups *viz.*, fully methylated, unmethylated, incompletely methylated and compositely methylated (Yamada *et al.*, 2004).

So far the work done in this sphere revealed that the methylated and unmethylated CpG Islands perceptibly diverge with regard to their sequence context *viz.*, sequence pattern, repeat frequency and particular structure. Apart from these sequence aspects there are certain other remarkable properties associated with DNA sequences that can assist in resolving the matter. The present work is an attempt to explore some of these parameters. A DNA sequence based approach has been applied to uncover the fine differences between CpG Islands that are resistant or sensitive to methylation. In this report the sequence analyses have been performed on the sensitive and resistant CpG Island sequences derived from the work of Yamada *et al*., in 2004.

It has been already known that there are certain sequence components that form the basis for the difference between methylated and unmethylated CpG Islands. A similar approach is applied here to explain the fact. The sequence components can be determined by comparing the oligonucleotide frequencies of length two, three and four in the two classes of CpG Islands.

The methylation of DNA is an enzyme mediated process the activity of enzyme can be largely influenced by the local structural conformations in the structure of DNA. The structural alterations based on the DNA sequence maybe manifested in DNA folding and also thereby in inverted repeat frequency. The double stranded DNA has a tendency to acquire a folded conformation based on occurrence of inverted repeats and free energy associated with DNA folding. The fine as well as coarse variations in DNA structures can influence methylation propensity.

It has been found that the CpG dinucleotide distribution in genomic DNA is not uniform. It is found to occur with more frequency in CpG Islands as compared to bulk genome and again within CpG Islands itself, the distribution of CpG dinucleotides varies. The dissimilarity of sequences in methylated and unmethylated CpG Islands may be attributed to differences in CpG distribution. In this report, CpG distribution has been analyzed in both the classes of CpG Islands and compared for difference in their methylation sensitivity.

CHAPTER 3

Literature review

Literature review

DNA methylation is an influential event leading to epigenetic control of gene expression and it usually occurs at CpG dinucleotide. The extent of CpG methylation varies with respect to cell types and pathological situation. Gardiner-Garden and Frommer, in 1987, screened vertebrate genes for the presence of CpG Islands. Each CpG Island was analyzed in terms of length, nucleotide composition, frequency of CpG dinucleotide and location relative to associated gene. Most CpG Islands were found to be associated with 5' end of all housekeeping genes and many tissue specific genes and 3' end of some tissue specific genes. In certain genes both 5' and 3' CpG Islands are present and are separated by several thousand base pair of CpG depleted DNA. CpG Islands were found in same position relative to the transcription unit of equivalent gene in different species with some prominent exceptions. G/C boxes (GGGCGG) or its reverse complement (CCGCCC) was investigated with respect to location of CpG Islands. It is found to be rare in CpG depleted DNA and plentiful in CpG Islands. The G/C boxes are located in both upstream and downstream from Transcription start site with 5' CpG Islands. These are a feature of CpG Islands in general rather than promoter region of housekeeping genes. They defined CpG Islands as 200 base pair or longer stretch of DNA with C+G to be greater than or equal to 50% and observed over expected CpG frequency of 0.6 or more. This definition of CpG Islands was framed before the sequencing of mammalian genomes (Gardiner-Garden and Frommer, 1987). Takai and Jones, in 2001, improved this definition based on the study of chromosome 21 and 22. They first used the definition of Gardiner-Garden to derive some contigs from chromosome 21 and 22 by using GenBank database, but found it wrong. Majority of the CpG Islands corresponds to Alu repeats. They redefined CpG Islands as stretches of more than or equal to 500 base pairs of DNA with G+C more than or equal to 55% and observed over expected CpG frequency of 0.65 or more (Takai and Jones, 2002). It has been found that the CpG dinucleotide occurs less frequently than expected in many animals (Josse *et al.*, 1961 and Swartz *et al.*, 1962). The deficiency of CpG dinucleotide is related to DNA methylation (Salser, 1977). The probable cause of CpG deficiency is mutation of 5mC to T. Nearest neighbor dinucleotide frequency and level of DNA methylation was investigated and it was found that CpG deficiency is related to mutation (Bird, 1980). Due to this phenomenon, the methylated and unmethylated CpG Islands perceptibly diverge with regard to their abundance in genome and methylation affinity. Feltus *et*

al. in 2003, attempted to answer this question that why some genes give way to this deviant event. They over expressed DNMT1 and analyzed the susceptibility of 1749 CpG Islands to *de novo* methylation. Methylation of CpG Islands increased in cells with DNMT1 over expression but it was found that all the sites were differently influenced by it. Majority, about 69.9%, were found to be resistant to methylation irrespective of DNMT1 over expression and a very small fraction of CpG Islands, about 3.8%, was consistently influenced by DNMT1 over expression. Also, the parameters like C+G content, CpG frequency and chromosomal location were found to be similar in both the domains. The susceptibility of *de novo* CpG Islands methylation varies and this difference can be explained on the basis of sequence context. Seven novel sequence patterns (TCCCCCNC; TTTCCTNC; TCCNCCNCCC; GGAGNAAG; GAGANAAG; GCCACCCC; and GAGGAGGNNG) were determined that can very well discriminate between the two domains with 82% accuracy (Feltus *et al*., 2003). In 2006 the same phenomenon was attempted to answer but in a different manner. There are certain CpG islands that remain methylation resistant and certain are methylation sensitive. They used MEME (Motif-based sequence analysis tool) and MAST (Motif Alignment and Search Tool) algorithm and derived 5 motifs from methylation prone sequences and 8 motifs from methylation resistant sequences. These motifs successfully discriminated between methylation prone and methylation resistant CpG Islands with 87% accuracy. The motifs derived from methylation resistant CpG Islands were found to be strongly associated with *Alu* and are randomly distributed across the genome. Whereas, the methylation prone motifs were found to be selectively associated with CpG Islands. The motifs associated with methylation prone sequences are: GGCTGCGGGGGCAGCAGCTG; AAGAAGGGAGAGAAGGAGGAA; TCCTCTCCCTTGTCTTCCTCCTCCTCCTC; GGGGTGGGGGAGGGGGAGGAG; CTCTCCCAAGC and the motifs associated with methylation resistant sequences are: TTTTTTTTTTTTTTTTTGAGACAGAGTCT; GTCAGGAGTTTGAGA; GCCCAGGCTGGAGTGCAGTGG; GCCTGTAATCCCAGCTACTCAGGAGGCTGAGGCAGGA; AAAGTGCTGGGATTACAGGCGTGAGCCA; CTCACTGCAACCTCCGCCTCCCGGGTTCA; TGATCCGCCCGCCTCGGCCTC; CCAGCCTGGCCAACATGGTGA.

(Feltus *et al*., 2006). Further in 2006, Bock *et al.* also attempted to discriminate between CpG Islands that are prone to methylation from those that remain unmethylated by using a

computational epigenetic approach. They used 8 attributes to distinguish between the two classes of CpG Islands *viz.,* single nucleotide polymorphism, evolutionary conservation, predicted transcription factor binding site, gene and exon distribution, predicted DNA structure, CpG Island frequency and distribution, DNA sequence properties and patterns, repeat frequency and distribution. These attributes were scored on 132 CpG Islands across the entire human Chromosome 21. The dataset was obtained from the work of Yamada *et al.* 2004. Three groups of DNA attributes that were found highly correlated with CpG Island methylation are sequence patterns, specific DNA repeats and particular DNA structure (Bock *et al.*, 2006). The methylation of human genomic DNA is mediated by DNA methytransferases that methylates cytosine residue within CpG dinucleotide. The data of human epigenome project was statistically analyzed to determine the flanking sequence upto ± four base pair surrounding the central CpG site *i.e.,* (5′-CTTGCpGCAAG-3′) and (5′-TGTTCpGGTGG-3′). The former sequence element is found to be associated with high level of methylation whereas the latter is associated with low level of methylation. A set of synthetic oligonucleotide substrates were used to investigate the influence of flanking sequences on catalytic activity of DNMT3a and DNMT3b. More than 13-fold difference is revealed between the preferred (RCGY) and disfavored (YCGR) ±1 flanking base pair. It revealed the intense flanking sequence preference of DNMT3a and DNMT3b and it could be involved in origin of CpG Islands (Handa and Jeltsch, 2005).

The methylation status of CpG Islands on human chromosome 21q was determined in 2004 by Yamada *et al.*. They repeat masked the sequence of human chromosome 21 and computationally identified all non-repetitive CpG Islands. The parameters for CpG Island identification includes: GC content above 50%, ratio of observed versus expected number of CpG dinucleotide above 0.6 and minimum length 400 base pairs. Their dataset encompasses the methylation status of 149 CpG Islands belonging to four categories: fully methylated, unmethylated, incompletely methylated and compositely/differentially methylated (Yamada *et al.,* 2004). This dataset forms the basis of present work.

CHAPTER 4

Materials and Method

Materials and method

Data source- The dataset comprises of DNA sequences comprising of two different kinds of CpG Islands *viz.*, fully methylated (M) and unmethylated (U). These two sets of DNA sequences are obtained from the published work of Yamada *et al.*. The report classified CpG Island sequences into four categories *viz.*, fully methylated, unmethylated, incomplete and compositely methylated (Yamada *et al.*, 2004). Two of these four categories *i.e.*, fully methylated and unmethylated, have been analyzed in the present work.

Sequences and regions	Chromosome	Length	Methylation status
(NT_002836.4 740746-742525)	21	1273	fully methylated
(NT_002836.4 798428-799837)	21	1410	fully methylated
(NT_002836.4 1099894-1101335)	21	1178	fully methylated
(NT_002836.4 2100365-2102278)	21	1914	unmethylated
(NT_002836.4 2765740-2767861)	21	2122	unmethylated
(NT_002836.4 4548804-4550994)	21	2191	unmethylated
(NT_002836.4 4648389-4650557)	21	2169	unmethylated
(NT_002836.4 4854926-4857042)	21	1959	unmethylated
(NT_002836.4 12511562-12513060)	21	1516	unmethylated
(NT_002836.4 12684911-12686522)	21	1612	unmethylated
(NT_002836.4 13119202-13121651)	21	2450	unmethylated
(NT_002836.4 13793868-13795756)	21	1889	unmethylated
(NT_002836.4 13915193-13916938)	21	1746	unmethylated
(NT_002836.4 13917154-13918663)	21	2000	unmethylated
(NT_002836.4 15834745-15836237)	21	1287	unmethylated
(NT_002836.4 16246692-16248676)	21	1985	unmethylated
(NT_002836.4 18607855-18609856)	21	1480	unmethylated
(NT_002836.4 18679791-18682057)	21	2267	unmethylated
(NT_002836.4 18821198-18823391)	21	2194	unmethylated
(NT_002836.4 19227134-19228707)	21	2000	unmethylated
(NT_002836.4 19248734-19250388)	21	1196	fully methylated
(NT_002836.4 19359982-19362161)	21	2052	unmethylated
(NT_002836.4 19561116-19562518)	21	975	unmethylated
(NT_002836.4 19675812-19677620)	21	1809	unmethylated
(NT_002836.4 19719511-19721351)	21	1673	unmethylated
(NT_002836.4 19972011-19973417)	21	2000	unmethylated
(NT_002836.4 19975198-19977216)	21	2019	unmethylated
(NT_002836.4 20018303-20020533)	21	1773	unmethylated
(NT_002836.4 20178164-20179883)	21	1639	unmethylated
(NT_002836.4 20273122-20274794)	21	1426	unmethylated
(NT_002836.4 20351550-20353202)	21	1274	unmethylated
(NT_002836.4 20427755-20429718)	21	1661	unmethylated
(NT_002836.4 20536702-20538471)	21	1770	unmethylated
(NT_002836.4 20590829-20592722)	21	1894	unmethylated
(NT_002836.4 21021661-21023784)	21	2124	unmethylated

(NT_002836.4 21323771-21325344)	21	1146	unmethylated
(NT_002836.4 21562903-21564940)	21	2038	unmethylated
(NT_002836.4 21617860-21620045)	21	2186	unmethylated
(NT_002836.4 21740221-21742136)	21	1916	fully methylated
(NT_002836.4 21835266-21836717)	21	1452	unmethylated
(NT_002836.4 22835347-22836774)	21	1428	fully methylated
(NT_002836.4 23008324-23010418)	21	1161	unmethylated
(NT_002836.4 23018419-23020042)	21	1284	unmethylated
(NT_002836.4 23083543-23085110)	21	1568	unmethylated
(NT_002836.4 23104500-23106945)	21	2446	unmethylated
(NT_002836.4 23268373-23270063)	21	1493	unmethylated
(NT_002836.4 23333456-23335039)	21	1584	unmethylated
(NT_002836.4 23646833-23649367)	21	2535	unmethylated
(NT_002836.4 23648887-23650365)	21	2000	unmethylated
(NT_002836.4 23657131-23658548)	21	1416	unmethylated
(NT_002836.4 23695691-23697656)	21	1439	unmethylated
(NT_002836.4 23914335-23916288)	21	1954	unmethylated
(NT_002836.4 23938177-23939758)	21	956	unmethylated
(NT_002836.4 24021123-24022718)	21	1596	unmethylated
(NT_002836.4 24215237-24217598)	21	2124	unmethylated
(NT_002836.4 24314183-24317285)	21	3103	unmethylated
(NT_002836.4 24511883-24513479)	21	1155	unmethylated
(NT_002836.4 25608278-25609973)	21	1696	unmethylated
(NT_002836.4 25753617-25755681)	21	2065	unmethylated
(NT_002836.4 26130818-26133115)	21	924	unmethylated
(NT_002836.4 26259167-26262261)	21	1064	unmethylated
(NT_002836.4 26295243-26297124)	21	1773	unmethylated
(NT_002836.4 26390768-26392267)	21	1270	unmethylated
(NT_002836.4 26556991-26558660)	21	1356	unmethylated
(NT_002836.4 27791317-27792866)	21	1550	unmethylated
(NT_002836.4 28112154-28114765)	21	2612	unmethylated
(NT_002836.4 28452242-28454382)	21	2141	unmethylated
(NT_003545.2 122339-124366)	21	2028	unmethylated
(NT_003545.2 178197-181181)	21	2945	unmethylated
(NT_003545.2 387867-389371)	21	1364	unmethylated
(NT_003545.2 403601-405007)	21	1227	unmethylated
(NT_003545.2 665089-666509)	21	1421	unmethylated
(NT_003545.2 682430-684731)	21	2302	unmethylated
(NT_003545.2 756256-760387)	21	4132	fully methylated
(NT_003545.2 822395-823837)	21	1443	unmethylated
(NT_003545.2 854975-856401)	21	766	fully methylated
(NT_003545.2 1017136-1019217)	21	1947	fully methylated
(NT_003545.2 1047902-1049365)	21	1437	unmethylated
(NT_003545.2 1142956-1145447)	21	2492	unmethylated
(NT_003545.2 1275925-1277810)	21	1132	unmethylated
(NT_002835.3 158046-161231)	21	1096	unmethylated
(NT_002835.3 297286-298768)	21	514	fully methylated
(NT_002835.3 389981-391414)	21	1430	fully methylated

(NT_002835.3 391442-393133)	21	1329	unmethylated
(NT_002835.3 473197-474680)	21	1484	fully methylated
(NT_002835.3 508085-509942)	21	1513	unmethylated
(NT_002835.3 521399-523051)	21	931	unmethylated
(NT_002835.3 596448-599173)	21	1008	unmethylated
(NT_002835.3 743918-746243)	21	1863	unmethylated
(NT_002835.3 839100-840858)	21	1941	unmethylated
(NT_002835.3 865543-867041)	21	1204	unmethylated
(NT_002835.3 972548-974376)	21	776	unmethylated
(NT_002835.3 974194-975798)	21	1723	unmethylated
(NT_002835.3 1031958-1033740)	21	1783	unmethylated
(NT_002835.3 1070778-1072793)	21	1706	unmethylated
(NT_002835.3 1082093-1084166)	21	2074	unmethylated
(NT_002835.3 1101487-1103181)	21	1695	fully methylated
(NT_002835.3 1187382-1188869)	21	1408	unmethylated
(NT_002835.3 1441617-1443042)	21	1426	unmethylated
(NT_002835.3 1533742-1535201)	21	1460	unmethylated
(NT_002835.3 1549573-1551858)	21	2154	unmethylated
(NT_002835.3 1604933-1607527)	21	2245	unmethylated
(NT_002835.3 1664660-1666157)	21	1498	unmethylated
(NT_002835.3 1671866-1673671)	21	1349	unmethylated
(NT_002835.3 1680257-1681681)	21	1118	fully methylated
(NT_002835.3 1710450-1712127)	21	1177	fully methylated
(NT_002835.3 1750441-1752578)	21	1749	unmethylated
(NT_002835.3 1806394-1808604)	21	1674	unmethylated
(NT_002835.3 1865795-1867783)	21	1392	fully methylated
(NT_002835.3 2020037-2021768)	21	1609	unmethylated
(NT_002835.3 2077997-2079517)	21	1060	fully methylated
(NT_002835.3 2112166-2114090)	21	1925	fully methylated
(NT_002835.3 2136861-2139459)	21	2599	unmethylated
(NT_002835.3 2286418-2288635)	21	1687	fully methylated
(NT_002835.3 2364595-2366958)	21	541	fully methylated
(NT_002835.3 2371534-2374424)	21	2891	unmethylated
(NT_002835.3 2475023-2477222)	21	2200	fully methylated
(NT_002835.3 2653007-2655508)	21	1994	fully methylated
(NT_002835.3 2716963-2718791)	21	1829	fully methylated
(NT_002835.3 2732562-2734239)	21	1678	fully methylated
(NT_002835.3 2735023-2736682)	21	1660	fully methylated
(NT_002835.3 2827157-2828726)	21	1570	unmethylated
(NT_002835.3 2841396-2842933)	21	1538	fully methylated
(NT_002835.3 2855101-2856552)	21	1452	fully methylated
(NT_002835.3 2861362-2862966)	21	1606	fully methylated
(NT_002835.3 2957438-2959877)	21	2440	unmethylated
(NT_002835.3 3014936-3016783)	21	1848	unmethylated
(NT_002835.3 3052859-3055543)	21	2263	unmethylated
(NT_002835.3 3132299-3134479)	21	2099	fully methylated
(NT_002835.3 3187879-3189962)	21	1902	unmethylated
(NT_002835.3 3364581-3366420)	21	1608	unmethylated

(NT_002835.3 3396443-3398312)	21	1561	unmethylated

Table 1. Nucleotide sequences of chromosome 21, their length and methylation status.

A total of 132 sequences were selected, out of which 29 are fully methylated and 103 are unmethylated.

Procedure- The sequences were acquired from NCBI (http://www.ncbi.nlm.nih.gov/) by using nucleotide database and all the sequences were downloaded in Fasta format. The CpG Island sequences were identified using CpG Island searcher (http://cpgislands.usc.edu/). The CpG Island searcher was designed originally by D. Takai and rewritten by S. Catherall. The CpG Island searcher provides a provision to select the lower limit of the parameters like %GC, ObsCpG/ExpCpG, length of CpG Island and Gap between adjacent Islands. The %GC was kept above 50%, ObsCpG/ExpCpG more than 0.6, minimum length of CpG Island 400 base pairs and minimum gap between adjacent islands 100 base pairs.

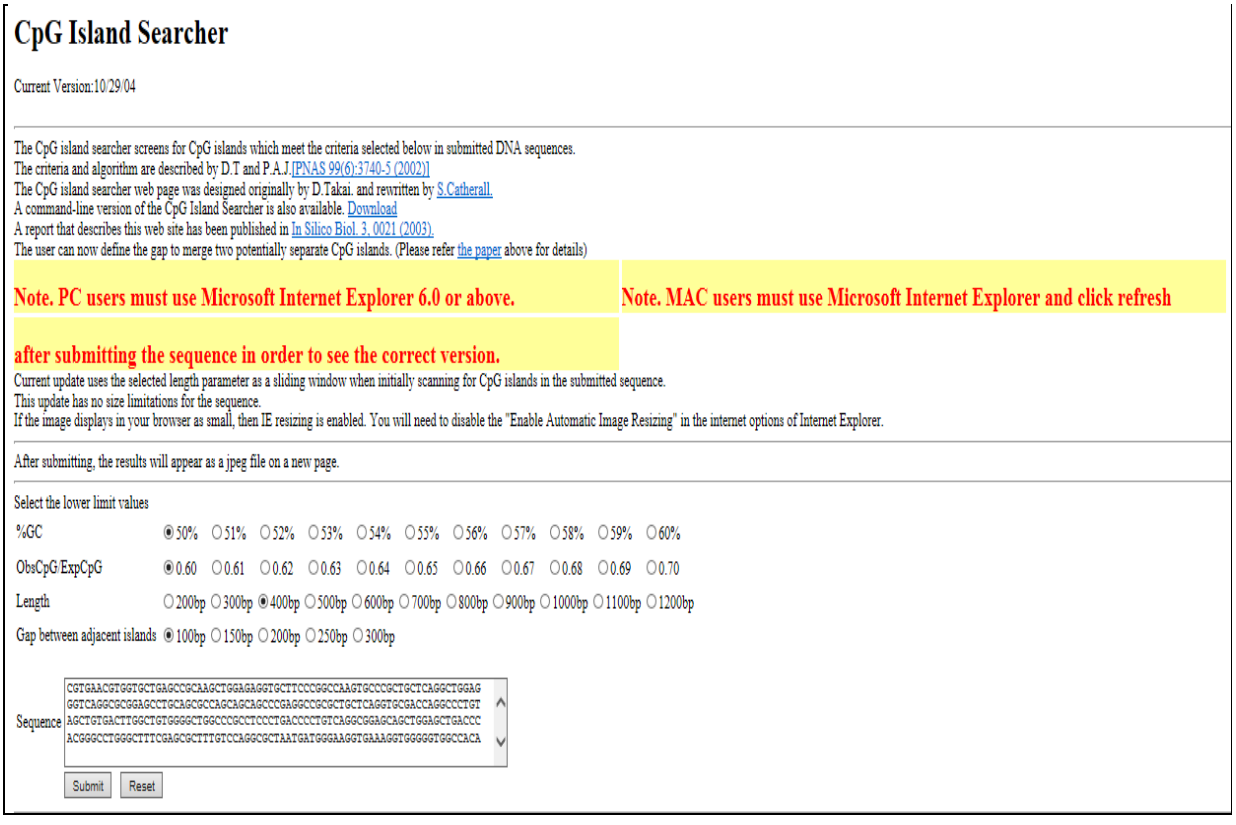


Fig. 4. CpG Island searcher: %GC ≥ 50%, ObsCpG/ExpCpG ≥ 0.60, Length ≥ 400bp, Gap between adjacent Islands ≥ 100bp.

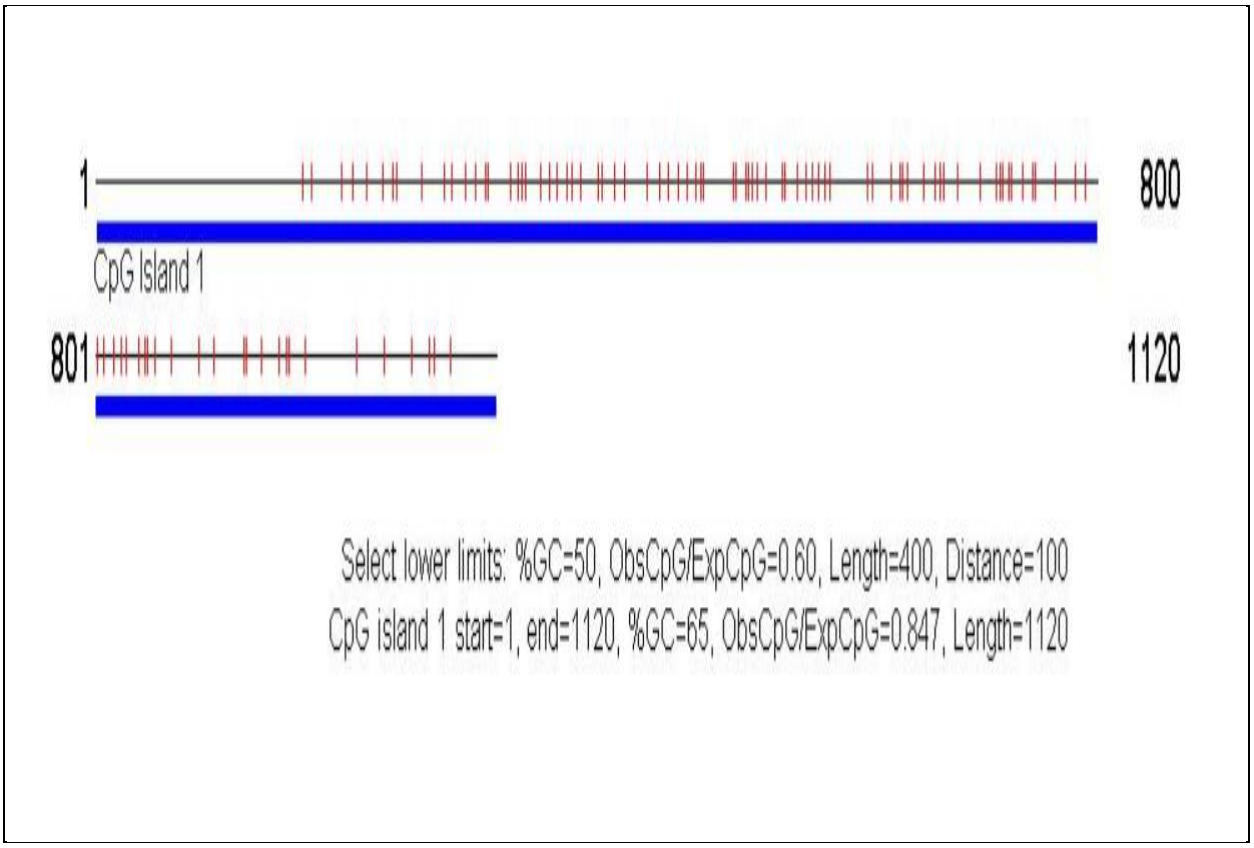


Fig. 5. CpG Island represented by bold blue lines.

Oligonucleotide frequency determination and its analysis

Web based database was used to find out the oligonucleotide frequencies *viz.*, FCGR (Chaos game representation of frequencies) (http://www.biophp.org/minitools/chaos_game_representation/demo.php?FCGR) designed by Deschavanne in 1991. FCGR represents frequencies of all the possible oligonucleotides. The frequencies were computed for upper strand only. All the possible permutations of mononucleotide, dinucleotide, trinucleotide and tetranucleotide were taken and the frequency of occurrence of each kind of permutation was calculated for both fully methylated and unmethylated sequences and the data was statistically analyzed to find some patterns that are repeated significantly more often in one group than other. The frequency of each permutation was added together and their ratio was calculated. The over-represented and under-represented permutations were selected as above 2×mean and below mean/2 respectively. Thus we got two groups of data for methylated and unmethylated sequences. The two were normalized and their

poisson counts were compared by using normal approximation of poisson distribution using method proposed by Best, 1975 (Best, 1957). The formula used for one-tailed testing is:

$$Z = \sqrt{2X_1 + \frac{3}{4}} - \sqrt{2X_2 + \frac{3}{4}}$$

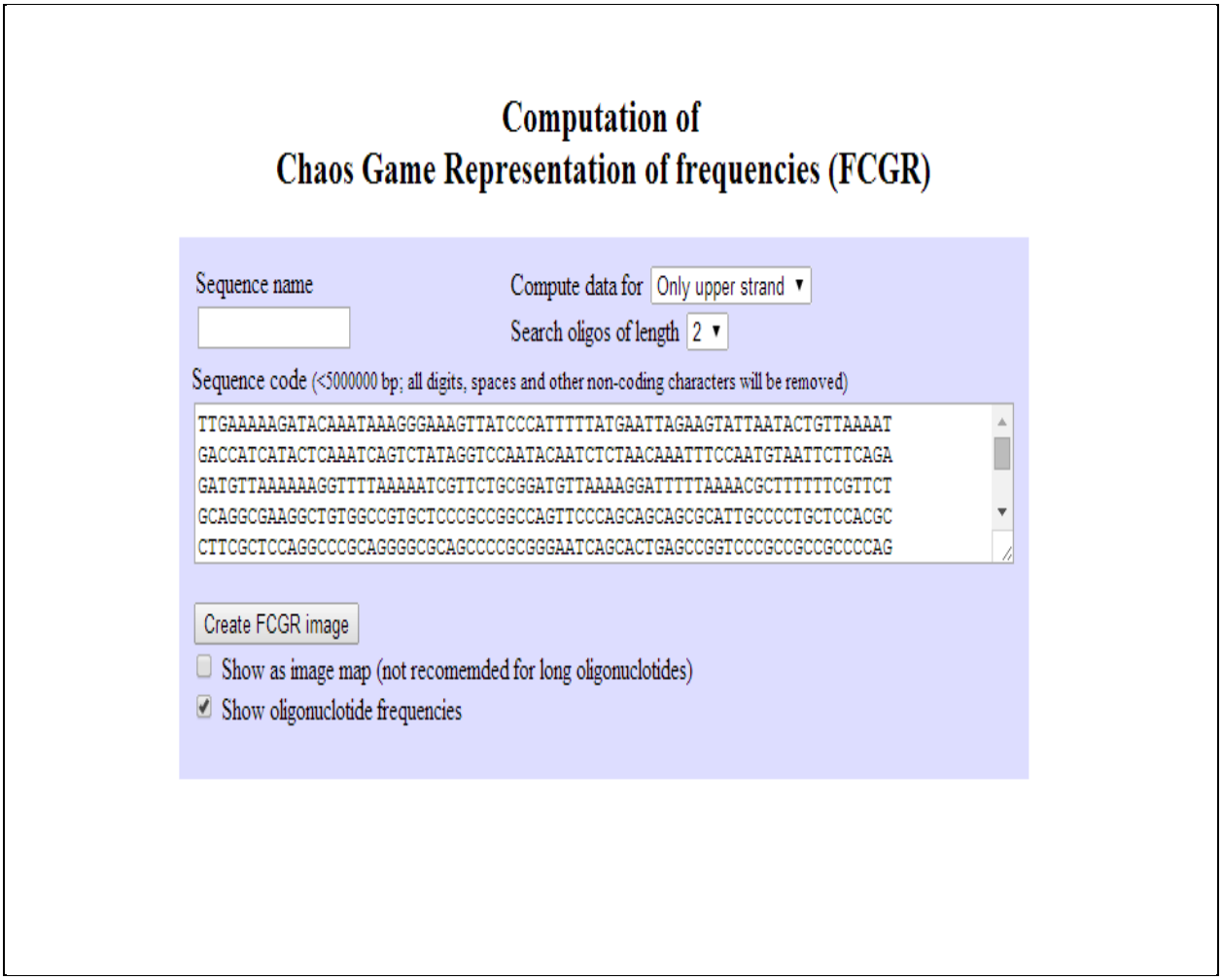


Fig. 6. FCGR data input screen.

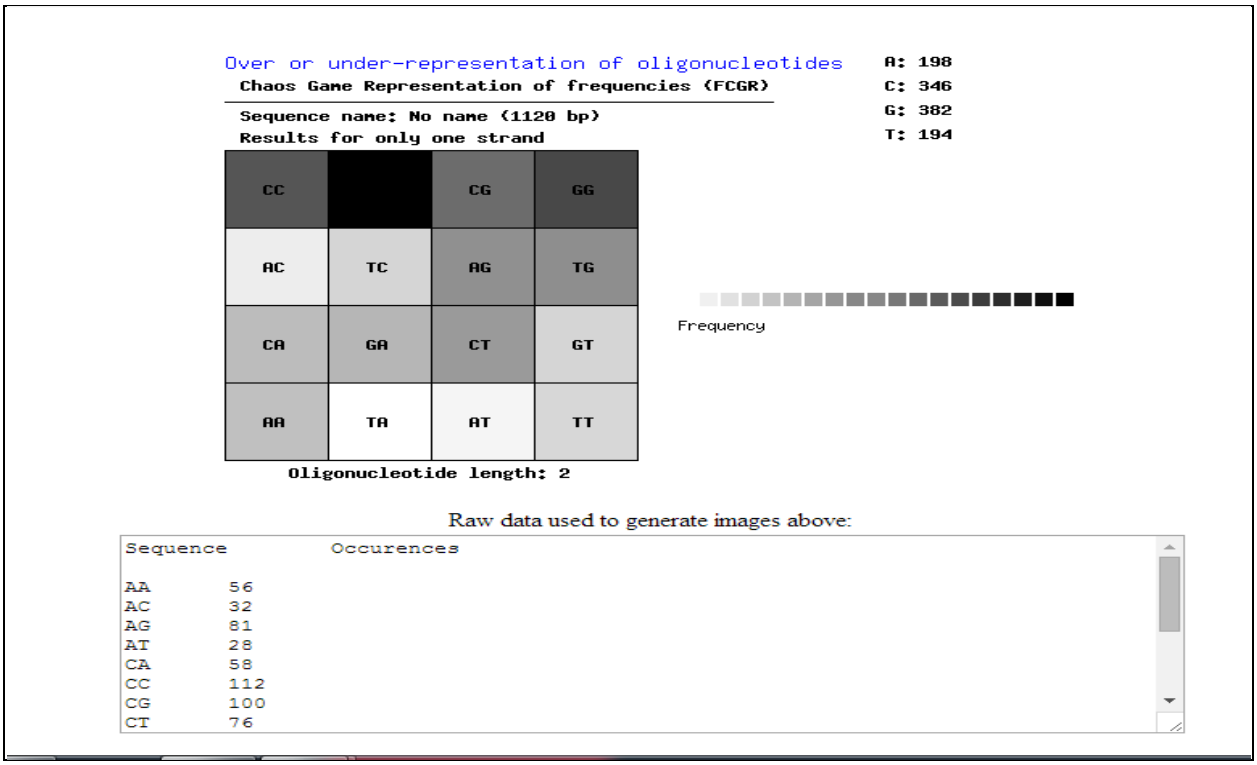


Fig. 7. Oligonucleotide frequencies representation by FCGR.

Gibb's free energy difference for DNA folding

The mfold web server (http://mfold.rna.albany.edu/?q=mfold) was used to find the ΔG folding of CpG Island sequences of both the classes. It is one of the oldest web servers in computational molecular biology introduced at Washington University's School of Medicine and created by Michael Zuker, professor of mathematical sciences in 2003. It helps in predicting the nucleic acid folding and hybridization. The parameters like ionic condition, folding temperature, percent sub optimality, upper bound, window and maximum distance between the paired bases can be assigned as per requirement. The ionic condition was maintained at: $[Na^{+}]$ = 12 mM and $[Mg^{++}]$ = 0.5 mM, which corresponds to normal physiological concentration of mammalian cells. The rest of the conditions were kept as default. Gibb's free energy per unit length was determined and analyzed by using Student's t-test.

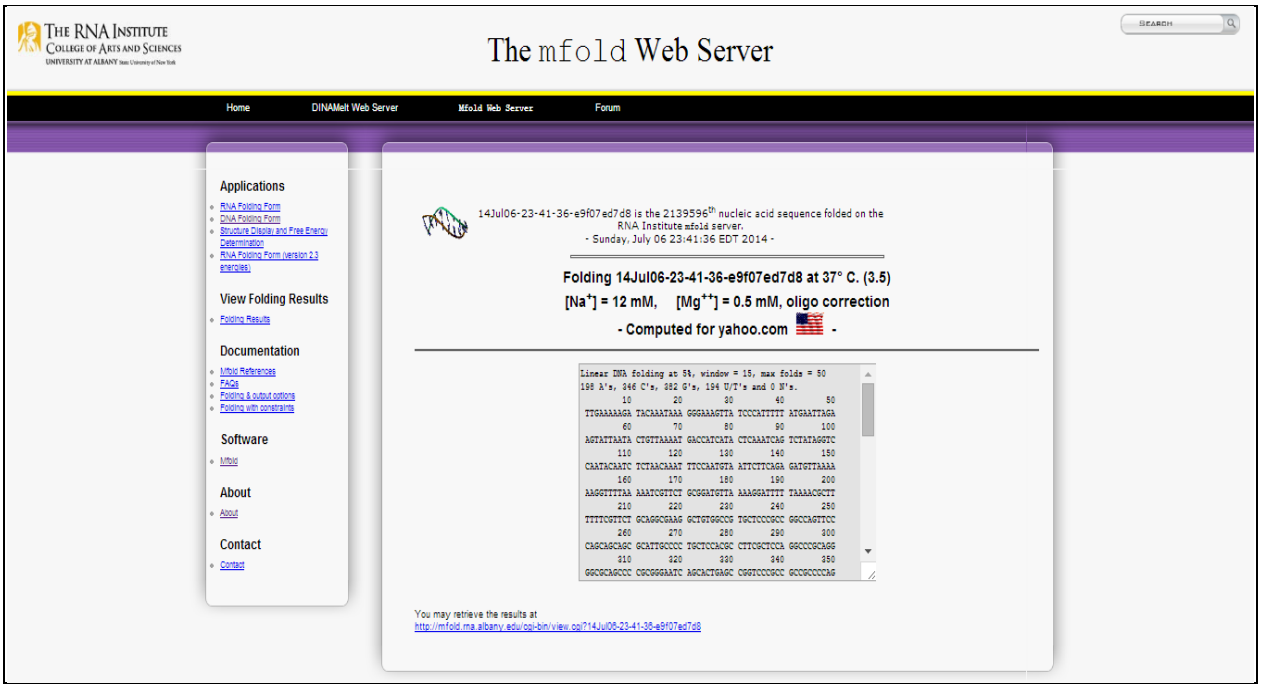


Fig. 8. Output file of The mfold web server.

Inverted repeat frequency

The inverted repeats were determined by using EMBOSS: palindrome from EMBOSS (European Molecular Biology Open Software Suite) explorer (http://emboss.bioinformatics.nl/CpGi-bin/emboss/palindrome), again a web based database. It helps in finding the inverted repeats (stem loops) in a nucleotide sequence that include a proportion of mismatches and gaps by considering the specified conditions of minimum and maximum length of palindrome, maximum gap between repeated regions and number of mismatches allowed. It was designed by Mark Faller in 1999. The minimum length of palindrome was kept at 4 base pairs and maximum at 6 base pairs with 0 and 1 mismatches allowed. Three types of Inverted repeats were considered *viz.*, 8 and 9 base pair long, 10 and 11 base pair long, 12 and 13 base pair long. The frequency of occurrence of each kind of Inverted repeats was calculated and analyzed by comparing their Poisson count as proposed by Best in 1975.

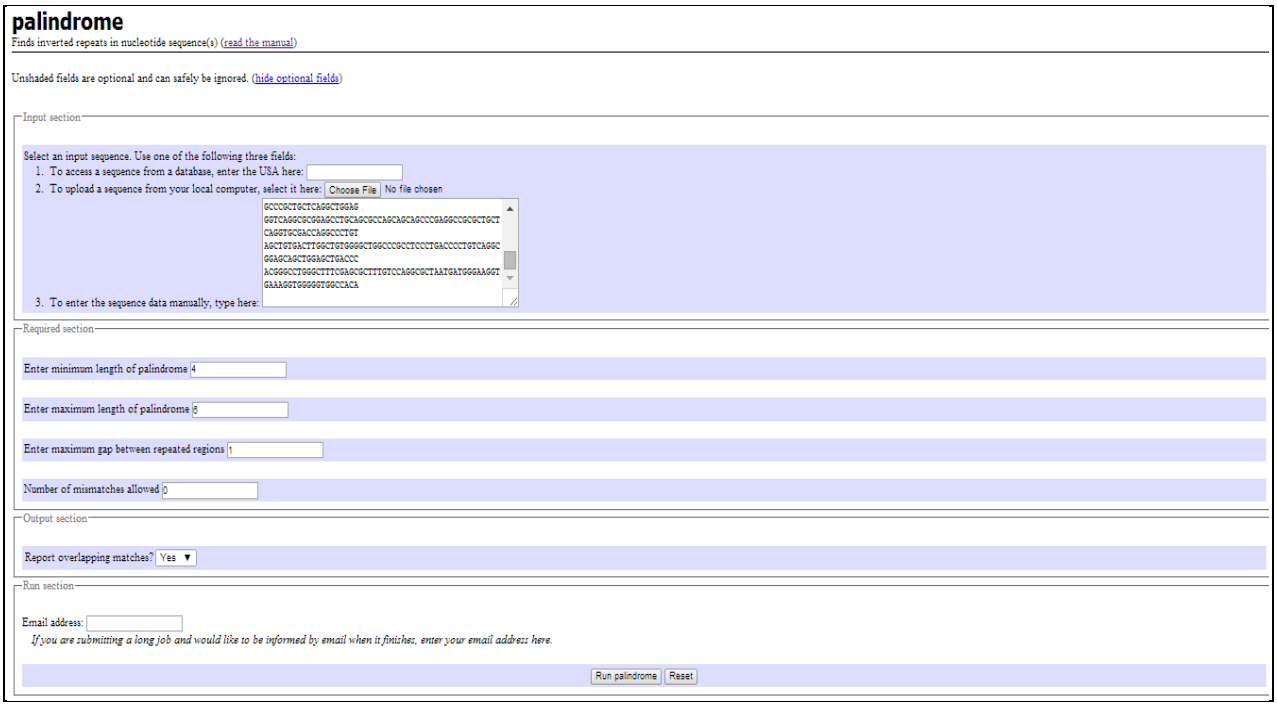


Fig. 9. Homepage of EMBOSS: palindrome database.

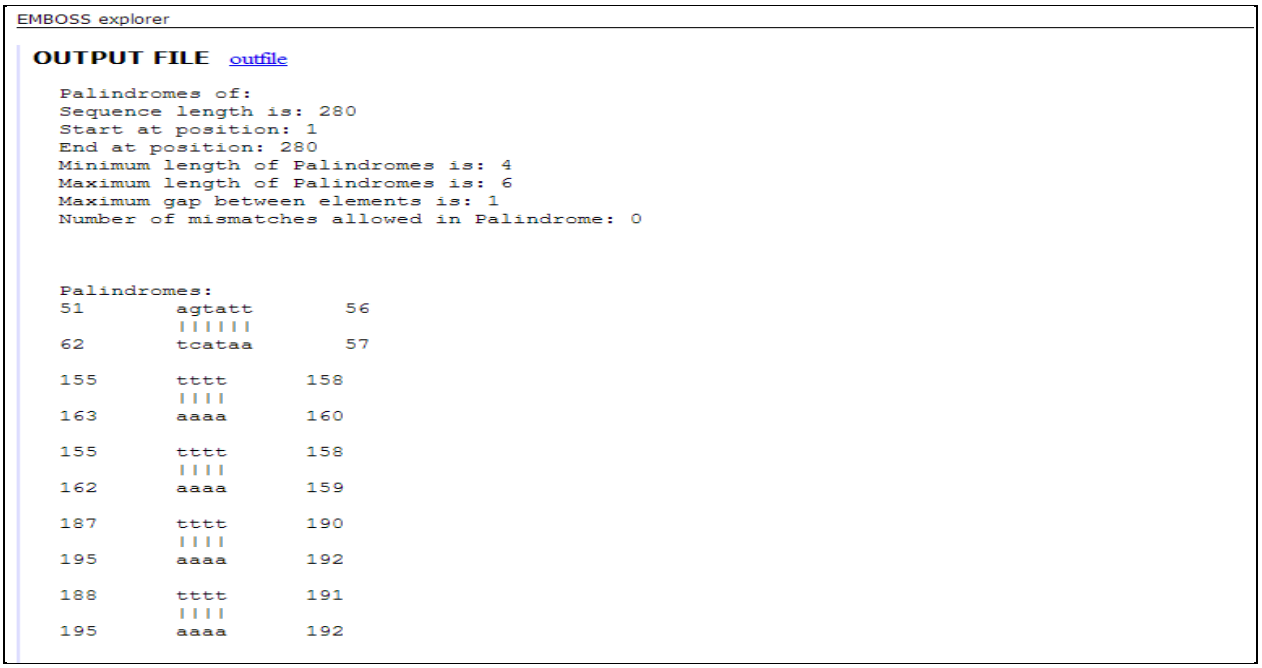


Fig. 10. Output file of EMBOSS:palindrome.

CpG gap length determination

DEV C^{++} computer program was used to find the positions of CpG dinucleotide in the CpG Islands. It is a C^{++} based program designed by one of the coworker Japnjyot Saini and is

unpublished. The gap lengths were calculated and to compare the magnitudes of CpG gaps in the two kinds of CpG Islands, the datasets were subjected to one tailed Student's t-test. The distribution of CpGs was studied by comparing their dispersion indices.

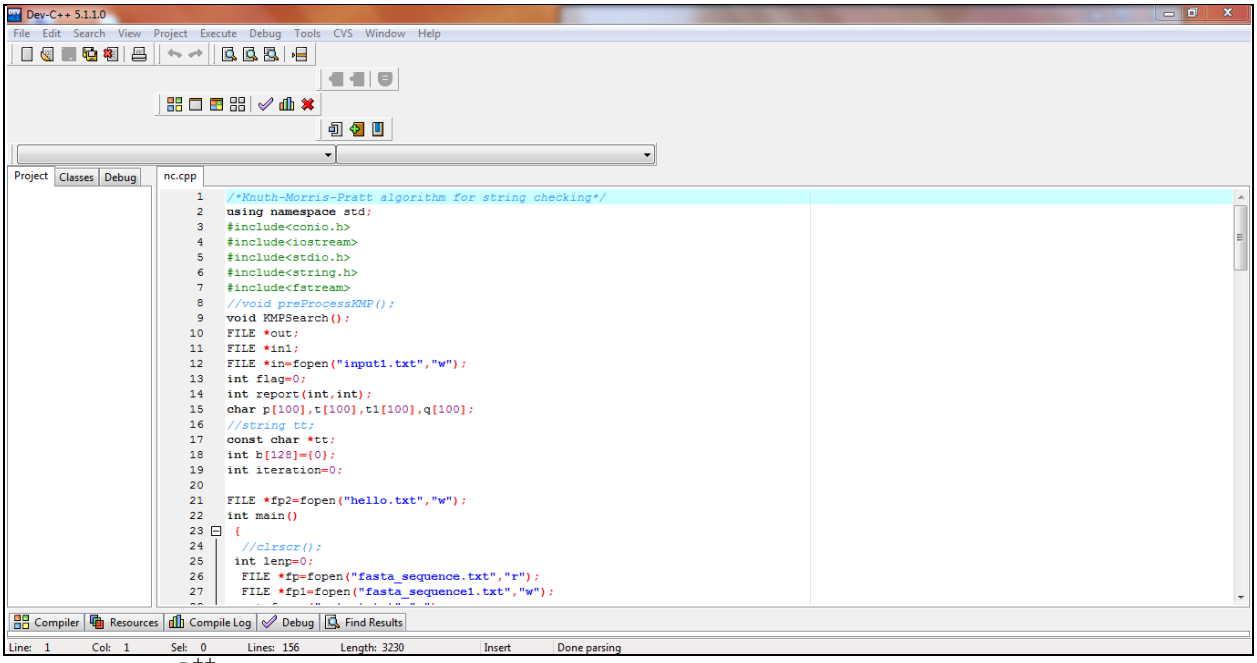


Fig. 11. Dev C^{++} computer program.

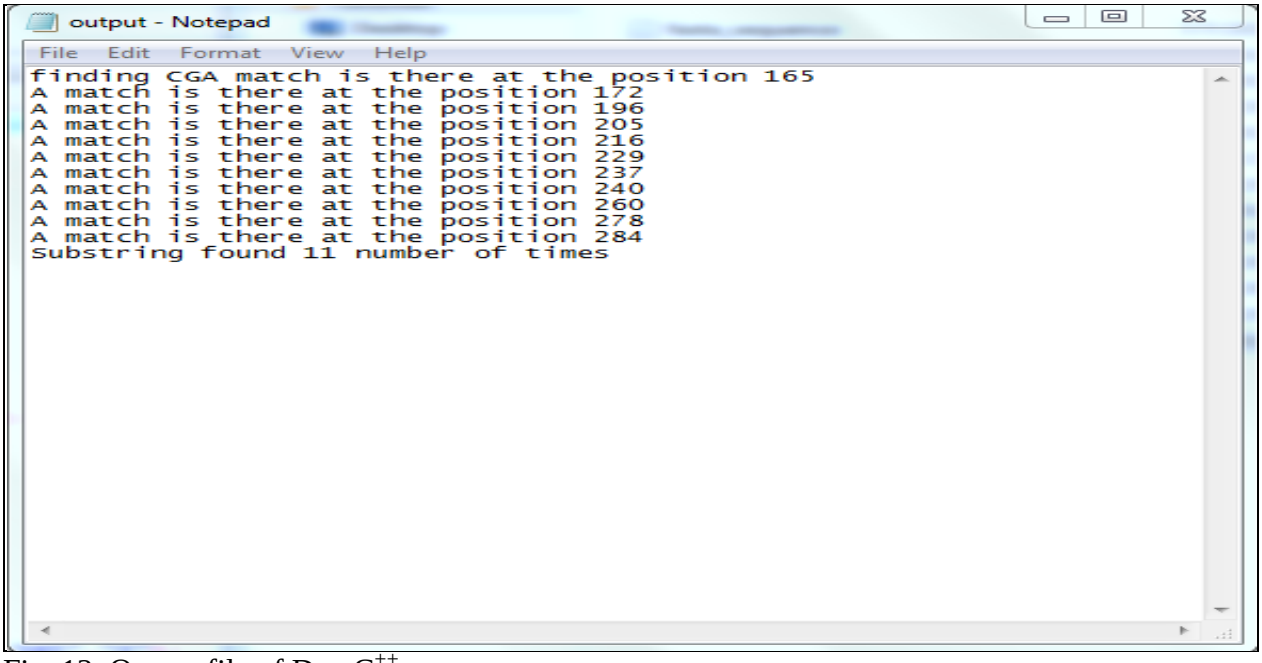


Fig. 12. Output file of Dev C^{++}.

The statistical tools used to analyze the raw data, includes Standard deviation, Variance, Arithmetic mean, Variance mean ratio, Student's t-test, Poisson distribution and Normal distribution.

CHAPTER 5

Result

Results

Mammalian DNA shows enormous diversity in its genome structure and organization. There are certain regions in genome that show a high frequency for the dinucleotide CpG, called CpG Islands. These regions are further assorted into methylated CpG Islands and non methylated CpG Islands. The methylated CpG Islands are found to occur with fewer propensities than non methylated ones and the methylation of CpG Islands plays an important role in epigenetic control of genes. Here we attempted to find out why certain CpG Islands are an easy target for methylation. For this, two groups of CpG Islands are studied *viz.*, fully methylated and unmethylated. The dataset for the study was derived from the work of Yamada *et al.* in 2004. They developed HpaII–McrBC PCR method for discriminating full, null, incomplete, and composite methylation patterns. They applied it to all computationally identified CpG Islands on human Chromosome 21q. It was found that most CpG Islands (103 out of 149) escape methylation, but a sizable fraction (31 out of 149) are fully methylated and seven CpG Islands showed composite methylation. The fully methylated and unmethylated data sets are considered here and were analyzed for oligonucleotide frequency, ΔG folding, inverted repeats and CpG distribution.

Oligonucleotide frequency

It has been found that in genome the regions rich in CpG are found to be less methylated than rest of genome and propensity of CpG rich DNA to get methylated had been explained on basis of sequence patterns. Here is an attempt to expand this idea by using all the possible permutations of mononucleotide, dinucleotide, trinucleotide and tetranucleotide. Certain novel sequence patterns were exposed that were found to diverge significantly in the two kinds of sequences. The sequence elements corresponding to the ratio with yellow text color were more in unmethylated and those corresponding to blue text color were more in fully methylated sequences as shown in Table 2.

Ratio of total length of M and U	Total Length		Normalization factor
0.227	39344	173382	4.407

Motif	∑M/∑U	∑M	∑U	M normalized	U Normalized	Z'	Z	p - value
CCGC	0.080	200	2512	881.4	570.0	28.89	-13.76	0
CGGA	0.092	79	856	348.1	194.2	14.98	-7.13	5.04E-13
CGCG	0.105	218	2070	960.7	469.7	20.51	-9.76	0
CTTT	0.113	72	635	317.3	144.1	10.44	-4.97	3.4E-07
GCCG	0.115	239	2079	1053.2	471.8	18.58	-8.85	0
CGCC	0.118	281	2373	1238.3	538.5	19.12	-9.11	0
TACT	0.120	26	216	114.6	49.0	5.64	-2.68	0.003727
AAAG	0.121	70	579	308.5	131.4	9.19	-4.37	6.22E-06
CCCC	0.126	309	2461	1361.7	558.5	17.97	-8.56	0
CCGG	0.127	260	2044	1145.8	463.8	16.07	-7.65	1.01E-14
TTCT	0.130	74	571	326.1	129.6	8.25	-3.92	4.34E-05
AGAT	0.532	149	280	656.6	63.5	-12.57	5.98	1.13E-09
CGTA	0.533	65	122	286.4	27.7	-8.31	3.94	4.02E-05
ACGT	0.542	143	264	630.2	59.9	-12.52	5.95	1.31E-09
CGTT	0.547	158	289	696.3	65.6	-13.27	6.31	1.38E-10
TTAC	0.562	123	219	542.0	49.7	-11.99	5.70	5.95E-09
CATC	0.585	196	335	863.7	76.0	-15.67	7.46	4.42E-14
ATAT	0.590	59	100	260.0	22.7	-8.65	4.11	2.02E-05
TATA	0.630	58	92	255.6	20.9	-9.03	4.29	9.11E-06
CACA	0.633	341	539	1502.7	122.3	-21.98	10.47	0
CACG	0.644	356	553	1568.8	125.5	-22.75	10.83	0
ACAC	0.728	316	434	1392.6	98.5	-23.31	11.09	0
TGTG	0.741	450	607	1983.1	137.7	-28.13	13.39	0
GTGT	0.894	473	529	2084.4	120.0	-32.03	15.25	0

Table 2. Z test of oligonucleotide frequencies. ∑M/∑U is ratio of sum of methylated and unmethylated frequency corresponding to its sequence element, sum M and sum U represents sum of frequencies for methylated and unmethylated sequence corresponding to its sequence element, , Z and Z' is the comparison of two poisson counts by using normal approximation of poisson distribution using method proposed by Best, 1975.

$\Delta G_{folding}$

The Gibb's free energy of folding for each sequence is determined by using the mfold web server. Gibb's free energy per unit length was calculated by dividing the computed energy value by length of the sequence analyzed and finally Student's t-test was applied. The p value was found to be 0.0059. The difference in free energy of the two kinds of sequences ascertained that the two were significantly different with respect to $\Delta G_{folding}$.

Sequence (M)	KCAL/MOL (E)	Length (L)	E/L	Sequence (U)	KCAL/MOL (E)	Length (L)	E/L	P(E/L)
M1	-69.18	1149	-0.0602	U1	-114.65	1826	-0.0628	0.0059
M2	-54.91	1112	-0.0494	U2	-100.51	1633	-0.0615	
M3	-53.61	1442	-0.0372	U3	-88.25	1952	-0.0452	
M4	-24.88	910	-0.0273	U4	-101.65	1566	-0.0649	
M5	-64.43	1314	-0.0490	U5	-71.73	1271	-0.0564	
M6	-19.01	930	-0.0204	U6	-37.39	1375	-0.0272	
M7	-96.23	3585	-0.0268	U7	-41.03	2195	-0.0187	
M8	-67.67	1171	-0.0578	U8	-90.61	1855	-0.0488	
M9	-75.01	1778	-0.0422	U9	-68.62	1531	-0.0448	
M10	-17.64	870	-0.0203	U10	-68.64	1200	-0.0572	
M11	-43.39	1335	-0.0325	U11	-34.58	1200	-0.0288	
M12	-40.53	924	-0.0439	U12	-30.99	854	-0.0363	
M13	-30.34	1110	-0.0273	U13	-106.45	1949	-0.0546	
M14	-27.82	838	-0.0332	U14	-63.98	1452	-0.0441	
M15	-42.04	1172	-0.0359	U15	-113.86	2049	-0.0556	
M16	-40.62	1355	-0.0300	U16	-132.25	2094	-0.0632	
M17	-38.13	1376	-0.0277	U17	-50.83	1342	-0.0379	
M18	-36.81	1524	-0.0242	U18	-111.26	1949	-0.0571	
M19	-106.17	2218	-0.0479	U19	-43.67	1403	-0.0311	
M20	-65.99	1764	-0.0374	U20	-79.19	1584	-0.0500	
M21	-50.08	1592	-0.0315	U21	-76.99	1522	-0.0506	
M22	-52.02	1694	-0.0307	U22	-52.09	1245	-0.0418	
M23	-22.2	1092	-0.0203	U23	-92.59	2011	-0.0460	
M24	-51.77	1257	-0.0412	U24	-122.44	1929	-0.0635	
M25	-62.58	1023	-0.0612	U25	-65.75	1454	-0.0452	
M26	-40.29	1070	-0.0377	U26	-50.72	1599	-0.0317	
M27	-51.26	1114	-0.0460	U27	-87.42	1424	-0.0614	
M28	-50.36	1193	-0.0422	U28	-85.91	1900	-0.0452	
M29	-66.99	1382	-0.0485	U29	-79.51	1489	-0.0534	
				U30	-94.51	1836	-0.0515	
Sum	-1461.96	39294	-1.089769808	U31	-135.52	1985	-0.0683	
Average	-50.41241379	1354.965517	-0.037578269	U32	-51.47	1189	-0.0433	
				U33	-96.61	1724	-0.0560	
				U34	-130.24	2183	-0.0597	
				U35	-29.54	1148	-0.0257	
				U36	-47.96	1853	-0.0259	
				U37	-48.62	1060	-0.0459	
				U38	-64.94	1397	-0.0465	
				U39	-99.37	2023	-0.0491	
				U40	-60.24	1523	-0.0396	
				U41	-73.17	1361	-0.0538	
				U42	-749.38	2224	-0.3370	
				U43	-51.75	1186	-0.0436	
				U44	-45.87	1181	-0.0388	
				U45	-98.84	1581	-0.0625	
				U46	-95.78	1480	-0.0647	
				U47	-67.8	1420	-0.0477	
				U48	-85.03	1596	-0.0533	
				U49	-94.08	1863	-0.0505	
				U50	-219.81	2913	-0.0755	
				U51	-53.34	1091	-0.0489	
				U52	-97.88	1629	-0.0601	
				U53	-74.19	1778	-0.0417	
				U54	-55.93	1719	-0.0325	
				U55	-129.82	2572	-0.0505	
				U56	-103.74	1719	-0.0603	
				U57	-45.79	1164	-0.0393	
				U58	-50.18	1248	-0.0402	
				U59	-48.84	1283	-0.0381	

U60	-112.47	2220	-0.0507
U61	-110.6	1972	-0.0561
U62	-122.42	2015	-0.0608
U63	-155.48	2667	-0.0583
U64	-70.26	1287	-0.0546
U65	-55.5	1128	-0.0492
U66	-48.18	1377	-0.0350
U67	-121.82	1855	-0.0657
U68	-92.7	1443	-0.0642
U69	-47.3	1379	-0.0343
U70	-149.43	2175	-0.0687
U71	-67.96	1638	-0.0415
U72	-167.49	2724	-0.0615
U73	-84.41	1692	-0.0499
U74	-89.88	1823	-0.0493
U75	-79.25	1559	-0.0508
U76	-168.85	2523	-0.0669
U77	-115.62	1944	-0.0595
U78	-76.77	1537	-0.0499
U79	-57.44	1499	-0.0383
U80	-101.13	1524	-0.0664
U81	-87.01	1374	-0.0633
U82	-107.74	1561	-0.0690
U83	-73.57	1543	-0.0477
U84	-92.38	1554	-0.0594
U85	-70.75	1196	-0.0592
U86	-32.87	868	-0.0379
U87	-60.22	1370	-0.0440
U88	-168.28	2286	-0.0736
U89	-118.93	2254	-0.0528
U90	-69.24	1150	-0.0602
U91	-66.55	1595	-0.0417
U92	-110.52	1842	-0.0600
U93	-107.28	1286	-0.0834
U94	-69.2	1655	-0.0418
U95	-153.84	2198	-0.0700
U96	-193.07	2891	-0.0668
U97	-108.02	1351	-0.0800
U98	-96.12	2015	-0.0477
U99	-59.63	1459	-0.0409
U100	-134.88	2627	-0.0513
U101	-90.95	1735	-0.0524
U102	-60.06	1487	-0.0404
U103	-64.19	1512	-0.0425

Sum	-9654.06	172572	-5.55300985
Average	-93.72873786	1675.456311	-0.053912717

Table 3. Statistical analysis of $\Delta G_{folding}$. Energy in kCal/mol**,** length in basepairs, E/L represents energy per unit length, P(E/L) represents p value of Student's t test.

Inverted repeats

In continuity with Gibb's free energy, another related feature was examined *i.e.*, inverted repeat frequency. Inverted repeats are sequences that consist of a base sequence followed by its reverse complement about the axis of symmetry and thus also aids in formation of DNA structures such

as cruciform and hairpins. The sequences with higher occurrence of inverted repeats have been found to be significantly more in unmethylated. Further, as we go for longer inverted repeats their abundance increases in methylated ones.

		M	U	∑M	∑U	M normalized	U normalized	Z'	Z	*p* - value
	FOUR	278	1652			1632.34	281.35	0.343	-0.141	0.444
	FIVE	71	459			416.89	78.17	1.423	-0.586	0.279
IR	SIX	33	132			193.77	22.48	-3.434	1.409	0.079
		23	198	108	873	853.06	110.52	0.480	-0.170	0.432
		72	526							
		2	110							
	FOUR	11	39							
		16	58	35	289	276.46	36.59	0.527	-0.187	0.426
		18	153							
		1	43							
	FIVE	0	35							
		2	15	15	86	118.48	10.89	-2.274	0.799	0.212
		5	47							
CGs in		1	9							
IR	SIX	7	15							

Table 4. Inverted repeat data analysis. M represents total number of inverted repeats of length four, five and six respectively and total number of CpGs in inverted repeats in fully methylated sequences, U represents total number of inverted repeats of length four, five and six respectively and total number of CpGs in inverted repeats in unmethylated sequences, ∑M is sum of all CpGs in inverted repeats of length four, five and six in fully methylated sequences, ∑U is sum of all CpGs in inverted repeats of length four, five and six in unmethylated sequences, Z and Z' is the comparison of two poisson counts by using normal approximation of poisson distribution using method proposed by Best, 1975.

CpG distribution

It has already been shown that CpG prevalence in sequences adversely affects their methylation. In order to investigate if distribution of CpGs such as their even, random or clustered distribution is having any bearing on methylation, we determined the length of sequence between each adjacent pair of CpGs and named it as CpG gap. The sequences were analyzed using a C^{++} based

programme which computed positions of CpGs in a given sequence from which the CpG gaps were calculated. Their statistical analysis revealed a significant difference with a p-value of 5.5 x 10^{-15}. The result clearly shows that the mean CpG gap in methylated CpG Islands is greater than unmethylated ones. The distribution of CpGs in the two classes of CpG Islands was studied by calculating dispersion index. Dispersion index is the ratio of variance to arithmetic mean of a given data. If the value is <1 it is considered to be under dispersed, =1 is random distribution and > 1 is over-dispersed or clustered data. The dispersion index value of methylated CpG Islands is less than unmethylated ones which shows that CpG are less clustered in methylated CpG Islands.

Mean (Gap Length)	
M	**U**
12.60715514	10.61873876
Standard deviation	
M	**U**
13.28587233	12.62147106
Variance/Mean (VMR)	
M	**U**
14.00112885	15.00192587
VMR/Mean	
M	**U**
1.110570045	1.412778505
T-test	
5.58165E-15	

Table 5. CpG distribution analysis. Mean, standard deviation, variance by mean, vmr by mean and T- test of gap lengths of fully methylated and unmethylated sequences.

CHAPTER 6

Discussion

Discussion

The purpose of the present work is to improve our understanding of DNA methylation. To be more precise our work is an attempt to find out why certain CpG Islands are more vulnerable to methylation than others. We have attempted to explore some novel DNA related attributes and relate them with CpG Island methylation. The attributes that are considered here are Oligonucleotide frequency, $\Delta G_{folding}$, Inverted repeats and CpG gaps.

The frequencies of oligonucleotides were checked in both the kinds of sequences. Among the four oligonucleotides *i.e.,* mononucleotide, dinucleotide, trinucleotide and tetranucleotide, only tetranucleotide accounted for the disparity in the two kinds of sequences. Certain tetranucleotide elements were found to be significantly different with respect to their occurrence in fully methylated and unmethylated sequence sets. The sequence elements corresponding to the ratio with yellow text color were more in unmethylated and those corresponding to blue text color were more in fully methylated. It depicts some biasness in the occurrence of these oligonucleotide elements and can be presumed to effect the CpG Island methylation. It is proposed that higher of lower abundance of certain tetranucleotide permutations of bases could affect the methylation of CpGs in their vicinity by altering the structure of DNA or directly affecting the catalytic efficiency of DNA methyltransferases.

DNA methylation is an enzyme mediated process and is largely influenced by the structural orientation of DNA sequence. The data from Gibb's free energy of folding and inverted repeats can helps in resolving this speculation. The $\Delta G_{folding}$ in the two groups were largely different in two groups. A p-value of 0.0059 illustrates about 95% dissimilarity. Thus it can't be considered a matter of chance. There must be some vital influence of $\Delta G_{folding}$ on CpG Island methylation. It may be inferred that certain subtle changes in DNA structure affecting its methyation by DNA methyltransferases is manifested by $\Delta G_{folding}$.

The analysis of inverted repeats frequency in the two classes of CpG Islands is serving the similar purpose. It mediates the formation of certain cruciform structures in DNA. These structures assist in DNA folding and thus can influence the enzymology of DNA methylation. The inverted repeats were present more in unmethylated CpG Islands and less in methylated

ones. The effect of inverted repeats was found to be more significant when only those inverted repeats were studied that contained at least one CpG dinucleotide in their sequence.

It has been found that the CpG gaps are significantly larger in methylated CpG Islands than unmethylated one. It is found to be in agreement with known fact that higher Obs/Exp ratio of CpGs leads to lower methylation (Saxonov *et al.*, 2006). The mean CpG gap in methylated CpG Islands is found to be greater than unmethylated ones with a p-value of 5.5×10^{-15}. Thus it can be concluded that the length of CpG gaps have an influential impact on methylation of CpG Islands. The results indicate that it is not merely CpG density expressed as $CpG_{Obs/Exp}$ ratio that influences DNA methylation but distribution of CpGs in a sequence also contributes to it. The VMR for unmethylated CpG Islands is found to be greater than methylated ones. It shows that higher clustering of CpG dinucleotides in CpG Islands resists methylation.

In the present study four different sequence attributes have been found that exhibit difference in methylated and unmethylated CpG Islands. Varying abundance of certain tetranucleotide sequences indicates their influence on local DNA structure or even direct interaction with DNA methyltransferase enzymes during their sliding mode of action. On the other hand frequency of inverted repeats and $\Delta G_{folding}$ per unit length of sequence manifest putative effect of DNA structure on DNA methylation. Investigation based on CpG gap analysis adds further dimension of distribution of CpG to the already known factor of CpG density influencing methytlation of DNA. These results further improves our understanding of DNA methylation especially somewhat odd phenomenon of methylation of CpG Islands. The interpretation is based on statistical analysis but on a sufficiently large sample size. Further work at biochemical level may help to understand the involved mechanism.

CHAPTER 7

References

References

[1]. Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev* **16**: 6–21.

[2]. Gardiner-Garden, M., Frommer, M. (1987). CpG Islands in vertebrate genomes. *J Mol Biol* **196**: 261–282.

[3]. Yamada, Y., Watanabe, H., Miura, F., Soejima, H., Uchiyama, M, *et al.* (2004). A comprehensive analysis of allelic methylation status of CpG Islands on human Chromosome 21q. *Genome Res* **14**: 247–266.

[4]. Handa, V., Jeltsch, A,. (2005). Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome. *J Mol Biol* **348**: 1103–1112.

[5]. Feltus, F. A., Lee, E. K., Costello, J. F., Plass, C., Vertino, P. M. (2003). Predicting aberrant CpG Island methylation. *Proc Natl Acad Sci U S A* **100**: 12253–12258.

[6]. Bock, C., Paulsen, M., Tireling, S., Mikeska, T., Lengauer, T., Walter, J. (2006). CpG Island methylation in human lymphocytesis highly correlated with DNA sequence, repeats, and predicted DNA structure. *Plos genetics* **2**: e26.

[7]. Salser, W. (1977). Cold Spring Harbour Synp. Quant. Biol. XLII, 98-1103.

[8]. SantaLucia, Jr. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA* **95**: 1460-1465.

[9]. Simmons, D. (2008). Epigenetic Influences and Disease. *Nature Education.* 1(1):6.

[10]. Hotchkiss, R. D. (1948). The quantitative separation of purines, pyrimidines and nucleosides by paper chromatography. *J. Biol. Chem*. **175**: 315–332.

[11]. Das, P. M., Singal, R. (2004). DNA Methylation and Cancer. *J Clin Oncol* **22**:4632-4642.

[12]. Jeltsch, A. (2002). Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases. *Chembiochem*. **3**: 274–293.

[13]. Roberts, R. J., Belfort, M., Bestor, T., Bhagwat, A. S., Bickle, T. A., Bitinaite, J. *et al.* (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* **31**: 1805–1812.

[14]. Herman, J.G., *et al*. (1994). Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma, *Proc. Natl. Acad. Sci. USA* **91**: 9700–9714.

[15]. Tatematsu, K. I., Yamazaki, T. and 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 Cells* **5**: 677–688.

[16]. Bird, A. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research.* **8**: 1499-1504.

[17]. Takai, D., Jones, P. A. (2001). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *PNAS.* **99**: 3740–3745.

[18]. 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**: 572–579.

[19]. Josse, J., Kaiser, A.A. and Kacnberg, A. (1961). *J. Biol. Chen*. **236**: 864-875.

[20]. Swartz, M.N., Trautner, T.A. and Kornberg, A. (1962). *J. Biol. Chem.* 237: 1961-1967.

[21]. Hermanna, A., Gowhera H., Jeltscha, A. (2004). Biochemistry and biology of mammalian DNA methyltransferases. *Cell. Mol. Life Sci*. **61**: 2571–2587.

[22]. Craig, J. M., Bickmore, W. A. (1994). The distribution of CpG Islands in mammalian chromosomes. *Nature genet.* **7**: 376-381.

[23]. Wu, J. C. and Santi, D. V. (1985). On the mechanism and inhibition of DNA cytosine methyltransferases. *Prog. Clin. Biol. Res.* **198**: 119–129.

[24]. Saxonov, S., Berg, P., Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *PNAS.* **103**: 1412-1417.

[25]. Matsuo, K., Clay, O., Takahashi, T., Silke, J., Schaffner, W. (1993). Evidence for erosion of mouse CpG islands during mammalian evolution. *Somatic Cell and Molecular Genetics.* **19**: 543-555.

[26]. Best, D. J. 1975. The difference between two Poission expectations. *Austral. J. Statist.* **17:** 29-33.