Studies on the effect of flanking bases on the distribution of CGs in methylated genomes

A dissertation report

submitted in partial fulfilment of the requirement for

the award of degree of

Master of Technology
In
Biotechnology

Under the guidance of
Dr.Vikas Handa
Assistant Professor

Submitted by
Japnjot Kaur
Roll No. 601204009

DEPARTMENT OF BIOTECHNOLOGY
THAPAR UNIVERSITY
PATIALA-147004
July 2014

CANDIDATE'S DECLARATION

I hereby declare that the work being presented in the M.Tech dissertation entitled "**Studies on the effect of flanking bases on the distribution of CGs in the methylated genomes**" in partial fulfilment of the requirement for the award of Degree of Masters in Technology in Biotechnology to Thapar University, Patiala is my own work during the period of July 2013 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: 18.07.2014

Japnjot Kaur
Roll No. 601204009

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

Date: 18.07.2014

Dr. Vikas Handa
Assistant Professor
Department of Biotechnology

CERTIFICATE

This is to certify that the work reported in M.Tech dissertation entitled **"Studies on the effect of flanking bases on the distribution of CGs in the methylated genomes"** submitted by Japnjot Kaur in partial fulfilment of the requirement for the award of Degree of Masters in Technology 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

ACKNOWLEDGEMENT

I would never have been able to finish my dissertation without the guidance of my advisor, help from friends, and support from my family.

I express my deepest gratitude towards my guide ***Dr. Vikas Handa****, Assistant Professor, Department of Biotechnology, Thapar University, Patiala for his valuable guidance, support and constant encouragement. He has been very kind and patient while correcting my mistakes and clearing my doubts throughout the project. Blessings, help and guidance given by him from time to time shall carry me a long way in the journey of life on which I am about to embark.*

I would also like to extend my thankfulness towards ***Dr. Dinesh Goyal****, Head of Department and* ***Dr. Niranjan Das****, P.G coordinator Thapar University, for their support, kind cooperation and encouragement. I am really pleased to acknowledge the kind help, cooperation and moral support which i have received throughout my dissertation from of all the teaching as well as non teaching faculty members of Department of Biotechnology, which helped me a lot in completion of this work.*

I am really thankful to my friends Amanjyoti and Vipin for their kind help and support. I would like to express my utmost gratitude to my parents, for their unconditional affection and support and towards my dear friends for giving me support, friendly environment and unforgettable moments in the Thapar University.

At last, I would like to thanks my Almighty God for his constant blessings without which any task would be impossible.

Date: 18.07.2014
Place: Patiala

Japnjot Saini
Japnjot Kaur
(601204009)

TABLE OF CONTENTS

ABBREVATIONS

A	Adenine
B	Not A (G, T and C)
C	Cytosine
C^5	Carbon at 5^{th} position
CpA	Phosphodiester bond between cytosine and adenine
CpG	Phosphodiester bond between cytosine and guanine
CpT	Phosphodiester bond between cytosine and thymine
CV	Co efficient of variance
D	Not C (G, A, and T)
DNA	Deoxyribonucleic acid
Dnmt 1	DNA methyltransferase 1
Dnmt 3a	DNA (cytosine-5)-methyltransferase 3A
Dnmt 3b	DNA (cytosine-5)-methyltransferase 3 beta
Dnmt 3L	DNA (cytosine-5)-methyltransferase 3-like
Exp CpG	Expected frequency of CpG dinucleotides
G	Guanine
H	Not G (A, C and T)
K	Keto (G or T)
5mC	5 methyl cytosine
M	Amino (A or C)
N^4	Nitrogen at 4^{th} position
N^6	Nitrogen at 6^{th} position
R	Purine
S	Strong (G or C)

T	Thymine
TpG	Phosphodiester bond between thymine and guanine
V	Not T or U (G, C and A)
W	Weak (A or T)
Y	Pyrimidine

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

DNA methylation is an epigenetic modification that plays very important role in vertebrate genomes as it is involved in number of important events that include gene regulation, gene imprinting, X-chromosome inactivation and even in diseases like cancer. In vertebrate genomes DNA methylation occurs at CG sites in both the strands of DNA. In the present studies the effect of flanking bases on the distribution of CG dinucleotides in the methylated as well as non methylated genomes is studied. In order to analyse the effect of flanking bases i.e. flanking sequence preference on the distribution of CG dinucleotides so as to understand the effect on rate of methylation we have designed various permutation sets ranging from a simple sets of NCG, CGN, NCGN, NCGNN, NNCGN, NNCG, CGNN, to complex ones i.e. NNCGNN, NNNCGNNN and NNNNCGNNNN. The consensuses sequences are derived based on NNNNCGNNNN (65536 permutations). Results of consensus sequences derived from high methylation group and low methylation group agrees with YCGR and RCGY observations repectively. There is appreciable overlap observed between the frequency based consensus sequences and those obtained by physiological data based on methylation levels which is very exhilarating.

CHAPTER 1

INTRODUCTION

1. INTRODUCTION

1.1 Background

The genome animatedly responds to the environment and various parameters belonging to it like stress, diet, behaviour, toxins etc. have the ability to activate chemical switches which regulate gene expression. All such kind of interactions come under the study known as epigenetics. Conrad Waddington coined the term epigenetics in 1942 and defined epigenetics as "the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being" (Waddington, 1940). Thus in simple words epigenetics can be defined as the study of heritable changes in gene expression that are not caused by changes in DNA sequence (Slatkin, 2009). Epigenetic changes are standard events and these events are very well predisposed by numerous factors including age, environment, disease state *etc*.. The above statement can be supported by the following example according to which epigenetic modifications can manifest as normally as the manner in which cells terminally differentiate to end up as skin cells, liver cells, brain cells, etc. or these epigenetic changes can have detrimental effects that can lead to diseases like cancer. Three systems including DNA methylation, histone modification and non-coding RNA associated gene silencing are currently considered to initiate and sustain epigenetic change. Recent ongoing research is continuously revealing the role of epigenetics in a variety of human disorders and fatal diseases (Issa, 2002).

1.2 DNA Methylation

DNA methylation is a biochemical process relating the addition of a methyl group to either cytosine or adenine DNA nucleotides in both prokaryotic as well as eukaryotic organisms. In prokaryotic organisms DNA methylation occurs at N^6 position of adenine and N^4 and C^5 position of cytosine while in eukaryotes only cytosine gets methylated at C^5 position and this cytosine methylation takes place predominantly at palindromic CG dinucleotides in both strands of the DNA. DNA methylation is one of the most intensely studied epigenetic modifications. This cytosine-5 methylation in mammals plays an significant key role in number of event that include embryonic development, gene imprinting, X chromosome inactivation, regulation of chromatin structure, silencing of transposons and endogenous retroviruses, genetic diseases and cancer biology (Ehrlich, 2003). The role of DNA methylation in relation to cancer is burning topic of research now days. DNA methylation silences tumor repressor genes.

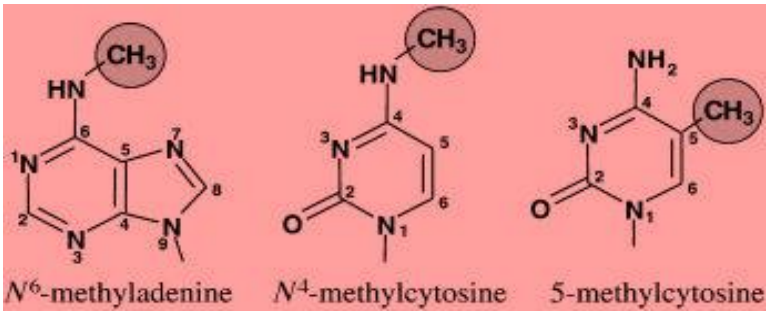


Fig 1: Structures of methylated bases occurring in DNA

Source: Jeltsch A, Chem BioChem 2002, 3, 274 -293

Apart from above mentioned events DNA methylation also plays important role in assuring proper regulation of gene expression along with stable gene silencing and histone modifications. The interface of these epigenetic modifications is very important as it is critical to regulate the functioning of the genome by changing chromatin architecture (James Hagood, 2014). This can be further explained by taking example of tumor-suppressor genes, these genes inactivation can occurs as a outcome of hypermethylation within the promoter regions and numerous studies carried out till date supports the fact that a broad range of genes are silenced by DNA methylation in diverse cancer types while on the other hand, global hypomethylation, inducing genomic instability, also contributes to cell transformation. DNA methylation shows potential in recognized translational use in patients and hypermethylated promoters thus may serve as biomarkers. Unlike genetic alterations, DNA methylation is reversible which makes it tremendously appealing for various kinds of therapies (Riggs, 1975). DNA methylation can be detected in the body fluids as well which is very useful in the early detection of tumors and is also very helpful in the determination of the prognosis.

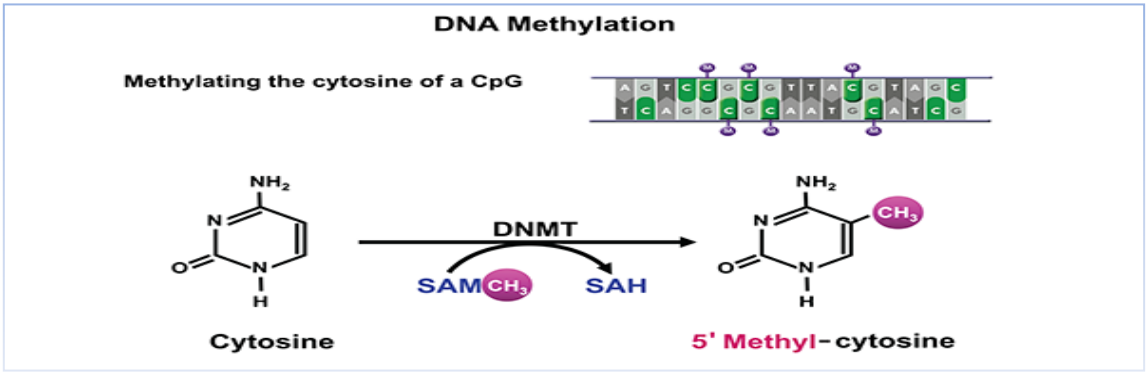


Fig 2: Conversion of Cytosine to 5-methyl Cytosine

Source: Zakhari S, Alcohol Research: Current Reviews 2013, *35, 8*

1.3 DNA Methylation Enzymes

As we know the process of methylation is enzymatically driven, DNA methyltransferases are enzymes responsible for establishing and maintenance of methylation pattern. DNA methyltransferase 1 abbreviated as Dnmt1 is responsible for inheritance of methylation pattern to daughter genomes produced during the process of DNA replication. Dnmt 1 enzyme have many fold high preference for a hemimethylated DNA substrate when compared to unmethylated DNA thought it is also competent of performing *de novo* methylation of unmethylated substrates in vitro (Zucker, 1983). The *de novo* DNA methyltransferases are responsible for establishing genomic methylation pattern during gametogenesis in a sex specific fashion followed by extensive demethylation of the genome, during embryogenesis (Li, 2002 & Meehan, 2003). The *de novo* methyltransferases are of further two types Dnmt3a and Dnmt3b, which methylate unmethylated as well as hemimethylated DNA without any preference (Okano *et al.*, 1998). In embryonic stem cells, early embryos and developing germ cells there is high expression of these *de novo* methyltransferases.

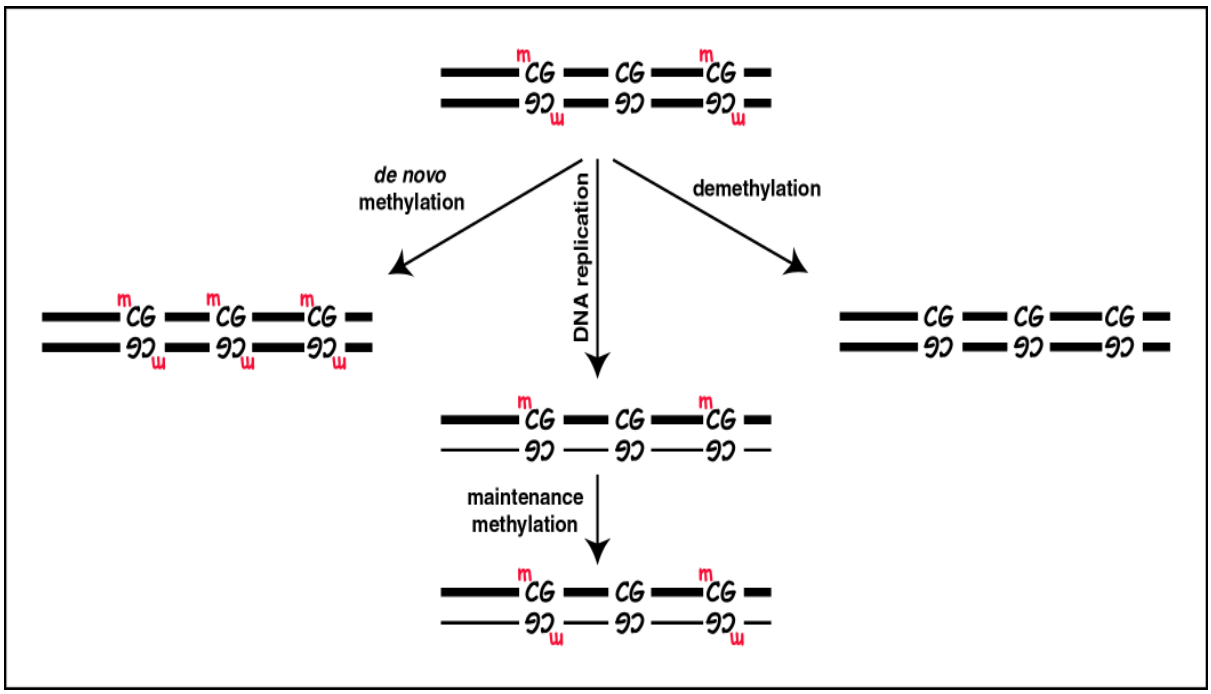


Fig 3: Role of *de novo* methylation and maintenance methylation enzymes

Source: Hariharan, 1974

In vertebrates DNA methylation occurs mainly in CpG dinucleotides depicted in above diagram as CG. Methyl residues are depicted as 'm'. New methylation patterns are

established by the process of *de novo* methylation (left, Fig 3). Existing methylation pattern can be erased by demethylation (right, Fig 3). During DNA replication (centre, Fig 3), the newly synthesized DNA strand (thin line) is unmethylated while the parent strands (thick line) retains its methylation pattern. The methylation pattern from the parent strand is copied on to the daughter strand by maintenance methyltransferase.

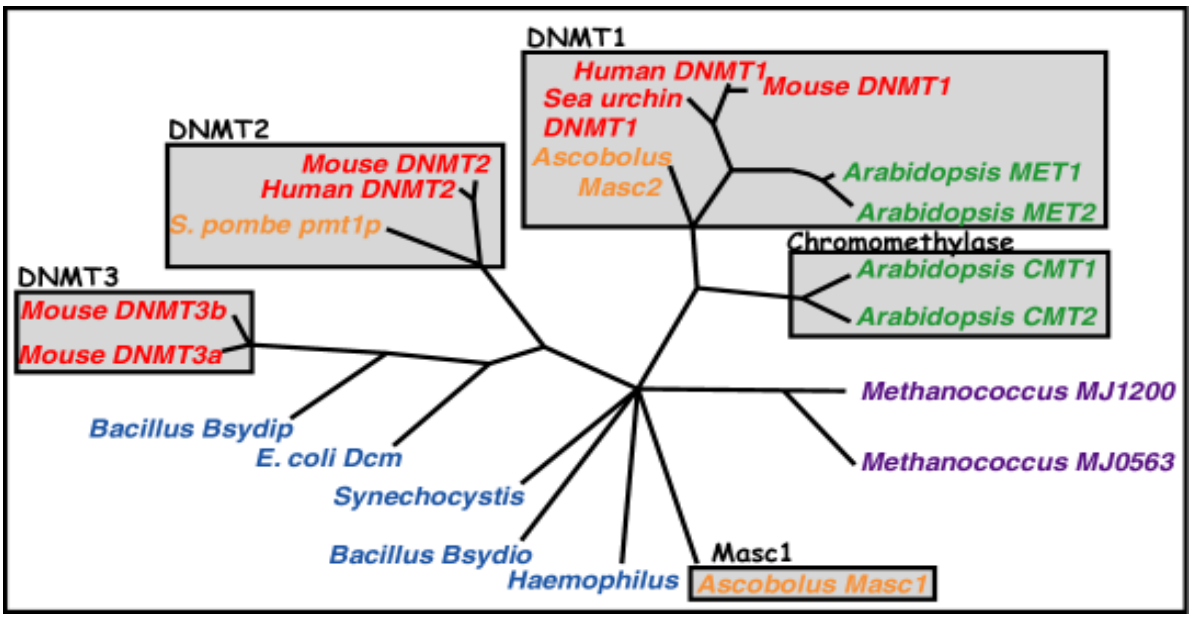


Fig 4: Families of eukaryotic MTases

Source: Colot and Rossignol, 1999

1.4 DNA Methylation in Eukaryotes

In eukaryotes there is no uniform pattern of methylation, there are methylated regions which are interspersed with unmethylated regions. In the course of evolution there is progressive elimination of CG dinucleotides especially in higher eukaryotes. DNA methylation in eukaryotes chiefly occur at CG dinucleotides resulting in conversion of cytosine into 5-methyl cytosine (5mC) which occasionally get hydrolytically deaminated to thymidine which being one of the natural base of DNA is not efficiently removed by repair system resulting in CG-deficiency in the genome.

However, contrary to the mutagenic properties of methyl cytosine in context to CG dinucleotides, unmethylated cytosine (CpA or any other CpH) if deaminated result in conversion to form Uracil, which being unnatural base for DNA is efficiently removed by the uracil DNA glycosylase to restore original pair against mismatched pair. This spontaneous deamination of 5mC to thymine leads to TpG and CpA mutations (Coulondre *et al.,* 1978). Therefore, this CG are globally underrepresented in genomes of vertebrate species and leads to uneven distribution of CGs in eukaryotic methylated genomes.

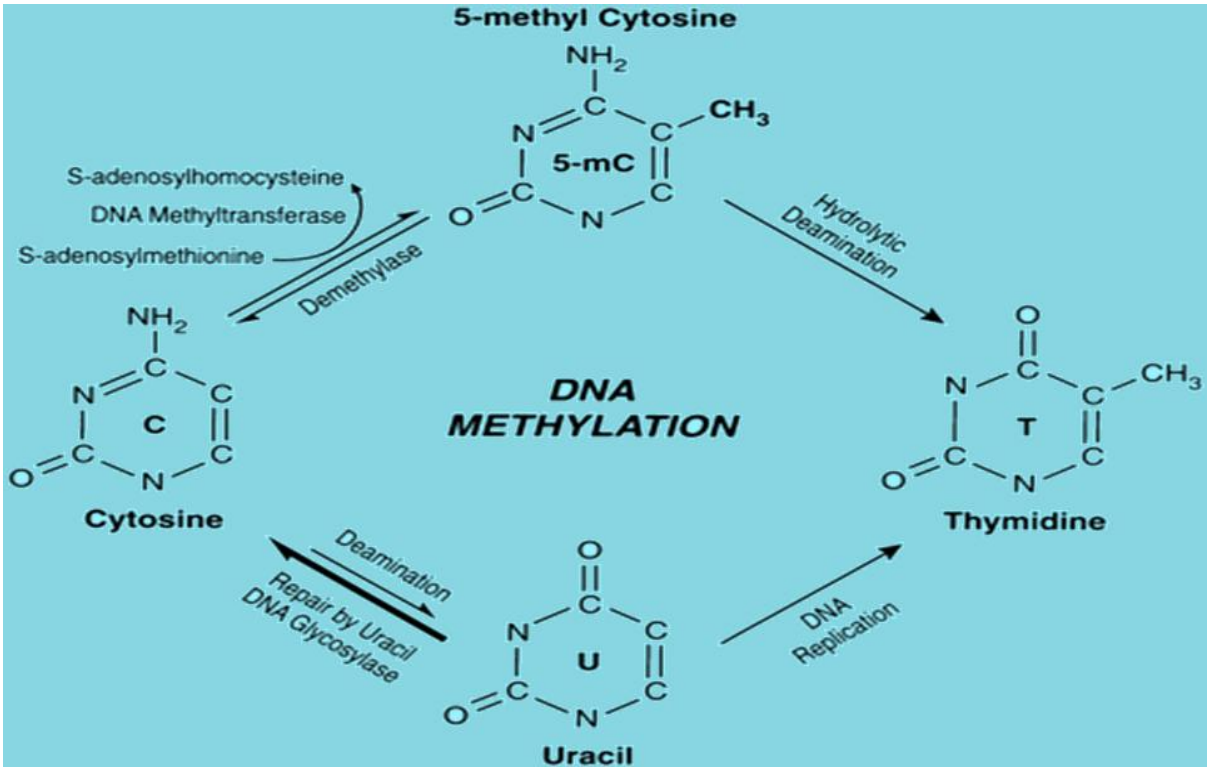


Fig 5: Biochemical pathway for cytosine methylation, demethylation and mutagenesis cytosine and 5-mC.

Source: Singhal and Ginder 2013

1.5 CpG Island

The uneven distribution of CG dinucleotides is exemplified by the existence of CpG islands. The CpG islands exhibit higher representation of CG dinucleotides when compared to rest of the genome in vertebrates. The mammalian genomes contain 60 million CG dinucleotides and 70–80% of those are modified in a non-random pattern (Handa and Jeltsch, 2005). In these CpG islands the CG dinucleotides are usually found to be unmethylated and and are associated with transcriptionally active housekeeping genes (near promoter or first exon) unlike rest of the genome where majority of CG dinucleotides are mostly methylated.

CpG islands are usually unmethylated and methylation of CpG islands of certain genes have also been found to be associated with cancer (Shimizu et al., 1997). CpG island density depends upon various genomic features of organism like chromosome size, chromosome number and GC% (Leng Hang *et al.*, 2008). CpG island density is high in telomere region of chromosomes. There are two most accepted definitions of these CpG islands which were proposed by Gardiner-Garden and Frommer and Takai and Jones.

Gardiner-Garden and Frommer (1987) definition states that these CpG islands are at least 200 base pair long with a minimum G + C content of at least 50% and Observed/expected CpG ratio at least 0.6 (Gardiner-Garden *et al.*, 1987). In 2002, Takai and Jones analyzed human Chromosome 21 and 22 and proposed an improved definition of CpG islands taking considerations of pseudo genes and alu repeats into account.

Takai and Jones definition states that these CpG islands are at least 500 base pair long with a minimum G + C content of 55% and Observed/expected CpG ratio at least 0.65 (Takai and Jones, 2002).

1.6 Effect of the flanking sequences on the distribution of CG dinucleotides

CG dinucleotide distribution in methylated genomes is uneven. This unevenness is reflected from the existence of CpG islands. CG dinucleotide distribution pattern further shows variations near the 5` end of genes in vertebrates, invertebrates, plants and bacteria (Shimizu *et al.*,1997). This unevenness in distribution of CG dinucleotide can be explained considering the fact that some CG dinucleotides are more prone to methylation reducing their frequency in the genome compared to others CG dinucleotides thus, resulting in the decrement of CGs in the genome of methylated organisms. One reason for this preference could be explained on the basis of flanking sequence preference for the DNA methyltransferases. CGs having flanking sequences which are preferred by the DNA methyltransferases are more prone to get methylated compared to those CGs which are having flanking sequences which are either partially preferred or totally not preferred by DNA methyltransferases. In present work CG distribution studies has been carried out by analysis of flanking sequences (4 positions upstream and 4 positions downstream) i.e. NNNNCGNNNN primarily along with smaller flanking sequences as well to study the effect of flanking sequences on the distribution of CGs in the methylated genomes. The plan of the study is based on the following logic. The flanks responsible for high or low propensity of CGs to get methylated will occur in the genome with lower and higher frequencies respectively. Higher the propensity of methylation, higher will be probability of methylation and proportionately higher will be the probability of deamination of methylated cytosine leading to its mutation.

CHAPTER 2

REVIEW LITERATURE

2. REVIEW OF LITERATURE

Research carried out in 1980s onwards focused on significance of flanking bases surrounding CG dinucleotides to understand the probability of CG dinucleotide to get mutated to TG/CA. Studies carried out in 1980 clearly established the fact that 5mC tends to mutate thymine which in turn result in CpG deficiency in heavily methylated genomes. This finding was supported by the analysis of nearest neighbour dinucleotides frequencies (Bird, 1980).

During 1985 studies were carried out to study the effect of flanking sequences on the *de novo* methylation of CG pairs by the human DNA methylase. In this study synthetic oligodeoxynucleotides were used and this research demonstrated clearly that flanking DNA sequences can be critical in determining whether a CG site can be methylated or not (Weissbach *et al*., 1985).

CG under-representation is seen in vertebrates, many diverse protist genomes, dicot plants, metazoan mitochondrial genomes, almost all vertebrate small viral genomes, many examples of thermophilic bacteria and some exceptional bacterial species, e.g. *Borrelia burgdorferi* and *Mycoplasma capricolum.*CG suppression across vertebrates is commonly described to the classical methylation-deamination mutation scenario (Karlin, 1995).

Another research were carried during early 1996 which showed that the CG which is accompanied by pyrimidine at 5' and one or more purine downstream is highly susceptible to get mutated. This study proposed the finding that TCGA is the most common mutation site. As per ollika *et al.* studies mutation appear in all tetra nucleotides but significant mutation was seen in YCGR combination. ollika *et al.* also extended their studies from tetra nucleotides to hex nucleotides because earlier studies with tetra nucleotides established clear sequence preference which further opened rooms for further investigation to analyse longer sequences as well so as to quantify the effects of longer flanks surrounding CG. On the basis of this NNCGNN studies ollika *et al.* proposed that YYCGRR and YYCGRY are combinations which are highly prone to get mutated. Both these sequences contain YCGR pattern. They also proposed that asymmetric sequences i.e. mixed occurrence of purine and pyrimidine on one or both sides is very rare and underrepresented (ollika *et al.*, 1996).

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 sequences. They further proposed that there are distinct statistically significant consensus sequences flanking CpG site

which induce 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 yet well understood (Handa & Jeltsch *et al*., 2005).

In one of the study carried during 2007 by Kim *et al*., CG site specific methylation information was used to characterize CG site methylation vulnerability. Findings of this research showed that there was significant difference in DNA character composition between methylation susceptible and resistant sequences. Comparison of methylation susceptible and resistant sequences using the two sample logo technique showed that over-represented characters in methylation susceptible sequences are in harmony with the analysis by Handa and Jeltsch showing CG flanking sequence specificity for methylation susceptibility. Furthermore CG flanking sequences was used to build predictive models for methylation susceptibility and achieved over 75% prediction accuracy in 10 fold cross validation tests (Kim et al., 2007).

Studies carried out during 2008 by Zhang *et al.* proposed that A/T flanks correlated with DNA methylation. They reported that this tendency was not uniform at all flanking sequences. The other important finding of zhang *et al.* studies include that all neighbouring base pairs are of not equal importance for DNA methylation. They suggested that correlation between DNA methylation and DNA sequence could be flanking sequence preference for DNA methyltransferase (Zhang *et al.*, 2008).

Studies carried out during 2008 by Bethany *et al*. mapped DNA methylation patterns of 190 gene promoter regions on chromosome 21 using bisulphite conversion followed by sequencing in five human cell types after sub cloning. Bethany *et al*. sequenced 28,626 clones using Sanger sequencing resulting in the measurement of the DNA methylation state of 580427 CG sites. Their results showed that average DNA methylation levels are distributed bimodally. Furthermore within CG-rich sequences, DNA methylation was found to be anti-correlated with CG dinucleotide density and GC content, and methylated CpGs are more likely to be flanked by AT rich sequences. Yingying *et al*. observed over-representation of CG sites in distances of 9, 18, and 27 bps in highly methylated amplicons and DNA methylation in promoter regions is strongly correlated with the absence of gene expression and low levels of activating epigenetic marks like H3K4 methylation and H3K9 and K14 acetylation (Yingying *et al*., 2008).

Bethany *et al.* studies showed that human DNMT3A and DNMT3B possess significant as

well as distinct flanking sequence preferences for CG sites. Base composition at the -2 and +2 positions flanking the CG site for DNMT3A and at -1 and +1 position for DNMT3B were used for selection of high and low efficiency sites. This leads to the formation of specific *de novo* methylation patterns characterized by up to 34-fold variations in the efficiency of DNA methylation at individual sites. Furthermore, analysis of the allocation of signature methylation hotspot as well as coldspot motifs supports the fact that DNMT flanking sequence preference has contributed to determining the composition of CpG islands in the human genome. Furthermore DNMT3L stimulatory factor modulates the formation of *de novo* methylation patterns in two ways. Firstly DNMT3L selectively focuses the DNA methylation machinery on properly chromatinized DNA templates. Secondly DNMT3L attenuates the impact of the intrinsic DNMT flanking sequence preference by providing a much greater heighten to the methylation of poorly methylated sites, promoting the formation of broader and more uniform methylation patterns (Bethany *et al*., 2009).

DNA methylation and its role in cancer is most widely researched topic now days. Alterations in DNA methylation are frequent in a variety of tumors as abnormal DNA methylation in the CG context is frequently observed in cancer cells and it is known that aberrant DNA methylation silences tumor repressor genes. Hypermethylation, which represses transcription of the promoter regions of tumor suppressor genes leading to gene silencing, has been most extensively studied in the recent times.

Global hypomethylation has also been recognized as a cause of oncogenesis. Information regarding the mechanism of methylation and its control has helped a lot in the discovery of many regulatory proteins and enzymes as potential therapeutics. Methylation occurs early and can be detected in body fluids; making it of very significant use in early detection of tumors and for determining the prognosis. The flanking sequences surrounding the CG dinucleotides are used as a measure to various cancer types.

Recent work of Jaehyun *et al*., 2013 related to Genome wide analysis and modelling of DNA methylation susceptibility in 30 breast cancer cell lines by using CG flanking sequences was one of the step towards understanding a relationship between DNA methylation and cancer. In this study they investigated differences in nucleotides composition around CG sites where high levels of methylation are observed and in turn they used this information for modelling DNA methylation susceptibility. In this study they observed that DNA methylation is not uniform in the whole genome region and character composition of CG flanking sequences are

significantly different between hypermethylated groups and hypomethylated groups. There is enrichment of adenine and thymine in specific positions around hyper methylated sites and there is increase in methylation adenine and thymine content in specific positions around the hypermethylated sites are increased while adenine and thymine content in other positions around hypermethylated sites is decremented. The genome wide analysis of Jaehyun *et al*., showed that nucleotides around CG sites contains information for cytosine methylation (Jaehyun *et al*., 2013).

CHAPTER 3

ORIGIN OF PROBLEM
&
SCOPE OF STUDY

3. ORIGIN OF PROBLEM AND SCOPE OF STUDY

CG flanking bases influence methylation propensity, the effect has not been studied at genome level and moreover the work done on sequence preference by DNA methyltransferases has been done in biochemical systems (*in vitro* methylation) or by studying physiological conditions (bisulfate sequencing of genomic DNA of somatic cells) but not in germ lines. Tetra, hepta and decanucleotide flanking sequence analyses indicated clear overall sequence preference for pyrimidines 5' and purines 3' to the methylated 5-methylcytosine (RCGY). CG islands are usually not methylated and exhibit abundant presence of CGs. Additionally, flanking sequences appear to influence methylation of CGs and implicitly their mutation into TpG/CpA.

Studies on the effect of flanking sequences on the distribution of CGs in the methylated genomes is preliminary work to understand how flanking sequences surrounding CG on both upstream and downstream regions are effecting the methylation of these CG dinucleotides. NCG, CGN, NCGN, NCGNN, NNCGN, NNCG, CGNN, NNCGNN, NNNCGNNN, NNNNCGNNNN (where N represent G, A, T and C) possible permutations have been designed to study effect of these flanking bases on CG distribution. The CG those flank is preferred by DNA methyltransferase is more prone to methylation and hence represented with low frequency in the genome. This work focuses on understanding the relationship between CG distribution and their methylation and furthermore CG/CG → TG/CA mutations may be better understood in the genome in context of their flanking bases. Overall this study will give insight into the contributory role of DNA methylation in evolutionary changes in genomes structure of eukaryotes in which DNA methylation occurs.

.

CHAPTER 4

OBJECTIVES

4. OBJECTIVE

- Coding of a computer program to determine frequency of permutations in given contig sequence.
- To create a dataset of all possible permutations of dinucleotide set, NCG, CGN, NCGN, NCGNN, NNCGN, NNCG, CGNN, NNCGNN, NNNCGNNN, NNNNCGNNNN
- Determination of frequency of each NNNNCGNNNN and other data set permutations in human chromosomes 21.
- Statistical analysis of the frequency of distribution of the possible permutations.
- Comparison of consensus sequence of high frequency and low frequency groups.
- To extend similar work in other organisms.

CHAPTER 5

MATERIAL AND METHODS

5. MATERIALS AND METHODS

5.1 Data Source

DNA sequences were downloaded from National Centre for Biotechnology information (NCBI) (URL: www.ncbi.nlm.nih.gov). To study the effect of flanking sequences on the distribution of CG dinucleotides in methylated genomes, six different species DNA sequences was selected randomly and downloaded from NCBI.

Table 1: DNA sequences of different species taken from Gen Bank to perform distribution study.

S.NO	Species	Common name	Chromosome number	Size
1	*Homo sapiens*	Human	Chromosome 21	48,129,895 bp
2	*Pan troglodytes*	Chimpanzee	Chromosome 22	193495092 bp
3	*Mus musculus*	Mouse	Chromosome 19	98207768 bp
4	*Danio rerio*	Zebra fish	Chromosome 21	445,440,65 bp
5	*Drosophila melanogester*	Fruit fly	Chromosome 3	23011544 bp
6	*Caenorhabditis elegans*	Roundworm	Chromosome 3	13783700bp

5.2 Permutations

A dataset of all the possible NNNNCGNNNN sequence permutations was created by considering all possibilities at -1,-2,-3,-4 and +1,+2,+3,+4 positions. 65536 (4^8) different NNNNCGNNNN permutations were made and the occurrence of these individual motifs was checked using C++ program in human chromosomes and the genomes of other related organisms so as to see the effect of flanking sequences on the distribution of CG in the methylated genomes.

5.2.1 Nomenclature of Permutations

Throughout this work, the bases flanking the central CG sites are designated as illustrated below:

Table 2: Nomenclature of CG flanking bases

-4	**-3**	**-2**	**-1**			**+1**	**+2**	**+3**	**+4**
N	N	N	N	**C**	**G**	N	N	N	N

Where N = G, A, T or C

5.2.2 Calculation for number of possible permutations

As it is well known that DNA contains four bases which are adenine (A), thymine (T), cytosine (C) and guanine (G) and we are allowing a base to be used more than once at each position there are 4*4*4*4*4*4*4*4 = 65536 possible 8 letter permutations with additional central CG with our four character alphabet which are A, G, T, C. The 65536 sequences each 10 bp long and containing CG in the centre, as shown in above table, were constructed using MS Excel spreadsheets.

5.3 Sequence Analysis Tools

1. Microsoft Word

Microsoft word was used to analyze the DNA sequences with the help of tools such as find, replace and recording macros.

2. Microsoft Excel

Microsoft Excel spreadsheet was used for computation and statistical analysis of data of analyzed sequence.

3. Notepad ++

In order to calculate the A+G+T+C content of the sequence notepad ++ was used. Notepad ++ allowed to select a particular range of the sequence using its functions like “Go to particular line number option”, “Begin selection” and “end selection” to enable the sequence to be analysed using Genomics GC content calculator more over this tool can allow be used to find number of A, G, T, C in the pasted sequence but it a bit longer time and in case the sequence is very big this tool is not able to handle the operation.

4. Genomics %G~C Content Calculator

Genomics %G~C Content Calculator is an online software which is used to screen DNA sequences and calculate its total base count, adenine count, thymine count, guanine count cytosine count as well as %G~C content by pasting the DNA sequence to be analyzed into the box. The calculator recognizes sequences in FASTA format and the URL of the tool is http://www.sciencebuddies.org/science-fair-projects/project_ideas/genomm_GC_Calculator.
One limitation with this tool is size constrain it can comfortably without hanging takes up to approximately 700,000 for analysis in one single attempt and for the longer sequences user needs to break the sequence into parts using tools like notepad ++.

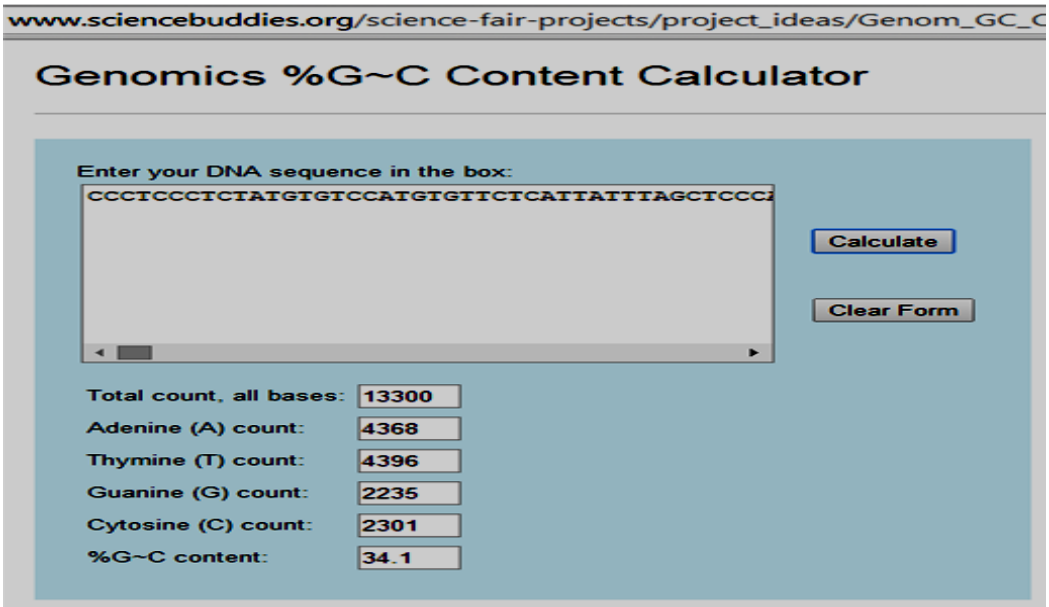


Fig 6: Genomics G~C Content Calculator interface

5.3.1 C++ Program

A C++ program was coded to search each of individual motif (65536) in target sequences. Files created for the use in program are as below:

(i) Fasta sequence: This is a notepad file in which the fasta sequence downloaded from NCBI can be pasted by the user and once the work is finished with the pasted sequence that sequence can be deleted to allow the user to paste the new sequence for analysis.

(ii) Fasta sequence 1: This is a notepad file which is generated automatically and requires no input by the user. This file is generated to get a continuous sequence without any gaps and is finally used by the C++ compiler to perform analysis. While pasting a new sequence user is required to delete sequence from the file Fasta sequence and paste new fresh sequence unlike in this file user need not to perform such operation as the compiler will automatically delete the previous file and will create the new one at the time of compilation.

(iii) Input file: To allow the user to provide a large number of input strings in a single ‘go’ this file is created. This saves the time and effort of the user as it allows the user to copy and paste the input strings in this particular file which saves the time to be required to enter the input strings manually using the compiler interface which would have been very tedious job especially when the data set is as large as 65,536 input strings. The motifs to be searched are pasted in the input file to allow compiler to search the motifs in the Fasta sequence 1 file.

(iv) Output file: The result of the motif search is directed to file named output from which

results can be copied directly and pasted in excel for further analysis which again saves time as it is not possible for the user to directly copy and paste the results from the compiler interface and leaves the user with a option of manually extracting the results from the compiler interface which is again a very tedious job especially when the data set is as large as 65,536 input strings.

5.3.2 Compiler

The program is compiled using Dev C++ compiler. Dev-C++ is a free integrated development environment (IDE). It is a very user friendly compiler which is distributed under the GNU which is General Public License for programming in languages like C and C++. Dev C++ is provided with a free in build compiler MinGW. The integrated development environment is written in Delphi. In this project Bloodshed Dev-C++ 5.5.1 compiler is used. Bloodshed Dev-C++ is a full-featured Integrated Development Environment for the programming languages C and C++. It uses the MinGW or TDM-GCC 64bit port of the GCC as its compiler for compilation and is hosted by Source Forge. Dev-C++ runs exclusively on Microsoft Windows and therefore generally Dev-C++ is considered as a Windows-only program, but there are attempts to create a Linux version as well. Colin Laplace is credited as the developer of Dev C++.

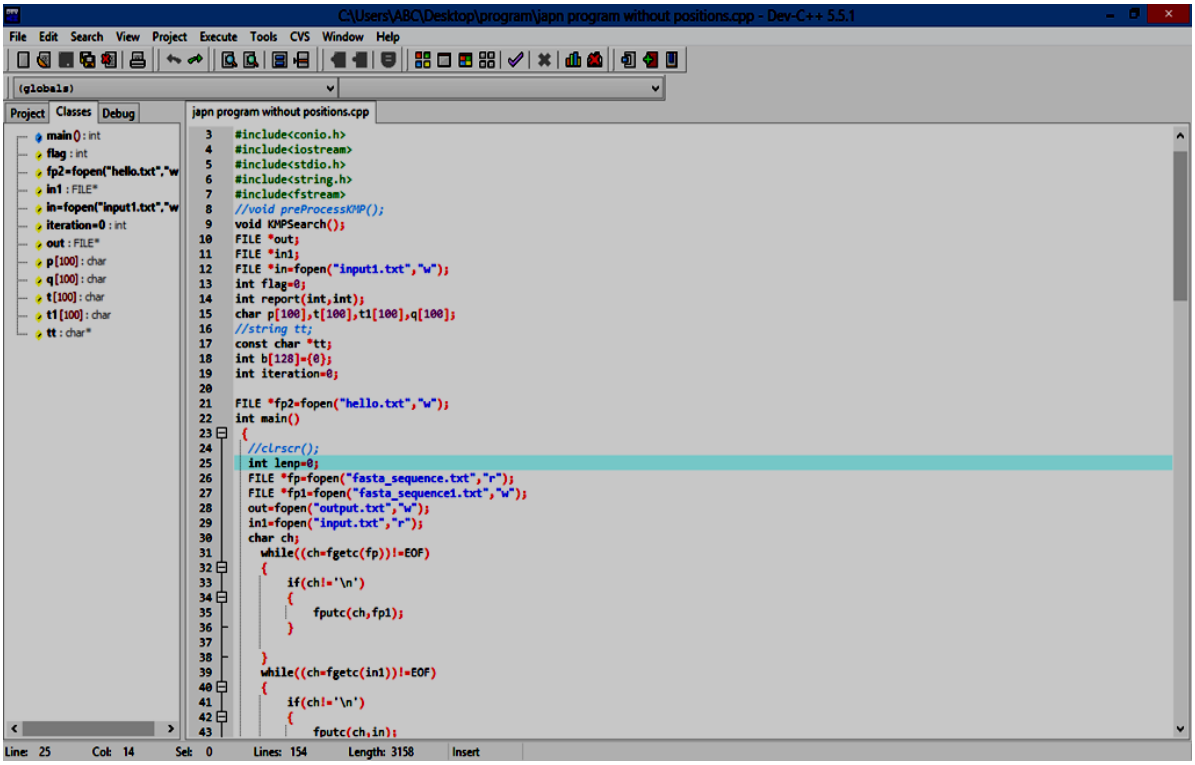


Fig 7: Dev C++ compiler interface showing browsing options

5.3.3 Algorithm

fopen (“File name”,“mode”) will open the file in read mode from which data is to be read and using fgetc ( ) function will read the file and furthermore remove the blank spaces from file. After removing the blank spaces the modified data is written to a new file using fputc ( ) function. Each time program will get 90 characters from file for matching it with pattern by calling KMPSearch ( ) function which is described latter. KMPSearch identifies any pattern matching in the 90 characters each time in a loop and after executing KMPSearch each time, go back by the length of pattern to search again with next 90 characters. Finally output is redirected to another file using fputc ( ) function. KMP search function is discussed on next page.

5.3.4 KMP Search( )

1. Match pattern with 90 characters read from file, character by character

2. If match occur:

a. If (j=m) i.e. length of pattern , then we found a match

- Report this match on screen.
- Make pointer position decremented by (length of pattern -1) for next search.
- Set j=0 for matching pattern again with remaining string.

b. Else increment both pointers i.e. for string and pattern, by 1.

3. Else set i = i-j+1 and j=0

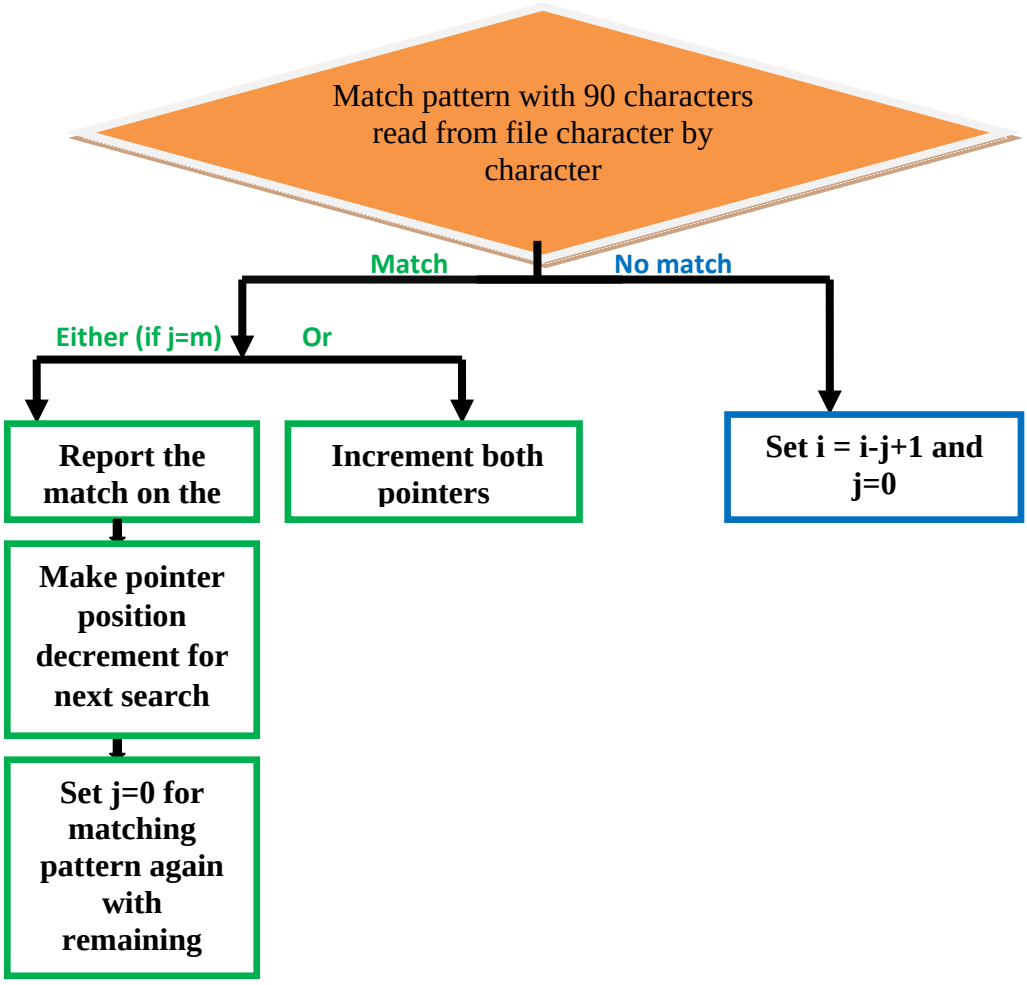
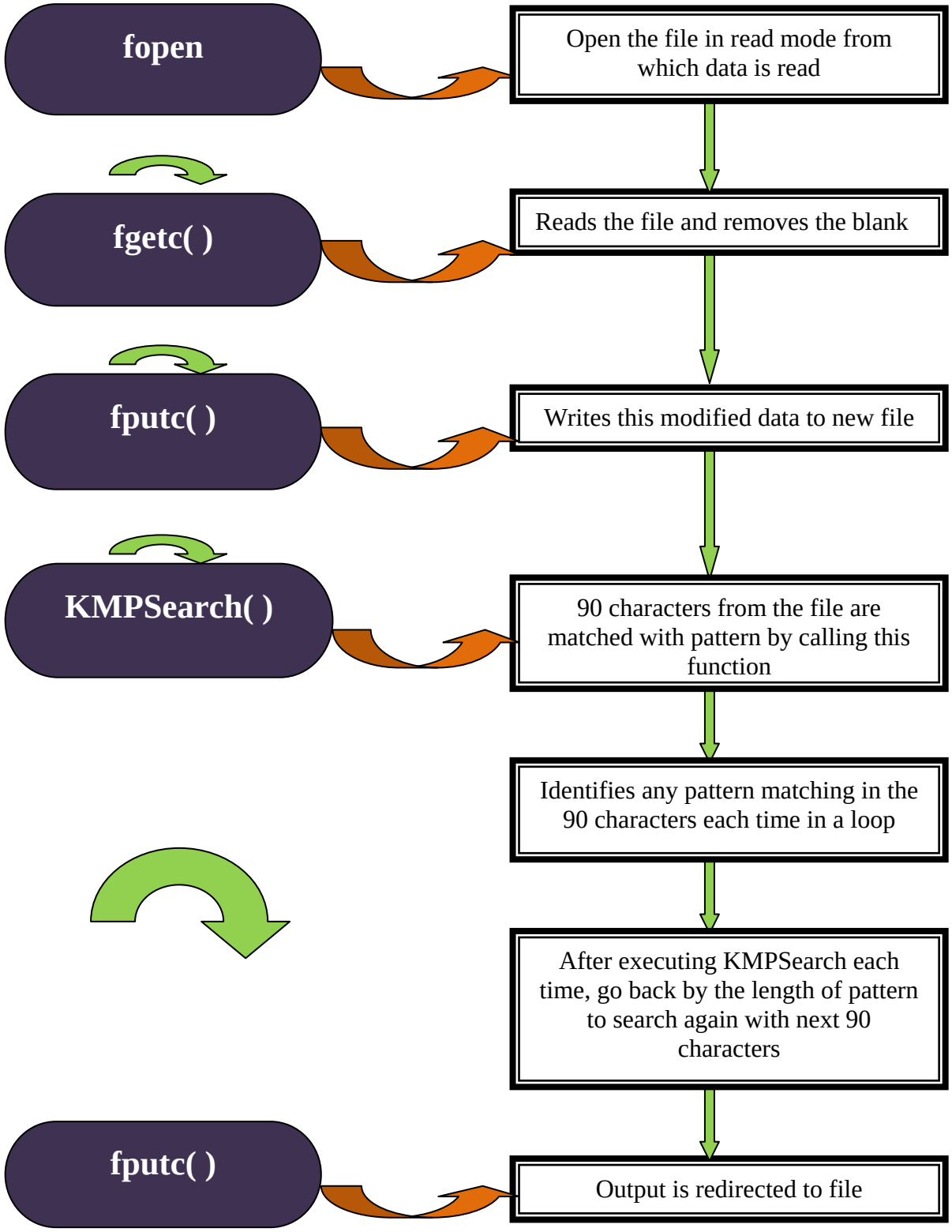


Fig 8: Schematic representation KMP Search algorithm

5.3.5 Flowchart of algorithm



5.3.6 Program code

```
using namespace std;
#include<conio.h>
#include<iostream>
#include<stdio.h>
#include<string.h>
#include<fstream>
//void preProcessKMP();
void KMPSearch();
FILE *out;
FILE *in1;
FILE *in=fopen("input1.txt","w");
int flag=0;
int report(int,int);
char p[100],t[100],t1[100],q[100];
//string tt;
const char *tt;
int b[128]={0};
int iteration=0;

FILE *fp2=fopen("hello.txt","w");
int main()
 {
  //clrscr();
  int lenp=0;
  FILE *fp=fopen("fasta_sequence.txt","r");
  FILE *fp1=fopen("fasta_sequence1.txt","w");
  out=fopen("output.txt","w");
  in1=fopen("input.txt","r");
  char ch;
        while((ch=fgetc(fp))!=EOF)
        {
                if(ch!='\n')
                {
                        fputc(ch,fp1);
                }

        }
        while((ch=fgetc(in1))!=EOF)
        {
                if(ch!='\n')
                {
                        fputc(ch,in);
                }

        }
        fclose(fp);
        fclose(fp1);
        fclose(in);
```

```
        fclose(in1);
        int n=0;
        out=fopen("output.txt","a");
        in=fopen("input1.txt","r");
//      in1=fopen("input.txt","a");
//      std::cout<<"Enter the number, how many times you want to search\n";
//      fprintf(out,"Enter the number, how many times you want to search\t");
//      std::cin>>n;
//      fflush(stdin);
//      fprintf(out,"%d\n",n);
 lenp=10;
 fflush(stdin);
  fp=fopen("fasta_sequence1.txt","r");
 //while(std::getline(in1,p))
 while(fgets(p,11,in)!=NULL)
 {
        //cout<<"hello1";
 /*for(int l=0;l<10&&q[l]!='\n';l++)
 {
        cout<<"helloloop";
        p[l]=q[l];
 }*/
 //cout<<"The pattern is\t"<<p[];
 puts(p);
 fprintf(out,"finding %s",p);
 //cout<<"hello";
//      getline(in1,p);
          fp=fopen("fasta_sequence1.txt","r");
          //std::cout<<"Enter the pattern string\t";
          //fprintf(out,"Enter the pattern string\t");
         // fflush(stdin);
         // gets(p);
         fflush(stdin);
          //fprintf(out,"%s",p);
         // fprintf(out,"\n");
         // fflush(stdin);
         // lenp=strlen(p);
          flag=0;
          int k=0;
          while(fgets(t1,90,fp)!=NULL)
          {
                strcat(t,t1);
                ++iteration;
                KMPSearch();
                k=0;
                for(int i=90-lenp;i<=89;i++)
                {
                        t[k]=t1[i];
                        k++;
                }
```

```
                t[k]='\0';
            }
            if(flag)
             {
               std::cout<<"Substring found "<<flag<<" number of times\n";
               fprintf(out," %d \n",flag);
               fflush(stdin);
              //  tt="Substring found "+flag+" number of times\n";
               //fputs(tt,fp2);
            }
            else if(flag==0)
            {

             std::cout <<"Not a substring\n";
             fprintf(out,"Not a substring\n");
             fflush(stdin);
            }
            fclose(fp);
  fgets(p,1,in1);
}
 fclose(out);
 fclose(fp2);
 }//end of main

void KMPSearch()
 {
  int i=0,j=0,m=strlen(p),n=strlen(t);
  const char *pp;
  while(i<n)
   {
    if(t[i]==p[j])
     {
      if(j==m-1)
       {
          //std::cout<<"hello";
       flag++;
       //std::cout<<"A match is there at the position "<<((iteration-1)*90)+(i-j)<<"\n";
       //fprintf(out,"A match is there at the position %d \n",((iteration-1)*90)+(i-j));
       fflush(stdin);
          //pp="A match is there at the position "+((iteration-1)*90)+(i-j)+"\n";
          //        fputs(pp,fp2);
       i=i-j+1;
                   j=0;
      }
            j++;
     i++;
     }
          else
          {
           i=i-j+1;
```

```
            j=0;

            }
        }
    }
```

5.3.7 Data Set Tables of All Possible Permutations

Table 3: Dataset table of dinucleotide set, NCG, CGN, NCGN, NNCG, CGNN

Dinucleotide Set		
CG		
TG		
CA		
AC		
GT		
GC		
NCG Set	**CGN Set**	
GCG	CGG	
CCG	CGC	
ACG	CGA	
TCG	CGT	
NCGN	**NNCG**	**CGNN**
GCGG	TTCG	CGGC
GCGC	ATCG	CGGG
GCGA	TCCG	CGGT
GCGT	AACG	CGGA
CCGG	CTCG	CGCC
CCGA	GCCG	CGCG
CCGT	GACG	CGCT
CCGC	GTCG	CGCA
ACGG	CACG	CGAC
ACGC	ACCG	CGAG
ACGT	TACG	CGAT
ACGA	TGCG	CGAA
TCGG	GGCG	CGTC
TCGC	AGCG	CGTG
TCGA	CCCG	CGTT
TCGT	CGCG	CGTA

Table 4: Dataset table of NNCGN

NNCGN Set		
AACGA	GCCGA	TACGA
AACGG	GCCGG	TACGG
AACGC	GCCGC	TACGC
AACGT	GCCGT	TACGT
AGCGA	GTCGA	TGCGA
AGCGG	GTCGG	TGCGG
AGCGC	GTCGC	TGCGC
AGCGT	GTCGT	TGCGT
ACCGA	CACGA	TCCGA
ACCGG	CACGG	TCCGG
ACCGC	CACGC	TCCGC
ACCGT	CACGT	TCCGT
ATCGA	CGCGA	TTCGA
ATCGG	CGCGG	TTCGG
ATCGC	CGCGC	TTCGC
ATCGT	CGCGT	TTCGT
GACGA	CCCGA	
GACGG	CCCGG	
GACGC	CCCGC	
GACGT	CCCGT	
GGCGA	CTCGA	
GGCGG	CTCGG	
GGCGC	CTCGC	
GGCGT	CTCGT	

Table 5: Dataset table of NNCGN

NCGNN Set		
ACGAA	TCGGG	CCGCT
GCGAA	ACGGC	TCGCT
CCGAA	GCGGC	ACGTA
TCGAA	CCGGC	GCGTA
ACGAG	TCGGC	CCGTA
GCGAG	ACGGT	TCGTA
CCGAG	GCGGT	ACGTG
TCGAG	CCGGT	GCGTG
ACGAC	TCGGT	CCGTG
GCGAC	ACGCA	TCGTG
CCGAC	GCGCA	ACGTC
TCGAC	CCGCA	GCGTC
ACGAT	TCGCA	CCGTC
GCGAT	ACGCG	TCGTC
CCGAT	GCGCG	ACGTT
TCGAT	CCGCG	GCGTT
ACGGA	TCGCG	CCGTT
GCGGA	ACGCC	TCGTT
CCGGA	GCGCC	
TCGGA	CCGCC	
ACGGG	TCGCC	
GCGGG	ACGCT	
CCGGG	GCGCT	

NNCGNN								
AACGAA	AGCGTT	ATCGTC	GGCGCC	GTCGCG	CGCGCA	CTCGCC	TGCGCG	TTCGCA
AACGAG	ACCGAA	ATCGTT	GGCGCT	GTCGCC	CGCGCG	CTCGCT	TGCGCC	TTCGCG
AACGAC	ACCGAG	GACGAA	GGCGTA	GTCGCT	CGCGCC	CTCGTA	TGCGCT	TTCGCC
AACGAT	ACCGAC	GACGAG	GGCGTG	GTCGTA	CGCGCT	CTCGTG	TGCGTA	TTCGCT
AACGGA	ACCGAT	GACGAC	GGCGTC	GTCGTG	CGCGTA	CTCGTC	TGCGTG	TTCGTA
AACGGG	ACCGGA	GACGAT	GGCGTT	GTCGTC	CGCGTG	CTCGTT	TGCGTC	TTCGTG
AACGGC	ACCGGG	GACGGA	GCCGAA	GTCGTT	CGCGTC	TACGAA	TGCGTT	TTCGTC
AACGGT	ACCGGC	GACGGG	GCCGAG	CACGAA	CGCGTT	TACGAG	TCCGAA	TTCGTT
AACGCA	ACCGGT	GACGGC	GCCGAC	CACGAG	CCCGAA	TACGAC	TCCGAG	
AACGCG	ACCGCA	GACGGT	GCCGAT	CACGAC	CCCGAG	TACGAT	TCCGAC	
AACGCC	ACCGCG	GACGCA	GCCGGA	CACGAT	CCCGAC	TACGGA	TCCGAT	
AACGCT	ACCGCC	GACGCG	GCCGGG	CACGGA	CCCGAT	TACGGG	TCCGGA	
AACGTA	ACCGCT	GACGCC	GCCGGC	CACGGG	CCCGGA	TACGGC	TCCGGG	
AACGTG	ACCGTA	GACGCT	GCCGGT	CACGGC	CCCGGG	TACGGT	TCCGGC	
AACGTC	ACCGTG	GACGTA	GCCGCA	CACGGT	CCCGGC	TACGCA	TCCGGT	
AACGTT	ACCGTC	GACGTG	GCCGCG	CACGCA	CCCGGT	TACGCG	TCCGCA	
AGCGAA	ACCGTT	GACGTC	GCCGCC	CACGCG	CCCGCA	TACGCC	TCCGCG	
AGCGAG	ATCGAA	GACGTT	GCCGCT	CACGCC	CCCGTA	TACGCT	TCCGCC	
AGCGAC	ATCGAG	GGCGAA	GCCGTA	CACGCT	CCCGTG	TACGTA	TCCGCT	
AGCGAT	ATCGAC	GGCGAG	GCCGTG	CACGTA	CCCGTC	TACGTG	TCCGTA	
AGCGGA	ATCGAT	GGCGAC	GCCGTC	CACGTG	CCCGTT	TACGTC	TCCGTG	
AGCGGG	ATCGGA	GGCGAT	GCCGTT	CACGTC	CTCGAA	TACGTT	TCCGTC	
AGCGGC	ATCGGG	GGCGGA	GTCGAA	CACGTT	CTCGAG	TGCGAA	TCCGTT	
AGCGGT	ATCGGC	GGCGGG	GTCGAG	CGCGAA	CTCGAC	TGCGAG	TTCGAA	
AGCGCA	ATCGGT	CCCGCT	GTCGAC	CGCGAG	CTCGAT	TGCGAC	TTCGAG	
AGCGCG	ATCGCA	CCCGCG	GTCGAT	CGCGAC	CTCGGA	TGCGAT	TTCGAC	
AGCGCC	ATCGCG	CCCGCC	GTCGGA	CGCGAT	CTCGGG	TGCGGA	TTCGAT	
AGCGCT	ATCGCC	GGCGGC	GTCGGG	CGCGGA	CTCGGC	TGCGGG	TTCGGA	
AGCGTA	ATCGCT	GGCGGT	GTCGGC	CGCGGG	CTCGGT	TGCGGC	TTCGGG	
AGCGTG	ATCGTA	GGCGCA	GTCGGT	CGCGGC	CTCGCA	TGCGGT	TTCGGC	
AGCGTC	ATCGTG	GGCGCG	GTCGCA	CGCGGT	CTCGCG	TGCGCA	TTCGGT	

The data set tables of all possible permutations of NNNCGNNN (4096) and NNNNCGNNNN (65536) are too big so that are not mentioned here.

5.4 Methods

5.4.1 Effect of flanking sequences on CG distribution

To perform CG distribution studies with respect to the flanking bases, DNA sequences of selected chromosomes of six organisms were downloaded from NCBI.

Downloaded DNA sequence of particular organism was pasted in program file named as fasta_sequence1.

The motif dataset (e.g. NNNNCGNNNN set) was pasted in program file named as input. After pasting the DNA sequence of particular chromosome in fasta_sequence1 file and target motif sequences to be searched in the input file the C++ program named motif search program was compiled and executed using Dev C++.

The output was retrieved from the output file of program and output which is searched motifs was then pasted in microsoft excel file along with their respective frequencies sorted from high to low range to analyse them further.

Consensus sequence was derived for all six organisms under study by making matrices using cut-off values obtained by using Poisson distribution and using these cut-off values high and low methylation groups were assigned.

	-4	-3	-2	-1	C	G	1	2	3	4
G										
A										
T										
C										
R										
Y										
S										
W										
K										
M										
H										
B										
V										
D										

Fig 10: Layout of high and low frequency group matrix

Probability matrixes were derived using following approach:

The entire segmented permutation set was pasted in excel with their respective frequencies four times. In first set using find and replace function G is replaced by 1 while other three bases which are A, T, G with zero. Similarly in second, third and fourth set A, T, C bases are replaced with 1 and rest bases with zero and used "count if" function of MS Excel spreadsheets to get count of each base in respective columns. The value in each respective column i.e. -1,-2,-3,-4 and 1,2,3,4 is further multiplied with their respective frequency and finally at each -1, -2, -3, -4, 1, 2, 3, 4 position sum is calculated resulting in four 1x8 matrices.

	-4	-3	-2	-1	C	G	1	2	3	4
ΣG										
ΣA										
ΣT										
ΣC										

1X4 Matrices of G, A, T, C

Fig 11: Layout of summation matrix obtained after multiplying each position with frequency

Probability Matrix										
	-4	-3	-2	-1	C	G	1	2	3	4
G										
A										
T										
C										

Fig 12: Layout of Probability matrix

By dividing each position of above 4x8 matrix with sum value i.e. ΣG+ΣA+ΣT+ΣC, a probability matrix was obtained which is used further to create PSSM i.e. position specific scoring matrix which was derived from PWM i.e. position weight matrix by dividing each position value of PWM with each respective corresponding value of probability matrix which is obtained earlier. Initially a PWM was obtained by counting the number of G, A, T, C at each of the -1, -2, -3, -4, 1, 2, 3, 4 positions using "count if " function of Microsoft excel for both high and low frequency set. Now a 10 x 4 matrix is created, which looks similar to the one shown in fig 12 and now dividing each position this matrix with sum (G+A+T+C) at each respective position that is -1, -2, -3, -4, 1, 2, 3, 4 PWM is derived. From PSSM consensus sequence was derived. The base with highest value was selected. In order to see the influence of more than one base R, Y, K, M, S, W, B, D, H and V were also considered.

Final probability matrix was made to include other standard nucleotides as well, layout of PWM and PSSM were also same as shown in fig 13.

Final Probability Matrix										
	-4	-3	-2	-1	C	G	1	2	3	4
G										
A										
T										
C										
R										
Y										
S										
W										
K										
M										
H										
B										
V										
D										

Fig 13: Final Probability matrix

5.4.2 CG distribution in methylated and non methylated genomes

All the six organisms DNA sequence were screened for the individual datasets using C++ program with following parameters as mentioned in the table below:

Table 7: Parameters for screening various permutations using C++ program

S.No	Dataset	Number of Permutations	Lenp value
1	NCG	4	3
2	CGN	4	3
3	NCGN	16	4
4	NCGNN	64	5
5	NNCGN	64	5
6	NNCG	16	4
7	CGNN	16	4
8	NNCGNN	256	6
9	NNNCGNNN	4096	8
10	NNNNCGNNNN	65536	10

CHAPTER 6

RESULTS AND DISCUSSION

6. RESULTS

6.1 Effect of flanking sequences on the distribution of CGs in methylated genomes

To perform the distribution studies of CGs with respect to the flanking sequences, the DNA sequences of the organisms under study were taken from NCBI. The effect of flanking sequences on CG distribution was studied in differently methylated genomes which include *Homo sapiens, Pan troglodytes, Mus musculus* and *Danio rerio* representing species which are heavily methylated and *Drosophila melanogester, Caenorhabditis elegans* representative of species with low or even no methylation respectively. The frequency of various permutation of flanking bases was determined using C++ program with conditions as mentioned in the Table 7 in the methods. NCG, CGN, NCGN, NCGNN, NNCGN, NNCG, CGNN, NNCGNN, NNNCGNNN and NNNNCGNNNN (where N represent G, A, T or C) possible permutations have been designed to study effect of these flanking bases on CG distribution. Statistical analysis of NNNNCGNNNN was performed on the data obtained after screening each respective permutation in targeted species DNA and finally consensus sequences were obtained. Comparison of consensus sequences clearly indicates that the data reflect certain key features of CG dinucleotide distribution with respect to the flanking bases.

6.2 Distribution of dinucleotides in various genomes

In order to verify under representation in methylated genomes, CG frequency was determined in the six genomes with varying methylation levels. As CG/CG is mutated to TG/CA, frequency of TG and CA were also determined. For comparison frequencies of GC, GT and AC were also obtained as these three dinucleotides have identical combination of bases for CG, TG and CA respectively. Finally ratio of frequencies of CG & GC, TG & GT and CA & AC were calculated. CG over GC ratio was found to be lowest amongst mammalian genomes which are highly methylated, followed by *Danio rerio*, *Drosophila melanogester* and *Caenorhabditis elegans* genomes which as moderately, sparingly and not methylated respectively. This shows good correspondance between methylation of genomes and CG under-representation in them. The effect was also evident in corresponding higher TG to GT and CA to AC ratios in methylated genomes in general.

Thus comparing CG dinucleotide count across various selected genomes reflects the varying degree of underrepresentation of CG dinucleotides. Comparing CG count in *Homo sapiens, Pan troglodytes, Mus musculus, Danio rerio* as one group with that of *Caenorhabditis elegans* and *Drosophila melanogester* supports the fact of CG dinucleotide underrepresentation in methylated genomes. Comparison of *Homo sapiens* TG and CA content with its CG content clearly reflects the abundance of other two dinucleotides when compared to CG. Similar trend is in seen in *Pan troglodytes, Mus musculus,* and the other two methylated genomes. Interestingly TG and CA dinucleotides count is more when compared to GT, GC and AC dinucleotides in methylated genomes that is very well explained by moderate overrepresentation of TG and CA dinucleotides resulting from loss CGs. Comparing CG/GC ratio across *Homo sapiens, Pan troglodytes, Mus musculus, Danio rerio, Caenorhabditis elegans* and *Drosophila melanogester* reflects a clear picture of CG underrepresentation. CG/GC ratio being 0.99, the highest value obtained in *Caenorhabditis elegans* signifies the fact that non methylated genomes are not CG poor unlike methylated or partially methylated genomes.

Table 8: Comparison of CG, GC, TG, GT, CA, AC, CG/GC, TG/GT & CA/AC count in various organisms

Organism	CG	GC	$\frac{CG}{GC}$	TG	GT	$\frac{TG}{GT}$	CA	AC	$\frac{CA}{AC}$
Homo sapiens	536861	2038794	0.26	3421029	2584606	1.32	3636982	2316685	1.57
Pan troglodytes	543211	2017823	0.25	3420018	2474602	1.26	3547993	2337697	1.64
Mus musculus	741392	3269795	0.23	5633872	4317565	1.3	5834126	3963970	1.47
Danio rerio	585516	590800	0.47	1039318	1065215	1.02	1428894	777293	1.54
Drosophila melanogester	1004265	2137678	0.71	3965646	3875208	1.14	4731113	3063550	1.6
Caenorhabditis elegans	2794081	3931032	0.99	4727559	4139334	0.98	5449324	3407348	1.84

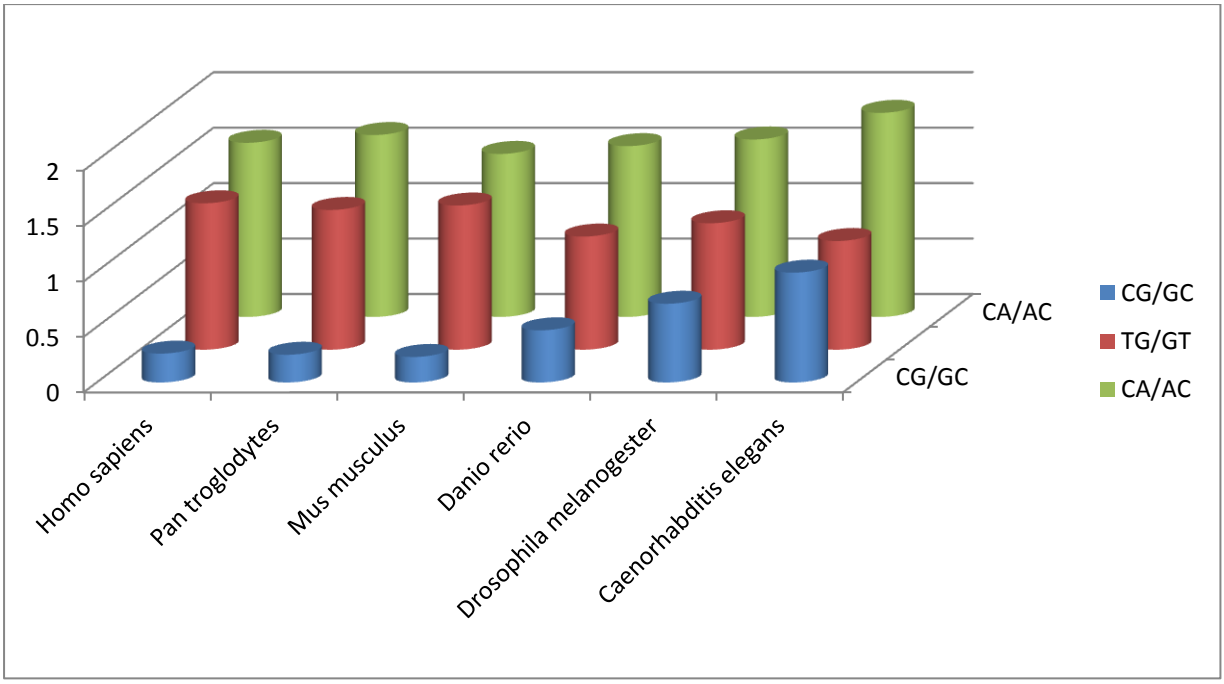


Fig 14: Comparison of CG/GC, TG/GT and CA/AC ratio in different organisms

The graphs clearly indicate that CGs are underrepresented in methylated genomes compared to non methylated or partially methylated genomes. This is in agreement with earlier reports of underrepresentation of CGs and moderate overrepresentation of TGs and CAs in methylated genomes. In order to investigate if there is any effect of flanking bases on occurrence of CGs, frequencies of CGs with differently sized flanks was determined as shown in the table on next page.

Table 9: Sequences of various flanks of NCG, CGN, NCGN, NNCGN, NCGNN, NNCGNN, NNCG, CGNN and NNNNCGNNNN

	H. Sapiens		*M. musculus*		*D. rerio*		*D. melongester*		*C. elegans*	
	High	Low	High	Low	High	Low	High	Low	High	Low
NCG	CCG	TCG	ACG	GCG	ACG	CCG	TCG	ACG	TCG	GCG
CGN	CGG	CGA	CGG	CGC	CGT	CGA	CGA	CGT	CGA	CGC
NCGN	CCGG	TCGC	ACGT	TCGC	ACGT	CCGG	TCGA	ACGG	TCGA	GCGC
NNCGN	CCCGG	CGCGA	CACGT	CGCGA	AACGT	CGCGG	TTCGA	CGCGG	TTCGA	CCCGG
NCGNN	ACGTG	TCGCG	ACGTG	TCGCG	ACGTT	CCGCG	TCGAA	CCGCG	TCGAA	CCGGG
NNCGNN	CACGCC	TACGCG	CACGTG	TACGCG	AACGTT	CGCGTA	AACGAA	TACGGG	TTCGAA	GCCGGG
NNCG	CACG	CGCG	CACG	CGCG	AACG	CGCG	TTCG	CGCG	TTCG	CGCG
CGNN	CGGG	CGCG	CGGG	CGCG	CGTT	CGCG	CGAA	CGCGG	CGAA	CGCGA
NNNCGNNN	CCTCGGCC	CGCCGATA	GTTCGAGG	TTTCGCGA	ACACGCAC	CTACGACG	AAACGAAA	TACCGGTC	TTTCGAAA	GTGCGGGC

The table 9 lists the various CGs with different types of flanks having highest or lowest frequency in differently methylated genomes of five organisms. It can be discerned from the data that there is little common in the sequence of any flank among different organisms. Even highly methylated genomes have different flanking sequences. Moreover inclusion of an additional base in the flanks for the same organism also changes the sequence. In may be inferred from the data that the effect of bases at different positions in the flanks of CGs on it methylation and consequently its frequency is largely combinatorial in nature rather than having individual influence. Further, flanks were expanded up to position -4 to +4 resulting in 65536 permutations. It may be considered a long enough to take care of combinatorial effect of the flanking bases. The table 10 represents the general statistical information on the data of CG, TG and CA flanks in five different genomes including minimum and maximum frequencies observed, total number of CGs, TGs and CAs ($\sum f_i$), mean frequency in each category, standard deviation and coefficient of variance of the frequencies.

Table 10: Statistical analysis of frequency data of NNNNCGNNNN, NNNNTGNNNN, NNNNCANNNN sets

Organism	Data set	Min value	Max value	Σfi	Mean	SD	CV
Homo sapiens	NNNN**CG**NNNN	0	1232	380445	5.8	21.0	362.8
	NNNN**TG**NNNN	0	8125	2552330	38.9	73.1	187.8
	NNNN**CA**NNNN	0	7597	2558762	39.0	71.6	183.4
Mus musculus	NNNN**CG**NNNN	0	956	535771	8.2	12.5	152.3
	NNNN**TG**NNNN	0	50427	4204271	64.2	212.1	330.6
	NNNN**CA**NNNN	0	48251	4191167	64.0	279.2	436.5
Danio rerio	NNNN**CG**NNNN	0	939	27.21	194.884	19.89	10.21
	NNNN**TG**NNNN	0	46372	3252556	49.63	49.63	100
	NNNN**CA**NNNN	0	2103	826846	12.95	27.71	213.98
Drosophila melanogester	NNNN**CG**NNNN	0	956	536935	12.8	8.19	63.98
	NNNN**TG**NNNN	0	8056	3615681	55.17	63.96	115.93
	NNNN**CA**NNNN	0	48251	4244380	204.8	64.76	31.62
Caenorhabditis elegans	NNNN**CG**NNNN	0	1101	459683	17.94	7.01	39.08
	NNNN**TG**NNNN	0	2285	831458	28.20	12.68	44.97
	NNNN**CA**NNNN	0	2157	3267326	195.11	49.86	25.55

The notable observation is that dispersion of frequency of CG flanks (measured as coefficient of variance) is much higher in methylated genomes in comparison with moderately, weakly and unmethylated genomes.

6.3 Distribution pattern of flanking bases around CG dinucleotide at upstream and downstream position

To study any relationship between flanking bases and frequency of different permutations, a subset of data each from highest frequency and lowest frequency was selected based on 0.003 fraction on either tail of Poisson distribution. The subsets were subjected to analysis for determining consensus base at each of the eight positions. The consensus sequences were determined with a procedure mentioned in methods.

Table 11: High Frequency Group Consensus Sequences

Organisms	High Frequency Group Consensus Sequences									
	-4	-3	-2	-1			1	2	3	4
Homo sapiens	T	T	G	Y	C	G	A	T	G	G
Pan troglodytes	C	C	G	C	C	G	G	C	G	G
Mus musculus	C	C	T	T	C	G	A	C	G	G
Danio rerio	A	A	C	G	C	G	A	T	G	T
Drosophila melanogester	A	G	T	G	C	G	G	T	A	T
Caenorhabditis elegans	T	T	C	A	C	G	G	A	A	A

contribution independent evolutionary paths of the different organisms originating from the common ancestors.

If we focus on the consensus sequences obtained for human genomics sequence only, we find that they are distinct for high and low frequency groups. Moreover, in particular in humans and in general in mouse and chimpanzee, the high frequency consensus has largely pyrimidine at -1 and purine at +1 position while low frequency consensus has largely purine at -1 and pyrimidine at +1 position which correspond to low methylation propensity and high methylation propensity flanks reported earlier (Handa and Jeltsch, 2005). This is a good correspondence between the two results from which it may be inferred that the methylation propensity of CGs with different flanks is manifested in their corresponding abundance in the methylated genomes with expected inverse relationship.

Table 12: Low Frequency Group Consensus Sequences

Organisms	Low Frequency Group Consensus Sequences									
	-4	-3	-2	-1			1	2	3	4
Homo sapiens	C	T	A	G	C	G	C	A	C	G
Pan troglodytes	A	T	G	A	C	G	T	G	G	A
Mus musculus	G	C	C	G	C	G	K	C	G	T
Danio rerio	T	T	C	G	C	G	T	C	A	G
Drosophila melanogester	G	C	T	C	C	G	C	C	G	G
Caenorhabditis elegans	C	G	C	T	C	G	G	C	C	G

Table 9 and 10 show the consensus sequences for CG flanks of high frequency and low frequency permutations in genomic sequence samples of six different organisms. There is moderate overlap in the high frequency group consensus sequences among highly methylated genomes of human, mouse and chimpanzee however poor overlap was observed in the low frequency group consensus sequences. The variation amongst consensus sequences of different organisms may be explained by the fact that overall CG distribution and CpG island distribution are different in the different organisms. This variation may also include the contribution independent evolutionary paths of the different organisms originating from the common ancestors.

6.4 DISCUSSION

In mammals not all the CGs are methylated. The best example is CpG islands which have high CG content yet they are not usually methylated. Methylated CGs may undergo spontaneous deamination resulting in mutation (CG/CG ⟶TG/CA) if they escape mismatch repair. This mutation if inherited (via natural selection) leads to loss of CGs in the genome. CpG islands evolved because they have high GC% and CG content and thus do not undergo methylation and remained CG rich. CGs distribution is highly uneven, CpG islands are an example. Since CpG mutation is linked to methylation and methylation propensity has been correlated with flanking base sequences, the proposed hypothesis states that the flanking sequences may have been caused higher or lower rates of mutation of CG/CG to TG/CA. In other words surviving frequency of CGs with different flanks may be affected by these flanking bases themselves.

The interpretation of reported data clearly indicates that the flanking bases do have a profound effect on the distribution of CGs in the methylated genomes. This study reiterates the earlier findings of CG dinucleotide under representation in the methylated genomes and on the other hand corresponding moderate overrepresentation of TG and CA dinucleotide in the methylated genomes which goes well when we talk about the loss of CG dinucleotide on the global level in the methylated genomes. It may be inferred that CG distribution in methylated genomes is designed by at least two factors, evolutionary pressure and sequence dependent factors affecting methylation propensity, for example role of flanking bases in the CG distribution resulting in differential methylation of CG. The CG flank preferred by DNA methyltransferase is more prone to methylation has higher probability of getting mutated and hence expected to have low frequency in the genome unlike other with high frequency. Thus data clearly indicates that CG distribution is non-random in methylated genome. This fact was hitherto based on existence of CpG islands with relatively very high CpG density when compared to rest of the genome.

The present study has attempted to understand the distribution of CGs with respect to the flanking bases in methylated genomes, particularly. Analysing the consensus sequences obtained in the high frequency group clearly goes well with the reported fact that flanks having pyrimidine on -1 position at 5'end and purine on +1 position at the 3' end are disfavoured flanking bases i.e. 5' YCGR 3'. High frequency group contain the consensus of motifs that occurred with high frequency which means either they are not methylated and

hence their number does not decremented in the genome while certain flanks do exist with exceptionally high value which can be explained as these motifs probably are repeats. Looking at the high frequency consensus of *Homo sapiens, Pan troglodytes, Mus musculus, Danio rerio* clearly goes well with the fact that 5'YCGR3' is a disfavoured flank while this cannot be explained in the *C.elegans* where a reverse pattern was observed while in *Drosophila melanogester* there is agreement at +1 position and disagreement at -1 position which explains pretty well as these two organisms are not the one which are representative of methylation family. Low frequency group contain motifs which are occurring with low frequency as their number has been decremented in the genome owing to their methylation. It has been reported that 5'RCGY3'is a favoured flank of methylation enzymes and consensus sequence of the low frequency group of *Homo sapiens, Pan troglodytes, Mus musculus* represents 5'RCGY3'pattern unlike in the *C.elegans* and *Drosophila melanogester* where there is no agreement with above related fact and there is partial agreement with *Danio rerio* consensus which again goes pretty well because *Danio rerio* is not a heavily methylated genome while *C.elegans* and *Drosophila melanogester* are not methylated genomes. No further distinct pattern could be observed which can be explained well as members of high frequency group can be repeat sequences. Furthermore CpG islands don't get methylated and thus are expected to have conflicting frequency of flanks. There are at least five steps after methylation that affect the frequency which include deamination, evading repair and selection of mutation ,amplification of non mutated CGs in tandem and dispersed repeats and horizontal transfer of genetic material which affects CGs flank composition. In spite of above factors there is appreciable overlap observed between the frequency based consensus sequences and those obtained by physiological data based on methylation levels from - 4 to +4 position, in humans (Handa and Jeltsch, 2005) which is very exciting. Overlap of high group consensus is not as good as low frequency group which again goes well with the fact that there is high probability of amplification resulting in repeats as shown in the table below. Having compared frequency data with the methylation data in the high frequency group (poor methylation) 4 out of 8 positions match perfectly, 2 out of 8 partially and 2 out of 8 don't match while in low methylation group consensus (high methylation) 6 out of 8 positions match well while 2 out of 8 positions don't match. Different organisms have different base composition as well as different distribution of permutations of bases. That could be responsible for weak overlap in frequency consensus sequences.

Frequency Methylation Data Consensus

	High Frequency Group Consensus										Low Frequency Group Consensus									
	-4	-3	-2	-1			1	2	3	4	-4	-3	-2	-1			1	2	3	4
Homo sapiens	T	T	G	Y	C	G	A	T	G	G	C	T	A	G	C	G	C	A	C	G
Pan troglodytes	C	C	G	C	C	G	G	C	G	G	A	T	G	A	C	G	T	G	G	A
Mus musculus	C	C	T	T	C	G	A	C	G	G	G	C	C	G	C	G	K	C	G	T
Danio rerio	A	A	C	G	C	G	A	T	G	T	T	T	C	G	C	G	T	C	A	G
Drosophila melanogester	A	G	T	G	C	G	G	T	A	T	G	C	T	C	C	G	C	C	G	G
Caenorhabditis elegans	T	T	C	A	C	G	G	A	A	A	C	G	C	T	C	G	G	C	C	G

Physiological Methylation Data Consensus

	Low methylation Consensus Sequence										High methylation Consensus Sequence									
Homo sapiens	T	G	T	T	C	G	G	T	G	G	C	T	T	G	C	G	C	A	A	G
	T	G	T	C	C	G	G	T	G	G	C	C	T	C	C	G	C	A	A	G
	T	G	G	C	C	G	G	T	G	G	C	C	T	G	C	G	C	A	A	G
	T	G	G	S	C	G	G	T	G	G	C	T	T	G	C	G	C	A	A	C
	T	G	T	G	C	G	G	T	G	S	M	T	G	G	C	G	C	A	T	C
	T	G	T	Y	C	G	G	T	G	C	C	T	K	A	C	G	C	A	A	S

Fig 15: Comparison of frequency methylation data consensus with physiological methylation data consensus

Light green color represents match, yellow color represents partial match in the high frequency group while dark green color represents match in the low frequency group while non colored bases represent mismatch.

This work is novel in away because the studies related to the CG flanking bases influence on the methylation propensity has not been studied at genome level and moreover the work done on sequence preference by DNA methyltransferases has been done in biochemical systems (*in vitro* methylation) or by studying physiological conditions (bisulfate sequencing of genomic DNA of somatic cells) but not in germ lines. We may be bold enough to claim that our data is based on much larger sample that accuracy is expected to be high.

CHAPTER 7

CONCLUSION

7. CONCLUSION

The current study of distribution of CGs in the methylated genomes with respect to their flanking bases was aimed to understand if there exists any influence of flanking bases on the distribution of CGs in the methylated genomes and as the numbers of bases in the flanking sequences are increased how the distribution of CG dinucleotides is being influenced. For this it was required to create a dataset of all possible permutations ranging from NCG, CGN, NCGN, NCGNN, NNCGN, NNCG, CGNN, NNCGNN, NNNCGNNN, NNNNCGNNNN and consensus are derived based on NNNNCGNNNN (65536 permutations). As the dataset size was too big the first task was to code a computer program that can search the individual sequence in the target genome. Existing tools performing similar job could not be employed owing to the magnitude of the number of sequences to be searched. It was clearly indicated in the results which we got that CG dinucleotides are underrepresented. R**CG**Y is a preferred flank base composition which is preferred by methylation enzymes and therefore there is decrement of such flanks in the genome, this observation fits well with the consensus sequences obtained in the low methylation group while Y**CG**R is a disfavoured flank and therefore the flanks with this composition are present in high number in genome as they don't get methylated as this too agrees with the consensus sequences obtained in the high methylation group. There is appreciable overlap observed between the frequency based consensus sequences and those obtained by physiological data based on methylation levels as shown in earlier report (Handa & Jeltsch *et al.*, 2005) which is very exciting. Overlap of high group consensus is not as good as low frequency group which again goes well with the fact that there is high probability of amplification resulting in repeats. It may be concluded that flanking bases do have profound effect on the distribution of CGs in the methylated genomes.

CHAPTER 8

REFERENCES

8. REFERENCES

- Bird, A. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research* **8**, 1499-1504 (1980).
- Cardon, L. *et al*. Pervasive CpG suppression in the animal mitochondrial genomes. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 3799-3803 (1994).
- Coulondre, C. *et al*. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274**, 775-780 (1978).
- Ehrlich, M. Expression of various genes is controlled by DNA methylation during mammalian development. *Journal of Cellular Biochemistry* **88,** 3387-3901 (2003).
- Francisco, A. & Bird, A. Number of CpG islands and genes in human and mouse. *Proceedings of the National Academy of Sciences* **90**, 11995-11999 (1993).
- Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *Journal of Molecular Biology* **196**, 261-282 (1987).
- Gowher, H. & Jeltsch, A. Enzymatic properties of recombinant Dnmt3a DNA methyltransferase from mouse: the enzyme modifies DNA in a non processive manner and also methylates non-CpG (correction of non-CpA) sites. *Journal of Molecular Biology* **309**, 1201–1208 (2001).
- Hagood, J. S. Beyond the Genome: Epigenetic Mechanisms in Lung Remodeling. *American Journal of Physiology* **29**, 177-185 (2014).
- Han, L. *et al*. CpG island density and its correlations with genomic features in mammalian genomes. *Genome biology* **9**, R79 (2008).
- Handa, V. & Jeltsch, A. Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome. *Journal of molecular biology* **348**, 1103-1112 (2005).
- Hermann, H. & Jeltsch, A. Biochemistry and biology of mammalian DNA methyltransferases. *Cellular and Molecular Life Sciences* **61**, 2571-2587 (2004).
- Issa, J. CpG-island methylation in aging and cancer. *Current Topics in Microbiology and Immunology* **249**, 101-118 (2000).

- Jeltsch, A. DNA methylation and molecular enzymology of DNA Methyltransferases. *ChemBioChem* **3**, 274-293 (2002).
- Jones, P. A. & Takai, D. The role of DNA methylation in mammalian epigenetics. *Science* **293**, 1068–1070 (2001).
- Jones, P. A. & Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics* **3,** 415-428 (2002).
- Ollila, J., Lappalainen, I. & Vihinen, M. Sequence specificity in CpG mutation hotspots. *Federation of European Biochemical Societies* **396**, 119-122 (1996).
- Kim, S. *et al*. Predicting DNA methylation susceptibility using CpG flanking sequences. *Proceedings of the National Academy of Sciences* **13,** 315-326 (2008).
- Klose, R. & Bird, A. Genomic DNA methylation: the mark and its mediators. *Trends in Biochemical Sciences* **31**, 89-97 (2006).
- Li, E. Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Reviews Genetics* **3**, 662–673 (2002).
- Okano, M., Xie, S. & Li, E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nature Reviews Genetics* **19**, 219–220 (1998).
- Meehan, R. R. DNA methylation in animal development. *Seminars in Cell and Developmental Biology* **14**, 53–65 (2003).
- Razin, A. DNA Methylation and Gene Expression. *Microbiological Reviews* **55**, 451-458 (1991).
- Riggs, A. D. X inactivation, differentiation, and DNA methylation. *Cytogenetics and Cell Genetic* **14**, 9-25 (1975).
- Rudolf J. & Bird, A. Epigenetic regulation of gene expression. *Nature genetics supplement* **33**, 246-251 (2003).
- Slatkin, M. Epigenetic Inheritance and the Missing Heritability Problem. *Genetics* **182**, 845-850 (2009).
- Han, S., & Brune, A. Histone methylation makes its mark on longevity. *Trends in Cell Biology* **22,** 42- 49 (2012).

- Takai, D. & Jones, P. A. Comprehensive analysis of CpG islands in the human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 3740-3745 (2002).
- Waddington, C. H. Genetic control of wing development in Drosophila. *Journal of Genetics* **41**, 75-139 (1940).
- Weissbach A, N. C., Ward, C. A., Bolden, A. H. The effect of flanking sequences on the de novo methylation of C-G pairs by the human DNA methylase. *Progress in clinical and biological research* **198**, 79-94 (1985).
- Yoder, J. A., Walsh, C. P. & Bestor, H. Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics* **13**, 335-340 (1997).
- Zucker, K. E. *et al.* Purification of human DNA (cytosine-5-) methyltransferase. *Journal of Cellular Biochemistry* **29,** 337–349 (1983).