Comparison of NCGN tetranucleotide frequencies in vertebrate and invertebrate genomes

Submitted in partial fulfilment of the requirements of the

Degree of

MASTER OF SCIENCE IN BIOTECHNOLOGY

Under the guidance of:

Dr. Vikas Handa

Assistant Professor

Submitted By:

Ravinder Singh

Roll No. 300901010

DEPARTMENT OF BIOTECHNOLOGY AND ENVIRONMENTAL SCIENCES

THAPAR UNIVERSITY

Patiala -147004

JUNE, 2011

CANDIDATE'S DECLARATION

I, hereby declare that the work presented in the dissertation entitled **"Comparison of NCGN tetranucleotide frequencies in vertebrate and invertebrate genomes"** in partial fulfilment of the requirement for the award of the degree of Master in Biotechnology, Department of Biotechnology and Environmental Sciences, Thapar University, Patiala, is an authentic record of my own work during the period of six months from January 2011 to June 2011, under the supervision of **Dr. Vikas Handa**, Assistant Professor, Department of Biotechnology & Environmental Sciences, Thapar University. The report has not been submitted for the award of any other degree or certificate in this or any other university.

Place: Patiala

Date: 01.08.2011

Ravinder Singh
(Roll No. 300901010)

CERTIFICATE

This is to certify that the thesis entitled **"Comparison of NCGN tetranucleotide frequencies in vertebrate and invertebrate genomes"** submitted by Ravinder Singh in partial fulfilment of the requirements for the award of Degree of Masters of Science in Biotechnology to Thapar University, Patiala, is a record of student's own work carried out by him 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 or Institute.

Dr. Vikas Handa
Supervisor
DBTES, TU
Patiala

Dr. M. S. Reddy
Head
DBTES, TU
Patiala

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

ACKNOWLEDGEMENT

In pursuit of this academic endeavour, I feel that I have been singularly fortunate because inspiration, guidance, direction, cooperation, love and care - all came in my way in abundance and it seems almost an impossible task for me to acknowledge the same in adequate terms.

Yes, I shall be failing in my duty if I do not record my profound sense of indebtedness and heartfelt gratitude to my guide, **Dr. Vikas Handa**, Assistant Professor, Department of Biotechnology and Environmental Sciences, T.U., Patiala, who guided and inspired me in pursuance of this work. His association with this endeavour of mine will remain a beacon light to me throughout my life.

I am sincerely thankful to **Dr. M. S. Reddy,** Head**,** Department of Biotechnology and Environmental Sciences for his immense concern throughout the project work. I wish to acknowledge the kind help, cooperation and moral support of all the faculty members of DBTES. Their suggestions and constructive criticism were highly useful.

Life at Thapar University, Patiala has been enjoyable with Amit, Avom, Kamal, Sheetal, Sumit and Veni. I thank them all for their great company and support. I also want to thank Bhupinder, Mamta, Nivedita, Prabhjot, Rakhee, Sapna and Vishal for giving me support and unforgettable moments. Above all, I'm grateful to almighty God and my parents for blessing me to complete this work successfully.

Date: 01.08.2011 (Ravinder Singh)

Place: Patiala

DEDICATED

TO

MY PARENTS

CONTENTS

LIST OF ABBREVATIONS

A – adenine

AdoMet - S-adenosyl-L-methionine

C – cytosine

C^5 – carbon at 5^{th} position

Dam - DNA adenine methyltransferase

Dnmt – DNA methyltransferase

Dnmt3a – DNA methyltransferase 3a

Dnmt3b – DNA methyltransferase 3b

Exp_{CpG} - expected frequency of CpG dinucleotide

FCGR - Chaos Game Representation of Frequencies

G – guanine

Gadd45a - Growth arrest and DNA-damage-inducible protein 45 alpha

H3K9ac/H3K14ac - histone H3 lysine 9 and 14 acetylation

HEP - The Human Epigenome Project

IFN – interferon

IL – interleukin

^mA – methylated adenine

MHC - major histocompatibility complex

N – any nucleotide (A or C or G or T)

N^4 – nitrogen at 4^{th} position

N^6 – nitrogen at 6^{th} position

Obs_{cpg} - observed frequency of CpG dinucleotide

ObsCpg/ExpCpG – ratio of observed frequency of Cpg and expected frequency of CpG dinucleotide

P (N) – probability of (N)

R – purine

RIV – relative initial velocity

SNP – single nucleotide polymorphism

T – thymine

Y – pyrimidine

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

DNA methylation, a DNA modification which occurs at the N^6 position of adenine and the N^4 and C^5 positions of cytosine in prokaryotes while only C^5 methylation is found in higher eukaryotes. It is an epigenetic mechanism which plays critical role in gene silencing, X chromosome inactivation, imprinting and silencing of intragenomic parasites. In mammals DNA cytosines are methylated by DNA methyltransferases in CpG dinucleotide context and the flanking regions of CpG dinucleotides affect the activity of the enzymes. CG methylation results in mutation (CG→TG/CA) which is responsible for CG dinucleotide suppression in vertebrate genome. In this study, we have investigated if flanking base preference of DNA methyltransferases is reflected in frequency distribution of NCGN frequencies of different genomes as a result of mutations of NCGN. We have compared tetranucleotide frequencies in randomly selected representative genomic sequences of different organisms with initial relative velocities of DNA methylation by Dnmt3a.

CHAPTER 1

INTRODUCTION

1. INTRODUCTION

Epigenetics is the study of heritable changes in phenotype or gene expression caused by mechanisms other than changes in the related DNA sequence. The epigenetic modification includes covalent modifications of histone tails (acetylation, methylation, phosphorylation, ubiquitination) or DNA methylation. Both the modifications suppress gene expression without altering the sequence of the silenced genes.

DNA methylation is essential for normal development and is associated with a number of key processes including genomic imprinting, X-chromosome inactivation, suppression of repetitive elements, and carcinogenesis. It occurs at the N^6 position of adenine residues and the N^4 and C^5 positions of cytosine residues in prokaryotes while only C^5 methylation is found in higher eukaryotes. Methylation is also a part of the restriction modification system in many bacteria. A methylase enzyme recognizes a specific sequence in genome and methylates one of the bases in or near that sequence. Foreign DNAs which are not methylated in this manner when introduced into the cell are cleaved by sequence-specific restriction enzymes. Thus DNA methylation protects bacteria from infection by bacteriophages by degrading their genomic DNA (Kruger and Bickle, 1983).

E. coli DNA adenine methyltransferase (Dam) is an enzyme that plays several key roles in bacterial processes, including mismatch repair, the timing of DNA replication, and gene expression. The recognition site for *E. coli* Dam is GATC, as the methylation occurs at the N^6 position of the adenine in this sequence (G^mATC). The status of GATC sites in the *E. coli* genome changes from fully methylated to hemi-methylated. This is because adenine introduced into the new DNA strand is un-methylated. Re-methylation occurs within two to four seconds, during which time replication errors in the new strand are repaired. It has been shown that altering Dam activity in bacteria results in increased spontaneous mutation rate (Palmer and Marinus, 1994).

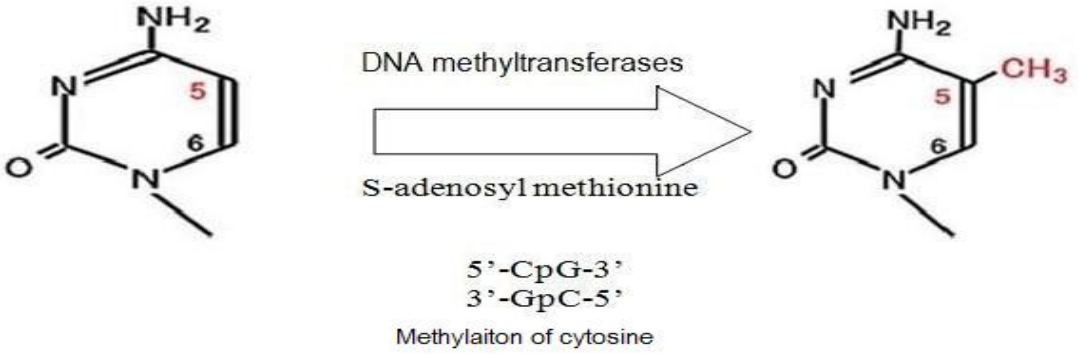

Figure 01: Methylation of cytosine

DNA methylation is carried out by enzymes called DNA methyltransferases. All DNA methyltransferases use S-adenosyl-L-methionine (AdoMet) as the source of the methyl group which is transferred to the DNA bases. Dnmt1, Dnmt3a and Dnmt3b are major DNA methyltransferases found in mammals that methylate cytosine in 5'-CpG-3' context. Nearly 60% to 90% of all CpGs are methylated in mammals. Methylated cytosines spontaneously deaminate to form thymines thus CpG dinucleotides steadily mutate to TpG/CpA dinucleotides, which is evidenced by the under-representation of CpG dinucleotides in the human genome (they occur at only 21% of the expected frequency) and corresponding over representation of TpG and CpA (Bird, 1980). On the other hand, spontaneous deamination of un-methylated C residues gives rise to U residues, a mismatch that is quickly recognized and repaired by the cell.

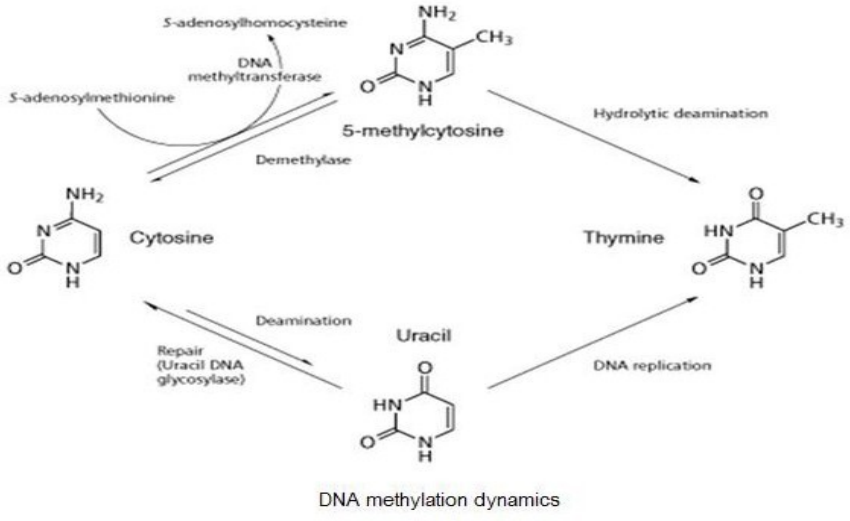

Figure 02: DNA methylation dynamics

Though majority of CpGs are methylated in mammalian genome, some clustered CpGs are found un-methylated in certain GC rich regions. These regions are known as CpG islands and are characterized by higher observed/expected frequency of CpGs when compared to global average in addition to GC rich sequences. As per the latest and more accurate description, CpG island is defined as a region with at least 500 bp with a GC content >55% and observed CpG/expected CpG value of >0.65. The CpG islands are usually not methylated and are associated with the 5' regions of nearly half of the genes (Bird, 1980; Takai and Jones, 2001).

In many diseases such as cancer, CpG islands in the promoter region of related genes acquire abnormal hyper-methylation, which results in transcriptional silencing that can be inherited by daughter cells following cell division. Alterations of DNA methylation have been recognized as an important component of cancer development. Hypo-methylation is linked to chromosomal instability and loss of imprinting, whereas hyper-methylation is associated with promoters and can give rise to gene silencing (eg tumour suppressor), but induced DNA methylation changes might be a target for epigenetic therapy. For example inhibitors of enzymes controlling epigenetic modifications, specifically DNA methyltransferases and histone de-acetylases may be used to produce anti-tumorigenic effects for some malignancies (Novik *et al*, 2002). Two groups of drugs are being developed in epigenetic therapy are one which inhibits DNA methyltransferases (DNMTs) useful in treatment of cancer where hyper-methylation of tumor suppressor genes is known to lead to silencing of these genes. Other group of drugs inhibits histone de-acetylases (HDACs) resulting in the accumulation of acetylated histones which are thought to mediate the anticancer effects of these drugs (Peedicayil, 2005). Since many tumor suppressor genes are silenced by DNA methylation during carcinogenesis, there have been attempts to re-express these genes by inhibiting the DNMTs. 5-Aza-2'-deoxycytidine (decitabine) is a nucleoside analog that inhibits DNMTs by trapping them in a covalent complex on DNA by preventing the β-elimination step of catalysis, thus resulting in degradation of the enzymes (Hagemann *et al*, 2011).

DNA methylation may affect the transcription of genes in two ways. The methylation of DNA itself may physically impede the binding of regulatory proteins to the binding sites.

The other mechanism involves methyl-CpG-binding domain proteins (MBDs) which specifically bind to methylated DNA. MBD proteins then recruit additional proteins to the locus, such as histone de-acetylases and other chromatin remodeling proteins that can modify histones, thereby forming compact, inactive chromatin, termed as silent chromatin. This link between DNA methylation and chromatin structure is very important. In particular, loss of methyl-CpG-binding protein 2 (MeCP2) has been implicated in Rett syndrome which is a neurodevelopmental disorder of the grey matter of the brain and methyl-CpG-binding domain protein 2 (MBD2) mediates the transcriptional silencing of hyper-methylated genes in cancer.

DNA methyltransferases

In mammalian cells, DNA methylation occurs at the C^5 position of CpG dinucleotides and is carried out by two general classes of enzymatic activities – maintenance methylation and *de novo* methylation. Maintenance methylation activity is necessary to preserve DNA methylation after every cellular DNA replication cycle. Semi-conservative DNA replication results in loss of methylation in the newly synthesized daughter strand of DNA leading to hemi-methylated state of DNA. If the hemi-methylated DNA undergoes another round of replication then one of the daughter molecules will be hemi methylated while the other will

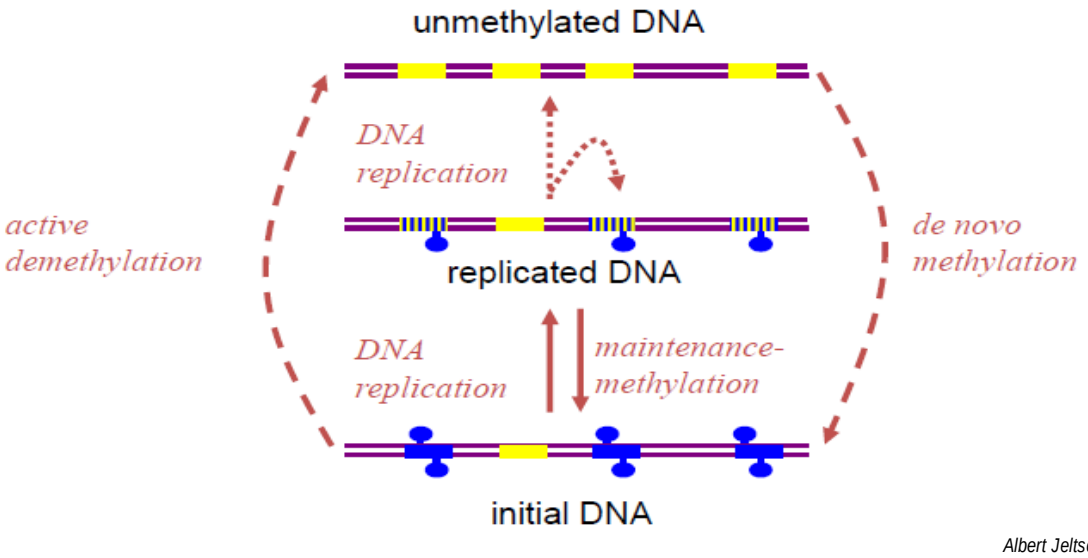

Figure 03: DNA methylation dynamics

be un-methylated and over the time, would lead to passive de-methylation.

Dnmt1 is the proposed maintenance methyltransferase that is responsible for copying DNA methylation patterns to the daughter strands during DNA replication. Dnmt1 exhibits several fold higher preference for hemi-methylated CpG sites in comparison to un-methylated sites. Thus Dnmt1 methylates the hemi-methylated daughter DNA molecule soon after replication. Mouse models with both copies of Dnmt1 knocked out leads to death of the embryo at approximately day 9, due to the requirement of Dnmt1 activity for development in mammalian cells.

It is thought that Dnmt3a and Dnmt3b are the *de novo* methyltransferases as they do not discriminate between hemi- methylated and un-methylated CpGs. The two enzymes play important role in setting up DNA methylation patterns early in development. When expression profiles of Dnmt3a and Dnmt3b were analyzed it was found that later is over-expressed in tumors in comparison to normal cells which proves important role played by Dnmt3b in tumorigenesis. Dnmt3L is a protein that is homologous to the other Dnmt3s but has no catalytic activity. Instead, Dnmt3L assists the *de novo* methyltransferases mainly Dnmt3a by increasing their ability to bind to DNA and stimulating their activity. It is expressed during gametogenesis and embryonic stages. Finally, Dnmt2 has been identified as a DNA methyltransferase homolog, containing conserved sequence motifs common to all DNA methyltransferases; however, Dnmt2 displays very low methylation activity on DNA and found in all eukaryotes that show methylation as well as those that lack detectable methylation (Jeltsch *et al*, 2004).

DNA methylation is also associated with certain genetic disorders. Rett syndrome is caused by mutations in the gene MECP2 encoding methyl-CpG-binding protein-2 (Amir *et al*, 1999). It is a neurodevelopmental disorder of the grey matter of the brain that affects females more commonly than male. Another disorder is or Immunodeficiency, Centromere instability and Facial anomalies syndrome (ICF syndrome) - a very rare autosomal recessive immune disorder caused by a mutation in the DNA-methyltransferase-3b (Dnmt3b) gene, located on chromosome 20q11.2. It is characterized by variable reductions in serum immunoglobulin levels which make most ICF patients prone to infectious diseases before

adulthood. ICF syndrome patients exhibit facial anomalies which include hyper-telorism, low-set ears, epicanthal folds and macroglossia (Jiang *et al*, 2005).

DNA methylation dynamics

It is reported in mice germ cells and early embryonic stages, there is a wave of de-methylation which de-methylates complete genome and re-methylation is initiated by *de novo* methyltransferases. After fertilization most of the paternal genome is rapidly de-methylated before replication starts. On the other hand, the maternal genome undergoes apparently passive, replication dependent de-methylation during subsequent cleavage divisions. After implantation, a wave of global *de novo* methylation re-establishes the DNA methylation patterns that will be maintained in somatic cells. Genome wide de-methylation also occur in primordial germ cells around embryonic days 11.5-12.5 and then *de novo* methylation establishes a gamete specific methylation pattern which is different for egg and sperm. Besides it, gene specific *de novo* methylation and de-methylation also occur during lineage-specific differentiation such as during differentiation of hematopoietic progenitors (Zhao-xia Chen and Arthur D. Riggs, 2005).

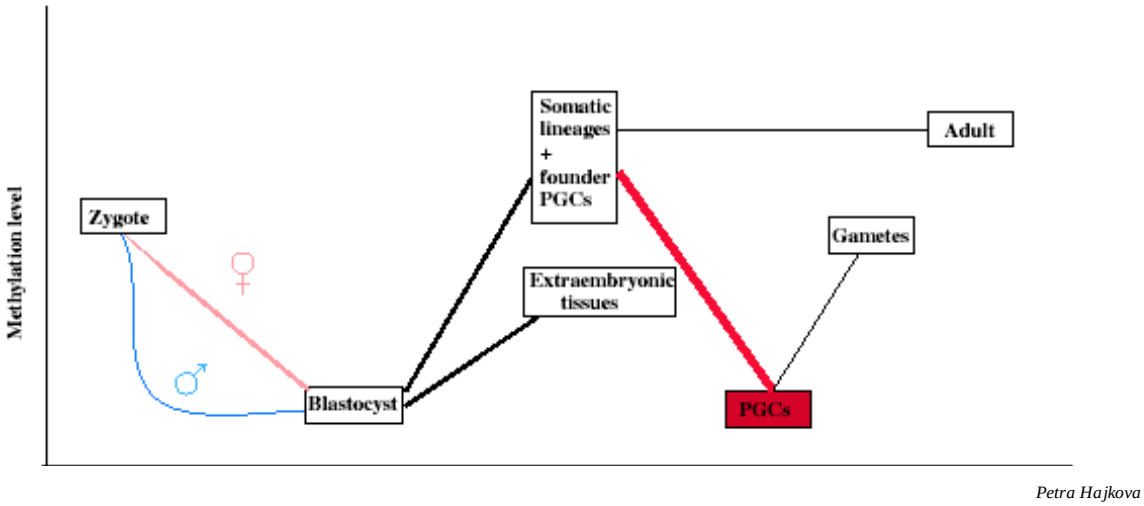

Figure 04: Gamete specific methylation by *de novo* methyltransferases

Gadd45a (Growth arrest and DNA-damage-inducible protein 45 alpha) plays an important role in DNA de-methylation. It is a nuclear protein involved in maintenance of genomic stability, DNA repair and suppression of cell growth. Over expression of Gadd45a activates

methylation-silenced reporter plasmids and promotes global DNA de-methylation. Gadd45a plays important role in active de-methylation of DNA via nucleotide excision repair mechanism. It relieves epigenetic gene silencing by promoting DNA repair which erases methylation marks. Gadd45a knockdown silences gene expression and leads to DNA hyper-methylation. It was reported that Gadd45a is specifically recruited to the site of de-methylation during active de-methylation of Oct4 in *Xenopus laevis* oocytes (Barreto *et al*, 2007).

CHAPTER 2

REVIEW OF LITERATURE

2. REVIEW OF LITERATURE

2.1 Epigenetics

Epigenetics may be defined as "study of the changes in gene expression that are mitotically and/or meiotically heritable and do not involve a change in the DNA sequence" (Wu and Morris, 2001). The two main components of Epigenetic effects include DNA methylation and Histone modifications that gene regulation via chromatin remodeling complexes. DNA methylation leads to gene silencing via binding of methyl cytosine binding proteins and their interaction with histone de-acetylases. Histone modification is important mechanism of epigenetic regulation of gene expression. DNA methylation is correlated with these modifications. There is a strong correlation between presence of histone H3 lysine 9 and 14 acetylation (H3K9ac/H3K14ac) and histone H3 lysine 4 trimethylation (H3K4me3) and absence of DNA methylation which are markers of gene activation while histone de-acetylation and H3K9 and H3K27 methylation marks gene silencing (Jaenisch and Bird, 2003 & Zhang *et al*, 2009). There is a dynamic transition between different chromatin states which are influenced by de-/methylation of DNA and various modifications of histones (mainly involving de-/acetylation and methylation of histone 3). Epigenetic mechanisms are responsible for several phenomena including: X-inactivation in which the random silencing of one of the X chromosomes in every normal somatic cell of female mammals takes place (Park and Kuroda, 2001) and genomic imprinting which is the expression or repression of certain genes according to their parental origin (Ferguson-Smith and Surani, 2001).

Recent advances in epigenetics include studying genome wide methylation phenomena termed to as epigenomics. Epigenomics is the study of the effects of chromatin structure including the higher order of chromatin folding and attachment to the nuclear matrix, packaging of DNA around nucleosomes, covalent modifications of histone tails (acetylation, methylation, phosphorylation, ubiquitination) and DNA methylation (Murrell *et al*, 2005).

2.2 DNA Methylation

DNA methylation is carried out by enzymes called DNA methyltransferases on cytosine and adenine bases. All DNA methyltransferases use *S-adenosyl*-L-methionine (AdoMet) as the source of the methyl group being transferred to the DNA bases. Prokaryotic

cytosine and adenine methylation can influence gene transcription, affect cell viability, play important role in mismatch repair of DNA and also serve the restriction-modification systems that protect the bacterial host DNA from cleavage by specific endonucleases (Kahng and Shapiro, 2001). Only cytosines are methylated at postion 5 in eukaryotes (mainly in vertebrate genomes). Dnmt1, Dnmt3a and Dnmt3b are major DNA methyltransferases found in mammals and methylate cytosine at position 5 in 5'-CG-3' context (Santi *et al*, 1983; Bird, 1992). The majority of 5-methylcytosine (5-MeC) occurs in the context of the dinucleotide CpG, although CpNG, CC(a/t)GG, CpA and CpT can also be methylated at a very low frequency (Clark *et al*, 1995; Woodcock *et al*, 1997; Lorincz and Groudine, 2001). In plants and filamentous fungi, cytosine methylation at non-CpG sites occurs much more frequently than in animals (Martienssen and Colot, 2001).

2.3 CpG islands

Though most of the vertebrate genome CpGs are methylated, there are some GC rich regions in the vertebrate genomes that remain un-methylated usually. These regions have higher CpG frequency in comparison to the expected frequency and they are known as CpG islands. CpG islands are located in promoter and exonic regions of nearly half of the mammalian genes and play important role in tissue specific gene expression and carcinogenesis. The first large-scale computational analysis of CpG islands using vertebrate sequences in GenBank concluded in defining a CpG island as a 200-bp or longer region of DNA with a high G+C content (greater than 50%) and $observed_{CpG}/expected_{CpG}$ ratio of greater or equal to 0.6 value (Gardiner-Garden and Frommer, 1986). Later study based on genomic analysis of Human, Arabidopsis and some non vertebrate genomes redefined CpG islands with more stringent values of the above mentioned parameters i.e. regions of DNA of greater than 500 bp with a G+C equal to or greater than 55% and $observed_{CpG}/expected_{CpG}$ of 0.65. In addition the gap between two adjacent CpG islands should be at least 100 bp. The new definition reduced the number of CpG islands in the genomes but on the other hand the newly defined islands were more likely to be associated with the 5'regions of genes and this definition excluded most *Alu*-repetitive elements (Takai and Jones, 2002). A correlation between CpG island density and few genomic features such as number of chromosome pairs, chromosome size and recombination rate was found in mammalian genomes. Increase in

number of chromosomes increases the rate of recombination which elevates G+C content preventing the loss of CpG islands and maintaining their density in genome (Leng *et al*, 2008).

2.4 CpG dinucleotides

CpG dinucleotide is a short palindromic sequence that is recognized and methylated by methyltransferases in vertebrate genomes. They are non-randomly distributed across the mammalian genomes and occur less frequently than their expected frequency. This CpG deficiency in genomes is due to methylation of cytosine at fifth position. This is evident from the fact that highly methylated genomes such as of vertebrates have significant deficiency of CpGs whereas poorly methylated genomes (i.e. insect type) display no significant CpG deficiency. Cytosine gets methylated enzymatically to 5-methylcytosine which has tendency to undergo spontaneous de-amination leading to its conversion into thymine (5-methylurasil) frequently and thus CpG mutates to TpG and CpA. Since CpG is palindromic sequence, its mutation causes loss of two CpG sites with gain of TpG and CpA. Therefore, genomes with low CpG sites (vertebrates) are rich in TpG and CpA frequencies and genomes with high CpG sites (insects) have normal TpG and CpA frequencies. The process seems to have repeated in entire expanse of genomes (except CpG islands probably) all through the course of evolution. This is the basis of under-representation of CpG dinucleotides in vertebrate genomes and shows that excess of TpG plus CpA in vertebrates is inversely proportional to CpG plus CpG in DNA sequence (Bird, 1987). CpG dinucleotides are hot spot for mutation. In order to investigate the effect of these hot spots on SNPs in humans, CpGs were analyzed considering CpGs in CpG islands and outside the CpG islands to check whether distribution of sequence surrounding the SNPs is affected by mutagenesis or not. The distribution of polymorphic alleles (C/G) at CpGs in CpG islands is also significantly different from that in non-island regions. CpG is the most abundant dinucleotide at polymorphic positions and highest observed/expected ratio of any sequence at SNP sites. It was found that CpGs outside the CpG islands were abundant than expected frequency in polymorphic sites where mutation takes place. CpGs present in CpG islands' polymorphic sites were under represented than their expected frequency. This analysis showed that CpG islands are devoid of methylation of Cytosine and mutation of cytosine plays important role in generation of SNPs and suggests

that there is not significant variation (polymorphic) at CpGs in CpG islands (Tomso and Bell, 2003; Zhao and Zhang, 2006). Theoretically dinucleotide should have the probability of nearly equal frequency in genome of any organism i.e. there expected frequencies should follow normal distribution but CpG dinucleotides are present in DNA less frequently than there expected base composition due to high tendency of mutation. This non- uniformity of frequencies of dinucleotides affects various biological phenomena. This includes effect of dinucleotide relative abundance deviations affecting DNA duplex curvature, supercoiling and other high order DNA features. The A, B and Z forms of DNA appear to depend upon base sequences. For example, the dinucleotide CC or GG favors an A form helix, whereas AA or TT exclusively adopts a B form helix. Poly CG can be best accommodated by Z form of DNA (Karlin *et al*, 1994). DNA methylation is linked with immunogenic responses also. Un-methylated CpGs in DNA sequence are immunogenic in mammals. These CpG sites stimulate B cells to produce IL-6 and IL-12, CD41 T cells to produce IL-6 and IFN-γ, and NK cells to produce IFN-γ (Kreig *et al*, 2002 and Rui *et al*, 2003).

2.5 Human epigenome project (HEP)

Genome level analysis gives much deeper and generalized insight into various biological phenomena as it is evident in improved definition of CpG islands by Takai and Jones. DNA methylation plays important role in gene expression and cellular development. To study the effect of DNA methylation on global expression level, on similar lines attempt is being made to determine DNA methylation information at genomic level. This has given rise to The Human Epigenome Project (HEP) - a joint effort by an international collaboration, which was established in 1999 with the aim to identify, catalogue and interpret genome-wide DNA methylation patterns and profiles of all human genes in all major tissues (www.epigenome.org). For the pilot study DNA methylation profiling was carried out on the major histocompatibility complex (MHC), a gene rich region on chromosome 6 in human genome. High-throughput methylation was analyzed by bisulfate sequencing. DNA methylation levels within regulatory, exonic and intronic regions associated with 90 genes (i.e. 70% of all expressed genes within the MHC) were analyzed in seven human tissues—adipose, brain, breast, liver, lung, muscle and prostate—with multiple samples from different individuals. For the DNA methylation profiling of the human MHC, regions with potential

regulatory functions and CpGs dense regions of a gene were selected for sequencing (Murell *et al*, 2005). In continuation of this attempt a similar study was carried out on human chromosome 21. An analysis of the DNA methylation pattern of 297 amplicons from 190 gene promoters was performed using bisulfite conversion, subcloning and sequencing in five cell types- human peripheral blood (mainly leucocytes), fibroblast, the human embryo kidney cell line HEK293, the human hepatocellular liver carcinoma cell line HepG2 and fibroblast cells derived from a patient with Down syndrome (trisomic 21) and methylation levels of all cell types were found out. DNA methylation levels in individual cells in one tissue are very similar and there are differences in methylation levels in different cell types. There is a bimodal distribution of methylation in the amplicons under investigation. There was differential DNA methylation within different parts of single CpG island as methylation of amplicons gradually decreased when approaching the transcription start site of the respective gene both from upstream and downstream. There was an inverse correlation of DNA methylation with gene expression. Also, correlation was there between absence of DNA methylation and presence of histone modification such as H3K4me3, H3K9ac and H3K14ac which are markers for active state of chromatin. This experimental study provided a huge knowledge about relationship of DNA methylation and genetic and epigenetic processes (Zhang *et al*, 2009).

2.6 Effects of flanking sequences on DNA methylation

Dnmt3a and Dnmt3b have a considerable influence on the methylation pattern of human genomic DNA. These *de novo* methyltransferases play important role in setting the pattern of methylation in mammalian genome. Dnmt1 has ability to discriminate hemi-methylated DNA produced after replication but Dnmt3a and Dnmt3b enzymes do not have such ability and methylate un-methylated DNA. However they seem to play important role in establishing *de novo* methylation patterns especially during gametogenesis and early embryonic development. Flanking sequence around CpGs have been reported to influences the catalytic activity of the DNA methyltransferases. The substrate of DNA methyltransferases, CpG is very short sequences and it is understandable that flanking bases would affect the interaction of DNA with the enzyme(s). The flanking sequence preferences of Dnmt3a were first detected by Lin *et al*. There is a strong preference for a CG site flanked

by pyrimidine bases (YNCGY) that is pyrimidines at the -2 position upstream and the +1 position downstream of CG. This enzyme does not have any preference for hemi-methylated DNA unlike Dnmt1 (Lin *et al, 2001).* Owing to the significant preference of Dnmt3a and Dnmt3b for CpGs based on flanking bases around them, there could be a bias in selecting CpGs for methylation. Methylation kinetics experiments based on methylation of double stranded DNA oligonucleotides having one CpG site by Dnmt3a revealed a significant difference in relative initial velocity between the preferred (ACGC/GCGT) and disfavored (TCGG/CCGA) flanking base-pairs. This may be inferred that purines and pyrimidines are preferred at -1 and +1 while they are disfavored at +1 and -1 positions. Dnmt3a and Dnmt3b have very similar flanking sequence preferences. Based on computational analysis as well as confirmation by methylation kinetics experiments, the *de novo* mammalian DNA MTases have been found to be exhibiting profound preference for bases flanking a CG site (5'-CTTACGCAAG- 3' consensus sequence) and show almost no activity for some flanking sequences (5'-TGTTCGGTGG-3' consensus sequence). That means presence of specific bases at upstream and downstream of CpG sites affect the enzymatic activity of *de novo* methyltransferases. This preference can lead to set a methylation pattern in the mammalian genome which is inherited to next generation. In addition it has been found that AT-rich flanks are preferred over GC-rich ones. Overall there is a >500-fold difference in the methylation rates of the preferred and disfavored sequences. The methylation pattern of human DNA is suggested due to the flanking sequence preferences of the *de novo* DNA methyltransferases and that flanking sequence preferences could be involved in the origin of CpG islands. Such preferences may have other implications also as is evident from the following correlation. The CpG containing DNA sequences causing high immunogenicity have the lowest probability to be un-methylated in the human DNA on the basis of flanking bases, which minimizes the risk of an auto-immune response generated from self DNA (Kreig *et al*, 2002 and Rui *et al*, 2003). This indicates a possibility of co-evolution of enzymatic activity of the *de novo* methyltransferases and effect of CpG flanks on their immunogenic effects. Similar kind of effect may be extrapolated to be having impact on reshaping the genome of mammals. The basic idea is to investigate if preferred flanks which have higher chances of getting methylated and subsequently mutated (CG/CG → TG/CA)

have lower frequency in the genomes when compared to those which are not preferred by *de novo* methyltransferases.

CHAPTER 3
OBJECTIVES

3. OBJECTIVES

1. To verify suppression of CpG dinucleotide frequencies in different organisms.
2. Comparison of NCGN tetranucleotide frequencies in genomes of different organisms.
3. Determine the correlation between NCGN tetranucleotide frequencies on genomes of different organisms with initial relative velocities of DNA methylation by Dnmt3a to investigate the effect of differential methylation of CpGs with different flanks on their frequencies in the genome.
4. Investigate the relative abundance of decanucleotides containing CpG flanked by very highly preferred, highly preferred, moderately preferred and poorly preferred bases in methylated genomes.

CHAPTER 4

MATERIALS AND METHODS

4. MATERIALS AND METHODS

4.1 Selection of sequences

Following organisms were selected for determining below mentioned oligonucleotide frequencies (Table 10)

1. *Homo sapiens*
2. *Mus musculus* strain C57BL/6J
3. *Danio rerio* strain Tuebingen
4. *Arabidopsis thaliana*
5. *Drosophila melanogaster*
6. *Caenorhabditis elegans*
7. *Plasmodium falciparum* 3D7
8. *Saccharomyces cerevisiae* S288c
9. *Escherichia coli* str. K-12

Owing to the limitation of length of sequence to be analyzed for tetranucleotide frequencies (100,000 bp), 10 such sequences were selected at random from each organism genome. Thus a total of 10^6 bp sequence was analyzed for each organism. The details of the sequences are as following:

1. *Homo sapiens*

Table 01: DNA sequences of *Homo sapiens*

S. No.	Chromosome	Accession no.	From	To
1	1	NT_077402.2	1	100000
2	2	NT_034508.2	1	100000
3	4	NT_037622.5	1	100000
4	6	NT_007299.13	1	100000
5	12	NT_024477.14	1	100000
6	13	NT_027140.6	100001	200000
7	14	NT_026437.12	800001	900000
8	15	NT_077631.1	100001	200000
9	19	NT_011295.11	15600001	15700000
10	X	NT_167196.1	500001	600000

2. *Mus musculus*

Table 02: DNA sequence of *Mus musculus*

S. No.	Chromosome	Accession no.	From	To
1	1	NT_039169.7	200001	300000
2	2	NT_166284.1	1	100000
3	5	NT_039306.7	300001	400000
4	6	NT_039340.7	1	100000
5	9	NT_039471.7	500001	600000
6	12	NT_039548.7	600001	700000
7	15	NT_039617.7	400001	500000
8	17	NT_039636.7	1	100000
9	19	NT_082868.6	200001	300000
10	X	NT_039699.7	1	100000

3. *Danio rerio*

Table 03: DNA sequences of *Danio rerio*

S. No.	Chromosome	Accession no.	From	To
1	1	NW_001884377.2	200001	300000
2	3	NW_00303929.1	300001	400000
3	4	NW_003039371.1	300001	400000
4	8	NW_003039725.1	1	100000
5	9	NW_001877082.2	1	100000
6	10	NW_001877101.2	400001	500000
7	12	NW_003040014.1	1	100000
8	16	NW_003040364.1	100001	200000
9	22	NW_001878303.2	400001	500000
10	24	NW_003040888.1	200001	300000

4. *Saccharomyces cerevisiae* S288c

Table 04: DNA sequences of *Saccharomyces cerevisiae* S288c

S. No.	Chromosome	Accession no.	From	To
1	1	NC_001133.8	1	100000
2	2	NC_001134.7	200001	300000
3	4	NC_001136.8	300001	400000
4	5	NC_001137.2	1	100000
5	8	NC_001140.5	400001	500000
6	9	NC_001141.1	200001	300000
7	11	NC_001143.7	300001	400000
8	14	NC_001146.6	1	100000
9	15	NC_001147.5	500001	600000
10	16	NC_001148.3	200001	300000

5. *Drosophila melanogaster*

Table 05: DNA sequences of *Drosophila melanogaster*

S. No.	Chromosome	Accession no.	From	To
1	X	NC_004354.3	400001	500000
2	X	NC_004354.3	20400001	20500000
3	2L	NC_033779.4	300001	400000
4	2L	NC_033779.4	22100001	22200000
5	2R	NT_033778.3	100001	200000
6	2R	NT_033778.3	15100001	15200000
7	3L	NT_037436.3	200001	300000
8	3L	NT_037436.3	4100001	4200000
9	3R	NT_033777.2	1100001	1200000
10	4	NC_004353.3	400001	500000

6. *Caenorhabditis elegans*

Table 06: DNA sequences of *Caenorhabditis elegans*

S. No.	Chromosome	Accession no.	From	To
1	I	NC_003279.6	400001	500000
2	II	NC_003280.7	1	100000
3	II	NC_003280.7	12100001	1220000
4	III	NC_003281.8	300001	400000
5	IV	NC_003282.5	1	100000
6	IV	NC_003282.5	12100001	12200000
7	V	NC_003283.8	500001	600000
8	V	NC_003283.8	1100001	1200000
9	X	NC_003284.7	1	100000
10	X	NC_003284.7	700001	800000

7. *Plasmodium falciparum*

Table 07: DNA sequences of *Plasmodium falciparum*

S. No.	Chromosome	Accession no.	From	To
1	1	NC_004325.1	300001	400000
2	2	NC_000910.2	200001	300000
3	3	NC_000521.3	1	100000
4	4	NC_004318.1	500001	600000
5	5	NC_004326.1	400001	500000
6	6	NC_004327.2	1100001	1200000
7	7	NC_004328.1	500001	600000
8	8	NC_004329.1	100001	200000
9	9	NC_004330.1	1200001	1300000
10	13	NC_004331.1	300001	400000

8. *Escherichia coli* str. K-12

Table 08: DNA sequences of *Escherichia coli* str. K-12

S. No.	Chromosome	Accession no.	From	To
1	-	NC_010473.1	1	100000
2	-	NC_010473.1	200001	300000
3	-	NC_010473.1	300001	400000
4	-	NC_010473.1	400001	500000
5	-	NC_010473.1	500001	600000
6	-	NC_010473.1	600001	700000
7	-	NC_010473.1	700001	800000
8	-	NC_010473.1	800001	900000
9	-	NC_010473.1	900001	1000000
10	-	NC_010473.1	1100001	1200000

9. *Arabidopsis thaliana*

Table 09: DNA sequences of *Arabidopsis thaliana*

S. No.	Chromosome	Accession no.	From	To
1	1	NC_003070.9	1900001	2000000
2	1	NC_003071.7	2100001	2200000
3	2	NC_003071.7	100001	200000
4	2	NC_003074.8	200001	300000
5	3	NC_003074.8	300001	400000
6	3	NC_003075.7	700001	800000
7	4	NC_003075.7	400001	500000
8	4	NC_003076.8	1100001	1200000
9	5	NC_003076.8	2100001	2200000
10	5	NC_003076.8	1100001	1200000

4.2 List of oligonucleotides for which frequency in different organisms was determined

Table 10: List of oligonucleotides

N	NN	NNNN	GNNC	Decamers	Complementary decamers
A	AA	ACGA	GAAC	CTTACGCAAG	CTTGCGTAAG
C	AC	ACGC	GACC	CTTATGCAAG	CTTGTGTAAG
G	AG	ACGG	GAGC	CTTACACAAG	CTTGCATAAG
T	AT	ACGT	GATC	TGTACGCTGG	CCAGCGTACA
	CA	CCGA	GCAC	TGTATGCTGG	CCAGTGTACA
	CC	CCGC	GCCC	TGTACACTGG	CCAGCATACA
	CG	CCGG	GCGC	CTTTCGGAAG	CTTCCGAAAG
	CT	CCGT	GCTG	CTTTTGGAAG	CTTCTGAAAG
	GA	GCGA	GGAC	CTTTCAGAAG	CTTCCAAAAG
	GC	GCGC	GGCC	TGTTCGGTGG	CCACCGAACA
	GG	GCGG	GGGC	TGTTTGGTGG	CCACTGAACA
	GT	GCGT	GGTC	TGTTCAGTGG	CCACCAAACA
	TA	TCGA	GTAC		
	TC	TCGC	GTCC		
	TG	TCGG	GTGC		
	TT	TCGT	GTTC		

Highly preferred and **poorly preferred** bases

4.3 Sequence analysis for determining tetranucleotide frequencies

Base, dinucleotide and tetranucleotide frequencies were determined using the online sequence analysis tool at BioPHP site - Chaos Game Representation of DNA (http://www.biophp.org/minitools/chaos_game_representation/demo.php)

Chaos Game Representation of Frequencies (FCGR), as described by Deschavanne *et al.* (1999), which calculates frequencies of oligonucleotides of various lengths such as mono-, di-, tri-, tetra-, penta-, hexa- and hepta-nucleotide frequencies. Frequencies of oligonucleotides in a given DNA sequence can be obtained by this tool in single strand or in double strand of the DNA sequence however in the present work the frequencies were determined in single strand only.

The expected frequencies of oligonucleotides were determined by multiplying the probabilities of the individual nucleotide bases consisted in it multiplied by the total length of

the sequence. For example, in a 1 Mbp sequence the expected frequency of ACGG was determined by following formula:

$$P(A) \times P(C) \times P(G) \times P(G) \times 1000000$$

where

P(A) = frequency of A in the 1 Mbp sequence / 1000000

P(C) = frequency of C in the 1 Mbp sequence / 1000000

P(G) = frequency of G in the 1 Mbp sequence / 1000000

P(T) = frequency of T in the 1 Mbp sequence / 1000000

The frequencies of relevant oligonucleotides were classified out of the comprehensive list of frequencies for further analysis. Based on frequency of individual bases, the expected frequencies of the classified oligonucleotides were also calculated. In addition to that, all the sequences were subjected to an algorithm for producing random sequences with identical bases composition using http://www.bioinformatics.org/sms2/shuffle_dna.html web site. Entire operation of classification of frequencies of relevant oligonucleotides was performed for randomized sequences. Finally Pearson's coefficient of correlation was obtained for the relative initial velocity of Dnmt3a enzyme kinetics using various tetranucleotides and their respective frequencies.

4.4 Determination of CpG islands in the sequences of *H. sapiens, M. musculus, D. rerio* and *D. melanogaster*

Screening of four of the above mentioned sequences for CpG islands was done using a web based resource, CpG islands searcher (www.cpgislands.com). The parameters used for the screen were as following:

1. GC%	55%
2. Obs_{CpG}/Exp_{CpG}	0.65
3. Minimum length of the CpG Island	500 bp
4. Minimum gap between two adjacent CpG Islands	100 bp

4.5 Determination of frequency of four deca-nucleotides (4bp most preferred and least preferred flank around CpG) in large contigs of *H. sapiens, M. musculus, D. rerio* and *D. melanogaster*

Randomly selected 100 Mbp DNA sequences of *H. sapiens, M. musculus, D. rerio* and *D. melanogaster* with following details were downloaded:

1. *Homo sapiens* (100 Mbp)

Table 11: 100 Mbp DNA sequences of *Homo sapiens*

S. No.	Accession no.	Length	From	To
1	NT_032977.9	80 Mbp	1	80000000
2	NT_004487.19	20 Mbp	1	20000000

2. *Mus musculus* (100 Mbp)

Table 12: 100 Mbp DNA sequences of *Mus musculus*

S. No.	Accession no.	Length	From	To
1	NT-039170.7	50 Mbp	1	50000000
2	NT_078297.6	50 Mbp	1	50000000

3. *Drosophila melanogaster* (100 Mbp)

Table 13: 100 Mbp DNA sequence of *Drosophila melanogaster*

S. No.	Accession no.	Length	From	To
1	NC_033779.4	23 Mbp	1	23000000
2	NC_004354.3	22 Mbp	1	22000000
3	NC_033778.3	21 Mbp	1	21000000
4	NC_037436.3	24 Mbp	1	24000000
5	NC_033777.2	10 Mbp	1	10000000

4. *Danio rerio* (50 Mbp)

Table 14: 100 Mbp DA sequence of *Danio rerio*

S. No.	Accession no.	Length	From	To
1	NW_003039148.2	1.2	1	1200000
2	NW_001878094.3	0.8	1	800000
3	NW_001878646.3	1.5	1	1500000
4	NW_001878649.3	2.0	1	2000000
5	NW_003334160.1	6.7	1	6700000
6	NW_00333465.1	4.2	1	4200000
7	NW_001878800.3	4.3	1	4300000
8	NW_001878801.3	4.8	1	4800000
9	NW_001878804.3	5.5	1	5500000
10	NW_001878810.3	2.6	1	2600000
11	NW_001878787.3	2.3	1	2300000
12	NW_001878785.3	3.7	1	3700000
13	NW_001878903.3	4.1	1	4100000
14	NW_001878907.3	3.6	1	3600000
15	NW_00187891.3	2.7	1	2700000

Each of the sequence was subjected to analysis for determining the frequencies of following oligonucleotides using web based sequence analysis resource, FUZZNUC (http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::fuzznuc) (Rice *et al*, 1999).

CHAPTER 5

RESULTS

5. RESULTS

DNA methylation leads to suppression of CpGs in vertebrate genome. The under-representation of CpGs is non-uniformly distributed in genome due to CpG islands which usually remain unmethylated and have higher CpG representation. Further it has been shown that methylation of CpGs is discriminated by mammalian *de novo* DNA methyltransferases as they exhibit preference for certain flanking bases around CpGs (Lin *et al, 2001;* Handa and Jeltch,*2005).* It is interesting indeed to investigate the effect of this discrimination on suppression of CpGs in context of different flanking sequences in various methylated genomes. It is hypothesized that CpGs with preferred flanking bases are expected to undergo higher levels of methylation and as a result would have higher probability to get mutated to TpG/CpA while other CpGs may be poorly methylated, consequently will survive. Thus flanking bases preference of Dnmt3a and Dnnmt3b might have affected the genome structure in the course of evolution. In order to test the hypothesis, randomly selected genomic sequences of ~1 Mbp length each of several organisms have been analyzed. The selected organisms are *Homo sapiens*, *Mus musculus* strain C57BL/6J, *Danio rerio* strain Tuebingen, *Arabidopsis thaliana, Drosophila melanogaster*, *Caenorhabditis elegans, Plasmodium falciparum* 3D7, *Saccharomyces cerevisiae* S288c and *Escherichia coli* str. K-12. The sequences were fragmented into regions of 100 Kbp length segments and were analyzed for determining frequencies of mono-, di-, tri- and tetra-nucleotide sequences. The frequencies of tetranuclotides NCGN were correlated against relative initial velocities (RIV) of Dnmt3a enzyme kinetics with corresponding tetra-nucleotide substrates (Handa and Jeltsch, 2005). A significant negative correlation was expected with methylated genomes (*H. sapiens, M. musculus, D. rerio* and *A. thaliana*) while weak correlations with unmethylated genomes (*D. melanogaster, C. elegans, P. falciparum, S. cerevisiae* and *E. coli).*

When Obs_{CpG}/Exp_{CpG} frequencies were compared, a strong suppression of CpGs was observed in *H. sapiens and M. musculus* sequences while weak suppression was seen in *D. rerio* and *A.thaliana* (fig 05).

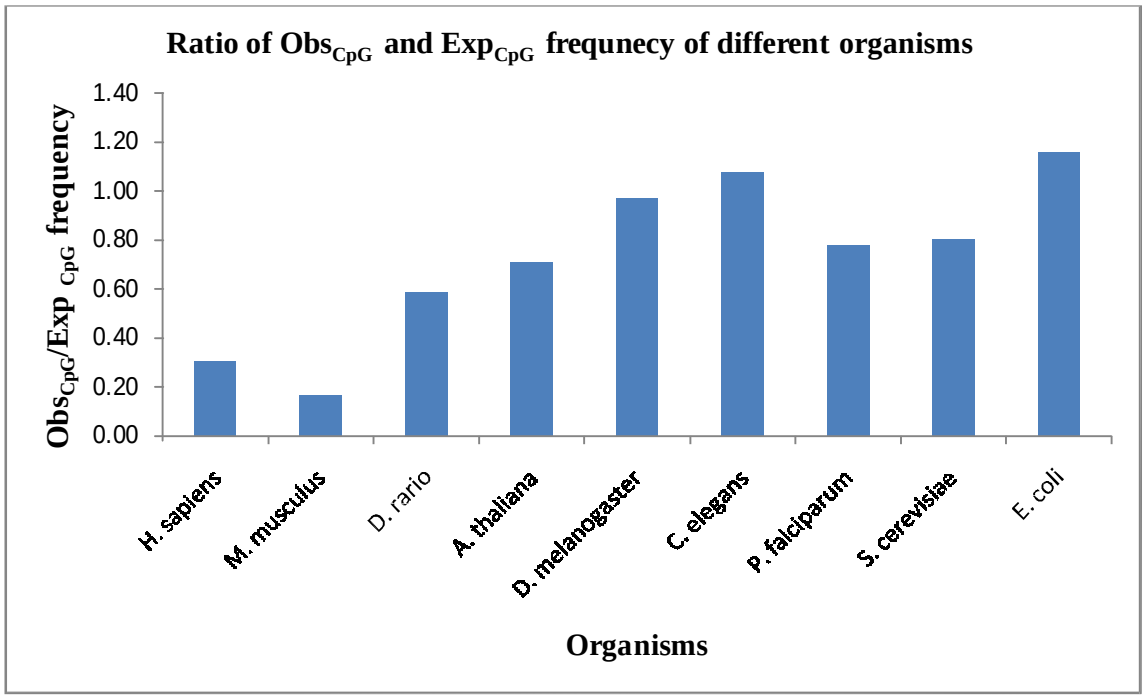

Figure 05: Ratio of Obs_{CpG} and Exp_{CpG} frequency of different organisms

This result largely is in accordance with the earlier reports of CpG suppression in vertebrate genomes that undergo DNA methylation at CpG sites (Bird 1983; Karlin 1994).

In order to investigate the effect of flanking bases at position -1 and +1 of CpGs, the ratio of frequencies of NCGN and GNNC were determined. The GNNC frequency was used as control to normalize the results as well as neutralize the bias in representation of individual bases. Pearson's coefficient of correlation was determined between NCGN/GNNC and RIV of Dnmt3a for corresponding NCGN for each organism (Fig. 06).

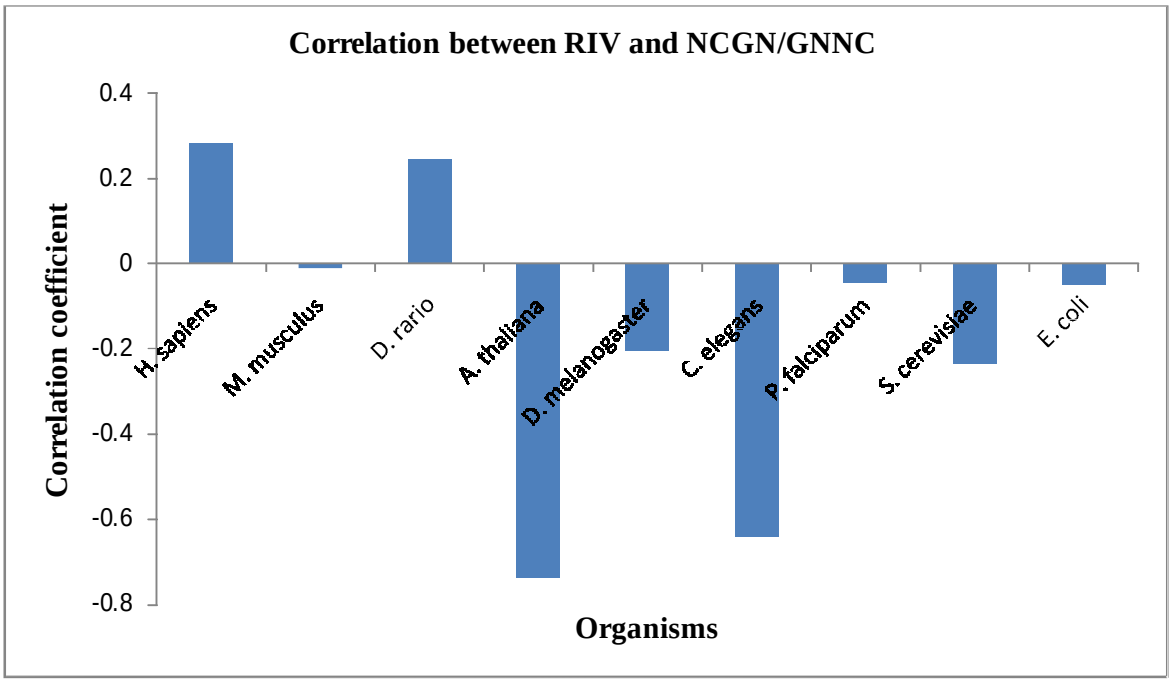

Figure 06: Correlation between RIV and NCGN/GNNC

The correlation values did not follow the expected trend. The normalizing control was improved by taking the ratio of observed and expected frequencies of each of the tetranucleotides (NCGN) (Fig. 07). With this expression negative correlations were obtained for human and mouse sequences but stronger values were observed for most of the other genomes. In another attempt, Monte Carlo approach was used to generate more appropriate control. Random sequences with base composition identical to the genomic sequences were determined by using a shuffle tool provided by website www.bioinformatics.org/sms2/shuffle_dna.html. The randomized sequences were once again analyzed for oligonucleotide frequencies. The basic idea behind this operation was to wash off the biased representation of di-, tri-, tetranucleotides and higher oligonucleotides in the sequence while maintaining the basic composition of bases. Correlation coefficient was determined between RIV and ratio of tetranucleotides (NCGN) frequencies in genomic and corresponding randomized sequences. The result was very similar to the one in which $obs_{NCGN}/expected_{NCGN}$ ratio was used (Fig 08).

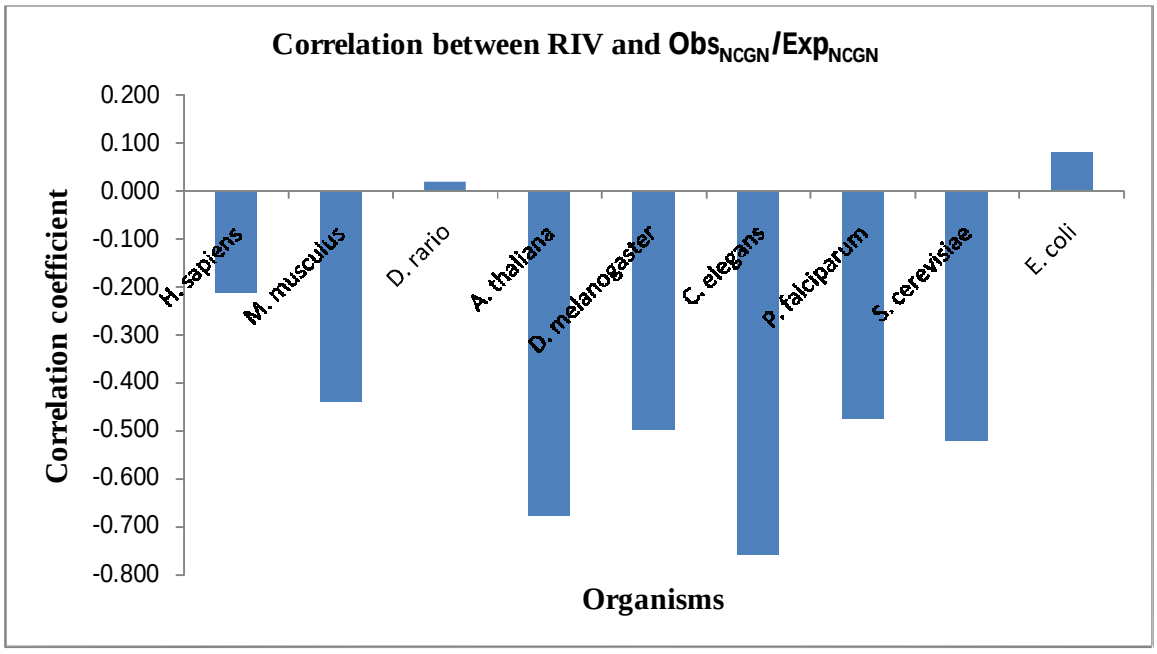

Figure 07: Correlation between RIV and Obs_{NCGN}/Exp_{NCGN}

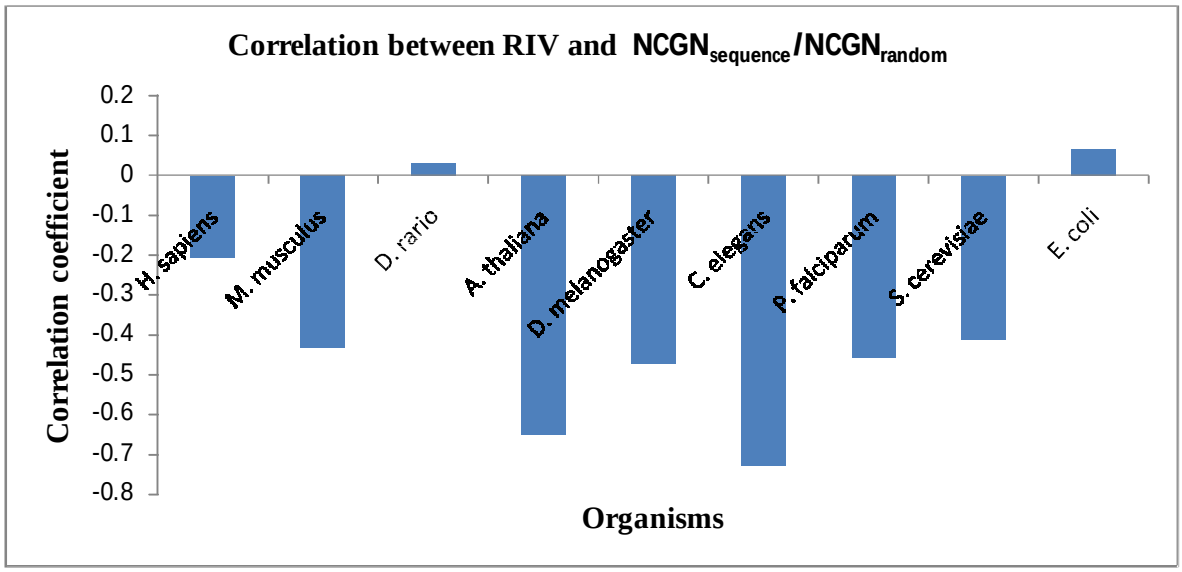

Figure 08: Correlation between RIV and $NCGN_{sequence}$ /$NCGN_{random}$

The control was further attempted to improve by using product of expected frequency of CpG and the two flanking bases (Obs_{CpG} x P(N) x P(N)). This expression was used to

neutralize the effect of underrepresentation of CpGs and to diminish the effect of NC and GN dinucleotides' bias in case it existed. However little change in the trend of correlation was achieved. Though stronger negative correlations were observed in *H. sapiens* and *M. musculus* sequences, more significant correlations were observed in *D. melanogaster* and *C. elegans* (Fig. 09).

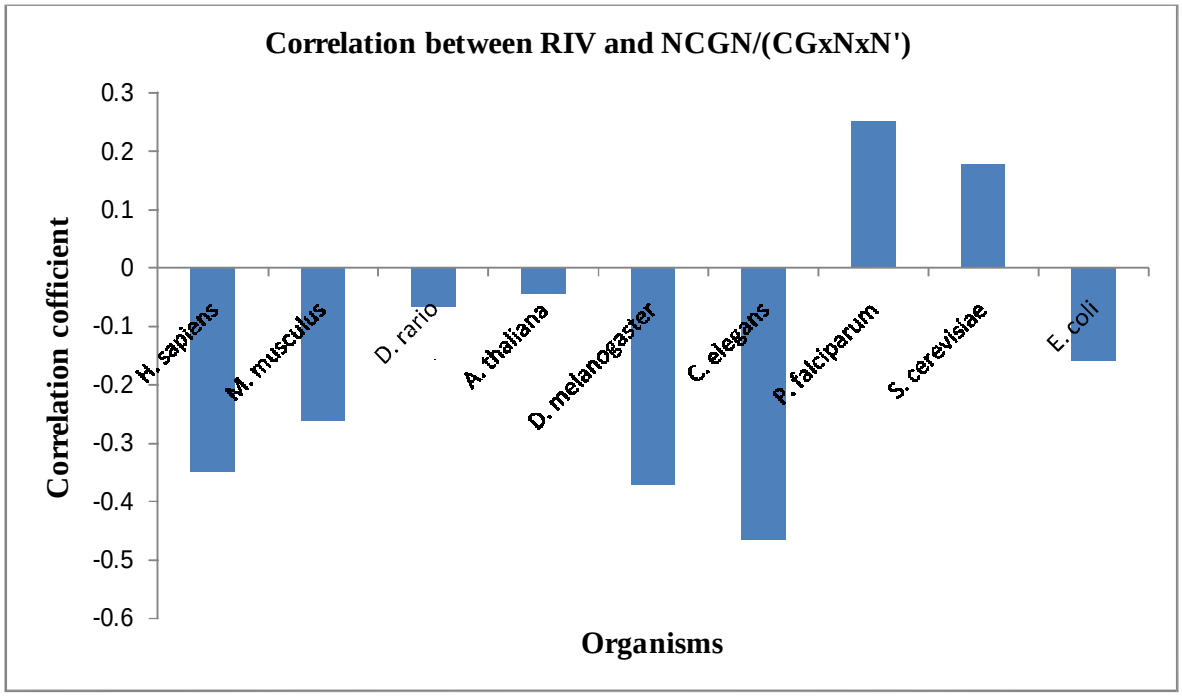

Figure 09: Correlation between RIV and NCGN/(CGxNxN)

Mutation of methylated CpGs generates TpG/CpA. It implies that suppression of CpGs in methylated genome should bring corresponding increase in TpGs and CpAs which has been shown in several reports (Bird 1983; Karlin 1994). Similar effect is expected in NCGN, NTGN and NCAN frequencies too. The frequencies of NTGN and NCAN were included in the analysis to enhance the effect of biased representation of the tetranucleotides in the methylated genomes. The correlation analysis was performed using NCGN-NTGN-$NCAN_{sequence}$/NCGN-NTGN-$NCAN_{random}$ and (NCGN-NTGN-NCAN)obs/(NCGN-NTGN-NCAN)exp ratios. Stronger negative correlations were observed for human and mouse sequences but similar significant negative correlations were observed in many unmethylated genomes as well in both the analyses.

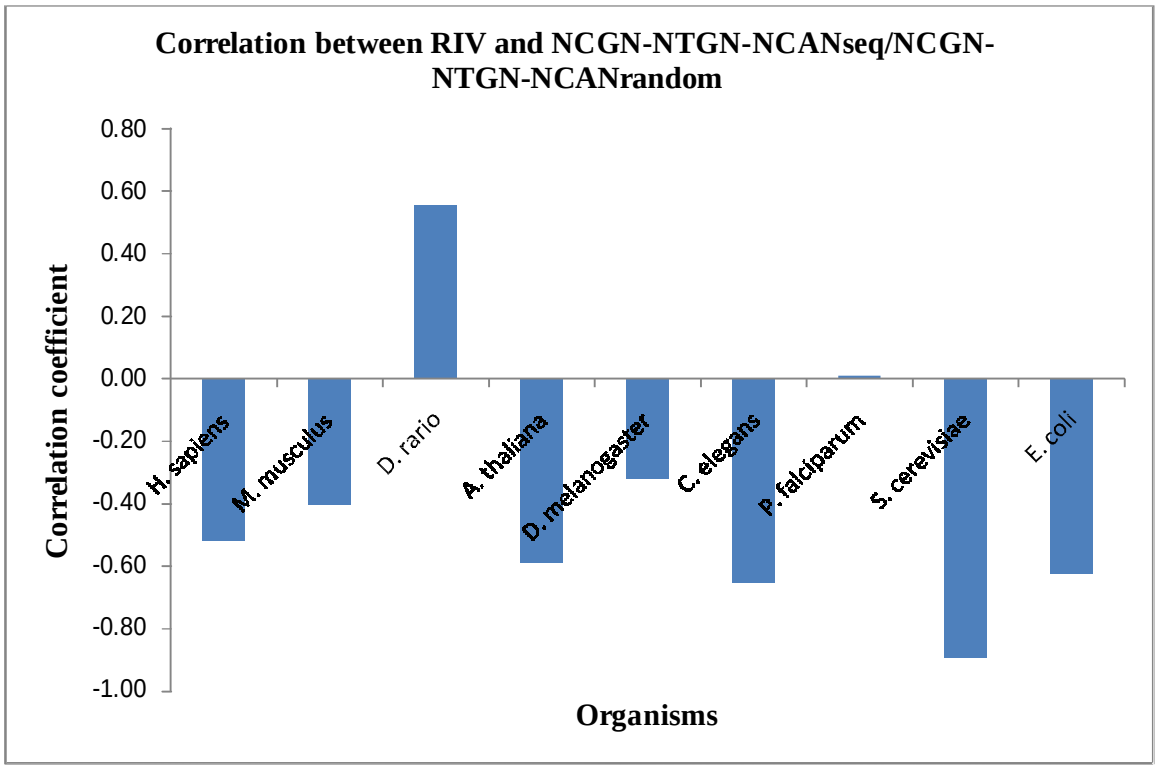

Figure 10: Correlation between RIV and NCGN-NTGN-$NCAN_{sequence}$/NCGN-NTGN-$NCAN_{random}$

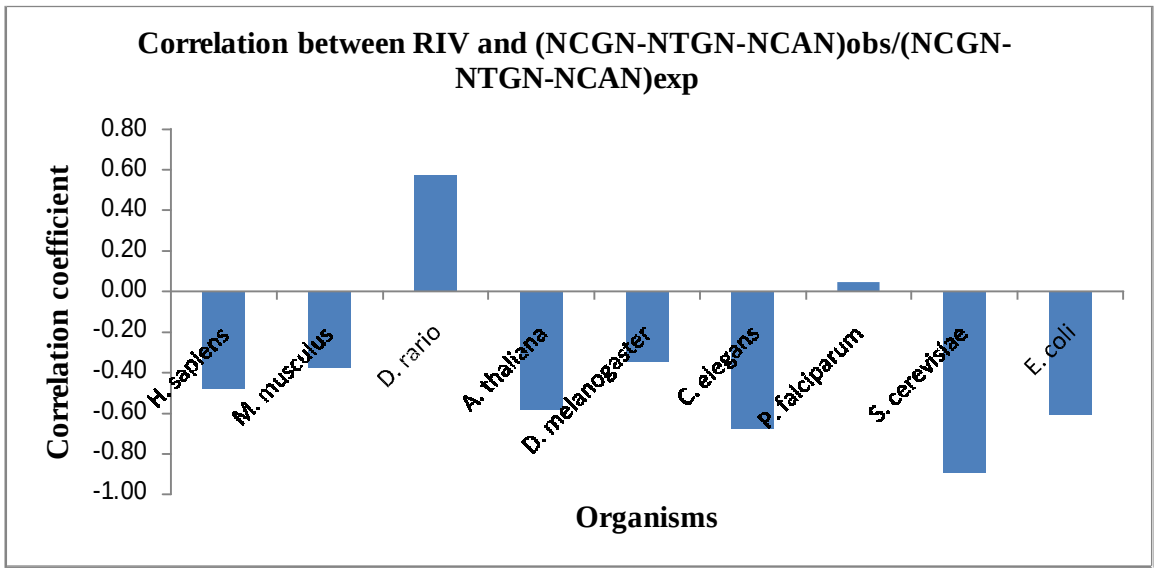

Figure 11: Correlation between RIV and (NCGN-NTGN-NCAN)obs/(NCGN-NTGN-NCAN)exp

Analysis of genomic sequences of methylated genomes for CpG islands

The expected results were not observed in the above mentioned experiments. There could be several reasons that were masking the effect of flanks on abundance of NCGNs. One of the apparent reasons could be presence of CpG islands in the sequences of methylated genomes. Since otherwise observed suppression of CpGs in not observed effectively in CpG islands owing to their usual unmethylated state, the presence of the islands in the sequences could be diluting the effect of flanking bases on the tetra-nucleotide frequencies. Then the sequences of methylated genomes (*H. sapiens, M. musculus, D. rerio and A. thaliana)* were subjected to analysis for CpG islands. CpG islands in the sequences of following organisms were screened to determine the number of CpG dinucleotides present in CpG islands and total CpG present in entire DNA sequence. A significant fraction of CpGs was found to be present in CpG islands (fig. 11). This could be one of the major reasons that diminished the effect of flanking bases on frequency of the tetra-nucleotides.

Table 15: CpG islands analysis

S.No.	**Organism**	**Length of sequence**	**CpG in CpG islands**	**CpG in total Sequence**
1	*Homo sapiens*	1000000	2843	14028
2	*Mus musculus*	1000000	400	7008
3	*Danio rerio*	1000000	1709	20826
4	*Drosophila melanogaster*	1000000	8827	42446

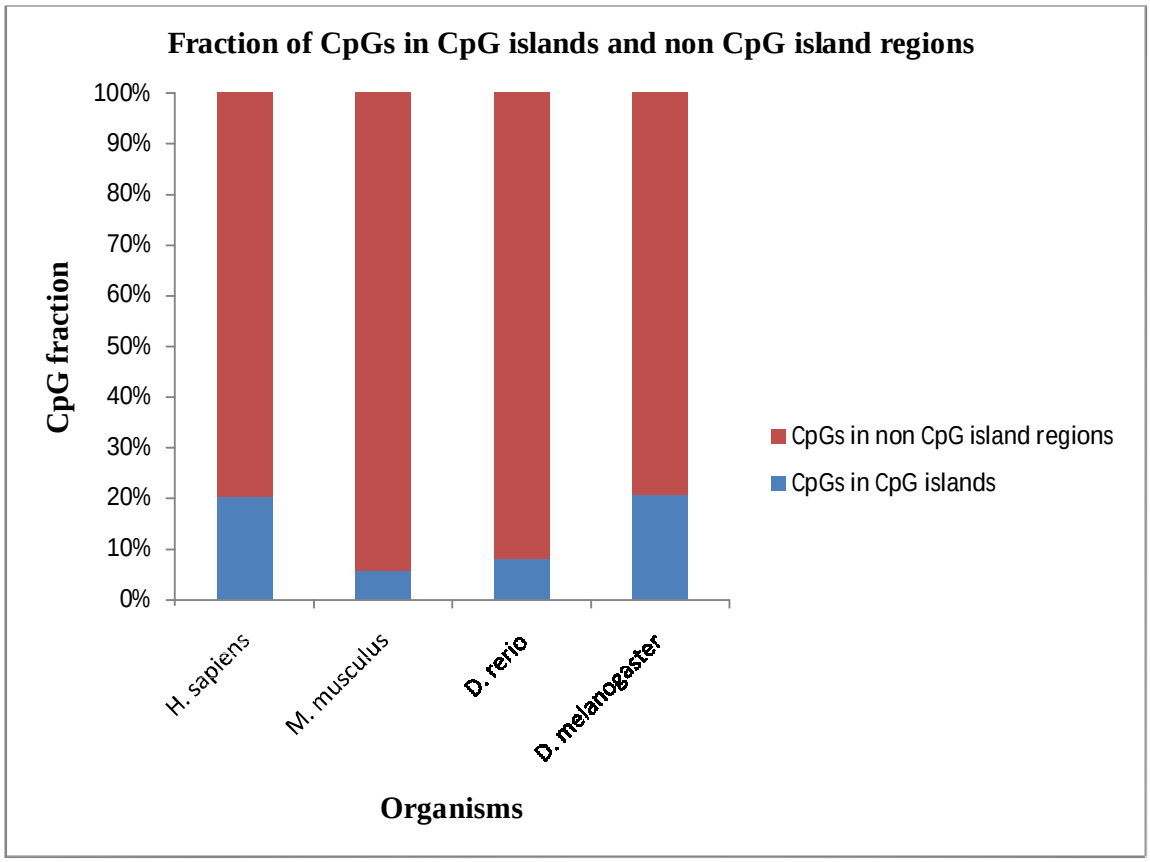

Figure 12: Fraction of CpG in CpG islands and non CpG island regions

Comparison of frequencies of decanucleotides containing CpGs flanked by highly preferred and poorly preferred bases upto -4 and +4 position

The effect of most preferred and least preferred flanking bases at -1 and +1 position of NCGN on methylation kinetics of Dnmt3a is only 13 fold (Handa and Jeltsch, 2005). DNA sequences were analyzed for decamer frequencies for *H.sapiens*, *M. musculus*, *D. rerio* and *D. melanogaster* (table 16). If the preference of flanking bases is extended up to -4 and +4 position, the fold difference in the relative initial velocity of kinetics increases more than an order of magnitude. Keeping this in view, frequencies of the deca-nucleotides with very high, high, moderate and poor preference was determined in 100 Mbp sequences of *H.sapiens*, *M. musculus* and *D. melanogaster* and 50 Mbp sequence of *D. rerio* using 'Fuzznuc' search tool (searches for a specified pattern of short length in nucleotide sequences) of EMBOSS.

Table 16: Analysis of decamer frequencies for following organisms

S.No.	Decanucleotide		Frequencies							
			H. sapiens (100 Mbp)		*M. musculus* (100 Mbp)		*D. rerio* (50 Mbp)		*D. melanogaster* (100 Mbp)	
			Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp
1	Very highly preferred	CTTACGCAAG	6	80	4	81	7	31	33	85
2		CTTGCGTAAG	13	80	7	80	3	31	35	85
3		CTTATGCAAG	78	116	78	168	27	56	77	115
4		CTTGCATAAG	85	116	84	116	42	56	98	115
5		CTTACACAAG	88	117	100	117	28	55	77	115
6		CTTGTGTAAG	92	116	114	116	37	56	94	115
7	Highly preferred	TGTACGCTGG	8	54	9	55	5	18	34	62
8		TGTATGCTGG	4	55	11	56	10	18	28	62
9		CCAGCATACA	72	79	140	80	57	32	81	85
10		CCAGCATACA	92	80	118	81	55	31	88	85
11		TGTACACTGG	66	80	111	80	47	31	80	85
12		CCAGTGTACA	66	80	112	81	35	31	67	85
13	Moderately preferred	CTTTCGGAAG	16	80	15	80	10	31	57	85
14		CTTCCGAAAG	19	80	29	81	19	31	51	85
15		CTTTTGGAAG	210	116	207	116	69	56	128	115
16		CTTCCAAAAG	193	117	218	117	71	55	128	115
17		CTTTCAGAAG	246	116	302	116	96	56	111	115
18		CTTCTGAAAG	238	116	282	116	88	56	145	115
19	Poorly preferred	TGTTCGGTGG	10	54	10	55	9	18	107	62
20		CCACCGAACA	6	55	19	56	6	18	24	62
21		TGTTTGGTGG	187	79	267	80	80	32	182	85
22		CCACCAAACA	165	80	285	81	86	31	120	85
23		TGTTCAGTGG	122	79	152	80	63	32	89	85
24		CCACTGAACA	131	80	140	81	55	31	88	85

Comparing the $(CpG-(TpG+CpA))_{obs}/(CpG-(TpG+CpA))_{exp}$ ratio of frequencies of the 4 deca-nucleotides an inverse relationship was observed between the ratio and the relative initial velocity in *H.sapiens*, *M. musculus* and *D. rerio* while a weaker similar trend was observed in *D. melanogaster* (Fig. 13) which may be explained by the fact that *D. melanogaster* has only Dnmt2 DNA methyltransferase.

Table 17: $CpG-(TpG+CpA))_{obs}/(CpG-(TpG+CpA))_{exp}$ ratio of frequencies of the 4 deca-nucleotides

Decamers	***H. sapiens***	***M. musculus***	***D. rerio***	***D. melanogaster***
Very Highly preferred	1.06	1.03	0.78	0.96
Highly preferred	1.36	2.19	1.99	1.19
Moderately preferred	2.79	3.18	1.84	1.39
Poorly preferred	2.81	3.87	2.98	1.62

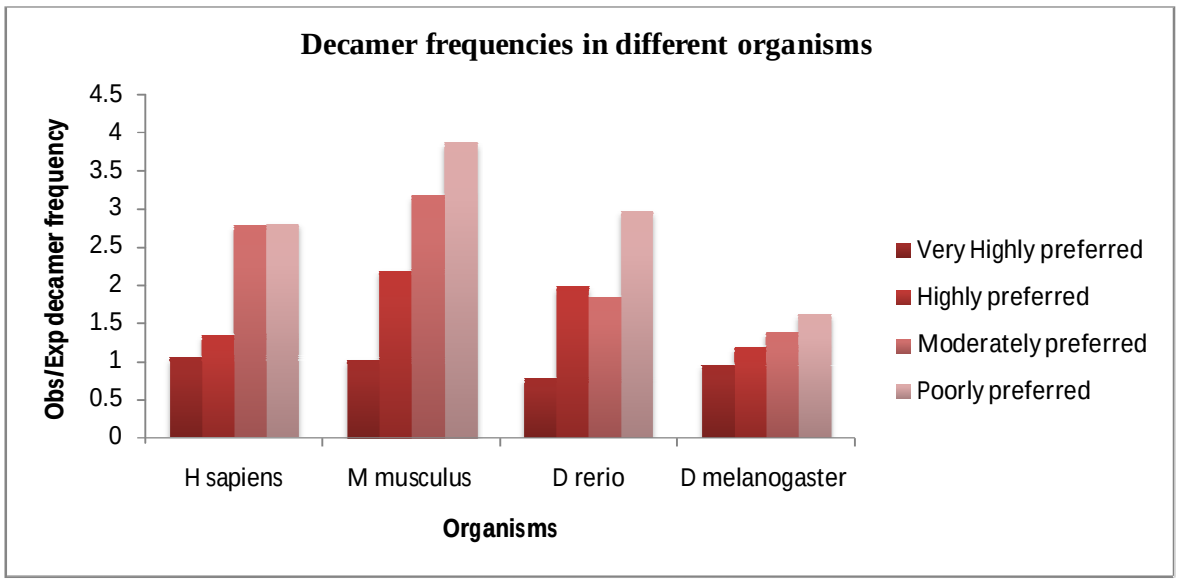

Figure 13: Decamer frequencies in different organisms

This result proves the hypothesis that the occurance of CpGs with different flanks is influenced by the preferences of *de novo* methyltransferases for the flanking sequences. The effect is not very strong however distinctly leaves its mark on the genome structure.

CHAPTER 6
DISCUSSION

6. DISCUSSION

In eukaryotes the DNA methylation is found in vertebrates at CpG nucleotides and plays important role in gene silencing. In addition to that methylated CpGs have tendency to be mutated to TpG/CpA that has reduced the representation of CpGs in methylated genomes in comparison to other dinucleotides. In the present study where ~1 Mbp representative sequences of various methylated and unmthylated genomes were analyzed to determine frequencies of mono-, di- and tetra-nucleotides and it was observed that in the methylated genomes there was indeed suppression of CpGs. The comparison was made by comparing the Obs/Exp ratios of CpGs in various genomes. Since Dnmt3a and Dnmt3b enzymes that are involved in CpG methylation have been found to have bias for substrate CpGs based on their flanking sequences, it is interesting to investigate if this preference is reflected in the occurance of CpG with differentially preferred flanks in the genomes survived in the course of evolution. It is reported that purines (R) at -1 and pyrimidines (Y) at +1 position of CpGs affect methylation preference of *de novo* methyltrasferases positively i.e. these sequences are highly favoured for methylation. Thus RCGY are highly preferred while YCGR are poorly preferred by *de novo* Dnmts.

Since only methylated CpGs may undergo occasional spontaneous deamination that mutates it to TpG/CpA, it is expected that owing to the preference of *de novo* Dnmts, RCGYs will have higher chance of getting methylated and as a result mutated when compared to YCGRs. This should further be reflected in the frequency of these tetra-nucleotides in methylated genomes. Since Handa and Jeltsch have reported relative initial velocity of Dnmt3a enzyme kinetics for exhaustive set of CpG -1 and +1 flanks (NCGN) which quantitatively represents the preference the enzyme, it was possible to compare it with the frequency of all theses tetranucleotides in different genomes. In order to minimize the effect of bias of individual bases based base composition of the sequence analyzed the frequencies of tetra-nucleotides (NCGN) were normalized by dividing them with frequency of another distinct tetra-nucleotide with identical base composition but lacking any CpG in it (GNNC). Pearson's coefficent of correlation was determined between the relative initial velocity of NCGN and corresponding NCGN/GNNC ratio for different organisms. The expected significant negative value of correlation in methylated genomes and non-significant correlation in

unmethylated genomes was not observed. In an attempt to further improve normalization which could be a possible reason diluting the effect of flanking bases on the occurance of CpGs, the correlation was performed between RIV and ratio of Obs_{CpG}/Exp_{CpG}. No significant negative correlation for methylated genomes was found in this attempt either. For further improvement in normalization, Monte Carlo approach was used and all the sequences were subjected to radomization ie to create random sequences of same length with same base composition. Then these randomized sequences were analyzed to determine frequencies of mono-, di- and tetra-nucleotides as was done in the case of original sequences. Pearson's coefficient of correlation was determined between realative initial velocity and the ratio of frequencies of tetranucleotides in original sequence and the randomized squence. Once again no expected significant negative correlation was found in methylated genomes.

To enhance the bias in representation of NCGNs in the genomes, the frequencies of NTGN and NCAN, which are mutation product of methylated NCGN, were also taken into consideration and the analysis was repeated with ratios of NCGN-NTGN-$NCAN_{sequence}$/NCGN-NTGN-$NCAN_{random}$ and (NCGN-NTGN-NCAN)obs/(NCGN-NTGN-NCAN)exp against RIV. No significant improvement in the trend of correlation values was observed in either case.

One of the factors that might be responsible for masking the hypothesized effect could be presence of significant fraction of CpGs in the CpG islands in the methylated sequences. In order to check it, the sequences of methylated genomes were subjected to screening for CpG islands. There indeed was significant fraction of CpGs present in the CpG islands. Since CpG islands are usually not methylated, it is expected that CpGs present in these regions would not experience mutations. This could be affecting the overall frequencies of the tetra-nucleotides.

There might be other reasons responsible for weaker correlations. For example, it has been studied that certain specific sequences flanking CpGs are preferred by methyltransferases but if there is any influence of such flanking sequences on deamination of methylated CpG is not known. Moreover evolutionary selection pressure can also influence the rate of mutations in

different regions of the genome. Moreover, though there is appreciable overlap between biochemical data and physiological data as reported by Handa and Jeltesch, relative initial velocity might not be quantitatively very accurate reflection of methylation preference under physiological conditions. Several factors may affect the relative initial velocity of DNA methylation by Dnmt3a and Dnmt3b such as wrapping of DNA around histone proteins might be influencing the preference of methylation of CpG. If DNA is tightly bound to histone proteins, it may not be accessible to *de novo* methyltransferases for methylation. Other DNA methylatrasferases present in the cells, such as Dnmt1and Dnmt3L may influence the activity of *de novo* methyltransferases as well as their flanking base preference. Methyl-CpG-binding domain proteins (MBD) might interact with Dnmt3a and Dnmt3b and affect their activity. Since these protein bind to methylated CpG, deamination of these methylated sites could be influenced.

Presence of transcription factors and other DNA binding proteins as well as protein constituting chromatin might affect the methylation or deamination processes. Moreover, methylation is followed by deamination which results in inherited changes in DNA sequence occurring in germ cells. But results of Handa and Jeltsch are entirely based on either biochemical studies or physiological data derived from non germ cell lines. We know that in germ cells, there is extensive dynamics of DNA methylation as well as active de-methylation.

Since Handa and Jeltsch have reported stronger bias in the enzyme kinectics of Dnmt3a when flanks of 4 bp upstream and downstream of CpG sites are considered when compared to the range of bias for flanks of one bp on either side of CpGs, the frequency of deca-nucleotides showing very high, moderately high, moderately low and very low relative initial velocity was determined. The ratio of difference of frequencies of these deca-nuclotides and their mutation products with their expected frequencies $(CpG-(TpG+CpA))_{obs}/(CpG-(TpG+CpA))_{exp}$) showed negative relation with relative intial velocity for *H. sapiens* as well as *M. musculus* and *D. rerio*. Even *D. melanogaster* which possesses only one DNA methyltransferase (Dnmt2) exhibited similar trend of ratio of frequencies though with mild effect . This clearly indicates that in human the sequences which are highly preferred by

Dnmt3a and Dnmt3b have high chances of getting methylated and as a result they have been mutated more often than poorly preferred CpGs in the course of evolution. This effect could not be observed with tetranucleotide frequencies maybe because the effect is not very prominent and is probably masked by various factors mentioned above. However the effect shows that enzymatic preference of Dnmt3a and Dnmt3b for CpGs with different flanking bases has been influencing the genome structure in a deeper manner than mere suppression of CpGs in methylated genomes.

CHAPTER 7
CONCLUSION

7. CONCLUSION

In the present study genomic sequences of nine different organisms have been analyzed to study the effect of CpG flanks on the genome structure via differential methylation followed by its mutation to TpG/CpA. The suppression of CpGs in methylated genomes was confirmed when Obs/Exp ratio of frequency of CpGs was determined and compared amongst different methylated and unmethylated genomes.

The effect of bases at -1 & +1 position flanking CpG that influence the substrate preference of *de novo* DNA methyltransferases on the frequency of corresponding NCGNs due to differential methylation was not detected. It may be attributed to several reasons including influence of CpGs as part of CpG islands which usually remain unmethylated and are not expected to get mutated.

The desired effect of inverse relationship between frequencies of CpG containing sequences and enzymatic preference of Dnmt3a for them as substrates was detected when frequency of four deca-nucleotides, each containing a CpG flanked by four bases on either side was compared to the RIV of Dnmt3a acting on them. The result clearly indicates that flanking base preference of *de novo* DNA methyltransferase has been gradually influencing the genome structure of methylated genomes in the course of evolution.

CHAPTER 8

REFERENCES

REFERENCES

1. Amir R, Veyver I, Wan M, Tran C, Francke U & Zoghbi H- Rett syndrome is caused by mutations in X-linked *MECP2*, encoding methyl-CpG-binding protein 2-*Nature Genetics*, 1999, **23**:185–188.
2. Barreto G, Schafer A, Marhold J, Stach D, Swaminathan S, Handa V, Derlein G, Maltry N, Wu W, Lyko F & Niehrs C- Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation-*Nature*, 2007, **445**:671-675.
3. Biemont C and Vieira C- Junk DNA as an evolutionary force-*Nature*, 2006, **443**(5):522-524
4. Bird AP- DNA methylation and the frequency of CpG in animal DNA-*Nucleic Acids Res.*, 1980, **8**(7):1499-504.
5. Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T, Walter J- CpG Island Methylation in Human Lymphocytes Is Highly Correlated with DNA Sequence, Repeats, and Predicted DNA Structure-*PLoS Genetics*, 2006, **2**(3):0243-0252.
6. Brena R, Huang T& Plass C- Toward a human epigenome-*Nature Genetics*, 2006, **38**(12):1359-1360.
7. Cardon L, Burge C, Clayton D and Karlin S- Pervasive CpG suppression in animal mitochondrial genomes-*Proc. Nati. Acad. Sci. USA*, 1994, 91:3799-3803.
8. Chen ZX, Riggs AD – Maintenance and regulation of DNA methylation patterns in mammals *Biochem Cell Biol.* 2005, **83**(4): 438-48.
9. Chen Z and Riggs A- DNA Methylation and Demethylation in Mammals-*J Biol Chem.*, 2011, **286**(21):18347–18353.
10. Clark S, Harrison J and Frommer M- CpNpG methylation in mammalian cells-*Nature Genetics*, 1995, **10**:20-27.
11. Daniel J. Tomso and Douglas A. Bell - Sequence Context at Human Single Nucleotide Polymorphisms: Overrepresentation of CpG Dinucleotide at Polymorphic Sites and Suppression of Variation in CpG Islands-*J. Mol. Biol.*, 2003, **327**:303–308.
12. Das R, Dimitrova N, Xuan Z, Rollins R, Haghighi F, Edwards J, Ju J, Bestor T, and Zhang M-Computational prediction of methylation status in human genomic sequences-*PNAS*, 2006, **103**(28):10713–10716.

13. Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B- Genomic signature: characterization and classification of species assessed by chaos game representation of sequences-*Mol Biol Evol*, 1999, **16**(10):1391-1399.

14. Duncan BK, Miller JH - Mutagenic deamination of cytosine residues in DNA- *Nature*, 1980, **287**(5782):560-561.

15. Feltus F, Lee E, Costello J, Plass C, and Vertino P- Predicting aberrant CpG island methylation-*PNAS*, 2003, **100**(21):12253–12258.

16. Feltus F, Lee E, Costello J, Plass C, Vertino P- DNA motifs associated with aberrant CpG island methylation-*Genomics*, 2006, **87**:572–579.

17. Ferguson-Smith AC, Surani MA- Imprinting and the epigenetic asymmetry between parental genomes-*Science*, 2001, **293**(5532):1086-1089.

18. Fryxell K and Moon W- CpG Mutation Rates in the Human Genome Are Highly Dependent on Local GC Content-*Mol. Biol.*, 2005, **22**(3):650–658.

19. Gardiner-Garden M, Frommer M- CpG islands in vertebrate genomes-*J Mol Biol.*, 1987, **196**(2):261-282.

20. Hagemann S, Heil O, Lyko F, Brueckner B- Azacytidine and Decitabine Induce Gene-Specific and Non-Random DNA Demethylation in Human Cancer Cell Lines- *Plos One*, 2011, **6**(3):e17388.

21. Han L, Su B, Li W and Zhao Z - CpG island density and its correlations with genomic features in mammalian genomes-*Genome Biology*, 2008, **9**:R79.

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

23. Hermann A, Gowher H, Jeltsch A- Biochemistry and biology of mammalian DNA methyltransferases-*Cell Mol Life Sci.*, 2004, **61**(19-20):2571-2587.

24. Jabbaria K and Bernardi G- Cytosine methylation and CpG, TpG (CpA) and TpA frequencies-*Gene*, 2004, **333**:143–149.

25. Jaenisch R, Bird A- Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals- *Nat Genet.*, 2003, **33**:245-254.

26. Jiang YL, Rigolet M, Bourc'his D, Nigon F, Bokesoy I, Fryns JP, Hultén M, Jonveaux P, Maraschio P, Mégarbané A, Moncla A, Viegas-Péquignot E- DNMT3B mutations and DNA methylation defect define two types of ICF syndrome-*Hum Mutat.*, 2005, **25**(1):56-63.

27. Kahng LS, Shapiro L- The CcrM DNA methyltransferase of Agrobacterium tumefaciens is essential, and its activity is cell cycle regulated- *J Bacteriol.*, 2001, **183**(10):3065-3075.

28. Karlin S, Ladunga I, Blaisdell BE- Heterogeneity of genomes: measures and values-*Proc Natl Acad Sci U S A.*, 1994, **91**(26):12837-12841.

29. Krieg AM- CpG motifs in bacterial DNA and their immune effects-*Immunol.*, 2002, **20**:709-760.

30. Kruger D and Bickle T- Bacteriophage Survival: Multiple Mechanisms for Avoiding the Deoxyribonucleic Acid Restriction Systems of Their Hosts-*Microbiological reviews*, 1983:345-360.

31. Lin IG, Han L, Taghva A, O'Brien LE, Hsieh CL- Murine de novo methyltransferase Dnmt3a demonstrates strand asymmetry and site preference in the methylation of DNA in vitro-*Mol Cell Biol.*, 2002, **22**(3):704-723.

32. Lorincz MC, Groudine M- C(m)C(a/t)GG methylation: a new epigenetic mark in mammalian DNA?-*Proc Natl Acad Sci U S A.*, 2001, **98**(18):10034-10036.

33. Lorincz MC, Schubeler D, Groudine M- Methylation-mediated proviral silencing is associated with MeCP2 recruitment and localized histone H3 deacetylation-*Mol Cell Biol.*, 2001, **21**(23):7913-7922.

34. Martienssen RA, Colot V- DNA methylation and epigenetic inheritance in plants and filamentous fungi-*Science*, 2001, **293**(5532):1070-1074.

35. Murrell A, Rakyan VK, Beck S- From genome to epigenome-*Hum Mol Genet.*, 2005, **14(**1):R3-R10.

36. Novik KL, Nimmrich I, Genc B, Maier S, Piepenbrock C, Olek A, Beck S- Epigenomics: genome-wide study of methylation phenomena-*Mol Biol.*, 2002, **4**(4):111-128.

37. Oller O, Cabre M, Montero M, Paternain J, Romeu A- Specific gene hypomethylation and cancer: New insights into coding region feature trends- *Bioinformation*, 2009, **3**(8): 340-343.

38. Palmer BR, Marinus MG - The dam and dcm strains of Escherichia coli--a review- *Gene*, 1994, **143**(1):1-12.

39. Park Y, Kuroda MI - Epigenetic aspects of X-chromosome dosage compensation, *Science*, 2001, **293**(5532):1083-1085.

40. Peedicayil J-Epigenetic therapy- a new development in pharmacology-*Indian J Med Res,* 2006, **123**:17-24.

41. Rakyan VK, Hildmann T, Novik KL, Lewin J, Tost J, Cox AV, Andrews TD, Howe KL, Otto T, Olek A, Fischer J, Gut IG, Berlin K, Beck S- DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project-*PLoS Biol.*, 2004, **2**(12):e405.
42. Rice P, Longden I, Bleasby A- EMBOSS: the European Molecular Biology Open Software Suite- *Trends Genet.*, 2000, 16(6):276-277.
43. Santi DV, Garrett CE, Barr PJ- On the mechanism of inhibition of DNA-cytosine methyltransferases by cytosine analogs-*Cell*, 1983, **33**(1):9-10.
44. Takai D, Jones PA- Comprehensive analysis of CpG islands in human chromosomes 21 and 22-*Proc Natl Acad Sci U S A.*, 2002, **99**(6):3740-3745.
45. Woodcock DM, Lawler CB, Linsenmeyer ME, Doherty JP, Warren WD- Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon-*J Biol Chem.*, 1997, 272(12):7810-7816.
46. Wu Ct, Morris JR- Genes, genetics, and epigenetics: a correspondence-*Science*, 2001, **293**(5532):1103-1105.
47. Zhang F, Zhao Z- The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs-*Genomics*, 2004, **84**: 785–795.
48. Zhang Y, Rohde C, Tierling S, Jurkowski T, Bock C, Santacruz D, Ragozin S, Reinhardt R, Groth M, Walter J, Jeltsch A- DNA Methylation Analysis of Chromosome 21 Gene Promoters at Single Base Pair and Single Allele Resolution-*PLoS Genetics*, 2009, **5**(3):e1000438.
49. Zhao Z and Zhang F- Sequence context analysis in the mouse genome: Single nucleotide polymorphisms and CpG island sequences-*Genomics*, 2006, **87**:68–74.