Studies on CpG distribution in genomes

A dissertation report

Submitted in partial fulfillment of the requirement for

The award of degree of

Master of Science in Biotechnology

Under the guidance of

Dr. Vikas Handa
Assistant Professor

Submitted by:

Bhupinder Singh

(301101008)

Department of Biotechnology and Environmental Sciences

THAPAR UNIVERSITY

PATIALA

July 2013

CERTIFICATE

This is to certify that the thesis entitled **"Studies on CpG distribution in genomes"** submitted by Bhupinder Singh in partial fulfillment of the requirement for the award of Degree of Masters in Science in Biotechnology to Thapar University, Patiala, is a record of student's own work carried out by 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.

15.07.2013

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

CANDIDATE'S DECLARATION

I hereby declare that the work being presented in the thesis entitled "**Studies on CpG distribution in genomes**" in partial fulfilment of the requirements for the award of degree of Masters in Biotechnology, Department of Biotechnology and Environmental Sciences, Thapar University, Patiala is my own work during the period of January 2013 to June 2013, under the conception and supervision of Dr. Vikas Handa, Assistant Professor, Department of Biotechnology and Environmental Sciences (DBTES), Thapar University, Patiala. I have not submitted the matter embodied in this thesis for the award of any other degree.

Patiala
Date: July 15th, 2013

Bhupinder Singh
Bhupinder Singh
Roll No. 301101008

This is to certify that the above statement made by the candidate is correct and true to the best of our knowledge.

15.07.2013

Dr. Vikas Handa
Assistant Professor / Supervisor
DBTES, Thapar University

Dr. M.S. Reddy
Professor & Head
DBTES, Thapar University

Department of Biotechnology & Environmental Sciences
Thapar University, Patiala 147004

ACKNOWLEDGEMENT

It would not have been possible without the kind support, inspiration, guidance, direction, cooperation, love, care, and help of many individuals and organization. I would like to extend my sincere thanks to all of them.

I am highly indebted to my guide ***Dr. Vikas Handa****, Assistant Professor, Department of Biotechnology and Environmental Sciences, Thapar University for his guidance and constant supervision as well as for providing necessary information regarding the thesis work & also for his support in completing the thesis work. His association with this Endeavour of mine will remain a beacon light to my throughout my life.*

I sincerely thankful to ***Dr. M.S. Reddy, Head****, Department of Biotechnology and Environmental Sciences, Thapar University 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.*

I would like to express my gratitude towards my parents & members of Thapar University for their kind co-operation and encouragement, which help me in completion of this work.

I have great pleasure to thanks Mahiti Gupta, Rajnesh Verma and all my other friends who have been always been encouraging and provided me constant support during my project work.

I would like to express my special gratitude and thanks to my friends for giving me support, friendly environment and unforgettable moments in the Thapar University.

At last, I would like to thanks THE ALMIGHTY for his constant blessings without which any task would be impossible.

Date: July 15, 2013, **Bhupinder Singh**

Place: Patiala **(301101008)**

TABLE OF CONTENTS

ABBREVATIONS

A	Adenine
ApC	Adenine and Cytosine
AM	Arithmetic mean
C	Cytosine
C^5	Carbon at 5^{th} position
CpA	Cytosine and Adenine
CpG	Cytosine and Guanine
CGI	CpG island
CV	Coefficient of variance
DI	Dispersion index
Dnmt	DNA methyltransferase
Dnmt 1	DNA methyltransferase 1
Dnmt3a	DNA methyltransferase 3a
Dnmt3b	DNA methyltransferase 3b
ExpCpG	Expected frequency of CpG dinucleotides
G	Guanine
GpC	Guanine and Cytosine
GpT	Guanine and Thiamine
N^5	Nitrogen at 6^{th} position

N^4	Nitrogen at 4^{th} position
ObsCpG	Observed frequency of CpG dinucleotides
ObsCpG/ExpCpG	Ratio of observed frequency of CpG over the expected frequency of CpG dinucleotide
SD	Standard deviation
TpG	Thymine and Guanine

List of Figures

List of Tables

ABSTRACT

DNA methylation is an epigenetic modification that plays very important role in vertebrate genomes. In vertebrate genomes, DNA methylation occurs at CpG sites and leads to gene regulation, gene imprinting, X-chromosome inactivation and cancer. DNA methylation has resulted in depletion in CpG sites in vertebrate genome. CpG islands are the clusters of CpGs in regions of high GC content and are usually unmethylated. However all the clusters of CpGs are not CpG islands but seem to affects the DNA methylation levels in vertebrate genomes. The present study has used CpG gap size (number of nucleotides between two adjacent CpG) to investigate the distribution of CpGs in genomes. Mean CpG gap sizes have been observed to be larger in methylated genome while smaller values are associated with poorly and non-methylated genomes. Similarly, dispersion index of CpG gap values also shows higher value for methylated genomes and lower value in poorly or unmethylated genomes indicating relationship between DNA methylation and clustered distribution of CpGs in the genomes. A similar result was obtained when exon, intron and intergenic region of different organisms were studied separately. Further CpG gap analysis were compared against five other related dinucleotides gaps (GpC, TpG, GpT, CpA and ApC) in differently methylated genomes and higher value of mean CG gaps and dispersion index was found in methylated human genome in a pronounced manner.

CHAPTER 1

INTRODUCTION

Introduction

Epigenetics is defined as mitotically and meiotically heritable changes in gene expression that do not involve a change in the DNA sequence. These changes are DNA methylation and histone modifications. These two processes plays important role in regulation of gene expression and various other biological processes like X-Chromosome inactivation, suppression of retroviral genes, development of tumor and cellular differentiation in vertebrate genomes. Histone modifications comprise of different modifications of histone proteins such as acetylation, methylation, phosphorylation and ubiquitination (Peter A. Jones *et al.*, 2007).

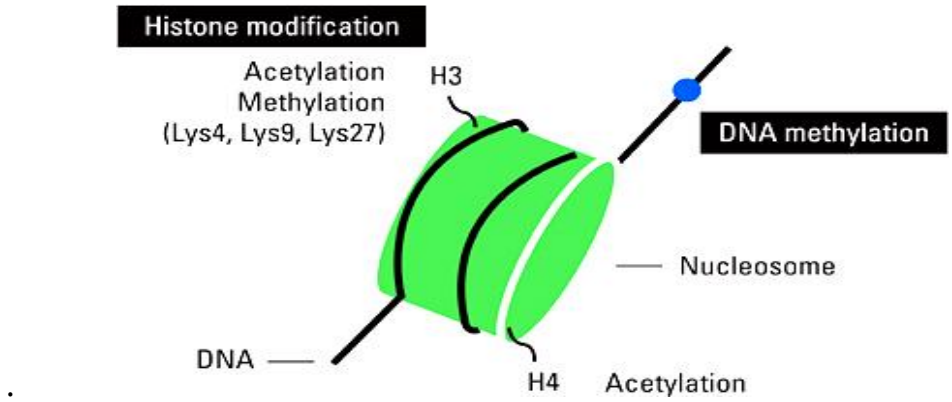

Fig. 1 Histone modification and DNA methylation (Lee *et al.*, 2009)

DNA methylation also takes place in prokaryotes. DNA methylation occurs at N-6 position of adenine and N-4 and C-5 position of cytosine in case of prokaryotes. In prokaryotes, DNA repair and restriction modification system are some of the well studied phenomena associated with DNA methylation. DNA methylation is an enzymatic process performed by enzyme family of methyltransferases. All DNA methyltransferases use *S*-adenosyl-L-methionine (AdoMet) as the source of the methyl group. Dmnt1, Dmnt2 and Dmnt 3a, Dnmt 3b are the methyltransfeases found in vergebrates and exhibit high sequence conservation. Dnmt1 has much higher specficity for hemimethylated sites and is responsible for maintenance of methylation in a post-replication manner whereas Dmt3a and Dnmt3b are the *de novo* methyltransferases and do not discriminated

between hemi- and unmethylated CpG substrates. In case of vertebrates 5-8% cytosine are methylated. Some times methylated cytosine is also called 5th base pair of vertebrate genome.

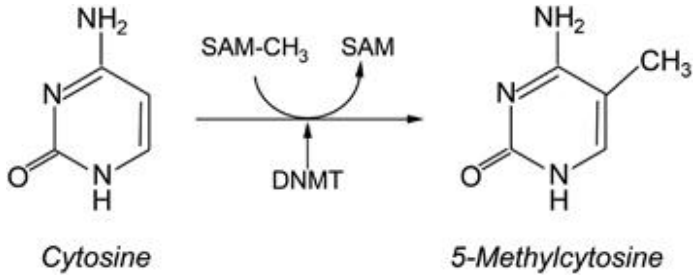

Fig. 2 Methylation at C^5 position of cytosine by methytransferases (Gibney *et al.*, 2010)

DNA methylation is a biochemical and enzymatic process which plays an important role in normal devolpment in higher organism. In X-chromosome inactivation DNA methylation plays very important role by in activation of genes on X-chromosome. In vertebrates, female have two X chromosomes and there is a need of inactivation of one X-chromosome out of two copies for the inactivation of the X-chromosome it is packed in transcriptionally inactive structure called heterochromatin. This X-chromosome remain inactive throughout the life time of cell (Gartler *et al*., 2001).

In eukaryotes, DNA methylation occurs at carbon 5 (C5) position of the cytosine ring by addition of methyl group catalyzed by DNA methyltransAferases. This generally occurs at the sequence 5' CG 3' which are also referred to as a CpG dinucleotide. As the cytosine is methylated, it may lead to change in gene expression as methylation may interfere with binding of sequence specific DNA binding proteins involved in regulation of transcription initiation. Gene regulation is also affected by chromatin remodeling induced by DNA methylation. Less frequently (typically during gametogenesis and early embryogenesis) reverse reaction also occurs causing demethylation of cytosine which may be caused by active DNA demethylation.

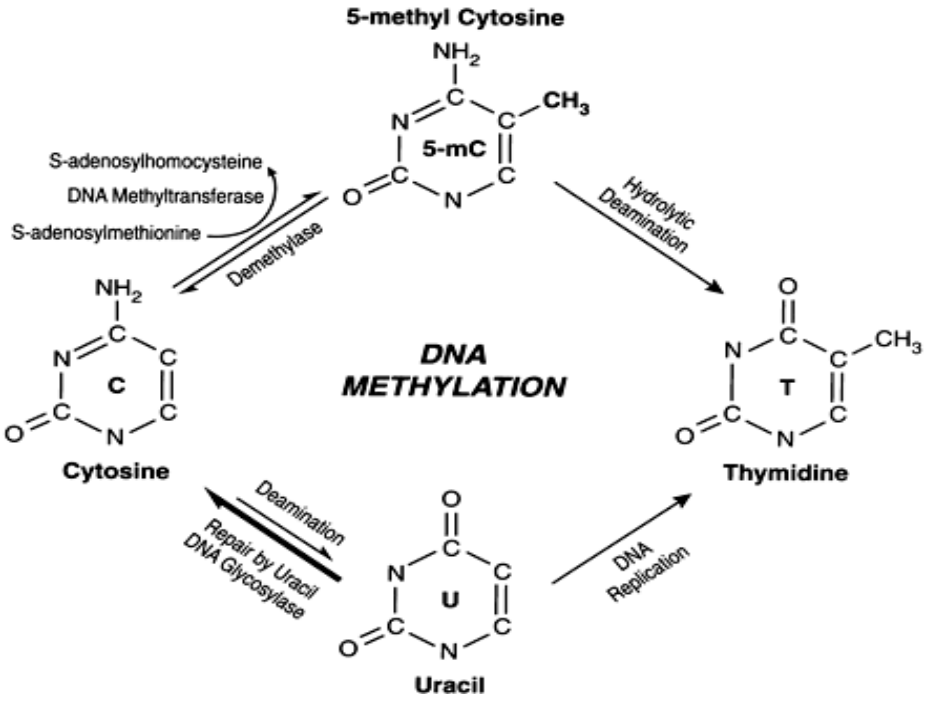

Fig. 3 Schematic representation of the biochemical pathways for cytosine methylation, demethylation, mutagenesis of cytosine and 5-metC (Singhal and Ginder 2013**)**

Methylated cytosine in CpG dinucleotides are highly mutable so CpG/CpG sites in genomic sequences can mutate into TpG/CpA sites. Due to this gradual but perpetual loss of CpG during the course of evolution, CpG sites are under representation in methylated eukaryotic genomes (A. Hermann *et al*., 2004). Methylated cytosine may change to thymidine by hydrolytic deamination (Fig. 3). Mutation at CpG occurs because 5meC is more susceptible than CpH because after deamination 5 methyl cytosine is converted into a thymine (a natural base in DNA) while cytosine is converted into uracil (not a natural base of DNA) and is replaced by cytosine via repair mechanism (Singhal and Ginder, 2013).

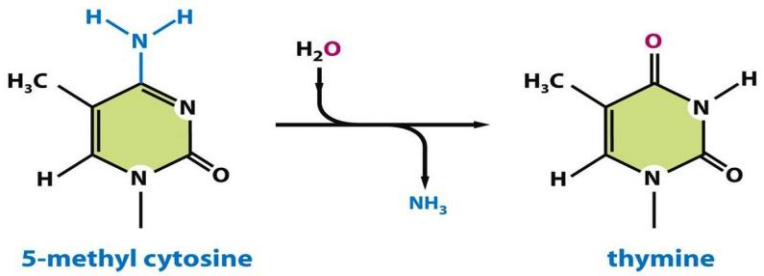

Fig. 4 Deamination of 5- methyl cytosine to thymine (Lea *et al.*, 2012)

Regulation of genes is necessary for development in higher eukaryotes. Methylation plays an important role in regulation of housekeeping and some tissue specific genes. Promoter region of most of the genes are associated with CpG islands and methylation of these CpG islands leads to gene turn off as shown in fig. 5. It has been observed that proper methylation is essential for cellular differentiation and embryonic development. There is significant difference in methylation level exist between different tissue types.

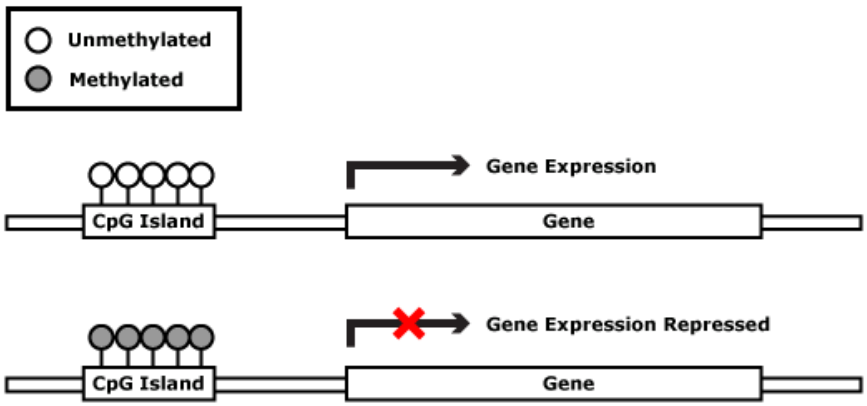

Fig. 5 DNA methylation role in gene expression (picture courtesy neurorexia.wordpress.com**)**

DNA methylation also has an important role in cancer and tumor development. In normal cells tumor suppressor genes are associated with hypomethylated CpG Island but due to hypermethylation of these areas transcriptional repression of TSG (tumor specific gene) is lost and this leads to cancer. It has been found that there is significant difference in DNA methylation level existing between normal cells and cancer cells. In cancer cells, methylation of CpG islands occurs and due to this transcriptional silencing of growth regulatory genes takes place and this leads to cancer.

Histone modification is also an epigenetic change, which helps in regulation of genes without altering DNA sequence. Histones are most abundant and highly conserved proteins in eukaryotes and used for nucleosome formation. Histone proteins undergo many types of modifications like acetylation, methylation, phosphorylation and ubiquitination. These modifications occurs in all familes of histones i.e. in H3, H4, H2A and H2B. These modifications lead to various biological processes like chromatin condensation, transcriptional activation, chromosomal stabilty, DNA repair, spermatogenesis, mitosis and meosis as shown in (fig. 6) (Jayani *et al.*, 2010)

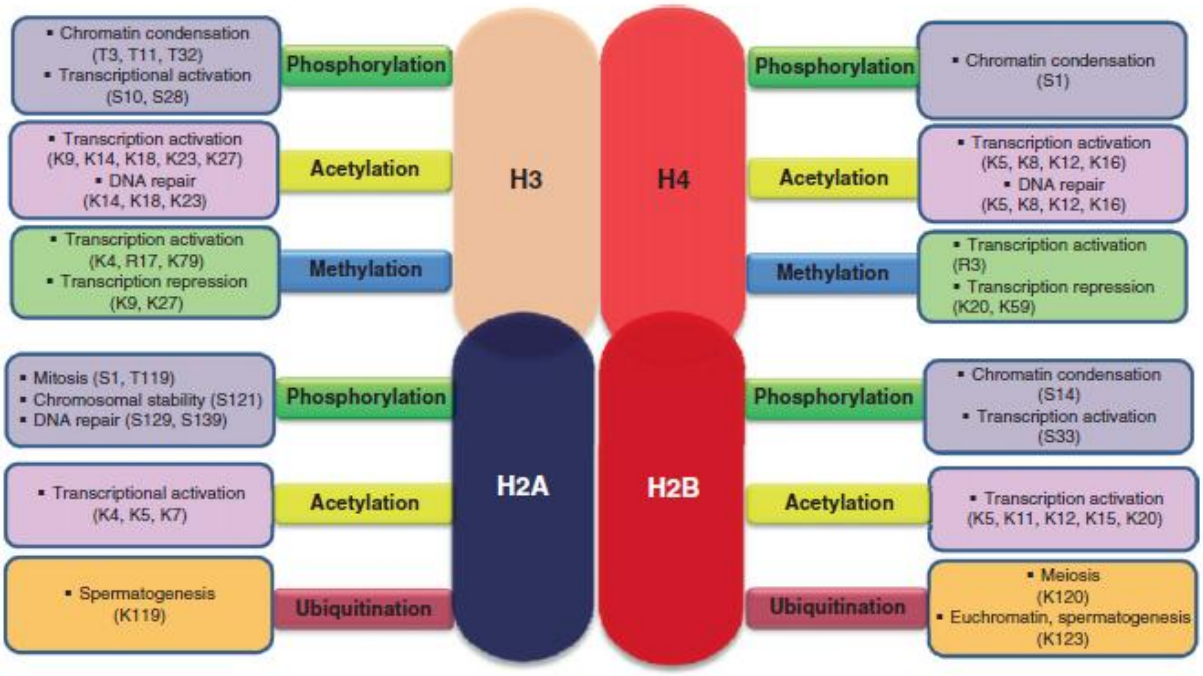

Fig. 6 Biological role of various histone modifications (Jayani *et al.*, 2010)

CpG islands: There are certain regions of DNA where CpGs are found in high abundance as compare to other region of genome. These regions are called **CpG islands**. CpG islands have high GC% and Observe/Expected (Obs/Exp) ratio of CpGs. 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.

CpG islands were initially defined by Gardiner-Garden et al. as $\geq$ 200 bp in length, having GC% $\geq$ 50% and $CpG_{Obs/Exp} \geq 0.60$ (Gardiner-Garden *et al.*,. 1987). A more accurate of description of CpG islands was later given as $\geq$ 500 bp long stretch of DNA having GC $\geq$ 55% and $CpG_{Obs/Exp} \geq 0.65$ with at least 100 bp distance between two adjacent CpG islands (Takai and Jones. 2001). Another algorithm has lately been used to define CpG island, a distance based algorithm for prediction of CpG islands. This algorithm is based on the physical distance between neighboring CpGs in DNA i.e. the distance between each CG in bulk DNA, in CpG island is different (Hackenberg et al., 2006). This algorithm is not depending upon the parameters like GC% or $CpG_{Obs/Exp}$. CpG islands are mostly associated with promoter region of house keeping genes and

some tissue specific genes. CpG islands helps in expression of genes and methylation of CpG islands leads to silencing of the gene expression.

CpG distribution

It is observed that CpGs are underrepresention in genomic sequences of those organism in which DNA methylation takes place. Additioally CpG distribution pattern near the 5` end of genes is different among vertebrates, invertebrates, plants and bacteria (Shimizu *et al.*,1997). CpG distribution is uneven in genomes of the species whose genomes undergo DNA methylation. The unevenness of CpG distribution is largely attributed to the CpG islands. However the distribution of CpGs may also vary at a finner level both within CpG islands (CGI) as well as in the non-CpG island regions (nCGI) i.e. rest of the genome. In present work the CG ditribution studies has been carried out by analysis of gaps (distance) between adjacent CGs in the genomic sequence using statistical methods, similar to the approach used by Hackenberg et al., 2006. The gap size on one hand represents (inverse of) frequency of CpGs and on the other hand enables one to study distribution of CpGs in 1-Dimensional space.

CHAPTER 2

REVIEW LITERATURE

Review of Literature

Epigenetics: Dr Alan Wolffe defined the term epigenetics as "heritable changes in gene expression that occur without a change in DNA sequence" (Wolffe and Matzke 1999). Epigenetics can be further described as stable alteration in gene expression that takes place during devoloplment and cell proliferation without any change in DNA sequence. Two processes are involved for the change in gene expression without altering DNA sequence i.e. DNA methylation and Histone modifications. DNA methylation is a very important epigenetic change which plays an important role in many biological processes like silencing the gene expression, X-chromosome inactivation and developmental process. DNA methylation also plays a role in uncontrolled growth of cells. Methylation generally occurs at C^5 position of cytosine in higher eukaryotes and N^6 position of adenine in prokaryotes. Histone modifications also change the expression of genes by acetylation, phosphorylation, ubiquitination and methylation of histone proteins (Paulsen and Smith 2001).

DNA methylation is a process in which methyl group is added to N-6 position of adenine, N-4 and C-5 position of cytosine. The enzymes that add methyl group to adenine or cytosine bases are called as DNA methyltransferases (Dnmt). All methyltransferases use S-adenosyl-L-methionine as source of methyl group being transferred to the DNA base. In case of prokaryotes DNA methylation takes place by three ways i.e methylation occurs at N-6 position of adenine, and N-4 and C-5 position of cytosine. In prokaryotes methylation has three major biological role: 1) distiction of self and non self DNA by adding methyl group to bacterial DNA, 2) direction of post replicative mismatch repair 3) control of DNA repliction and cell cycle. Doerfler *et al.,* 1997 proposed that *de novo* methylation constitutes a cellular defense mechanism to silence integrated foreign DNA or genes. In case of bacteria DNA methylation occurs at C-5 or N-4 position of cytosine and N-6 position of Adenine. In bacteria DNA methylation plays important role in protecting bacteria from foreign DNA such as viral DNA by restriction modification system (David A. Low *et al.,* 2001). Restriction and Modification (R-M) systems are bacterial protective systems that reduce horizontal transfer of DNA by recognizing incoming DNA as foreign while marking the host DNA as self (James *et al.,* 2011). This marking is done by adding methyl group to bacterial genome and this methyl group prevents the DNA from endonucleases that cleaves non methylated DNA of bacteriophage (fig. 7).

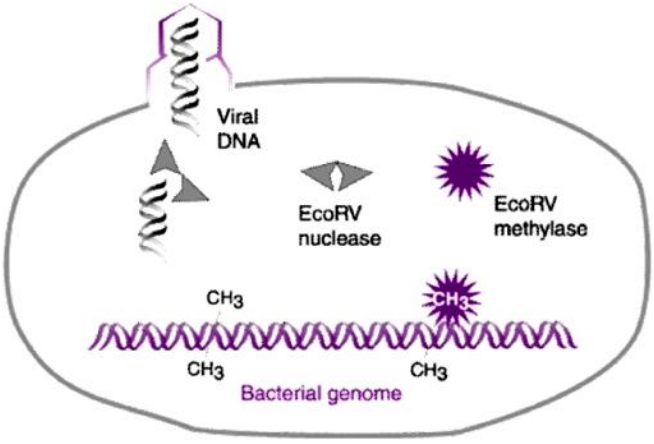

Fig. 7. The process of restriction and modification in bacteria (James *et al.*, 2010)

In eukaryotes DNA methylation takes place only on cytosines at position 5 in CpG sequence context. DNA methylation level is varying in all the eukaryotic species. It ranges from very low level in arthropods, intermediate in some protozoa and high level of DNA methylation in vertebrates. In vertebrate genomes the DNA methylation is carried out by family of enzymes, Dnmt1, Dnmt2, Dnmt3a, Dnmt3b. Dnmt1 is the maintainance methyltransferase that is responsible for coping DNA methylation pattern to the daughter stands during DNA replication (Hermann *et al.*, 2004).

Bestor (1992) and Pradhan et al., (1999) gave the fact that Dnmt 1 prefers to methylate only those new CpGs whose partners on parental stand already carry a methyl group. Thus a pattern of methylated and non methylated CpGs along a DNA strand tends to be copied, and this provides a way of passing epigenetic information between cell generations. Mammalian DNA methylation patterns are established in early development by de novo methyltransferases Dnmt 3a and Dnmat 3b (Okano et al., 1998) and then copied to somatic cells by the maintainance DNA methyl transferase Dnmt 1 (Hsieh *et al.*, 1999).

DNA methylation is of two types one is maintance methylation and another is *de novo* methylation. Okano et al., showed that de novo methylation takes place with the help of two enzymes known as Dnmt 3a and Dnmt 3b are highly expressed in early embryonic cells, and at this stage that most programmed de novo methylation events occur. Maintance methylation take place by the enzyme Dnmt 1. Maintenance methylation is the process which reproduces the DNA methylation patterns between cell generations. The mechanism of maintenance methylation depends upon semi conservative type of replication. This fact was given by Holliday and Pugh (1975).

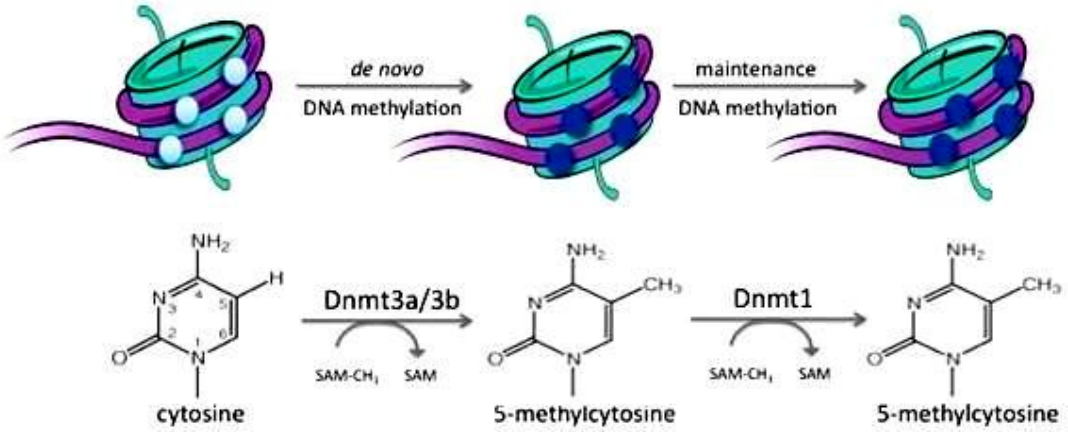

Fig. 8 Mechnanism *de novo* and mantenance methylation by DNA methyl transferase

During evolution, the dinucleotide CpG has been progressively eliminated from the genome of higher eukaryotes and is present at only 5% to 10% of its predicted frequency (Antequera and Bird, 1993). Cytosine methylation appears to have played a major role in this process, because most CpG sites lost represent the conversion through deamination of methylcytosines to thymines. Approximately 70% to 80% of the remaining CpG sites contain methylated cytosines in most vertebrates, including humans (Bird AP, 1995)

Epigenetic changes i.e. DNA methylation and histone modifications helps in gene regulation process. DNA methylation generally represses the expression of genes by altering protein-DNA interactions and via chromatin reomodelling. In eukaryotes DNA methylation plays very important role in various biological processes like genetic imprinting, X-chromosome inactivation, regulation of gene expression and in disease process (A. Razin and Cedar. 1991).
Adrian P Bird in 1980 suggested that in gene regulation, methylation in promoter region of gene supress or turn off gene expression. CpG island density is high in promoter region of gene and methylation of these CpG islands leads to turn off the gene expression.
Cancer is uncontrolled mitotic divison of cells and formation of tumor cause cancer. Formation of tumor is controlled by some genes known as tumor suppressor genes. Promoter of these genes are associated with CpG islands i.e CpG density is high in promoter region of TSG (tumor supperssor genes) genes. Methylation of these CpG islands cause the gene turn off and this leads to cancer. Alteration in DNA methylation is commen in a varity of tumors. All the epigenetic changes like hypermethylation of TSG genes leads to gene silencing and as a result formation of

tumor. The other reason of cancer or formation of tumor is hypomethylation of heterochromatin region. Hetrochromatin region of chromosome is generaly hypermethylated i.e. level of methylation is high in heterochromatin region of chromosomes. The hypomethylation of heterochromatin region cause genomic instability and expression of recessive genes which leads to tumor formation. Partha M. Das and Rakesh Singal gave this fact in 2004.

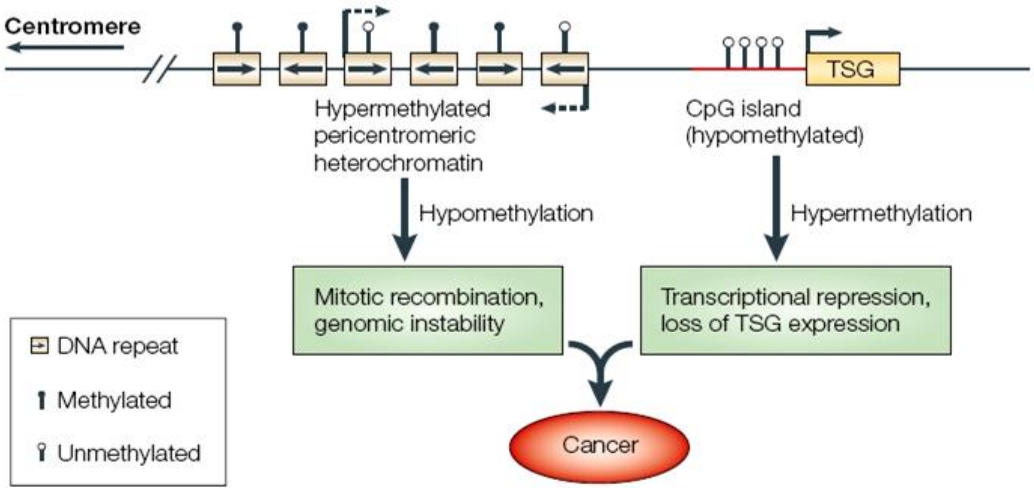

Fig. 9 Hypermethylation of CpG island leads to transcriptional repression and loss of TSG expression leads to cancer

Decreased levels of overall genomic methylation are common findings in tumorigenesis (Gama-Sosa 1983). This decrease in global methylation appears to begin early and before the development of frank tumor formation (Christman *et al.*, 1993). Apart from the overall genomic hypomethylation, specific oncogenes have been observed to be hypomethylated in human tumors.

Hanada *et al.*, 1993 observed a good inverse correlation between methylation and gene expression in the antiapoptotic bcl-2 gene in B-cell chronic lymphocytic leukemia and for the k-ras proto-oncogene in lung and colon carcinomas.

Methylation occurs at Cytosine nucleotide (C^5 position) of CpG dinucleotides of DNA in eukaryotes and this leads to defficiency of CpGs (Adrian P. Bird 1980). The extent of methylation is high in higher organisms as compared to lower organisms and due to this CpG dinucleotides, frequency is low in higher organisms as comparative to lower organisms. Methylation is the process in which methyl group is added to the cytosine and methylated cytosine is called as 5-methycytosine. In mammlian DNA 3 to 5% cytosine are methylated and

due to this sometimes 5-methylcytosine is considered as 5th base of mammalian DNA (Jeltsch *et al.*, 2004). CpG methylation leads to depletion in CpGs in DNA. The extent of CpG depletion show a positive correlation between CpG methylation and extent of TpG and CpA elevation (Smith *et al.*, 1983). There are some areas of genome which are non methylated and high CpG dinucleotides frequency; these areas are called as CpG islands. CpG dinucleotides are underrepresented in genome sequences of organisms, which methylated their DNA (M. Gardiner-Garden 1987). In these organisms, most of the cytosine residues are methylated and this methylation cause depletion of CpG dinucleotides over the course of evolution (Schorderet and Gartler, 1992). 5-methylecytosine residues are mutational hotspots so these are responsible for depletion in cytosine residue in vertebrate species over the course of evolution.

CpG island are those areas of genome which are non methylated and having GC% up to 55% and $CpG_{Obs/Exp}$ value is 0.65 (Takai and Jones, 2001). CpG islands are found most often near the 5` end of genes (Gardiner-Garden and Frommer, 1987) and often include gene regulatory elements. Discovery of CpG islands and their location in genome i.e. at 5` end of gene provided the evidence in support of hypothesis that DNA methylation is involved in gene regulation. Promoter region of genes having higher CpG density and methylation of cytosine in promoter region may lead to gene silencing. Methylation of cytosine is crucial process for regulation of genes in vertebrates and other higher organisms. CpG islands initially discovered on the basis of discordant patterns of digestion of genomic DNA by restriction enzymes that differed only in their sensitivity to cytosine methylation. The enzymes MspI and HpaII digest or cut the DNA at its recognition sequence and gives sticky ends but HpaII is sensitive to methylation and this enzyme will not digest the DNA at the restriction site if the cytosine is methylated (CC*) as shown in fig. 10. CpG islands are those areas of genome which are unmethylated and due to this reason by the activity of restriction enzymes, CpG islands are detected.

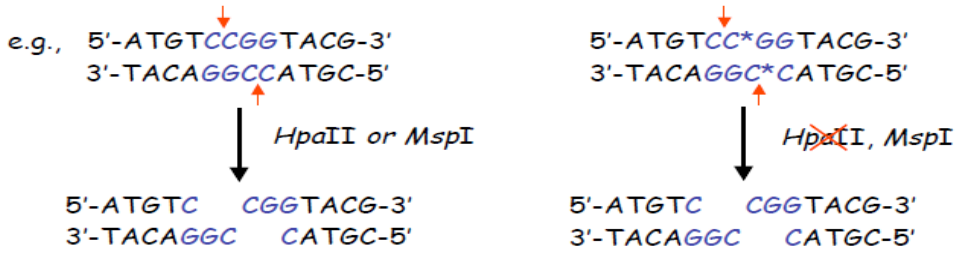

Fig. 10 Restriction digestion sensitivity of MspI and HpaII toward methylated cytosine

Leng Han *et al.,* (2004) gave the concept that CpG density in genomic DNA depends upon various genomic features like chromosome number, chromosome size, GC content of chromosome, $CpG_{Obs/Exp}$ ratio and recombination rate. It has been observed that there is a positive correlation between CpG island density and GC content, O/E ratio and chromosome pair. However as chromosome size increases the CpG island density tends to decrease. CGI density also correlated with recombination rate and other genetic factors like body temperature and life span of an organism. There is a positive correlation exist between CpG island density and recombination rate. Recombination rate increases as we move centromeric region to telomeric region of chromosome and due to this is the reason of high CpG island density in telomeric region of chromosome. CpG island density also depends on genetic factors of an organism like body temperature and life span. It has been observed that there is significant correlation exists between CpG island density and body temperature of organism and insignificant correlation exist between CpG island density and life span of an organism. CpG distribution is uneven or random in genomes of all the species. Approximately 60–90% of all CpG sequences in the genome are methylated, while unmethylated CpG dinucleotides are mainly clustered in the CpG rich sequence, termed CpG island, of the gene promoter region (Ng and Bird, 1999). The distribution affects the DNA methylation level in all species. CpG distribution pattern in methylated and non methylated genomes was studied on the basis of $CpG_{Obs/Exp}$ ratio (Shimizu *et al.,*1997). The $CpG_{Obs/Exp}$ value is calculated by the formula ((No. of CpG/ (Num of C × No. of G)) × Total number of nucleotides in the sequence). Lower value of $CpG_{Obs/Exp}$ ratio shows that CpGs are under repressed in genomic DNA and due to this fact DNA methylation level is more in these genomes. The $CpG_{Obs/Exp}$ ratio ranges from very low value in vertebrates, through intermediate range in protozoa to high value in *E. coli*. The genomes having very low $CpG_{Obs/Exp}$ ratio are heavily methylated and level of methylation goes on decreasing as $CpG_{Obs/Exp}$ ratio increases.

CHAPTER 3

SCOPE OF STUDY

Scope of study

DNA methylation is playing significant role in very important biological phenomena including tumourigenesis via epigenetic gene regulation. Another interesting phenomenon associated with DNA methylation is gradual but continuous change in genome fine structure via loss of CpGs (CpG/CpG → TpG/CpA). Methylation of cytosine in vertebrate genomes causes depletion of cytosine base and due to this reason, CpGs are under repressed in vertebrate genome. Over evolutionary periods the loss of CpG with higher propensity to get methylated probably lead to certain unmethylated regions relatively richer in CpGs and are known as CpG islands.

As it is well known that CpG islands are associated with 5' regions of genes and methylation of promoters may lead to gene silencing it is interesting to learn why certain CpGs are methylated more often than certain others. Several reports have attempted to understand this paradox. Similarly it is interesting to know that a different kind of CpG distribution (high GC% and high $CpG_{Obs/Exp}$ values) found in CpG islands and they are usually not methylated. Based on these facts it has been attempted to study CpG distributions in different genomes (methylated as well as non-methylated). It is important to learn it because, CpG distribution pattern may also affects the DNA methylation level in genome. CpG distribution in the genome can be present either in the form of clusters, for example CpG islands or dispersed in genome. It is not necessary that all clusters of CpGs are CpG Island. The study of CpG distributions in genomes is a preliminary work to understand paradoxical cause and effect relationship between CpG distribution and their methylation.

CHAPTER 4

OBJECTIVES

Objectives

- Comparison of CpG gap size with gap size of other related dinucleotides
- Comparison of CpG gap size in differently methylated genomes
- To investigate CpG gap size difference in CGI and nCGI regions and thereby study the CpG distributions of these regions
- Comparison of CpG gap size in human and chimpanzee genomes in context of exon, intron and intergenic regions

CHAPTER 5

MATERIAL AND METHODS

Material and Methods

DATA source

DNA sequences were downloaded from National Centre for Biotechnology information (NCBI) (www.ncbi.nlm.nih.gov). To study Distribution of CpGs in genome, eleven different species' DNA sequences of length 3.3×10^6 bp (except S. *cerevisiae* which was 1.3×10^6 bp) were selected randomly from GenBank database. DNA sequences was taken randomly from NCBI for different species.

Data: DNA sequences:

Table: 1 DNA sequences of different species taken from GenBank to perform distribution study

Species	Accesion No.	Chromosome No.	Start position (bp)	End position (bp)	Total length of sequence (bp)	Misssed sequence (bp)
Homosapiens	NT_004487.19	Chr. 1	10000000	13300000	3300000	0
Pan troglodytes	NT_006471.3	Chr. 4	20000000	23326985	3300000	26985
Mus musculus	NC_000082.6	Chr. 16	30000000	33300000	3300000	0
Rattus norvegicus	NC_005100.3	Chr. 1	20000000	23514624	3300000	214624
Danio rerio	NC_007123.5	Chr. 12	10000000	13300000	3300000	0
Takifugu rubripes	HE602535.1	Chr. 1	10000000	13300000	3300000	0
Drosophila melanogaster	NT_033779.4	Chr. 2	3000000	6300000	3300000	0
Anopheles gambiae	NT_078265.2	Chr. 2	10000000	13310232	3300000	10232
Caenorhabditis elegans	NC_003281.9	Chr. 3	2000000	5300000	3300000	0
Saccharomyces cerevisiae	NC_001136.10	Chr. 4	100000	140000	1300000	0
Escherichia coli	NC_010473.1	__	1000000	4300000	3300000	0

Sequence analysis tools:

- **MS Word:** Microsoft word was used to analyze the DNA sequences with the help of tools such as 'find,' 'replace' and recording macros.

- **MS Excel:** Microsoft Excel spreadsheet was used for computation and statistical analysis of data of analyzed sequence. Apart from standard mathematical and statistical functions the 'LEN' (text function) was used to count the number of characters in a cell of spreadsheet.

- **CpG island searcher (http://cpgislands.usc.edu/):** CpG island searcher is an online software which is used to screen DNA sequences for CpG islands. This software detect the CpG islands in the CpG islands from given genomic sequence. This software detects the CpG islands on the basis of GC% and CpGobs/exp values of DNA sequence. The CpG island searcher was designed D.Takai (Takai and Jones, 2001).

Methods

Distribution of CpGs in different species

- To perform CpG distribution study, DNA sequences of length 3.3 million bp (base pair) was downloaded from Genbank database for elevan different species and paste in MS Word. The sequences were continuous and selected at random.

- Some of the sequences carring stretches of 'N' (regions not sequenced) were removed and additional equal sized sequence was considered at the 3' end.

- Using the algorithm mentioned below all the CpG gap sizes (number of nucleotides between two adjacent CpGs) in genomic sequences were determined.

- In case of sequences containing stretch of N, the CpG gap size between last CpG (before N stretch) and following CpGs was not considered.

Algorithm

The DNA sequence was pasted in MS word file and converted into one single word of length equal to the number of base pair in the sequence. This process was performed by recording a macro consisting of repeated steps of shifting the cursor to the end of each existing word (line of sequence) followed by delete function to bring the first character of the next word (line of sequence) next to the last character of the word. For subsequent processing of the sequence another macro was recorded which consist of replacement of all CGs by 1 followed by replacement of G, A, T and C with 0 in separate steps. The replacements convert the DNA sequence into a string of '1's and '0's where '1's represent CGs and 0s represented bases of the sequence flanking all the CpGs.

Then another macro was recorded for cutting the string of ‘1’s and ‘0’s just upstream of each ‘1’. This resulted in cutting the string into ‘n+1’ number of smaller strings, where n = number of ‘1’s present in the uncut string. The resultant cut strings each beginning with ‘1’ followed by ≥ 0 ‘0’s represented a CG followed by all the bases reprented by ‘0’s till the next CG. This was achieved by recording macro with following MS word functions in given sequence:

- Find ‘1’
- Shift the cursor one step back just upstream of the 1
- Enter

The cut strings beginning with ‘1’ followed by ≥ 0 ‘0’s were copied and pasted into a MS excel spreadsheet column. Using ‘LEN’ function in ‘Text’ category (returns the number of characters in a text string) total number of character in each string were determined. Subtracting 1 from numbers (for ‘1’s representing CGs) the number of bases present between the CG and its neighbouring CG were calculated. This is how CpG gap sizes were computed.

Algorithm

ATGCGTACGTAGCTAATCGATAGGACGCTA

(DNA Sequence)

ATG1TA1TAGCTAAT1ATAGGA1CTA

(CGs replaced by '1')

000100100000000100000001000

(Other nucleotides replaced by '0')

000

100

100000000

1000000

1000

(Cutting the string)

3

3

9

7

4

(Counting characters in MS Excel using 'LEN' function)

Statistical analysis

Coefficient of variance: A statistical measure of the dispersion of data points in a data series around the mean. It is calculated as follows:

$$\text{Coefficient of Variation} = \frac{\text{Standard Deviation}}{\text{Expected Return}}$$

Coefficient of variance = σ / μ

The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from each other.

Dispersion index (D): It is also known as variance by mean ratio or coefficient of dispersion. It is a measure of dispersion of a probability distribution. Dispersion index is used to quantify whether the set of observance are dispersed or clustered.

Dispersion index is defined as ratio of variance by mean.

Dispersion index = variance / mean

Or

$$D = \sigma^2 / \mu$$

CpG distribution in CpG islands and non CpG island areas of genome

All the sequences were screened for CpG islands using CpG island searcher with following conditions.

- Length of CpG island ≥ 500 bp
- GC% ≥ 55%
- $CpG_{obs/exp}$ ≥ 0.65
- Gap between adjacent CpG island ≥ 100 bp

The maximum intake of software (CpG island searcher) is 5×10^5 bp. Therefore all the sequences were broken into 10,000 bp overlapping fragment of 5×10^5 bp size so as to avoid missing CpG islands at the cut site generating the fragments. Having classified this sequence into CGI and nCGI sequences, bith the classes of the sequences were used to determine CpG gap sizes seperately as mentioned above.

Distribution of CpGs in Intron, Exon and intergenic region of genes

Three genomic sequence stretches of human and 23 genomic sequence stretches of chimpanzee (details of genes and number of introns and exons in them are shown in Table 2) were used for generating data of CpG gap size in introns, exons and intergenic region. The sequences of both the species were classified into exons, introns and intergenic (entire sequence except introns and exons) using information of start and end position of exons, introns and intergenic regions obtained from evidence viewer. As mentioned above, the sequences before first and after last CpGs were ignored from each intron, exon and intergenic region before computing CpG gap sizes.

Table. 2 Genes of *Homo sapiens* studying distribution in exon, intron and intergenic regions

Sr. No	Gene name	Chromosome No.	Accesion. No.	Total length (bp)	Start position (bp)	End position (bp)	No. of Exon	No. of intron
1	LOC100133331	1	NT_004350.19	5393	144763	139371	3	2
2	FAM41C	1	NT_004350.19	9532	291214	281683	3	2
3	SAMD11	1	NT_004350.19	19641	339553	358993	12	11
4	HES4	1	NT_004350.19	2011	414584	412574	3	2
5	C1orf159	1	NT_004350.19	35339	495430	530768	14	13
6	TNERSF18	1	NT_004350.19	4002	621121	617120	5	4
7	FAM132A	1	NT_004350.19	5077	661134	656058	8	7
8	UBE2J2	1	NT_004350.19	20743	688266	667524	10	9
9	SCNN1D	1	NT_004350.19	12394	694048	706441	15	14
10	MXRA8	1	NT_004350.19	6645	772947	766303	9	8
11	AURKAIP1	1	NT_004350.19	2509	789850	787342	4	3
12	ANKRD65	1	NT_004350.19	3825	835856	832032	4	3
13	TMEM88B	1	NT_004350.19	2460	839740	842199	2	1
14	VWA 1	1	NT_004350.19	8160	849135	857294	3	2
15	ATAD3C	1	NT_004350.19	21270	863301	884570	13	12
16	SSU72	1	NT_004350.19	34010	989294	955285	6	5
17	MMP23B	1	NT_004350.19	3271	1045792	1049062	6	5
18	MMP23A	1	NT_004350.19	2670	1109610	1112279	6	5
19	C1orf222	1	NT_004350.19	6779	1338400	1331622	10	9
20	GABRD	1	NT_004350.19	12225	1429000	1441224	9	8
21	SKI	1	NT_004350.19	82319	1638366	1720684	9	8
22	LOC115110	1	NT_004350.19	3726	1963316	1959591	3	2
23	TTC34	1	NT_004350.19	134224	2185262	2185262	7	6
24	SMIM1	1	NT_004350.19	3995	3167584	3171578	4	3
25	LRRC47	1	NT_004350.19	17085	3192100	3175016	7	6
26	LOC728716	1	NT_021937.19	12772	5004	17775	4	3
27	LOC284661	1	NT_021937.19	13434	476443	489876	6	5
28	AJAP1	1	NT_021937.19	129547	719437	848983	7	6
29	RPL22	1	NT_021937.19	15400	2264811	2249412	4	3
30	LINC00337	1	NT_021937.19	2434	2302200	2304636	2	1
31	HES3	1	NT_021937.19	2187	2308584	2310770	4	3
32	GPR153	1	NT_021937.19	14430	2326167	2311738	6	5
33	HES2	1	NT_021937.19	5488	2479624	2485111	4	3
34	KLHL21	1	NT_021937.19	12946	2668061	2655116	4	3
35	PHF13	1	NT_021937.19	11138	2678088	2689225	4	3
36	THAP3	1	NT_021937.19	11522	2689257	2700778	7	6

37	VAMP3	1	NT_021937.19	10964	3835661	3846624	5	4
38	UTS 2	1	NT_021937.19	6680	3912004	3918683	5	4
39	PARK7	1	NT_021937.19	24429	4026046	4050474	7	6
40	ERRFI1	1	NT_021937.19	15415	4060697	4076111	3	2
41	SLC45A1	1	NT_021937.19	20638	4388722	4409359	8	7
42	ENO1-AS1	1	NT_021937.19	1851	4943226	4945076	2	1
43	GPR157	1	NT_021937.19	25554	5194361	5168808	4	3
44	H6PD	1	NT_021937.19	37334	5299195	5336528	6	5
45	SPSB1	1	NT_021937.19	77451	5357273	5434723	5	4
46	CTNNBIP1	1	NT_021937.19	62783	5975448	5912666	7	6
47	LZIC	1	NT_021937.19	13851	6007958	5994108	7	6
48	NMNAT1	1	NT_021937.19	42871	6007818	6050688	6	5
49	RBP7	1	NT_021937.19	19624	6061587	6081210	5	4
50	DFFA	1	NT_021937.19	12826	6537745	6524920	5	4
51	TARDBP	1	NT_021937.19	13671	7077011	7090681	7	6
52	AGTRAP	1	NT_021937.19	15487	7800474	7815960	6	5
53	NPPB	1	NT_021937.19	2272	7924124	7921853	3	2
54	KIAA2013	1	NT_021937.19	7641	7991617	7983977	3	2
55	AADACL4	1	NT_021937.19	23332	8708898	8732229	4	3
56	AADACL3	1	NT_021937.19	13409	8780450	8793858	4	3
57	C1orf158	1	NT_021937.19	15740	8810495	8826234	4	3
58	PRAMEF12	1	NT_021937.19	3866	8839316	8843181	3	2
59	PRAMEF1	1	NT_021937.19	6032	8855878	8861909	5	4
60	PRAMEF11	1	NT_021937.19	7597	8896396	8888800	4	3
61	HNRNPCL1	1	NT_021937.19	2118	8913710	8911593	2	1
62	PRAMEF2	1	NT_021937.19	5624	8921273	8926896	4	3
63	PRAMEF4	1	NT_021937.19	7793	8951157	8943365	4	3
64	PRAMEF10	1	NT_021937.19	6168	8963226	8957059	4	3
65	PRAMEF7	1	NT_021937.19	3521	8981845	8985365	4	3
66	PRAMEF22	1	NT_021937.19	3639	9039875	9043513	3	2
67	LOC440563	1	NT_021937.19	2167	9189458	9187292	2	1
68	PRAMEF3	1	NT_004610.19	4297	12180	7884	3	2
69	PRAMEF5	1	NT_004610.19	10039	39507	49545	4	3
70	PRAMEF8	1	NT_004610.19	4920	71253	66334	4	3
71	PRAMEF9	1	NT_004610.19	7816	100864	108679	4	3
72	PRAMEF13	1	NT_004610.19	6043	133144	127102	4	3
73	PRAMEF18	1	NT_004610.19	4317	158057	153741	3	2
74	PRAMEF16	1	NT_004610.19	3807	174942	178748	3	2
75	PRAMEF21	1	NT_004610.19	5781	201651	207431	3	2
76	PRAMEF15	1	NT_004610.19	7815	321661	329475	4	3

77	PRAMEF14	1	NT_004610.19	6043	353999	347957	4	3
78	PRAMEF19	1	NT_004610.19	4317	378893	374577	3	2
79	PRAMEF17	1	NT_004610.19	3777	395776	399552	3	2
80	PRAMEF20	1	NT_004610.19	11697	416595	428291	4	3
81	LRRC38	1	NT_004610.19	39598	520730	481133	2	1
82	EFHD2	1	NT_004610.19	21249	2416079	2437327	4	3
83	FLJ37453	1	NT_004610.19	14733	2855130	2840398	2	1
84	C1orf64	1	NT_004610.19	3250	3010419	3013668	2	1
85	HSPB7	1	NT_004610.19	5563	3025773	3020211	2	1
86	REG1	1	NT_004610.19	6278	3244147	3237870	5	4
87	SZRD1	1	NT_004610.19	31858	3373271	3405128	5	4
88	ACTL8	1	NT_004610.19	72551	4761496	4834046	3	2
89	TAS1R2	1	NT_004610.19	20863	5866643	5845781	6	5
90	HTR6	1	NT_004610.19	15076	6671468	6686543	2	1
91	PLA2G2E	1	NT_004610.19	4111	6930598	6926488	4	3
92	PLA2G5	1	NT_004610.19	22494	7076389	7098882	3	2
93	PLA2G2F	1	NT_004610.19	11857	7145511	7157367	4	3
94	PLA2G2C	1	NT_004610.19	12004	7182175	7170172	3	2
95	UBXN10	1	NT_004610.19	8165	7192266	7200430	2	1
96	CAMK2N1	1	NT_004610.19	4645	7493216	7488572	2	1
97	MUL1	1	NT_004610.19	9534	755162	7505629	4	3
98	CDA	1	NT_004610.19	30758	7595132	7625889	5	4
99	LOC100506801	1	NT_004610.19	7380	8299471	8306850	3	2
100	LINC00339	1	NT_004610.19	6810	9031395	9038204	3	2

Table. 3 Genes of *Pan troglodytes* to study distribution in exon, intron and intergenic regions

Sr. no	Gene	Chromosome no.	Accession no.	Total length (bp)	Start position (bp)	End position (bp)	No. of exons	No. of introns
1	HES4	1	NW_003456453.1	1958	63798	61841	4	3
2	ISG15	1	NW_003456453.1	1869	79580	81448	2	1
3	TNFRSF18	1	NW_003456453.1	10140	71976	61837	5	4
4	LOC100611076	1	NW_003456453.1	2620	223321	220702	3	2
5	PRDM16	1	NW_003456463.1	377617	44304	421920	3	2
6	LOC100611664	1	NW_003456464.1	13346	482397	495742	6	5
7	LOC100612135	1	NW_003456466.1	11740	293251	281512	2	1
8	LOC100612505	1	NW_003456469.1	16174	387187	371014	4	3
9	VAMP3	1	NW_003456469.1	10893	1973248	1984140	5	4
10	ERRFI1	1	NW_003456469.1	15520	2231725	2216206	3	2
11	LOC749268	1	NW_003456469.1	14793	3341810	332708	3	2
12	SPSB1	1	NW_003456470.1	14652	181214	195865	2	1
13	RBP7	1	NW_003456472.1	20123	612077	632199	5	4
14	ANGPTL7	1	NW_003456473.1	7614	412059	419672	5	4
15	UBIAD1	1	NW_003456473.1	16502	495811	512312	2	1
16	LOC737686	1	NW_003456477.1	1893	1893	1	2	1
17	LOC737847	1	NW_003456482.1	3450	3144	6593	3	2
18	LOC100611253	1	NW_003456495.1	2390	1	2390	2	1
19	PRAMEF19	1	NW_003456496.1	4371	10167	5797	3	2
20	PRAMEF17	1	NW_003456496.1	3745	27713	31457	4	3
21	LOC100611522	1	NW_003456496.1	7426	54736	62164	4	3
22	LRRC38	1	NW_003456496.1	41649	155178	113530	2	1
23	PRDM2	1	NW_003456496.1	17414	413438	430581	4	3
24	LOC740552	1	NW_003456497.1	5373	1537422	1542794	2	1
25	HSPB7	1	NW_003456497.1	6361	1553524	1547164	3	1
26	LOC747486	1	NW_003456497.1	6540	1575424	1568885	3	1
27	LOC456475	1	NW_003456497.1	32363	1908921	1941283	4	3
28	ACTL8	1	NW_003456502.1	4847	938754	943600	2	1
29	PLA2G2E	1	NW_003456506.1	4111	1901818	1897708	4	3
30	PLA2G2D	1	NW_003456506.1	8418	2099317	2090900	4	3
31	PLA2G2F	1	NW_003456506.1	11842	2118413	2130254	4	3
32	PLA2G2C	1	NW_003456506.1	11807	2155013	2143207	4	3
33	UBXN10	1	NW_003456506.1	10896	2165798	2176693	2	1
34	CAMK2N1	1	NW_003456507.1	4631	102350	97720	2	1
35	MUL1	1	NW_003456507.1	9536	124554	115019	4	3
36	LOC749172	1	NW_003456509.1	12858	584911	597768	3	1

37	C1QA	1	NW_003456509.1	8167	1208320	1216486	3	2
38	C1QC	1	NW_003456509.1	5736	1219627	1225362	2	1
39	C1QB	1	NW_003456509.1	10488	1229856	1240343	3	2
40	LOC100612472	1	NW_003456509.1	6356	1973534	1967179	3	2
41	LOC746089	1	NW_003456511.1	54590	551621	606210	5	4
42	LOC735630	1	NW_003456513.1	10752	259289	270040	5	4
43	LOC735740	1	NW_003456513.1	25307	296513	271207	3	2
44	PAQR7	1	NW_003456513.1	16674	314500	297827	2	1
45	LOC736983	1	NW_003456513.1	15212	557520	572731	5	4
46	GRRP1	1	NW_003456513.1	5207	606145	611351	3	2
47	ZNF593	1	NW_003456513.1	2597	617309	619905	3	2
48	CD52	1	NW_003456513.1	3608	770297	773904	4	3
49	NR0B2	1	NW_003456513.1	3391	1361100	1357710	2	1
50	LOC737621	1	NW_003456513.1	11015	1407696	1396682	4	3
51	LOC456668	1	NW_003456513.1	9249	1462195	1452947	2	1
52	TMEM222	1	NW_003456514.1	10086	149993	160078	2	1
53	GPR3	1	NW_003456514.1	3971	222800	226770	2	1
54	G1P3	1	NW_003456514.1	6777	504224	497448	5	4
55	LOC456679	1	NW_003456514.1	16379	703589	719967	6	5
56	XKR8	1	NW_003456514.1	9379	796543	805921	5	4
57	ATPIF	1	NW_003456514.1	2836	1079287	1082122	2	1
58	MED18	1	NW_003456514.1	8189	1173542	1181730	4	3
59	SNHG	1	NW_003456514.1	4093	1362170	1366262	3	2
60	LOC740014	1	NW_003456514.1	5289	1444088	1438800	3	2
61	LOC469251	1	NW_003456514.1	3782	1453148	1456929	3	2
62	OPRD1	1	NW_003456514.1	53223	1688382	1741604	3	2
63	SDC3	1	NW_003456518.1	10387	70359	59973	5	4
64	NKAIN1	1	NW_003456518.1	9838	398002	388165	6	5
65	FABP3	1	NW_003456518.1	27442	582036	5544595	4	3
66	PEF1	1	NW_003456518.1	16138	861231	845094	6	5
67	PTP42	1	NW_003456518.1	32333	1161871	1161871	6	5
68	LOC742435	1	NW_003456519.1	2448	52625	55072	2	1
69	ZBTB8B	1	NW_003456522.1	25577	35033	60609	4	3
70	ZBTB8A	1	NW_003456522.1	12320	169079	181398	4	3
71	HPCA	1	NW_003456524.1	9296	164639	173934	4	3
72	LOC100612077	1	NW_003456524.1	9209	183853	174645	5	4
73	MT1A	1	NW_003456524.1	160041	7150707	7134667	4	3
74	GJB5	1	NW_003456524.1	4278	2032666	2036943	2	1
75	GJB4	1	NW_003456524.1	7145	2037376	2044520	3	2
76	GJA4	1	NW_003456524.1	4175	2068282	2072456	2	1

77	LOC744133	1	NW_003456524.1	6198	2140530	2134333	3	2
78	LOC100612862	1	NW_003456524.1	4644	2264355	2259712	3	2
79	LOC100612967	1	NW_003456524.1	16041	7150707	7134667	4	3
80	TFAP2E	1	NW_003456524.1	22034	2867803	2889836	2	1
81	LOC456744	1	NW_003456524.1	6840	3015369	3008530	2	1
82	COL8A2	1	NW_003456524.1	5810	3399351	3393542	2	1
83	LSM10	1	NW_003456524.1	5338	3708068	3702731	2	1
84	OSCP1	1	NW_003456524.1	24160	3757042	3732883	5	4
85	LOC748170	1	NW_003456524.1	20529	4794837	4774309	3	2
86	SNIP1	1	NW_003456524.1	20692	4878751	4858060	4	3
87	DNALI1	1	NW_003456524.1	11343	4880540	4891882	4	3
88	LOC456761	1	NW_003456524.1	10598	5017280	5006683	4	3
89	LOC100615930	1	NW_003456524.1	16041	7150707	7134667	4	3
90	LOC100608031	1	NW_003456524.1	5540	5093073	5087534	3	2
91	MANEAL	1	NW_003456524.1	9511	5121796	5131306	4	3
92	LOC456765	1	NW_003456524.1	2462	5136445	5128906	2	1
93	FHL3	1	NW_003456524.1	9674	5336210	5326537	5	4
94	UTP11L	1	NW_003456524.1	12413	5342643	5326537	8	7
95	RRAGC	1	NW_003456524.1	22460	6189589	6167130	7	6
96	LOC100609944	1	NW_003456524.1	8665	6211589	6202925	2	1
97	AKIRIN1	1	NW_003456524.1	19695	6333806	6353500	5	4
98	NDUFS5	1	NW_003456524.1	6010	6374709	6380718	2	1
99	HEYL	1	NW_003456524.1	14714	7013736	6999013	6	5
100	HPCAL4	1	NW_003456524.1	13924	704476	7052553	5	4

CHAPTER 6

RESULTS AND DISCUSSION

Results

Comparison of CpG gap size with gap size of other related dinucleotides

To perform the distribution study of dinucleotides, the DNA sequence of 3.3×10^6 bp was taken from GenBank database. The dinucleotide gap sizes i.e. number of nucleotides between adjacent dinucleotides, for example, CpG were determined by using algorithm mentioned in the methods. Gap sizes of CpG and five other related dinucleotides, GpC, TpG, GpT, CpA and ApC were determined for three differently methylated genomes, *Homo sapiens, Drosophila melanogaster* and *C. elegans*. The other five dinucleotides were selected for comparison as methylated CpGs are mutated into TpGs and CpA (CpG/CpG → TpG/CpA) and GpC, GpT and ApC are the dinucleotides with same base composition when compared to CpG, TpG and CpA respectively. Statistical analysis was performed on the data obtained from gap lengths of mentioned dinucleotides. Similarly arithmetic mean, coefficient of variance and index of dispersion was calculated from the gap size data obtained from analyzed sequence.

Table. 4 Statistical analysis of distribution of dinucleotide in *Homo sapiens*

DNA dinucleotides	Arithmetic Mean	Coefficient of variance	Dispersion Index (DI)
CpGs	96.58	161.46	251.81
GpCs	21	121.12	30.90
GpTs	18.68	114.39	24.45
ApCs	17.76	113.63	22.94
TpGs	12.24	111.54	15.23
CpAs	11.74	111.91	14.71

Table. 5 Statistical analysis of distribution of dinucleotide in *Drosophila melanogaster*

DNA dinucleotides	Arithmetic Mean	Coefficient of variance	Dispersion Index (DI)
CpGs	21.86	126.32	34.89
GpCs	15.12	121.62	22.37
GpTs	17.45	114.67	22.95
ApCs	17.24	115.04	22.82
TpGs	12.52	113.37	16.09
CpAs	12.37	112.88	15.76

Table. 6 Statistical analysis of distribution of dinucleotide in *C. elegans*

DNA dinucleotides	Arithmetic Mean	Coefficient of variance	Dispersion Index (DI)
CGs	28.83	127.90	47.177
GCs	28.19	127.68	45.97
GTs	20.44	125.56	32.24
ACs	20.26	118.88	28.64
TGs	14.00	110.82	18.39
CAs	15.00	113	19.18

The mean CpG gap size was found to be nearly five times higher than mean GpC gap size in human (methylated genome) while the difference was only moderate in *D. melanogaster* (poorly methylated genome) and insignificant difference in *C. elegans* (unmethylated genome). This clearly reflects varying degree of underrepresentation of CpGs in human and *D. melanogaster* in

comparison to *C.elegans* where CpG is not underrepresentated. Similar trends can be observed for other four dinucleotides gap sizes also. Interestingly TpG and CpA gap sizes are lower when compared to GpT and ApC (same base composition) gap sizes that may be explained by moderate overrepresentation of TpGs and CpAs resulting from loss of CpGs. Though this observation fits well with methylated genomes but cannot be explained in *C. elegans*. Mean values do not represent degree of variation in data, so coefficient of variance was calculated for all the dinucleotide gap sizes and only human genome CpG gap sizes exhibited significantly higher scattering of data about the mean value in comparison with other dinucleotides. This clearly indicates loss of CpG in a non-random pattern in methylated genomes. Further dispersion index (DI) was calculated for the data to determine degree of dispersion. The DI of CpG gap size in human sequence was more than eight fold higher when compared to any of the other five dinucleotides while only moderately (1.5 fold) higher in case of poorly methylated genome of *D. melanogaster* and the difference was insignificant in unmethylated genome of *C. elegans*.

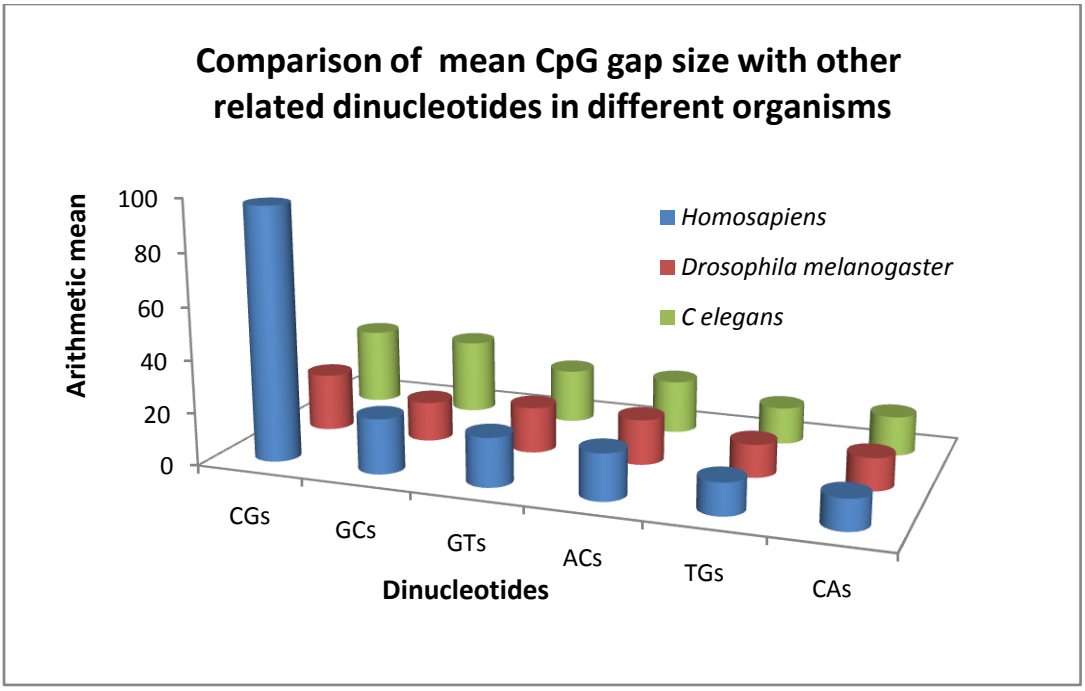

Fig. 11 Mean of CpG gap size with other related dinucleotides in different organisms

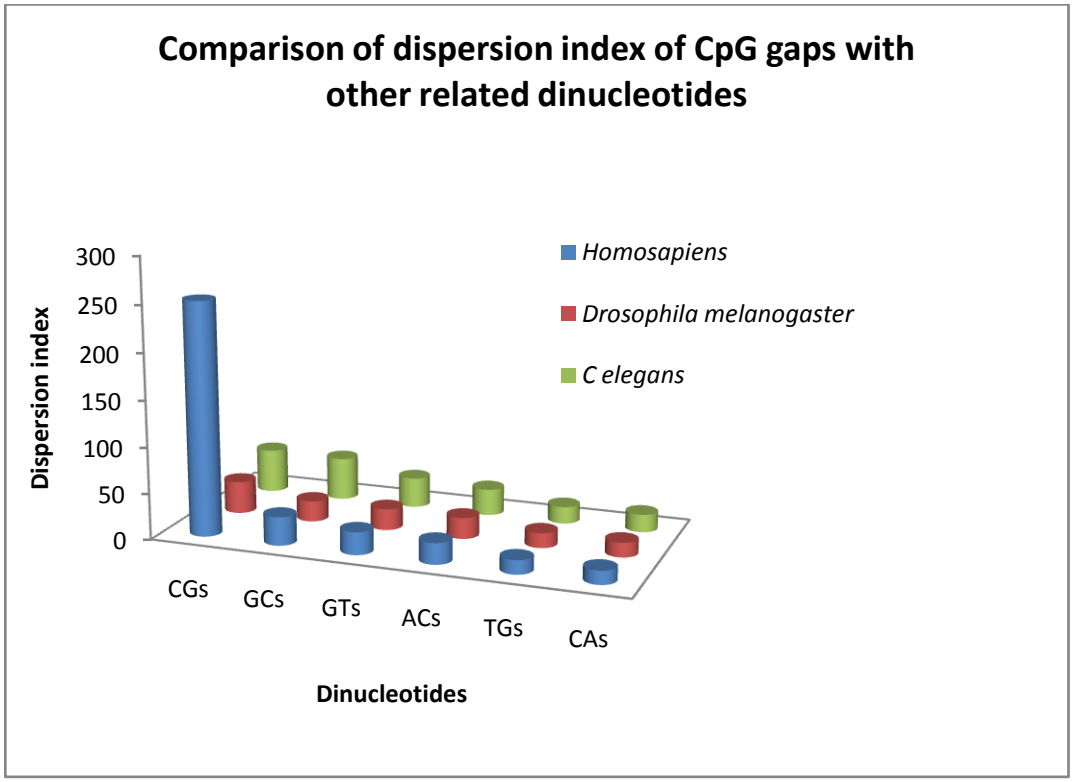

Fig. 12 Dispersion index of CpG gap size with other related dinucleotides in different organisms

So, these graphs clearly show that CpGs are more dispersed as compared to other related nucleotides in Homosapiens rather than any other species. The value of the arithmetic mean and dispersion index is maximum in Homosapiens.

Distribution pattern of CpGs in differently methylated genomes

Comparison of CpG gap sizes with that of other five related dinucleotides clearly show that the data reflects CpG density and distribution in the genome. Comparison of three differently methylated genomes showed the impact of methylation on CpG abundance and distribution in the genomes. To ascertain the later observation more lucidly, CpG gap size data was analyzed for more organisms exhibiting different levels of methylation in their genomes. Eleven species' genomic sequences were analyzed and CpG gap size data was generated. The species included four mammals, two non-mammal vertebrates (fishes), one each of insect, worm, protozoan, fungus and bacterium. DNA sequences of length 3.3×10^6 bp taken from GenBank database for the eleven species. The gaps between each CpG were determined using algorithm mentioned in the methods. The average distance (Arithmetic mean), coefficient of variance, and dispersion index (DI) was calculated for data, obtained from CpG gap size values. The mean CpG gap size for mammals was found to be highest followed by the two fishes, D. melanogaster, and *C. elegans* showed lowest values. The fungal and protozoan exhibited moderately higher values, which cannot be explained on the basis of methylation of genome. Another contributing factor of GC% of sequence could be responsible to some extent. Similar pattern is found in nCGI regions while CGIs have comparable gap size values which is acceptable owing to the fact that CGIs have been defined on the basis of predecided values of GC% and $CpG_{Obs/Exp}$ values.

In order to compare the distribution of CpGs in these genomes their DI values of CpG gap sizes were determined. It was observed that in a fashion similar to the mean gap size values, the DI also was highest for primate genomes followed by murine genomes, which was followed by the two fish genomes. The poorly methylated and unmethylated genomes of insect, worm, protozoan, fungus and bacterium had very low DI values. A similar pattern was observed for nCGI regions while CGIs had comparatively very low values. These observations indicated that the gap size values of CpGs is comparable in CGIs (owing to the identically defined parameters of CGI screening in all the species) while in nCGIs the DI values are largely dependent to the methylation status of the genomes. Higher DI values of nCGI regions of methylated genomes explicitly indicate that CpG distribution is clustered even outside the CGIs. The DI values of

nCGI of methylated genomes are also higher than that of total genomes of poorly or unmethylated genomes. It show that CpGs in methylated genomes are highly over dispersed and clustered in comparison with poorly or unmethylated genomes.

Table: 6 Represents the arithmetic mean of gaps in CpG dinucleotides

Species	Arithmetic mean		
	CGI	nCGI	Total sequence
Homosapiens	12	114	96
Pan troglodytes	12	154	149
Mus musculus	13	93	86.2
Rattus norvegicus	13	82	78.3
Danio rerio	12	57	54.3
Takifugu rubripes	15	46	42.3
Drosophila melanogaster	13	24	21.86
Anopheles gambiae	9.6	21	18.58
Caenorhabditis elegans	12	29	28.84
Saccharomyces cerevisiae	18	34	33.7
Escherichia coli	_	_	11.72

Table: 7 Represents the coefficient of variance and index of dispersion of CpG/ non CpG islands and total sequence

Species	Coefficient of variance			Index of dispersion or variance by mean ratio		
	CGI	nCGI	Total sequence	CGI	nCGI	Total sequence
Homosapiens	157	145	161.5	29	241	251.8
Pan troglodytes	157	125	128	30	239	243
Mus musculus	161	134	140	34	166	170
Rattus norvegicus	128	134	137	21	146	147
Danio rerio	134	149	153	22	126	126
Takifugu rubripes	135	137	142	27	86	85.6
Drosophila melanogaster	115	122	126.3	18	36	34.89
Anopheles gambiae	127	127	136.8	16	34	34.78
Caenorhabditis elegans	105	127	127.9	14	47	47.18
Saccharomyces cerevisiae	102	106	107	19	39	38.8
Escherichia coli	_	_	115	_	_	15.5

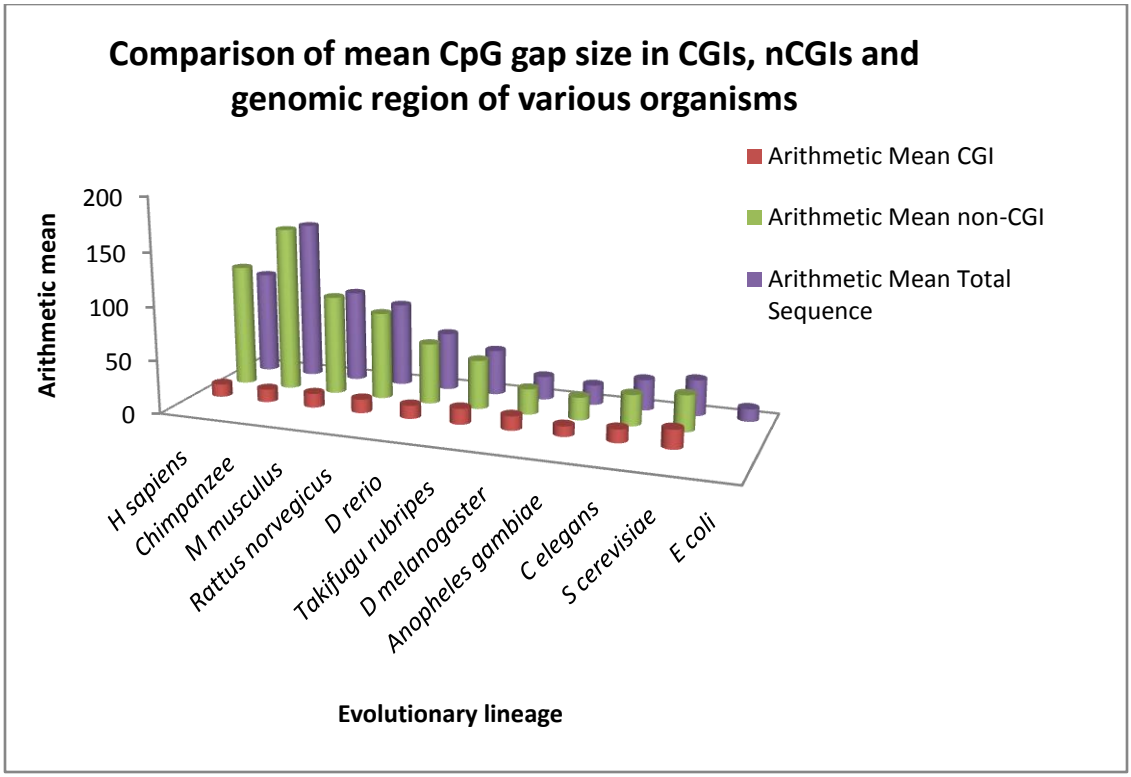

Fig. 13 Mean of CpG gap size in CGIs, nCGIs and genomic region of various organisms

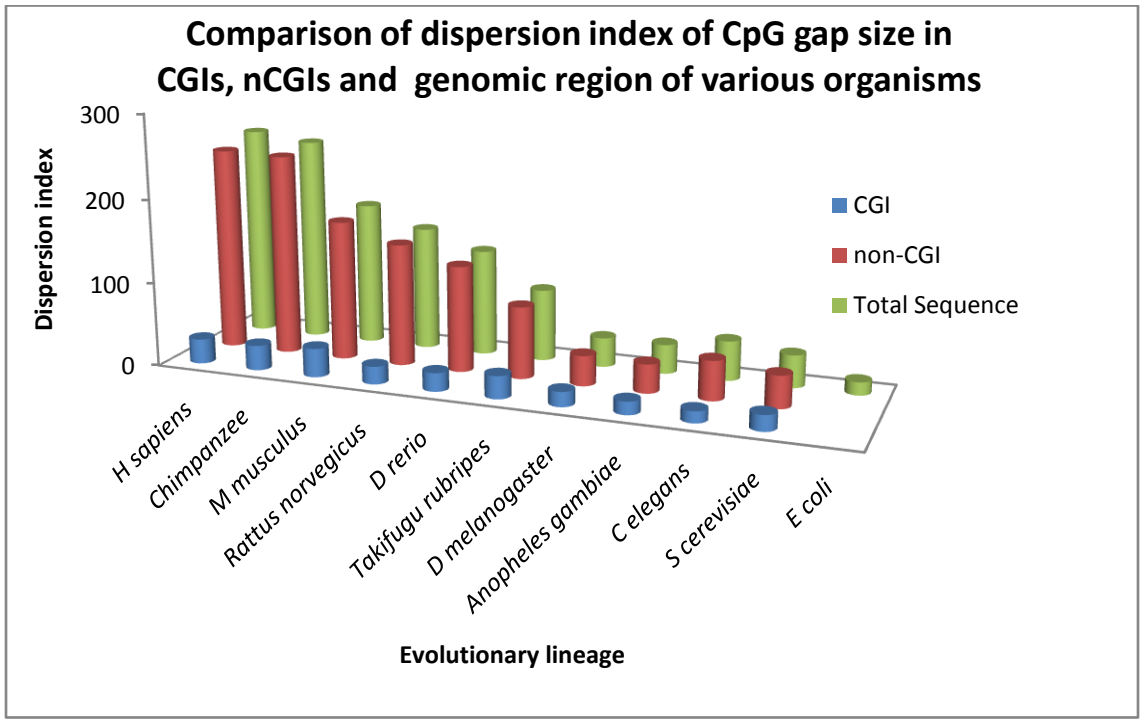

Fig. 14 Dispersion index of CpG gap size in CGIs, nCGIs and genomic region of various organism

Distribution of CpGs in genes

After comparing CpG distribution in CGI and nCGI regions of differently methylated genomes, it was attempted to understand the distribution of CpGs in coding and other parts of genomes. For this study 100 gene (few cluster of genes) each of two primate genomes were selected at random and their sequences were classified into exons, introns and intergenic regions (including UTRs, promoters and regulatory regions). The sequences of each of the three classes were used to determine CpG gap sizes as mentioned in methods. The mean gap size values were found to be lowest for exons followed by introns and highest for intergenic regions for both the genomes. In implies that CpGs have highest abundance in coding regions and least in intergenic regions. This shows the evolutionary pressure against loss of CpGs and proves that the loss in non random in nature resulting in the non random distribution

To have better understanding of distribution, DI values of gap sizes were determined and it was found that DI value was least for exons and highest for intergenic regions. The apparent correlation between mean and DI values of gap sizes in both the methylated primate genomes shows that methylation induced loss of CpGs only is responsible for non random (over dispersed and clustered) distributions of CpGs in the genomes

Table 8: Represents arithmetic and standard deviation of distribution of CpG in exon/intron/intergenic region

Species	Arithmetic mean			Standard deviation		
Homosapiens	exons	introns	Intergenic regions	exons	introns	Intergenic regions
	25.26	40.69	42.22	43.79	62.09	71.88
Pan troglodytes	30.68	44.39	52.48	55.61	68.86	77.05

Table: 9 Represents the coefficient of variance and index of dispersion of distribution of CpGs in exon/intron/intergenic region

Organism	Coefficient of variance			Index of dispersion		
	exons	introns	Intergenic region	exons	introns	Intergenic region
Homosapiens	173.29	152.57	170.23	75.88	94.74	122.37
Pan troglodytes	181.26	155.11	146.82	100	106.82	113.13

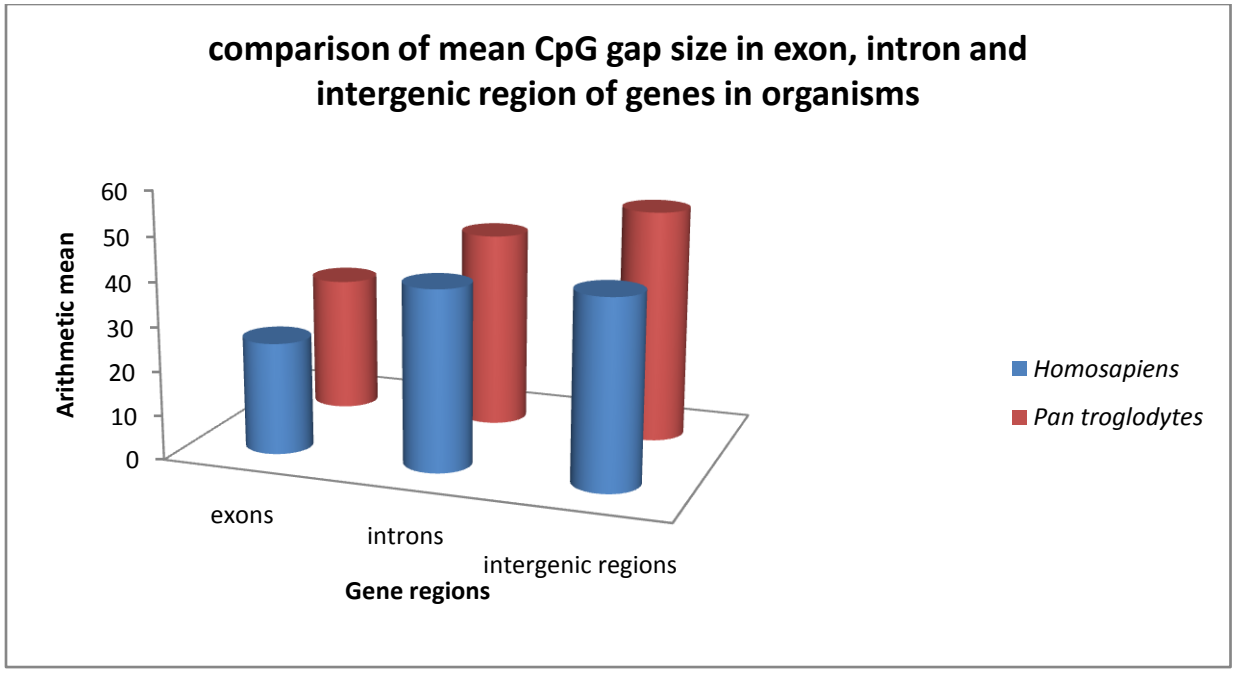

Fig. 15 Mean of CpG gap size in exon, intron and intergenic region of genes in organisms

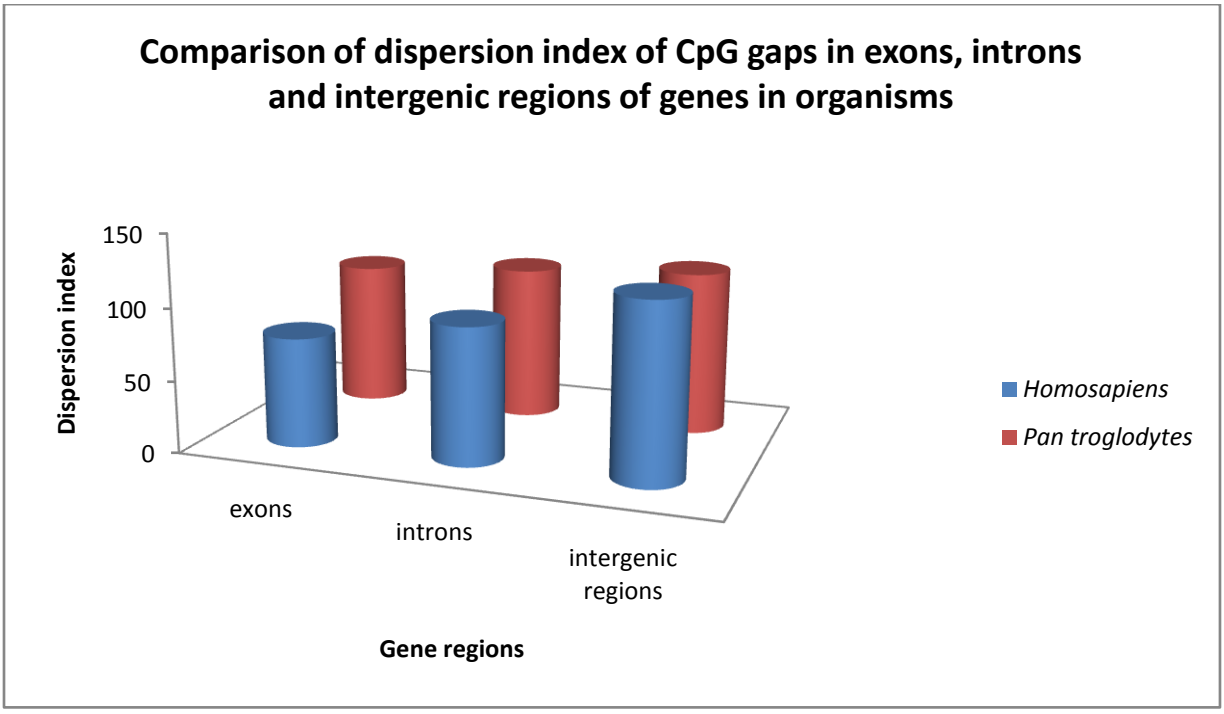

Fig. 16 Dispersion index of CpG gap size in exon, intron and intergenic region of genes in organisms

Discussion

The interpretation of above data shows that CpG gap size can be used as a tool to study CpG abundance (density) as well as non random distribution of these dinucleotides in the (methylated) genomes. This study reiterates the earlier findings of CpG underrepresentation of methylated genomes and the effect of evolutionary pressure on the loss of CpGs. It may be inferred that CpG distribution in methylated genomes is designed by at least two factors, evolutionary pressure and sequence dependent or independent factors resulting in differential methylation of CpGs. CpG 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. However CpGs in non-CpG islands regions also play very important role. Majority of them remain methylated in contrast to usually unmethylated CpGs of CpG islands but some of them are differentially methylated depending on spatial, termporal or other factors.

The present study has attempted to understand the distribution of CpGs of nCGI regions in comparison with CGI regions or overall genome. Dispersion index of CpG gap sizes was determined for this purpose. Dispersion index is ratio of variance and mean. DI < 1 represent ordered distribution, DI = 0 is random distribution whereas DI > 1 represents over disperesion and clustering of data in space. The DI values of CpG gap sizes thereby represent their distribution in one dimensional space. Relatively very high DI values of methylated genomes as well as their nCGI regions mean that the clustering of CpG distribution is not merely because of CpG islands and nCGI regions but also within nCGI regions. So in addition of the factors such as high GC% and $CpG_{Obs/Exp}$ which play significant role in differential methylation in genome, distribution of CpG gap sizes might also be contributing to this phenomenon and be responsible for differential methylation of CpGs in nCGI regions.

Comparing DI values in exons, introns and intergenic regions also support the above hypothesis. Exons and introns with lower DI values are usually not methylated and intergenic regions which largely contain repeat sequences (usually remain methylated) have highest DI values. Based on these preliminary results it may be inferred that the the study may be expanded to improve the understanding of CpG distributions in methyalted genomes and as a result interesting biological phenomena based on differential methylation.

CHAPTER 7

CONCLUSION

Conclusion

This study of CpG distributions in genomes of widely varying evolutionary lineages was aimed to understand if there exists non-randomness of CpG distribution as finer level than classifying genomes in merely into CpG islands and non CpG islands categories. For this purpose initially it was confirmed that gap size between adjacent CpG dinucleotides can be used as a tool to study their distribution by comparing the their mean, standard deviation and dispersion index with five other related dinucleotides. It was found that CpG gap size based statistics are distinct in comparison to rest all five dinucleotides. Then CpG gap size statistical analysis was compared amongst eleven different species. It was concluded that CpGs are more distance apart, as expected due to their loss in course of evolution, and over dispersed in methylated genomes when compared to poorly or unmethylated genomes. This comparison was found to be occurring in a similar fashion in nCGI regions of these genomes also whereas CGIs of all the organisms exhibited similar mean and DI values. Then mean and DI values of CpG gap sizes was compared for exons, introns and intergenic regions of human and chimpanzee genes. It was observed that both mean and DI values are highest for intergenic regions and least for exon regions. It may be concluded that CpG occur in a rare and over dispersed fashion in intergenic regions while with higher density and relatively even distribution in coding regions.

CHAPTER 8

REFERENCES

References

1. Adrian Bird. (2012). "DNA methylation patterns and epigenetic memory." Gens & Development**16:** 6-21
2. Adrian P.Bird. (1980). "DNA methylation and the frequency of CpG in animal DNA." Nucleic Acids Research **8:** 1499-1504.
3. Aharon Razin and Howard Cedar. (1991). "DNA Methylation and Gene Expression." Microbiological reviews. **55**: 451-458.
4. Aissani, B. and Bernardi, G. (1991). "CpG islands, genes and isochores in the genomes of vertebrates." Gene **106**:185-195.
5. Annalisa Varriale, Giorgio Bernardi. (2010). "Distribution of DNA methylation, CpGs, and CpG islands in human isochors." Genomics. **95**: 25-28.
6. Bird, A. P., and Taggart, M. H.m. (1980). "Variable patterns of total DNA and rDNA methylation in animals." Nucleic Acids Res. **8:**1485-9.
7. Bird, A. P., Taggart, M. H., Frommer, M., Miller, J. M. and Macleod, M.A. (1985). "Fraction of the Mouse Genome That Is Derived from Islands of Nonmethylated, CpG-Rich DNA." (1985) Cell. **40:** 91-99.
8. Daniela Carotti, Salvatore Funiciello, Franco Palitti and Roberto Strom. (1997). "Influence of Pre-existing Methylation on the de Novo Activity of Eukaryotic DNA Methyltransferase." Biochemistry **37:** 1101-1108.
9. David A. Low, Nathan J. Weyand, Michal J. Mahan. (2001) "Roles of DNA Adenine Methylation in Regulating Bacterial Gene Expression and Virulence." Molecular,cellular and developmental biology **69**.
10. ER Gibney and CM Nolan. (2010). "Epigenetics and gene expression." Heredity **105:** 4-13.
11. Fazzari, M. J., and J. M. Greally. (2004). "Epigenomic: beyond CpG islands." Nature Reviews Genetics **5:** 446-455.

12. FrancIsco Antequera and Adrian Bird. (1993). “Number of CpG islands and genes in human and mouse.” Proc. Natl. Acad. Sci. USA **90:** 11995-11999.
13. Hermann H. Gowhera and A. Jeltsch. (2004). “Biochemistry and biology of mammalian DNA methyltransferases.” Cellular and Molecular Life Sciences **61:** 2571-2587.
14. Jean-Pierre Issa. (2004). “CpG island methylator phenotype in cancer.” Nature/review/cancer **4**: 988-93.
15. JODY C. CHUANG AND PETER A. JONES. (2007) “Epigenetics and MicroRNAs.” Pediatric research **5**: 24-29.
16. Jones, P. A., and S. B. Baylin. (2002). “The fundamental role of epigenetic events in cancer.” Nature Reviews, Genetics **3:** 415-428.
17. Kelly M. McGarvey, Leander Van Neste and Leslie Cope. (2008). “Defining a chromatin pattern that characterizes DNA-hypermethylated genes in colon cancer cells.” Cancer research **68:** 5753-5759.
18. Leng Han, Bing Su, Wen-Hsiung and Zhongming Zhao. (2008). “CpG island density and its correlations with genomic features in mammalian genomes.” Genome Biology. **9:** R79.
19. M. Gardiner-Garden and M. Frommer. (1987). “CpG islands in vertebrate genomes.” J. Mol. Biol **196:** 261-282.
20. Matsuo K., Clay, O., Takahashi, T., Silke, J. and Scha_Ner, W. (1993). “Evidence for Erosion of Mouse CpG Islands during Mammalian Evolution.” Som. Cell and Mol. Gen. **19**: 543-555.
21. Michael Hackenberg, Guillermo Barturen1, Pedro Carpena, Pedro L Luque-Escamilla, Christopher Previti and José L Oliver. (2010). “Prediction of CpG island function: CpG clustering vs. sliding window method.” BMC Genomics **11:**1471-2164.

22. Pradhan, S., Bacolla, A., Wells, R.D., and Roberts, R.J. (1999). "Recombinant human DNA (cytosine-5) methyltransferase. I. Expression, purification, and comparison of de novo and maintenance methylation." J. Biol. Chem. **274**: 33002–33010.
23. Rakesh Singal and Gordon D. Ginder. (1999) "DNA methylation." American Society of Hematology review **93:** 4059-4070.
24. Robert S. Illingworth, Ulrike Gruenewald-Schneider, Shaun Webb1, Alastair R. W. Kerr, Keith D. James, Daniel J. Turner, Colin Smith, David J. Harrison, Robert Andrews, Adrian P. Bird. (2010). "Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome." PLOS genetics **6.**
25. Rudolf J., A. Bird. (2003) "Epigenetic regulation of gene expression." Nature genetics supplement **33**: 246-251.
26. Serge Saxonov, Paul Bergt and Douglas L. Brutlag. (2006). "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters." BioMedical Informatics Program **103:** 1412-1417.
27. Shuo Han and Anne Brunet. (2012) "Histone methylation makes its mark on longevity." Trends in cell biology. **22:** 42-49.
28. Sved J, Bird A. "The expected equilibrium of the CpG dinucleotides in vertebrate genomes under a mutation model." Proc Natl Acad Sci USA. **87**: 4692-6.
29. Takai, D., and P. A. Jones. (2002). "Comprehensive analysis of CpG islands in human chromosomes 21 and 22." Proc. Natl. Acad. Sci. USA **99**: 3740-3745.
30. Tom Shimizu, Kouichi Takahashi and Masaru Tomita. (1997). "CpG Dinucleotide Distribution and DNA Methylation." Laboratory for Bioinformatics.
31. Yoder, J. A., C. P. Walsh, and T. H. Bestor. (1997). "Cytosine methylation and the ecology of intragenomic parasites." Trends in Genetics **13**: 335-340.
32. Zhongming Zhao and Leng Hana. (2010). "CpG islands: algorithms and applications in methylation studies." Biochem Biophys Res Commun **382:** 643-645.