Study of CpG distribution in coding and non - coding regions of different 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:

Tajeshwar Preet Kaur
(301201021)

Department of Biotechnology and Environmental Sciences

THAPAR UNIVERSITY
PATIALA

July 2014

CERTIFICATE

This is to certify that the report entitled "**Study of CpG distribution in coding and non-coding regions of different genomes**" submitted by Tajeshwar Preet Kaur (301201021) in partial fulfilment of the requirement for the award of Degree of Masters in Science in Biotechnology to Thapar University, Patiala is a record of student's own work carried out by her under my supervision and guidance. The report has not been submitted for the award of any other degree or certificate in this or any other university.

Dr. Vikas Handa
Assistant Professor
Department of Biotechnology

Dr. Dinesh Goyal
Head of Department
Department of Biotechnology

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

CANDIDATE'S DECLARATION

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

Tajeshwar Preet Kaur

Date: 21 July, 2014

Tajeshwar Preet Kaur
Roll No. 301201021

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

Date: 21 July, 2014

Dr. Vikas Handa
Assistant Professor
Department of Biotechnology

ACKNOWLEDGEMENT

This project bears on imprint of many people. I would like to gratefully acknowledge the enthusiastic supervision of my guide ***Dr. Vikas Handa****, Assistant Professor, Department of* ***Biotechnology*** *and Environmental Sciences, Thapar University, Patiala for his guidance as well as providing necessary information regarding the thesis work. His willingness to motivate contributed tremendously to my project. I perceive as this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way.*

I am sincerely thankful to ***Dr. Dinesh Goyal****, 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 thanks to my friends Japnjyot Saini, Shivangi and Bhupinder Singh for their kind co-operation and encouragement which helped me materialize this project.

At last, I would like to make an honourable mention of my parents and THE ALMIGHTY for his constant blessings.

Date: July 18, 2014 **Tajeshwar Preet Kaur**
Place: Patiala **(301201021)**

TABLE OF CONTENTS

ABBREVATIONS

A	Adenine
AM	Arithmetic mean
C	Cytosine
C_5	Carbon at 5_{th} position
CDS	Coding Sequence
CpG	Cytosine and Guanine
CGI	CpG island
CpG_{Exp}	Expected frequency of CpG dinucleotides
CpG_{Obs}	Observed frequency of CpG dinucleotides
$CpG_{Obs/Exp}$	Ratio of observed frequency of CpG over the expected frequency of CpG dinucleotide
CV	Coefficient of variance
DI	Dispersion index
Dnmt	DNA methyltransferase
Dnmt 1	DNA methyltransferase 1
Dnmt3a	DNA methyltransferase 3a
Dnmt3b	DNA methyltransferase 3b
G	Guanine

MBDs	Methyl- CpG Binding Proteins
N^5	Nitrogen at 6^{th} position
N^4	Nitrogen at 4^{th} position
SD	Standard deviation
TSG	Tumor Specific Gene
TSS	Transcription Start Site

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 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. Histone modifications comprise of different modifications of histone proteins such as acetylation, methylation, phosphorylation and ubiquitination (Peter A. Jones *et al.*, 2007).

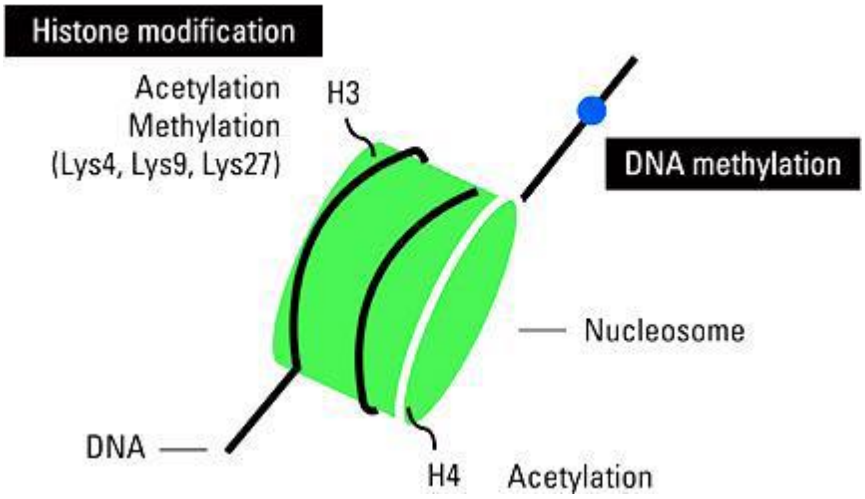

Figure 1:- Histone modification and DNA methylation (Lee *et al.*, 2009)

DNA methylation is a biochemical process that is important for normal development in higher organism. DNA methylation has a variety of important functionsin mammals, such as gene repression, control of cellular differentiation and development, preservation of chromosomal integrity, parental imprinting and X-chromosome inactivation. In addition DNA methylation silences endogenous retroviruses and suppresses homologous recombination. DNA methylation occurs at N-6 position of adenine and N-4 and C-5 position of cytosine in case of prokaryotes. It involves the addition of a methyl group to the 5 position of the Cytosine pyrimidine ring in eukaryotes.

CpGdinucleotides are CG dinuleotides of DNA where a cytosine followed by a guanine is methylated at position 5 in vertebrate genomes. The "p" in CpG refers to the phospho-diester bond between the cytosine and the guanine Cytosines in some of the CpGdinucleotides can be enzymatically methylated.Enzymes that add a methyl group to cytosine are called DNA methyltransferases. Dmnt1, Dmnt2 and Dmnt3a, Dnmt3b are the methyltransferasesfound in vertebrates.All DNA methyltransferases use *S*-adenosyl-L-methionine (AdoMet) as the source of the methyl group.

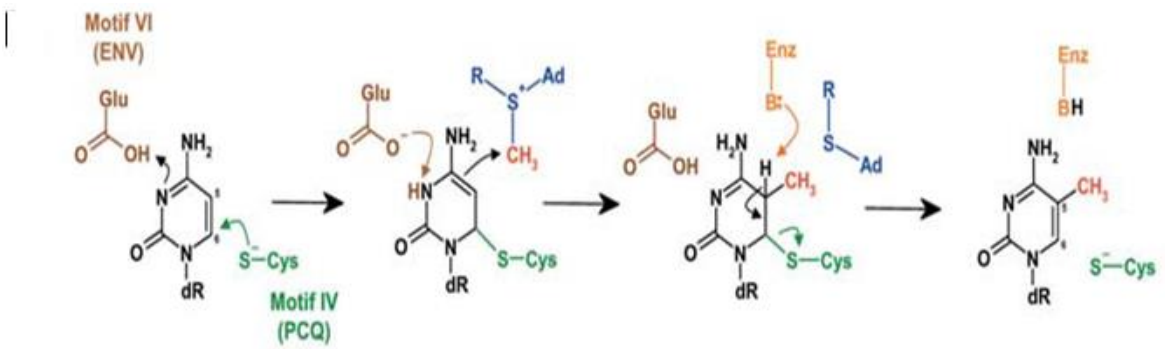

Figure 2:- Chemistry of the methylation reaction of DNA-(cytosine-C5)-MTases.

The methylation reaction catalysed by DNA MTases involves the formation of a dihydrocytosine intermediate that is covalently bound to the enzyme and releases S- adenosyl -L-homocysteine as product. In the second step of the reaction, the covalent bond is broken and the methylated cytosine released (Biochemistry and biology of mammalian DNAMethyltransferases, A. Hermann , H. Gowhera and A. Jeltsch, Cellular and Molecular Life Sciences 61(2004) : 2571-2587).

DNMT1predominantly methylateshemimethylatedCpG di-nucleotides in the mammalian genome. DNA methylation in vertebrates typically occurs at CpG sites, that is where a Cytosine is immediately followed by a Guanine at 3`end in the DNA sequences.Methylated CpGs are mutable (CpG/CpG→(methylation)→TpG/CpA). The deamination of methylated cytosines results in an $^{m}C \rightarrow T$ transition mutation. In contrast, deamination of unmethylated cytosine leads to uracil in the DNA, which can be recognized and repaired efficiently by the uracil-deglycosylase pathway. As a result CG sites are major mutational hotspots. Hence due to gradual loss of CpGs during evolution,CpG sitesare globally under-representated in genomes of vertebrate species.

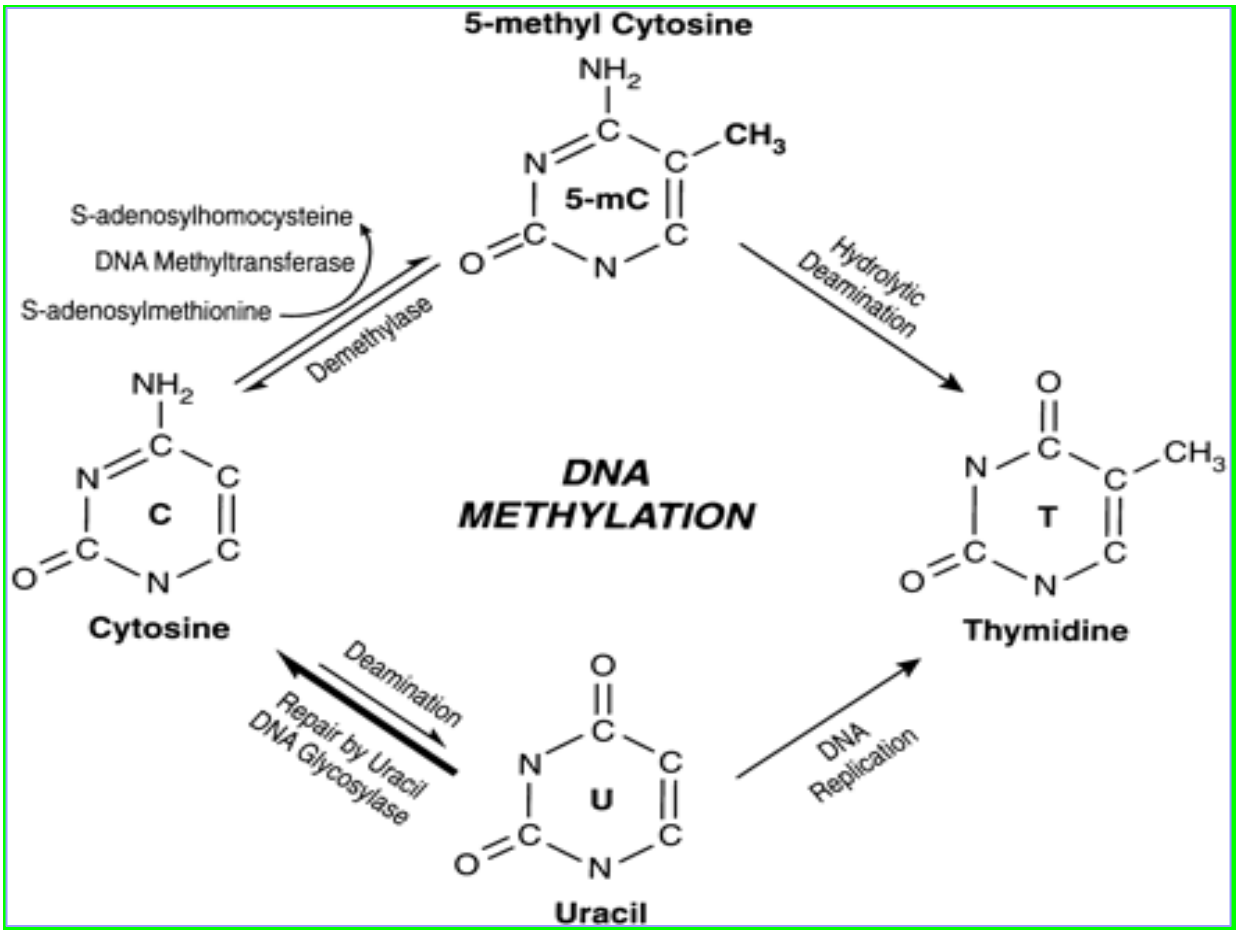

Figure 3:- Schematic representation of the biochemical pathways for cytosine methylation,demethylation, mutagenesis of cytosine and 5-metC (Singhal and Ginder 2013)

Approximately 60–90% of all CpGdinucleotides in the genome are methylated, while unmethylatedCpGdinucleotides are mainly clustered in the CpG rich sequences, terrmedas CpGisland, of the gene promoter region. The distribution may affect the DNA methylation levels in all the methylated genomes. CpG islands have higher representation of CpGs when compared to rest of the genome in vertebrates.CpG distribution is uneven in genomes of all methylated eukaryotic genomes which is also evident in CpG islands.

CODING REGION

The coding region of a gene, also known as the **coding sequence** or **CDS** (from **Coding DNA Sequence**), is that portion of a gene's DNA or RNA, composed of exons, that codes for protein. The region begins at the 5' end by a start codon and ends at the 3' end with a stop codon in prokaryotic genes while consists of the sequence of spliced exons in eukaryotes. The coding region in mRNA is bounded by the 5’ untranslated region and the 3’ untranslated region, which are also parts of the exons.The coding region of an organism is the sum total of the organism's genome that is composed of protein coding regions.

Coding sequence annotation

While identification of open reading frames within a DNA sequence is straightforward, identifying coding sequences is not, because the cell translates only a subset of all open reading frames to proteins. Currently CDS prediction uses sampling and sequencing of mRNA from cells, although there is still the problem of determining which parts of a given mRNA are actually translated to protein. CDS prediction is a subset of geneprediction, The latter also including prediction of DNA sequences that code not only for protein but also for other functional elements such as RNA genes and regulatory sequences.

NON CODING REGION

Noncoding DNA sequences are components of an organism's DNA that do not encode proteinsequences. Some noncoding DNA is transcribed into functional non-coding RNA molecules (e.g. transfer RNA, ribosomal RNA, regulatory RNAs and introns), while others are not transcribed or give rise to RNA transcripts of unknown function. The amount of non-coding DNA varies greatly among species.Non-coding regions are interspersed throughout DNA.

Non- coding DNA sequences include:

Noncoding functional RNA,*Cis*- and *Trans*-regulatory elements, Introns, Pseudogenes, Repeat sequences, transposons and viral elements, Telomeres.

Promoters are the key regulatory elements known to regulate the initiation oftranscription in both eukaryotes and prokaryotes. In a DNA sequence, promoter is the region that appears immediately

upstream of a Transcription Start Site (TSS) thatperforms a crucial role in initiation of transcription and is responsible for binding of RNA polymerase. The promoter region is divided into three parts: (1) The Core promoter, (2) The Proximal promoter and (3) The Distant promoter. A few core promoter elements have been detected, of which the most common elements are CpG Island, TATA box, Initiator (Inr), downstream promoter element (DPE) and TFIIB recognition element (BRE).60% of the human core promoter falls near a CpG island a short stretch of DNA having a high G+C content and a high frequency of GC dinucleotides compared to the bulk DNA. Certaincombinatorial nature exists among CpGIsland and certainpromoter elements. Such as TATAbox are more common inpromoters lacking the CpG island nearby whereas BREs are morecommon in the promoters that are associated with CpGIsland.

5- Methyl cytosine (5-mC) is most common DNA modification found in eukaryotic genomes.

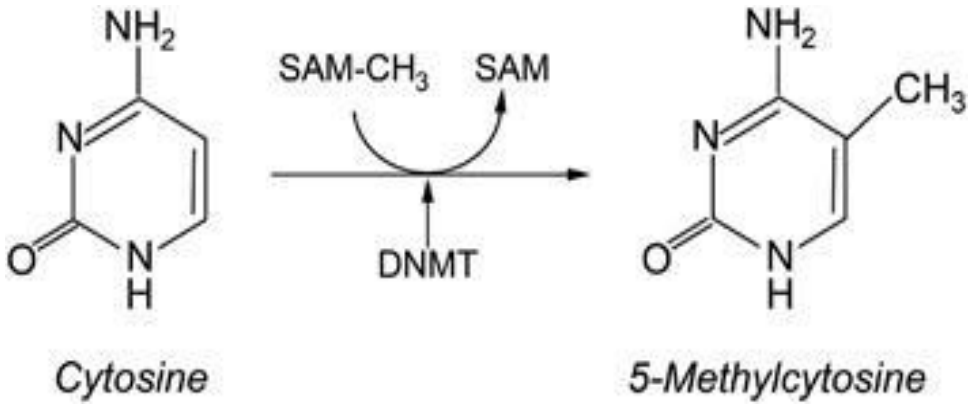

Figure 4:- Methylation at C5 position of cytosine by methytransferases (Gibney*et al.*, 2010)

In vertebrates genome,some 90%of 5-mC occurs within the dinucleotidesCpG(Grippo*et al.*1968), the majority of which is thought to be methylated(reviewed by Cooper, 1983). Methylation within the promoter region can turn the gene off. CpG islands in or near promoter acquire abnormal hypermethylation which results in silencing of genes which can be inherited by daughter cells. Trancripition of genes is affected in two ways: 1) DNA methylation prevents binding of transcriptional proteins to genes. 2) Methylated DNA bind to proteins known as Methyl – CpG Binding Proteins (MBDs)**.** MBDsrecruit additional protein such as histone deacetylases, chromatin remodeling proteins

that can modify histones which form compact inactive chromatin called heterochromatin. This leads to silencing of genes. DNA methylation also has an important role in cancer and tumor development. In normal cells tumor suppressor genes are associated with hypomethylatedCpG 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.

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. It has been observed that proper methylation is essential for cellular differentiation and embryonic development. There is significant difference in methylation level that exist between different tissue types.

The promoter region of a gene is usually located between the 5'end of the CpGisland and the gene's transcription initiation site (Antequera, 2003). Chromatin structure suggests the possibility that CpG islands might serve as replication origins of DNA (Antequera, 2003).The most important feature of CpG islands, besides their overlapping the promoter region of genes, is the fact that they remain unmethylated when about 80% of the CpGdinucleotides in the genome succumb to methylation. The methylated CpG is susceptible to deamination, and once deaminated, the mutation rate is 10-50 times higher than the mutation rate in other dinucleotides. Methylation is important: it can prevent transcription of the DNA of invading molecular parasites, and also prevent recombination between repetitive sequences ofDNA, thus keeping the genome stable (Antequera 2003).

CpG islands

The regions of DNA where CpGs are found in high abundance ascompared to other region of genome are called **CpGislands**. CpG islands have high GC% and Observed/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). CpGislanddepends 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. asDNA sequence regions> 200 bp in length, having GC% > 50% and $CpG_{Obs/Exp}$> 0.60 (Gardiner-Garden *et al.*,1987). A more accurate description of CpG islands was later given as > 500 bp long stretch of DNA having GC > 55% and $CpG_{Obs/Exp}$> 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 forprediction 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, 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 housekeeping 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 under-representated in genomic sequences of those organisms in whichDNA methylation takes place. Additionally CpGdistribution patterns near the 5` end of genes isdifferent among vertebrates, invertebrates, plants and bacteria (Shimizu *et al.*,1997). CpGdistribution is uneven in genomes of the species whose genomes undergo DNA methylation. Theunevenness of CpG distribution is largely attributed to the CpG islands. However the distributionof CpGs may also vary at a finer level both within CpG islands (CGI) as well as in the non-CpGisland regions (nCGI) i.e. rest of the genome. In present work the CG distribution studies has been carried out by analysis of gaps (distance) between adjacent CGs in the genomic sequence usingstatistical methods, similar to the approach used by Hackenberg et al., 2006. The gap size on onehand represents (inverse of) frequency of CpGs and on the other hand enables one to studydistribution of CpGs in 1-Dimensional space.

CHAPTER 2
REVIEW LITERATURE

Review of Literature

Epigenetic is defined as the inheritance of changes in gene function without changing the DNA sequence (Zhang, Rohde et al. 2009). Epigenetic signals comprise methylation of cytosine bases of the DNA and chemical modifications of the histone proteins. DNA methylation plays important roles in development and disease processes. Epigenetic modification includes covalent modification of histone tails (acetylation, phosphorylation, ubiquitination and methylation). All the modifications change the gene expression without altering the sequence of silenced gene.

Methylation is the process in which methyl group is added to the cytosine and methylated cytosine is called as 5-methycytosine. DNA methylation is carried out by enzymes called DNA methyltransferases on Cytosine and Adenine bases. All DNA methyltransferases use S-adenosyl-L-methonine(AdoMet) as the source of methyl group being transferred to DNA bases. Prokaryotic Cytosine and Adenine methylation can influence gene transcription, cell viability, play important role in mismatch repair of DNA and also serve the restriction-modification system that protects the bacterial host DNA from cleavage by specific endonucleases (Kahng and Shapiro 2001). Only Cytosines are methylated at position 5 in eukaryotes (mainly in vertebrate genomes). DNA methylation is carried out by DNA methyltransferases, Dnmt1, Dnmt3a, Dnmt3b and Dnmt2 found in mammals. DNA methyltransferases have two different modes of methylation processes: *de novo* methylation establishes the methylation state; maintenance methylation copies it onto daughter DNA strands after DNA replication. The first mammalian DNA methyltransferasediscovered was Dnmt1, which is highly conserved among eukaryotes and is responsible for maintaining methylation patterns in the DNA after replication. Later, Dnmt2, Dnmt3a and Dnmt3b were discovered. . DNA methyltransferases use S-adenosyl-L-methionine (AdoMet) as the source of the methyl group being transferred to the DNA bases. All the known mammalian DNA methyltransferases have a common structure of the catalytic domain which resembles the prokaryotic enzymes and is characterized by the 10 conserved amino acid motifs implicated in the catalytic function. In addition, the Dnmt1 and Dnmt3 enzymes contain a large N-terminal regulatory domain. Methylation occurs at Cytosine nucleotide (C5 position) of CpGdinucleotides of DNA in eukaryotes and this leads to deficiency of CpGs (Bird, 1980). The extent of methylation is high in higher organisms as compared to lower organisms and due to this CpGdinucleotides, frequency is low in higher organisms as comparative to lower organisms. In mammlian DNA, 3 to 5% cytosine are methylated

anddue to this sometimes 5-methylcytosine is considered as 5th base of mammalian DNA (Jeltsch*etal.*, 2004). CpG methylation leads to depletion in CpGs in DNA invertebrate genomes. Non- methylated and high CpGdinucleotides frequency; these areas are called as CpG islands. CpGdinucleotides areunderrepresented in genome sequences of organisms, which methylate their DNA (Gardiner-Garden and Frommer, 1987). 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, Obs/Exp 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. Approximately 60–90% of all CpG sequences in the genome are methylated, while unmethylatedCpGdinucleotides are mainly clustered in the CpG rich sequence, termed CpGisland, of the gene promoter region (Ng and Bird, 1999). The distribution affects the DNA methylation level in all species.

CpG island are those areas of genome which are non methylated and having GC% 55% and or moreand $CpG_{Obs/Exp}$value is at least 0.65 (Takai and Jones, 2001). CpG islands are found most often near the5` 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 providedthe 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 ofgenes 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 (CCGG)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 ($C^{m}CGG$). 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.CpG islands have been found at 5' end of almost all of housekeeping genes and large proportion of genes with a tissue restricted pattern of expression. (Craigand Bickmore,1994).Scarano*et al.* proposed that the CpG deficiency is due to an increased vulnerability of methylcytosines to spontaneously deaminate to

thymine in genomes with CpG cytosine methylation.Recent work has uncovered a large class of CGIs that are remote from annotated transcription start sites (TSSs), but nevertheless show evidence for promoter function (Illingworth et al. 2010; Maunakea et al. 2010). These findings emphasize the strong correlation between CGIs and transcription initiation.Anteguera and Bird, 1993 reported that CpGs are present at only 5% -10% of its predicted frequency in higher eukaryotes. Jabbari K, Bernardi G, May 2004 found that in mammals, 70% to 80% of CpGcytosines are methylated. Based on an extensive search on the complete sequences of human chromosomes 21 and 22, DNA regions greater than 500 base pair were found more likely to be the "true" CpG islands associated with the 5' regions of genes if they had GC content greater than 55%, and an observed-to-expected CpG ratio of 65%.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. 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 a 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. CpGdistribution is uneven or random in genomes of all the species. Approximately 60–90% of allCpG sequences in the genome are methylated, while unmethylatedCpGdinucleotides are mainly clustered in the CpG rich sequence, termed CpGisland, of the gene promoter region (Ng andBird, 1999).

CHAPTER 3
SCOPE OF STUDY

Scope of study

DNA methylation plays 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 fine structure of genomes 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 probablyhas lead tocertain unmethylated regions, relatively richer in CpGswhich 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 the reason certain CpGs are methylatedmore 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 asnon-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 of relationship between CpG distribution and their methylation.

CHAPTER 4
OBJECTIVES

Objectives

- Analysis of distribution of CGs in exons, introns and intergenic region for six genomes

- 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 six 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). Six different species were taken to study distribution of CpGs in coding and non- coding regions of different genomes.A region spanning 100 genes was selected at random in each genome to analyze distribution of CpGs in exons, introns and intergenic regions. The details of the sequences are as following:

Data: DNA sequences:

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

Species	Accession No.	Start position (bp)	End position (bp)	Total length of sequence (bp)
Homosapiens	NT_004487.19	10000000	13300000	3300000
Pan troglodytes	NC_006471.3	20000000	50000000	30000000
Musmusculus	NC_000082.6	30000000	50000000	20000000
Rattusnorvegicus	NC_005100.3	20000000	50000000	30000000
Daniorerio	NC_007123.5	10000000	20000000	10000000
Drosophila melanogaster	NT_033779.4	3000000	6300000	3300000

Sequence analysis tools:

MS Word: Microsoft word was used to analyze the DNA sequences with the help of tools such as find, replace.

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

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 theCpG islands in the CpG islands from given genomic sequence. This software detects theCpG 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 were downloaded from Genbankdatabase for six different species and pasted in MS Word.100 genes from each organism were taken.Each gene of each organism was analysed in NCBI Map Viewer.Positions of exons and intron boundaries of genes were noted.MS Excel format was used to store the information.Sequences were downloaded from NCBI in fasta format for exon boundaries of all the exons of each gene of all organisms.Sequences were pasted in MS Word. For evaluating position of introns in a gene, the position of exons only were used. The sequences for intron boundaries were again downloaded from NCBI and pasted in MS Word.Intergenic region sequences were taken from NCBI and pasted in MS Word. The sequences of each gene`s exons were joined by X character to form a continuous sequence. This was donefor all the organisms.Similar procedure was done for introns boundaries and intergenic sequences for all the organisms.Now the sequences of all the organisms were analyzed for position of CGs and X using a program designed by Japnjot(Unpublished) to determinethe positions. The positions were noted down in MS Excel spreadsheet.The positions of X were eliminated to find positions of CGs in the sequences. Variance to average ratio was calculated for all the positions of all the organisms. The position of each CG in each was used to find out CpG gap Length.

Statistical analysis

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

$$\boldsymbol{Coefficient\ of\ variance} = \frac{\boldsymbol{Standard\ deviation}}{\boldsymbol{Arithmetic\ mean}} = \frac{\boldsymbol{\sigma}}{\boldsymbol{\mu}}$$

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 dispersion from one data series to another, even if the means are drastically different from each other.

Dispersion index (DI): It is also known as variance by mean ratio. Dispersion index is used to quantify whether the set of observance are evenly distributed, randomly distributed, or clustered.

Dispersion index is defined as ratio of variance by mean.

Dispersion index = variance / mean

DI = σ^2/ μ

Distribution of CpGin CpG islands and non CpG island areas of genomes

Sequences of all the organisms were taken from the region mentioned in materials. The sequences were analyzed by taking 600kb fragments for sequences of each organism. Each sequence fragmentwas searched for number and position of CpG island using CpG island Searcher (**cpgislands**.usc.edu). The number of CpGisland and their respective start and endpositions were determined. Using the start and end position for each CGI, the sequences were extracted from NCBI Nucleotide database. Each sequence was evaluated for no. of CG`s in each island. Having classified this sequence into CGI and nCGI sequences, both the classes of the sequences were used to determine CpG gap sizesseparately.

CHAPTER 6
RESULTS AND DISCUSSION

RESULTS

Comparison of CpG gap size

To perform the distribution study of dinucleotides, the DNA sequences were taken from GenBank database. The dinucleotide gap sizes i.e. number of nucleotides between adjacent dinucleotides, for CpG were determined by using method mentioned in the methods.
Gap sizes of CpG were determined for six differently methylated genomes, *Homo sapiens, Pan troglodytes, Musmusculus, Rattusnorvegicus, Daniorerio andDrosophila melanogaster.*

Statistical analysis was performed on the data obtained from gap lengths of CGs. Similarly arithmetic mean, variance to average ratio, coefficient of variance and index of dispersion wascalculated from the gap size data obtained from analyzed sequence for exon, intron and intergenic region separately.

Table. 2 Statistical analysis of distribution of CGs in exons, introns and intergenic regions

	Variance/Average Ratio of CGs		
ORGANISM	EXON	INTRON	INTERGENIC REGION
Homo sapiens	64.21	85.27	144.04
Pan troglodytes	93.63	101.35	109.3
Mus musculus	198.35	190.95	222.9
Drosophila melanogaster	21.3	31.13	33.72
Danio rerio	64.38	140.14	156.93
Ratus norvegicus	244.91	205.36	162.52

Figure 5:- Variance/Average Ratio of CGs in exons, introns and intergenic region of organisms

Meanvalues do not represent degree of variation in data, so coefficient of variance was calculated forall the dinucleotide gap sizes and human genome CpG gap sizes exhibited significantlyhigher scattering of data about the mean value in comparison with other organisms.

Table. 3 Represents Arithmetic Mean and Standard Deviation of distribution of CpG gaps in exons

Organism	Arithmetic Mean	Standard deviation
Homo sapiens	28.35	43.79
Pan troglodytes	34.87	55.61
Mus musculus	116.85	160.95
Rattus norvegicus	295.7	676.97
Danio rerio	50.34	76.61
Drosophila melanogaster	19.67	21.09

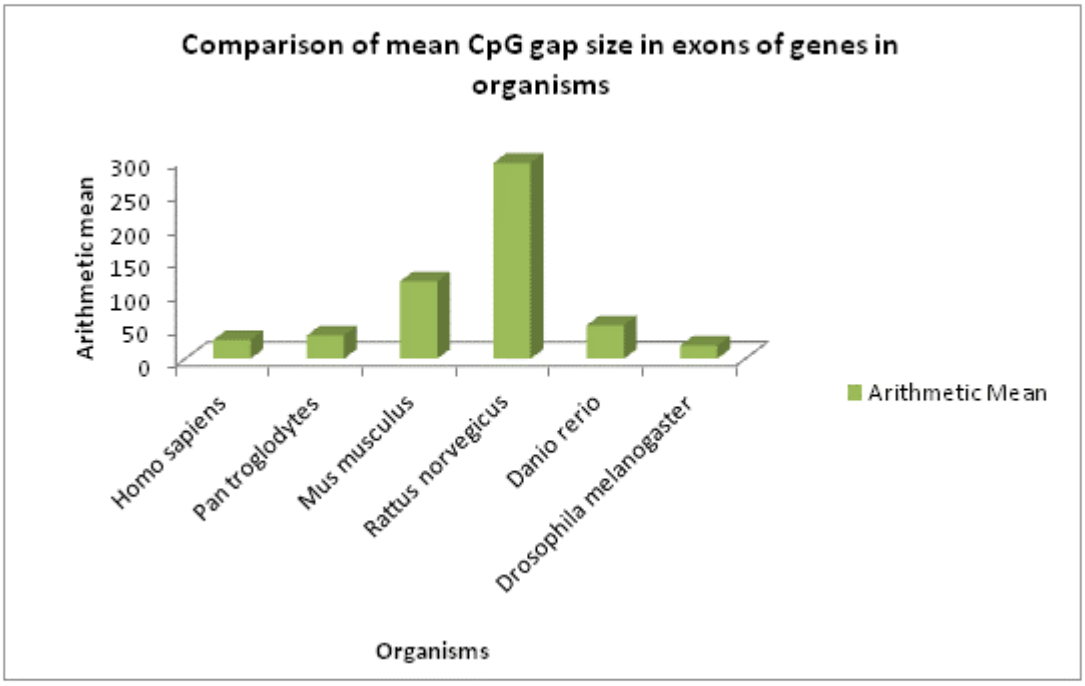

Figure 6: - Arithmetic Mean of CpG gap size in exons of genes in organisms

Table.4 Represents the Coefficient of Variance and Index of Dispersion of distribution of CpGs in exons

Organism	Coefficient of Variance	Dispersion Index
Homo sapiens	154.44	67.63
Pan troglodytes	159.45	88.67
Mus musculus	137.73	221.68
Rattus norvegicus	228.93	1549.81
Danio rerio	152.19	116.6
Drosophila melanogaster	107.22	22.61

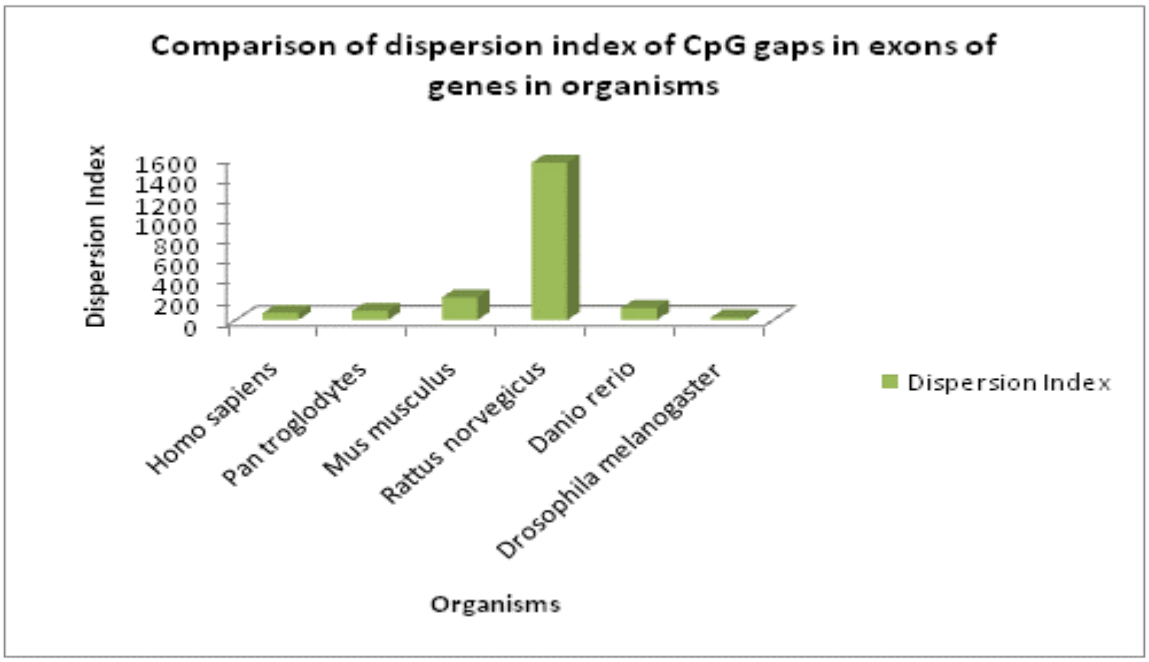

Figure 7:- Dispersion Index of CpG gap size in exons of genes of organisms

Table.5 Represents Arithmetic Mean and Standard Deviation of distribution of CpG in Introns

Organism	Arithmetic Mean	Standard deviation
Homo sapiens	40.87	62.09
Pan troglodytes	46.65	68.86
Mus musculus	99.79	108.7
Rattus norvegicus	278.2	23038.04
Danio rerio	52.97	88.18
Drosophila melanogaster	23.44	29.22

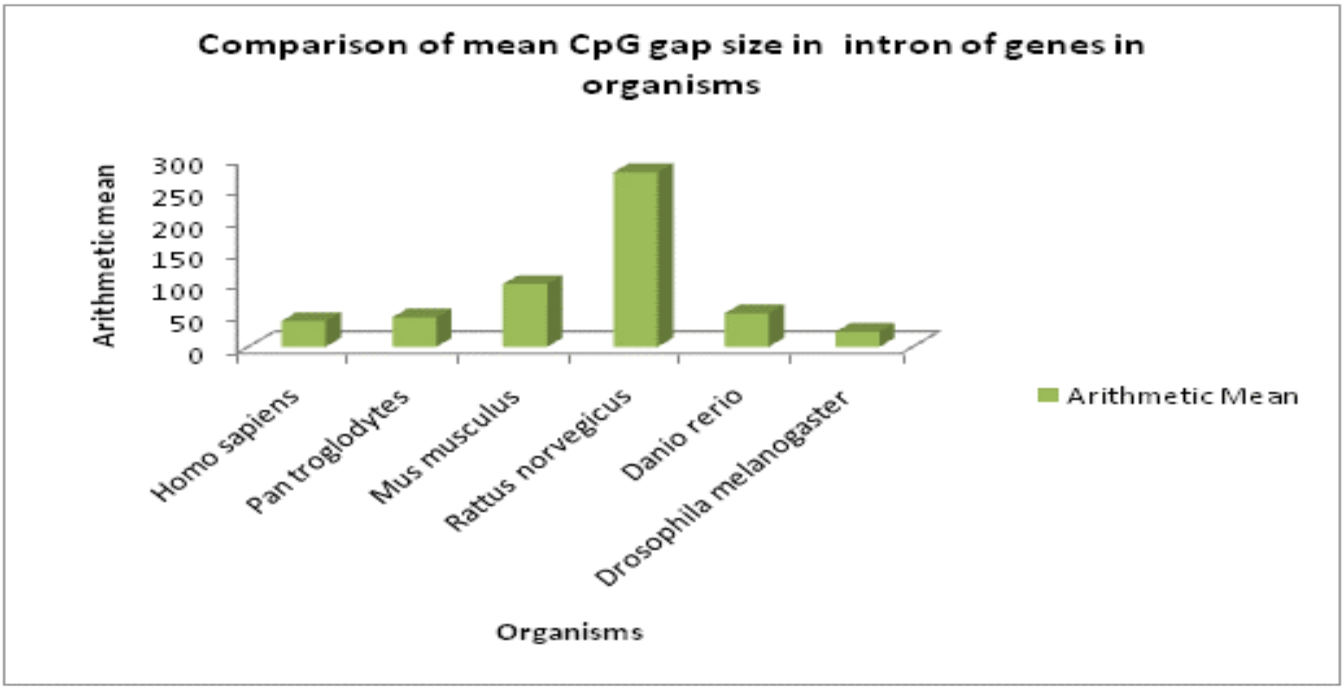

Figure8:- Arithmetic Mean of CpG gap size in introns of genes in organisms

Table. 6 Represents the Coefficient of Variance and Index of Dispersion of distribution of CpGs in introns

Organism	Coefficient of Variance	Dispersion Index
Homo sapiens	151.91	94.32
Pan troglodytes	147.6	101.63
Mus musculus	108.92	118.4
Rattus norvegicus	128.34	198.8
Danio rerio	166.45	146.78
Drosophila melanogaster	124.64	36.42

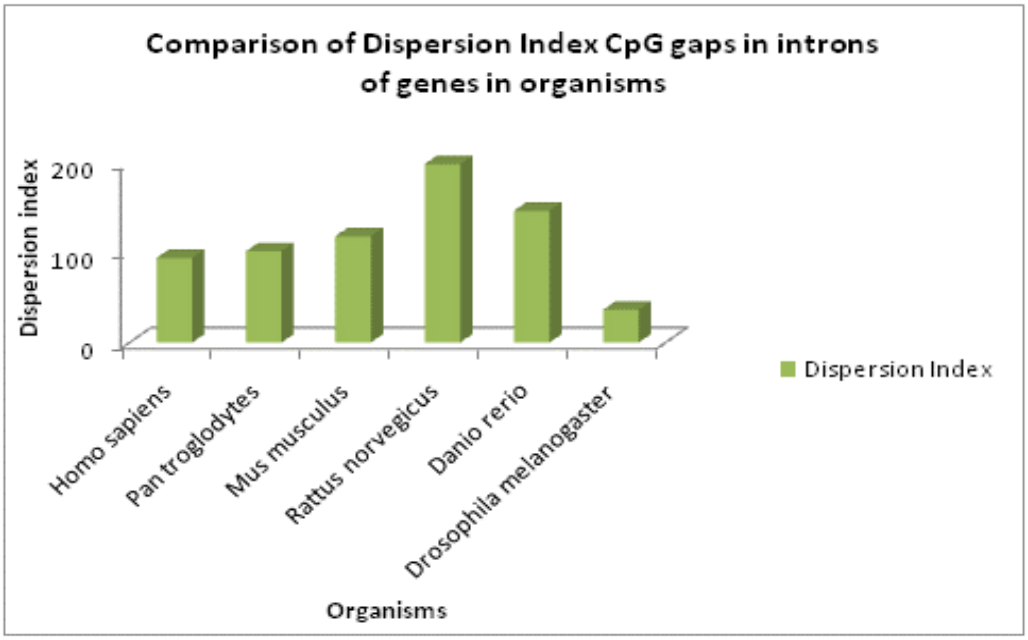

Figure 9: - Dispersion Index of CpG gap size in intons of genes of organisms.

Table.7 Represents Arithmetic Mean and Standard Deviation of distribution of CpG in Intergenic region

Organism	Arithmetic Mean	Standard deviation
Homo sapiens	53.45	71.88
Pan troglodytes	56.25	77.05
Mus musculus	140.67	147.03
Rattus norvegicus	309.11	9394.34
Danio rerio	57.06	96.7
Drosophila melanogaster	25.19	27.49

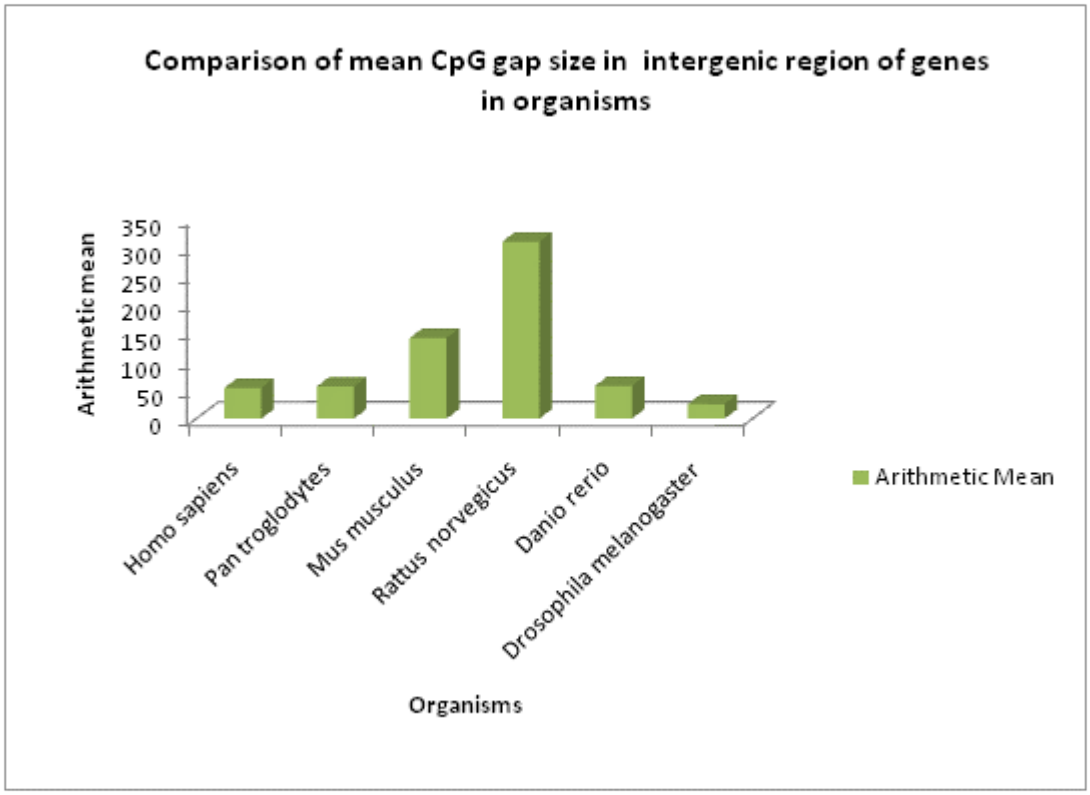

Figure 10:- Arithmetic Mean of CpG gap size in intergenic region of genes inorganisms

Table. 8 Represents the Coefficient of Variance and Index of Dispersion of distribution of CpGs inintergenic regions

Organism	Coefficient of Variance	Dispersion Index
Homo sapiens	134.47	96.66
Pan troglodytes	136.96	105.53
Mus musculus	104.51	153.67
Rattus norvegicus	3039.15	288.91
Danio rerio	169.45	163.86
Drosophila melanogaster	109.12	30

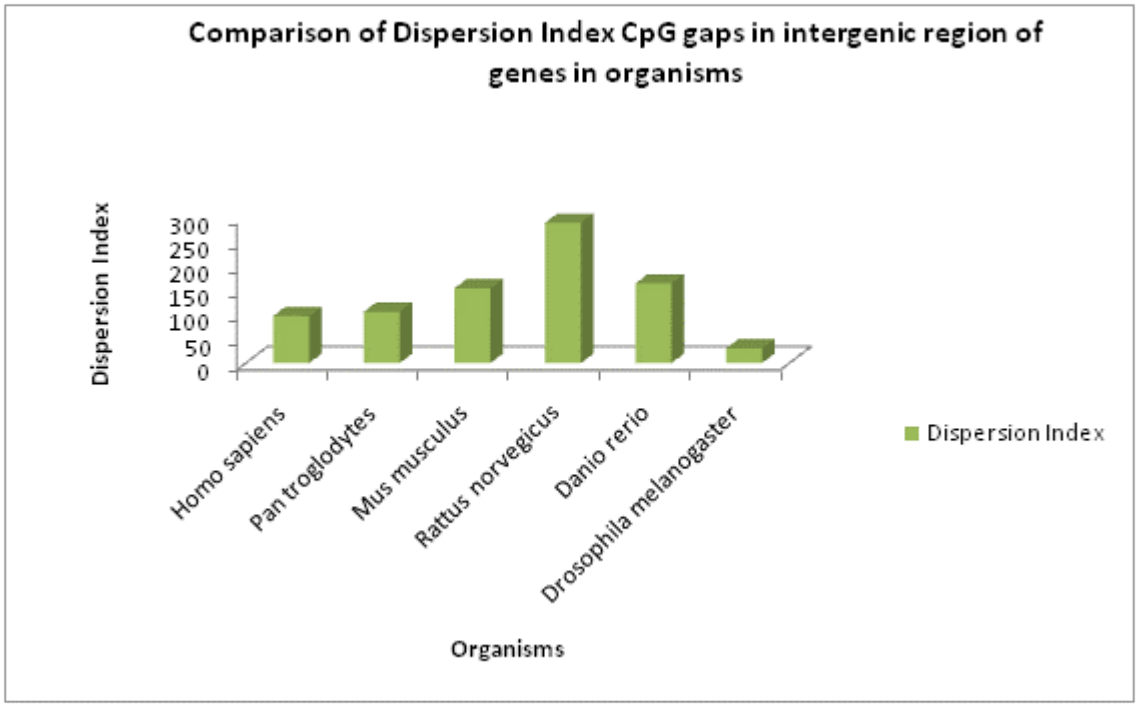

Figure11:- Dispersion Index of CpG gap size in intergenic region of genes of organisms

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. With an exception of *Rattusnorvegicus* genome sequence, all the organisms exhibited higher DI values for intergenic regions when compared to exons or introns in general. Largely DI value in introns was marginally greater than exons. This shows that CGs are more overdispersed or clustered in intergenic regions when compared to exons or introns in general. Since exons are most conserved and intergenic regions have lowest conservation, it may be inferred that the clustering of CGs in the genomes correspond well with the decrease in evolutionary pressure.

Distribution pattern of CpGs in CpG Islands and non CpG Island regions

After comparing CpG distribution in exons, introns and intergenic regions of differently methylated genomes, it was attempted to understand the effect of distribution of CpGs in CpG islands and non CpG island regions. It was found that selected sequences spanning 100 genes each for the six genomes had maximum number of CpG islands in humans but there was no correlation between degree of genome methylation and number of CpG islands. Neither any relationship could be established between degree of genome methylation and fraction of CpGs as part of CpG islands.

Table.9Represents the statistical analysis of CGI

Organism	No. of CGI	CGI/10000 bp	No. of CGs in CGI	Total no. of CGs	Fraction of CGs in CGI
Homo sapiens	234	0.023	11894	290555	0.041
Pan troglodytes	47	0.005	620	34849	0.018
Mus musculus	52	0.005	3032	122317	0.025
Rattus norvegicus	86	0.009	2255	52756	0.043
Danio rerio	121	0.012	2460	74584	0.033

Table. 10 Represents the Arithmetic mean of gaps in CpGdinucleotides

Organism	Arithmetic mean		
	CGI	nCGI	Total sequence
Homo sapiens	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
Drosophila melanogaster	13	24	21.86

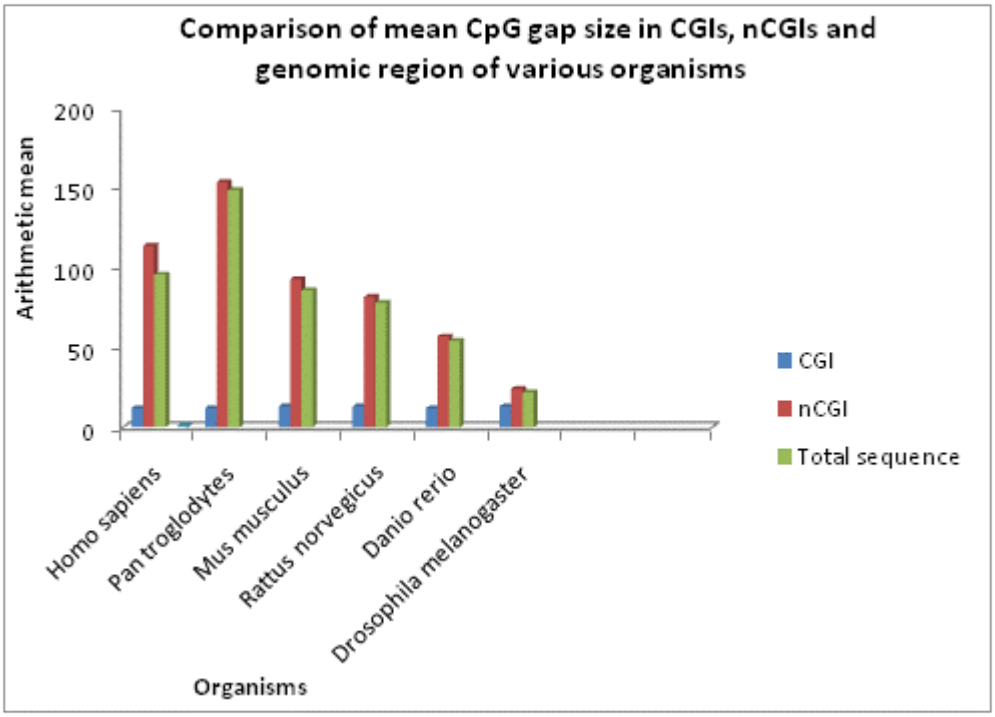

Figure 12:- Arithmetic Mean of CpG gap size in CGIs, nCGIs and genomic region of various organisms

Table.11 Represents the Coefficient of Variance and Index of Dispersion of CpG/ non CpG islands and total sequence

Organism	Coefficient of variance			Index of dispersion		
	CGI	nCGI	Total sequence	CGI	nCGI	Total sequence
Homo sapiens	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
Drosophila melanogaster	115	122	126.3	18	36	34.89

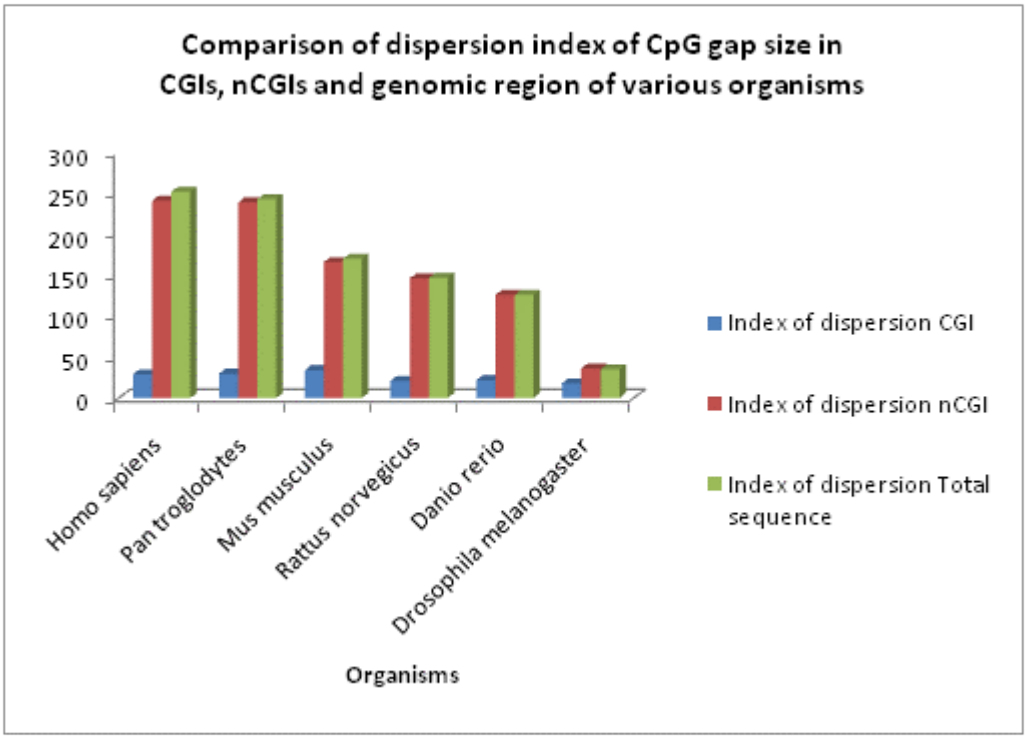

Figure14:- Dispersion Index of CpG gap size in CGIs, nCGIs and genomic region of variousorganisms

Discussion

CpGdinucleotides are known to have different types of distributions in methylated eukaryotic genomes when compared to the unmethylated ones due to mutagenicity of the methylated CpGdinucleotides. CpG islands of methylated genomes diplaying high abundance of CpGs are good examples to show the difference in distributions as rest of the genome has underrepresentation of CpGs. CpG gap size has been used as a tool to study CpG distribution in the (methylated) genomes. This study corroborates the earlier findings of CpG underrepresentation of methylated genomes and the effect of evolution on the loss of CpGs. It may be inferred that CpG distribution in methylated genomes is anticipated by factors such as 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 until now based on existence of CpG islands with relatively very high CpG density when compared to rest of the genome. However it has been found that CpGs in non-CpG islands regions also play very important role. Greater part of them remain methylated in contrast to usually unmethylatedCpGs of CpG islands but some of them are differentially methylated depending on spatial, termporal or other factors.

In this study the distribution of CpGs of non CpG island regions in comparison with CGI regions and overall genome classified into coding (exons) and noncoding regions (introns and intergenic regions) has been evaluated. The distributions of CpGs in one dimensional space can be even, random or clustered. Dispersion index of CpG gap sizes has been used to evaluate degree of clustering of CpGs and was determined for this purpose. Dispersion index is ratio of variance and mean. DI < 1 represent even distribution, DI = 0 is random distribution whereas DI > 1 represents over dispersion and clustering 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 indicate that the clustering of CpGdistribution is not merely because of CpG islands and nCGI regions but also within nCGI regions. Thus it can be inferred that despite 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 largely containing repeat sequences (usually remain methylated) have highest DI values. Based on these preliminary results it may be inferred that the study may be expanded to improve the understanding of CpG distributions in methyalted genomes.

Comparison of CpG gap sizes of all the organisms reflects CpG density and distribution in the genome. Comparison of six 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 organisms exhibiting different levels of methylation in their genomes. Genomic sequences of six species were analyzed and CpG gap size data was generated. The species included four mammals, one non-mammal vertebrates (fish), one insect. DNA sequences were taken from GenBank database for all the species. The gaps between each CpG were determined using method 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.

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 pattern similar to the mean gap size values, the DI also was highest for primate genomes followed by murine genomes, which was followed by the fish genomes. The poorly methylated and unmethylated genomes 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 clearly indicate that CpG distribution is clustered even outside the CGIs. The DI values of nCGIof 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.

From this study it may be inferred that CpG clustering is higher in non coding regions such as intergenic regions and introns when compared to coding regions of exons in general. The effect is largely more pronounced in methylated genome. The result indicate that mutagenicity of methylated CpGs leads to non random loss of CpGs in nonCpG islands thereby increasing the overdispersion or clustering of CpGs

CHAPTER 7
CONCLUSION

Conclusion

This study of CpG distributions in genomes of different evolutionary lineages was aimed to compare distribution of CpGs in coding and non coding regions of methylated genomes. For this purpose initially it was confirmed that gap size between CpGdinucleotides can be used as a tool to study their distribution by comparing their mean, standard deviation and dispersion index of six organisms. Then CpG gap size statistical analysis was compared amongst six different species. It was concluded that CpGs are more clustered, as expected due to their loss in course of evolution, and over dispersed in intergenic regions when compared to exons or even introns. The effect was more pronounced in methylated genomes. A similar pattern was observed in nCGI regions of these genomes also whereas CGIs of all the organisms

exhibited similar mean and DI values. 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. AharonRazin 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, andCpG 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, CpGRich 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. FrancIscoAntequera 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 CpGdinucleotides 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 CpGdinucleotides in vertebrate genomes under a mutation model.” ProcNatlAcadSci 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." BiochemBiophys Res Commun**382:** 643-645.