Studies on relation between DNA Methylation levels and sequence characteristics of mammalian DNA

Submitted in partial fulfilment of the requirements of the

Degree of

MASTER OF SCIENCE IN BIOTECHNOLOGY

Under the guidance of:
Dr. Vikas Handa
Assistant Professor

Submitted by:
Jasbir kaur
Roll no. 301001012

DEPARTMENT OF BIOTECHNOLOGY AND ENVIRONMENTAL SCIENCES
THAPAR UNIVERSITY, PATIALA

July, 2012

CANDIDATE'S DECLARATION

I, hereby declare that the work presented in the dissertation entitled **"Studies on relation between DNA Methylation levels and sequence characteristics of mammalian DNA"** in the partial fulfilment of the requirement for the award of the degree of Master in Biotechnology, Department of Biotechnology and Environment Sciences, Thapar University, Patiala is an authentic record of my own work during the period of six months from January 2012 to June 2012, under the supervision of **Dr. Vikas Handa**, Assistant Professor, Department of Biotechnology and Environment Sciences, Thapar University. The report has not been submitted for the award of any other degree or certificate in this or any other university.

Jasbir Kaur

Place: Patiala

Date: July 18, 2012

Jasbir Kaur

301001012

CERTIFICATE

This is to certify that the thesis entitled **"Studies on relation between DNA Methylation levels and sequence characteristics of mammalian DNA"** submitted by Jasbir Kaur 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 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.

18/07/2012

Dr. Vikas Handa
Supervisor
DBTES, TU
Patiala

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

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

ACKNOWLEDGEMENT

It would not have been possible without the kind support, inspiration, guidance, direction, cooperation, love and 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 their guidance and constant supervision as well as for providing necessary information regarding the thesis work & also for their 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.

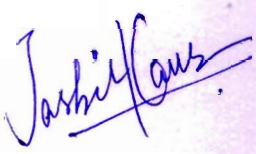
I would like to express my gratitude towards my parents & members of Thapar University for their kind co-operation and encouragement which help me in completion of this work.

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

Date: July 18, 2012

Place: Patiala

Jasbir Kaur



CONTENTS

Page no.

ABBREVATIONS

A -- Adenine

AdoMet -- S-Adenosyl-L-Methonine

C -- Cytosine

C^5 -- carbon at 6^{th} position

Dnmt -- DNA methyltransferase

Dnmt3a -- DNA methyltransferase 3a

Dnmt3b -- DNA methyltransferase 3b

ExpCpG -- expected frequency of CpG dinucleotide

G -- Guanine

N^5 -- nitrogen at 5^{th} position

N^4 -- nitrogen at 4^{th} position

ObsCpg -- Observed frequency of CpG dinucleotide

ObsCpG/ExpCpg -- ratio of observed frequency of CpG over the expected frequency of CpG dinucleotide

Poly B -- non Adenine sequence

Poly D -- non Cytosine sequence

Poly H -- non Guanine sequence

Poly V -- non Thymine sequence

Poly R -- Polypurines in the sequence

Poly Y – Polypyrimidines stretches

S – Strong G/C regions

W – Weak A/T regions.

LISTS OF TABLES

Page no.

LISTS OF FIGURES

Page no.

CHAPTER 1

ABTRACT

ABSTRACT

DNA methylation is an epigenetic modification that occurs at N^6 position of Adenine and N^4 and C^5 positions of Cytosine in prokaryotes while, in higher eukaryotes it occurs at C^5 position of Cytosine. DNA methylation is performed by DNA methyltransferases, maintenance methyltransferases Dnmt1 that has preference for hemimethylated CpG sites and de novo methyltransferases Dnmt3a Dnmt3b which do not discriminate between hemimethylated and unmethylated CpG sites. Target CpG sites are known to be differentially methylated in the genome and the phenomenon has been shown to be influenced by the local DNA sequence characteristics. In the present study it has been attempted to investigate effects of sequence composition such as homopolymeric stretches of purines, pyrimidines, poly H, V, D and B on DNA methylation. Recently published bisulphite sequence analysis based methylation data of 297 amplicones of chromosome 21 of 5 human cell types has been used to compare mean methylation levels with above mentioned DNA sequence characteristics.

CHAPTER 2

INTRODUCTION

INTRODUCTION

"Epigenetics" is a heritable patterns of gene expression or gene function that initiated and maintained without changing the underlying DNA sequence (Zhang, C.Rohde et al. 2009). It refers to functionally relevant modifications to the genome that do not involve a change in the nucleotide sequence. Examples of epigenetics changes are DNA methylation and histone modification, both of which serve to regulate gene expression without altering the underlying DNA sequence. The epigenetic changes comprise the different modification of the histones proteins including acetylation, ubiquitination, phosphorylation and methylation.

DNA methylation is a biochemical process that is important for normal development in higher organism. It involves the addition of a methyl group to the 5 position of the Cytosine pyrimidine ring or the number 6 nitrogen of the Adenine purine ring (H. Gowher and Jeltsch. 2004). DNA methylation ranges from very low level in arthropods, through intermediate levels in many non-arthropod intermediates, to high levels in vertebrates. The methylation pattern established by the enzymes in the CpG dinucleotides regardless of the sequence setting, as the susceptibility of the CpG towards methylation appears to be independent of the nucleotide sequence (Bird 1980). DNA methylation in vertebrates typically occurs at CpG sites (Cytosine-phosphate-Guanine) sites, that is, where a Cytosine is immediately followed by a Guanine at 3' end in the DNA sequence. The methylation of Cytosines results in the conversion of the Cytosine to 5-methylCytosine. CpG dinucleotides are the "hotspots" for mutation in the vertebrate genomes as a result of the modification of the 5' Cytosine by cellular DNA methyltransferases and their spontaneous deamination to thymidine (N.Cooper and krawczak. 1982). This results in conversion of Cytosine into Uracil, through a process known as hydrolytic deamination. When this happens, the Guanine that was initially bound to the Cytosine molecule is left opposite to Uracil (Uracil normally base pairs with Adenine). When the cell next replicates its DNA, the position opposite the Uracil molecule would be taken up by an Adenine instead of the Guanine that should be there, altering the message that this DNA encodes. Thus a CpG/CpG is converted into TpG/CpA.

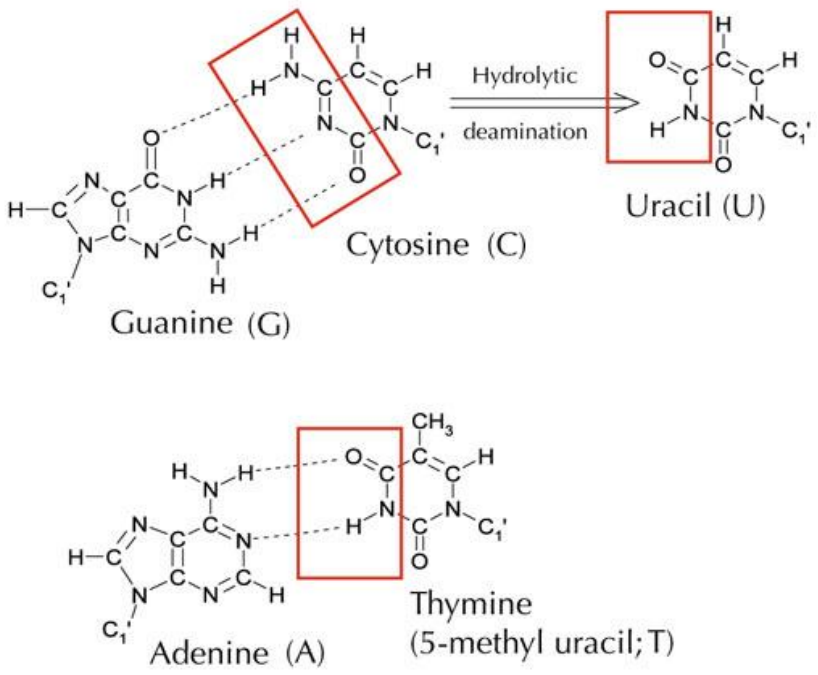


Figure 1. AT & GC Base pairing.

A Cytosine : Guanine pair is directly mutated to a Uracil : Guanine mismatch. As Uracil residue do not found normally in the DNA, as it is RNA base, therefore, to maintain the integrity of the DNA genome these mutations must be repaired by high-fidelity DNA mismatch repair enzymes. The Uracil bases are removed by the repair enzyme, Uracil-DNA glycosylase. The Uracil which is labeled with methyl group is paired with Adenine – resulting in Thymine. This way, if the cell machinery found a Uracil, it cut it out and repaired it, but if it found a Uracil with a methyl label – a Thymine (5-m-Uracil) (Liu and M.Schatz, 2009).

The other prominent epigenetic modification takes place in histones. Native cell within the human genome is packaged with histones and other proteins into chromatin. Many studies over recent years in mammals have identified a wide range of post-translational modifications to the N-terminal tails of the histones in chromatin. Chromatin, is the physiological template of all eukaryotic genetic information, is subject to a diverse array of posttranslational modifications that largely impinge on histone amino termini, thereby regulating access to the underlying DNA. Distinct histone amino-terminal modifications can generate synergistic or antagonistic interaction affinities for chromatin-associated proteins, which in turn dictate dynamic transitions

between transcriptionally active or transcriptionally silent chromatin states (Jenuwein and Allis. 2001); (Martin and Zhang. 2005). These include a series of methylations and acetylations at defined lysine and arginine residues. Methylated histones have been implicated in heterochromatic repression, promoter regulation and the propagation of a repressed state via DNA methylation (Kouzarides. 2002). A growing literature is defining the mechanisms for addition and removal of the modifications catalyzed by a range of methyl (Kouzarides. 2002);(Martin and Zhang. 2005) and acetyl-transferases (Roth, Denu et al. 2001) deacetylases, (Kurdistani and Grunstein. 2003), and most recently demethylases (Shi Y, Lan et al. 2004).

DNA methylation reaction is catalyzed by the enzymes called DNA methyltransferases. All methyltransferases use S-adenosyl-L-methonine (AdoMet) as the source of methyl group which is transferred to the substrate (DNA bases in the case of DNA methyltransferases). The methyl group of AdoMet is bound to the sulphonium atom, which thermodynamically destabilize the DNA molecule (H. Gowher and Jeltsch. 2004). In mammalian cells, DNA methylation is carried out by two general classes of enzymatic activities – maintenance methylation and *de novo* methylation. Maintenance methylation activity is necessary to preserve DNA methylation pattern after every cellular DNA replication cycle. By virtue of its strong preference for hemi-methylated CpGs as its substrates against unmethylated CpGs, Dnmt1 plays the role of maintenance methyltransferase. As illustrated in the figure 2, due to semi-conservative mode of DNA replication, the methylated DNA gives rise to two hemi-methylated daughter DNA molecules. The parent strand is methylated while the newly synthesized daughter strand is unmethylated in both the DNA molecules. Immediately after replication, Dnmt1 acts on the newly formed DNA molecules methylating hemi-methylated sites (that were also methylated in the parent DNA molecule) while ignores the unmethylated sites (that were unmethylated in parent DNA molecule). Thus, the enzyme is responsible for copying DNA methylation patterns to the daughter strands after DNA replication cycle (R.Taby and J.Pierrej, 2010).

Dnmt3a and Dnmt3b are the *de novo* methyltransferases (do not exhibit any substrate preference for hemi-methylated or unmethylated CpGs) that set up DNA methylation patterns early in development. Dnmt2 (TRDmt1) has been identified as a DNA methyltransferase homolog, containing all 10 sequence motifs common to all DNA methyltransferases; however, Dnmt2 (TRDmt1) does not methylate DNA but instead methylates Cytosine-38 in the anticodon

loop of aspartic acid transfer RNA. Dnmt3L is the third member of Dnmt family. Its N-terminus comprises of plant homedomain. Dnmt3L folds like a DNA methyltransferases but cannot have any catalytic activity. Dnmt3L is expressed during gametogenesis and embryonic stages, showing an expression pattern similar to Dnmt3a and Dnmt3b (H. Gowher and Jeltsch. 2004).

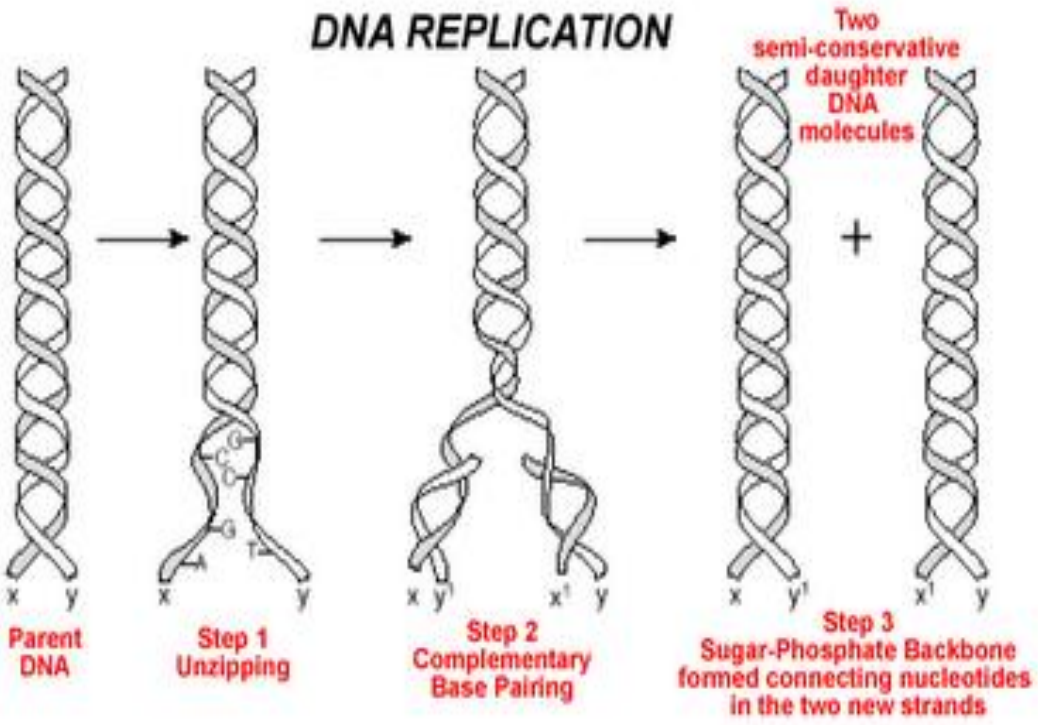


Figure 2. DNA replication.

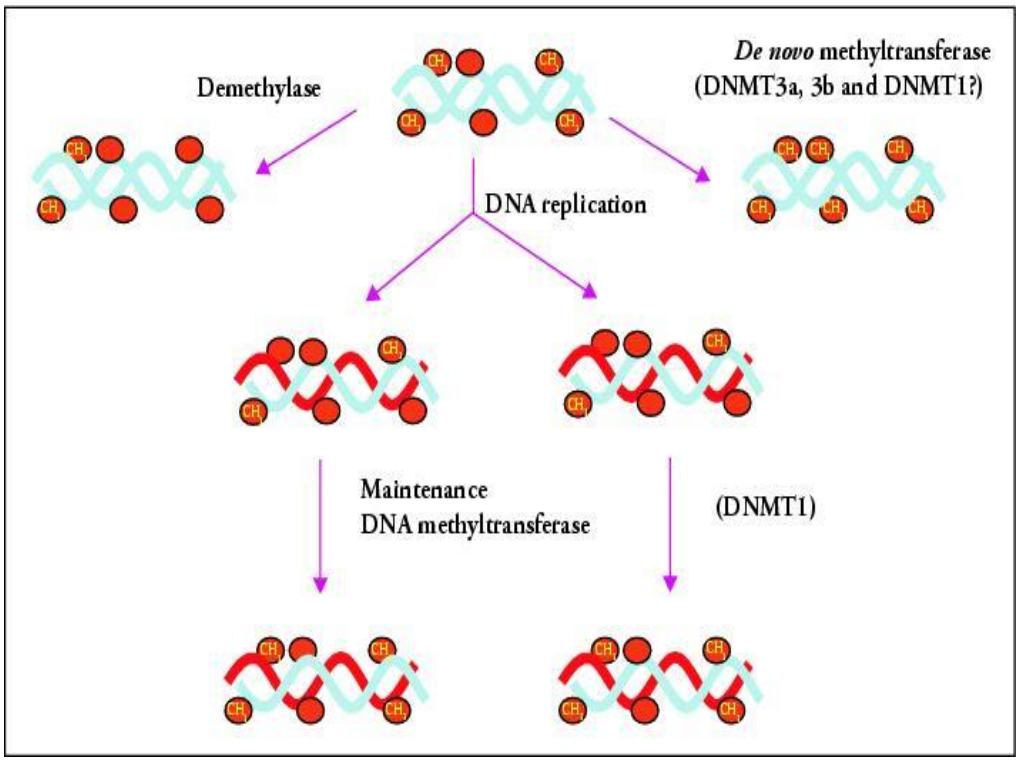


Figure3. Activity of DNA methyltransferases.

DNA methylation plays important role in embryonic development, gametogenesis, X chromosome inactivation of female mammals, regulation of chromatin structure, silencing of transposons and endogenous retroviruses, cancer and gene imprinting. DNA methylation is a crucial part of normal organismal development and cellular differentiation in higher organisms. DNA methylation is typically removed during zygote formation and re-established through successive cell divisions during development. Genomic imprinting implies the two parental chromosomes are not equivalent and show either maternal or paternal-specific expression at a subset of genes in the genome. These patterns are set up by differential DNA methylation marking at the imprinting control regions in male and female germ line (Strogantsev R and AC. 2012). Female mammalian cells silence one of their two X chromosomes, resulting in equal expression levels of X-encoded genes in female XX and male XY cells. During developmental, imprinted or random X chromosome inactivation (XCI) is

initiated and both processes lead to an inactive X chromosome that is clonally inherited (Barakat TS and J. 2012). X-chromosome inactivation (XCI) results in one inactive X chromosome (Xi) and one active X chromosome (Xa). Based on studies using somatic cell hybrids the majority of X- linked genes have been determined to be subject to XCI and are only expressed from the Xa(active), while approximately 15% of genes escape XCI and are expressed from both the Xa and the Xi (Carrel L and Willard. 2005). Error in the DNA methylation contributes to development of human cancer and multifactorial diseases. Complexity of human carcinogenesis cannot define by genetic alterations alone, but cancer also involves epigenetic changes in processes such as DNA methylation. DNA Methylation play important role in activation of oncogenic or inactivation of tumor suppressor pathways that results in tumorgenesis. DNA methylation has secondary role in gene inactivation. Methylation has a role in maintenance of gene repression via inappropriate activation of cancerous genes which are strongly correlated with abnormal CpG methylation patterns (S.Pennings, J.Allan et al. 2004). Methylation is an alternate way of silencing tumors suppressor genes, is equivalent to genetic mutations. DNA methylation also plays an important role in gene silencing as it regulates the chromatin structure.

DNA methylation may affect the transcription of genes in two ways. First, the methylation of DNA itself may physically impede the binding of transcriptional proteins to the gene, and second and likely more important, methylated DNA may be bound by proteins known as methyl-CpG-binding domain proteins (MBDs). MBD proteins then recruit additional proteins to the locus, such as histone deacetylase and other chromatin remodeling proteins that can modify histones, thereby forming compact, inactive chromatin, termed as heterochromatin.

In mammals, DNA methylation predominantly occurs at CpG dinucleotides, which are methylated under normal cell conditions. In addition to that due to mutation of CpG sites, they are underrepresented in the vertebrate genomes. However there are GC rich sequences containing clustered unmethylated CpGs known as CpG-islands. They overlap with the transcription start site (TSS) of about 70% of all human genes and these CpG islands are unmethylated in the normal differentiated cells (Zhang, C.Rohde et al. 2009). CpG islands are associates with the 5’ ends of all housekeeping genes and with many tissue specific genes. The 5’ CpG islands extends through 5’ flanking DNA, exons and introns, whereas most of the 3’ CpG islands appeared to be associated with exons. CpG islands are present in the promoter and

exonic regions of approximately 40% of mammalian genes. CpG islands plays an important role in gene silencing, genomic imprinting, X- chromosome inactivation, silencing of intragenomic parasites and carcinogenesis (G.Gardiner and Frommer. 1986; D.Takai and P.Jones. 2001) These regions are 200 to 300 bp in length, having high GC content around ≥55% and having ObsCpG /ExpCpG ≥0.65 (D.Takai and P.Jones. 2001), (G.Gardiner and Frommer. 1986).

As methylation is an enzymatic process, it is naturally interesting to understand the mechanisms in order to investigate various effects of DNA methylation. It has been demonstrated that DNA methyltransferases pull their target base out of the DNA helix prior to methylation, a process called *base flipping*, the target Cytosine is no longer buried in the double helix but is rotated about its flanking sugar-phosphate bonds so that it projects out the catalytic pocket of the enzyme. The base-pairing hydrogen bonds are broken, and the stacking interactions with the adjacent base pairs are lost during this process. Base flipping is an phenomenon that facilitates deformation of a double-stranded DNA fragment, initiated by the rupture of the Watson and Crick hydrogen bonds at a target base pair and followed by the turning of one of the bases to an extrahelical position, where it subsequently becomes exposed to chemical attack from its environment (B.Bouvier and H.Grubmüller. 2007) This process involves the pushing of the base out of the helix, this push must take place on the sugar phosphate backbone but not on the base. Thus rotation of the DNA backbone is probably the key to base flipping(Horton JR, Roberts RJ et al. 1998). Alteration of base, recruit the action of repair and modification enzyme on specific position. In fact, it is assumed that these enzymes facilitate the flipping mechanism upon binding. X-ray diffraction structures of these enzymes in complex with base-flipped DNA strands have been published (B.Bouvier and H.Grubmüller. 2007). Experimental evidence suggests two different opening mechanisms: Uracil glycosylase facilitates the flipping of its target Uracil through the major groove of the DNA double strand, whereas Cytosine-5 methyltransferase favors a minor groove pathway for its target Cytosine.

Since DNA structure and its distortions are associated with base flipping, the very mechanism of DNA methylation, it is interesting to investigate the factors responsible for various types of distortions in the double helical structure of DNA. One such distortion is bending of DNA. DNA bending occurs in Adenine and Thymine tracts (A/T-tracts) and they also play

significant role in DNA recognition by gene-regulatory proteins (J.Hizver, H.Rozenberg et al. 2001). The bending of the DNA structure occur in the major groove while minor groove exhibit slight bending (X.Yong and M.Sundaralingam. 2000).

Another unusual DNA structure is triple helix that involves poly-purine/poly-pyrimidine stretches. For the Pyrimidine.purine, Pyrimidine motif, all eight combinational strands are composed of either DNA or RNA. The chemical nature of sugars has a dramatic influence on triple helix stability. For each double helix composition, a more stable triple helix was formed when the third strand was RNA rather than DNA. No stable triple helix was detected when the polypurine sequence was made of RNA with a third strand made of DNA. The interactions between the 2'-hydroxyl group of the third strand and the phosphates of the polypurine strand play an important role in determining the relative stabilities of triple-helical structures in which the polypyrimidine third strand is oriented parallel to the polypurine sequence. These interactions are not allowed when the third strand adopts an antiparallel orientation with respect to the target polypurine sequence, as the third strand contains G and A or G and T/U. Transcriptions is inhibited by triple helix formation in eukaryotes include polypyrimidine stretches involved in termination of tRNA transcripts (C.Escudé, J C François et al. 1993).

CHAPTER 3

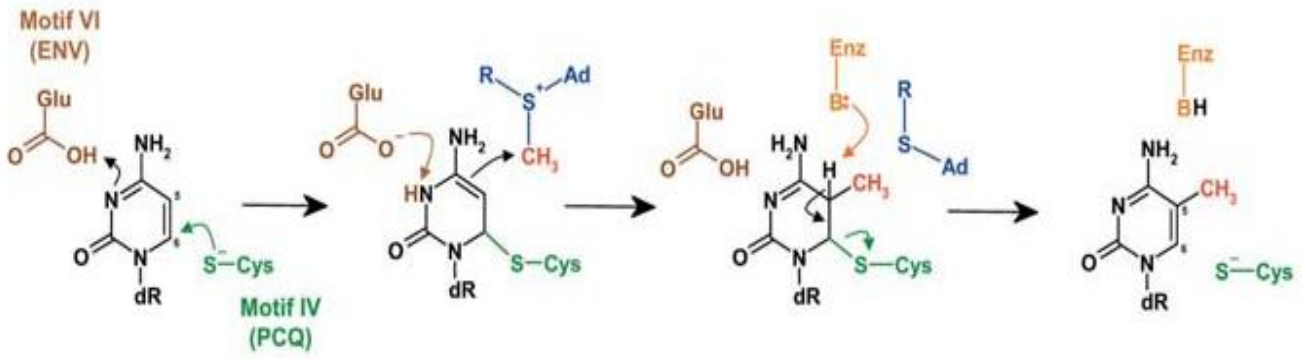
LITERATURE REVIEW

LITERATURE REVIEW

Epigenetics: 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). Both the modification suppresses gene expression without altering the sequence of silenced gene.

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.

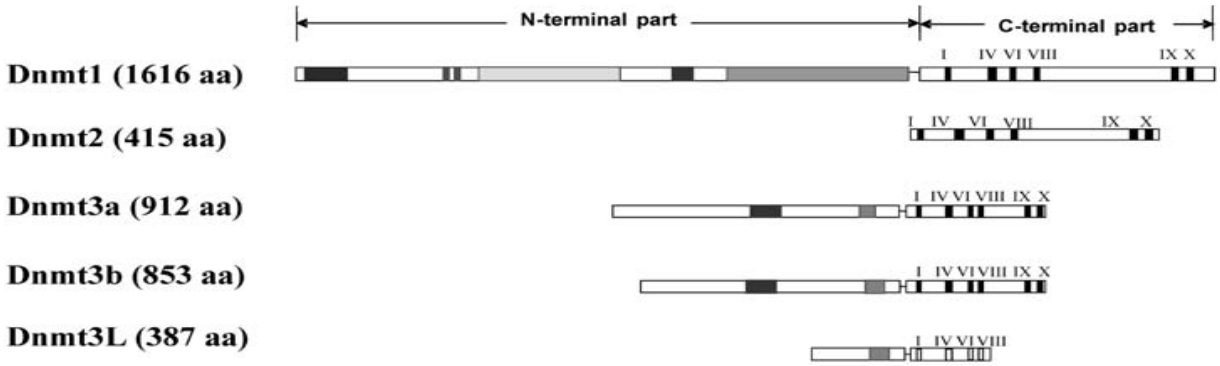
DNA methyltransferases have two different modes of methylation processes: *de novo* methylation establishes the methylation state; maintenance methylation copies it onto daughter DNA strands after DNA replication. The first mammalian DNA methyltransferase discovered was Dnmt1, which is highly conserved among eukaryotes and is responsible for maintaining methylation patterns in the DNA after replication. Later, Dnmt2 and the 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. The methyl group of AdoMet is bound to a sulphonium atom, which thermodynamically destabilizes the molecule and makes the relatively inert methylthiol of the methionine moiety very reactive towards nucleophilic attack by nitrogen, oxygen and sulphur atoms or activated C atoms (carbanions).



(Hermanna 2004)

Figure4: Chemistry of the methylation reaction of DNA-(Cytosine-C5)-DNA methyltransferases. The methylation reaction catalyzed by DNA DNA methyltransferases 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.

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.



(Hermanna 2004)

Figure4. Domain organization of the mammalian Dnmts. The mammalian methyltransferases are divided into an N-terminal regulatory part and a C-terminal catalytic part. The C-terminal part shows strong amino acid sequence homology to prokaryotic DNA-(Cytosine-C5) - DNA methyltransferase and contains all the conserved catalytic amino acid motifs (denoted in Roman numerals) defined for the prokaryotic enzymes. The boxes shown in the DNA methyltransferase sequence indicate the various domains, structural and sequence motifs identified in these proteins.

The Dnmt1 enzyme comprises a large N-terminal domain with regulatory function and a smaller C-terminal catalytic domain (V.Handa and A.Jeltsch. 2005). The regulatory domain harbors different motifs, like a charge-rich domain that interacts with the Dmap1 transcriptional repressor and contains different start codons; a nuclear localization signal, a PCNA (proliferating cell nuclear antigen) interacting domain, a replication foci targeting region and a cysteine-rich Zn^{2+} binding domain of the CXXC type also found in the HRX protein. The zinc domain comprises eight conserved cysteine residues in two CXXCXXC clusters and two isolated cysteines. The essential function of Dnmt1 in the mammalian cell is demonstrated by the observation that mice deficient for Dnmt1 die in midgestation with significantly reduced levels of DNA methylation. Dnmt1 has high preference towards hemimethylated and unmethylated sites (Goyal, Reinhardt et al. 2006)

Though maintenance DNA methyltransferase Dnmt1 plays central role in inheriting methylation patterns in the vertebrate genomes, *de* novo DNA methyltransferases are also indispensable as they perform methylation during early embryogenesis and gametogenesis. The indispensability is evident from the reports of Dnmt3a and Dnmt3b knock-out mice during embryogenesis. Mutations in human Dnmt3b are also linked with a syndrome called ICF

(immunodeficiency, centromeric instability, facial abnormalities), a rare recessive autosomal disorder characterized by hypomethylation at pericentromeric satellite regions. The Dnmt3 enzymes were identified in mouse and human expressed sequence tag (EST) databases by their homology to the bacterial 5mC DNA methyltransferases. The Dnmt3 family consists of Dnmt3a and Dnmt3b, which are highly related to one another but encoded by separate genes. Dnmt3a shows preference to methylate sites that are flanked by pyrimidines rather than purines. Dnmt3b is known to methylate the pericentromeric repeats carrying high CpG content; a processive reaction mechanism supports methylation of this DNA region. Dnmt3a cannot replace Dnmt3b in this function, possibly because of its distribution mechanism, which is less efficient in methylating highly CpG rich DNA. Dnmt3a has strong preference towards CpG rich flanking sequences, who found a strong preference for CpG site flanked by pyrimidine bases and a loose consensus sequence of YNCpGY (Lin, Han et al. 2002).

Effect of flanking sequences on methylation of CpG has been reported on the basis of methylation kinetic studies performed on Dnmt3a and 10 different DNA substrate molecules representing the exhaustive set of $\pm$ 1 flanks and computational analysis of methylation data from human epigenome database. It has been reported that there is more than 13 fold difference in the rate of methylation by Dnmt3a between the most favored and least favored substrates with consensus sequences of RCGY and YCGR respectively. The computational analysis in combination with kinetics studies revealed much higher (~500 fold) difference between the most favored and the least favored substrates when upto $\pm$ 4 flanks were taken into consideration. This was a significant evidence of influence of DNA sequence on CpG methylation. It was also found that the methylation is distributed in a bimodal fashion in the genomes (Handa & Jeltsch, 2005).

Different DNA-related attributes differentiate the methylated and unmethylated CpG islands. These include DNA sequence properties and patterns, repeat frequency and distribution, and predicted DNA structure, most significant attributes the frequencies of GC-rich and CpG-rich DNA sequence patterns, which are overrepresented in unmethylated CpG islands. Non-strand-specific patterns and patterns that are strand-specific relative to the chromosomal plus-strand occur with similar frequency and composition. Several attributes that refer to repetitive DNA are more frequent in methylated CpG islands (such as segmental duplications, self chain

alignments, and tandem repeats) CpG islands that consistently deviate from their default methylation state due to monoallelic methylation (imprinting, X-chromosome inactivation) are characterized by a medium degree of methylation propensity or whether the underlying biological processes are so strong that basically every CpG island can become differentially methylated independently of its DNA sequence (Bock, Walter et al. 2006).

DNA methylation levels are distributed bimodally with enrichment of highly methylated and unmethylated sequences, both for amplicons and individual subclones, which represent single alleles from individual cells. Within CpG-rich sequences, DNA methylation was found to be anti-correlated with CpG dinucleotide density and GC content, and methylated CpGs are more likely to be flanked by AT-rich sequences. The methyaltion pattern of the amplicones gradually decreased when approaching TTS of the respective gene both from upstream and downstream (Zhang, C.Rohde et al. 2009)

Several factors have been studied to identify their effect on DNA methylation. Continuing the trend it is interesting to investigate some of the DNA sequence based attributes that might affect the enzyme activity. One such attribute can be Polypurines and polypyrimidine stretches. Such sequences are known to be involved in formation of unusual DNA structure of triple helix. (Felsenfeld and Rich, 1957). Triplex DNA is formed when a third DNA strand binds into the major groove of a double helix via Hoogsteen hydrogen bonding. Triplex DNA has been used as an powerful tool in the genetic manipulation as triplex DNA can inhibit DNA transcription and replication, generate site-specific mutations, cleave DNA, and induce homologous recombination (Chan and Glazer. 1997). DNA triplexes can naturally occur, co-localize and interact with many other regulatory DNA elements (e.g. G-quadruplex (G4) DNA motifs), specific DNA-binding proteins (e.g. transcription factors (TFs), and micro-RNA (miRNA) precursors (Jenjaroenpun and A Kuznetsov 2009). Frequencies of purines and pyrimidines stretches of 10 bp or more have been associated with several genomic and DNA structural features. For instance, purine and pyrimidine patterns have been found to be better conserved than base composition in all domains of life. Runs of YR stretches tend to form Z-DNA helices in GC-rich sequences, and some purine tracts are associated with A-DNA helices (Bohlin., Hardy et al. 2009) (Y Zhaoyang , R. V. Guntaka et al. 2007). It is known that highly

AT rich base sequences can cause intrinsic bending in DNA. The bending of DNA is expected to affect the double helix structure of DNA which in turn may influence the base flipping mechanism (Koo and Crothers. 1986). (Handa and Jeltsch 2005).

CHAPTER 4

OBJECTIVE

OBJECTIVES

To study DNA sequence attributes affecting the levels of methylation

- To investigate the effect of Polypurine and polypyrimidine sequences on DNA methylation.
- To investigate the effect of Poly V, Poly H, Poly B and Poly D sequences on DNA methylation.
- Study the combinational role of GC and AT rich sequences on the methylation levels.

CHAPTER 5

METHODOLOGY

METHODOLOGY

Data source

All methylation data obtained here are presented in an integrated web platform (http://biochem.jacobs-university.de/name21/) from paper DNA Methylation Analysis of Chromosome 21 Gene Promoters at Single Base Pair and Single Allele Resolution by Yingying Zhang and Albert Jeltsch, 2009. The methylation values of 297 amplicones of 5 different tissues were taken from there to study different aspects of the methylation.

Procedure

The 297 sequences of the 5 tissues were picked from above mention website. This data is analyzed on the Microsoft excel using the applications of spreadsheet. Using algorithm mentioned below and applying macros, the frequency of different lengths of R, Y, H, B, V, D, S and W were determined for each sequence. The sequences, one at a time were copied from the data and pasted on Microsoft word. By using the tool 'Replace' the desired bases were converted into '0' while rest of all the bases were converted into 'x'. Then the 'x0' pairs were replaced with '10' and finally all the 'x' were removed. This process converted the sequence into 1 followed by 0s representing the stretches of desired base(s). These '1' and '0' consisting strings were subjected to line breaks just before each '1' giving smaller strings each beginning with '1' followed by '0's only. The data was transferred to Excel spread sheet where $\log_{10}$ values delivered the number of '0's which actually is the length of repeat of desired base(s).

To study the impact of stretches of different bases, Pearson coefficient of correlation was determined between the mean methylation of sequences and the frequency of stretches of the different base(s) using Microsoft Excel spreadsheet.

CHAPTER 6

RESULTS

RESULTS

Effect of polypurine and polypyrimidine sequences on DNA methylation

Polypurine and polypyrimidine sequences of varying lengths were identified in all the 297 sequences mentioned above. The frequency of these sequences was determined and compared against the mean methylation value. No significant Pearson coefficient of correlation value was obtained for polypurine or polypyrimidine sequences of varying lengths (R_2 to R_{30} & Y_2 to Y_{24}) as well as the sums of their frequencies at different lengths. It may be inferred that no such polypurine or polypyrimidine stretches were there that could affect the structure of DNA influencing the methylation of CpG sites in these 297 sequences.

Effect of polyH, polyB, polyV and polyD sequences on DNA methylation

Genomes are significantly heterogeneous in base sequence composition and it has significant influence on the structural and functional aspects of genetic information. This makes it interesting to look into effect of different levels of complexity of genetic information in general and on DNA methylation in our case. To study this effect, one of our attempts was to study effect of polyH, polyB, polyV and polyD sequences on DNA methylation. We carried out a study similar to that of Polypurines and Polypyrimidines as mentioned above.

PolyH, polyB, polyV and polyD sequences of varying lengths were identified in all the 297 sequences mentioned above. The frequency of these sequences was determined and compared against the mean methylation value. In addition to that sum of frequencies of any two, any three and all four were also used for correlation studies. No significant Pearson coefficient of correlation value was obtained for any type of the sequences of lower complexity or their combinations. However some weak correlation was found in the case of some of frequencies of polyH and polyV. As the length increases there is an upward trend of correlation value that peaks at (H_7 + V_7) and again decreases. Similarly for (polyD + polyV) and (polyB + polyD +polyV) also weak positive correlations for very short lengths and as the length increases the correlation decreases. Though the none of the correlation value goes beyond 0.3, the number of

sequences (297) involved makes it statistically significant. However as it is combination of H & V or D & V or B, D & V, it is difficult to infer any further useful information from the data from the current results.

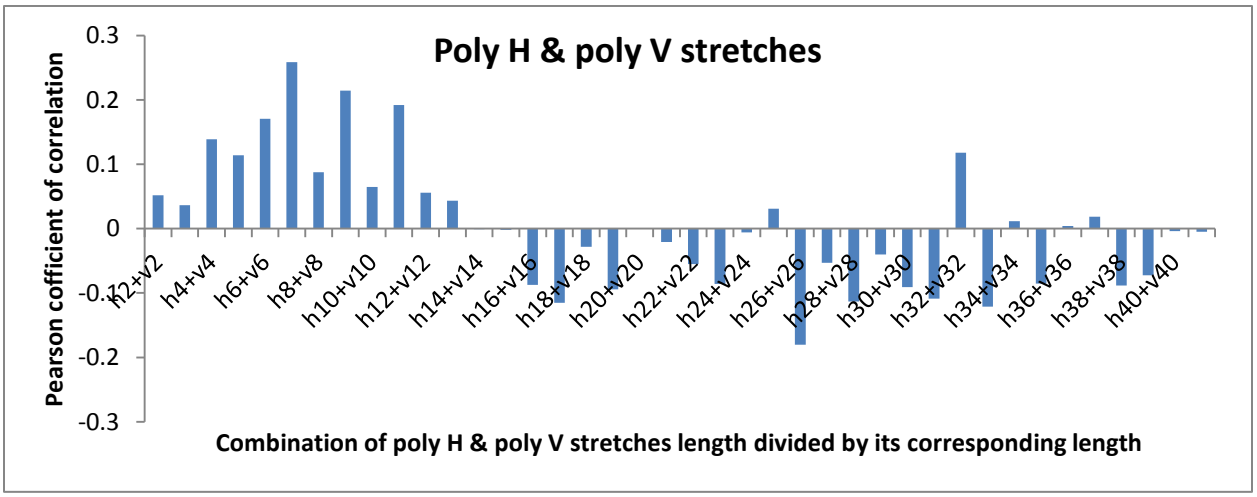


Table 1.

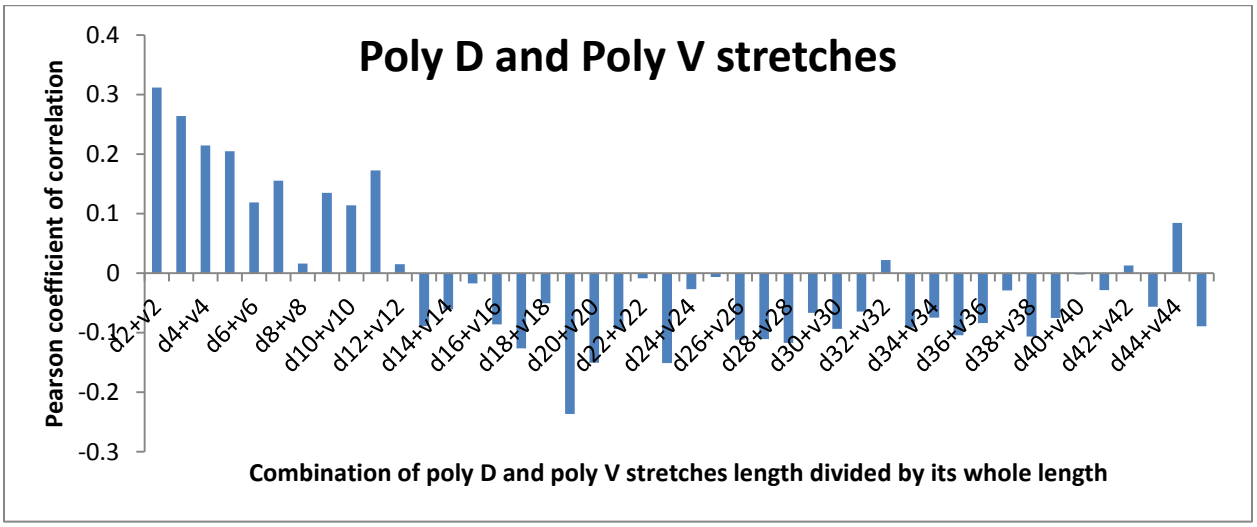


Table 2

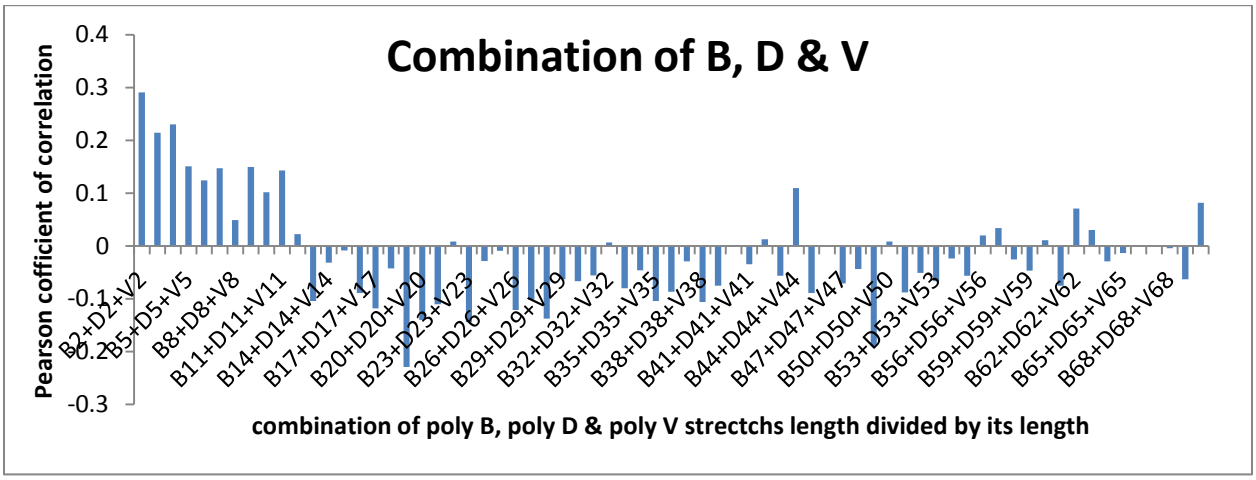


Table 3.

Effect of polyA/T and polyG/C sequences on DNA methylation

Since A/T stretches are known to cause intrinsic bending of DNA it may be logical to expect it to effect DNA methylation as it involves base flipping. In order to investigate any such effect, correlation studies were carried out comparing frequencies of A/T or G/C stretches and the mean methylation values of the 297 sequences used for previous studies also. No significant correlations were obtained in the case of G/C stretches' frequency while significant positive correlation of (r = 0.40) was obtained for W_2 sequences' frequencies and methylation levels. This correlation value followed a downward trend along the length of the poly A/T stretches and reached a minimum of (r = -0.25) for W_9 stretches. It is already a known fact that high AT content in the regions increases the propensity of DNA methylation and significant positive correlation value for W_2 frequencies may be a reflection of that fact only. But as the length of A/T stretches increases, the correlation is sharply lost and rather a weak negative correlation peaks at W_9 length.

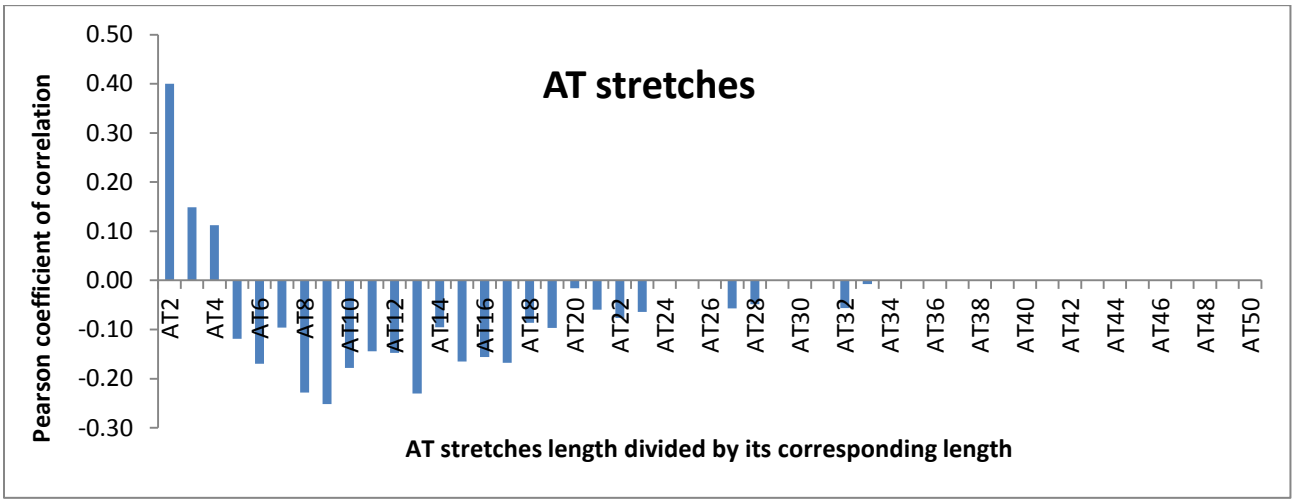

AT stretches
Pearson coefficient of correlation
0.50
0.40
0.30
0.20
0.10
0.00
-0.10
-0.20
-0.30
AT2
AT4
AT6
AT8
AT10
AT12
AT14
AT16
AT18
AT20
AT22
AT24
AT26
AT28
AT30
AT32
AT34
AT36
AT38
AT40
AT42
AT44
AT46
AT48
AT50
AT stretches length divided by its corresponding length

Table 4.

CHAPTER 7

DISCUSSION

DISCUSSION

DNA methylation is an enzymatic process that involves intimate interaction between enzymes Dnmts and substrate DNA. Though a canonical sequence of CpG is the substrate requirement of DNA methyltransferases, owing to the small size of CpG substrate and in turn fairly large sized enzymes Dnmts methylating them, it is expected that flanking bases might also have effect on the catalysis. This question has been addressed in previous reports in which effect of flanking bases has been studied. In recent such report based on biochemical as well as computational analysis of physiological data it has been shown that RCGY consensus sequence is preferred by *de novo* DNA methyltrasferases and YCGR is least preferred substrates. More over farther lying flanking bases' effect has also been studied to yield 10 bp long highly preferred and unpreferred sequences as substrates. In addition to that several other attributes of DNA sequence have also been studied such as GC content, CpG frequencies, distance between adjacent CpGs and certain other repeat or non-repeat sequences influencing methylation of DNA.

Following similar lines, it has been attempted to look into some more unexplored DNA sequence bases attributes for their effect on DNA methylation. The investigation was initiated with exploring the effect of poly purine or poly pyrimidine stretches on DNA methylation. A correlation analysis carried out between frequencies of poly purine stretches or poly pyrimidine stretches of varying lengths. No significant Pearson coefficient of correlation value (r) was obtained. Poly purine or poly pyrimidine stretches were selected for this study because of their involvement in formation of triplex DNA and effect on DNA structure. Next step was to study the influence of sequence complexity on DNA methylation for which polyH, polyB, polyV and polyD stretches were studied. Neither any of these four sequences nor did their combinations yield any significant 'r' value with a few exceptions. Theses exceptions are weak valued and largely stand out of context. It may be inferred that we need to use some more sophisticated approach to study the influence of sequence complexity on DNA methylation.

In the third experiment similar analysis was performed but with polyA/T (W) or polyG/C (S) stretches. In this case a weak but significant positive correlation was obtained with W_2 frequencies with mean methylation values. The correlation also followed a trend and weakened with increasing length of W_n. At the other end a weak significant negative correlation was

obtained for W_9 stretch frequencies. The positive correlation may be explained as more or less reflection of AT content of the sequences that has been reported to promote DNA methylation. But negative correlation between longer AT stretches indicates their role in deforming the regular double helical structure of DNA adversely affecting DNA methylation. It is a well established fact that polyW regions bend the DNA in absence of any external factor (eg protein binding). Any deformity in the regular structure of DNA may affect ease with which base flipping can take place which is the key step of DNA methylation catalysis. So it may be concluded that AT rich regions promote DNA methylation but longer A/T stretches may adversely affect the same phenomenon owing to their effect on the DNA structure. The inference may be confirmed by experimental evidence.

CHAPTER 8

CONCLUSION

CONCLUSION

DNA methylation is an important epigenetic modification that influenced several vital processes. DNA methylation is found to be heterogeneous in the genome and thus proves that it is a regulated process. This enzymatic process is known to be influenced by several factors related to DNA sequence. In the present study some new DNA sequence based attributes have been studied using approach of correlation studies. It was found that poly A/T stretches in the sequence have dual effect on DNA methylation. Very small AT stretches (As, Ts, ATs & TAs) are positively correlated with DNA methylation while longer stretches (eg W_9) are negatively correlated with DNA methylation. The latter observation may be reasoned by the fact that poly A/T stretches are know to bend the DNA thereby affecting its regular double helical structure.

CHAPTER 9

REFERENCES

REFERENCES

1. Bouvier and H.Grubmüller. (2007). "A Molecular Dynamics Study of Slow Base Flipping in DNA using Conformational Flooding." Biophys **3**: 770-786.

2. Barakat (2012). "X chromosome inactivation in the cycle of life." Development **139**: 2085-2089.

3. Bird, P. (1980). "DNA methylation and the frequency of CpG in animal DNA." Nucleic Acids Research **8**.

4. Bock, Walter, et al. (2006). "CpG Island Methylation in Human Lymphocytes Is Highly Correlated with DNA Sequence, Repeats, and Predicted DNA Structure" Plos genetics .

5. Bohlin, Hardy et al, (2009). "Stretches of alternating pyrimidine/purines and purines are respectively linked with pathogenicity and growth temperature in prokaryotes." BMC Genomics **10**.

6. C.Escudé, François, et al. (1993). "Stability of triple helices containing RNA and DNA strands: experimental and molecular modeling studies." Nucleic Acids Research **21**: 5547-5553.

7. Carrel L and Willard (2005). "X-inactivation profile reveals extensive variability in X-linked gene expression in females." Nature **434**: 400-404.

8. Chan and Glazer ,(1997). "Triplex DNA: fundamentals, advances, and potential applications for gene therapy." Mol Med **75**: 267-282.

9. D.Takai and P.Jones. (2001). "Comprehensive analysis of CpG islands in human chromosomes 21 and 22." Biochemistry.

10. G.Gardiner and M. Frommer. (1986). "CpG Islands in Vertebrate Genomes" J. Mol. Biol. **196**: 261-282.

11. Goyal, Reinhardt et al, (2006). "Accuracy of DNA methylation pattern preservation by the Dnmt1 methyltransferase." Nucleic Acids Research **34**: 1182–1188.

12. H. Gowher and A. Jeltsch. (2004). "Biochemistry and biology of mammalian DNA methyltransferases." cell and molecular life sciences. **61**: 2571–2587.

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

14. Hermanna, A. Jeltsch, Gowher (2004). "Biochemistry and biology of mammalian DNA methyltransferases. ." Cell. Mol. Life Sci **61**: 2571–2587.

15. Horton JR, Roberts RJ, et al. (1998). "Structures of HhaI methyltransferase complexed with substrates containing mismatches at the target base." Structral Biol. **10**: 872-877.

16. J.Hizver, H.Rozenberg, et al. (2001). "DNA bending by an adenine–thymine tract and its role in gene regulation." structral Biology.

17. Jenjaroenpun, and A. Kuznetsov (2009). "TTS Mapping: integrative WEB tool for analysis of triplex formation target DNA Sequences, G-quadruplets and non-protein coding regulatory DNA elements in the human genome " BMC genomics **10**: 1471-2164-1410-S1473-S1479.

18. Jenuwein and Allis. (2001). "Translating the histone code." Science **293**: 1074-1080.

19. Kahng, and L. Shapiro (2001). "The CcrM DNA Methyltransferase of Agrobacterium tumefaciens Is Essential, and Its Activity Is Cell Cycle Regulated. ." Bacteriol. **183**: 3065-3075

20. Klimasauskas, Kumar et al (1994). "HhaI methyltransferase flips its target base out of the DNA helix" "Cell. **2**

21. Koo and Crothers. (1986). "DNA bending at adenine . thymine tracts." Nature: 5011-5016.

22. Kouzarides. (2002). "Histone methylation in transcriptional control." Curr Opin Genet Dev **2**: 198-209.

23. Kurdistani and Grunstein. (2003). "Histone acetylation and deacetylation in yeast." Mol Cell Biol **4**:276-284.

24. Lin, Han et al. (2002). "Murine de novo methyltransferase Dnmt3a demonstrates strand asymmetry and site preference in the methylation of DNA in vitro. ." Mol. Cell. Biol. **22**: 704– 723.

25. Martin and Zhang. (2005). "The diverse functions of histone lysine methylation." Mol Cell Biol **11**: 838-849.

26. N.Cooper and M. krawczak. (1982). "Cytosine methylation and the fate of CpG dinucleotides in vertebrates genomes." Human genentics **83**: 181-188.

27. Roth, Denu, et al. (2001). "Histone acetyltransferases." Biochemistry **70**: 81-120.

28. S.Pennings, J.Allan, et al. (2004). "DNA methylation ,nucleosome formation and positioning." genomics ang proteomics **3**: 351-361.

29. Shi Y, Lan et al. (2004). "Histone demethylation mediated by the nuclear amine oxidase homolog LSD1." cell **119**: 941-953.

30. Strogantsev (2012). "Proteins involved in establishment and maintenance of imprinted methylation marks." Genomics. **3**: 227-239.

31. V.Handa and A.Jeltsch. (2005). "Profound Flanking Sequence Preference of Dnmt3a and Dnmt3b Mammalian DNA Methyltransferases Shape the Human Epigenome." Mol Bio **348**: 1103–1112.

32. X.Yong and M.Sundaralingam. (2000). "Crystal structure of a DNA:RNA hybrid duplex with a polypurine RNA and a complementary polypyrimidine DNA." Nucleic Acids Research **28**: 2171-2176.

33. Zhaoyang , R. V. Guntaka, et al. (2007). "Sequence-specific Triple Helix Formation with Genomic DNA." Biochemistry **40**: 11240–11252.

34. Zhang, C.Rohde, et al. (2009). "DNA Methylation Analysis of Chromosome 21 Gene Promoters at Single Base Pair and Single Allele Resolution." Plos genetics **5**.

35. Zhang, Y., C. Rohde, et al. (2009). "DNA methylation analysis by bisulfite conversion, cloning, and sequencing of individual clones. ." Methods Mol Biol **507**: 177–187.