Effects of DNA Methylation on Adenovirus genome evolution

A Dissertation Report

Submitted in partial fulfilment of the requirement

For the award of degree of

Masters of Technology

In

Biotechnology

Under the guidance of

Dr. Vikas Handa

Assistant Professor

Submitted by

Meera Sharma

Roll no. 601504006

DEPARTMENT OF BIOTECHNOLOGY

THAPAR UNIVERSITY

PATIALA-147004

July 2017

CANDIDATE DECLARATION

I hereby declare that the work being presented in the M.Tech dissertation entitled "**Effects of DNA Methylation on Adenovirus genome evolution**" has been carried out by me during the period of July 2017 to July 2018, under the guidance of Dr. Vikas Handa, Associate Professor, Department Of Biotechnology, Thapar University, Patiala. Further, I declare that I have not submitted the matter embodied in this dissertation for the award of any other degree or any other qualification of any university or examining body in India/elsewhere.

Meera

Meera Sharma
M.Tech Biotechnology
Roll No. 601504006

Date: 21 August 2017
Place: Thapar University
Patiala.

CERTIFICATE

This is to certify that dissertation entitled "**Effects of DNA Methylation on Adenovirus genome evolution**" submitted by **Meera Sharma (601504006)** in partial fulfilment of the requirements for the award of Masters in technology in Biotechnology to Thapar University, Patiala is an authentic work carried out by her under my supervision and guidance.

To the best of our knowledge, the matter embodied in this dissertation has not been submitted to award of any Degree or certificate in any other university/institute.

Dr. Vikas Handa
Assistant Professor
Department of Biotechnology
Thapar University
Patiala

Meera Sharma
(601504006)
M-Tech (Biotechnology)
Thapar University
Patiala

ACKNOWLEDGEMENT

First of all I would like to thanks Almighty God for his constant blessings, who has guided me to work on the right path of the life.

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

I express my special gratitude to Dr. Moushmi Gosh, Head, Department of Biotechnology, Thapar University, Patiala, for all his possible support in various facilities of the department for this work. I am really pleased to acknowledge the kind help, cooperation and moral support which I have received throughout my dissertation from of all the teaching as well as non teaching faculty members of Department of Biotechnology, which helped me a lot in completion of this work.

I am really thankful to Dr. Rana for their guidance and valuable advice throughout my work. With heartiest reverence I admire confidence bestowed on me by my parents. The untiring pains taking dedicated help, affection and blessing received from them to bring me to this level, it is beyond my capacity to express in words.

Meera

Meera Sharma

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

DNA Methylation is a form of epigenetic modification which plays important role in regulating many cellular processes including gametogenesis, early embryogenesis, cellular differentiation and development, genomic imprinting in the mammals. Vertebrates have been reported to show CpG suppression. Viruses infecting these vertebrates have co-evolved with their hosts and thus show similar under-representation of CpGs within their genomes. We attempted to study genomes of Adenoviral strains causing infection in humans. Our data shows that in Human Adenoviruses, CpG dinucleotides are underrepresented while TpG and CpA being over-represented. The Adenovirus genome sequences were subjected to Multiple Sequence Alignment for identifying the CG dinucleotide conserved positions. Mutations of CGs at conserved positions provide strong evidence that CpG mutations have bias for TpG/CpA in comparison with any other dinucleotide sequence thereby indicating a strong association with DNA methylation in Adenoviral genome.

Keywords: DNA Methylation, Adenoviral genome, CpG suppression, Proteome

CHAPTER 1

INTRODUCTION

The term 'Epigenetics' given by the biologist Conrad Waddington in 1940 described it as "the gene interaction with their environment which bring the phenotype into being". Epigenetics is primarily focused on studying the heritable changes rather than the mutations in the primary sequence of DNA. Since no mutations are introduced in the DNA sequence, the genotype remains intact but variations occur in gene expression and activity, causing changes in the phenotype (Goldberg *et al.*, 2007). The two primary processes that are responsible for introducing epigenetic modifications to the genome are DNA Methylation and Histone Modifications.

DNA Methylation is a modification involving addition of methyl group ($-CH_3$) at N^4 and C^5 position of Cytosine and N^6 position of Adenine in prokaryotes, whereas at C^5 position of Cytosine in eukaryotes (Hoelzer *et al*., 2008).

Histone Modifications is a process involving winding of DNA around the protein called 'Histone' and is then further compacted into chromosomes. Histone Modifications occur due to alteration in the extent to which a DNA is wrapped around the histone proteins, thereby affecting the availability of genes for activation (Handy *et al*., 2011). These modifications occur in response to the binding of epigenetic factors (carried as chemical tags) to the histone tails, affecting the DNA wrapping and gene activation (Strahl and Aliis, 2000).

1.1 DNA Methylation

DNA Methylation is a primary process that results in the epigenetic modification of the genomes. It is a mechanism through which sustainable transmission of epigenetic information is done through multiple cycles of DNA replication and cell division (Hermann *et al*., 2004). The methylation of DNA in mammals plays a crucial role in the development of embryo, genomic imprinting, inactivation of X chromosome, regulating the structure of chromatin, silencing of transposons and endogenous retroviruses, genetic diseases and cancer biology (Bird, 2002, Li, 2002 & Feinberg and Tycko, 2004). This process involves a covalent modification by addition of a methyl group to the DNA strand in such a way that the Watson-

Crick base pairing capacity is not interfered. In prokaryotes, DNA Methylation occurs at N4 and C5 position of Cytosine residues and N6 position of Adenine residues, whereas in eukaryotes it occurs at C5 position of Cytosine, mainly in context of CpG dinucleotides (Bird, 1980).

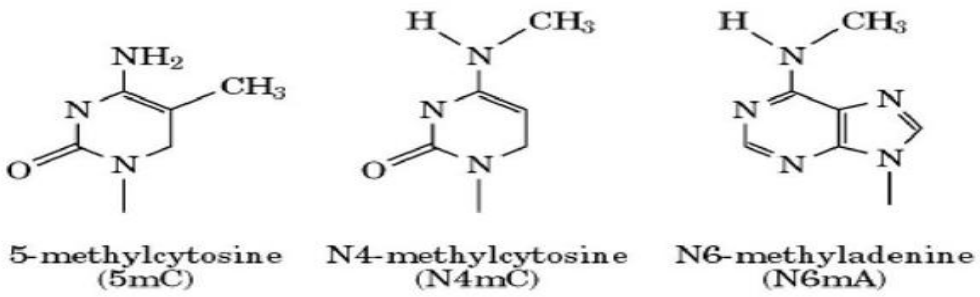
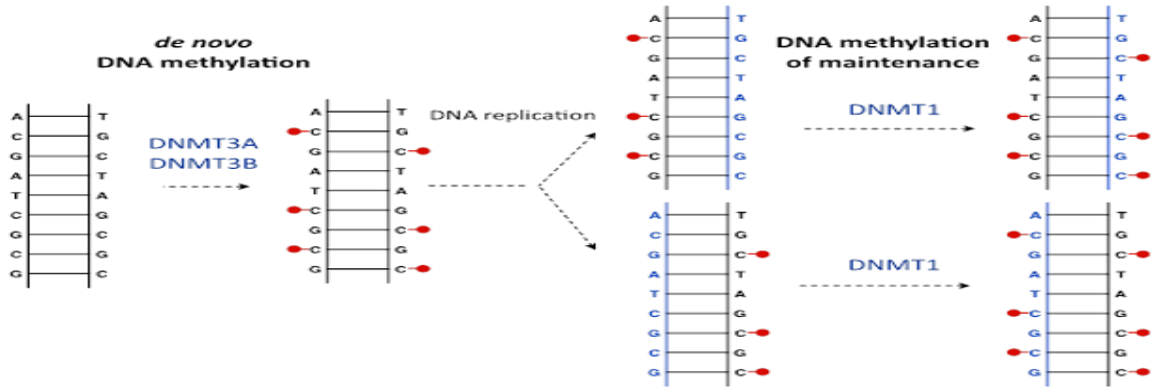


Figure 1: Methylation positions at Cytosine and Adenine base

The Cytosine within 5'-CpG-3' dinucleotides serves as a site for covalent modification of DNA methylation, where a methyl group is transferred to the 5th position of Cytosine to generate 5-methyl Cytosine (5-mC) in genomic DNA (Phillips, 2008). A molecule of S-adenosylmethionine (SAM)^{6} serves as a methyl group donor for enzymatically driven methylation process. Apart from 5-methylcytosine (5-mC), an oxidative modified form of 5-mC is also present in the mammalian genomic DNA which is 5-hydroxymethylcytosine (5hmc) and is termed as the sixth base within the DNA (Ratel *et al*., 2006 & Wu *et al*., 2016).

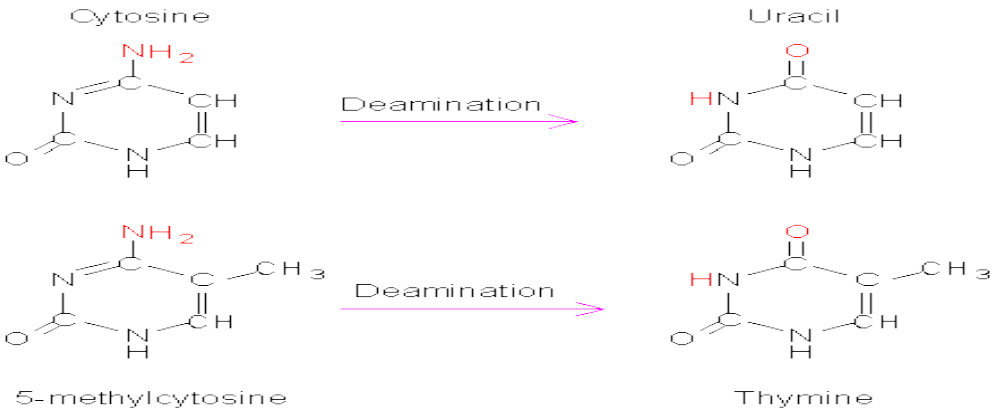
The enzymes that are responsible for carrying out the covalent modification of cytosine to 5-methylcytosine are known as DNA Methyltransferases (DNMT). On the basis of their activity, they are further classified as: *De-novo* Methyltransferases which includes DNMT 3a and 3b that are responsible for setting up of methylation patterns early during embryogenesis and later during gametogenesis (Flynn *et al*., 1998 & Hoelzer *et al*., 2008). Maintenance Methyltransferases have strong preference for hemi-methylated DNA and thus responsible for inheriting these patterns to daughter strands during DNA replication (Pradhan *et al*., 1999).



Source: Moison *et al.*, 2014

Figure 2: For DNA methylation maintenance DNMT1 is mostly involved whereas for *de novo* methyltransferases DNMT3a and DNMT3b involved. Red lollipop indicated the methyl group on the CpG site**.**

The methylated Cytosine undergoes spontaneous deamination giving rise to Thymine. This transition mutation is not recognized by cellular DNA repair machinery and thus not usually repaired making this conversion irreversible. Deamination of unmethylated Cytosine base results in formation of Uracil, which is recognized by cellular Uracil-DNA glycosylase pathway, and is therefore repaired as shown in figure (Cooper and Youssoufian, 1988 & Hoelzer *et al.*, 2008). Thus the phenomena of Cytosine getting converted to Thymine causes mutations that are not repaired and are left as it is which accounts in under-representation of CpG dinucleotides leading to their suppression in genome of organism.



Source: http://www.web-books.com/MoBio

Figure 3: Spontaneous deamination of Cytosine leads to Uracil whereas of methylated cytosine to thymine.

It has been observed that the mammalian genome has uneven distribution of CpG dinucleotides, making some regions within the genome CpG rich and other being CpG poor. The high frequency CpG rich regions within the mammalian genome are termed as "CpG islands", which make punctuations in the DNA sequences. These regions remain unmethylated and are generally associated with constitutive gene promoters and having independent state of expression (Jones *et al*., 2009). Earlier studies have shown that both vertebrate and invertebrates shows suppressive counts of CpG in their genomic sequences. Also, nucleotide substitutions occurring at average rate have been reported to be most rapid at dinucleotides involving CpG. Further studies have also been carried out in the viruses infecting vertebrates (Cardon *et al.*, 1994). Observations inferring lower CpG relative abundance among viral genomes are reported, indicating that these dinucleotides are also methylated in viruses and play important role in the evolution of their genomes. Significant levels of suppression in CpG dinucleotides have occurred in small viral genomes having vertebral hosts. The reason behind this suppression is that the loss of CpG dinucleotides results in minimization of stimulation of toll like receptor 9 and thus lowering immunogenic response against viruses. Also, the CpG methylation renders epigenetic gene silencing in the viral genome (Karlin *et al*., 1994).

Earlier studies provide evidence for the co-evolution of the viral genome along with the host. Analysis on the relative abundance of the dinucleotides on large DNA gives information about various host related factors that play role in shaping evolution of viruses. Along with this the evolutionary pressures i.e. whether translational selection or mutational pressure contributes in the evolution of viruses is also studied. Analysis of the codon usage bias patterns along with the genomic GC content strongly supports that among the DNA viruses, genome wide mutational pressure is the primary factor that determines codon usage, rather than natural selection of coding triplet codons (Upadhyay *et al*., 2014).

1.2 Adenovirus

Adenovirus belong to a family of linear, double stranded DNA, non-enveloped viruses having medium size range (90-100 nm). The GC content of virus lies within the range of 48 - 56%. The non-segmented genome of Adenovirus ranges in size between 26 to 48 kbp. The classification of Adenovirus family is as follows (Davison *et al.*, 2003).

Genus	Species
Atadenovirus	Ovine
Aviadenovirus	Fowl
Mastadenovirus	Mammals
Siadenovirus	Frog
Unclassified	

Table 1: Classification of Adenovirus

Human Adenovirus possesses remarkable capacity to spread infections and cause a wide range of illness. Adenovirus accounts for causing acute respiratory infections and can lead to bronchitis, pneumonia follicular conjunctivitis and multi organ disease among those with weakened immune response. Transmission occurs primarily via droplets of ocular and respiratory secretions. Host defences include neutralizing antibodies as well as cytotoxic-T lymphocytes which are activated in response to viral infection (Saha *et al.,* 2014).

Variations in genome organization exist depending on the species and genera, but generally the Adenoviral transcription unit is divided into Early (E) and Late (L) regions, depending on whether they are expressed before or after DNA replication.

1.2.1 Early Region

This region is transcribed first during viral replication and encodes for proteins that are essential for early adenoviral infection cycle. This region promotes viral infection by altering the cellular environment (Stone *et al.*, 2003). Early region is classified as:

Regions	Function
E1 (E1a/E2b)	Stimulates activation of other genes and induce mitosis in host cell
E2 (E2a/E2b)	Initiates transcription and codes for proteins necessary for viral replication
E3	Essential for modulation of host function, therefore altering host immune responses during disease pathogenesis
E4	Alters cell signalling pathways in host cells

Table 2: Early gene regions with their functions

1.2.2 Late Region

Late region genes are activated after the expression of early function genes i.e. following viral DNA replication and synthesis. This region encodes viral structural proteins primarily (Doerfler, 1996).

Regions	Functions
L1	Encodes protein that helps in facilitating the packaging of virus by aiding in capsid assembly
L2	Polypeptide produced play role in virus internalization via integrins
L3	Polypeptide encoded is a component of viral capsid where it bridges viral capsid and core components
L4	Is responsible for selectively activating the regions that encode for late viral protein synthesis as well as facilitates hexon assembly
L5	Produces polypeptide which trimerize to produce virus fiber

Table 3: Late gene regions with their functions

Adenoviruses while infecting humans having methylated genomes are probably getting methylated and therefore show co-evolution with their host genomes. Due to this, the CpG dinucleotides of Adenoviral genome are expected to be suppressed in comparison to other dinucleotides abundance (Karlin *et al*., 1994). Present study is to investigate the effect of methylation on Adenoviral genome evolution via CpG suppression.

Chapter 2

Review of Literature

2.1 DNA Methylation

It is a process of covalent modification in genomic DNA by addition of a methyl group resulting in modulation of the gene expression. The features of CpG Methylation process including heritability and reversibility makes it a dynamic system suitable for regulating the development of organism (Egger *et al.*, 2004). Also, understanding the role of DNA Methylation in cellular regulation has provided potential for a new archetype of disease intervention and treatment (Crider *et al.*, 2012). The mammalian genome undergoes reprogramming of DNA Methylation patterns. Initiation of this process starts with complete de-methylation of DNA in the pre-implanted embryos and is also known to occur in germ cells. This step is crucial for erasing the existing epigenetic information and for resetting the system for a new developmental cycle (Watters, 2006). Various developmental pathological conditions as well as diseases in the later stages of life are known to occur due to aberration in these methylation patterns (Robertson, 2005). Since DNA Methylation patterns follows a dynamic state, two different methylation processes exists which are: *De-novo* Methylation and Maintenance Methylation. The former one is responsible for establishing the methylation state or patterns early in development whereas the latter is required for copying the methylation patterns onto daughter DNA strands after DNA replication has occurred (Hermann *et al.*, 2004).

2.1.1 Mechanism of DNA Methylation

The process of DNA Methylation involves a covalent modification on the cytosine base. This modification occurs mostly in a 5'-CG-3' dinucleotides pattern by adding a methyl group to Carbon-5 position of the Cytosine pyrimidine ring (Ma *et al*., 2013). The enzyme family that are involved in this process are termed as DNA Methyltransferases (DNMTs). S-adenosyl-L-methionine (Adomet) serves as a common methyl group ($-CH_3$) donor to the DNA bases for all DNA Methyltransferases.

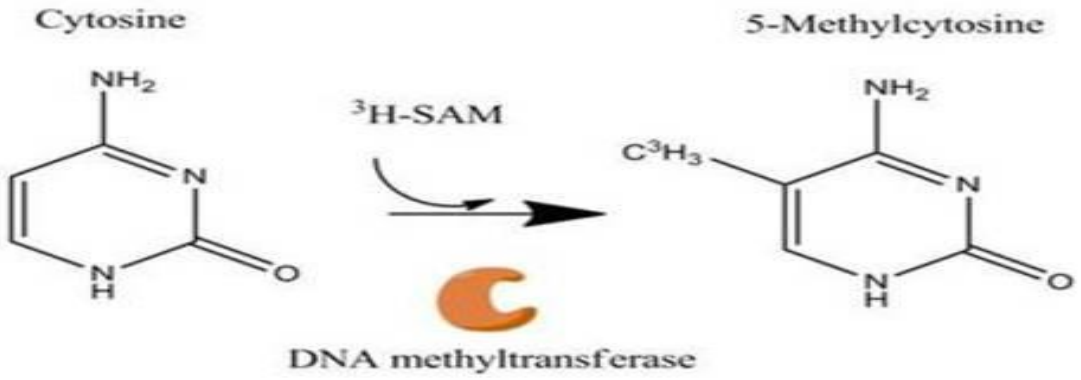


Figure 4: Methylation of Cytosine base carried out by DNA Methyltransferase (DNMT) using S-Adenosylmethionone (SAM) as methyl donor.

Thermodynamic de-stability is introduced by the methyl group of Adomet molecule which is bound to a sulphonium atom, making it highly reactive for a nucleophilic attack by oxygen, nitrogen, sulphur and by activated carbon atoms (Cedar *et al*., 2012). The prokaryotic DNMT M.HhaI was the first enzyme for which the mechanism of DNA Methylation was analyzed. This enzyme specifically recognizes the 5'-GCGC-3' sequence, causing methylation of the first cytosine lying within this sequence (Portela *et al.,* 2010).

2.1.2 DNA Methyltransferases

The process of DNA Methylation in organisms is an enzymatically driven process. The enzymes which carry out the process of adding a methyl group to DNA are called as DNA Methyltransferases (DNMTs) and are responsible for carrying out the epigenetic modification of DNA Methylation. On the basis of structure and function, mammalian DNMTs are classified majorly into two families of *de-novo* methyltransferases (DNMT3a and DNMT 3b) and maintenance methyltransferases (DNMT1).

2.1.2.1 DNA Methyltransferase 1 (Dnmt 1)

Dnmt 1, the first discovered DNA Methyltransferase is accountable for the inheritance of methylation patterns to daughter cells during the process of DNA replication, therefore is known as 'Maintenance methyltransferase' (Doerfler, 2008). *In-vitro* studies have shown that this enzyme has strong preference for hemi-methylated DNA over the un-methylated ones,

owing to its functionality for maintaining the existing methylation patterns after DNA replication. Therefore, during replication DNMT 1 is required for copying the information to the newly synthesized strands of DNA (Jin *et al*., 2011).

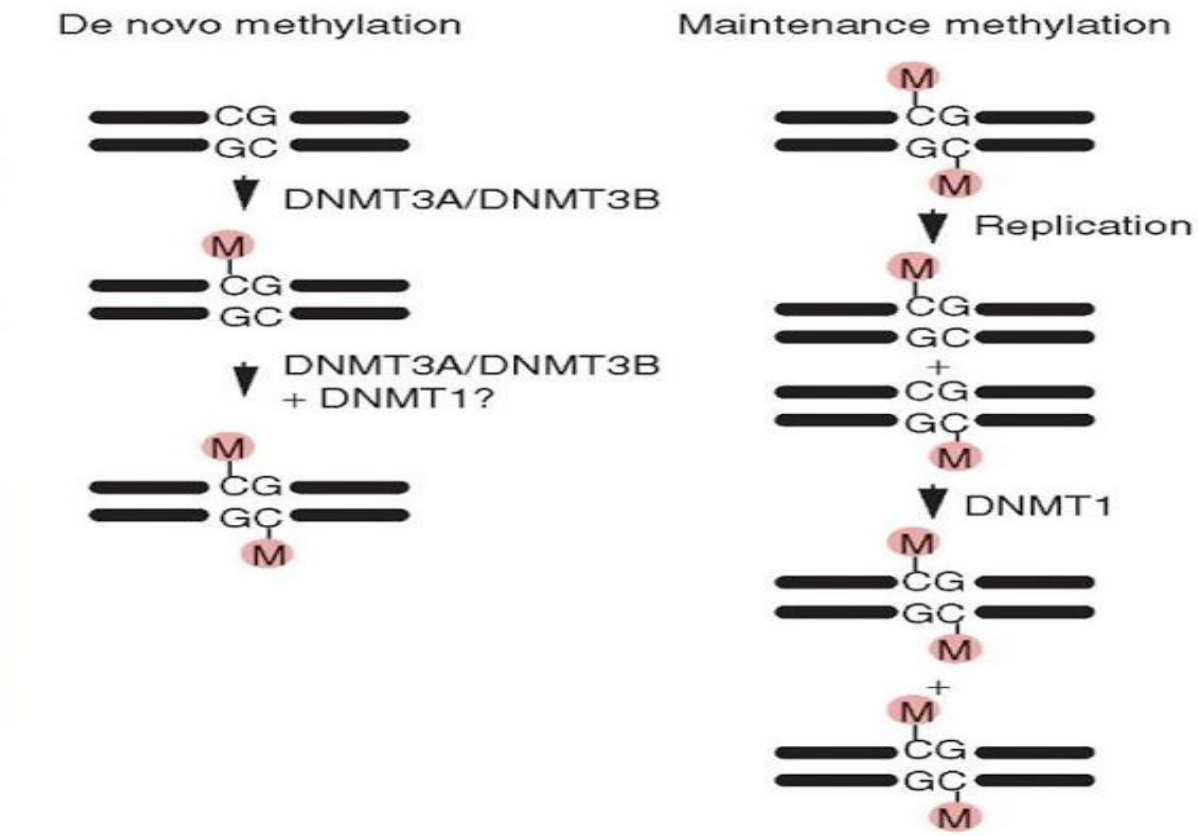


Figure 5: Schematic representation of de novo methylation and maintenance methylation of DNA.

2.1.2.2 Family of DNA Methyltransferase 3 (DNMT3a and DNMT3b)

DNMT 3 family consists of DNMT 3a and DNMT 3b which are accountable for setting up of genomic methylation patterns during embryogenesis as well as gametogenesis, therefore known as '*de-novo* methyltransferase'. These enzymes are thus involved in establishing the patterns of CpG Methylation during embryonic development (Zucker *et al*., 1985). This family also includes DNMT like- protein (DNMT 3L) which shows no catalytic activity but is involved in the physical association of DNMT 3a and DNMT 3b and have a role in modulating their activity (Li, 2002 & Meehan, 2003).

2.1.2.3 DNA Methyltransferase 2 (DNMT2)

DNMT 2 shares structural similarity with the rest of the DNMTs. It has been reported to play role in the methylation of RNA, therefore acts as a RNA Methyltransferase (Hermann *et al.,* 2004). Significance of DNMT 2 is observed in the methylation of tRNA where it specifically methylates the Cytosine base at 38^{th} position of transfer RNA^{Asp}. Since methylation of tRNA has effect on folding of protein and stability of its structure, therefore might have a protective function (Goll *et al*., 2006).

2.1.3 CpG suppression

Most eukaryotes shows suppression of CpG's in their genome. This CpG frequency suppression is highly variable among the species and correlates negatively with the extent and presence of methylated Cytosine in the genome (Hoelzer *et al.,* 2008). While considering the GC content of mammalian genomes, CpG dinucleotides shows under representation. On the basis of base composition, only 25% of CpG abundance is observed of what is expected. The fraction of G+C content in the human DNA is 0.4, it is expected that the frequency of occurrence of the CpG dinucleotides is 0.2 x 0.2 = 0.4, but in contrast the frequency observed is about 0.008 (Bird, 1980).

Also, the CpG dinucleotides show uneven distribution in the genome, giving rise to clusters called as CpG islands (Cardon *et al*., 1993). After complete genome sequence analysis of human chromosome 21 and 22, the CpG islands are defined as the regions of DNA having >500 bp with a GC content of >55% and observed/expected CpG dinucleotides ratio of 0.65 (Takai *et al.,* 2002). These islands are responsible for regulating the expression of the genes since they are known to occur at or near about 40% of the promoters in a mammalian genome and are generally unmethylated (Bird, 2002).

The CpG represent only one third to one fourth of the expected frequencies in a vertebrate genome; the reason behind this can be the higher stacking energy of the nucleotide base Cytosine and Guanine in comparison to Adenine and Thymine bases. Therefore, structural constrain is an important factor that may lead to CpG avoidance in the genome (Deichmann, 2016). The transcription efficiency of the CpG codons are lower since the proportion of tRNA's that contain CpG in their anticodon are lower in comparison to the tRNA's of any other di-nucleotide. Also, the presence of large number of unmethylated CpG in the genome

stimulates innate immune response reactions if not methylated (Upadhyay *et al*., 2014). At last, the methylated Cytosine shows high propensity of undergoing deamination which also accounts for CpG depletion. Cellular DNA repair machinery is capable of correcting the transitions of unmethylated Cytosine to Uracil base but there is no such mechanism present for the correction of transitions of methylated Cytosine to Thymine; making this an irreversible process and therefore accounts for elevated mutational frequency in the methylated genomes (Jones *et al*., 2009). The methylated Cytosine causes loss of two CpG (one from each strand of DNA), adding one TpG and CpA. This results in deficiency of CpG/CpG and overrepresentation of TpG/CpA as shown in figure (Bird, 1980).

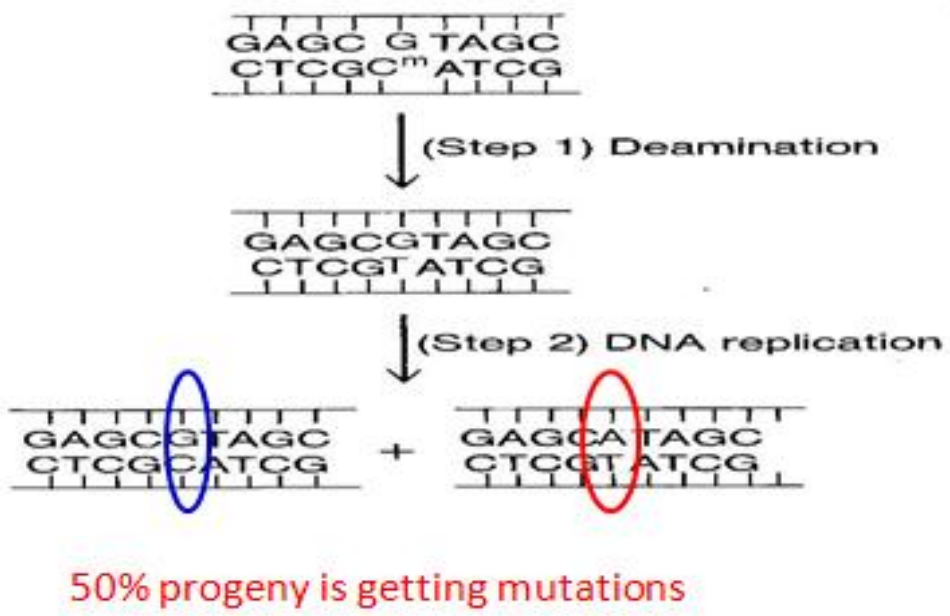


Figure 6: 5-methylcytosine cause loss of two CpG and gain of one TpG and CpA.

Along with CpG, TA is also significantly under-repressed in the eukaryotic and prokaryotic genomes (Karlin *et al*., 1997). The eukaryotic chromosomes show under-representation in the range of 0.61 to 0.81, whereas CpG in vertebrates shows drastic suppression in the range of 0.28 to 0.37 (Karlin *et al*., 1998). The possible reasons underlying the under-representation of TA di-nucleotide is of having the lowest thermodynamic stacking energy among all di-nucleotides; it forms a part of many regulatory signals such as TATA box, transcription terminators and high preference of ribonucleotides for degradation of UA dinucleotides (in mRNA). Thus the suppression of TA confers avoidance for the binding of inappropriate regulatory factors (Karlin *et al*., 1997).

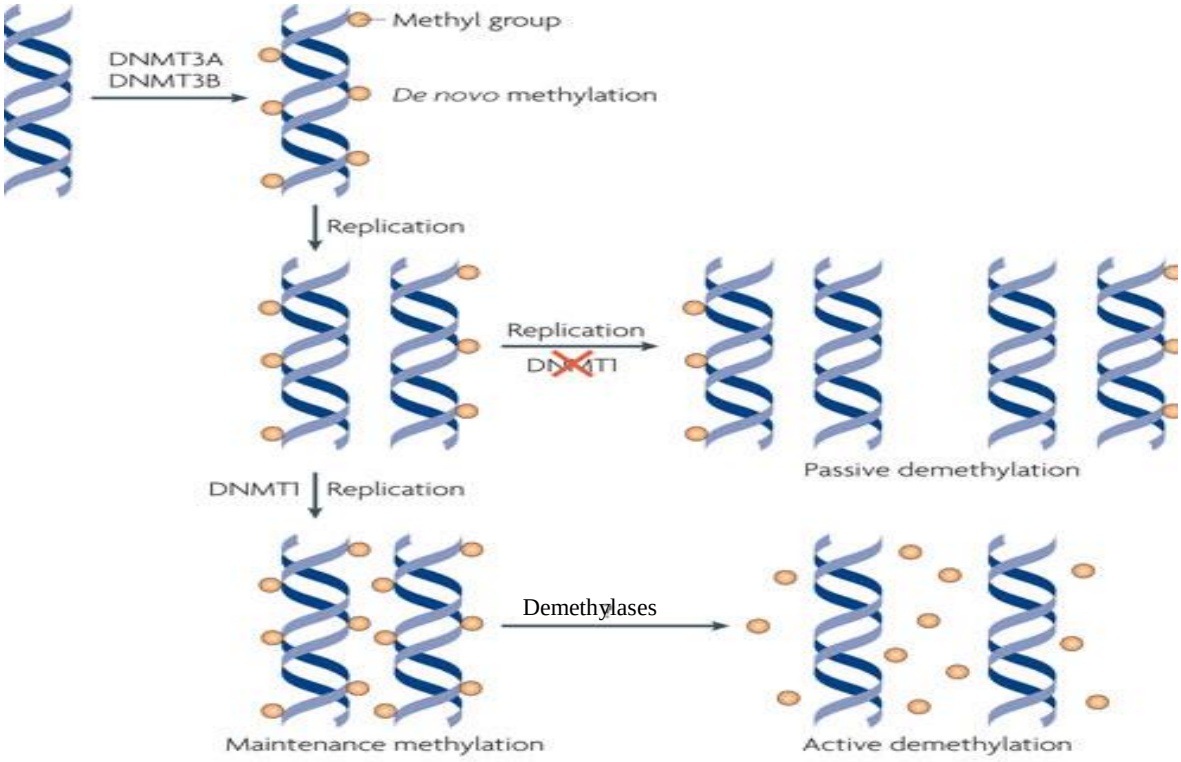
2.2 DNA Demethylation

Earlier DNA methylation was considered to be an irreversible, permanent and unidirectional change, but it is not so. This epigenetic modification is a form of stable and reversible change (Barreto *et al*., 2007). It has been observed that genome wide demethylation of the DNA occurs after fertilisation, wherein the existing patterns of DNA methylation from both maternal and paternal genomes are erased and are then reset in an early embryo, establishing the epigenetic state (Hafner *et al*., 2016). Also during successive rounds of replication, a cell with maintenance methylation loses the patterns of CpG methylation in 50% progeny of its daughter cells. The process of removal or loss of methylation marks is done in two different ways:

2.2.1 Active Demethyaltion

It is an enzymatic process that involves active removal of methyl groups or 5mC by breaking C-C bond. Various proposed mechanisms involved in active demethylation are (Pan *et al.*, 2012):

A. Removal of methyl group directly from 5mC enzymatically
B. Nucleotide replacement reaction
C. DNA glycosylase mediated active demethylation involving modification of 5mC followed by Base Excision Repair (BER) system



Source- Nature Reviews Molecular Cell Biology, 2010

Figure 7: The patterns of Methylation are initially established by De-novo Methyltransferases (DNMT 3a and DNMT 3b) during early development of organism. Even after successive rounds of DNA replication and cell division, these patterns are subsequently maintained by maintenance methyltransferase (DNMT 1) which shows high preference for hemi-methylated DNA. During the process of cell division, if DNMT 1 is either absent or inhibited, the newly synthesized strands of DNA will not inherit the methylation patterns, resulting in Passive Demethylation over successive rounds of replication. In contrast, the process of Active Demethylation occurs when 5-methylcytosine (5mC) is replaced with Cytosine enzymatically.

2.2.2 Passive Demethylation

It involves loss of methyl group from 5mC when DNMT1 is either absent or inhibited in the successive rounds of replication, resulting in lack of DNA methylation in the newly synthesized strands of DNA (Rasmussen *et al.*, 2016).

2.3 Adenovirus

Adenoviruses represent medium sized (26-46 kb long) ds DNA and linear genomic viruses that are capable of replicating in the nucleus of vertebrate cells using host replication machinery. This virus belongs to a family of '*Adenoviridae'* which can be further classified into *Mastadenovirus* (includes all Human Adenoviruses), *Aviadenovirus, Atadenovirus, Siadenovirus* and *Ichtadenovirus* (Shaw *et al.*, 2008). Human Adenovirus cause wide range of illness; the most common being respiratory infections, eye infections, gastroenteritis to life threatening multi organ disease in people with weakened immune response (Gerba *et al*., 2008). For studying the course of host driven evolution in viruses, understanding the relative abundance of dinucleotides as well as the extent of codon usage bias are important parameters that need to be considered. One of the factors that contribute in shaping the viral evolution in context to CpG usage is DNA Methylation.

2.3.1 Role of Methylation in host viral interactions

The DNA viruses infecting animals show significant variations in the nucleotide composition, but the role of evolutionary pressures and biological mechanisms, which are responsible for driving these patterns, remains indistinct. The location of viral replication within the host cell along with intracellular trafficking route specifically involved in the pathway affects the susceptibility of viral genome undergoing methylation and immune recognition within the host system (Hoelzer *et al*., 2008). Among viruses evolutionary studies can be done based on the differences occurring in the relative abundance of dinucleotides. The most extensively studied among all are the CpG whose depletion is reported. The reasons underlying this depletion are linked with evolution of virus, translational selection as well as mutational pressures. Apart from these, virus genome size and the type of genetic material (whether DNA or RNA) are also crucial factors shaping the viral evolution (Upadhyay *et al*., 2015). Earlier, genome of the Herpes virus has been reported to be significantly suppressed in context to CpG and shows excess of CpA/TpG nucleotides relatively (Karlin *et al*., 1994). Methylation also plays important role in HBV gene expression by down regulating it. During chronic viral infection, Dnmt's are up-regulated in host as a mechanism of host defence system. The Hepatocytes of the host response to HBV infection by increased expression of Dnmt's, as a result causing methylation of viral DNA and thus leading to inhibition of viral replication and its expression of genes (Vivekanandan *et al*., 2010).

Evidence which suggests role of methylation in the regulation of viral protein production is provided by the CpG islands of HBV DNA which are methylated in the human tissue (Vivekanandan *et al*., 2009). CpG Methylation at low densities regulates viral DNA, mRNA as well as protein expression, therefore reducing the production of protein encoded by virus. EBV genome shows strong under-representation of CpG in comparison to Herpes Simplex virus which shows relative abundance (Vivekanandan *et al*., 2008). In contrast, Cytomegalovirus shows over representation of CpG dinucleotides. The observed low frequency count in the EBV is due to high probability of 5-methylCytosine undergoing spontaneous deamination during the course of evolution (Burge *et al*., 1992). The suggested hypothesis for this CpG suppression is that the peripheral blood mononuclear cells are able to detect the methylated genome of EBV. In response, the genome is susceptible to mutagenesis by methyl Cytosine deamination which becomes a major contributor in shaping the viral genome over evolutionary time (Ambinder *et al*., 1999).

2.4 Evolution of Viruses

Viruses representing a class of ubiquitous and diverse infectious organisms have DNA or RNA as their genetic elements and require host cell for their replication (Domingo and Perales, 2014). Early studies have shown that *Polyomavirus, Papillomavirus* and *Parovirus* representing group of small DNA viruses are species specific, genetically stable and show relation of being co-evolved with their host species (Shadan and Villarreal, 1995). The vertebrate infecting viral species having small size genomes ($\leq$ 30 kb) are observed to be significantly underrepresented in context to CpG dinucleotides. This suppression prevails irrespective of genomic organization in virus and its morphology. Also, no correlation was found in the GC content of virus with the measure of relative abundance of P^*_{CG} values for dinucleotides. For viruses of larger genomic size having vertebral hosts, suppression of CpG is observed to be inconsistent. Gammaherpesvirus class of virus shows neither over nor under representation of CpG's whereas, Adenoviruses tend to be on the lower side of CpG dinucleotides frequency representation within its genome (Karlin *et al*., 1994). DNA viruses with large genome size emerged from an ancient viral ancestors carrying a smaller subset of 30-35 genes that are involved in encoding the proteins essential for viral structure assembly and replication (Koonin *et al*., 2015).

DNA methylation based CpG suppression in the genomes appears to be and additional factor causing diversity among eukaryotic viruses infecting hosts experiencing extensive genome methylation.

CHAPTER 3

SCOPE OF STUDY

Study Adenoviral genome to see the effect of DNA Methylation. Early studies reveal CpG dinucleotides suppression in vertebrate infecting viruses. In our case we try to use a novel approach by taking advantage of comparative genome analysis of Adenovirus. The genomic sequences of virus isolates infecting humans were selected for studying different kind of mutations that have occurred during the course of evolution along with the host. Such analysis is expected to throw light on CpG loss in Adenovirus.

CHAPTER 4

OBJECTIVES

- Data mining for genomic sequence of various isolates of Adenovirus.
- Comparative genome analysis to study effect of DNA methylation in Adenoviral genome.
- Assessment of effect of CpG mutations on Adenoviral proteome.

CHAPTER 5

Materials and Methods

5.1 Retrieval of Sequences

Complete genomic DNA sequence of the Human Adenoviral strain was searched from National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/). BLAST (Basic Local Alignment Tool) was then run for this sequence. In BLAST, within the algorithm parameters, the maximum target sequences were set to 500 to get maximum number of isolates of Adenovirus that infects humans. A total of 210 sequences were then selected where viral host was human. The sequences of selected strains were then downloaded in fasta format.

Adenovirus was chosen as target organism for studying the effect of DNA Methylation (an epigenetic trait) in the evolution of viral genome. The criterion for selection includes:

- The organism they infect should have a methylated genome.
- The virus should have a dsDNA genome (or dsDNA as its replication intermediate).
- The virus should be known to cause a commonly found disease in humans. Since it will be a human infecting virus, whole genomic sequences of large number of isolates and strains would be available and thus very well studied.

A total of 210 complete genomic sequences of Human Adenoviral strain were used for analysis. Accession number of each genomic isolate of Adenovirus is given in the table below.

S.No.	Accession Number	S.No.	Accession Number	S.No.	Accession Number
1	JO1917.1	71	JN226754.1	141	AB605241.1
2	JX173077.1	72	AP012285.1	142	FJ824826.1
3	KF268130.1	73	JN162671.1	103	AY601636.1
4	KF268310.1	74	JN226760.1	104	JN226759.1
5	JX173084.1	75	JN226758.1	105	HQ910407.1
6	KR699642.1	76	JN226755.1	106	KF268206.1
7	JX173079.1	77	JN935766.1	107	JN226756.1
8	JX173081.1	78	HQ883276.1	108	JN226749.1
9	HQ003817.1	79	AB562587.1	109	JN162672.1
10	KF268129.1	80	KX827426.1	110	KF268315.1

11	FJ349096.1	81	AB695622.1	111	LC066535.1
12	KF951595.1	82	KF268333.1	112	KF268213.1
13	HQ413315.1	83	KF268313.1	113	KF268203.1
14	JX173083.1	84	KF268209.1	114	KF268122.1
15	JX173086.1	85	JN226764.1	115	JN226747.1
16	JX173085.1	86	JN226746.1	116	JQ326209.1
17	JX173078.1	87	AB695621.1	117	JQ326208.1
18	JX173080.1	88	HM770721.2	118	JQ326207.1
19	AF534906.1	89	AB562588.1	119	FJ619037.1
20	JX173082.1	90	KJ626292.2	120	FJ404771.1
21	AY339865.1	91	KJ626291.1	121	EF121005.1
22	KF268127.1	92	KF268335.1	122	AB448776.1
23	AY601635.1	93	KF268329.1	123	AB448775.1
24	KF429754.1	94	KF268204.1	124	AB448774.1
25	M73260.1	95	JN226751.1	125	AB448773,1
26	JX423389.1	96	AB562586.1	126	AB448772.1
27	KF268199.1	97	GQ384080.1	127	DQ900900.1
28	AY487947.1	98	FJ169625.1	128	JN860678.1
29	AY594253.1	99	KF268322.1	129	AB448778.1
30	AY594254.1	100	JN226748.1	130	AB448777.1
31	EF371058.1	101	AB605242.1	131	AY37798.1
32	JN226763.1	102	KF268334.1	132	JN860680.1
33	KF268201.1	103	KF268320.1	133	AF108105.1
34	JN226757.1	104	KF268312.1	134	AB448771.1
35	AB765926.1	105	KF268207.1	135	HQ659699.1
36	JN226752.1	106	KC529648.1	136	GQ478341.1
37	EF153473.1	107	JN226762.1	137	AY594255.1
38	KF268332.1	108	JN226761.1	138	KF268205.1
39	AP012302.1	109	JF99911.1	139	KF006344.1
40	KP641339.1	110	AB605245.1	140	AY599837.1
41	KF268211.1	111	AB605244.1	141	AB333801.2
42	KF268325.1	112	AB605243.1	142	AY599825.1
43	AJ854486.1	113	AB605240.1	143	KF268321.1
44	KF279629.1	114	KF268330.1	144	KF528688.1
45	JN226753.1	115	JN226750.1	145	AY601633.1
46	EF153474.1	116	DQ393829.1	146	KP670856.2
47	KF268327.1	117	AY875648.1	147	KP670855.2
48	K268324.1	118	KF268197.1	148	KP670860.1
49	KF268319.1	119	JN226765.1	149	KP670858.1
50	KF268208.1	120	HQ007053.1	150	KP670857.1
51	KJ364590.1	121	KF577595.1	191	KF577597.1
52	KJ364589.1	122	KF268134.1	192	KF577593.1
53	KJ364587.1	123	JX625134.1	193	KF268316.1
54	KJ364584.1	124	JF800905.1	194	KF268314.1
55	KJ364582.1	125	AY594256.1	195	KF268135.1
56	KJ364581.1	126	KU361344.1	196	KF268117.1
57	KJ36457.1	127	KT963081.1	197	JX423383.1

58	KJ364577.1	128	KJ364592.1	198	KJ364575.1
59	KJ364576.1	129	KJ364591.1	199	KJ364574.1
60	KJ364573.1	130	KJ364588.1	200	KF577598.1
61	KJ019888.1	131	KJ364586.1	201	JN860677.1
62	KJ0198887.1	132	KJ364585.1	202	KF268212.1
63	KJ019886.1	133	LJ364583.1	203	KF268125.1
64	KJ019885.1	134	KJ364580.1	204	JX423387.1
65	KJ019883.1	135	KJ364578.1	205	AY599834.1
66	KJ019882.1	136	KJ019884.1	206	DQ086466.1
67	KJ019881.1	137	KC440171.1	207	KF268210.1
68	KJ019880.1	138	KF938575.1	208	KX423388.1
69	KJ01987.1	139	KF802426.1	209	AY601634.1
70	KC857700.1	140	KF802425.1	210	KF633445.1

Table 4: Accession Number of 210 Adenoviral genomic isolates

5.2 Sequence Analysis Tools

5.2.1 Multiple Sequence Alignment (MSA)

Multiple Sequence Alignment of 210 human infecting Adenoviral strains was carried out. The tool that was used for carrying out alignment of 210 sequences of Adenovirus was "R-Studio" (version 1.0.136.0, 64 bit). The library used under this tool was 'Bioconductor' along with the package 'Decipher' for carrying out alignment.

5.2.2 Microsoft Excel

For carrying out statistical and computational analysis on the obtained sequence data, Microsoft Excel Spreadsheet was used. The various operations along with the functions used were:

Operation Name	Class	Function
IF	Logical	makes logical comparisons and checks whether a condition is met; returns a value depending on whether the test is TRUE or FALSE.
COUNTIF	Statistical	counts the number of cells within a range that meet a specific given criteria
CONCATENATE	Text	allow to combine or join text and values to create a single combined string

Table 5: Microsoft Excel operations

5.2.3 Mega 7

MEGA stands for 'Molecular Evolutionary Genetic Analysis' which is a software suite for analysis of DNA as well as protein sequences from various populations and species. This tool is also used for carrying out molecular evolutionary studies as well as phylogenetic tree construction.

5.2.4 Notepad ++

Notepad ++ is a source code editor which has attractive features such as recording and running macros. Macro recording was employed in manipulation and analysis of large DNA sequences. Manipulation of the sequences involves conversion of sequence lines to a single string continuous string. Macro recording was implicated for execution of simple sequence manipulations which were required to be repeated several times such as joining DNA sequences into one single string.

5.3 Methods

5.3.1 Genomic sequence search for Adenovirus

For Adenovirus, genomic sequences were acquired from NCBI (http://www.ncbi.nlm.nih.gov/) in fasta format. The sequences were then subjected to BLAST to get more genomic isolates of the virus. A total of 210 complete genomic sequences of Adenovirus were available which infects humans. Only human infecting strains of virus were selected.

5.3.2 Analysis of mono and dinucleotide frequencies within individual Adenoviral genome.

Counts of A, T, G and C mono-nucleotides were computed for each selected 210 Adenoviral genome. Dinucleotide combinations were then formed within these genomes by using concatenate function in MS Excel. Counts of CpG, TpG and CpA were then computed in these genomes along with the frequencies of GpC, GpT and ApC having similar base composition but different sequence. Observed and expected frequencies of these dinucleotide bases in each sequence were then calculated.

Probabilities of expected CpGs were calculated as follows:

$$P(CG) = P(G) \text{ X } P(C)$$

Where P(C) = Total number of Cs/Total length of sequence (i.e. G+A+T+C)

P(G) = Total number of Gs/Total length of sequence (i.e. G+A+T+C)

Expected number of CpG in a given sequence:

= P(CG) x Length of Sequence

= P(CG) x (G+A+T+C)

5.3.3 Multiple Sequence Alignment of the genomic sequences

Multiple Sequence Alignment (MSA) of 210 sequences was performed by using R studio programming tool. The library used was 'Bioconductor' and the package used under this library was 'Decipher'. The set of commands that were used in R-studio for alignment were as shown in the picture.

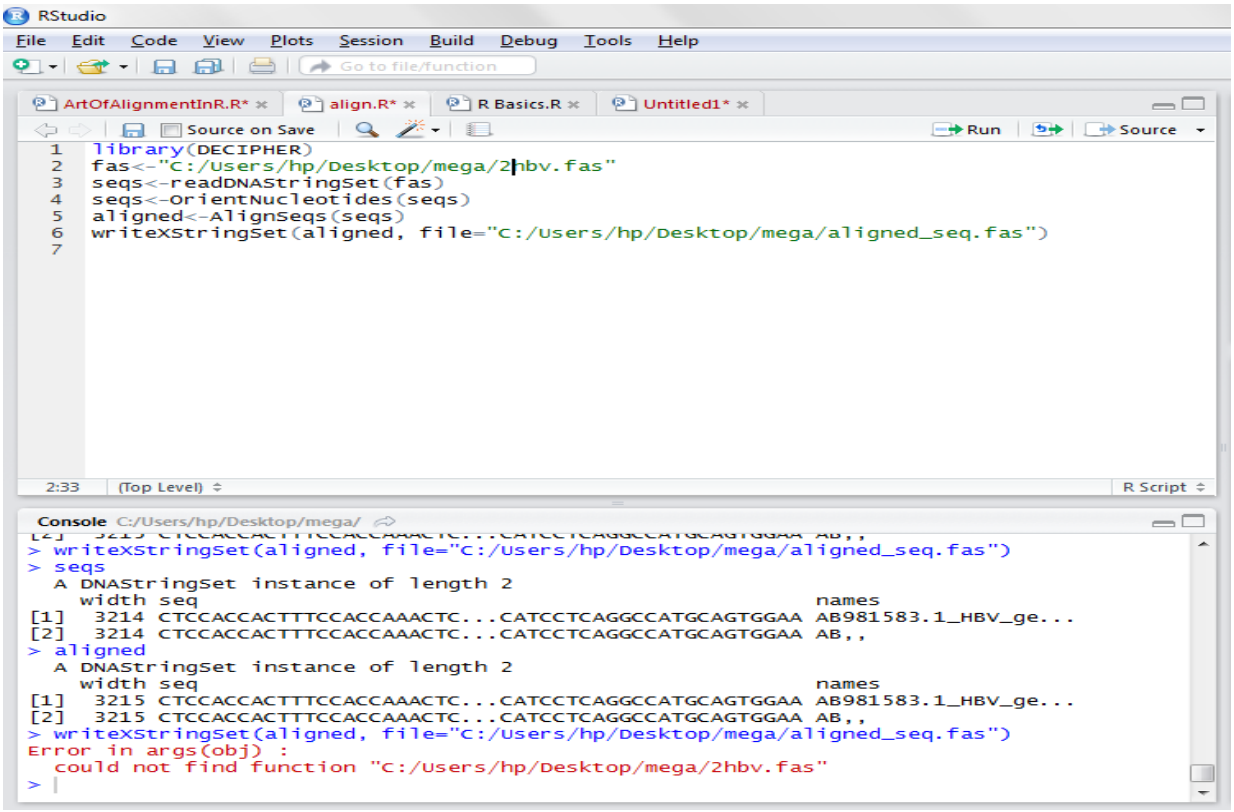


Figure 8: Set of commands used in 'R-studio' for carrying out Multiple Sequence Alignment

Mega 7 software (version 7.0.21_win64) was used to view the output file obtained after alignment which was in .fas format.

5.3.4 Analysis in MS Excel

As further computational and statistical analysis has to be done in MS Excel, the data should be in a format that can be viewed under MS Excel. For this each aligned sequence was copied in Notepad ++ and Macro was run for the conversion of sequence lines to a single string.

Similar procedure was repeated for each 210 sequences and these vertically stringed sequences were then copied to MS Excel worksheet.

Software Used	File Extension Required
R studio	.fas
Notepad ++	.txt
Microsoft Excel	.xlsx/.xls
Mega 7	.meg/.fas

Table 6: File extensions required for different software's

Probabilities of expected TpG+CpA and rest other dinucleotides (which include AG, AT, AA, AC, TA, TT, TC, CC, CT, GA, GT, GC, and GG) were computed as follows:

Expected number of TpG+CpA = (Net total – CG count) x 2/15

Where, - = gaps in aligned sequence

Net total = 210 - (-N, N-) – (--)

-N = frequencies of –A/-T/-G/-C

N- = frequencies of A-/T-/G-/C-

-- = frequencies of double gaps

Factor of 2/15 was taken instead of 2/16 since CpG dinucleotides were not considered.

Expected number of Rest dinucleotides = (Net total – CG count) x 13/15

Where, - = gaps in aligned sequence

Net total = 210 – (-N, N-) – (--)

-N = frequencies of –A/-T/-G/-C

N- = frequencies of A-/T-/G-/C-

-- = frequencies of double gaps

Factor of 13/15 was taken instead of 13/15 since CpG dinucleotides were not considered.

5.3.5 Data Fragmentation

As the genome size of Adenovirus was too large (approx 36 kbp), for easy handling of the aligned data, the file containing the 210 sequences obtained after alignment was fragmented into eight parts. The total length of the sequence obtained after alignment is 40431. Each of the 8 fragmented parts has 210 sequences of 5053 nucleotide length.

5.3.6 Dinucleotide frequency calculation

The number of A/T/G/C/- were counted within the horizontal rows having 210 sequences by using COUNTIF function (MS Excel).

5.3.6.1 Making di-nucleotide combinations

For analyzing the di-nucleotide frequencies, di-nucleotide combinations were made using CONCATENATE function. Each cell was concatenated with the one lying below, giving di-nucleotide combination. This step is repeated for each 210 sequences.

For calculation of frequencies, all 16 possible di-nucleotide combinations were analyzed along with the combinations that include gaps '-'. A total of 25 combinations were made which were:

'A' nucleotide combinations	'T' nucleotide combinations	'G' nucleotide combinations	'C' nucleotide combinations	'-' combinations
AA	TT	GG	CC	--
AT	TA	GA	CA	G-
AG	TG	GT	CT	A-
AC	TC	GC	CG	T-
				C-
				-G
				-A
				-T
				-C

Table 7: Possible dinucleotide combinations

5.3.6.2 Counting dinucleotide frequencies

The function 'COUNTIF' was used while specifying the range of the horizontal cells having di-nucleotide combinations of the viral strains whose frequency has to be calculated. Similarly, the function was also used to calculate the frequencies of all 25 combinations that were stated earlier.

5.3.6.3 Determining positions in multiple sequence alignment having maximum CpG count and calculating TpG and CpA correspondingly at those positions

Positions having maximum CpG were separated from the rest by using 'IF' function. This returns the count of CpG at specific position only when CpG count is maximum (otherwise J) within the range of dinucleotides frequency counts. For TpG and CpA frequency, IF function was used which returns the frequency of TpG and CpA only when maximum CpG count value is present in the previous cell, otherwise J. All the J's were replaced with nothing and

blank cells were deleted, giving maximum CpG count along with TpG and CpA count at respective positions.

5.3.6.4 Applying Chi-Test and obtaining p-value at max CpG positions

Chi-test was applied on the positions that were sorted earlier having maximum CpG count. P-values were obtained to check the level of significance. For this, the parameters that need to be calculated were:

Sum of TpG and CpA frequency

Frequency of rest of the dinucleotides (other than TG, CA, CG, N-,-N,--) → Actual Range

Sum of expected TpG and CpA frequency

Expected frequency of rest of the dinucleotides (other than TG, CA, CG, N-,-N,--) → Expected Range

Calculation of p-values at these positions depicts the significance level of CG getting mutated majorly to TpG and CpA. Observed and expected frequencies of TpG + CpA dinucleotides were compared with the rest 13 dinucleotide combinations in overall genomic sequence as well as at max count CpG position and p-values were computed. This shows that TpG + CpA were significantly over-represented than rest any other dinucleotide.

5.3.6.5 Computing C's and G's single transitions and transversions in overall genome of Adenovirus

Probabilities of C's and G's were calculated at each position of multiple sequence aligned Adenoviral genomes by using binomial distribution function (in MS Excel). Significant positions were selected on the basis of p-values (< 0.05) and frequency of C's or G's at a particular position greater than the average i.e. >58 computed for both C's and G's in the

genomes of Adenovirus. Only single transitions (C→T, G→A) along with single transversions (C→G, C→A, G→T, G→C) were considered for analysis.

5.3.7 Studying the effect of CpG mutations on Adenoviral proteome

For studying the effect of CpG mutations on Adenoviral proteome, the proteins coded by an Adenovirus were mapped on the viral genome. A total of 28 proteins coded by a Human Adenoviral strain (reference AC_J01917.1) were mapped.

5.3.7.1 Translating viral nucleotide genomic sequences to amino acids

For this the mega file having the alignment of all 210 sequences were translated into six reading frames by using various options of Mega 7.

Frame 1: All nucleotide sequences were selected and translated using translate/untranslate option.

Frame 2: The first line having the first nucleotide base of each sequence was deleted which results in shifting of frame. After deletion, the sequences were selected and then translated.

Frame 3: The first two lines having the two nucleotides bases of each sequence were deleted, resulting in shifting of the frame and the sequences were then translated to amino acids.

The next three frames were of reverse complementary strand. These frames were created by using the options:

Frame 4: All the sequences were selected and reverse complemented by using reverse complement option in Mega 7. These sequences were then transcribed into amino acids using translate option.

Frame 5: After deleting the first nucleotide base of each sequence, they were revere complemented and transcribed to amino acids.

Frame 6: By deletion of first two nucleotide bases the sequences were reverse complemented and then transcribed.

All the six reading frames that were created in Mega were then exported to excel sheet.

5.3.7.2 Analysing CpG positions on protein sequences

CpG positions that were selected earlier in the multiple sequence aligned Adenoviral genomes were then mapped on proteins as:

$$\text{Approximate Nucleotide position} = \text{Protein position} \times 3$$

The codons having CpG can be categorized in three classes, namely CGN, NCG or NNCGNN. In the last CpG is split into two successive codons. A sample of 30 CpG position in coding regions of protein 32 kD protein (gene E1a) were analyzed. Also, the counts of all 20 amino acids were computed at these positions to study the effect of methylation on amino acids.

CHAPTER 6

RESULTS

The effect of DNA Methylation on the genome of higher eukaryotes is observed as under-representation of CpG dinucleotides and corresponding overrepresentation of TpG and CpA dinucleotides. Earlier reports regarding analysis of CpG dinucleotides frequencies in the genomes of vertebrates and invertebrates have shown this suppression (Cardon *et al*., 1993). Since viruses infects the vertebral hosts, it has been reported that methylation has also been observed in the genomes of viruses leading to CpG suppression and giving evidence of co-evolution of vertebrate infecting viruses with their hosts in the past (Galvan *et al*., 2011). Many DNA viruses have been analyzed for studying the effects of methylation on the evolution of viral genomes.

6.1 Genome Analysis

In this report, Adenoviral genomes infecting humans have been analyzed for studying the under-representation of CpG's. Data mining for this analysis was done from NCBI from where human Adenoviral strains were selected. A total of 210 human Adenoviral genomic sequences were downloaded in fasta format. Analysis of each Adenoviral genomes was based on frequencies of mono-nucleotides (A, T, G and C) and CpG, TpG and CpA dinucleotides. Additionally GpC, GpT and ApC dinucleotide frequencies were also determined as they were used as controls with same base composition. Expected frequencies of CpG, TpG and CpA were computed as given in methods.

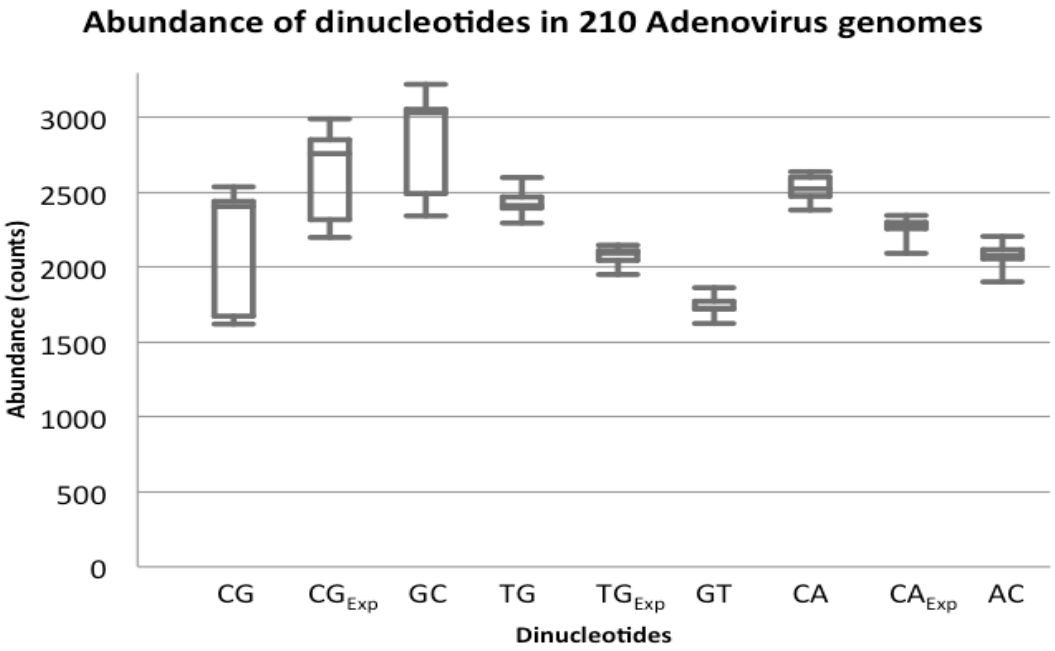


Figure 9: Graph showing comparison of observed frequencies of CpG, TpG and CpA with observed frequencies of GpC, GpT and ApC along with expected frequencies of CpG, TpG and CpA

In all the 210 Adenoviral genomes, CpG frequencies were found to be lower than expected frequency as well as GpC frequencies. This shows CpG underrepresentation based on base composition as well as the control, in all the strains. Similarly all the 210 strains had overrepresentation of TpG and CpA when compared to their respective expected as well as control frequencies. It was inferred that CpG is underrepresented in Adenoviral genomes while TpG and CpA are overrepresented. It indicates that observed variation in abundance of these dinucleotides is contributed at least partially due to CpG/CpG mutation to TpG/CpA. Further it may be extrapolated to methylation of Adenoviral genomes.

Genomic analysis was done by computing base compositions frequencies in all 210 sequences of Adenovirus (Table 8). Along with this, the frequencies of dinucleotides combinations were also computed.

Nucleotide	Observed Frequency	P(N)
G	2017321	0.272
A	1751955	0.236
T	1612091	0.217
C	2032505	0.274
Total	7413872	

Table 8: Total counts of Nucleotides in Adenoviral genome.

The observed frequency counts for all 16 possible dinucleotides in entire Adenoviral genome were also determined as shown in Table 9.

Dinucleotides	Observed Frequencies	Dinucleotides	Observed Frequencies
GG	563282	**TG**	**505541**
GA	476341	TA	274026
GT	363356	TT	397495
GC	596575	TC	414240
AG	480689	**CG**	**447811**
AA	449404	**CA**	**526189**
AT	362472	CT	470030
AC	433649	CC	565273

Table 9: Total counts of Dinucleotides in Adenoviral genome.

Observed and Expected frequencies of CpG, TpG and CpA have been computed as given in methods. In a similar fashion observed and expected frequencies were also computed for GC, GT and AC dinucleotides which were used as controls with identical base compositions and thereby same expected frequency. The observed frequencies of these dinucleotides have been compared as shown in graph (Figure 9).

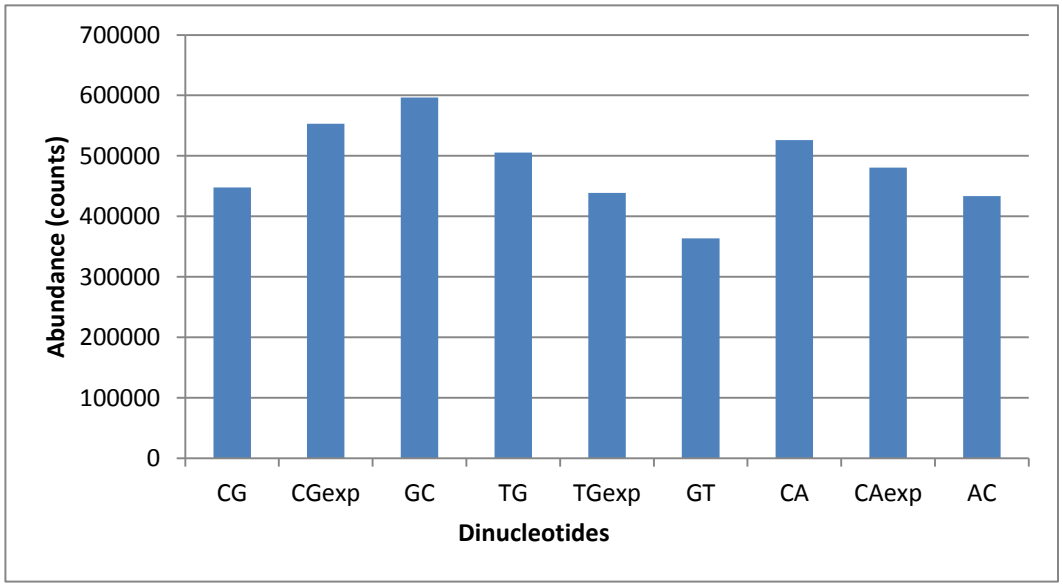


Figure 10: Graph showing comparison of observed frequencies between CpG/GpC, TpG/GpT and TpA/ApT.

The graph shows lower count of CpG dinucleotides in Adenoviral genomes with respect to TpG and CpA dinucleotides. To show that difference between CpG counts and TpG & CpA counts is not relative and it is actually due to loss of CpGs and corresponding gain of TpGs and CpAs, they have been compared with dinucleotide counts having different sequence but

same base composition. CpGs are fewer than GpCs while TpGs and CpAs are more numerous than GpTs and ApCs respectively. The picture becomes clearer when we compare the observed/expected frequency ratios of CpG, TpG and CpA. When the observed and expected frequencies of three sets of dinucleotides (CpG vs. GpC, TpG vs. GpT and CpA vs. ApC) were subjected to Chi-square test, a significant difference between the observed frequencies and their theoretical distributions was observed with a p-value approaching zero as shown in Table 10.

Dinucleotides	Observed Frequencies	Expected Frequencies	O/E Ratios	p-Value
CG	447811	553046.37	0.81	0.00
GC	596575	553046.37	1.08	
TG	505541	438651.36	1.15	0.00
GT	363356	438651.36	0.83	
CA	526189	480296.57	1.10	0.00
AC	433649	480296.57	0.90	

Table 10: p-values and O/E Ratios for dinucleotides

An overall effect of CpG/CpG to TpG/CpA mutations was assessed by comparing TpG and CpA dinucleotide frequencies against rest of all other 14 possible dinucleotides. To study the overall increase of TpG + CpA count in the Adenoviral genome, observed and expected frequencies of these dinucleotides were compared with the observed and expected frequencies of rest 14 dinucleotides. An overall increase of ~1.14 fold is observed for TpG + CpA in comparison to the rest with a p-value approaching zero (Table 11).

	TpG+CpA	Rest 14 Dinucleotides	p-value
Observed Frequencies	1031730	6294643	0.00
Expected Frequencies	918947.93	6389688.7	
RATIO	1.14		

Table 11: Frequency comparison of TpG + CpA with rest 14 dinucleotides.

Similarly, on comparing observed and expected frequencies of CpG dinucleotides with rest 14 dinucleotides, a ~1.22 fold decrease was observed with p- value approaching zero within the overall genome of Adenovirus (Table 12).

	CpG	REST 15 Dinucleotides	p-value
Observed frequencies	447811	6878562	0.00
Expected frequencies	553046.37	6973607.70	
RATIO	0.82		

Table 12: Frequency comparison of CpG dinucleotide with rest 15.

6.2 Multiple Sequence Alignment

Our earlier analysis was based on computing the frequencies of nucleotides in the entire Human Adenoviral genome. These values give an overall view about the suppression of CpG whereas gain of TpG and CpA within the genome. This analysis lacks specificity in study of CpG to TpG and CpA. For example it is not possible to distinguish if a given a dinucleotide sequence has same ancestral alleles or it is resulting from a mutation.

To overcome this problem a different approach based on Multiple Sequence Alignment (MSA) was adopted to study CpG methylation in virus. MSA of the 210 selected human infecting Adenoviral genomic sequences was performed. Figure below depicts 210 aligned Adenoviral genomes which were then further analyzed for studying the effects of methylation.

Figure 11: Alignment of 210 human infecting Adenoviral genome sequences.

Dinucleotides combinations were then created in MS Excel and frequency counts were then computed. Max CpG positions were determined where they were present predominantly in comparison to rest any other dinucleotides species. At these positions along with CpG frequencies, counts of TpG + CpA and rest were also computed to determine the conversion

of CpG/CpG to TpG/CpA and rest any other dinucleotide. Expected values of TpG + CpA and rest dinucleotides were calculated as given in methods.

Comparison for significant difference between observed and expected frequencies of TpG + CpA and rest of the dinucleotides (excluding CpG) was performed based on Chi-square test and P-values were calculated for all the positions in multiple sequence alignment where CpG was the most frequent dinucleotide (Table 13).

S.No.	Alignment position	Max CpG counts	TpG+CpA	Rest (excluding CG,TG,CA,-N,N-,--)	Expected TpG+CpA	Expected rest	p-value
1	92	168	6	5	1.47	9.53	5.79702E-05
2	94	71	19	89	14.40	93.60	0.19287595
3	101	92	87	1	11.73	76.27	3.5961E-123
4	112	107	92	5	12.93	84.07	2.6202E-123
5	119	74	73	59	17.60	114.40	1.13569E-45
6	123	76	7	123	17.33	112.67	0.007674119
7	139	203	5	0	0.67	4.33	1.19193E-08
8	144	131	38	36	9.87	64.13	6.5339E-22
9	147	101	64	44	14.40	93.60	8.84187E-45
10	151	101	0	108	14.40	93.60	4.57782E-05
11	160	100	1	108	14.53	94.47	0.000137147
12	164	174	9	26	4.67	30.33	0.031183049
13	167	105	0	103	13.73	89.27	6.8706E-05
14	169	100	31	77	14.40	93.60	2.61503E-06
15	187	98	3	82	11.33	73.67	0.007837932
16	201	127	46	36	10.93	71.07	4.59429E-30
17	218	210	0	0	0.00	0.00	
18	230	100	83	27	14.67	95.33	7.06126E-82
19	243	205	5	0	0.67	4.33	1.19193E-08
20	249	89	12	109	16.13	104.87	0.268994351
21	262	126	84	0	11.20	72.80	9.3359E-121
22	264	208	2	0	0.27	1.73	0.000311491
23	275	101	0	109	14.53	94.47	4.22123E-05
24	296	203	7	0	0.93	6.07	1.52639E-11
25	323	106	2	101	13.73	89.27	0.000671384
26	325	209	0	0	0.00	0.00	
27	327	154	0	55	7.33	47.67	0.003627349
28	367	148	33	28	8.13	52.87	7.5286E-21
29	374	82	1	126	16.93	110.07	3.1937E-05
30	383	126	79	4	11.07	71.93	1.19E-106
31	395	180	0	29	3.87	25.13	0.034666267
32	402	206	3	0	0.40	2.60	1.006E-05
33	417	209	0	1	0.13	0.87	0.694886602
34	424	209	0	1	0.13	0.87	0.694886602

35	435	183	0	27	3.60	23.40	0.041540072
36	441	101	0	109	14.53	94.47	4.22123E-05
37	454	99	0	83	11.07	71.93	0.000352368
38	456	92	6	112	15.73	102.27	0.008392009
39	464	210	0	0	0.00	0.00	
40	466	210	0	0	0.00	0.00	
41	471	127	1	82	11.07	71.93	0.001151991
42	483	108	0	102	13.60	88.40	7.45266E-05
43	533	150	50	10	8.00	52.00	2.81969E-57
44	540	176	0	34	4.53	29.47	0.022190719
45	568	210	0	0	0.00	0.00	
46	592	206	4	0	0.53	3.47	3.41417E-07
47	597	104	5	101	14.13	91.87	0.009063698
48	635	173	3	34	4.93	32.07	0.349789579
49	676	78	29	103	17.60	114.40	0.003512443
50	715	133	50	27	10.27	66.73	1.76382E-40
51	742	131	79	0	10.53	68.47	1.0983E-113
52	763	108	1	101	13.60	88.40	0.000242478
53	799	108	0	0	0.00	0.00	
54	808	133	77	0	10.27	66.73	7.3991E-111
55	868	86	45	78	16.40	106.60	3.297E-14
56	1004	178	5	27	4.27	27.73	0.70293882

Table 13: Positions within Adenoviral genome having maximum CpG counts with respect to rest

The positions with highest frequency of CpG and significant difference in TpG + CpA and rest of the 13 dinucleotide frequencies were considered to have undergone CpG methylation based mutations from CpG/CpG to TpG/CpA. It implies that such positions had disproportionately higher TpG + CpA as mutation products in comparison with rest of the 13 dinucleotides.

In similar fashion, positions for rest of the 15 dinucleotides were also determined where a dinucleotide was having maximum count in comparison to the rest as well as the counts of positions where there was 100% conservation of a dinucleotide i.e. a position had only a single dinucleotide across all 210 genomes (Table 14).

Dinucleotides	total positions having maximum dinucleotide count	total positions having 210 dinucleotide count	conservation percentage
CG	**2353**	**277**	**11.77**
TG	2415	438	18.14
CA	2444	445	18.21
GC	3054	441	14.44
GT	1639	335	20.44
AC	1970	366	18.58
GG	2828	636	22.49
GA	2314	428	18.50
AG	2258	278	12.31
AA	1874	386	20.60
AT	2892	474	16.39
TA	1055	226	21.42
TT	1636	359	21.94
TC	1909	367	19.22
CT	2208	243	11.01
CC	2892	474	16.39

Table 14: Total positions having maximum dinucleotide counts and 100% conservation.

It has been inferred from above data that CpG dinucleotide was one of the least conserved dinucleotide (showing only ~11% CpGs in the genome exhibited 100% conservation) among rest all other dinucleotides. This shows that CpG dinucleotides were one of the most frequently mutated bases in the genome of Adenovirus during the course of evolution when compared to the rest.

6.3 Analysis at max CpG count positions

Analysis of maximum CG positions was done. Since the modification involving methylation of CpG/CpG results in formation of TpG and CpA dinucleotides, their frequencies were also computed at respective positions of CpG max. TpG + CpA counts were then compared with rest other dinucleotides to check over-representation of these dinucleotides within the viral genome at specific CpG max sites that will be majorly due to the effect of methylation at these sites. The observed and expected frequencies were computed for TpG + CpA dinucleotides in comparison to rest other 13 dinucleotides only at CpG max positions (Table 15).

	TpG+CpA dinucleotide	Rest 13 dinucleotides	p-value
Observed Frequencies	47075	99573	0.00
Expected Frequencies	65319.6	424577.4	
RATIO	3.07		

Table 15: Frequency comparison of TpG+CpA with rest 13 dinucleotides at CpG max positions.

On comparison of TpG + CpA frequencies with rest other dinucleotides, it has been observed that there was a ~3.073 fold abundance of TpG + CpA dinucleotides in comparison to rest. As TpG + CpA were the products of methylated CpG sites, therefore their over-representation was determined to be the result of methylation phenomena, whereas the rest were due to any other mutations. There was a significant increase of TpG + CpA frequency from ~0.88 to ~0.33 folds at overall genomic and at specific CpG max sites respectively.

This increase could also be because of biased base composition within the genome of Adenovirus. Therefore to eliminate this factor, an analysis of TpG + CpA dinucleotide frequency was done in comparison to the rest while considering base composition (Table 16).

	TG+CA dinucleotide	Rest 13 Dinucleotides	p-value
Observed Frequencies	47075	99573	0.00
Expected Frequencies	18176.99	117531.63	
RATIO	3.06		

Table 16: Frequency comparison of TpG+CpA with rest 13 dinucleotides at CpG max positions taking base composition into account.

A fold value of ~0.326 was computed on comparing TpG + CpA frequencies with rest other dinucleotides while taking base composition into account. It has been observed that even after considering base composition, only marginal difference of 0.001 was observed in the fold ratio of TpG + CpA frequency at CpG max positions. This supports our result inferring that majority of TpG and CpA frequencies were arising as a result of methylation at CpG sites. However the mutations in CpG dinucleotides may mutate to rest of 15 possible dinucleotides because of single transition mutations, single transversion mutations, double transition mutations, double transversion mutations or double mixed mutations.

Type of Mutation (in context to CpG)	Dinucleotides
Single Transitions	TG, CA
Single Transversions	GG, AG, CT, CC
Double Transitions	TA, TC
Double Transversions	GT, GC, AT, AC
Double Mixed Mutations	GA, AA, AT

Table 17: Mutations of Dinucleotides in context to CpG.

It is known that transitions occur more often than transitions. In order to take these factors into consideration, analysis was performed again based on total number of transitions and transversions of C and G. On the other hand only those mutation products were taken into consideration which result out of single mutations (TpG, CpA, CpC, CpT, ApG & GpG). A 2 x 2 contingency table was obtained for this data as following:

	At CpG sites	In rest of the Adenoviral genomes
Transitions (C→T) & (G→A)	47075	444918
Transversions (C→G), (C→A), (G→T) & (G→C)	27179	669312

Table 18: Transitions and Transversions at CpG sites and in overall genome of Adenovirus.

Both the possible transitions at CpG sites result in TpG and CpA which may be caused by mutation as a result of CpG methylation or some other mechanisms while all the possible transversions giving rise to CpC, CpT, GpG and ApG cannot be the result of CpG methylation lead mutation. The Odds ratio of getting TpG and CpA against rest of the four dinucleotides at CpG sites in comparison with rest of the result is 2.6056. This odds ratio is statistically significant with a p-value of <0.0001.

In this analysis the factor of higher propensity of transition mutations has been taken into consideration and yet a >2.6 fold higher abundance of TpG and CpA has been found against rest of the four dinucleotides. This disproportionate abundance of TpG and CpAs at CpG sites strongly indicates involvement of CpG methylation as a major cause of these mutations.

Thus it may be inferred from this data that Adenoviral genome has been getting methylated during the course of evolution whenever it came in contact with the host cell system.

6.4 Assessment of effect of CpG mutations on Adenoviral proteome.

The CpG positions that were determined earlier in multiple sequence alignment of 210 Adenoviral genomes were mapped on protein sequences of Adenovirus in order to investigate the effect of their mutations to TpG and CpA. A CpG can be part of codons in the coding regions and when mutated to TpG or CpA as a result of CpG methylation, can cause synonymous, miss-sense or nonsense mutations.

The codons having CpG can be categorized in three classes, namely CGN, NCG or NNCGNN. In the last CpG is split into two successive codons. A sample of 30 CpG position in coding regions of 32 kD protein (gene E1a) were analyzed. It was observed that CpGs were present as NNCGNN i.e. split CG more frequently than as CGN or NCG at these positions. Assuming equal probability of a CpG to be part of the above mentioned three classes of codon, a Chi-square test was performed to test if observed frequency of the three classes conforms to the expected frequency. The observed frequencies were found to be significantly different from theoretically determined frequency with a p-value of 0.006.

	CGN	NCG	NNCGNN
Observed positions	8	4	18
Expected positions	10	10	10
p-value	0.006		

Table 19: Relative abundance of codons involving CpG dinucleotides.

Based on binomial distribution of NNCGNN was found to be overrepresented when compared to CGN and NCG. This observation suggests a role of evolutionary pressure in minimising the effect of CpG mutations on protein sequences.

6.4.1 Amino Acid counts at mapped CpG positions in 32 kD protein (gene E1a).

Counts of all the different amino acids present at the mapped CpG positions were determined to study the effects of methylation on proteins where CpG was present as NCG or CGN. The analysis of these positions having CGN and NCG codon is shown in Table below.

Alignment Position	CGN or NCG	Single Transitions	Single Transversions	Double Mutations
635	CGN	3 x H	-	27 x I
1004	CGN	4 x H 1 x Q	1 x Q (may arise from single transition or single transversion)	27 x T
1166	CGN	-	-	-
1599	CGN	-	-	-
1677	CGN	-	74 x L	1x I
1923	CGN	-	-	2 x K
1926	CGN	8 x Q	8 x Q (may arise from single transition or single transversion) 74 x L	27 x N
1955	CGN	-	-	34 x K
1522	NCG	7 x L	-	27 x N
1543	NCG	-	-	62 x G 2 x C 3 x N 34 x D
1547	NCG	27 x V	-	1 x N 33 x D 67 x E
1549	NCG	62 x V	-	-
Total		112	157	347

Table 20: Positions having CGN or NCG codons and coded amino acids

Double mutations were predominantly observed more frequently than single mutations at mapped positions of proteins having CGN and NCG codons involved in coding amino acids. Based on binomial distribution of CGN and NCG, it was found that single transitions were favoured over single transversions within the codons involving CpG dinucleotides.

Also double mutations were present predominantly at the analyzed positions of protein; a possible reason lying behind this may be the existence of CpG dinucleotides in the form of NNCGNN, giving amino acids bases that were a result of double mutations. During the course of evolution, NNCGNN were favoured over CGN or NCG combinations giving rise to double mutations in proteins while showing lower values for mutations involving single transitions and transversions.

CHAPTER 7

DISCUSSION

Heterogeneity in relation to relative dinucleotide abundance within the genomes of vertebrate and invertebrates has been observed. Since vertebrates serves as host for many DNA viruses, the nucleotide composition is also found to vary in the viral genomes, but evolutionary pressure and biological mechanisms lashing these patterns are indistinct. Earlier studies have reported suppression of CpG dinucleotides in vertebrates as well as in all genomic sequences of animal mitochondria (Burge *et al*., 1991). Genomic DNA viruses infecting vertebrates have been reported to undergo epigenetic modification at 5-Cytosine, leading to methylation of CpG dinucleotide. Viral genomes are considered to have loss of CpGs similarly to their host, due to deamination of methylated Cytosine. In DNA, presence of one methyl Cytosine would cause loss of two CpGs and add one TpG and one CpA, which is considered to be a prime reason of CpG deficiency and TpG/CpA excess.

Under-representation of CpG dinucleotides with relative abundance of TpG and CpA has been well studied in viral genomes. Short oligonucleotide extremes of most bacteriophage sequences have shown underrepresentation of TpA while overrepresentation of GpC dinucleotides. Small DNA viruses including Hepatitis B virus and class of papovaviruses have shown underrepresentation of CpG frequencies within their genome. Intermediate and large genome sized viruses have shown normal range of CpG relative abundance. An exception of viruses having large genome size but still show CpG suppression is the family of Gammaherpesviruses (including EBV and Bovine herpes virus) which have potently suppressed CpG dinucleotide and relatively higher abundance of TpG/CpA dinucleotides. Retrotransposons infecting eukaryotes having genome size within the range of ≥5 kbp often have lower CpG dinucleotide abundance (Karlin *et al*., 1994). There are some other possible reasons underlying suppression of CpG dinucleotides within a genome of organism. The CpG represent only one third to one fourth of the expected frequencies in a vertebrate genome; the reason behind this can be the higher stacking energy of the nucleotide base Cytosine and Guanine in comparison to Adenine and Thymine bases. Therefore, structural constrain is an important factor that may lead to CpG avoidance in the genome (Deichmann, 2016). The transcription efficiency of the CpG codons are lower since the proportion of tRNA's that

contain CpG in their anticodon are lower in comparison to the tRNA's of any other di-nucleotide. Also, the presence of large number of unmethylated CpG in the genome stimulates innate immune response reactions if not methylated (Upadhyay *et al*., 2014).

For viruses belonging to same family and having similar genome organization and life cycle, the relative abundance of CpG dinucleotides is dependent on the infected host cells. Infecting viral genomes shows strong correlation between the evolutionary lineage of the infected host and the extent of CpG dinucleotide reduction. Since viruses have short reproductive life cycle with higher rates of evolution, observable changes are expected in their genome over relatively shorter time periods (Upadhyay *et al*., 2013).

We attempted to study Human Adenoviral genomic strains having DNA genome, size ranging between 26 to 46kbp and causing wide range of respiratory infections among humans. Adenovirus is chosen as a target organism for this study since it is a double stranded virus which infects humans i.e. an organism with methylated genome. Additionally its genome size is much larger that the viruses used in the earlier similar studies providing an advantage of larger data size. Genomic sequences are selected to study the effect of methylation on the viral genome evolution. The sequences of viral isolates are analyzed and compared by Multiple Sequence Alignment tool, a novel approach which gives clue about the ancestral allele and the mutations which have occurred in the parent allele during the course of evolution.

Previous analysis of CpG dinucleotides in the genomes of Papillomaviruses and polyomaviruses have been reported to be underrepresented. The extent of suppression within these small double stranded DNA viruses is determined by the evolutionary lineage of the host in which virus is causing infection. Phenomena of methylation have been reported to be the primary cause of suppression within these DNA viruses. Also, it has been demonstrated that the depletion of CpG dinucleotides is more pronounced in human infecting strains and other mammals in comparison to those strains in which birds are the hosts for infecting viral species (Upadyay et al., 2015).

The study plan involving analysis of Adenoviral genome for methylation was based on a logic that these viruses having double stranded DNA genomes infects humans and are expected to undergo methylation and hence ultimately resulting in modification of CpG to TpG/CpA. A total of 210 genomic isolates of human Adenovirus are considered for carrying out methylation studies. Our analysis of selected Adenoviral sequences individually shows

significant suppression of CpG observed frequencies in comparison to observed frequencies of GpC and with those of TpG and CpA expected frequencies. Overall analysis of these genomes also showed significant under-representation of TpG + CpA frequency in comparison to rest other dinucleotides. These sequences are then subjected to Multiple Sequence Alignment tool, for analyzing the dinucleotide frequencies and their substitutions within the viral genome.

Since mutations of CpG dinucleotides due to methylation produces modified TpG/CpA dinucleotide bases, there frequencies are analyzed in the entire genome as well as at the positions where CpG counts are maximum when compared to the rest. It has been observed that 2353 positions in multiple sequence alignment are occupied by CpGs, where they are present predominantly in comparison to rest of the other dinucleotides and exhibit great extent to mutations.

Analysis of CpG and TpG+CpA dinucleotide frequency counts profoundly shows underrepresentation and overrepresentation of these dinucleotides at both CpG maximum count positions as well as in the overall genome of Adenovirus. This data is strongly favoured by calculating observed/Expected (O/E) Ratios. The deviation of observed frequencies and the expected frequencies were found to be statistically significant with a p-value approaching zero. A study of single transitions and single transversions involving overall C's and G's in the Adenoviral genome in comparison to single transitions and single transversions of CpG dinucleotides at CpG positions was performed. Single transitions of C's and G's (C→T & G→ A) will occur more often than single transversions of C's and G's (C→G, C→A, G→T, G→ C). Single transition mutations of CpG dinucleotide at CpG sites will result to TpG & CpA occurring due to methylation or any other mutation, whereas single transversions giving rise to CpC, CpT, GpG and ApG were not at all a result of methylation at CpG site. Odds ratio of getting TpG & CpA in comparison to rest four was higher by ~2.6 fold with a significant p-value approaching zero. It is also observed that TpG and CpA have relatively higher propensity with higher abundance in comparison to rest four dinucleotides. This data strongly supports that TpG & CpA are present in abundance at CpG sites indicating CpG methylation will be a major force shaping the Adenoviral genome during the course of evolution.

On analysing the effect of CpG methylation on Adenoviral proteome, it was observed that the existence of CpG involving codon in the form of NNCGNN i.e. split into two successive

codons was favoured in comparison to CGN or NCG forms of codons. This data is supported with a low p-value calculated by using chi-square test. While analysing the frequency of amino acids at CpG positions within a protein, it was observed that the amino acid counts resulting from double mutations were relatively higher than those which were a result of single transitions and transversions.

A disproportionate distribution of these mutations may be explained by the existence of CpG dinucleotide in split form while forming a codon in protein. Since NNCGNN gives rise to higher number of mutant amino acids in comparison to CGN or NCG, double mutations were observed in higher counts than single transitions and transversions.

So from the above result we say that during the course of evolution in Adenoviral genome, CpG are underrepresented and TpG and CpA overrepresentation due to the phenomena of methylation which results in conversion of CpG/CpG dinucleotide to TpG/CpA along withthe presence of CpG dinucleotide in the form of NNCGNN while forming a part of codon and predominantly resulting in double mutations.

CHAPTER 8

CONCLUSION

The genomes of the vertebrates are susceptible to methylation at the CpG dinucleotide sites, where a methylated Cytosine gives rise to TpG and CpA dinucleotide. Mutation of the CpG dinucleotide bases to TpG and CpA has lead to its suppression in higher eukaryotic organisms. Since these vertebrates serves as host for various viruses, it is expected that viruses have coevolved with their host and thus shows similar levels of CpG suppression within their genomes.

Current study is to investigate the role of methylation in evolution of Adenoviruses. On studying individual Adenoviral genome, CpG suppression is observed on comparing frequency of observed CpG with those of observed GpC as well as with expected CpG. Similarly relative abundance of TpG & CpA is observed on comparing observed frequencies of these dinucleotides with observed frequencies of GpT & ApC along with those of expected TpG & CpA in overall Adenoviral genome. TpG+CpA dinucleotides show abundance whereas CpG dinucleotides are deprived in comparison to rest 14 and 15 dinucleotides within the overall genome of Adenovirus respectively.

In this study we have used a novel approach by performing Multiple Sequence Alignment of 210 sequences of Adenoviral strains infecting humans. Following this approach leads to identification of CpG positions in the genomes that are present predominantly with respect to rest others. Analysis of aligned adenoviral sequences revealed that CpG is one of the dinucleotide that shows lowest percentage of conservation in the genome of Adenovirus.

For all CpG sites having maximum count, TpG+CpA dinucleotides frequencies are compared with rest other dinucleotides and three fold abundance is observed. It may be concluded that like the host, Adenoviral genomes also exhibit CpG suppression which is largely due to mutation of CpG/CpG to TpG/CpA arising from methylation of the CpG. Therefore, CpG sites are underrepresented due to accumulation of these mutations during the course of evolution. Comparison of single transitions and transversions of overall C's and G's with those of CpG dinucleotides at CpG sites showed that abundance of TpG and CpA dinucleotides which are a result of methylation are higher in comparison to rest with

significant p-value. Also, double mutated amino acid counts are higher than single transitions and transversions resulting amino acids, therefore favouring NNCGNN form of codons in comparison to CGN or NCG in a codon formation for a protein.

CHAPTER 9

REFERENCES

- Adrian Bird. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev*, 16(1): 6–21.
- Alan R. Shaw and Mark B. Feinberg. (2008). 92 - Vaccines, In *Clinical Immunology* (Third Edition), Pages 1353-1382, ISBN 9780323044042.
- Andrew J. Davison, Mária Benkő, Balázs Harrach. (2003). Genetic content and evolution of adenoviruses. *Journal of General Virology*, 84: 2895-2908.
- Anna Portela, Manel Esteller. (2010). Epigenetic modifications and human disease. *Nature Biotechnology*, 28(10): 1057–1068.
- Ambinder, R. F., Robertson, K. D., & Tao, Q. (1999). DNA methylation and the Epstein–Barr virus. In *Seminars in cancer biology*, *9*(5), 369-375.
- Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic acids research*, *8*(7), 1499-1504.
- Burge, C., Campbell, A. M., & Karlin, S. (1992). Over-and under-representation of short oligonucleotides in DNA sequences. *Proceedings of the National Academy of Sciences*, *89*(4), 1358-1362.
- Cardon, L. R., Burge, C., Clayton, D. A., & Karlin, S. (1994). Pervasive CpG suppression in animal mitochondrial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 91(9), 3799–3803.
- Cooper, D. N., & Youssoufian, H. (1988). The CpG dinucleotide and human genetic disease. *Human genetics*, 78(2), 151-155.
- C.P. Gerba and R.A. Rodriguez. (2008). Adenoviruses. In *International Encyclopedia of Public Health*, Academic Press, Oxford, Pages 28-33, ISBN 9780123739605.
- Crider, K. S., Yang, T. P., Berry, R. J., & Bailey, L. B. (2012). Folate and DNA Methylation: A Review of Molecular Mechanisms and the Evidence for Folate's Role. *Advances in Nutrition*, *3*(1), 21–38.
- Daiya Takai and Peter A. Jones. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. PNAS, 99 (6) 3740-3745.
- Daniel Stone, Anne Furthmann, Volker Sandig, André Lieber. (2003). The complete nucleotide sequence, genome organization, and origin of human adenovirus type 11. *Virology*, Volume 309, Issue 1, 2003, Pages 152-165, ISSN 0042-6822.

- Doerfler W. (1996). Adenoviruses. *Medical Microbiology*, 4th edition, Galveston (TX): University of Texas Medical Branch at Galveston; Chapter 67, PubMed.
- Domingo, E., & Perales, C. (2014). Virus evolution. *eLS*.
- Feinberg, A. P., & Tycko, B. (2004). The history of cancer epigenetics. *Nature Reviews Cancer*, *4*(2), 143-153.
- Flynn, J., Azzam, R., & Reich, N. (1998). DNA binding discrimination of the murine DNA cytosine-C 5 methyltransferase. *Journal of molecular biology*, *279*(1), 101-116.
- Galván, S. C., Martínez-Salazar, M., Galván, V. M., Méndez, R., Díaz-Contreras, G. T., Alvarado-Hermida, Moisés- Alcántara-Silva, Rogelio García-Carrancá, A. (2011). Analysis of CpG methylation sites and CGI among human papillomavirus DNA genomes. *BMC Genomics*. 12:580.
- Gerda Egger, Gangning Liang, Ana Aparicio, Peter A. Jones. (2004). Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429(6990): 457–463. doi: 10.1038/nature02625.
- Goldberg, A. D., Allis, C. D., & Bernstein, E. (2007). Epigenetics: a landscape takes shape. *Cell*, *128*(4), 635-638.
- Guillermo Barreto, Andrea Schäfer, Joachim Marhold, Dirk Stach, Suresh K Swaminathan, Vikas Handa, Gabi Döderlein, Nicole Maltry, Wei Wu, Frank Lyko, Christof Niehrs. (2007). Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation. *Nature*, 10.1038/nature05515.
- Handy, D. E., Castro, R., & Loscalzo, J. (2011). Epigenetic modifications basic mechanisms and role in cardiovascular disease. *Circulation*, *123*(19), 2145-2156.
- Hafner SJ, Lund AH. (2016). Great expectations e Epigenetics and the meandering path from bench to bedside. *Biomedical Journal*, Volume 39, Issue 3, 2016, Pages 166-176, ISSN 2319-4170.
- Hermann, A., Gowher, H., & Jeltsch, A. (2004). Biochemistry and biology of mammalian DNA methyltransferases. *Cellular and Molecular Life Sciences CMLS*, *61*(19-20), 2571-2587.
- Hoelzer, K., Shackelton, L. A., & Parrish, C. R. (2008). Presence and role of cytosine methylation in DNA viruses of animals. *Nucleic acids research*, *36*(9), 2825-2837.
- Howard Cedar, Yehudit Bergman. (2012). Programming of DNA methylation patterns. *Annual Review of Biochemistry*, 81: 97–117.

- Jones, P. A., & Liang, G. (2009). Rethinking how DNA Methylation Patterns are Maintained. *Nature Reviews. Genetics*, *10*(11), 805–811.
- Karlin, S., Doerfler, W., & Cardon, L. R. (1994). Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses?. *Journal of virology*, *68*(5), 2889-2897.
- Karlin, S., Mrazek, J., & Campbell, A. M. (1997). Compositional biases of bacterial genomes and evolutionary implications. *Journal of bacteriology*, *179*(12), 3899-3913.
- Karlin, S., Campbell, A. M., & Mrazek, J. (1998). Comparative DNA analysis across diverse genomes. *Annual review of genetics*, *32*(1), 185-225.
- Koonin, E. V., Dolja, V. V., & Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology*, *479*, 2-25.
- Jin, B., Li, Y., & Robertson, K. D. (2011). DNA Methylation Superior or Subordinate in the Epigenetic Hierarchy? *Genes & cancer*, *2*(6), 607-617.
- Lin, I. G., Han, L., Taghva, A., O'Brien, L. E., & Hsieh, C. L. (2002). Murine de novo methyltransferase Dnmt3a demonstrates strand asymmetry and site preference in the methylation of DNA in vitro. *Molecular and cellular biology*, *22*(3), 704-723
- Li, E. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Reviews Genetics*, *3*(9), 662-673.
- Ma, X., Wang, Y. W., Zhang, M. Q., & Gazdar, A. F. (2013). DNA methylation data analysis and its application to cancer research. *Epigenomics*, 5(3), 301-316.
- Mary Grace Goll, Finn Kirpekar, Keith A. Maggert, Jeffrey A. Yoder, Chih-Lin Hsieh, Xiaoyu Zhang, Kent G. Golic, Steven E. Jacobsen, Timothy H. Bestor. (2006). Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science*, 311(5759): 395–398.
- Meehan, R. R. (2003, February). DNA methylation in animal development. In *Seminars in cell & developmental biology*, 14(1) , 53-65.
- Phillips, T. (2008). The role of methylation in gene expression. *Nature Education*, 1(1), 116.
- Pradhan, S., Bacolla, A., Wells, R. D., & Roberts, R. J. (1999). Recombinant human DNA (cytosine-5) methyltransferase I. Expression, purification, and comparison of de novo and maintenance methylation. *Journal of Biological Chemistry*, *274*(46), 33002-33010.

- Rasmussen, K. D., & Helin, K. (2016). Role of TET enzymes in DNA methylation, development, and cancer. *Genes & Development*, *30*(7), 733–750.
- Robertson, K. D. (2005). DNA methylation and human disease. *Nature Reviews Genetics*, *6*(8), 597-610.
- Saha, B., Wong, C. M., & Parks, R. J. (2014). The Adenovirus Genome Contributes to the Structural Stability of the Virion. *Viruses*, *6*(9), 3563–3583.
- Shadan, F. F., & Villarreal, L. P. (1995). The evolution of small DNA viruses of eukaryotes: past and present considerations. *Virus Genes*, *11*(2-3), 239-257.
- Strahl, B. D., & Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, *403*(6765), 41-45.
- Ratel, D., Ravanat, J. L., Charles, M. P., Platet, N., Breuillaud, L., Lunardi, J., & Wion, D. (2006). Undetectable levels of N6-methyl adenine in mouse DNA:Cloning and analysis of PRED28, a gene coding for a putative mammalian DNA adenine methyltransferase. *FEBS letters*, *580*(13), 3179-3184.
- Upadhyay M, Sharma N, Vivekanandan P. (2014). Systematic CpT (ApG) Depletion and CpG Excess Are Unique Genomic Signatures of Large DNA Viruses Infecting Invertebrates. *PLoS ONE*, 9(11): e111793.
- Upadhyay M, Vivekanandan P. (2015). Depletion of CpG Dinucleotides in Papillomaviruses and Polyomaviruses: A Role for Divergent Evolutionary Pressures. *PLoS ONE*, 10(11): e0142368.
- Ute Deichmann. (2016). Epigenetics: The origins and evolution of a fashionable topic, *Developmental Biology*, Volume 416, Issue 1, Pages 249-254, ISSN 0012-1606.
- Vivekanandan, P., Daniel, H. D. J., Kannangai, R., Martinez-Murillo, F., & Torbenson, M. (2010). Hepatitis B virus replication induces methylation of both host and viral DNA. *Journal of virology*, *84*(9), 4321-4329.
- Vivekanandan, P., Kannangai, R., Ray, S. C., Thomas, D. L., & Torbenson, M. (2008). Comprehensive genetic and epigenetic analysis of occult hepatitis B from liver tissue samples. *Clinical infectious diseases*, *46*(8), 1227-1236.
- Vivekanandan, P., Thomas, D., & Torbenson, M. (2009). Methylation regulates hepatitis B viral protein expression. *Journal of Infectious Diseases*, *199*(9), 1286-1291.

- Wu, T. P., Wang, T., Seetin, M. G., Lai, Y., Zhu, S., Lin, K., & Tackett, A. (2016). DNA methylation on N6-adenine in mammalian embryonic stem cells. *Nature*, 532, 329–333.
- Watters E. 2006. DNA is Not Destiny. *Discover*, (27); 1-8.
- Walter Doerfler (2008) In pursuit of the first recognized epigenetic signal—DNA methylation: A 1976 to 2008 synopsis. *Epigenetics*, 3:3, 125-133.
- Xin Pan, Roger Smith and Tamas Zakar (2012). DNA Methylation in Development, Embryology - Updates and Highlights on Classic Topics, *InTech*, 10.5772/37696.
- Zucker, K. E., Riggs, A. D., & Smith, S. S. (1985). Purification of human DNA (cytosine-5-) -methyltransferase. *Journal of cellular biochemistry*, *29*(4), 337-349.

a

by A A

FILE	PLAG_CHECK.DOCX (964.38K)		
TIME SUBMITTED	17-JUL-2017 06:24AM	WORD COUNT	11066
SUBMISSION ID	831222950	CHARACTER COUNT	60306

CHAPTER 1

INTRODUCTION

The term 'Epigenetics' given by the biologist Conrad Waddington in 1940 described it as "the gene interaction with their environment which bring the phenotype into being". Epigenetics is primarily focused on studying the heritable changes rather than the mutations in the primary sequence of DNA. Since no mutations are introduced in the DNA sequence, the genotype remains intact but variations occur in gene expression and activity, causing changes in the phenotype (Goldberg *et al.*, 2007). The two primary processes that are responsible for introducing epigenetic modifications to the genome are DNA Methylation and Histone Modifications.

DNA Methylation is a modification involving addition of methyl group ($-CH_3$) at N^4 and C^5 position of Cytosine and N^6 position of Adenine in prokaryotes, whereas at C^5 position of Cytosine in eukaryotes (Hoelzer *et al.*, 2008).

Histone Modifications is a process involving winding of DNA around the protein called 'Histone' and is then further compacted into chromosomes. Histone Modifications occur due to alteration in the extent to which a DNA is wrapped around the histone proteins, thereby affecting the availability of genes for activation (Handy *et al.*, 2011). These modifications occur in response to the binding of epigenetic factors (carried as chemical tags) to the histone tails, affecting the DNA wrapping and gene activation (Strahl and Aliis, 2000).

1.1 DNA Methylation

DNA Methylation is a primary process that results in the epigenetic modification of the genomes. It is a mechanism through which sustainable transmission of epigenetic information is done through multiple cycles of DNA replication and cell division (Hermann *et al.*, 2004). The methylation of DNA in mammals plays a crucial role in the development of embryo, genomic imprinting, inactivation of X chromosome, regulating the structure of chromatin, silencing of transposons and endogenous retroviruses, genetic diseases and cancer biology (Bird, 2002, Li 2002 & Feinberg and Tycko, 2004). This process involves a covalent modification by addition of a methyl group to the DNA strand in such a way that the Watson-

Crick base pairing capacity is not interfered. In prokaryotes, DNA Methylation occurs at N4 and C5 position of Cytosine residues and N6 position of Adenine residues, whereas in eukaryotes it occurs at C5 position of Cytosine, mainly in context of CpG dinucleotides (Bird, 1980).

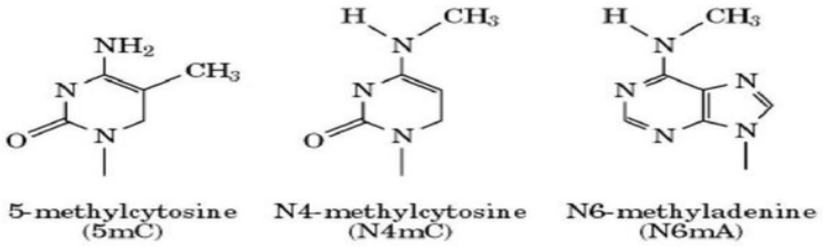
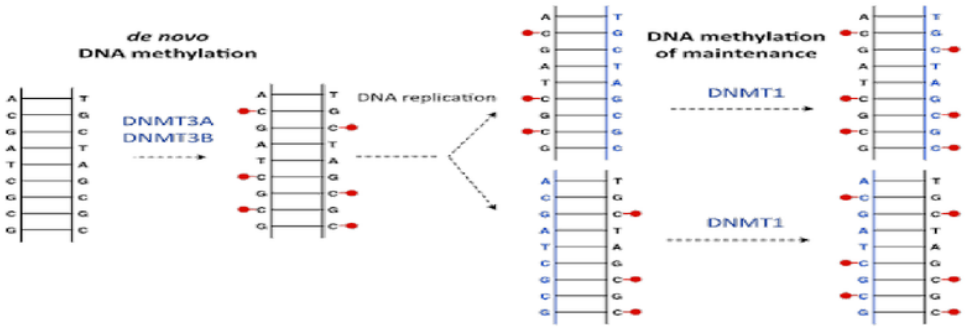


Figure 1: Methylation positions at Cytosine and Adenine base

The Cytosine within 5'-CpG-3' dinucleotides serves as a site for covalent modification of DNA methylation, where a methyl group is transferred to the 5th position of Cytosine to generate 5-methyl Cytosine (5-mC) in genomic DNA (Phillips, 2008). A molecule of S-adenosylmethionine (SAM)[6] serves as a methyl group donor for enzymatically driven methylation process. Apart from 5-methylcytosine (5-mC), an oxidative modified form of 5-mC is also present in the mammalian genomic DNA which is 5-hydroxymethylcytosine (5hmc) and is termed as the sixth base within the DNA (Ratel *et al*., 2006 & Wu *et al*., 2016).

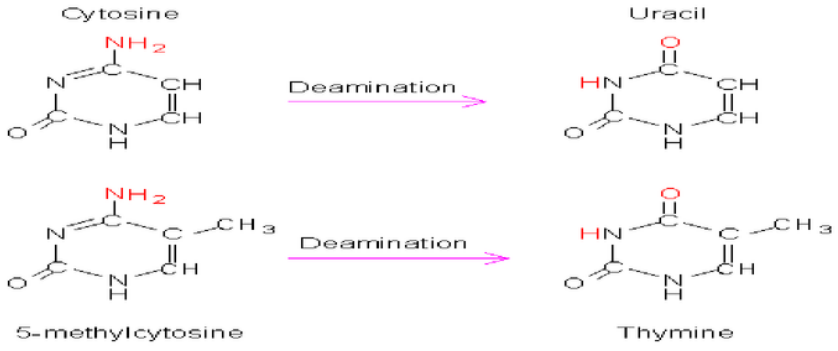
The enzymes that are responsible for carrying out the covalent modification of cytosine to 5-methylcytosine are known as DNA Methyltransferases (DNMT). On the basis of their activity, they are further classified as: *De-novo* Methyltransferases which includes DNMT 3a and 3b that are responsible for setting up of methylation patterns early during embryogenesis and later during gametogenesis (Flynn *et al*., 1998 & Hoelzer *et al*., 2008). Maintenance Methyltransferases have strong preference for hemi-methylated DNA and thus responsible for inheriting these patterns to daughter strands during DNA replication (Pradhan *et al*., 1999).



Source: Moison *et al.*, 2014

Figure 2: For DNA methylation maintenance DNMT1 is mostly involved whereas for *de novo* methyltransferases DNMT3a and DNMT3b involved. Red lollipop indicated the methyl group on the CpG site.

The methylated Cytosine undergoes spontaneous deamination giving rise to Thymine. This transition mutation is not recognized by cellular DNA repair machinery and thus not usually repaired making this conversion irreversible. Deamination of unmethylated Cytosine base results in formation of Uracil, which is recognized by cellular Uracil-DNA glycosylase pathway, and is therefore repaired as shown in figure (Cooper and Youssoufian, 1988 & Hoelzer *et al.*, 2008). Thus the phenomena of Cytosine getting converted to Thymine causes mutations that are not repaired and are left as it is which accounts in under-representation of CpG dinucleotides leading to their suppression in genome of organism.



Source: http://www.web-books.com/MoBio

Figure 3: Spontaneous deamination of Cytosine leads to Uracil whereas of methylated cytosine to thymine.

It has been observed that the mammalian genome has uneven distribution of CpG dinucleotides, making some regions within the genome CpG rich and other being CpG poor. The high frequency CpG rich regions within the mammalian genome are termed as "CpG islands", which make punctuations in the DNA sequences. These regions remain unmethylated and are generally associated with constitutive gene promoters and having independent state of expression (Jones *et al*., 2009). Earlier studies have shown that both vertebrate and invertebrates shows suppressive counts of CpG in their genomic sequences. Also, nucleotide substitutions occurring at average rate have been reported to be most rapid at dinucleotides involving CpG. Further studies have also been carried out in the viruses infecting vertebrates (Cardon *et al*, 1994). Observations inferring lower CpG relative abundance among viral genomes are reported, indicating that these dinucleotides are also methylated in viruses and play important role in the evolution of their genomes. Significant levels of suppression in CpG dinucleotides have occurred in small viral genomes having vertebral hosts. The reason behind this suppression is that the loss of CpG dinucleotides results in minimization of stimulation of toll like receptor 9 and thus lowering immunogenic response against viruses. Also, the CpG methylation renders epigenetic gene silencing in the viral genome (Karlin *et al*., 1994).

Earlier studies provide evidence for the co-evolution of the viral genome along with the host. Analysis on the relative abundance of the dinucleotides on large DNA gives information about various host related factors that play role in shaping evolution of viruses. Along with this the evolutionary pressures i.e. whether translational selection or mutational pressure contributes in the evolution of viruses is also studied. Analysis of the codon usage bias patterns along with the genomic GC content strongly supports that among the DNA viruses, genome wide mutational pressure is the primary factor that determines codon usage, rather than natural selection of coding triplet codons (Upadhyay *et al*., 2014).

1.2 Adenovirus

Adenovirus belong to a family of linear, double stranded DNA, non-enveloped viruses having medium size range (90-100 nm). The GC content of virus lies within the range of 48 - 56%. The non-segmented genome of Adenovirus ranges in size between 26 to 48 kbp. The classification of Adenovirus family is as follows (Davison *et al.*, 2003).

Genus	Species
Atadenovirus	Ovine
Aviadenovirus	Fowl
Mastadenovirus	Mammals
Siadenovirus	Frog
Unclassified	

Table 1: Classification of Adenovirus

Human Adenovirus possesses remarkable capacity to spread infections and cause a wide range of illness. Adenovirus accounts for causing acute respiratory infections and can lead to bronchitis, pneumonia follicular conjunctivitis and multi organ disease among those with weakened immune response. Transmission occurs primarily via droplets of ocular and respiratory secretions. Host defences include neutralizing antibodies as well as cytotoxic-T lymphocytes which are activated in response to viral infection (Saha *et al.,* 2014).

Variations in genome organization exist depending on the species and genera, but generally the Adenoviral transcription unit is divided into Early (E) and Late (L) regions, depending on whether they are expressed before or after DNA replication.

1.2.1 Early Region

This region is transcribed first during viral replication and encodes for proteins that are essential for early adenoviral infection cycle. This region promotes viral infection by altering the cellular environment (Stone *et al.*, 2003). Early region is classified as:

Regions	Function
E1 (E1a/E2b)	Stimulates activation of other genes and induce mitosis in host cell
E2 (E2a/E2b)	Initiates transcription and codes for proteins necessary for viral replication
E3	Essential for modulation of host function, therefore altering host immune responses during disease pathogenesis
E4	Alters cell signalling pathways in host cells

Table 2: Early gene regions with their functions

1.2.2 Late Region

Late region genes are activated after the expression of early function genes i.e. following viral DNA replication and synthesis. This region encodes viral structural proteins primarily (Doerfler, 1996).

Regions	Functions
L1	Encodes protein that helps in facilitating the packaging of virus by aiding in capsid assembly
L2	Polypeptide produced play role in virus internalization via integrins
L3	Polypeptide encoded is a component of viral capsid where it bridges viral capsid and core components
L4	Is responsible for selectively activating the regions that encode for late viral protein synthesis as well as facilitates hexon assembly
L5	Produces polypeptide which trimerize to produce virus fiber

Table 3: Late gene regions with their functions

Adenoviruses while infecting humans having methylated genomes are probably getting methylated and therefore show co-evolution with their host genomes. Due to this, the CpG dinucleotides of Adenoviral genome are expected to be suppressed in comparison to other dinucleotides abundance (Karlin *et al*., 1994). Present study is to investigate the effect of methylation on Adenoviral genome evolution via CpG suppression.

Chapter 2

Review of Literature

2.1 DNA Methylation

It is a process of covalent modification in genomic DNA by addition of a methyl group resulting in modulation of the gene expression. The features of CpG Methylation process including heritability and reversibility makes it a dynamic system suitable for regulating the development of organism (Egger *et al.,* 2004). Also, understanding the role of DNA Methylation in cellular regulation has provided potential for a new archetype of disease intervention and treatment (Crider *et al.,* 2012). The mammalian genome undergoes reprogramming of DNA Methylation patterns. Initiation of this process starts with complete de-methylation of DNA in the pre-implanted embryos and is also known to occur in germ cells. This step is crucial for erasing the existing epigenetic information and for resetting the system for a new developmental cycle (Watters, 2006). Various developmental pathological conditions as well as diseases in the later stages of life are known to occur due to aberration in these methylation patterns (Robertson, 2005). Since DNA Methylation patterns follows a dynamic state, two different methylation processes exists which are: *De-novo* Methylation and Maintenance Methylation. The former one is responsible for establishing the methylation state or patterns early in development whereas the latter is required for copying the methylation patterns onto daughter DNA strands after DNA replication has occurred (Hermann *et al.,* 2004).

2.1.1 Mechanism of DNA Methylation

The process of DNA Methylation involves a covalent modification on the cytosine base. This modification occurs mostly in a 5'-CG-3' dinucleotides pattern by adding a methyl group to Carbon-5 position of the Cytosine pyrimidine ring (Ma *et al.*, 2013). The enzyme family that are involved in this process are termed as DNA Methyltransferases (DNMTs). S-adenosyl-L-methionine (Adomet) serves as a common methyl group ($-CH_3$) donor to the DNA bases for all DNA Methyltransferases.

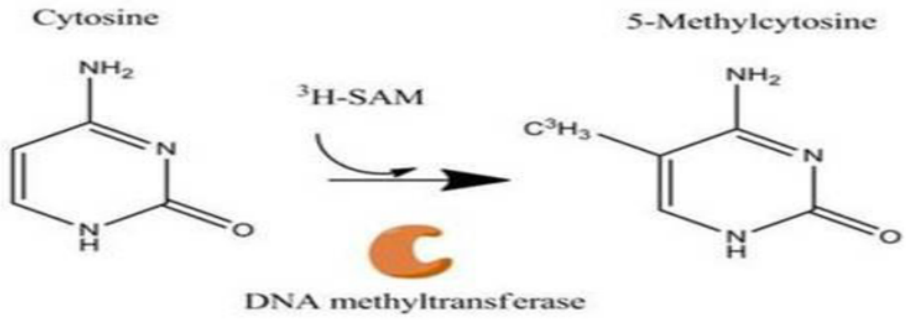


Figure 4: Methylation of Cytosine base carried out by DNA Methyltransferase (DNMT) using S-Adenosylmethionone (SAM) as methyl donor.

Thermodynamic de-stability is introduced by the methyl group of Adomet molecule which is bound to a sulphonium atom, making it highly reactive for a nucleophilic attack by oxygen, nitrogen, sulphur and by activated carbon atoms (Cedar *et al.*, 2012). The prokaryotic DNMT M.HhaI was the first enzyme for which the mechanism of DNA Methylation was analyzed. This enzyme specifically recognizes the 5'-GCGC-3' sequence, causing methylation of the first cytosine lying within this sequence (Portela *et al.*, 2010).

2.1.2 DNA Methyltransferases

The process of DNA Methylation in organisms is an enzymatically driven process. The enzymes which carry out the process of adding a methyl group to DNA are called as DNA Methyltransferases (DNMTs) and are responsible for carrying out the epigenetic modification of DNA Methylation. On the basis of structure and function, mammalian DNMTs are classified majorly into two families of *de-novo* methyltransferases (DNMT3a and DNMT 3b) and maintenance methyltransferases (DNMT1).

2.1.2.1 DNA Methyltransferase 1 (Dnmt 1)

Dnmt 1, the first discovered DNA Methyltransferase is accountable for the inheritance of methylation patterns to daughter cells during the process of DNA replication, therefore is known as 'Maintenance methyltransferase' (Doerfler, 2008). *In-vitro* studies have shown that this enzyme has strong preference for hemi-methylated DNA over the un-methylated ones,

owing to its functionality for maintaining the existing methylation patterns after DNA replication. Therefore, during replication DNMT 1 is required for copying the information to the newly synthesized strands of DNA (Jin *et al*., 2011).

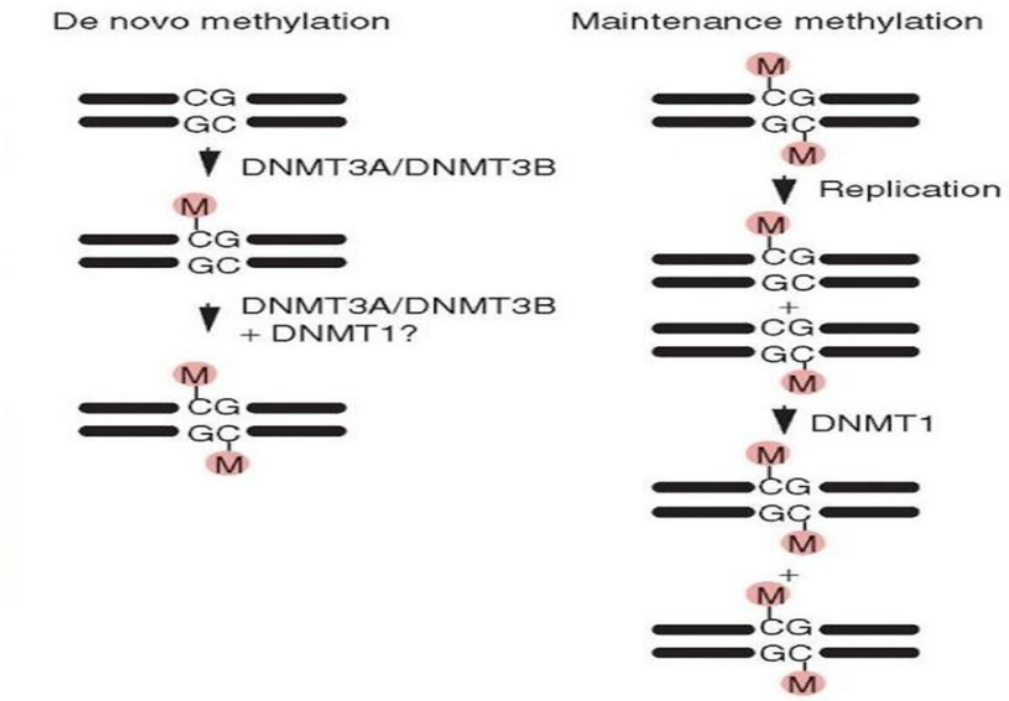


Figure 5: Schematic representation of de novo methylation and maintenance methylation of DNA.

2.1.2.2 Family of DNA Methyltransferase 3 (DNMT3a and DNMT3b)

DNMT 3 family consists of DNMT 3a and DNMT 3b which are accountable for setting up of genomic methylation patterns during embryogenesis as well as gametogenesis, therefore known as '*de-novo* methyltransferase'. These enzymes are thus involved in establishing the patterns of CpG Methylation during embryonic development (Zucker *et al*., 1985). This family also includes DNMT like- protein (DNMT 3L) which shows no catalytic activity but is involved in the physical association of DNMT 3a and DNMT 3b and have a role in modulating their activity (Li, 2002 & Meehan, 2003).

2.1.2.3 DNA Methyltransferase 2 (DNMT2)

DNMT 2 shares structural similarity with the rest of the DNMTs. It has been reported to play role in the methylation of RNA, therefore acts as a RNA Methyltransferase (Hermann *et al.*, 2004). Significance of DNMT 2 is observed in the methylation of tRNA where it specifically methylates the Cytosine base at 38th position of transfer RNA^{Asp}. Since methylation of tRNA has effect on folding of protein and stability of its structure, therefore might have a protective function (Goll *et al.*, 2006).

2.1.3 CpG suppression

Most eukaryotes shows suppression of CpG's in their genome. This CpG frequency suppression is highly variable among the species and correlates negatively with the extent and presence of methylated Cytosine in the genome (Hoelzer *et al.*, 2008). While considering the GC content of mammalian genomes, CpG dinucleotides shows under representation. On the basis of base composition, only 25% of CpG abundance is observed of what is expected. The fraction of G+C content in the human DNA is 0.4, it is expected that the frequency of occurrence of the CpG dinucleotides is 0.2 x 0.2 = 0.4, but in contrast the frequency observed is about 0.008 (Bird, 1980).

Also, the CpG dinucleotides show uneven distribution in the genome, giving rise to clusters called as CpG islands (Cardon *et al.*, 1993). After complete genome sequence analysis of human chromosome 21 and 22, the CpG islands are defined as the regions of DNA having >500 bp with a GC content of >55% and observed/expected CpG dinucleotides ratio of 0.65 (Takai *et al*, 2002). These islands are responsible for regulating the expression of the genes since they are known to occur at or near about 40% of the promoters in a mammalian genome and are generally unmethylated (Bird, 2002).

The CpG represent only one third to one fourth of the expected frequencies in a vertebrate genome; the reason behind this can be the higher stacking energy of the nucleotide base Cytosine and Guanine in comparison to Adenine and Thymine bases. Therefore, structural constrain is an important factor that may lead to CpG avoidance in the genome (Deichmann, 2016). The transcription efficiency of the CpG codons are lower since the proportion of tRNA's that contain CpG in their anticodon are lower in comparison to the tRNA's of any other di-nucleotide. Also, the presence of large number of unmethylated CpG in the genome

stimulates innate immune response reactions if not methylated (Upadhyay *et al*., 2014). At last, the methylated Cytosine shows high propensity of undergoing deamination which also accounts for CpG depletion. Cellular DNA repair machinery is capable of correcting the transitions of unmethylated Cytosine to Uracil base but there is no such mechanism present for the correction of transitions of methylated Cytosine to Thymine; making this an irreversible process and therefore accounts for elevated mutational frequency in the methylated genomes (Jones *et al*., 2009). The methylated Cytosine causes loss of two CpG (one from each strand of DNA), adding one TpG and CpA. This results in deficiency of CpG/CpG and overrepresentation of TpG/CpA as shown in figure (Bird, 1980).

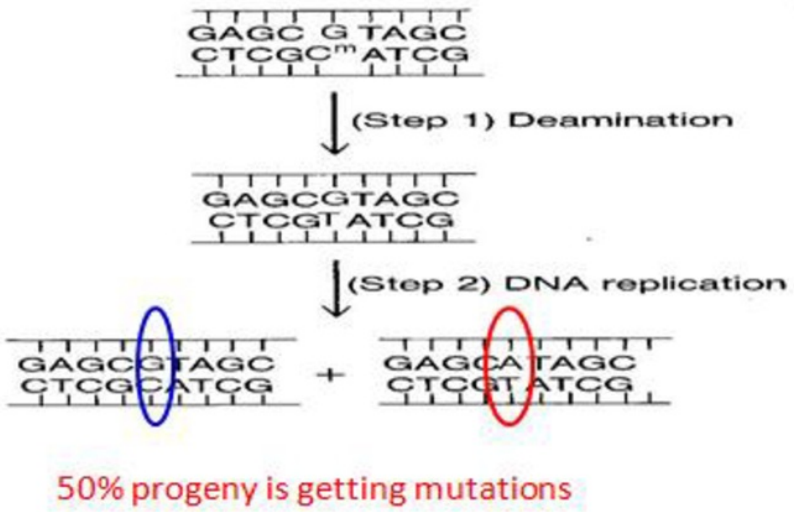


Figure 6: 5-methylcytosine cause loss of two CpG and gain of one TpG and CpA.

Along with CpG, TA is also significantly under-repressed in the eukaryotic and prokaryotic genomes (Karlin *et al*., 1997). The eukaryotic chromosomes show under-representation in the range of 0.61 to 0.81, whereas CpG in vertebrates shows drastic suppression in the range of 0.28 to 0.37 (Karlin *et al*., 1998). The possible reasons underlying the under-representation of TA di-nucleotide is of having the lowest thermodynamic stacking energy among all di-nucleotides; it forms a part of many regulatory signals such as TATA box, transcription terminators and high preference of ribonucleotides for degradation of UA dinucleotides (in mRNA). Thus the suppression of TA confers avoidance for the binding of inappropriate regulatory factors (Karlin *et al*., 1997).

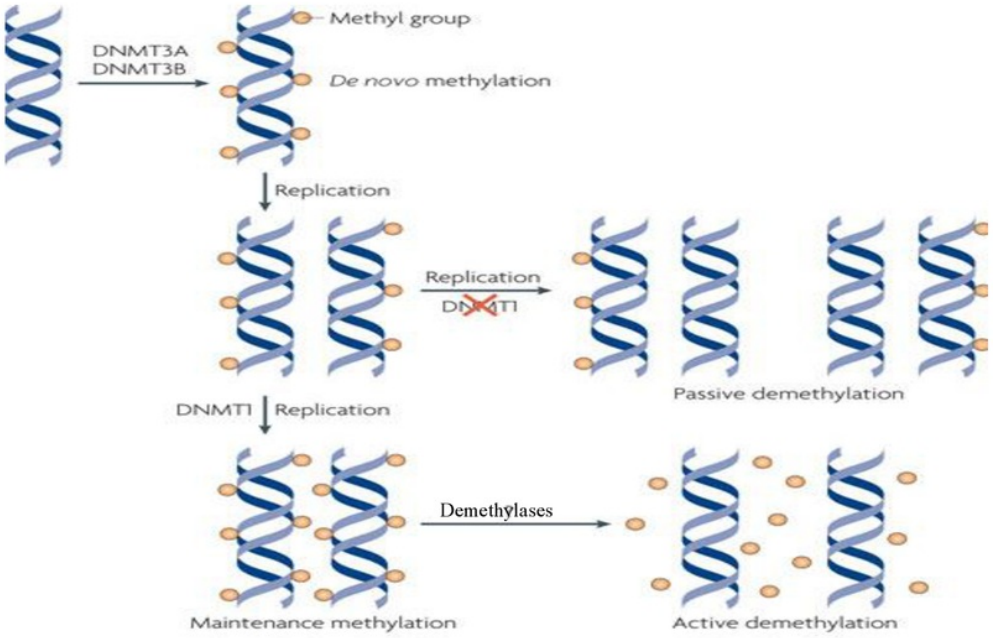
2.2 DNA Demethylation

Earlier DNA methylation was considered to be an irreversible, permanent and unidirectional change, but it is not so. This epigenetic modification is a form of stable and reversible change (Barreto *et al.*, 2007). It has been observed that genome wide demethylation of the DNA occurs after fertilisation, wherein the existing patterns of DNA methylation from both maternal and paternal genomes are erased and are then reset in an early embryo, establishing the epigenetic state (Hafner *et al.*, 2016). Also during successive rounds of replication, a cell with maintenance methylation loses the patterns of CpG methylation in 50% progeny of its daughter cells. The process of removal or loss of methylation marks is done in two different ways:

2.2.1 Active Demethyaltion

It is an enzymatic process that involves active removal of methyl groups or 5mC by breaking C-C bond. Various proposed mechanisms involved in active demethylation are (Pan *et al.*, 2012):

A. Removal of methyl group directly from 5mC enzymatically
B. Nucleotide replacement reaction
C. DNA glycosylase mediated active demethylation involving modification of 5mC followed by Base Excision Repair (BER) system



Source- Nature Reviews Molecular Cell Biology, 2010

Figure 7: The patterns of Methylation are initially established by De-novo Methyltransferases (DNMT 3a and DNMT 3b) during early development of organism. Even after successive rounds of DNA replication and cell division, these patterns are subsequently maintained by maintenance methyltransferase (DNMT 1) which shows high preference for hemi-methylated DNA. During the process of cell division, if DNMT 1 is either absent or inhibited, the newly synthesized strands of DNA will not inherit the methylation patterns, resulting in Passive Demethylation over successive rounds of replication. In contrast, the process of Active Demethylation occurs when 5-methylcytosine (5mC) is replaced with Cytosine enzymatically.

2.2.2 Passive Demethylation

It involves loss of methyl group from 5mC when DNMT1 is either absent or inhibited in the successive rounds of replication, resulting in lack of DNA methylation in the newly synthesized strands of DNA (Rasmussen *et al.*, 2016).

2.3 Adenovirus

Adenoviruses represent medium sized (26-46 kb long) ds DNA and linear genomic viruses that are capable of replicating in the nucleus of vertebrate cells using host replication machinery. This virus belongs to a family of '*Adenoviridae'* which can be further classified into *Mastadenovirus* (includes all Human Adenoviruses), *Aviadenovirus, Atadenovirus*, *Siadenovirus* and *Ichtadenovirus* (Shaw *et al.*, 2008). Human Adenovirus cause wide range of illness; the most common being respiratory infections, eye infections, gastroenteritis to life threatening multi organ disease in people with weakened immune response (Gerba *et al.*, 2008). For studying the course of host driven evolution in viruses, understanding the relative abundance of dinucleotides as well as the extent of codon usage bias are important parameters that need to be considered. One of the factors that contribute in shaping the viral evolution in context to CpG usage is DNA Methylation.

2.3.1 Role of Methylation in host viral interactions

The DNA viruses infecting animals show significant variations in the nucleotide composition, but the role of evolutionary pressures and biological mechanisms, which are responsible for driving these patterns, remains indistinct. The location of viral replication within the host cell along with intracellular trafficking route specifically involved in the pathway affects the susceptibility of viral genome undergoing methylation and immune recognition within the host system (Hoelzer *et al.*, 2008). Among viruses evolutionary studies can be done based on the differences occurring in the relative abundance of dinucleotides. The most extensively studied among all are the CpG whose depletion is reported. The reasons underlying this depletion are linked with evolution of virus, translational selection as well as mutational pressures. Apart from these, virus genome size and the type of genetic material (whether DNA or RNA) are also crucial factors shaping the viral evolution (Upadhyay *et al.*, 2015). Earlier, genome of the Herpes virus has been reported to be significantly suppressed in context to CpG and shows excess of CpA/TpG nucleotides relatively (Karlin *et al.*, 1994). Methylation also plays important role in HBV gene expression by down regulating it. During chronic viral infection, Dnmt's are up-regulated in host as a mechanism of host defence system. The Hepatocytes of the host response to HBV infection by increased expression of Dnmt's, as a result causing methylation of viral DNA and thus leading to inhibition of viral replication and its expression of genes (Vivekanandan *et al.*, 2010).

Evidence which suggests role of methylation in the regulation of viral protein production is provided by the CpG islands of HBV DNA which are methylated in the human tissue (Vivekanandan *et al*., 2009). CpG Methylation at low densities regulates viral DNA, mRNA as well as protein expression, therefore reducing the production of protein encoded by virus. EBV genome shows strong under-representation of CpG in comparison to Herpes Simplex virus which shows relative abundance (Vivekanandan *et al*., 2008). In contrast, Cytomegalovirus shows over representation of CpG dinucleotides. The observed low frequency count in the EBV is due to high probability of 5-methylCytosine undergoing spontaneous deamination during the course of evolution (Burge *et al*., 1992). The suggested hypothesis for this CpG suppression is that the peripheral blood mononuclear cells are able to detect the methylated genome of EBV. In response, the genome is susceptible to mutagenesis by methyl Cytosine deamination which becomes a major contributor in shaping the viral genome over evolutionary time (Ambinder *et al*., 1999).

2.4 Evolution of Viruses

Viruses representing a class of ubiquitous and diverse infectious organisms have DNA or RNA as their genetic elements and require host cell for their replication (Domingo and Perales, 2014). Early studies have shown that *Polyomavirus, Papillomavirus* and *Parovirus* representing group of small DNA viruses are species specific, genetically stable and show relation of being co-evolved with their host species (Shadan and Villarreal, 1995). The vertebrate infecting viral species having small size genomes ($\leq$ 30 kb) are observed to be significantly underrepresented in context to CpG dinucleotides. This suppression prevails irrespective of genomic organization in virus and its morphology. Also, no correlation was found in the GC content of virus with the measure of relative abundance of P^*_{CG} values for dinucleotides. For viruses of larger genomic size having vertebral hosts, suppression of CpG is observed to be inconsistent. Gammaherpesvirus class of virus shows neither over nor under representation of CpG's whereas, Adenoviruses tend to be on the lower side of CpG dinucleotides frequency representation within its genome (Karlin *et al*., 1994). DNA viruses with large genome size emerged from an ancient viral ancestors carrying a smaller subset of 30-35 genes that are involved in encoding the proteins essential for viral structure assembly and replication (Koonin *et al*., 2015).

DNA methylation based CpG suppression in the genomes appears to be and additional factor causing diversity among eukaryotic viruses infecting hosts experiencing extensive genome methylation.

CHAPTER 3

SCOPE OF STUDY

Study Adenoviral genome to see the effect of DNA Methylation. Early studies reveal CpG dinucleotides suppression in vertebrate infecting viruses. In our case we try to use a novel approach by taking advantage of comparative genome analysis of Adenovirus. The genomic sequences of virus isolates infecting humans were selected for studying different kind of mutations that have occurred during the course of evolution along with the host. Such analysis is expected to throw light on CpG loss in Adenovirus.

CHAPTER 4

OBJECTIVES

- Data mining for genomic sequence of various isolates of Adenovirus.
- Comparative genome analysis to study effect of DNA methylation in Adenoviral genome.

CHAPTER 5

Materials and Methods

5.1 Retrieval of Sequences

Complete genomic DNA sequence of the Human Adenoviral strain was searched from National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/). BLAST (Basic Local Alignment Tool) was then run for this sequence. In BLAST, within the algorithm parameters, the maximum target sequences were set to 500 to get maximum number of isolates of Adenovirus that infects humans. A total of 210 sequences were then selected where viral host was human. The sequences of selected strains were then downloaded in fasta format.

Adenovirus was chosen as target organism for studying the effect of DNA Methylation (an epigenetic trait) in the evolution of viral genome. The criterion for selection includes:

- The organism they infect should have a methylated genome.
- The virus should have a dsDNA genome (or dsDNA as its replication intermediate).
- The virus should be known to cause a commonly found disease in humans. Since it will be a human infecting virus, whole genomic sequences of large number of isolates and strains would be available and thus very well studied.

A total of 210 complete genomic sequences of Human Adenoviral strain were used for analysis. Accession number of each genomic isolate of Adenovirus is given in the table below.

S.No.	Accession Number	S.No.	Accession Number	S.No.	Accession Number
1	JO1917.1	71	JN226754.1	141	AB605241.1
2	JX173077.1	72	AP012285.1	142	FJ824826.1
3	KF268130.1	73	JN162671.1	103	AY601636.1
4	KF268310.1	74	JN226760.1	104	JN226759.1
5	JX173084.1	75	JN226758.1	105	HQ910407.1
6	KR699642.1	76	JN226755.1	106	KF268206.1
7	JX173079.1	77	JN935766.1	107	JN226756.1
8	JX173081.1	78	HQ883276.1	108	JN226749.1
9	HQ003817.1	79	AB562587.1	109	JN162672.1
10	KF268129.1	80	KX827426.1	110	KF268315.1

11	FJ349096.1	81	AB695622.1	111	LC066535.1
12	KF951595.1	82	KF268333.1	112	KF268213.1
13	HQ413315.1	83	KF268313.1	113	KF268203.1
14	JX173083.1	84	KF268209.1	114	KF268122.1
15	JX173086.1	85	JN226764.1	115	JN226747.1
16	JX173085.1	86	JN226746.1	116	JQ326209.1
17	JX173078.1	87	AB695621.1	117	JQ326208.1
18	JX173080.1	88	HM770721.2	118	JQ326207.1
19	AF534906.1	89	AB562588.1	119	FJ619037.1
20	JX173082.1	90	KJ626292.2	120	FJ404771.1
21	AY339865.1	91	KJ626291.1	121	EF121005.1
22	KF268127.1	92	KF268335.1	122	AB448776.1
23	AY601635.1	93	KF268329.1	123	AB448775.1
24	KF429754.1	94	KF268204.1	124	AB448774.1
25	M73260.1	95	JN226751.1	125	AB448773,1
26	JX423389.1	96	AB562586.1	126	AB448772.1
27	KF268199.1	97	GQ384080.1	127	DQ900900.1
28	AY487947.1	98	FJ169625.1	128	JN860678.1
29	AY594253.1	99	KF268322.1	129	AB448778.1
30	AY594254.1	100	JN226748.1	130	AB448777.1
31	EF371058.1	101	AB605242.1	131	AY37798.1
32	JN226763.1	102	KF268334.1	132	JN860680.1
33	KF268201.1	103	KF268320.1	133	AF108105.1
34	JN226757.1	104	KF268312.1	134	AB448771.1
35	AB765926.1	105	KF268207.1	135	HQ659699.1
36	JN226752.1	106	KC529648.1	136	GQ478341.1
37	EF153473.1	107	JN226762.1	137	AY594255.1
38	KF268332.1	108	JN226761.1	138	KF268205.1
39	AP012302.1	109	JF99911.1	139	KF006344.1
40	KP641339.1	110	AB605245.1	140	AY599837.1
41	KF268211.1	111	AB605244.1	141	AB333801.2
42	KF268325.1	112	AB605243.1	142	AY599825.1
43	AJ854486.1	113	AB605240.1	143	KF268321.1
44	KF279629.1	114	KF268330.1	144	KF528688.1
45	JN226753.1	115	JN226750.1	145	AY601633.1
46	EF153474.1	116	DQ393829.1	146	KP670856.2
47	KF268327.1	117	AY875648.1	147	KP670855.2
48	K268324.1	118	KF268197.1	148	KP670860.1
49	KF268319.1	119	JN226765.1	149	KP670858.1
50	KF268208.1	120	HQ007053.1	150	KP670857.1
51	KJ364590.1	121	KF577595.1	191	KF577597.1
52	KJ364589.1	122	KF268134.1	192	KF577593.1
53	KJ364587.1	123	JX625134.1	193	KF268316.1
54	KJ364584.1	124	JF800905.1	194	KF268314.1
55	KJ364582.1	125	AY594256.1	195	KF268135.1
56	KJ364581.1	126	KU361344.1	196	KF268117.1
57	KJ36457.1	127	KT963081.1	197	JX423383.1

58	KJ364577.1	128	KJ364592.1	198	KJ364575.1
59	KJ364576.1	129	KJ364591.1	199	KJ364574.1
60	KJ364573.1	130	KJ364588.1	200	KF577598.1
61	KJ019888.1	131	KJ364586.1	201	JN860677.1
62	KJ0198887.1	132	KJ364585.1	202	KF268212.1
63	KJ019886.1	133	LJ364583.1	203	KF268125.1
64	KJ019885.1	134	KJ364580.1	204	JX423387.1
65	KJ019883.1	135	KJ364578.1	205	AY599834.1
66	KJ019882.1	136	KJ019884.1	206	DQ086466.1
67	KJ019881.1	137	KC440171.1	207	KF268210.1
68	KJ019880.1	138	KF938575.1	208	KX423388.1
69	KJ01987.1	139	KF802426.1	209	AY601634.1
70	KC857700.1	140	KF802425.1	210	KF633445.1

Table 4: Accession Number of 210 Adenoviral genomic isolates

5.2 Sequence Analysis Tools

5.2.1 Multiple Sequence Alignment (MSA)

Multiple Sequence Alignment of 210 human infecting Adenoviral strains was carried out. The tool that was used for carrying out alignment of 210 sequences of Adenovirus was "R-Studio" (version 1.0.136.0, 64 bit). The library used under this tool was 'Bioconductor' along with the package 'Decipher' for carrying out alignment.

5.2.2 Microsoft Excel

For carrying out statistical and computational analysis on the obtained sequence data, Microsoft Excel Spreadsheet was used. The various operations along with the functions used were:

Operation Name	Class	Function
IF	Logical	makes logical comparisons and checks whether a condition is met; returns a value depending on whether the test is

		TRUE or FALSE.
COUNTIF	Statistical	counts the number of cells within a range that meet a specific given criteria
CONCATENATE	Text	allow to combine or join text and values to create a single combined string

Table 5: Microsoft Excel operations

5.2.3 Mega 7

MEGA stands for 'Molecular Evolutionary Genetic Analysis' which is a software suite for analysis of DNA as well as protein sequences from various populations and species. This tool is also used for carrying out molecular evolutionary studies as well as phylogenetic tree construction.

5.2.4 Notepad ++

Notepad ++ is a source code editor which has attractive features such as recording and running macros. Macro recording was employed in manipulation and analysis of large DNA sequences. Manipulation of the sequences involves conversion of sequence lines to a single string continuous string. Macro recording was implicated for execution of simple sequence manipulations which were required to be repeated several times such as joining DNA sequences into one single string.

5.3 Methods

5.3.1 Genomic sequence search for Adenovirus

For Adenovirus, genomic sequences were acquired from NCBI (http://www.ncbi.nlm.nih.gov/) in fasta format. The sequences were then subjected to BLAST to get more genomic isolates of the virus. A total of 210 complete genomic sequences of Adenovirus were available which infects humans. Only human infecting strains of virus were selected.

5.3.2 Analysis of mono and dinucleotide frequencies within individual Adenoviral genome.

Counts of A, T, G and C mono-nucleotides were computed for each selected 210 Adenoviral genome. Dinucleotide combinations were then formed within these genomes by using concatenate function in MS Excel. Counts of CpG, TpG and CpA were then computed in these genomes along with the frequencies of GpC, GpT and ApT having similar base composition but different sequence. Observed and expected frequencies of these dinucleotide bases in each sequence were then calculated.

5.3.3 Multiple Sequence Alignment of the genomic sequences

Multiple Sequence Alignment (MSA) of 210 sequences was performed by using R studio programming tool. The library used was 'Bioconductor' and the package used under this library was 'Decipher'. The set of commands that were used in R-studio for alignment were as shown in the picture.

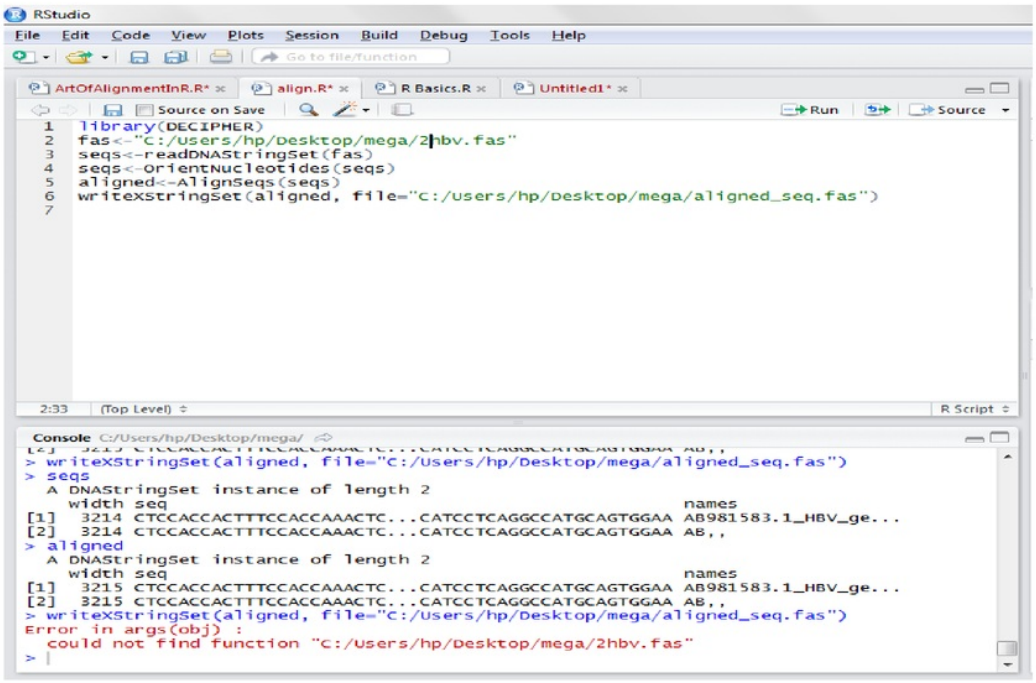


Figure 8: Set of commands used in 'R-studio' for carrying out Multiple Sequence Alignment

Mega 7 software (version 7.0.21_win64) was used to view the output file obtained after alignment which was in .fas format.

5.3.4 Analysis in MS Excel

As further computational and statistical analysis has to be done in MS Excel, the data should be in a format that can be viewed under MS Excel. For this each aligned sequence was copied in Notepad ++ and Macro was run for the conversion of sequence lines to a single string. Similar procedure was repeated for each 210 sequences and these vertically stringed sequences were then copied to MS Excel worksheet.

Software Used	File Extension Required
R studio	.fas
Notepad ++	.txt
Microsoft Excel	.xlsx/.xls
Mega 7	.meg/.fas

Table 6: File extensions required for different software's

Probabilities of expected CpGs were calculated as follows:

$$P(CG) = P(G) \text{ X } P(C)$$

Where P(C) = Total number of Cs/Total length of sequence (i.e. G+A+T+C)

P(G) = Total number of Gs/Total length of sequence (i.e. G+A+T+C)

Expected number of CpG in a given sequence:

= P(CG) x Length of Sequence

= P(CG) x (G+A+T+C)

Probabilities of expected TpG+CpA and rest other dinucleotides (which include AG, AT, AA, AC, TA, TT, TC, CC, CT, GA, GT, GC, and GG) were computed as follows:

Expected number of TpG+CpA = (Net total – CG count) x 2/15

Where, - = gaps in aligned sequence

Net total = 210 - (-N, N-) – (--)

-N = frequencies of –A/-T/-G/-C

N- = frequencies of A-/T-/G-/C-

-- = frequencies of double gaps

Factor of 2/15 was taken instead of 2/16 since CpG dinucleotides were not considered.

Expected number of Rest dinucleotides = (Net total – CG count) x 13/15

Where, - = gaps in aligned sequence

Net total = 210 – (-N, N-) – (--)

-N = frequencies of –A/-T/-G/-C

N- = frequencies of A-/T-/G-/C-

-- = frequencies of double gaps

Factor of 13/15 was taken instead of 13/15 since CpG dinucleotides were not considered.

5.3.5 Data Fragmentation

As the genome size of Adenovirus was too large (approx 36 kbp), for easy handling of the aligned data, the file containing the 210 sequences obtained after alignment was fragmented into eight parts. The total length of the sequence obtained after alignment is 40431. Each of the 8 fragmented parts has 210 sequences of 5053 nucleotide length.

5.3.6 Dinucleotide frequency calculation

The number of A/T/G/C/- were counted within the horizontal rows having 210 sequences by using COUNTIF function (MS Excel).

5.3.6.1 Making di-nucleotide combinations

For analyzing the di-nucleotide frequencies, di-nucleotide combinations were made using CONCATENATE function. Each cell was concatenated with the one lying below, giving di-nucleotide combination. This step is repeated for each 210 sequences.

For calculation of frequencies, all 16 possible di-nucleotide combinations were analyzed along with the combinations that include gaps '-'. A total of 25 combinations were made which were:

'A' nucleotide combinations	'T' nucleotide combinations	'G' nucleotide combinations	'C' nucleotide combinations	'-' combinations
AA AT AG AC	TT TA TG TC	GG GA GT GC	CC CA CT CG	G- A- T- C- -G -A -T -C

Table 7: Possible dinucleotide combinations

5.3.6.2 Counting dinucleotide frequencies

The function 'COUNTIF' was used while specifying the range of the horizontal cells having di-nucleotide combinations of the viral strains whose frequency has to be calculated. Similarly, the function was also used to calculate the frequencies of all 25 combinations that were stated earlier.

5.3.6.3 Determining positions in multiple sequence alignment having maximum CpG count and calculating TA and CA correspondingly at those positions

Positions having maximum CpG were separated from the rest by using 'IF' function. This returns the count of CpG at specific position only when CpG count is maximum (otherwise J) within the range of dinucleotides frequency counts. For TpG and CpA frequency, IF function was used which returns the frequency of TpG and CpA only when maximum CpG count

value is present in the previous cell, otherwise J. All the J's were replaced with nothing and blank cells were deleted, giving maximum CpG count along with TpG and CpA count at respective positions.

5.3.6.4 Applying Chi-Test and obtaining p-value at max CpG positions

Chi-test was applied on the positions that were sorted earlier having maximum CpG count. P-values were obtained to check the level of significance. For this, the parameters that need to be calculated were:

Sum of TpG and CpA frequency

Frequency of rest of the dinucleotides (other than TG, CA, CG, N-,-N,--)] → Actual Range

Sum of expected TpG and CpA frequency

Expected frequency of rest of the dinucleotides (other than TG, CA, CG, N-,-N,--)] → Expected Range

Calculation of p-values at these positions depicts the significance level of CG getting mutated majorly to TpG and CpA. Observed and expected frequencies of TpG + CpA dinucleotides were compared with the rest 13 dinucleotide combinations in overall genomic sequence as well as at max count CpG position and p-values were computed. This shows that TpG + CpA were significantly over-represented than rest any other dinucleotide.

5.3.6.5 Computing C's and G's single transitions and transversions in overall genome of Adenovirus

Probabilities of C's and G's were calculated at each position of multiple sequence aligned Adenoviral genomes by using binomial distribution function (in MS Excel). Significant positions were selected on the basis of p-values (< 0.05) and frequency of C's or G's at a particular position greater than the average i.e. >58 computed for both C's and G's in the

genomes of Adenovirus. Only single transitions (C→T, G→A) along with single transversions (C→G, C→A, G→T, G→C) were considered for analysis.

5.3.7 Mapping of proteins on viral genome

The proteins coded by an Adenovirus were mapped on the viral genome. A total of 28 proteins coded by a Human Adenoviral strain (reference AC_J01917.1) were mapped.

5.3.7.1 Translating viral nucleotide genomic sequences to amino acids

For this the mega file having the alignment of all 210 sequences were translated into six reading frames by using various options of Mega 7.

Frame 1: All nucleotide sequences were selected and translated using translate/untranslate option.

Frame 2: The first line having the first nucleotide base of each sequence was deleted which results in shifting of frame. After deletion, the sequences were selected and then translated.

Frame 3: The first two lines having the two nucleotides bases of each sequence were deleted, resulting in shifting of the frame and the sequences were then translated to amino acids.

The next three frames were of reverse complementary strand. These frames were created by using the options:

Frame 4: All the sequences were selected and reverse complemented by using reverse complement option in Mega 7. These sequences were then transcribed into amino acids using translate option.

Frame 5: After deleting the first nucleotide base of each sequence, they were revere complemented and transcribed to amino acids.

Frame 6: By deletion of first two nucleotide bases the sequences were reverse complemented and then transcribed.

All the six reading frames that were created in Mega were then exported to excel sheet.

5.3.7.2 Mapping of max CG positions on protein sequences

Positions were assigned to proteins coded by Adenoviral genome. CpG max positions were then mapped on proteins as:

Approximate Protein Position = Nucleotide position x 3

The three possible frames of CpG i.e. CGN, NCG or NNCGNN were then determined that which frame among these was involved in coding amino acid in a protein at the mapped positions. Also, the counts of all 20 amino acids were computed at these positions to study the effect of methylation on the coded amino acids. A total of 30 positions were analyzed.

CHAPTER 6

RESULTS

The effect of DNA Methylation on the genome of a higher eukaryotes is observed as under-representation of CpG dinucleotides and corresponding overrepresentation of TpG and CpA dinulcleotides. Earlier reports regarding analysis of CpG dinucleotides frequencies in the genomes of vertebrates and invertebrates have shown this suppression (Cardon *et al*., 1993). Since viruses infects the vertebral hosts, it has been reported that methylation has also been observed in the genomes of viruses leading to CpG suppression and giving evidence of co-evolution of vertebrate infecting viruses with their hosts in the past (Galvan *et al*., 2011). Many DNA viruses have been analyzed for studying the effects of methylation on the evolution of viral genomes.

6.1 Genome Analysis

In this report, Adenoviral genomes infecting humans have been analyzed for studying the under-representation of CpG's. Data mining for this analysis was done from NCBI from where human Adenoviral strains were selected. A total of 210 human Adenoviral genomic sequences were downloaded in fasta format. Analysis of each Adenoviral genomes was based on frequencies of mono-nucleotides (A, T, G and C) and CpG, TpG and CpA dinucleotides.

Additionally GpC, GpT and ApC dinucleotide frequencies were also determined as they were used as controls with same base composition. Expected frequencies of CpG, TpG and CpA were computed as given in methods.

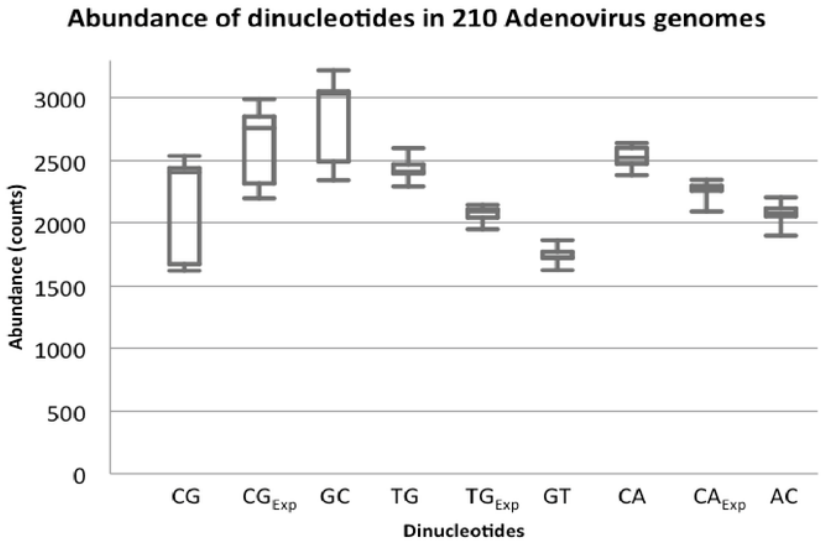


Figure: Graph showing comparison of observed frequencies of CpG, TpG and CpA with observed frequencies of GpC, GpT and ApC along with expected frequencies of CpG, TpG and CpA

In all the 210 Adenoviral genomes, CpG frequencies were found to be lower than expected frequency as well as GpC frequencies. This shows CpG underrepresentation based on base composition as well as the control, in all the strains. Similarly all the 210 strains had overrepresentation of TpG and CpA when compared to their respective expected as well as control frequencies. In was inferred that CpG is underrepresented in Adenoviral genomes while TpG and CpA are overrepresented. It indicates that observed variation in abundance of these dinucleotides is contributed at least partially due to CpG/CpG mutation to TpG/CpA. Further it may be extrapolated to methylation of Adenoviral genomes.

Genomic analysis was done by computing base compositions frequencies in all 210 sequences of Adenovirus (Table 8). Along with this, the frequencies of dinucleotides combinations were also computed.

Nucleotide	Observed Frequency	P(N)
G	2017321	0.272
A	1751955	0.236

T	1612091	0.217
C	2032505	0.274
Total	7413872	

Table 8: Total counts of Nucleotides in Adenoviral genome.

The observed frequency counts for all 16 possible dinucleotides in entire Adenoviral genome were also determined as shown in Table 9.

Dinucleotides	Observed Frequencies	Dinucleotides	Observed Frequencies
GG	563282	**TG**	**505541**
GA	476341	TA	274026
GT	363356	TT	397495
GC	596575	TC	414240
AG	480689	**CG**	**447811**
AA	449404	**CA**	**526189**
AT	362472	CT	470030
AC	433649	CC	565273

Table 9: Total counts of Dinucleotides in Adenoviral genome.

Observed and Expected frequencies of CpG, TpG and CpA have been computed as given in methods. In a similar fashion observed and expected frequencies were also computed for GC, GT and AC dinucleotides which were used as controls with identical base compositions and thereby same expected frequency. The observed frequencies of these dinucleotides have been compared as shown in graph (Figure 9).

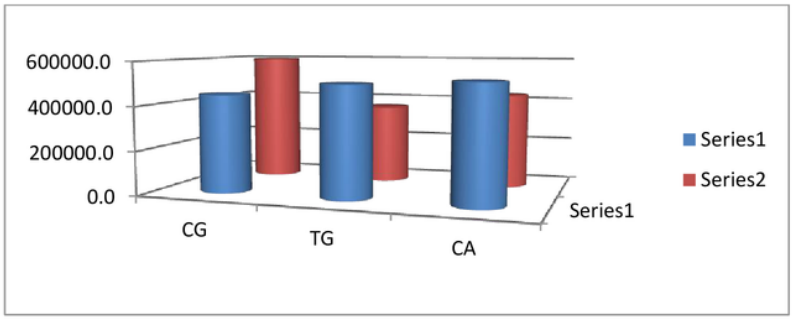


Figure 9: Graph showing comparison of observed frequencies between CpG/GpC, TpG/GpT and TpA/ApT.

The graph shows lower count of CpG dinucleotides in Adenoviral genomes with respect to TpG and CpA dinucleotides. To show that difference between CpG counts and TpG & CpA counts is not relative and it is actually due to loss of CpGs and corresponding gain of TpGs and CpAs, they have been compared with dinucleotide counts having different sequence but

same base composition. CpGs are fewer than GpCs while TpGs and CpAs are more numerous than GpTs and ApCs respectively. The picture becomes clearer when we compare the observed/expected frequency ratios of CpG, TpG and CpA. When the observed and expected frequencies of three sets of dinucleotides (CpG vs. GpC, TpG vs. GpT and CpA vs. ApC) were subjected to Chi-square test, a significant difference between the observed frequencies and their theoretical distributions was observed with a p-value approaching zero as shown in Table 10.

Dinucleotides	Observed Frequencies	Expected Frequencies	O/E Ratios	p-Value
CG	447811	553046.37	0.81	0.00
GC	596575	553046.37	1.08	
TG	505541	438651.36	1.15	0.00
GT	363356	438651.36	0.83	
CA	526189	480296.57	1.10	0.00
AC	433649	480296.57	0.90	

Table 10: p-values and O/E Ratios for dinucleotides

An overall effected of CpG/CpG to TpG/CpA mutations was assessed by comparing TpG and CpA dinucleotide frequencies against rest of all other 14 possible dinucleotides. To study the overall increase of TpG + CpA count in the Adenoviral genome, observed and expected frequencies of these dinucleotides were compared with the observed and expected frequencies of rest 14 dinucleotides. An overall increase of ~1.14 fold is observed for TpG + CpA in comparison to the rest with a p-value approaching zero (Table 11).

	TpG+CpA	Rest 14 Dinucleotides	p-value
Observed Frequencies	1031730	6294643	0.00
Expected Frequencies	918947.93	6389688.7	
RATIO	1.14		

Table 11: Frequency comparison of TpG + CpA with rest 14 dinucleotides.

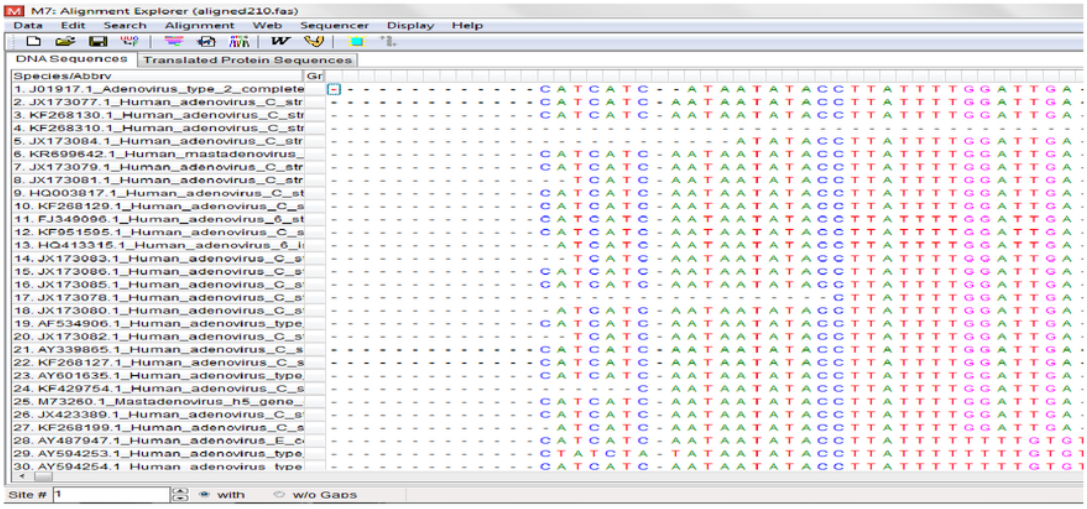
Similarly, on comparing observed and expected frequencies of CpG dinucleotides with rest 14 dinucleotides, a ~1.22 fold decrease was observed with p- value approaching zero within the overall genome of Adenovirus (Table 12).

	CpG	REST 15 Dinucleotides	p-value
Observed frequencies	447811	6878562	0.00
Expected frequencies	553046.37	6973607.70	
RATIO	0.82		

Table 12: Frequency comparison of CpG dinucleotide with rest 15.

6.2 Multiple Sequence Alignment

Our earlier analysis was based on computing the frequencies of nucleotides in the entire Human Adenoviral genome. These values give an overall view about the suppression of CpG whereas gain of TpG and CpA within the genome. This analysis lacks specificity in study of CpG to TpG and CpA. For example it is not possible to distinguish if a given a dinucleotide sequence has same ancestral alleles or it is resulting from a mutation. To overcome this problem a different approach based on Multiple Sequence Alignment (MSA) was adopted to study CpG methylation in virus. MSA of the 210 selected human infecting Adenoviral genomic sequences was performed. Figure below depicts 210 aligned Adenoviral genomes which were then further analyzed for studying the effects of methylation.



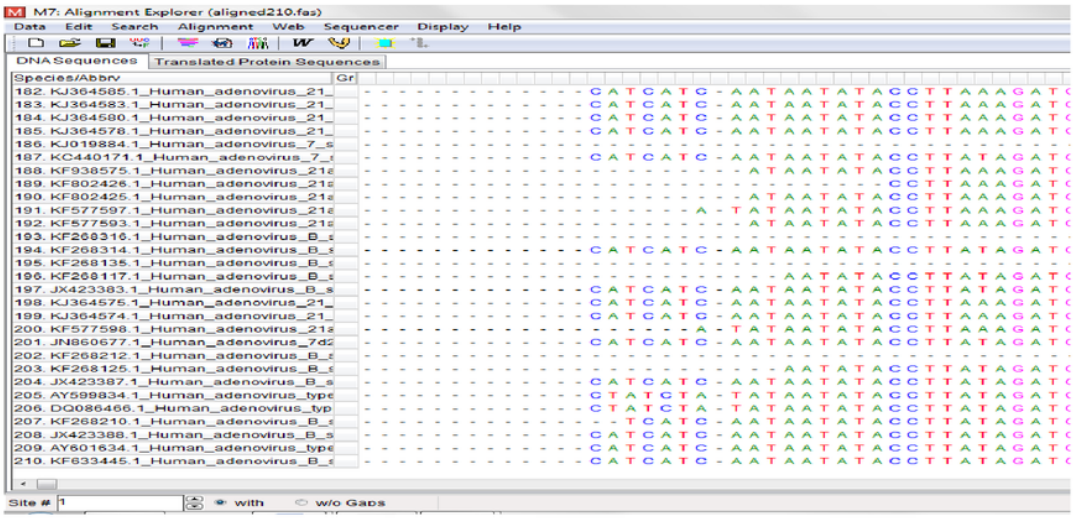


Figure 10: Alignment of 210 human infecting Adenoviral genome sequences.

Dinucleotides combinations were then created in MS Excel and frequency counts were then computed. Max CpG positions were determined where they were present predominantly in comparison to rest any other dinucleotides species. At these positions along with CpG frequencies, counts of TpG + CpA and rest were also computed to determine the conversion of CpG/CpG to TpG/CpA and rest any other dinucleotide. Expected values of TpG + CpA and rest dinucleotides were calculated as given in methods. Comparison for significant difference between observed and expected frequencies of TpG + CpA and rest of the dinucleotides (excluding CpG) was performed based on Chi-square test and P-values were calculated for all the positions in multiple sequence alignment where CpG was the most frequent dinucleotide (Table 13).

S.No.	Alignment position	Max CpG counts	TpG+CpA	Rest (excluding CG,TG,CA,-N,N-,--)	Expected TpG+CpA	Expected rest	p-value
1	92	168	6	5	1.47	9.53	5.79702E-05
2	94	71	19	89	14.40	93.60	0.19287595
3	101	92	87	1	11.73	76.27	3.5961E-123
4	112	107	92	5	12.93	84.07	2.6202E-123
5	119	74	73	59	17.60	114.40	1.13569E-45
6	123	76	7	123	17.33	112.67	0.007674119
7	139	203	5	0	0.67	4.33	1.19193E-08
8	144	131	38	36	9.87	64.13	6.5339E-22
9	147	101	64	44	14.40	93.60	8.84187E-45
10	151	101	0	108	14.40	93.60	4.57782E-05

11	160	100	1	108	14.53	94.47	0.000137147
12	164	174	9	26	4.67	30.33	0.031183049
13	167	105	0	103	13.73	89.27	6.8706E-05
14	169	100	31	77	14.40	93.60	2.61503E-06
15	187	98	3	82	11.33	73.67	0.007837932
16	201	127	46	36	10.93	71.07	4.59429E-30
17	218	210	0	0	0.00	0.00	
18	230	100	83	27	14.67	95.33	7.06126E-82
19	243	205	5	0	0.67	4.33	1.19193E-08
20	249	89	12	109	16.13	104.87	0.268994351
21	262	126	84	0	11.20	72.80	9.3359E-121
22	264	208	2	0	0.27	1.73	0.000311491
23	275	101	0	109	14.53	94.47	4.22123E-05
24	296	203	7	0	0.93	6.07	1.52639E-11
25	323	106	2	101	13.73	89.27	0.000671384
26	325	209	0	0	0.00	0.00	
27	327	154	0	55	7.33	47.67	0.003627349
28	367	148	33	28	8.13	52.87	7.5286E-21
29	374	82	1	126	16.93	110.07	3.1937E-05
30	383	126	79	4	11.07	71.93	1.19E-106
31	395	180	0	29	3.87	25.13	0.034666267
32	402	206	3	0	0.40	2.60	1.006E-05
33	417	209	0	1	0.13	0.87	0.694886602
34	424	209	0	1	0.13	0.87	0.694886602
35	435	183	0	27	3.60	23.40	0.041540072
36	441	101	0	109	14.53	94.47	4.22123E-05
37	454	99	0	83	11.07	71.93	0.000352368
38	456	92	6	112	15.73	102.27	0.008392009
39	464	210	0	0	0.00	0.00	
40	466	210	0	0	0.00	0.00	
41	471	127	1	82	11.07	71.93	0.001151991
42	483	108	0	102	13.60	88.40	7.45266E-05
43	533	150	50	10	8.00	52.00	2.81969E-57
44	540	176	0	34	4.53	29.47	0.022190719
45	568	210	0	0	0.00	0.00	
46	592	206	4	0	0.53	3.47	3.41417E-07
47	597	104	5	101	14.13	91.87	0.009063698
48	635	173	3	34	4.93	32.07	0.349789579
49	676	78	29	103	17.60	114.40	0.003512443
50	715	133	50	27	10.27	66.73	1.76382E-40
51	742	131	79	0	10.53	68.47	1.0983E-113
52	763	108	1	101	13.60	88.40	0.000242478
53	799	108	0	0	0.00	0.00	
54	808	133	77	0	10.27	66.73	7.3991E-111
55	868	86	45	78	16.40	106.60	3.297E-14
56	1004	178	5	27	4.27	27.73	0.70293882

Table 13: Positions within Adenoviral genome having maximum CpG counts with respect to rest

The positions with highest frequency of CpG and significant difference in TpG + CpA and rest of the 13 dinucleotide frequencies were considered to have undergone CpG methylation based mutations from CpG/CpG to TpG/CpA. It implies that such positions had disproportionately higher TpG + CpA as mutation products in comparison with rest of the 13 dinucleotides.

In similar fashion, positions for rest of the 15 dinucleotides were also determined where a dinucleotide was having maximum count in comparison to the rest as well as the counts of positions where there was 100% conservation of a dinucleotide i.e. a position had only a single dinucleotide across all 210 genomes (Table 14).

Dinucleotides	total positions having maximum dinucleotide count	total positions having 210 dinucleotide count	conservation percentage
CG	**2353**	**277**	**11.77**
TG	2415	438	18.14
CA	2444	445	18.21
GC	3054	441	14.44
GT	1639	335	20.44
AC	1970	366	18.58
GG	2828	636	22.49
GA	2314	428	18.50
AG	2258	278	12.31
AA	1874	386	20.60
AT	2892	474	16.39
TA	1055	226	21.42
TT	1636	359	21.94
TC	1909	367	19.22
CT	2208	243	11.01
CC	2892	474	16.39

Table 14: Total positions having maximum dinucleotide counts and 100% conservation.

It has been inferred from above data that CpG dinucleotide was one of the least conserved dinucleotide (showing only ~11% CpGs in the genome exhibited 100% conservation) among rest all other dinucleotides. This shows that CpG dinucleotides were one of the most frequently mutated bases in the genome of Adenovirus during the course of evolution when compared to the rest.

6.3 Analysis at max CpG count positions

Analysis of maximum CG positions was done. Since the modification involving methylation of CpG/CpG results in formation of TpG and CpA dinucleotides, their frequencies were also computed at respective positions of CpG max. TpG + CpA counts were then compared with rest other dinucleotides to check over-representation of these dinucleotides within the viral genome at specific CpG max sites that will be majorly due to the effect of methylation at these sites. The observed and expected frequencies were computed for TpG + CpA dinucleotides in comparison to rest other 13 dinucleotides only at CpG max positions (Table 15).

	TpG+CpA dinucleotide	Rest 13 dinucleotides	p-value
Observed Frequencies	47075	99573	0.00
Expected Frequencies	65319.6	424577.4	
RATIO	3.07		

Table 15: Frequency comparison of TpG+CpA with rest 13 dinucleotides at CpG max positions.

On comparison of TpG + CpA frequencies with rest other dinucleotides, it has been observed that there was a ~3.073 fold abundance of TpG + CpA dinucleotides in comparison to rest. As TpG + CpA were the products of methylated CpG sites, therefore their over-representation was determined to be the result of methylation phenomena, whereas the rest were due to any other mutations. There was a significant increase of TpG + CpA frequency from ~0.88 to ~0.33 folds at overall genomic and at specific CpG max sites respectively.

This increase could also be because of biased base composition within the genome of Adenovirus. Therefore to eliminate this factor, an analysis of TpG + CpA dinucleotide frequency was done in comparison to the rest while considering base composition (Table 16).

	TG+CA dinucleotide	Rest 13 Dinucleotides	p-value
Observed Frequencies	47075	99573	0.00
Expected Frequencies	18176.99	117531.63	
RATIO	3.06		

Table 16: Frequency comparison of TpG+CpA with rest 13 dinucleotides at CpG max positions taking base composition into account.

A fold value of ~0.326 was computed on comparing TpG + CpA frequencies with rest other dinucleotides while taking base composition into account. It has been observed that even after considering base composition, only marginal difference of 0.001 was observed in the fold

ratio of TpG + CpA frequency at CpG max positions. This supports our result inferring that majority of TpG and CpA frequencies were arising as a result of methylation at CpG sites. However the mutations in CpG dinucleotides may mutate to rest of 15 possible dinucleotides because of single transition mutations, single transversion mutations, double transition mutations, double transversion mutations or double mixed mutations.

Table of ss, sv, ds, dv, and dm

It is known that transitions occur more often than transitions. In order to take these factors into consideration, analysis was performed again based on total number of transitions and transversions of C and G. On the other hand only those mutation products were taken into consideration which result out of single mutations (TpG, CpA, CpC, CpT, ApG & GpG). A 2 x 2 contingency table was obtained for this data as following:

	At CpG sites	In rest of the Adenoviral genomes
Transitions (C→T) & (G→A)	47075	444918
Transversions (C→G), (C→A), (G→T) & (G→C)	27179	669312

Both the possible transitions at CpG sites result in TpG and CpA which may be caused by mutation as a result of CpG methylation or some other mechanisms while all the possible transversions giving rise to CpC, CpT, GpG and ApG cannot be the result of CpG methylation lead mutation. The Odds ratio of getting TpG and CpA against rest of the four dinucleotides at CpG sites in comparison with rest of the result is 2.6056. This odds ratio is statistically significant with a p-value of <0.0001. In this analysis the factor of higher propensity of transition mutations has been taken into consideration and yet a >2.6 fold higher abundance of TpG and CpA has been found against rest of the four dinucleotides. This

disproportionate abundance of TpG and CpAs at CpG sites strongly indicates involvement of CpG methylation as a major cause of these mutations. Thus it may be inferred from this data that Adenoviral genome has been getting methylated during the course of evolution whenever it came in contact with the host cell system.

6.4 Assessment of effect of CpG mutations on Adenoviral proteome.

The CpG positions that were determined earlier in multiple sequence alignment of 210 Adenoviral genomes were mapped on protein sequences of Adenovirus in order to investigate the effect of their mutations to TpG and CpA. A CpG can be part of codons in the coding regions and when mutated to TpG or CpA as a result of CpG methylation, can cause synonymous, miss-sense or nonsense mutations. The codons having CpG can be categorized in three classes, namely CGN, NCG or NNCGNN. In the last CpG is split into two successive codons. A sample of 30 CpG position in coding regions of protein [redacted] were analyzed. It was observed thatCpGs were present as NNCGNN i.e. split CG more frequently than as CGN or NCG at these positions. Assuming equal probability of a CpG to be part of the above mentioned three classes of codon, a Chi-square test was performed to test if observed frequency of the three classes conforms to the expected frequency. The obsevered frequencies were found to be significantly different from theoretically determined frequency with a p-value of 0.006.

	CGN	NCG	NNCGNN
Observed positions	8	4	18
Expected positions	10	10	10
p-value	0.006		

Based on binomial distribution of NNCGNN was found to be overrepresented when compared to CGN and NCG. This observation suggests a role of evolutionary pressure in minimising the effect of CpG mutations on protein sequences.

6.4.1 Amino Acid counts at mapped CpG positions in [redacted] protein

Counts of all the different amino acids present at the mapped CpG positions were determined to study the effects of methylation on proteins where CpG was present as NCG or CGN. The analysis of these positions having CGN and NCG codon are shown in Table below.

Alignment Position	Coded Amino Acids
635	all double mutations of I, rest R
1004	all double mutations of T and H with one mutation of Q which can be resulting from mutation of CG to CA or other single mutations
1166	all R
1599	all R
1677	all single mutations of L (not involving CG to TG/CA mutations), one double mutation of I
1923	two double mutations of K
1926	all single mutations of L with one mutation of Q, which can be resulting from mutation of CG to CA or other single mutations and 27 double mutations of N
1955	all double mutations of K

Table : Positions having CGN codon and coded amino acids

Alignment Position	Amino Acid Position	Coded Amino Acids
1522	507	double mutations of N, single mutations of CG to TG/CA resulting to L
1543	514	double mutations G,D,C and N
1547	515	double mutations of F,D,N and single mutation of CG to TG/CA resulting to V
1549	516	single mutations of CG to TG/CA resulting to V

Table: Positions having NCG codon and coded amino acids

Double mutations were predominantly observed more frequently than single mutations at mapped positions of proteins having CGN and NCG codons involved in coding amino acids.

CHAPTER 7

DISCUSSION

Heterogeneity in relation to relative dinucleotide abundance within the genomes of vertebrate and invertebrates has been observed. Since vertebrates serves as host for many DNA viruses, the nucleotide composition is also found to vary in the viral genomes, but evolutionary pressure and biological mechanisms lashing these patterns are indistinct. Earlier studies have reported suppression of CpG dinucleotides in vertebrates as well as in all genomic sequences of animal mitochondria (Burge *et al*., 1991). Genomic DNA viruses infecting vertebrates have been reported to undergo epigenetic modification at 5-Cytosine, leading to methylation of CpG dinucleotide. Viral genomes are considered to have loss of CpGs similarly to their host, due to deamination of methylated Cytosine. In DNA, presence of one methyl Cytosine would cause loss of two CpGs and add one TpG and one CpA, which is considered to be a prime reason of CpG deficiency and TpG/CpA excess.

Under-representation of CpG dinucleotides with relative abundance of TpG and CpA has been well studied in viral genomes. Short oligonucleotide extremes of most bacteriophage sequences have shown underrepresentation of TpA while overrepresentation of GpC dinucleotides. Small DNA viruses including Hepatitis B virus and class of papovaviruses have shown underrepresentation of CpG frequencies within their genome. Intermediate and large genome sized viruses have shown normal range of CpG relative abundance. An exception of viruses having large genome size but still show CpG suppression is the family of Gammaherpesviruses (including EBV and Bovine herpes virus) which have potently suppressed CpG dinucleotide and relatively higher abundance of TpG/CpA dinucleotides. Retrotransposons infecting eukaryotes having genome size within the range of ≥5 kbp often have lower CpG dinucleotide abundance (Karlin *et al*., 1994). There are some other possible reasons underlying suppression of CpG dinucleotides within a genome of organism. The CpG represent only one third to one fourth of the expected frequencies in a vertebrate genome; the reason behind this can be the higher stacking energy of the nucleotide base Cytosine and Guanine in comparison to Adenine and Thymine bases. Therefore, structural constrain is an important factor that may lead to CpG avoidance in the genome (Deichmann, 2016). The transcription efficiency of the CpG codons are lower since the proportion of tRNA's that

contain CpG in their anticodon are lower in comparison to the tRNA's of any other di-nucleotide. Also, the presence of large number of unmethylated CpG in the genome stimulates innate immune response reactions if not methylated (Upadhyay *et al*., 2014).

For viruses belonging to same family and having similar genome organization and life cycle, the relative abundance of CpG dinucleotides is dependent on the infected host cells. Infecting viral genomes shows strong correlation between the evolutionary lineage of the infected host and the extent of CpG dinucleotide reduction. Since viruses have short reproductive life cycle with higher rates of evolution, observable changes are expected in their genome over relatively shorter time periods ([redacted]).

We attempted to study Human Adenoviral genomic strains having DNA genome, size ranging between 26 to 46kbp and causing wide range of respiratory infections among humans. Adenovirus is chosen as a target organism for this study since it is a double stranded virus which infects humans i.e. an organism with methylated genome. Additionally its genome size is much larger that the viruses used in the earlier similar studies providing an advantage of larger data size. Genomic sequences are selected to study the effect of methylation on the viral genome evolution. The sequences of viral isolates are analyzed and compared by Multiple Sequence Alignment tool, a novel approach which gives clue about the ancestral allele and the mutations which have occurred in the parent allele during the course of evolution.

Previous analysis of CpG dinucleotides in the genomes of Papillomaviruses and polyomaviruses have been reported to be underrepresented. The extent of suppression within these small double stranded DNA viruses is determined by the evolutionary lineage of the host in which virus is causing infection. Phenomena of methylation have been reported to be the primary cause of suppression within these DNA viruses. Also, it has been demonstrated that the depletion of CpG dinucleotides is more pronounced in human infecting strains and other mammals in comparison to those strains in which birds are the hosts for infecting viral species (Upadyay et al., 2015).

The study plan involving analysis of Adenoviral genome for methylation was based on a logic that these viruses having double stranded DNA genomes infects humans and are expected to undergo methylation and hence ultimately resulting in modification of CpG to TpG/CpA. A total of 210 genomic isolates of human Adenovirus are considered for carrying out methylation studies. Our analysis of selected Adenoviral sequences individually shows

significant suppression of CpG observed frequencies in comparison to observed frequencies of GpC and with those of TpG and CpA expected frequencies. Overall analysis of these genomes also showed significant under-representation of TpG + CpA frequency in comparison to rest other dinucleotides. These sequences are then subjected to Multiple Sequence Alignment tool, for analyzing the dinucleotide frequencies and their substitutions within the viral genome.

Since mutations of CpG dinucleotides due to methylation produces modified TpG/CpA dinucleotide bases, there frequencies are analyzed in the entire genome as well as at the positions where CpG counts are maximum when compared to the rest. It has been observed that 2353 positions in multiple sequence alignment are occupied by CpGs, where they are present predominantly in comparison to rest of the other dinucleotides and exhibit great extent to mutations.

Analysis of CpG and TpG+CpA dinucleotide frequency counts profoundly shows underrepresentation and overrepresentation of these dinucleotides at both CpG maximum count positions as well as in the overall genome of Adenovirus. This data is strongly favoured by calculating observed/Expected (O/E) Ratios. The deviation of observed frequencies and the expected frequencies were found to be statistically significant with a p-value approaching zero.

So from the above result we say that during the course of evolution in Adenoviral genome, CpG are underrepresented and TpG and CpA overrepresentation due to the phenomena of methylation which results in conversion of CpG/CpG dinucleotide to TpG/CpA.

CHAPTER 8

CONCLUSION

The genomes of the vertebrates are susceptible to methylation at the CpG dinucleotide sites, where a methylated Cytosine gives rise to TpG and CpA dinucleotide. These methylation patterns are crucial for the development of embryo, genomic imprinting as well as for inactivation of X chromosome in mammals. Mutation of the CpG dinucleotide bases to TpG and CpA has lead to its suppression in higher eukaryotic organisms. Since these vertebrates serves as host for various viruses, it is expected that viruses have coevolved with their host and thus shows similar levels of CpG suppression within their genomes. CpG suppression in vertebrates infecting viral genomes has been previously observed for Papillomaviruses and Polyomaviruses which belong to a group of small double stranded DNA viruses infecting humans.

Current study is to investigate the role of methylation in evolution of Adenoviruses. On studying individual Adenoviral genome, CpG suppression is observed on comparing frequency of observed CpG with those of observed GpC as well as with expected CpG. Similarly relative abundance of TpG & CpA is observed on comparing observed frequencies of these dinucleotides with observed frequencies of GpT & ApC along with those of expected TpG & CpA. Also, the frequencies of dinucleotides are computed for CpG and TpG+CpA and are compared with rest 15 and 14 dinucleotides respectively in all the selected sequences of Adenoviral genome. TpG+CpA dinucleotides show 1.14 fold more abundance with respect to rest 13 dinucleotides whereas CpG dinucleotides are 0.82 fold deprived in comparison to rest 14 dinucleotides within the overall genome of Adenovirus.

CpG suppression can also be a result of higher stacking energies of Cytosine and Guanine bases in comparison to Adenine and Thymine bases, resulting in structural constrain within the DNA and thus leading to their avoidance. Also, immune response is stimulated in response to the presence of unmethylated CpG dinucleotides leading to their suppression. The major factors contributing in the suppression of CpG dinucleotides can be Cytosine

methylation, Post regulatory mechanisms, Random mutations and thermodynamic stability of CpG dinucleotides

Among these, the stability factors and random mutations do not play a predominant role in CpG suppression since it has been observed that there are many fold increase in the frequency of TpG + CpA dinucleotides in the viral genome with respect to rest other dinucleotides.

In this study we have used a novel approach by performing Multiple Sequence Alignment of 210 sequences of Adenoviral strains infecting humans. Following this approach leads to identification of CpG positions in the genomes that are present predominantly with respect to rest others. These positions are considered for further analysis since it can be assumed that any changes in the dinucleotides are most likely resulting from CpG to exhaustive set of rest other dinucleotides. Analysis of aligned adenoviral sequences revealed that CpG is one of the dinucleotide that shows lowest percentage of conservation in the genome of Adenovirus.

For all CpG sites having maximum count, TpG+CpA dinucleotides frequencies are compared with rest other dinucleotides and three fold abundance is observed. It may be concluded that like the host, Adenoviral genomes also exhibit CpG suppression which is largely due to mutation of CpG/CpG to TpG/CpA arising from methylation of the CpG. Therefore, CpG sites are underrepresented due to accumulation of these mutations during the course of evolution.

a

ORIGINALITY REPORT

%5	%1	%5	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Upadhyay, M., J. Samal, M. Kandpal, S. Vasaikar, B. Biswas, J. Gomes, and P. Vivekanandan. "CpG DINUCLEOTIDE FREQUENCIES REVEAL THE ROLE OF HOST METHYLATION CAPABILITIES IN PARVOVIRUS EVOLUTION", Journal of Virology, 2013. Publication	%1
2	Zakar, Tamas. "DNA methylation in development", Embryology - Updates and Highlights on Classic Topics, 2012. Publication	<%1
3	R. W. Honess. "Deviations from Expected Frequencies of CpG Dinucleotides in Herpesvirus DNAs May Be Diagnostic of Differences in the States of Their Latent Genomes", Journal of General Virology, 04/01/1989 Publication	<%1
4	Subcellular Biochemistry, 2013. Publication	<%1

5	A. Hermann. "Biochemistry and biology of mammalian DNA methyltransferases", Cellular and Molecular Life Sciences CMLS, 10/2004 Publication	<%1
6	Upadhyay, Mohita, and Perumal Vivekanandan. "Depletion of CpG Dinucleotides in Papillomaviruses and Polyomaviruses: A Role for Divergent Evolutionary Pressures", PLoS ONE, 2015. Publication	<%1
7	Nava-González, E.J., E.C. Gallegos-Cabriales, J.C. Lopez-Alvarenga, J.W. Kent, and R.A. Bastarrachea. "Recent Advances in Genomics of Body Composition, Adipose Tissue Metabolism, and Its Relation to the Development of Obesity", Pathobiology of Human Disease, 2014. Publication	<%1
8	Robert Ivarie. "Asymmetrical distribution of CpG in an 'average' mammalian gene", Nucleic Acids Research, 1982 Publication	<%1
9	Konsoula, Zacharoula, and Frank A. Barile. "Epigenetic Modifications and Stem Cell Toxicology: Searching for the Missing Link", Handbook of Nanotoxicology Nanomedicine and Stem Cell Use in Toxicology, 2014.	<%1

	Publication	
10	felicitysmoak.info.tm Internet Source	<%1
11	Renata Z. Jurkowska. "Silencing of Gene Expression by Targeted DNA Methylation: Concepts and Approaches", Methods in Molecular Biology, 2010 Publication	<%1
12	www.larapedia.com Internet Source	<%1
13	docs.di.fc.ul.pt Internet Source	<%1
14	Sandi, Chiranjeevi, Madhavi Sandi, Sara Anjomani Virmouni, Sahar Al-Mahdawi, and Mark A. Pook. "Epigenetic-based therapies for Friedreich ataxia", Frontiers in Genetics, 2014. Publication	<%1
15	Crider, K. S., T. P. Yang, R. J. Berry, and L. B. Bailey. "Folate and DNA Methylation: A Review of Molecular Mechanisms and the Evidence for Folate's Role", Advances in Nutrition An International Review Journal, 2012. Publication	<%1
16	M. Randić. "Highly compact 2D graphical representation of DNA sequences", SAR and QSAR in Environmental Research, 6/1/2004	<%1

	Publication	
17	gbe.oxfordjournals.org Internet Source	<%1
18	www.ideal-ageing.eu Internet Source	<%1
19	www.plosone.org Internet Source	<%1
20	ab.inf.uni-tuebingen.de Internet Source	<%1
21	Thomas J. Near. "Intraspecific phylogeography of Percina evides (Percidae: Etheostomatinae): an additional test of the Central Highlands pre-Pleistocene vicariance hypothesis", Molecular Ecology, 9/2001 Publication	<%1
22	Yongmei Xiao. "Age and gender affect DNMT3a and DNMT3b expression in human liver", Cell Biology and Toxicology, 06/2008 Publication	<%1
23	Tapader, Rima, Dipro Bose, Pallabi Basu, Moumita Mondal, Ayan Mondal, Nabendu Sekhar Chatterjee, Pujarini Dutta, Sulagna Basu, Rupak K. Bhadra, and Amit Pal. "Role in proinflammatory response of YghJ, a secreted metalloprotease from neonatal septicemic Escherichia coli", International Journal of	<%1

Medical Microbiology, 2016.
Publication

---

24 Schilling, Elmar. "Analysis of tissue-specific & allele-specific DNA methylation", Publikationsserver der Universität Regensburg, 2010. <%1
Publication

---

25 Pan Tao. "Analysis of synonymous codon usage in classical swine fever virus", Virus Genes, 02/2009 <%1
Publication

---

26 Ke-xin Wen, Jelena Miliç, Bassem El-Khodor, Klodian Dhana, Jana Nano, Tammy Pulido, Bledar Kraja, Asija Zaciragic, Wichor M. Bramer, John Troup, Rajiv Chowdhury, M. Arfam Ikram, Abbas Dehghan, Taulant Muka, Oscar H. Franco. "The Role of DNA Methylation and Histone Modifications in Neurodegenerative Diseases: A Systematic Review", PLOS ONE, 2016 <%1
Publication

---

27 Sau, K.. "Factors influencing synonymous codon and amino acid usage biases in Mimivirus", BioSystems, 200608 <%1
Publication

---

28 Koumbi, Lemonica, and Peter Karayiannis. "The Epigenetic Control of Hepatitis B Virus <%1

Modulates the Outcome of Infection", Frontiers in Microbiology, 2016.
Publication

29	eprints.utas.edu.au Internet Source	<%1
30	Butler, J.S.. "PAGE separation of hemi-methylated or unmethylated oligonucleotide substrates to distinguish between maintenance and de novo DNA methyltranferase activity", Journal of Biochemical and Biophysical Methods, 20061031 Publication	<%1
31	Epigenetics and Human Health, 2013. Publication	<%1
32	Upadhyay, Mohita, Neha Sharma, and Perumal Vivekanandan. "Systematic CpT (ApG) Depletion and CpG Excess Are Unique Genomic Signatures of Large DNA Viruses Infecting Invertebrates", PLoS ONE, 2014. Publication	<%1
33	Methods in Molecular Biology, 2009. Publication	<%1
34	Hiroshi Yamaki. "Molecular basis for the actions of Hsp90 inhibitors and cancer therapy", The Journal of Antibiotics, 09/2011 Publication	<%1

EXCLUDE QUOTES	ON	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	ON		