Studying epigenetic effect on evolution of virus genomes

A Dissertation Report

Submitted in partial fulfillment 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

Nehakshi Sharma

Roll no. 601404012

DEPARTMENT OF BIOTECHNOLOGY

THAPAR UNIVERSITY

PATIALA-147004

July 2016

CANDIDATE DECLARATION

I hereby declare that the work being presented in the M.Tech dissertation entitled "**Studying epigenetic effect on evolution of virus genomes**" has been carried out by me during the period of July 2015 to July 2016, 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.

N Sharma

Nehakshi Sharma
M.Tech Biotechnology
Roll. No. 601404012

Date: 15/7/2016
Place: Patiala

CERTIFICATE

This is to certify that dissertation entitled **"Studying epigenetic effect on evolution of virus genomes"** submitted by **Nehakshi Sharma (601404012)** 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.

15.07.2016

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

Dr. Dinesh Goyal
Head of Department
Department of Biotechnology
Thapar University
Patiala

Dr. S.S Bhatia
Dean (Academic Affairs)
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. Dinesh Goyal, 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 Neha mam, Supriya mam 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.

Lastly, I would also like to thank my friends Shivangi, Sonika, Akash, Ankit, Tania, Ananta, Charu di and Shikha di who supported me in writing, and incented me to strive towards my goal.

Nehakshi Sharma

TABLE OF CONTENTS

ABBREVATIONS

(O/E)	Observed/expected ratios
A	Adenine
B	Not A (G, T and C)
C	Cytosine
C^5	Carbon at 5^{th} position
CpA	Phosphodiester bond between cytosine and adenine
CpG	Phosphodiester bond between cytosine and guanine
CpT	Phosphodiester bond between cytosine and thymine
D	Not C (G, A and T)
DNA	Deoxyribonucleic acid
Dnmt	DNA methyltransferase
Dnmt1	DNA methyl transferase 1
Dnmt3a	DNA (cytosine-5)-methyltransferase 3A
Dnmt3b	DNA (cytosine-5)-methyltransferase 3 beta
dsDNA	Double-stranded DNA
dsRNA	Double-stranded RNA
EBV	Epstein-Barr virus
G	Guanine
H	Not G (A, C and T)
HBV	Hepatitis B virus
HIV	Human Immunodeficiency Virus
K	Keto (G or T)
M	Amino (A or C)
m6A	N6-methyladenine
MAFFT	Multiple Alignment using Fast Fourier Transform
^{m}C	Methyl Cytosine
mRNA	Messenger RNA

MSA	Multiple Sequence Alignment
MTases	Methyltransferases
N	Any nucleotide (A, C, G and T)
N^4	Nitrogen at 4^{th} position
N^6	Nitrogen at 6^{th} position
nBlast	Nucleotide blast
NCBI	National Centre for Biotechnology Information
OR	Odd Ratio
R	Purine
RNA	Ribonucleic acid
S	Strong (G or C)
SAM	S-adenosyl-L-methionine
ssDNA	Single-stranded DNA
ssRNA	Single-stranded RNA
T	Thymine
TpG	Phosphodiester bond between thymine and guanine
V	Not T or U (G, C and A)
W	Weak (A or T)
Y	Pyramidine

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

DNA methylation is an epigenetic mechanism which play imported role in vertebrate genome and which is used to regulate many cellular processes that include gametogenesis, early embryogenesis, cellular differentiation and development, genomic imprinting. Viruses have very short period of reproductive life cycle and higher rate of evolution; we should expect observable changes in genome over relatively shorter time periods. The changes are based on host compatibility. We attempted to study Human viruses causing common diseases and having DNA genome or DNA intermediate in there genome replication. They have been selected to study changes in genome structure over stipulated time. Our data present evidence that in small viruses (HBV and adenovirus) the CpG is under-representative and TpG and CpA is over- representative. In large DNA viruses (cytomegalovirus) the CpG is not under-representative except in EBV where CpG is under-representative. Our finding based on Multiple Sequence Alignment on evolution of HBV genome to support a novel perspective that CpG is mutated to TpG/CpA.

Keywords. Epigenetics, DNA methylation, CpG methylation, virus genome

CHAPTER 1
INTRODUCTION

The term epigenetic, given by Conrad Waddington in 1942 to describe that 'it shows interaction between genes and their products that have been studied which brings the phenotype into being' (Goldberg *et al*., 2007). It is a study of genetic change in gene expression and activity without vary in the DNA sequence *viz.* changes in phenotype and without change in genotype. Epigenetic change is directly influenced by both environment as well as individual life style. The two major epigenetic modifications are:

DNA methylation: DNA methylation occurs at N^4 and C^5 position of cytosine and N^6 position of adenine in prokaryotic organisms, while at C^5 position of cytosine in eukaryotes.

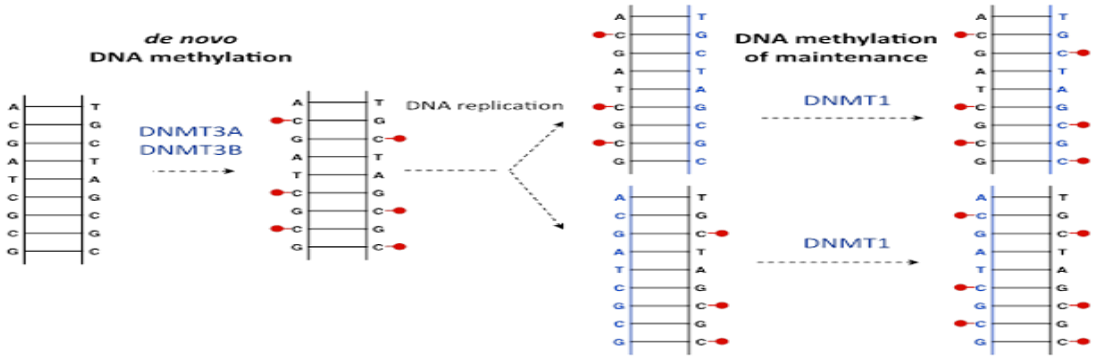
Histone modification: In chromatin eukaryotic DNA is packaged, folded and compacted. The building blocks of mammalian chromatin, Histones, are the basic proteins which can be covalently modified by methylation (at lysine and arginine residues), phosphorylation, acetylation (at lysine residues), ubiquitylation *etc*. at their unstructured and alkaline N-terminal tails (Strahl and Aliis, 2000). Compare to DNA methylation, which is moderately stable, histone modifications are more dynamic and complex to analyze (Handy *et al*., 2011).

In eukaryotes, there are number of mechanism to control gene expression but commonly used epigenetic signaling tool is DNA methylation which fixes genes in off position.

1.1 DNA methylation

DNA methylation is an epigenetic mechanism which is used to control gene expression by cells. DNA composed of cytosine, guanine, thymine and adenine. In 1948 Hotchkiss discovered DNA methylation in calf thymus DNA (Hermann *et al*., 2004). Addition of the methyl group to DNA strand is known to DNA methylation. DNA grooves are filled with methyl groups, such that they do not interfere with the Watson/Crick base pairing capacity. DNA methylation occurs at N^4 and C^5 position of cytosine residues and N^6

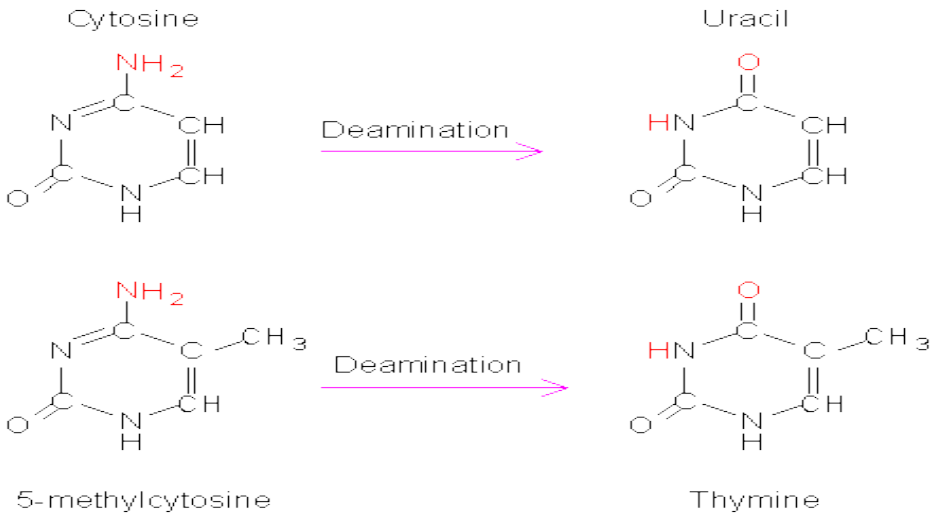
position of adenine residues in prokaryotes. In eukaryotes DNA methylation occurs only at C^5 position of cytosine, mostly in context of CG dinucleotides sequence (Karlin *et al*., 1994). In eukaryotes particularly in mouse embryonic stem cells, N6-methyladenine (m6A) is another form of DNA modification. In mammalian evolution m6A developed a new role in epigenetic silencing (Ratel *et al*., 2006 & Wu *et al*., 2016). Cytosine base is converted to 5-methylcytosine by DNA methyltransferase (DNMT) enzymes. These modified cytosine residues generally have a guanine base (CpG methylation) next to them in 3´ direction (Phillips, 2008). Thus cytosine methylation occurs at palindromic CpG dinucleotides in both DNA (Bird, 1980). DNA methylation in mammals is carried out by two classes of enzyme activities namely maintenance methyl transferases and *de novo* methyl transferases as shown in Figure 1. Maintenance DNA methyl transferase (DNMT1) has strong preference of hemi methylated methylation pattern (CG) in the DNA when compared to unmethylated substrate and thus inherit in the daughter strands of DNA during DNA replication (Pradhan *et al*., 1999). The *de novo* methyl transferases (DNMT3a and DNMT3b) do not have any preference for hemi-methylated or unmethylated DNA as substrate. These enzymes setup genomic methylation pattern during gametogenesis and later cellular methylation pattern during embryogenesis (Flynn *et al*., 1998 & Hoelzer *et al*., 2008)



Source: Moison *et al.*, 2014

Figure 1: **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.**

In mammals, DNA methylation play important role in cellular differentiation and development, X-chromosome inactivation, embryonic development, control of gene expression, preservation of chromosomal integrity, cancer biology, regulation of chromatin structure and genetic diseases (Bird, 2002, Li, 2002 & Feinberg and Tycko, 2004). Spontaneous deamination of methylated cytosine gives rise thymine. Such transition mutation (^{m}C-T) is usually not repaired. Deamination of unmethylated cytosine result into uracil, which is repaired by uracil-DNAglycosylase pathway, is shown in Figure 2 (Cooper and Youssoufian, 1988 & Hoelzer *et al*., 2008).



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

Figure 2: Spontaneous deamination may change the cytosine to uracil, or the methylated cytosine to thymine.

In most eukaryotic genome CpG are under-represented, but the frequency varies broadly among species and is negatively associated with occurrence and extend of cytosine methylation in the genome. CpG are highly methylated in vertebrate genome. 60-90% of genomic CpG are consideration to be in a methylated state, but methylation and frequency of CpG differ widely across a single vertebrate genome (Hoelzer *et al*., 2008). One methyl cytosine would cause loss of two CpG and add of one TpG and CpA, which could we a reason for CpG deficiency and TpG and CpA excess (Bird, 1980). Relative

abundance GCG and CGC trinucleotides and CCGG, CGCG and GCGC tetranucleotides are normal, signifying that CpG under-representation in certain higher order oligonucleotides are not a consequence of the reduced frequencies (Cardon *et al*., 1994). In dinucleotides, TA is generally under-representated, with the exemption of vertebrate mitochondrial genomes, and CG is strongly under-representated in both vertebrates and in mitochondrial genomes. In trinucleotides, GCA-TGC is under-representated in phage, human viral and eukaryotic sequences and CTA-TAG is under-representated in several prokaryotic, eukaryotic and viral sequences. The AC-GT doublet in eukaryotic and mitochondrial sequence is under-representated, whereas CA-TG doublet in eukaryotes is over-representated but in mitochondrial genomes it is under-represented (Burge *et al*., 1992).

1.2 Virus genome

Today we are in the middle of the genomic revolution which was spearheaded by many international projects, planning to sequence the whole genomes of organism ranging from bacteria to mammals. In these organisms many of genes have been recognized and advancement is being made towards understanding the roles of these genes in health and diseases (Primrose and Twyman, 2009). Latest discovery in environmental genomics show that viruses, especially bacteriophage are the widespread and copious biological entities on earth (Koonin *et al*., 2015). Viruses are the simplest form of life. They stand on the borderline between living and the in animated, non-biological world. Unlike as any genome, genome of virus has basic genes for virus replication. Viral genomes are the fastest growing entities in biology, because of their small replication time and the huge number of offspring released per cell infected. Viruses use host molecular machinery for their multiplication and growth. Host cells once infected by the viruses are engineered to construct specific viral proteins or gene products, which are used by them as efficient tools to direct host cell tasks for their own continued existence and replication (Kim, 2001). A small amount of eukaryotic genomes codes for molecules intricate in cellular function and structures, and remaining have a viral origin (Bamford *et al*., 2002). Viruses enter host cell through skin, gastrointestinal tract and respiratory tract. Animal viruses

mainly spread throughout the body by blood stream or nervous system. They counter host immune response by several methods like HIV destroy white blood cells, influenza viruses changes their genes by antigenic drift or antigenic shift. In antigenic drift, virus surface proteins are changed which results in the development of new virus strain which may not predictable by host antibodies. In antigenetic shift, through combination of genes from different viral strains, a new virus subtype is formed. While antigenic drift happens steadily over time, antigenetic shift occurs quickly. In DNA viruses connection between GC composition and codon usage bias is explained by overall nucleotide content. In DNA viruses, genome-wide mutational pressure is mainly significant feature shaping pattern of codon usage bias relatively. Nucleotide biases differ among closely correlated viruses (Shackelton *et al.*, 2006).

1.2.1 Types of virus genome

On the basis of replication-expression strategies and particularly on the form of nucleic acid that is in corporate into virions, viruses genomes are classified into 7 classes (Baltimore, 1971) as shown in table 1

Table 1: 7 classes of virus genome classified by Baltimore in 1971

GENOME	EXAMPLE
Positive-strand ssRNA virus(virions contain RNA of the same polarity as mRNA)	Picornaviruses, Togaviruses
Negative-strand ssRNA virus(virions contain RNA molecules complementary to the mRNA)	Orthomyxoviruses, Rhabdoviruses
dsRNA viruses	Reoviruses
Reverse-transcribing viruses with +ve strand ssRNA genomes	Retroviruses
Reverse-transcribing viruses with dsDNA genomes	Hepadnaviruses
ssDNA viruses	Parvoviruses
dsDNA viruses	Adenoviruses, Herpesviruses, Poxviruses

dsDNA viruses enter the host nucleus before it is able to replicate. These viruses are highly dependent upon the host cell cycle. ssDNA viruses replicate normally within the nucleus frequently via a rolling circle mechanism, forming double-stranded DNA intermediate in the process. In dsRNA virus replication is monocistronic and each of the genes codes for only one protein, contrasting other viruses that explain more complex translation. In both positive and negative sense ssRNA virus the replication of viruses occurs in the cytoplasm or nucleus. ssRNA viruses do not much depend upon host cell cycle (Koonin *et al*., 2015).

1.3 Methylation in viruses

The nucleotide composition significantly varies in DNA viruses of animals, but evolutionary pressure and biological mechanism lashing these patterns are indistinct. The frequency of CpG, in small DNA viruses is under-representated whereas in large DNA viruses it is not under-representated. Viral genome susceptibility to methylation and immune recognition may affected by the location of replication within the cell and the specific intracellular trafficking route (Hoelzer *et al*., 2008). In herpes virus genome CpG significantly under abundant and have relative excess of CpA/TpG (Karlin *et al*., 1994 a).The methylation patterns is different in occult and non occult HBV infection, which shows that epigenetic changes may be related to occult HBV(Vivekanandan *et al*., 2008).

The present work is an attempt to understand the effect of interaction between viruses and their methylated genomes on the genome evolution of the viruses.

CHAPTER 2

REVIEW OF LITERATURE

2.1 DNA methylation

DNA methylation is an epigenetic mechanism which is used to regulate many cellular processes. These include X-chromosome inactivation, gametogenesis, early embryogenesis, cellular differentiation and development, genomic imprinting, preservation of chromosomal integrity and transcription. Other than these roles, a growing number of human diseases have been associated with peculiar DNA methylation (Robertson, 2005 & Hermann *et al*., 2004). In DNA methylation, at C^5 position of cytosine nucleotides a methyl group is added to form methylcytosine by DNA methyltransferases (DNMTs) in mammalian cells. Such modifications mostly occur on CpG dinucleotides (Ma *et al*., 2013). In plants; methylation in cytosine occurs at CpG, CpHpG (symmetrical) and CpHpH (asymmetrical) site. DNMTs are classified as 'maintenance' methyltransferases (DNMT1) and '*de novo'* methyltransferases (DNMT3a and DNMT3b) (Jin *et al*., 2011). All DNMTs use S-adenosyl-L-methionine (SAM) as methyl donor to the DNA base. In SAM, methyl group is bound to a sulphonium atom which destabilizes the molecule thermodynamically (Hermann *et al*., 2004). DNMT1 is accountable for inheritance of methylation pattern to daughter cell during DNA replication process. This is based on its strong preference for hemi-methylated (Zucker *et al*., 1985). DNMT3a and DNMT3b accountable for genomic methylation pattern during gametogenesis and demethylation of genome during embryogenesis (Li, 2002 & Meehan, 2003).

2.1.1 Mutagenicity

In vertebrates CG suppression due to methylation /deamination /mutation causing mutation of CG to TG/CA. In all animal mitochondrial and mammalian genome, CC/GG is high but TG/CA is in the usual range, associate with a CG →CC/GG mutation bias. In

eukaryotic and prokaryotic genome TA is significantly underrepresented (Karlin *et al*., 1997). In eukaryotic genome TA is under-representated *i.e.* TA~0.61-0.81, whereas CG show drastic suppression *i.e.* CG~0.23-0.37(Karlin *et al*., 1998). TA is under-representative due to several reasons like it has the minimum thermodynamically constant DNA duplex of all dinucleotides and entailing flexibility of the TA site for unwrap the DNA double helix (Karlin *et al*., 1997). One methyl cytosine would cause loss of two CpG and add of one TpG and CpA, which could we a reason for CpG deficiency and TpG and CpA excess (Bird, 1980). Relative abundance GCG and CGC trinucleotides and CCGG, CGCG and GCGC tetranucleotides are normal, signifying that CpG under-representation in certain higher order oligonucleotides are not a consequence of the reduced frequencies (Cardon *et al*., 1994). In vertebrate genomes CG is considerably deficient. The AC-GT doublet in eukaryotic and mitochondrial sequence is under-representated, whereas CA-TG doublet in eukaryotes is overrepresentated but in mitochondrial genomes it is under-represented (Burge *et al*., 1992). There are chances that life cycle of DNA viruses can get greatly affected by CpG methylation. The effect varies from virus to virus and many other factors which include virus lifecycles, the flanking nucleotide motives, genomic location and the host species and infected tissues. Large and small DNA viruses CpG methylation may also differentiate the DNA lifecycle. In small DNA viruses CpG frequency is under represented than that of larger DNA. This might be because of partial contribution of cytosine methylation (Hoelzer *et al*., 2008). Among large DNA viruses there is a relative abundance of dinucleotides which affects several invertebrates and vertebrates hosts. Also, invertebrate hosts are infected by the depleting CpT dinucleotides sequence and excess CpG presence which is a unique genomic signature of large DNA virus. They are also affected by the pressure induced by hosts, different from translation pressure lead to CpT depletion and CpG excess in large DNA viruses and hence infecting hosts (Upadhyay *et al*., 2014).

2.1.2 Flanking bases on the distribution of CGs

Methylated genome has uneven CG dinucleotide distribution (that shows variation near the 5’ end of genes in vertebrates, invertebrates, plants and bacteria) because of the occurrence of CpG islands. CGs having flanking sequences which are chosen by the

DNA methyltransferase is more prone to get methylated compared to those CGs which are having flanking sequences which are moderately preferred or completely not preferred by DNA methyltransferases (Weissbach *et al*., 1984). Studies conducted by Ollila *et al*., was a good example showing that significant mutation seen in YCGR combination as well as mutation appear in all tetra-nucleotides. Ollila *et al*., conducted studies on hexa nucleotides because studies on tetra nucleotides presented great results and thus paved paths for studies on even longer nucleotides as well as to measure effects of long flanks surrounding CG nucleotide. Ollila *et al*., proposed that asymmetric sequences *viz.* mixed occurrence of pyramidine and purine in one or both sides are under-represented and very rare. They also proposed that YYCGRR and YYCGRY contain YCGR pattern and are highly prone to get mutated (Ollila *et al*., 1996). For DNMT1there is no preference for CG flanking sequences but they have a preference for GC-rich flanking sequences (Flynn *et al*., 1998). For DNMT3a there is a strong preference for CG site flanked by pyramidine bases and YNCGY free consensus sequence (Lin *et al*., 2002).Flanking sequence preference and Jeltsch proved the relation between the tendencies of a CpG site to undergo methylation. It was further proposed by them that there are discrete statistically consensus sequences flanking CpG site which stimulate various levels of methylation. The generation of DNA methylation pattern of mammalian genome is influenced by the intrinsic sequence inclination of *de novo* DNMTs but that process is not well understood (Handa and Jeltsch *et al*., 2005).

2.1.3 Role of methylation in virus-host interactions

The nucleotide composition significantly varies in DNA viruses of animals, but evolutionary pressure and biological mechanism lashing these patterns are indistinct. Viral genome susceptibility to methylation and immune recognition may affected by the location of replication within the cell and the specific intracellular trafficking route (Hoelzer *et al*., 2008).In herpes virus genome CpG significantly under abundant and have relative excess of CpA/TpG (Karlin *et al*., 1994). The HBV gene expression is down regulated by the methylation. Hepatocytes react to HBV infection by upregulating DNMTs, which can methylate HBV viral DNA, leading to abate viral replication and viral gene expression. Due to this, there is increased expression of DNMTs in chronic

viral expression (Vivekanandan *et al*., 2010). HBV DNA has been shown to contain CpG islands that are methylated in human tissue, which suggests a role for methylation in regulating viral protein production (Vivekanandan *et al*., 2009). Due to methylation of viral DNA, HBV mRNA and protein expression are also regulated and low densities of CpG methylation reduce production of viral protein (Vivekanandan *et al*., 2008). In EBV CpG is strongly under-representated compare to relative abundance in herpes simplex virus but in cytomegalovirus CpG is over-representated dinucleotides (Burge *et al*., 1992). Studies carried out by Ambinder *et al*., showed that in EBV CpG suppression is recognizing by peripheral blood mononuclear cells which are prone to mutagenesis methylcytosine deamination. Due to this virus are perpetuating over evolutionary time (Ambinder *et al*., 1999).

2.2 Virus evolution

Viruses carry infectious DNA or RNA genetic elements that require a cell for their multiplication. Viruses are very diverse and present everywhere in environment (Domingo and` Perales, 2014). Viruses are physical abundant entities in the biosphere and they have genetic diversity (Koonin *et al*., 2015). Small DNA viruses like polyomaviruses, papillomaviruses, parvoviruses *etc*. are species specific, stable genetically and co-evolved with their host species (Shadan and Villarreal, 1995). Large and giant DNA viruses emerge from ancient viral ancestors who have a small subset of 30–35 genes which is encoding for structural proteins and replication. The diversity and genome size of these viruses grow due to gene duplications which are lineage-specific, lateral gene transfers of cellular genes and growth of various families of movable genetic elements (Filée and Chandler, 2008).

CHAPTER 3

SCOPE OF STUDY

Study viral genome to see the effect of DNA methylation. Earlier studies reveal CG suppression. In our case we try to use novel approach taking advantage of comparative genome of virus. The genome sequences of virus isolates of same evolutionary lineage but differing by maximum possible time period may be compared to study different kind of mutations. Such comparison is expected to throw light on CG loss in the viruses.

CHAPTER 4

OBJECTIVES

1. Selection of viral genomes for sequence comparison.
2. Viral genome sequence analysis to study epigenetic effect on their genome structure.

CHAPTER 5

MATERIALS AND METHODS

5.1 Data source

DNA sequences were downloaded from the National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/). To study the effect of epigenetic on evolution of virus genome, 4 different double stranded DNA viruses was selected. These viruses are selected on the basis of:

a) They infect organism which has methylated genome.
b) There viral genome (or its replication intermediate) is dsDNA.
c) They cause commonly found diseases in humans. Therefore these viruses are very well studied *i.e.* whole genome sequence of large number of isolates and strains are available.

The earliest and latest sequence of HBV is taken from Devesa and Pujol, 2007 paper. For EPV, adenovirus and cytomegalovirus genome sequence are downloaded from NCBI.

Table 2: DNA sequences of different viruses taken from Gen Bank

Virus Name	Genotype/strain/serotype	Date	Accession no.
	Genotype A	15-Nov-1997	D50521
HBV	Genotype A	20-Jul-2007	D00331
	Genotype F	17-Jan-2001	AF223963
HBV	Genotype F	9-Feb-2010	AB166850
	Serotype 9	14-NOV-2006	AJ854486.1
Adenovirus	Serotype 15	31-Dec-2013	KF268201.1
	Serotype 48	18-Apr-2007	EF153473.1
Adenovirus	Serotype 44	4-Nov-2013	JN226763.1
	Strain Merlin	2-Jul-2013	AY446894.2
Cytomegalovirus	Strain CZ/1/2013	27-May-2015	KP745691.1
	Strain GD1	6-Jan-2006	AY961628.3
EBV	Isolate EBVaGC8	11-Jan-2016	KT273948.1

5.2 Sequence analysis tools

Multiple Sequence Alignment (MSA)

MSA is the alignment of the three or more biological sequences (proteins or nucleic acid) of similar length. From the output, homology could be inferred and the evolutionary relationships between the sequences were studied. Tools used for MSA were Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/) and MAFFT (Multiple Alignment using Fast Fourier Transform) (http://www.ebi.ac.uk/Tools/msa/mafft/).

Pairwise Sequence Alignment

Pairwise sequence alignment is to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences. Tool used for pairwise sequence alignment was Matcher (http://www.ebi.ac.uk/Tools/psa/emboss_matcher/nucleotide.html).

Microsoft Excel

Microsoft excel spreadsheet was used for computational and statistical analysis sequence data. The various tool with their function are shown in table 3

Table 3: Microsoft Excel Functions

Tool name	Class	Function
IF	Logical	Calculate certain conditions, and respond differently depending on whether the test is FALSE or TRUE.
COUNTIF	Statistical	Used to count the number of cells that meet the condition.
CONCATENATE	Text	Used to unite two or more string into one string.
RIGHT	Text	Used to return the particular number of characters from the last part of the string.
LEFT	Text	Used to return the particular number of the characters from the first part of the string.

Notepad ++

Notepad++ was primary used for recording macros to manipulate and analyze DNA sequences of large size. DNA sequence manipulation such as converting the sequence lines into a single line was performed by 'line operation' (ctrl J). Macro recording was applied to execute simple sequence manipulations which were required to be repeated several times.

Computation of Chaos Game Representation of frequencies (FCGR)

(http://www.biophp.org/minitools/chaos_game_representation/demo.php?FCGR)
This tool was used for determining mono and di nucleotide frequency of DNA sequences.

Odds ratio calculation

http://vassarstats.net/odds2x2.html

This web based tool was used for determining Odds Ratio (OR) and p-value. Odds ratio (OR) is one of the numerous statistics which is important decision-making and case control studies (Scotia, 2010 & McHugh, 2009**).** It measures the ratio of the odds that an event will happen to the odds of the event not happening.

	Standard treatment	New Treatment
Event Happens	A	B
Event does not happen	C	D

OR= (a/c)/ (b/d)

p- value or statistical probability reveals whether the findings in a research study are statistically significant, *i.e*. the findings are not likely to have occurred by chance. Before calculating the p-value, researchers specify α level (Forbes, 2012)
α = 0.10 it means 10% chances that the conclusion are wrong
α = 0.01 it means 1% chances that the conclusion are wrong

5.3 Methods

5.3.1 Genomic sequence search of viruses

HBV

The genome sequences of HBV having Genotype A, Genotype F were selected randomly based on Devesa and Pujol, 2007 work. The selected two isolates had 10 and 9 years of time gap respectively.

EBV, Adenovirus and Cytomegalovirus

For other three viruses (Adenovirus, EBV and cytomegalovirus) genome sequences were acquired from NCBI (http://www.ncbi.nlm.nih.gov/) in fasta format. For these three viruses, following sequences were subjected to nBLAST to get genome sequences of more isolates. MSA of randomly selected 250 sequences was performed by Clustal omega for Adenovirus and MAFFT for cytomegalovirus and EBV. The parameters selected in Clustal omega was

Dealign input sequences: No

Mbed-like clustering guide-tree: Yes

Mbed-like clustering iteration: Yes

Number of combined iterations: 0

Max guide tree iterations: Default

Max hmm iterations: Default

Order: Aligned

The parameters selected in MAFFT was

Gap extension penalty: 1.53

Gap open penalty: 0.123

Maxiterate: 2

Guide tree output: On

Tree rebuilding number: 2

Order: Aligned

Perform FFTS: None

From the results of the above Alignment the Phylogenic tree were acquired. From the phylogenic tree sequences, nodes were isolated on the basis of a significant time gap. (An example of the same is highlighted below in Figure 3)

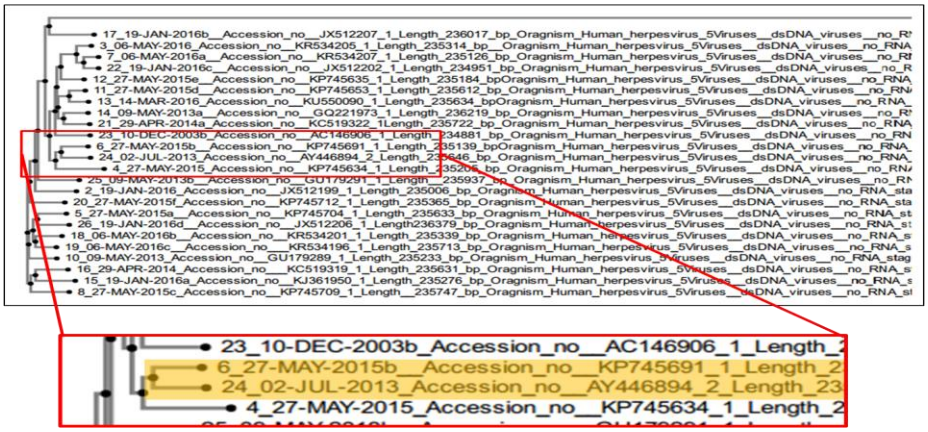


Figure 3: Latest and oldest sequence taken from the phylogenetic tree

5.3.2 Di-nucleotide Frequencies of both the nucleotides among the 4 Sequences

Fasta formats of the selected mono-nodal sequences were fed in to the Computation of Chaos Game Representation of frequencies (FCGR) Tool for the upper strand. The tool provided the frequency of Di-nucleotides (AA, CA, GA *etc.*) occurring in the sequence was found. In case of dinucleotide frequency Y_pZ $\left[(O/E)_{YpZ}\right]$ in single stranded sequences the observed/expected ratios (O/E) are calculated by using following formula:

$$(O/E)_{YpZ} = [f(YZ)/f(Y)f(Z)] * G$$

Where,

$f(YZ)$ = observed frequency of the dinucleotide

$f(Y)F(Z)$ = frequencies of the mononucleotides $f(Y)$ and $f(Z)$

G = length of the genome

5.3.3 Analysis of the CG dinucleotide loss using odds ratio (OR)

An indirect method was used for determine the CG loss in a given viral genome over a defined period of time. For calculation of CGs loss or gain, pair wise alignment of both (old and new) sequences from the same node (common ancestors) were done using EMBOSS Matcher Tool. For further analysis, the aligned sequences were converted into vertical arrays using following steps:

1. The aligned sequences were copied to notepad then exported to an Excel sheet such that the bases occupied separate cells as shown in Figure 4.

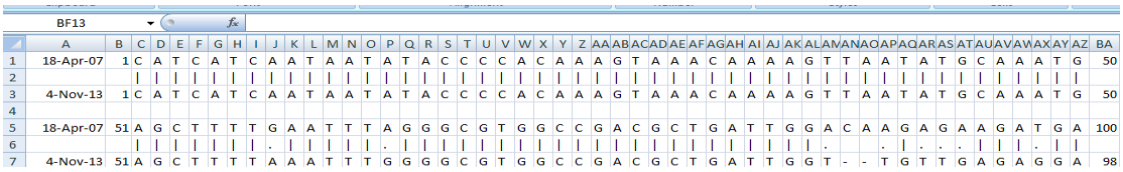
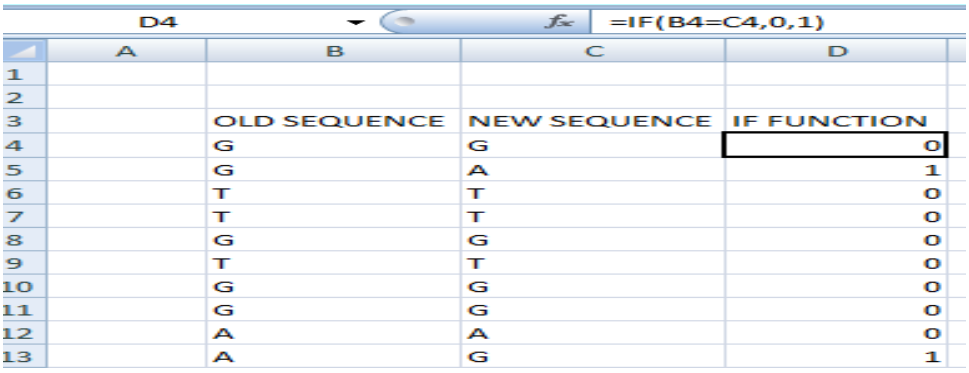


Figure 4: In aligned sequence the bases occupied separate cells

2. The sequence in each row was converted into one continuous string for characters using CONCATENATION function of MS Excel as shown in Figure 5. The concatenation of sequences of the two genomes (Old and new isolates) was performed in distinct adjacent columns

Figure 5: Old and New isolates in distinct adjacent columns (Column BN and BO)

3. Old and new sequence strings were copied separately and pasted to two different files in Notepad++.
4. Spaces in the sequences were removed after joining the strings using the JOIN LINES (ctrl J) line operations function in Notepad++. This resulted in a single continuous string of sequence with intermittent gaps shown as "–".
5. Both the sequences (Old and New) was converted into a vertical string by running a recorded macro which consisted of two following steps:
    a. Moving cursor one step right
    b. Enter
6. The vertical sequence strings of Old and New sequences were copied and pasted in adjacent columns in MS Excel sheet.
7. Vertical sequence strings of Old and New sequences are compared for corresponding nucleotide and if they are same then 0 was awarded 1 using IF operation as shown in Figure 6.



	A	B	C	D
1				
2				
3		OLD SEQUENCE	NEW SEQUENCE	IF FUNCTION
4		G	G	0
5		G	A	1
6		T	T	0
7		T	T	0
8		G	G	0
9		T	T	0
10		G	G	0
11		G	G	0
12		A	A	0
13		A	G	1

Figure 6: Old and new sequences are compared for corresponding nucleotide and if they are same then 0 else 1 awarded.

8. For vertical sequence string of Old sequence we find out -3,-2,-1 position in vertical direction, then +1 position in horizontal direction and then +1,+2 and +3 position in vertical direction. Name the centre most two columns as MM (bases involved mutation) as shown in Figure 7.

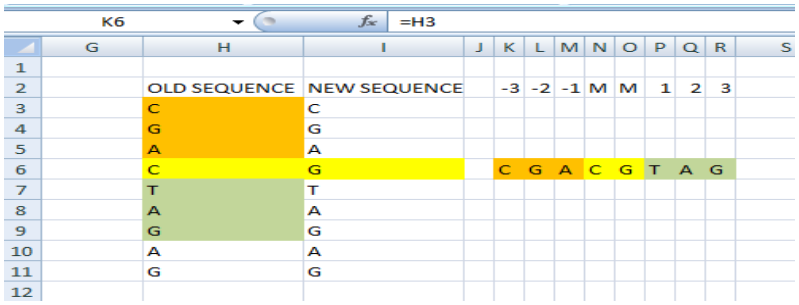


Figure 7: -3,-2,-1 position in vertical direction, +1 position in horizontal direction and +1, +2 and +3 position in vertical direction as reference of old sequence

9. Using CONCATENATE operation for -3,-2 and -1 position, MM columns and +1, +2 and +3 positions as shown in Figure 8.

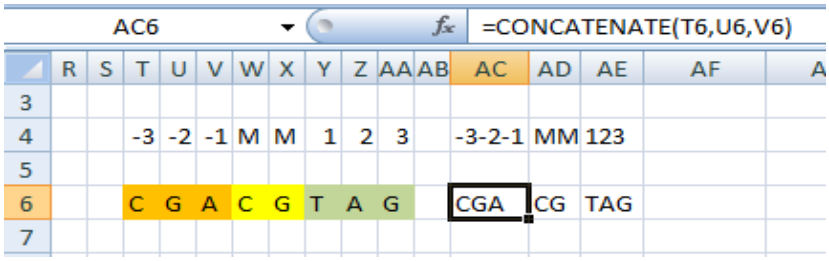


Figure 8: CONCATENATE operation for -3,-2 and -1 position, MM columns and +1, +2 and +3 positions.

10. Copy and paste concatenated sequences and remove gaps (-).
11. Using LEFT operation for -3,-2 and -1 concatenated column and RIGHT operation for +1, +2 and +3 concatenated column.
12. Using CONCATENATE operation for the column of LEFT operation, RIGHT operation and MM column.
13. Counting of the mutations was done using COUNTIF formula in excel. All possible substitutions G/A, G/T, G/C, A/G, A/T and A/C with all possible flanks at +1 and -1 position. This is in order to study mutation involving/not involving CG dinucleotides. N (N/N) N where N/N both are not same.

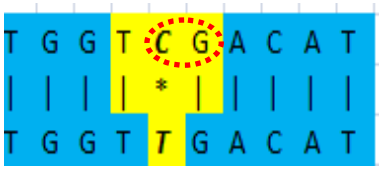
    Example

    AC/TG means A**CG** →A**TG** (it is a CG→TG mutation)

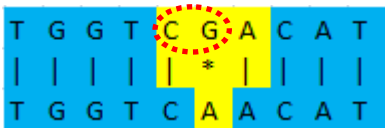
    CG/CA means **CG**A →**CA**A (it is a CG→CA mutation)

Example of

1. Mutation involving CG

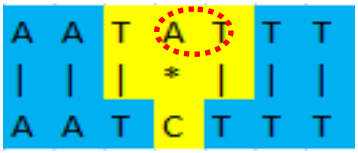


2. Mutation involving CA



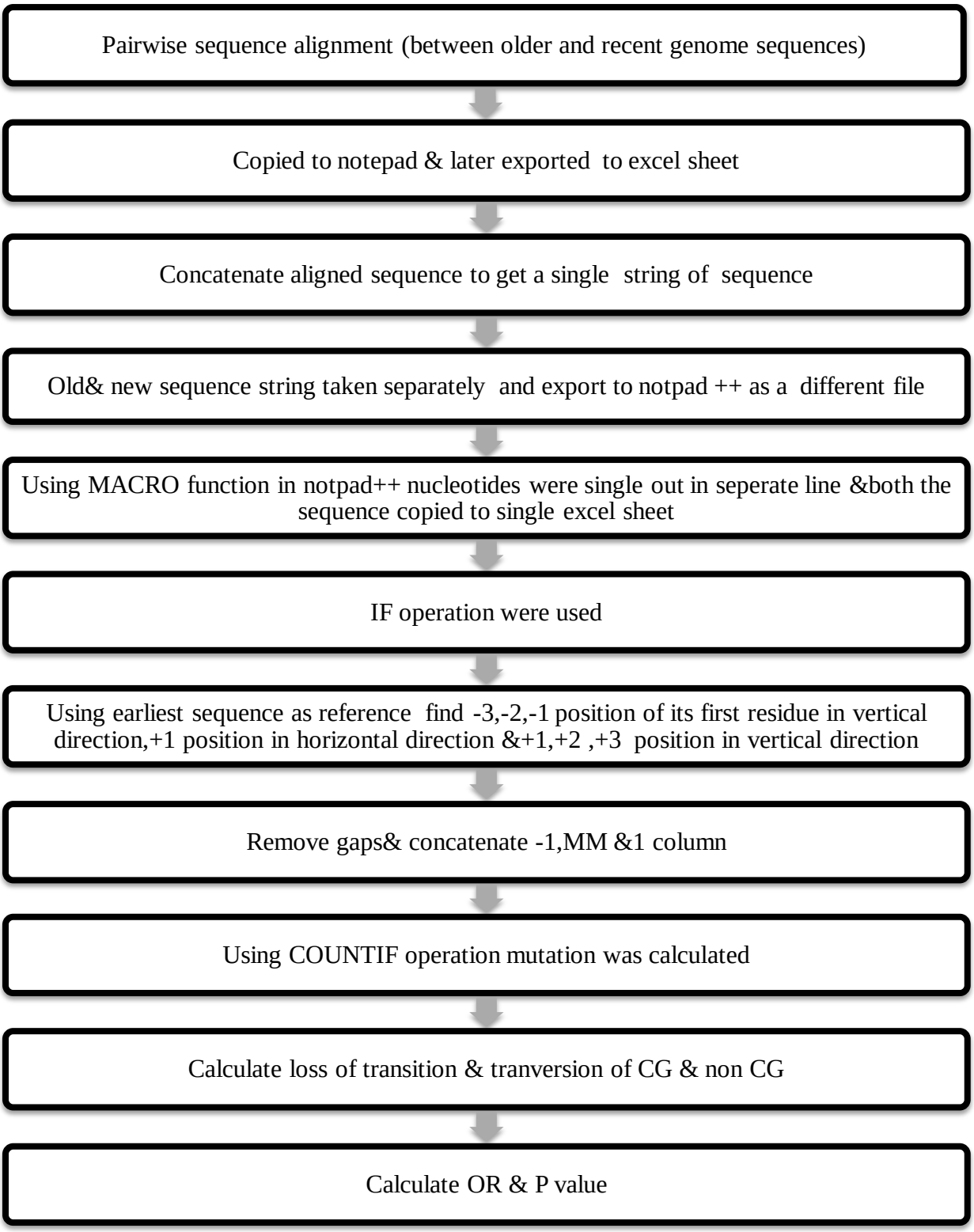
Example of

Mutation involving non-CG



14. Calculate the loss of transition and transversion CG.
15. Calculate the loss of transition and transversion non - CG.
16. OR and P-value (1 tail test) of the loss CG and non-CG were calculated using 2x2 Contingency Table Tool (http://vassarstats.net/odds 2x2. html)

Flowchart of algorithm



5.3.4 Analysis the +1 and -1 flanks of the CG/CG-TG/CA mutation frequency

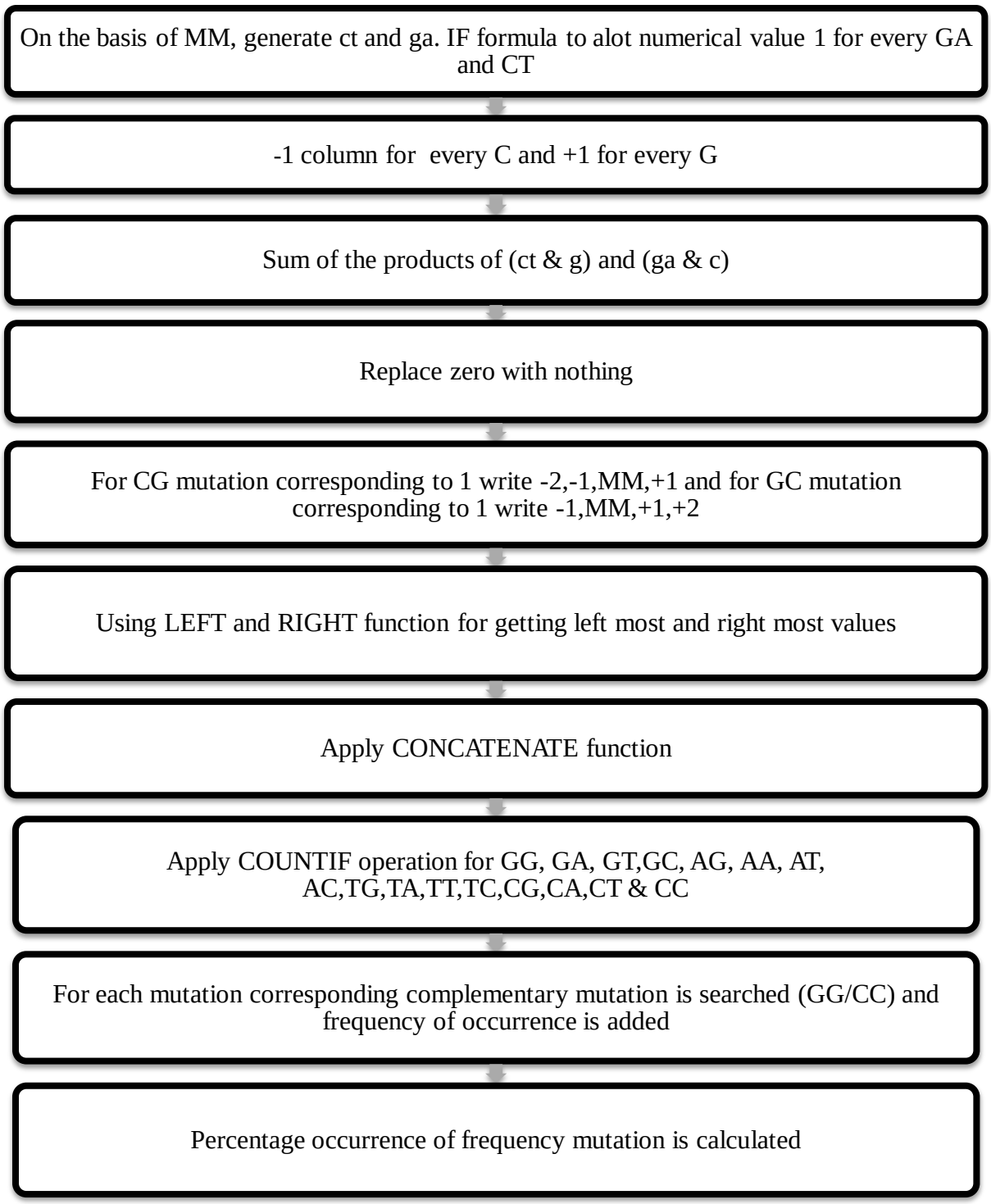
1. Continuing from previous algorithm, take the values of MM column, +1 and -1 column.
2. Numerical value of 1 for every CT and GA present in the column MM using IF operation in excel under the respective columns of ct and ga.
3. Column c and g had values for every C present in -1 column and G in every +1 column respectively.
4. Sum of products of column (ct and g) and (ga and c) were calculated.
5. In the column of sum of products, replace zero with nothing.
6. In the column of sum of products, for every CG mutation corresponding to 1 write -2,-1,MM,+1 and for every GC mutation corresponding to 1 write -1,MM,+1,+2 and name this column x.

   E.g.

A	CC	G	=	ACGAC
C	GA	C		

7. Using the LEFT and RIGHT operation in column x, two new column y and z are formed each consisting of left most and right most values of column x respectively.
8. CONCATENATE operation was applied for y and z in column u.
9. COUNTIF operation was applied to the column u for GG, GA, GT,GC, AG, AA, AT, AC,TG,TA,TT,TC,CG,CA,CT and CC.
10. After applying COUNTIF operation, for each mutation their corresponding complementary mutation is searched (GG/CC), (GA/TC), (GT/AC), (AG/CT), (AA/TT), (TG/CA) and frequency of occurrence was added (GG+CC), (GA+TC), (GT+AC), (AG+CT), (AA+TT), (TG+CA).
11. The Percentage occurrence of frequency mutation was calculated.

Flowchart of algorithm



CHAPTER 6

RESULTS

6.1 Finding the earliest and latest nucleotide sequence of 4 viruses:

For earliest and latest nucleotide sequence of HBV virus we used Devesa and Pujol, 2007 research paper as shown in Table 4.

Table 4: Earliest and latest nucleotide sequence in HBV

Virus	Accession No.	Date
	D50521	15-Nov-1997
HBV	D00331	20-Jul-2007
	AF223963	17-Jan-2001
HBV	AB166850	9-Feb-2010

For earliest and latest nucleotide sequence of EBV, cytomegalovirus and adenovirus we draw a phylogenetic tree. For adenovirus we draw phylogenetic tree from Clustal omega whereas for EBV and cytomegalovirus we draw phylogenetic tree from MAFFT.

Table 5: Earliest and latest nucleotide sequence of EBV, cytomegalovirus and adenovirus

Virus	Accession No.	Date
	AJ854486.1	14-Nov-2006
Adenovirus	KF268201.1	31-Dec-2013
	EF153473.1	18-Apr-2007
Adenovirus	JN226763.1	4-Nov-2013
	AY446894.2	2-Jul-2013
Cytomegalovirus	KP745691.1	27-May-2015
	AY961628.3	6-Jan-2006
EBV	KT273948.1	11-Jan-2016

6.2 Di-nucleotide frequencies of earliest and latest nucleotide sequence of 4 viruses:

Di-nucleotide frequencies and mono nucleotide are calculated by http://www.biophp.org/minitools/chaos_ game representation/demo.php?FCGR tool. The $(\boldsymbol{O/E})_{\boldsymbol{YpZ}}$ for dinucleotide frequencies in a single stranded sequence is calculated by

$$(O/E)_{YpZ} = [f(YZ)/f(Y)f(Z)]$$

Table 6: $(\boldsymbol{O/E})$ for CG, TG & CA frequencies in a single stranded sequence

Viruses	Accession no	$(\boldsymbol{O/E})_{\boldsymbol{CG}}$	$(\boldsymbol{O/E})_{\boldsymbol{TG}}$	$(\boldsymbol{O/E})_{\boldsymbol{CA}}$
	D50521	0.54	1.15	1.17
HBV	D00331	0.57	1.15	1.15
	AF223963	0.54	1.17	1.18
HBV	AB166850	0.53	1.18	1.18
	AJ854486.1	0.86	1.17	1.09
Adenovirus	KF268201.1	0.86	1.18	1.09
	EF153473.1	0.85	1.17	1.10
Adenovirus	JN226763.1	0.85	1.18	1.10
	AY446894.2	1.19	1.03	1.06
Cytomegalovirus	KP745691.1	1.19	1.03	1.06
	AY961628.3	0.62	1.10	1.17
EBV	KT273948.1	0.61	1.11	1.17

In small genome virus HBV and adenovirus(O/E) of CG is under-representative and(O/E) of TG and CA is over-representative. In large genome virus cytomegalovirus (O/E) of CG is over-representative and(O/E) of TG and CA is under-representative except in EBV where CG is under-representative and(O/E) of TG and CA is over-representative. The result of CG suppression in the four viral genomes is in agreement with the previous reports (Hoelzer *et al.*, 2008).

6.3 Odds ratio of CGs loss in different virus genomic sequence pairs differing in time:

All the substitutions in the pairwise sequence alignment were classified as transitions or transversion. In further classification substitutions of both the classes were divided into two categories, involving CG dinucleotides and those which involved any other dinucleotides.

	Transitions	Transversion
CGs	CG→TG or CA (mutation likely due to C^5 methylation)	CG→CC or CT or GG or AG (mutation unlikely to involve C^5 methylation)
Non-CGs	All transition but not involve CG, therefore unrelated to C^5 methylation	All transversion but not involve CG, therefore unrelated C^5 methylation.

To evaluate the odds of CG mutating to TG or CG, which is likely to be caused by methylation against single nucleotide mutation to other dinucleotides while considering the ratio of transitions to transversion, odds ratio was calculated for all the 6 pairwise sequence alignments. Odds ratio was expected to reveal the relative odd of transition due to methylation of CG in comparison to all other transitions.

CGs loss odds ratio and p-value is calculated by 2x2 contingency table for 6 pairwise alignments of all 4 viruses.

HBV (D50521/ D00331)

	Transition	**Transversion**
CG	10	6
Non-CG	116	72

OR = 1.034 **p-value** = 0.59

HBV (AF223963/ AB166850)

	Transition	Transversion
CG	5	7
Non-CG	115	66

OR = 0.410 **p-value** = 0.12

Adenovirus (AJ854486.1/ KF268201.1)

	Transition	Transversion
CG	48	40
Non-CG	380	292

OR = 0.922 **p-value** = 0.40

Adenovirus (EF153473.1/ JN226763.1)

	Transition	Transversion
CG	45	43
Non-CG	390	386

OR = 1.035 **p-value** = 0.48

Cytomegalovirus (AY446894.2/ KP745691.1)

	Transition	Transversion
CG	372	150
Non-CG	2260	824

OR = 0.904 **p-value** = 0.41

EBV (AY961628.3/KT273948.1)

	Transition	Transversion
CG	74	43
Non-CG	347	238

OR = 1.180 **p-value** = 0.24

Table 7: OR and p-value of HBV, Adenovirus, Cytomegalovirus and EBV

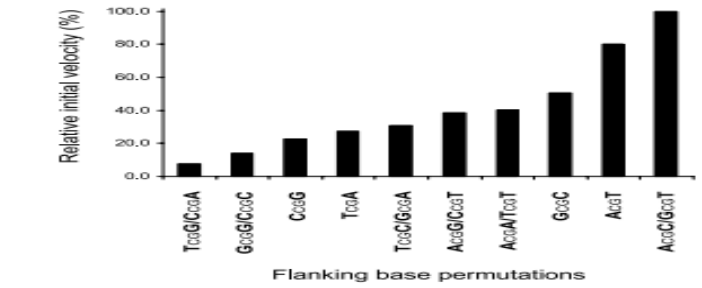
Virus	Accession no.	OR	p-value
HBV	D50521/ D00331	1.034	0.59
HBV	AF223963/ AB166850	0.410	0.12
Adenovirus	AJ854486.1/ KF268201.1	0.922	0.40
Adenovirus	EF153473.1/ JN226763.1	1.035	0.48
Cytomegalovirus	AY446894.2/KP745691.1	0.904	0.41
EBV	AY961628.3/KT273948.1	1.180	0.24

None of the sequence comparisons exhibited large OR value. Moreover p-value of all the OR was very high. This shows that the results are not significant in any of the 4 viruses. It implies that CG loss to TG or CA (transitions) over the mentioned time gaps cannot be proven to be significantly greater than loss of any other nucleotides which do not involved DNA methylation.

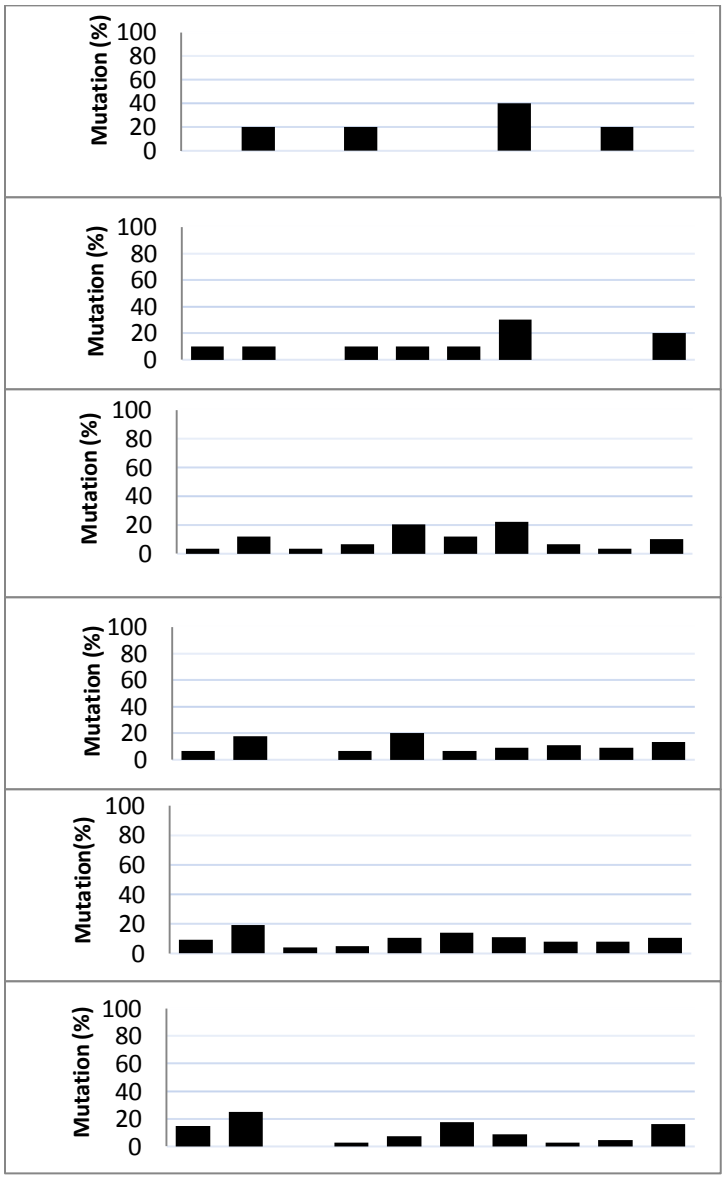
6.4 Effect of flanking bases on CG loss in viral genomes

There are reports that flanking bases around CG sites influence DNA methylation. To great extent it is attributed to the DNA methyltransferases' preference for the flanking bases in the substrate (CG dinucleotide). It has been shown that de novo methyltransferases prefer to methylate RCGY while YCGR dinucleotides exhibit poor methylation. Since methylated CGs only mutate to TG/CA, similar effect was expected to percolate in mutation phenomenon also. To investigate the effect of flanking bases at +1 and -1 position of CGs on

methylation and eventually on CG loss to TG/CA, all the CG/CG → TG/CA substitutions in pairwise sequence alignments were classified into ten possible flanks as shown in Figure 9.



Relative initial velocity of Dnmt3a measured with DNA substrates containing the CG site flanked given by Handa and Jeltsch *et al*., 2005



Mutation (%) in HBV (D50521/ D00331)

Mutation (%) in HBV (AF223963/ AB166850)

Mutation (%) in adenovirus (AJ854486.1/KF268201.1)

Mutation (%) in adenovirus (EF153473.1/ JN226763.1)

Mutation (%) in cytomegalovirus (AY446894.2/ KP745691.1)

Mutation (%) in EBV (AY961628.3/ KT273948.1)

Figure 9: Mutation % in HBV, Adenovirus, Cytomegalovirus and EBV.

The frequency of CG/CG → TG/CA substitution for each flank was compared against the methylation preference (Relative initial velocity of the DNA methyltransferases). For all the viral genomes, no discernible similarity between methylation preference of *de novo* methyltransferases and CG substitutions was observed for the ten flanking sequences (Figure 9).

6.5 CG loss analysis in HBV using multiple sequence alignment

Since no significant odds ratio were obtained for any of the four viral genome comparisons and we failed to infer if there were statistically significant loss of CGs due to DNA methylation, we tried another approach in one of the small viruses *i.e.* HBV virus. Multiple sequence alignment (MSA) of randomly selected 250 HBV genomic sequences selected randomly was analyzed for CG substitutions. It was observed that 389 positions in the MSA were occupied by CGs. None of the 250 sequences had CG>103 and average number of CG were ~97.3. The inflated figure of 389 positions exhibits the great extent to which CG is subjected to mutation.

For all the CG position in MSA, corresponding TG and CA frequencies were also determined. The CG positions where CG had the largest frequency in comparison with all the remaining dinucleotides were considered for the analysis. Such 52 positions CG frequencies were compared to those of TG and CA.

"E" (Total sequences) 250-CG frequency were considered as loss of CG to some other dinucleotides due to mutation. Since very large number of sequence as shown in table 8 have CG in comparison to rest of the 15 dinucleotide. It was assumed that the mutation directions were from CG to some other dinucleotide.

Table 8: Significant mutation of CG

Position (In MSA)	CG	TG	CA	Other remaining nucleotide	p-value
113	233	15	0	2	1.03753E-19
146	249	1	0	0	0.01078745
164	239	8	2	1	3.76691E-14

541	246	3	1	0	3.41417E-07
616	241	8	0	1	2.59404E-11
620	245	3	1	1	1.15831E-05
842	247	2	1	0	1.006E-05
1167	247	0	2	1	0.006578414
1229	249	0	1	0	0.01078745
1258	239	8	0	3	6.83776E-09
1393	243	6	0	1	1.76594E-08
1421	249	0	1	0	0.01078745
1452	232	0	10	8	1.36683E-07
1530	232	0	17	1	4.35576E-24
1545	249	0	1	0	0.01078745
1609	247	3	0	0	1.006E-05
1616	247	0	2	1	0.006578414
1635	249	1	0	0	0.01078745
1771	249	1	0	0	0.01078745
1839	245	3	1	1	1.15831E-05
2133	237	0	9	4	3.05122E-09
2502	242	8	0	0	5.55006E-13
2604	249	0	1	0	0.01078745
2624	230	0	17	3	4.16496E-21
2697	243	0	6	1	1.76594E-08
2735	247	0	2	1	0.006578414
2738	249	0	1	0	0.01078745
3204	245	0	3	2	0.002142717
3262	246	2	1	1	0.000285461
3382	243	0	7	0	1.52639E-11
3434	233	5	3	10	0.000103217
3482	245	3	0	2	0.002142717
3636	223	0	20	7	1.62206E-20
3540	249	1	0	0	0.01078745

3628	241	0	9	0	2.03291E-14
3762	201	1	19	29	1.51946E-08

Now 250-CG frequencies (the lost CG) were classified into TG or CA and rest of the 13 dinucleotide were compared to their respective expected frequency.

	TG+CA	Rest 13 nucleotide
Observation	TG+CA	E-TG-CA
Expected	$\frac{\mathbf{2}}{\mathbf{15}} \times \boldsymbol{E}$	$\frac{\mathbf{13}}{\mathbf{15}} \times \boldsymbol{E}$

E represents frequency of all the dinucleotide except at particular positions.

Since TG and CA are being consider out of all the 16 dinucleotide except CG=2 out of (16-1)

Chi square test was performed to find if the observation ratio is significantly different from expected.

Total position having ≥1 CG = 389

Total position having consequence = 52

CG mutated to TG/CA significantly more than rest 13 dinucleotide= 36

CG mutated TG/CA not significant more than 13 dinucleotide=16

$$\frac{\textbf{CG mutated to TG/CA significantly more than rest 13 dinucleotide}}{\textbf{Total position having consequence}} \times \mathbf{100}$$

$$\frac{\mathbf{36}}{\mathbf{52}} \times \mathbf{100} = 69.230$$

There are 69.23 % chance that CG is mutated to TG/CA more than 13 dinucleotide.

CHAPTER 7
DISCUSSION

The nucleotide composition significantly varies in DNA viruses of animals, but evolutionary pressure and biological mechanism lashing these patterns are indistinct. Viral genome susceptibility to methylation and immune recognition might have been affected by the location of replication within the cell and the specific intracellular trafficking route.

The genomes DNA viruses that infect vertebrates have been reported to undergo methylation at 5-position of cytosine in CG dinucleotides. Similar to their host genomes, viral genomes are also considered to lose CGs due to deamination of methylated cytosine. One methyl cytosine would cause loss of two CpG and add of one TpG and CpA, which could be a reason for CpG deficiency and TpG and CpA excess (Bird, 1980). Under-representation of CG relatively over representation of TG and CA has been studied well in viruses. The frequency of CpG, in small DNA viruses is under-representated whereas in large DNA viruses it is not under-representated (Karlin *et al*., 1998 & Hoelzer *et al*., 2008). Among viruses of the same family with similar genome organization and life cycle, the relative abundance of CpG dinucleotide is dependent on the infected host cells. In the infected virus genome there is strong connection between the evolutionary lineage of the infected host and the extent of reduction of CpG dinucleotide (Upadhyay *et al*., 2013). Since viruses have very short period of reproductive life cycle and higher rate of evolution, we should expect observable changes in genome over relatively shorter time periods. The changes are based on host compatibility. We attempted to study Human viruses causing common diseases and having DNA genome or DNA intermediate in there genome replication. They have been selected to study changes in genome structure over stipulated time. The sequences of viral isolates of a subtype with maximum possible time gap were compared by pairwise sequence alignment. There are various types of dsDNA viruses such as Adenovirus, Papillomaviridae, Hepatitis B virus (HBV), Smallpox virus, varicella-zoster virus, Epstein-Barr virus (EBV), Herpes simplex virus, cytomegalovirus *etc.* We are focusing on HBV, EBV, Adenovirus and Cytomegalovirus because these are

double stranded viruses which infect humans, an organism which has methylated genome.

Table 9: Brief description about HBV, EPV, Adenovirus & cytomegalovirus

VIRUS NAME	GENOME SIZE	DISEASES CAUSED
HBV	3.2 kb	Scarring of the organ, Liver failure, and Cancer
EBV	170kb	Burkett's lymphoma , Hodgkin's lymphoma, Gastric cancer and Lymphomatoid granulomatosis
Adenovirus	36kb	Upper respiratory infection, lower respiratory infection such as pneumonia
Cytomegalovirus	236kb	CMV mononucleosis , Gastrointestinal problems, Pneumonitis, CNS complications

The plan of the study was based on the logic that above double stranded DNA genome viruses infects humans and expected to undergo methylation (5mCG) and in turn some of the methylated CGs may mutate into TG/CA. Virus have high evolution rate, we may expect observable changes in genome over relatively short period of time, *viz.* few to several years.

We find the genome sequence of 12 isolates of the 4 viruses different at significance time period get belonging same evolution lineage. So the genome sequences of isolates of same or very similar evolutionary lineage (based on same subtypes or phylogenetic analysis) but varying in time of isolation was subjected to pairwise sequence alignment. Such 6 alignments were used to search for various types of substitutions. These substitutions were classified as:

CG→TG
CG→CA
} CG to TG/CA a mutation likely due to methylation

CG→AG
CG→GG
CG→CC
CG→CT
} CG loss but not due to methylation

Any other dinucleotides → any dinucleotides but different only at one of the two possible position by a transition. (TG →TA)

Any other dinucleotides → any dinucleotides but different only at one of the two possible position by a transversion. (CA→GA)

In small viruses (HBV and adenovirus) the CG is under-representative and TG and CA is over- representative. In large DNA viruses (cytomegalovirus) the CG is not under-representative except in EBV where CG is under-representative. An indirect method was used for determine the CG loss in a given 4 viral genome over a defined period of time. We find that in some virus there is loss of CG and in some virus there is gain of CG. Those viruses in which loss is CG found, we can say that there is gain of TG/CA due to mutation. Those viruses in which gain of CG was found, we can say that during evolution of those virus genome had that much lost of CG that they cannot further loss their CG and they started to have gain in CG. This data is strongly favored by calculating OR and p-value.

Earlier reports claim DNA methylation in viral genome is effected by flanking bases. We find the CG methylation flanks which have a high probability to get mutated. Higher rate of mutation is applied to our data; we look for mutation in different flanks. We compare our result with the flanks prefer by DNMT. But our result not conforming to the flanks prefer by DNMT. It simple mean the certain factor have more dominant role for which CG get mutated than their flanks. We have negative result but good observation.

The mutation of CG in 4 virus genome did not give favorable result. So we try novel approach for calculating the CG mutation in MSA of HBV genome. We took HBV genome because its genome size *viz.* 3.2 kb is very small as compare to other DNA virus. It was observed that 389 positions in the MSA were occupied by CG, which exhibits the great extent to mutation. Out of which CG mutated to TG/CA significantly more than rest 13 dinucleotide was found to be 26.We found the % conversion of CG mutated to TG/CA was 69.23. So from above result we say that in evolution of virus genome CG is mutated to TG/CA.

CHAPTER 8
CONCLUSION

In small DNA genome the CG is under-representative and TG and CA is over-representative. In large DNA genome the CG is over- representative except EBV where CG is under-representative.

We compared to observe different kind of mutation in genome sequence of 4 viruses isolates different at significance time period get belonging same evolution lineage. In genome of virus there is CpG deficiency because loss of CpG and gain to TpG and CpA due to mutation. Gain in CpG is due to entropy.

We compare -1 and +1 flanks of CG with the flanks prefer by DNMT. But our result not conforming to the flanks prefer by DNMT. It mean the certain factor have more dominant role for which CG get mutated than their flanks.

As earlier result not come significant we applied novel approach for finding mutation of CG to TG/CA in HBV genome. We find that there 69.23 chance that CG is mutated to TG/CA.

CHAPTER 9

REFERENCES

- 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.
- Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriological reviews*, *35*(3), 235.
- Bamford, D. H., Burnett, R. M., & Stuart, D. I. (2002). Evolution of viral structure. *Theoretical population biology*, *61*(4), 461-470.
- Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & development*, *16*(1), 6-21.
- 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*, *91*(9), 3799-3803.
- Cooper, D. N., & Youssoufian, H. (1988). The CpG dinucleotide and human genetic disease. *Human genetics*, *78*(2), 151-155.
- Devesa, M., & Pujol, F. H. (2007). Hepatitis B virus genetic diversity in Latin America. *Virus research*, *127*(2), 177-184.
- 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.
- Filée, J., & Chandler, M. (2008). Convergent mechanisms of genome evolution of large and giant DNA viruses. *Research in microbiology*, *159*(5), 325-331.

- 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.
- Forbes, D. A. (2012). What is a p value and what does it mean? *Evidence Based Nursing*, *15*(2), 34-34.
- Goldberg, A. D., Allis, C. D., & Bernstein, E. (2007). Epigenetics: a landscape takes shape. *Cell*, *128*(4), 635-638.
- Handa, V., & Jeltsch, A. (2005). Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome. *Journal of molecular biology*, *348*(5), 1103-1112.
- Handy, D. E., Castro, R., & Loscalzo, J. (2011). Epigenetic modifications basic mechanisms and role in cardiovascular disease. *Circulation*, *123*(19), 2145-2156.
- 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.
- Jin, B., Li, Y., & Robertson, K. D. (2011). DNA Methylation Superior or Subordinate in the Epigenetic Hierarchy? *Genes & cancer*, *2*(6), 607-617.
- Karlin, S., Campbell, A. M., & Mrazek, J. (1998). Comparative DNA analysis across diverse genomes. *Annual review of genetics*, *32*(1), 185-225.
- 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., Mocarski, E. S., & Schachtel, G. A. (1994a). Molecular evolution of herpesviruses: genomic and protein sequence comparisons. *Journal of virology*, *68*(3), 1886-1902.
- Karlin, S., Mrazek, J., & Campbell, A. M. (1997). Compositional biases of bacterial genomes and evolutionary implications. *Journal of bacteriology*, *179*(12), 3899-3913.

- Kim, J. J. (2001). Using viral genomics to develop viral gene products as a novel class of drugs to treat human ailments. *Biotechnology letters*, *23*(13), 1015-1020.
- Koonin, E. V., Dolja, V. V., & Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology*, *479*, 2-25.
- Li, E. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Reviews Genetics*, *3*(9), 662-673.
- 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.
- 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.
- McHugh, M. L. (2009). The odds ratio: calculation, usage, and interpretation.*Biochemia Medica*, *19*(2), 120-126.
- Meehan, R. R. (2003, February). DNA methylation in animal development. In*Seminars in cell & developmental biology*, 14(1) , 53-65.
- Moison, C., Guieysse-Peugeot, A. L., & Arimondo, P. B. (2014). DNA methylation in cancer. *Atlas of Genetics and Cytogenetics in Oncology and Haematology* , 18(4) , 285-292.
- Ollila, J., Lappalainen, I., & Vihinen, M. (1996). Sequence specificity in CpG mutation hotspots. *FEBS letters*, *396*(2-3), 119-122.
- 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
- Primrose, S. B., & Twyman, R. (2009). Principles of genome analysis and genomics. *John Wiley & Sons.*
- 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.

- Robertson, K. D. (2005). DNA methylation and human disease. *Nature Reviews Genetics*, *6*(8), 597-610.
- Scotia, N. (2010). Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry*, *19*, 227.
- Shackelton, L. A., Parrish, C. R., & Holmes, E. C. (2006). Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. *Journal of molecular evolution*, *62*(5), 551-563.
- 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.
- Upadhyay, M., Samal, J., Kandpal, M., Vasaikar, S., Biswas, B., Gomes, J., & Vivekanandan, P. (2013). CpG dinucleotide frequencies reveal the role of host methylation capabilities in parvovirus evolution. *Journal of virology*, *87*(24), 13816-13824.
- 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), 1-11.
- 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.

- Weissbach, A., Nalin, C. M., Ward, C. A., & Bolden, A. H. (1984). The effect of flanking sequences on the de novo methylation of CG pairs by the human DNA methylase. *Progress in clinical and biological research*, *198*, 79-94.
- 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.
- 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.