Effect of Epigenetics on Human Herpesvirus 4 genome

A Dissertation Report

Submitted in partial fulfilment of the requirement

For the award of degree of

Masters of Science

In

Biotechnology

Under the guidance of

Dr. Vikas Handa
Assistant Professor

Submitted by

Vardhan Chhabra

Roll no. 301601024

DEPARTMENT OF BIOTECHNOLOGY,

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY

PATIALA-147004

June 2018

CANDIDATE DECLARATION

I hereby declare that the research work being presented in M.Sc. project entitled "**Effect of Epigenetics on Human Herpesvirus 4 genome**" has been carried out by me during the period of January 2018 to June 2018, under the supervision and guidance of Dr. Vikas Handa, Assistant Professor, Department of biotechnology, Thapar Institute of Engineering and Technology, Patiala. Further, I declare that I have not submitted the matter embodied this dissertation for the award of any other degree or any other qualification of any other university or examining body in India/elsewhere.

Vardhan Chhabra

Date: 18 June, 2018

Vardhan Chhabra

Reg No. 301601024

M Sc. Biotechnology

Thapar Institute of Engineering and Technology,

Patiala- 147004

CERTIFICATE

This is to certify that this thesis entitled "**Effect of Epigenetics on Human Herpesvirus 4 genome**" submitted by Mr. Vardhan Chhabra (Roll. No. 301601024) in partial of the fulfilment requirement for the award of Master in Science Degree in Biotechnology from Department of Biotechnology, Thapar Institute of Engineering and technology, Patiala (Punjab), India.

It is an exclusive original record of candidate's own research work carried out by him under my supervision and guidance. This Thesis in part or full has not been submitted in any other institute for award of such kind of degree.

Dr. Vikas Handa

Assistant Professor

Department of biotechnology,

Thapar Institute of Engineering,

Patiala-147004

ACKNOWLEDGMENT

I am submitting my Thesis for the fulfilment of my M.Sc. degree. This work would not have been accomplished without the help, support and guidance of a large number of people. I express my deep gratitude and respect to my guide Dr. Vikas Handa, Assistant Professor, Department of Biotechnology for his strong motivation, trust and constant encouragement during the course of work. I thank him for his great patience, constructive criticism and for giving me the opportunity to undertake this project.

I also express my heartiest and special gratitude to Dr. Moushmi Ghosh, Head, Department of Biotechnology, for all her possible in various facilities of the department for this work. I am really pleased to acknowledge the kind help, cooperation which I have received throughout my dissertation from the entire teaching and non-teaching faculty member of Department of Biotechnology, which helped me a lot in completion of this work.

I am also really Thankful to Dr. Prashant Rana, Assistant Professor and Mr Rajesh Kondabala, Research Scholar, Computer Science and Engineering Department, for their guidance and valuable advice that helped me in completing this work.

With Heartiest reverence I admire the confidence bestowed on me by my parents. The untiring pains taking dedicated help, affection and blessings received from them to bring me to this level, it is beyond my capacity to express in words.

Vardhan Chhabra

Vardhan Chhabra

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

DNA Methylation is a process responsible for introducing epigenetic modification in a genome. Epigenetic modification plays important role in regulating many cellular processes including early embryogenesis, gametogenesis and cellular differentiation in the mammals. CG suppression has been reported in vertebrates. Viruses, whose host organisms are vertebrates, have also shown similar under-representation of CGs within their genomes because of their co-evolution with their hosts. We attempted to study genomes of Human Herpesvirus 4 causing infection in humans. Our data and results have shown that in Human Herpesvirus 4, CG dinucleotides are showing underrepresentation while TG and CA are being over-represented because of the corresponding gain from loss of CG. Herpesvirus 4 genome sequences were then subjected to Multiple Sequence Alignment to identify CG dinucleotide's conserved positions. Mutations of CGs were observed at conserved positions that provide us with strong evidence that CG mutations have been bias for TG/CA as compared to any other dinucleotide sequence so therefore it indicates a strong association with DNA methylation in Human Herpesvirus 4 genome.

Keywords: DNA methylation, Human Herpesvirus 4, CG suppression, Multiple sequence alignment.

CHAPTER 1
INTRODUCTION

Epigenetics is the study of heritable changes in gene function that do not encourage changes in the DNA sequence. “Epigenetics” term was coined by Conrad Waddington in the early 1942. According to him epigenetics is ‘‘the study of the interactions between genes and the products they form that bring the phenotype into being” (Goldberg *et al.,* 2007).

Epigenetics is defined as introducing changes in a chromosome which in result will affect the gene activity and its expression. Three processes that are responsible for introducing epigenetic modifications in a genome are - DNA methylation, Histone modification and non-coding RNA associated gene silencing.

1.1 DNA Methylation

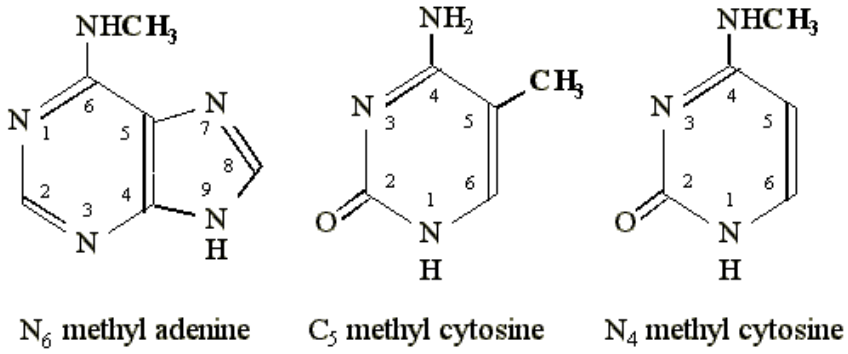
DNA methylation plays a vital role in regulation of gene expression. In this process methyl groups are added to the DNA molecule at N^4 and C^5 position of Cytosine and N^6 position of adenine(prokaryotes), whereas at C^5 position of Cytosine in eukaryotes (Hoelzer *et al.,* 2008).

DNA methylation is an epigenetic modification in which methyl ($-CH_3$) group is added to DNA to modifiy the function of the genes by affecting their expression. In higher higher eukaryotes, methyl group is added covalently at the 5-carbon of a cytosine which is followed by a guanine, resulting in a 5-methylcytosine. During DNA methylation epigenetic information is transmitted through multiple cycles of DNA replication and cell division (Hermann *et al*., 2004). DNA methylation plays a crucial role in the development of embryo, genomic imprinting, silencing of transposons and genetic diseases and cancer biology (Bird *et al.,* 2002). When a methyl group is added to cytosine, DNA methylation is controlled at various levels in cells and is performed by several enzymes called, DNA methyltransferase (DNMTs).

For DNA methylation to occur DNMTs (DNMT1, DNMT3a, DNMT3b & DNMT2) are responsible. However, DNMT2 is reported to methylate tRNA while DNMT1 is responsible for maintenance of DNA methylation by methylating hemimethylated DNA after DNA replication.

DNMT1 maintains the established patterns of DNA methylation, whereas DNMT3a and 3b are used for *de novo* DNA methylation patterns (Flynn *et al.*, 1998).

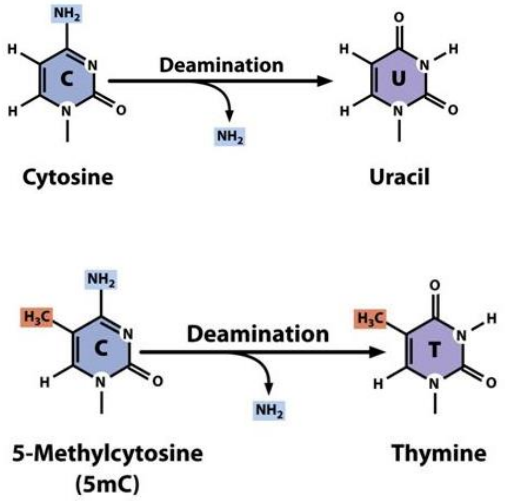
In eukaryotes, DNA Methylation occurs at C5 position of Cytosine, whereas in prokaryotes it occurs at N4 and C5 position of Cytosine residues and N6 position of Adenine residues, mainly in context of CG dinucleotides (Bird, 1980).



(Source: www.mikeblaber.org/oldwine/BCH4053I/Lecture10/Lecture10.htm)

Figure 1: Methylation positions at Cytosine and Adenine base

The cytosine-guanine (CG) dinucleotides are known to be a hotspot for pathological mutation in the human genomic sequences. To yield thymine 5-methylcytosine (5mC) goes under spontaneous deamination it is called hypermutability (Cooper *et al.*, 2010). DNA repair machinery of cell does not identify this mutation and because of this mutation is not usually repaired that makes this conversion irreversible.When unmethylated Cytosine goes for deamination its base composition results in formation of Uracil, which is recognized by cellular Uracil-DNA glycosylase pathway, and then it gets repaired (Fig. 3) (Cooper and Youssoufian, 1988 & Hoelzer *et al.*, 2008). So when cytosine is getting converted into thymine it induces mutations in the genome that does not get repaired and had to be left as it is which leads to under-representation of CG di-nucleotides and then their suppression in genome.



(Source: https://slideplayer.com/slide/8068761/)

Figure 2: Spontaneous deamination of Cytosine leads to Uracil whereas of methylated cytosine to thymine.

In mammals, distributions of CGs are uneven. It has been observed that in mammals, around 70% to 80% of CG dinucleotides are methylated.Those regions, where CGs are present in significant number, are called CG rich region and the other region where CGs are less in number are called CG poor region. In a genome, regions that have a high frequency of CGs are called as – "CG islands". These CG islands always remain unmethylated specially those that are associated to gene promoters (Jones *et al.,* 2009). Various studies and experiments have reported that vertebrate and invertebrates have shown under-representation of CG in their genomic sequences (Cardon *et al*, 1994). It's been a widely known that viral genomes have shows lower CG abundance, which indicates that these di-nucleotides gets methylated in viruses and plays a important role in the evolution of their genomes.

It has been observed that for the viral genome for their evolution co-evolve along with their host. When the di-nucleotides frequencies are compared and analyzed within an individual genome it can provide with information related to host and factors that helps in evolution of virus genome (Upadhyay *et al.,* 2014).

1.2 Human Herpesvirus 4

Human herpesvirus 4 (HHV-4) is also called The Epstein–Barr virus (EBV), HHV-4 is one of the eight human herpesvirus that are known. It is most commonly found virus in humans. The virus is approximately 122–180 nm in diameter and It has a double helix of DNA and is ~172,000 base pairs long and having 85 genes (Amon *et al.*, 2004). Double stranded DNA is surrounded by a protein nucleocapsid. This nucleocapsid is surrounded by a tegument made up of a protein, surrounded by an envelope containing lipids and surface projections of glycoproteins that helps in infecting the host cell (Odumade and Balfour, 2011).

Group:	Group I (dsDNA)
Order:	Herpesvirales
Family:	Herpesviridae
Subfamily:	Gammaherpesvirinae
Genus:	Lymphocryptovirus
Species:	**Human herpesvirus 4**(HHV-4)

Table 1 -Classification of Human Herpesvirus 4

Transmission and infection with HHV 4 occurs primarily via the oral transfer of saliva and genital secretions. Herpesvirus 4 causes infectious mononucleosis also known as glandular fever. It is also associated with cancer, gastric cancer, and several other conditions associated with human immunodeficiency virus (HIV) such as – central nervous system lymphomas and hairy leukoplakia.

Herpesvirus, when infects humans having methylated genomes, might get its genome also methylated by DNA methyltransferases present in the host cell. As a result of their co-evolution with the host genomes, the Herpesvirus is expected to have experienced suppression of CG dinucleotides. So the CG di-nucleotides of Herpesvirus 4 genome are expected to be less in number or under-representation when compared to other dinucleotides. The present work in this report is based on *in silico* analyses to investigate the effect of DNA methylation on Herpesvirus genome.

CHAPTER 2
REVIEW OF LITERATURE

2.1 DNA Methylation

DNA methylation is a covalent modification in which a methyl group is enzymatically transferred to the bases of genomic DNA and results in modulation of the gene expression. CG Methylation process is a dynamic system suitable for regulation and the development of the organism (Egger *et al.,* 2004). DNA Methylation follows a dynamic state in which two different methylation processes occurs that are: *De novo* Methylation and Maintenance Methylation.

De-novo methylation establishes the methylation patterns while the other one is required for copying those methylation patterns onto daughter strands of DNA (Hermann *et al.,* 2004).

Basic mechanism of DNA methylation can be divided into 3 parts-

- Writer
- Erasers
- Readers

Writers are those enzymes that catalyze the process of addition of methyl group onto cytosine. Erasers help in modifying and removing of methyl group. Readers recognize the methyl group and bind to them and help in gene expression. (Moore *et al.,* 2012)

2.2 Mechanism of DNA methylation

In DNA methylation cytosine base is covalently modified by adding a methyl group at Carbon-5 position with the help of DNMTs. Modification is mostly occurs at 5'-CG-3' di-nucleotide sequence at C-5 position of cytosine. Enzymes that are used for thus mechanism are called as DNA methyltransferase (DNMTs) SAM (S –adenosyl-L- methionine) is used as a methyl group donor for all DNMTs.

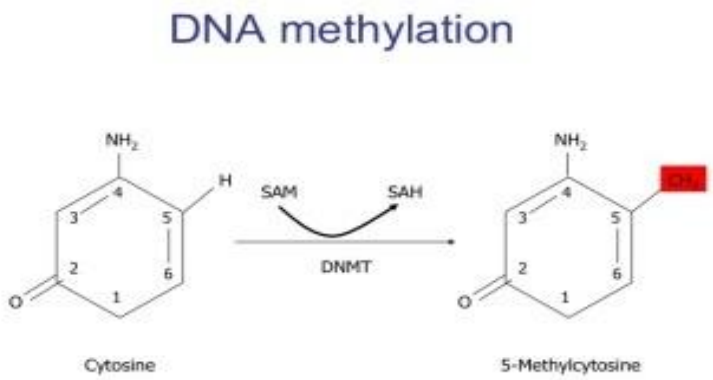


Figure 3: Methylation of Cytosine base carried out by DNA Methyltransferase (DNMT) using S adenosyl methionine as a methyl group donor.

When the methyl group is added it introduces the thermodynamic de-stability as it binds to sulphonium atom, turning it into a highly active atom to react with oxygen, nitrogen, sulphur and activated carbon atoms (Cedar *et al*., 2012). Mechanism of DNA Methylation was first analyzed for the prokaryotic HhaI methyltransferase. This enzyme recognizes only the 5'-GCGC-3' sequence and causes methylation in the very first cytosine of the sequence (Portela *et al*., 2010).

2.3 The DNTMs:

DNTMs family consists of 3 enzymes: DNMT1, DNMT 3a, and DNMT 3b. These enzymes help in catalyzing the process methyl group addition onto DNA. All these enzymes are having same structures but different functions and expression patterns. DNMT3a and DNMT3b have similar structure and function. When DNMT3a and DNMT3b are being over expressed they are capable of methylating the native and synthetic DNA both. Because of this reason, DNMT3a and DNMT3b are termed as *de novo* DNMT because of their ability of introducing methylation into naked DNA (Okano *et al*, 1999).

2.4 DNMT2 (DNA Methyltransferase 2)

DNMT2 also has a structure similar to the rest of the DNMTs enzymes. It plays a vital role in methylation of RNA that is why it can be referred as RNA Methyltransferase (Hermann *et al.*, 2004). DNMT 2 specifically methylates the Cytosine base at 38th position of tRNA. Because methylation of tRNA effects folding of protein and stability of its structure, therefore might have a protective function (Goll *et al*., 2006).

2.5 CG suppression

Most of the eukaryotic organisms have shown CG's suppression in their genome. Frequency suppression of CG is highly variable and negatively correlates with the presence of methylated Cytosine in the genome (Hoelzer and Shackelton, 2008). On the basis of base composition calculation, of expected frequencies only 25% of CG abundance is observed. The expected frequency of G+C content in human DNA is 0.4, whereas the frequency observed is about 0.008 (Bird, 1980).

CG dinucleotides have shown uneven distribution in the mammalian genome. Due to which there can be certain regions that might have higher frequency of CG sites are called as CG islands (Cardon *et al*., 1993). After the completion of genomic sequence analysis of human chromosome 21 and 22, CG islands can be re- defined as the DNA regions having >500 bp and GC content of >55% and Observed/Expected (O/E) CG dinucleotides ratio of 0.65 (Takai *et al*, 2002).

CG di-nucleotides represent only $1/3^{rd}$ to $1/4^{th}$ of the expected frequencies in vertebrate genomes. The main reason behind this is the higher stacking energy of Cytosine and Guanine as compared to Adenine and Thymine nucleotide bases.

Methylated Cytosine is known to show the high propensity of undergoing deamination which can also be referred as CG depletion. DNA repair machinery of cell can correct the transitions of unmethylated Cytosine into Uracil but transitions of methylated Cytosine to Thymine cannot be corrected which makes this an irreversible process (Jones, Liang. 2009). Cytosine when gets methylated it causes loss of two CG, converting into 1 TG and CA. This leads to under representation of CG/CG and overrepresentation of TG/CA (Bird, 1980).

2.6 Methylation and interactions between host and viral genome

The DNA viruses that infect vertebrates have shown significant variations in their base nucleotide composition (Hoelzer and Shackelton, 2008). Evolutionary studies, among viruses, can be done based on their differences in the relative abundance of di-nucleotides count. Among all the possible di-nucleotides most extensively studied di-nucleotide is CG whose depletion is reported in various viruses and their host genomes. Reasons for this decrease in CG counts can be inter-related with the evolution of viruses, mutations in di-nucleotides and DNA methylation. Apart from above all these factors, genome size of viruses and the type of genetic material (DNA/RNA) can also be a important factor in shaping the viral evolution (Upadhyay *et al*., 2015).

Earlier, Herpesvirus genome has also been reported for showing significant suppression in CG di-nucleotides and overrepresentation of CA/TG di-nucleotides relatively (Karlin *et al*., 1994). Methylation also plays important role in HBV gene expression by down regulating it. During chronic viral infection, DNMTs 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 DNMTs, as a result causing methylation of viral DNA and thus leading to inhibition of viral replication and expression of its genes (Vivekanandan *et al*., 2010).

Evidence which suggests role of methylation in the regulation of viral protein production is provided by the CG islands of HBV DNA which are methylated in the human tissue (Vivekanandan *et al*., 2009). CG 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 CG in comparison to Herpes Simplex virus which shows relative abundance (Vivekanandan *et al*., 2008). In contrast, Cytomegalovirus shows over representation of CG 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 CG 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).

CHAPTER 3
SCOPE OF STUDY

In this *in silico* analysis we want to see the effect of DNA Methylation on Herpesvirus 4 genome and its complete genome isolates. Earlier studies have established that viruses, that infect vertebrates, have shown CG dinucleotides suppression in their viral genome. In our case we are using a novel approach by taking advantage of comparative genome analysis of Herpesvirus 4. In this method complete genomic sequences of virus isolates, whose host is human, are selected for studying the effect of mutations due to DNA methylation and calculating the di-nucleotide freqencies in different genomic sequences to determine the di-nucleotide abundance in Herpesvirus 4 genome. Such type of analysis might help in determining the effect of DNA methylation on Herpesvirus 4 genome that could have occurred during the course of the evolution.

CHAPTER 4
OBJECTIVES

- DNA sequence analysis of various complete genomic isolates of Herpesvirus 4 genomes to study the effect of DNA methylation on CG dinucleotides.
- Comparative genomic analysis of Human Herpesvirus 4 genomes to study the effect of DNA methylation.
- Analysis of effect of methylated CG di-nucleotides causing mutations in Herpesvirus 4 genome.

CHAPTER 5
Materials and Methods

5.1 Retrieval of Sequences

DNA sequence of the Human Herpesvirus 3 was searched and downloaded in FASTA format from NCBI (National Centre for Biotechnology Information- www.ncbi.nlm.nih.gov/). For the same DNA sequence BLAST was performed. During BLAST, the maximum target sequences were set to 5000 to get maximum number of similar and complete genome isolates of Herpesvirus 3 that affect humans. Out of 2335 sequences a total of 69 sequences, that were complete genomes, were then selected. All the 69 sequences of selected strains were then downloaded in FASTA format.

For studying the effect of DNA methylation in the evolution of genome, Herpesvirus 3 was chosen as target organism. The criterion for selection includes:

- The host organism of virus should have a methylated genome.
- The virus should have a double stranded DNA genome.
- The virus should cause a commonly found disease in humans.

As it is a virus which causes commonly found disease in humans – Chickenpox and also it has a dsDNA that is 125kbp long, whole genomic sequences of large number of isolates and strains would be available and can be very well studied.

A total number of 69 complete genomic sequences of Human Herpesvirus 3 were used for analysis. Nucleotide frequencies of all 69 sequences were calculated by using online tool FCGR.

Nucleotide frequency calculation result did not show expected under-representation of CGs and over representation of TG & CA in HHV 3. So this led us to analysis of another virus, Human Herpesvirus 4.

DNA sequence of Human Herpesvirus 4 was searched from NCBI and downloaded in FASTA format. BLAST was performed. During BLAST, the maximum target sequences were set to 5000 to get maximum number of similar and complete genome isolates of Herpesvirus 4.

A total number of 51 complete genomic sequences of Human Herpesvirus 4 were used for further analysis. Following are the accession number of each genomic isolate of Human Herpesvirus4.

Table 2: Accession numbers of all 51 sequences

S.N.	Accession Number	S.N.	Accession Number	S.N.	Accession Number
1	DQ279927.1	18	KX674066.	35	AB850645.
2	MG021312.	19	MG021307.	36	AB850652.
3	AY961628.3	20	AB828191.	37	AB850658.
4	MG021308.	21	AP015015.	38	AB850660.
5	KF717093.1	22	LC150337.	39	AB850650.
6	MG021309.	23	LC150327.	40	AB850653.
7	NC_007605.	24	LC150741.	41	AB850651.
8	AB850654.1	25	LC150742.	42	AB850644.
9	MG021310.	26	LC150338.	43	AB850659.
10	MG021311.	27	LC137018.	44	AB850657.
11	KC207814.1	28	KX674064.	45	AB850649.
12	AP015016.1	29	LC150743.	46	AB850655.
13	MG021306.	30	LC149491.	47	AB850656.
14	MG021305.	31	AB850647.	48	AB850648.
15	KF373730.1	32	AB828190.	49	AB850646.
16	KP735248.1	33	HQ020558.	50	KX674067.
17	KC207813.1	34	AB850643.	51	AP015015.

5.2 Sequence Analysis Tools

5.2.1 Multiple Sequence Alignment (MSA)

DNA sequences of all the 51 genomic isolates of Herpesvirus 4 were aligned by performing Multiple Sequence Alignment. Tools that were used for MSA of all the sequences were CLUSTAL-W, CLUSTAL-OMEGA and CLUSTAL 2.1.

CLUSTAL-W: A multiple sequence alignment tool used for aligning nucleotide and protein sequences. It requires three or more sequences to calculate the global alignment.

CLUSTAL-OMEGA: It is a consistency based one of the fast multiple sequence alignment tool. It uses new HMM engine and has higher accuracy.

CLUSTAL 2.1: It is an updated version of CLUSTAL-X and CLUSTAL-W. It has higher accuracy and proficiency.

CLUSTAL-W and CLUSTAL-OMEGA could not perform MSA because of the large size of DNA sequence data (~172 kb each of 51 sequences). Later CLUSTAL-2.1 tool was used for Multiple Sequence alignment of all the sequences.

5.2.2 MS Excel spreadsheet

For further computational and statistical analysis on the sequence data that was obtained, Microsoft Excel was used. Various operations such as counting dinucleotide frequencies in the aligned sequences were performed using functions, shown in the table below:

Table 3: Microsoft excel functions

Function	Class	Function
CONCATENATE	Text	Allows different text values to join to form a combined single string
COUNTIF	Statistical	A statistical tool that is used for counting cells within a range of given specific criteria.
IF	Logical	A logical tool that is used to confirm if certain conditions are met or not.

5.3 Methods

5.3.1 Multiple Sequence Alignment

Multiple Sequence Alignment (MSA) of 51 sequences was performed by using CLUSTAL 2.1 multiple sequence alignment programming tool. Notepad++ and MS Excel were used to view the output file that was obtained after the alignment

.

Table 4: Software for output file

Software Used	File Extension Required
Clustal 2.1	.fas
Microsoft Excel	.xlsx
Notepad++	.txt

5.3.2 Analysis in MS Excel

For further computational and statistical analysis that was to be done in MS Excel, the data was converted in a format that can be worked upon MS Excel. Each aligned sequence was opened in Notepad ++ and Macro was run and sequence lines to 60 single strings. Process was repeated for each 51 sequences then all these vertically stringed sequences were copied to MS Excel worksheet.

Probabilitiy counts of expected CGs were calculated as follows:

$$P(CG) = P(G) * P(C)$$

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 CG in a given sequence:

= P(CG) x Length of Sequence

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

Probabilities of expected TG+CA and rest other dinucleotides (which include AG, AT, AA, AC, TA, TT, TC, CC, CT, GA, GT, GC, and GG) were also computed.

5.3.3 Data Fragmentation

Herpesvirus 4 has large genome of approx 172 kbp. After the MSA was performed, the file that was containing the 51 sequences was fragmented into ten parts for easy handling of the obtained data. The total length of the sequence obtained after alignment is 198433. Each of the 10 fragmented consists of 51 sequences of 19983 nucleotide length.

5.3.4 Dinucleotide frequency calculation

The number of A, T, G and C were calculated within the vertical columns having 51 sequences by using COUNTIF function (MS Excel).

5.3.5 Making di-nucleotide combinations

In MS Excel for analysis of the di-nucleotide frequencies, all possible dinucleotide combinations were then formed within these genomes by using CONCATENATE function. Observed and expected frequencies of all possible combinations were calculated.

All 16 possible di-nucleotide combinations were analyzed for calculation of frequencies, which were as follow:

Table 5: Possible di-nucleotide permutations in the multiple sequence alignment

Dinucleodtides with 'A'	**Dinuclodtides with 'T'**	**Dinuclodtides with 'G'**	**Dinuclodtides with 'C'**
AA	TT	GG	CC
AT	TA	GA	CA
AG	TG	GT	CT
AC	TC	GC	CG

5.3.6 Counting dinucleotide frequencies

'COUNTIF' is used by specifying the range of the horizontal cells that are having di-nucleotide combinations of the viral genome. Likewise, 'COUNTIF' is also used to compute the frequencies of all possible 16 combinations as stated above in the table.

5.3.7 Identification of CG positions in multiple sequence alignment

For each dinucleotide position in multiple sequence alignment, count of each possible dinucleotide was determined using COUNTIF function. The positions where >50% sequences had CG were selected for further analysis. This selection was performed to ensure that the ancestral alleles in the dinucleotide are CG only and occurrence of each dinucleotide is a result of mutation.

5.3.8 Chi-square test and calculation of p-value

Chi-test is performed on positions that having maximum CG count. P-values were calculated to check the level of significance. Parameters that need to be calculated were:

Sum of TG and CA frequency

Frequency of rest of the dinucleotides (excluding CG, CA, TG)

Sum of expected TG and CA frequency

Expected frequency of rest of the dinucleotides (excluding CG, CA, TG)

Observed and expected frequencies of TG + CA dinucleotides were compared with the remaining 13 di-nucleotide at max count CG position to perform Chi-square test for goodness of fit.

CHAPTER 6

RESULTS

Higher eukaryotes over the years have been reported to show under representation of CG di-nucleotide and overrepresentation of TG and CA di-nucleotides because of effect of the DNA methylation. Various reports related to analysis of CG di-nucleotides have shown genomic DNA of vertebrates and invertebrates are pervasively CG suppressed (Karlin *et al.*, 1993). As Vertebral hosts are infected by viruses it has been observed that viruses, that infects vertebrates and evolve with them, have also shown DNA methylation in their genomes leading to CG suppression (Galvan *et al.*, 2011).

6.1 Genome analysis

Human Herpesvirus 3 genomic sequence was downloaded from NCBI and subjected to nblast. Blast results were screened for complete genome sequences of Human Herpesvirus 3. The 69 complete genome sequences so found were analyzed for base frequencies by using FCGR tool.

A	B	C	D	E	F	G	H	I	J	K	L
	Accession number -	X04370.1	JN704693.1	JN704695.1	KC112914.1	JN704707.1	JN704708.1	JN704704.1	JN704705.1	JN704706.1	JN704696.1
	A	33792	33784	33785	33786	33779	33775	33770	33770	33760	33775
	C	29296	29296	29293	29311	29298	29283	29295	29298	29194	29293
	T	28178	28185	28166	28179	28172	28184	28164	28178	28184	28153
	G	33625	33625	33629	33616	33618	33610	33609	33606	33597	33627
	Sequence	Occurences	Occurences	Occurences	Occurences	Occurences	Occurences	Occurences	Occurences	Occurences	Occurences
	AA	10303	10302	10303	10299	10293	10293	10291	10294	10284	10301
	AC	8364	8366	8365	8354	8352	8352	8356	8353	8355	8364
	AG	5542	5537	5536	5541	5545	5543	5541	5553	5545	5531
	AT	9580	9578	9580	9590	9588	9585	9582	9570	9576	9579
	CA	8042	8033	8039	8040	8034	8026	8033	8029	8028	8038

Figure 4: Calculation of Nucleotide frequencies of 69 sequences using FCGR tool

The genome sequences did not show CG suppression and a corresponding over-representation of TG or CA. Therefore the focus of study was switched to Human Herpesvirus4.

Human Herpesvirus 4 infects humans and its genome has been analyzed to study the under-representation of CG di-nucleotides in this report. Genomic sequence of virus was searched from *Genome* database of NCBI. The sequence was subjected to nblast in search of genome sequence of more isolates. Total number of 51complete genomic sequences were selected and downloaded in FASTA format. All the 51 genomic sequences were analyzed to find the counts of the 4 bases (A, T, G & C) as well as 16 possible dinucleotides as shown in Table 2.

	A	B	C	D	E	F	G	H	I	J
1	**Accession number -**		**DQ279927.1**	**MG021312.1**	**AY961628.3**	**MG021308.1**	**KF717093.1**	**MG021309.1**	**NC_007605.1**	**AB850654.1**
2										
3										
4		**A**	34113	34127	33991	33974	32738	33981	33978	33987
5		**C**	51453	51435	51032	51035	49424	50993	51068	50809
6		**G**	51415	51409	51101	51087	49634	51104	51161	51410
7		**T**	35783	35793	35533	35597	34386	35627	35616	35416
8										
9		**N**	172764	172764	171657	171693	166182	171705	171823	171622
10										
11		**Sequence**	**Occurences**	**Occurences**	**Occurences**	**Occurences**	**Occurences**	**Occurences**	**Occurences**	**Occurences**
12										
13		**AA**	7208	7208	7201	7165	6916	7168	7214	7216
14		**AC**	8170	8165	8112	8130	7821	8119	8117	8128
15		**AG**	12132	12144	12056	12062	11672	12064	12026	12020

Figure 5: Calculation of Nucleotide frequencies of 51 sequences using FCGR tool

Based on the frequencies of bases expected frequency of dinucleotides (CG, TG and CA) was calculated and compared with the respective observed frequencies. Additionally observed frequencies were also compared with observed frequencies of other di-nucleotides with same base compositions respectively as shown in Table 5. This result showed that CG observed is less than CG expected as well as GC which has same base composition. Similarly corresponding higher frequencies were detected for TG obs and CA obs when compared to TG exp & GT and CA expected and AC respectively (Fig 6)

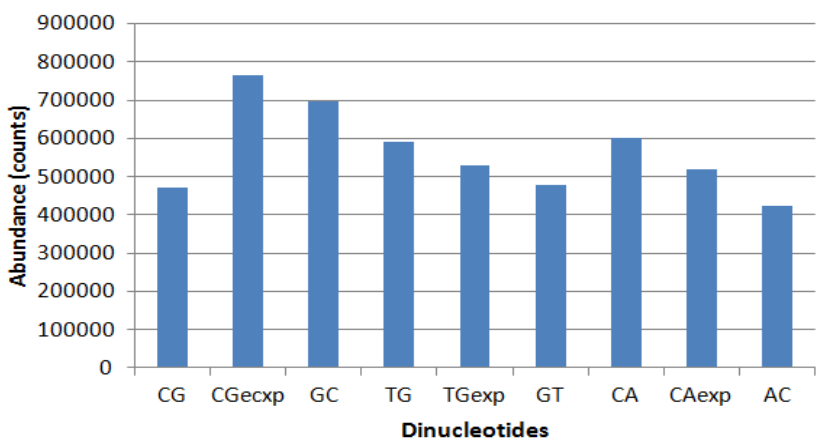


Figure 6: Comparison of observed and expected frequency of Di-nucleotides

Base composition frequencies of all the 51 sequences of Herpesvirus were computed for further genomic analysis.

Table 6: Mono-Nucleotide frequencies in All 51 Herpesvirus 4 genomes

Nucleotide	**Observed Frequency**	**P(N)**
G	2563641	0.294935
A	1739775	0.200153
T	1796661	0.206698
C	2592144	0.298214
Total	8692221	

Frequencies of all the possible 16 combinations of di-nucleotides were also calculated as shown in the table below:

Table 7: Di-Nucleotide Count in all 51 Herpesvirus 4 genomes

Di-Nucleotides	**Frequency**	**Di-Nucleotides**	**Frequency**
AA	373984	GA	490082
AC	423429	GC	697485
AG	605078	GG	899833
AT	337273	GT	477871
CA	602439	TA	273193
CC	938344	TC	533746
CG	470420	TG	589921
CT	581760	TT	399773

Observed and expected frequencies of CG, TG and CA were also compared with the computed frequencies of GC, GT and AC that were used as base control because of identical base composition. Below is a Box and Whisker graph showing comparison between the frequencies of the nucleotides.

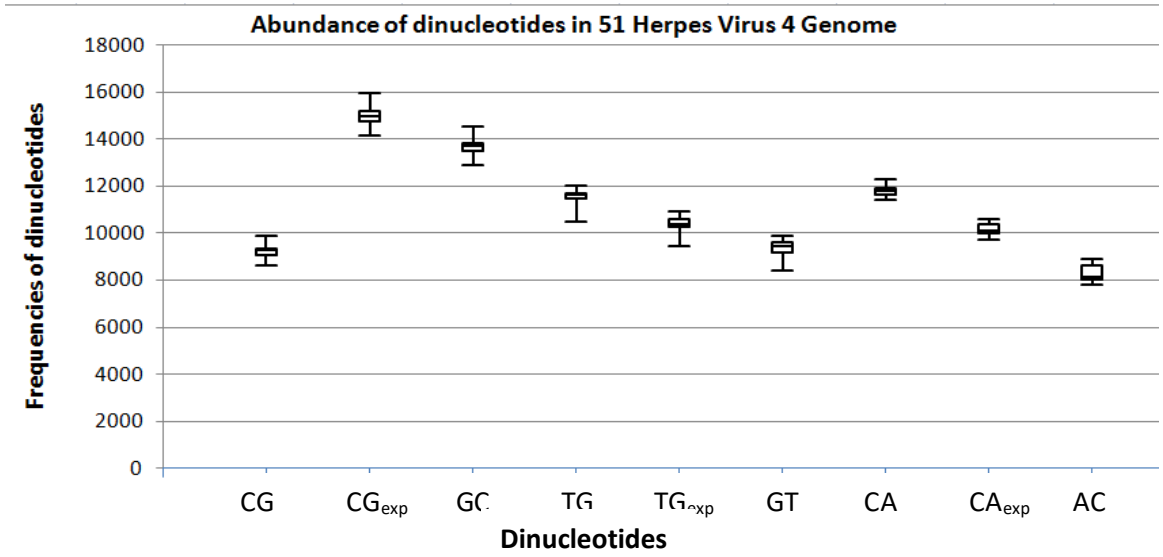


Figure 7: Box and Whisker plot showing comparison of observed and expected frequencies of CG, TG and CA with observed frequencies of GC, GT and AC

The graph shows each of the 51 Herpesvirus 4 genomes has shown lower frequencies of observed CG than the expected CG frequencies and observed GC frequencies. This confirms underrepresentation of CG based on base composition in each of the 51 virus genomes. Similarly in all the 51 genomes TG and CA were overrepresented as compared to their expected frequencies. Based on this it can be inferred that TG and CA are overrepresented whereas CG is underrepresented. This indicates that variation observed in abundance of these nucleotides can be partially due to mutation of CG/CG into TG/CA.

To determine, that the increase in count of TG and CA is due to the loss in CG count, frequencies of these three di-nucleotides were compared with other di-nucleotides of different sequences having identical base composition. CGs were found less in count than GCs where as TGs and CAs were higher in number than GTs and ACs respectively.

Observed/expected frequency ratios of CG vs GC, TG vs GT and CA vs AC were calculated and Chi- square test was performed and as p-value was almost zero significant difference was observed as shown in the table below

Table 8: Observed/Expected ratios of di-nucleotides and p-values

Di-nucleotides	Observed frequencies	Expected frequencies	O/E ratio	p-value
TG	589921	530075.83	1.112	0.0
GT	477871	530075.83	0.901	
CA	602439	518918.86	1.160	0.0
AC	423429	518918.86	0.815	
CG	470420	765075.09	0.614868	0.0
GC	697485	765075.09	0.911656	

Similarly O/E ratio of TG + CA was calculated and compared with rest of the 14 di-nucleotides and chi-test was performed with p-value observed 0 as shown in the table below

Table 9: Comparison of frequencies of TG + CA with rest 14 di-nucleotides

Di-nucleotides	Observed frequencies	Expected frequencies	O/E ratio	p-value
TG+CA	1192360	1048995	1.14	0.0
Rest 14	7502271	4423557		

O/E ratio of CG di-nucleotides was also calculated and compared with rest of the 15 di-nucleotides and after performing the chi-square test p-value was observed approaching 0 (Table 10)

Table 10: Comparison of frequencies of CG with rest 15 di-nucleotides

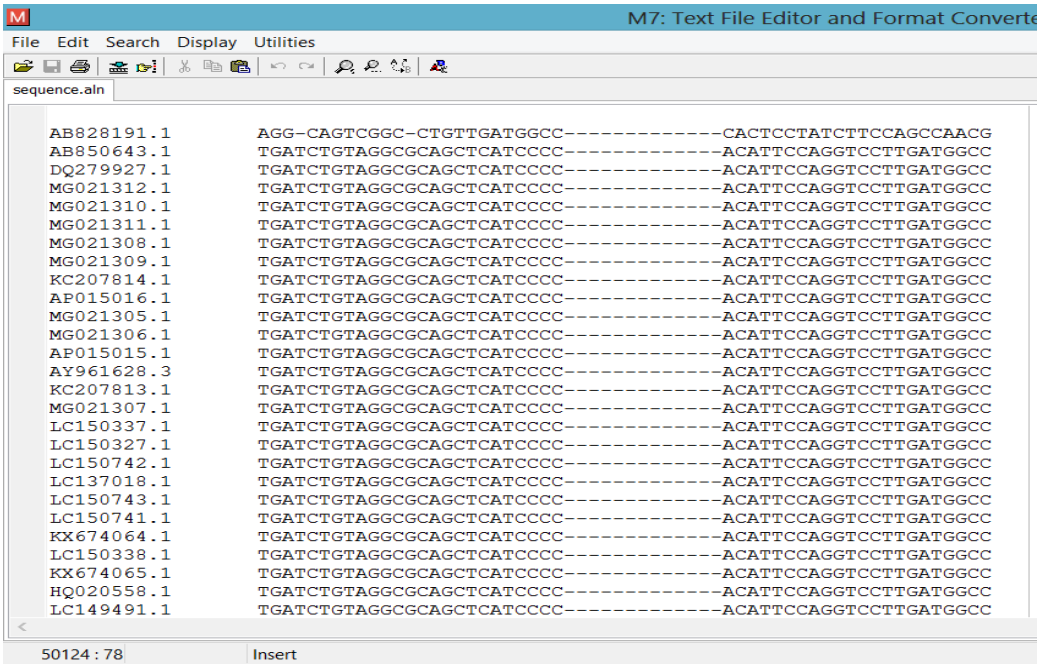
Di-nucleotides	Observed frequencies	Expected frequencies	O/E ratio	p-value
CG	470420	765075.1	0.61	0.0
Rest 15	8224211	4707477		

Based on the above results it can be inferred that TG + CA have shown the overall increase of ~1.14 fold in their count whereas CGs are decreasing by ~0.61 fold in overall genome of Herpesvirus 4. So it can be determined that increase in number of TG + CA and decrease in number of CG are due to mutation of CGs into TG/CA.

6.2 Multiple Sequence Alignment

The analysis that was done before was based on calculating and comparing the frequencies of di-nucleotides in all 51 genomic sequences of Herpesvirus 4. From this analysis it was inferred that there was a certain amount of increase in TGs and CAs count whereas CGs were found to be in suppression. But this analysis could not determine if the resulting increase in TG/CA from CGs were because of mutations or not.

In order to prove that direction of mutation is from CG/CG to TG/CA and the occurrence of such mutations are significantly higher than any other mutation. The genome sequences were subjected to multiple sequence alignment. Multiple Sequence Alignment method was used for studying the CG methylation in Herpesvirus 4 genome. Multiple sequence alignment of all 51 DNA sequences was performed using CLUSTAL 2.1 alignment tool shown in the figure below.



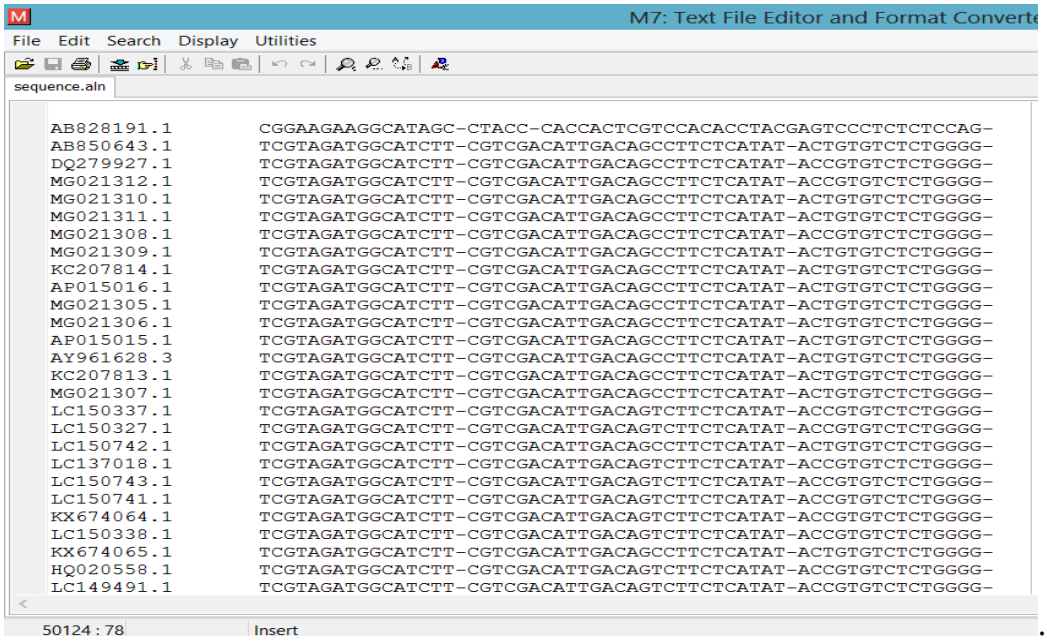


Figure 8: Multiple sequence alignment of 51 sequence of Herpesvirus 4 genome

The alignment positions which had >50% CGs (in 26 or more sequences out of 51 sequences) and had significantly higher TGs and CAs than other dincleotides were selected and considered to be having CG as ancestral alleles at those positions in the genome. A total of 2143 such positions were found in the multiple sequence alignment. These positions were further analyzed to study relative abundance of mutations leading to TG and CA or rest of the 13 dinucleotides. For all such positions, total frequencies of TG + CA and rest of the 13 dinucleotides were determined. The observed frequencies were compared to expected frequencies and chi-square test of goodness of fit was performed and a p-value of ~0.00 was obtained. The relative frequency of TG + CA was found to be 6.887 fold higher than rest of the 13 dinucleotides at these CG positions.

Table 11: Comparison of frequencies of TG+CA with rest 13 di-nucleotides

	TG+CA di-nucleotides	**Rest 13 di-nucleotides**	**P- value**
Observed Frequency	11476	10675	0.0
Expected Frequency	2953.46	19197.53	
Ratio	6.89772		

Based on above analysis it can be implied that TGs+CAs have increased by ~6.89 fold and since it's already been determined that TGs+CAs are formed by methylation in CG so this confirms that over-representation of TGs+CAs occur mainly due to methylation of CG sites.

This increase of ~6.89 fold frequency of TG + CA could also be because of biased base composition within the genome of Herpesvirus 4. So to eliminate this factor, an analysis of TG + CA dinucleotide frequency was done in comparison to the rest while considering base composition (Table 12).

Table 12: Comparison of frequencies of TG+CA with rest 13 di-nucleotides TG+CA with rest 13 dinucleotides at CG max positions taking base composition into account.

	TG+CA di-nucleotides	**Rest 13 di-nucleotides**	**P- value**
Observed Frequency	11476	10675	0.0
Expected Frequency	2953.46	18942.42	
Ratio	6.887		

A fold value of ~6.887 was computed on comparing TG + CA frequencies with rest other dinucleotides while taking base composition into account. It has been observed that even after considering base composition, marginal difference of 0.01 was observed in the fold ratio of TG + CA frequency at CG max positions.

The total number of MSA positions having CG occurring in >25 sequence is 7933 out of which 2143 are such positions where TG + CA are occurring in higher proportion in comparison with rest of the 13. When these numbers of positions were compared against expected number of positions (including the bias of base composition), a 2.37 fold higher numbers of CG position were detected which had TG + CA substitutions significantly greater than other 13 dinucleotides (Table 13).

Table 13: Comparison of frequencies of TG+CA with rest 13 di-nucleotides

	TG+CA di-nucleotides	**Rest 13 di-nucleotides**	**P- value**
Observed Frequency	2143	5790	0.0
Expected Frequency	1057.7333	6776.7395	
Ratio	2.371		

CHAPTER 7

DISCUSSION

DNA viruses, whose host is a vertebrate, have been reported to show under-representation of CGs in their genomes. DNA viruses that infect vertebrates have shown to undergo epigenetic modification at 5- cytosine position resulting in CG di-nucleotide methylation. Deamination of methylated cytosine causes loss of CGs in viral genomes as well as in their hosts. If only one methylated CG undergoes spontaneous deamination, it may cause mutation resulting in loss of 2 CGs and a corresponding gain of one TG and CA each. This is the main reason of suppression of CGs and over-representation of TGs and CAs.

Viruses with double stranded DNA genome of relatively small size show under-representation of CG whereas large genome size viruses exhibit normal range of CG dinucleotide. But the Gammaherpesviruses family, which has a large genome size, is an exception as it shows higher count of TG and CA and low frequency of CGs. Viruses that are from same family and have similar kind of genomic organization, the abundance of CGs are totally depended on their infected host vertebrates (Upadhyay *et al.,* 2014).

Here, in this study, we have performed *in silico* studies on Human Herpesvirus 4 genome which has a ds-DNA and size ~ 172 kbp. This virus causes various infections and disease in humans. Genome size of this virus is much larger than other viruses that were used in the earlier similar experiments. Complete isolate genome sequences were selected for studying the effect of methylation on the Herpesvirus 4 genome. Multiple Sequence Alignment is used to compare and analyze the genomic sequences of viral isolate. This approach can be used to find out ancestral allele and the mutations which may have occurred over the course of evolution.

The present work on Herpesvirus 4 genome is based on the hypothesis that this virus has ds-DNA genomes and infects humans and therefore, can undergo methylation. This will ultimately lead to conversion of CG to TG/CA. Total 51 complete genomic isolates of human Herpesvirus are selected for this study.

The analysis that we have done has shown decrease in count of CG compared to expected CG and GpC. Similarly a corresponding increase in the count of TG and CA observed when compared to the expected frequencies. These genomes have shown significant under-representation of CG. This observation has been made by analysis on all the 51 genomes studied together as well as when individual genomes were analyzed separately. Further position specific analysis of CG, TG, CA and other dinucleotide frequencies were performed with the help of Multiple Sequence Alignment. Multiple Sequence Alignment enabled us to focus analysis only on the conserved CG sites where mutations have converted them into TG or CA (possibly due to methylation) or other dinucleotides (most likely due to reasons other than methylation).

AT CG positions (where CG occurs in at least 26 out of the 51 genome sequences) TG + CA di-nucleotide frequencies were found to be 6.887 fold proportionately higher than those of other 13 dinucleotides. Further number of CG positions where TG + CA were proportionately higher than rest of the 13 dinucleotides and the number of positions where they were proportionately lesser were compare with the expected numbers of such positions. This analysis showed a 2.37 fold higher proportion of CGs where TG + CA relatively more abundant than the other 13 dinucleotides. This analysis is a strong evidence that CG methylation is a strong force leading to loss of CGs in Human Herpesvirus 4 due to DNA methylation as there is a corresponding over-representation TG + CA.

So from the above result it can be inferred that in Herpesvirus 4 genome, CGs are in suppression whereas TG and CAs are over-represented because of the occurrence of DNA methylation that converts CG di-nucleotide to TG or CA during the course of evolution.

CHAPTER 8
REFERENCES

- Adrian Bird (2002). DNA methylation patterns and epigenetic memory. *Genes Dev*, 16(1): 6–21.
- Amon, Wolfgang; Farrell (2004). "Reactivation of Epstein–Barr virus from latency". *Reviews in Medical Virology*. **15** (3): 149–56. PMID 15546128.
- 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*, 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*, 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.
- 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.
- Feinberg, A. P., & Tycko, B. (2004). The history of cancer epigenetics. *Nature Reviews Cancer*, *4*(2), 143-153.
- 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*.
- Goldberg, A. D., Allis, C. D., & Bernstein, E. (2007). Epigenetics: a landscape takes shape.

- 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.
- 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.
- 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.
- Odumade, O. A.; Hogquist, Balfour (2011). "Progress and Problems in Understanding and Managing Primary Epstein–Barr Virus Infections". *American Society for Microbiology*.
- 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.
- 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, 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.
- 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.

Thesis

ORIGINALITY REPORT

5%	3%	4%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	en.wikipedia.org Internet Source	1%
2	Moore, Lisa D, Thuc Le, and Guoping Fan. "DNA Methylation and Its Basic Function", Neuropsychopharmacology, 2012. Publication	1%
3	Zakar, Tamas. "DNA methylation in development", Embryology - Updates and Highlights on Classic Topics, 2012. Publication	1%
4	www.intechopen.com Internet Source	<1%
5	edoc.ub.uni-muenchen.de Internet Source	<1%
6	Renata Z. Jurkowska. "Silencing of Gene Expression by Targeted DNA Methylation: Concepts and Approaches", Methods in Molecular Biology, 2010 Publication	<1%